

Delving deep into fetal ultrasound video analysis



Arijit Patra
Exeter College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary 2020

Acknowledgements

I would like to express my heartfelt gratitude towards my supervisor Professor Alison Noble for her support and guidance. I would like to thank my mentors Dr. Weilin Huang, Dr. Pierre Chatelain, Dr. Tapabrata Chakraborty, Dr. Yipeng Hu, Dr. Ana Namburete, Dr. Christopher Bridge, Dr. Ruobing Huang, Dr. Harshita Sharma, Dr. Lior Drukker, Dr. Pulkit Sharma and Dr. Vinayak Abrol for being constantly around to support with different aspects of my DPhil work, presentations and more generally, for making the journey a lot more meaningful. In the same vein, I would want to acknowledge my friends, colleagues and well-wishers for their constant optimism about my work and well-being. Lastly, I want to thank The Rhodes Trust for providing me an opportunity of a lifetime in the form of the Rhodes Scholarships, without which this DPhil would have remained a pipe dream.

Abstract

Over the last five decades, ultrasound imaging has revolutionized the quality of prenatal healthcare through the course of pregnancies. There have been significant achievements towards improving the design, reliability and versatility of ultrasound hardware. A key aspect that has been a bottleneck in the widespread adoption of ultrasound imaging is the requirement of a high degree of expertise for both performing an ultrasound scan and inferring meaningful diagnostic information from the same. With recent innovations in medical image analysis, it is high time that we explored the possibilities of automating the diagnostic understanding of fetal ultrasound videos for clinically useful diagnostics. Specifically, ultrasound image and video analysis can potentially benefit from the improvements in pattern recognition and image understanding brought upon by the recent proliferation of machine learning algorithms. This thesis explores facets of automated analysis of fetal ultrasound videos and frames for a breadth of clinical applications including view classification, anatomical structure detection, continually learning algorithms over variable data distributions and utilisation of human expert knowledge into automated ultrasound image analysis tasks. Concretely, the contributions of this thesis are:

1. Temporal encoding algorithms for spatiotemporal aggregation of fetal ultrasound video sequences towards visibility assessments, standard fetal cardiac viewing plane identification and cardiac region localization: We showed that aggregating spatial features over temporal windows in a hierarchical fashion allowed the identification of fetal standard cardiac viewing planes with an overall accuracy of 0.8358 along with localization of the fetal cardiac region. Localization of anatomical structures in the Four Chamber standard view of the fetal heart was also obtained over available video sequences.

2. Ultrasound video interpretation algorithms that enable novel classes or instances to be added over time without significantly compromising prediction abilities on prior learnt representations: We showed that using the proposed incremental learning pipelines controls forgetting in deep neural networks for ultrasound video analysis as compared to standard transfer learning methods, achieving average drop in accuracies of 0.011 between additions of three novel classes over subsequent learning sessions. This is in addition to generalization

to the new classes added with average accuracy on new classes of 0.6420 for the proposed pipeline compared to 0.4508 by naive transfer learning methods. Further, explainable continual learning pipelines for ultrasound image analysis were developed by extending concepts of activation maps to the class incremental learning problem so as to quantify metrics for assessment of multi-class incremental learning in terms of activation map quality.

3. Improving the efficiency of ultrasound image analysis models with sonographer gaze assisted distillation: We combine deep learning based methods with the expertise of sonographers by utilizing information obtained from eye-tracking data as a prior encoding domain expertise. We showed that it is possible to improve performance of reduced or memory-efficient models using sonographer gaze data with the best performing memory-efficient models attaining performance within 5% of conventional models $1000\times$ larger in size.

The ideas explored in the thesis are aimed at contributing to the knowledge of how to design efficient ultrasound video analysis with the reliability of high-fidelity knowledge retention across learning sessions, and to explore the feasibility of improving model performance through the inclusion of human expert knowledge and experience.

Contents

List of Figures	xi
List of Abbreviations	xv
1 Introduction	1
1.1 Motivation	3
1.2 Datasets	8
1.2.1 Fetal echocardiography dataset	9
1.2.2 Dataset of cardiac anatomies	11
1.2.3 PULSE data	14
1.3 Thesis structure	16
1.4 Originality	18
1.5 Publications	19
2 Literature Review	21
2.1 Introduction	22
2.2 Advances in deep learning	24
2.2.1 Introduction	24
2.2.2 Pre-2012 models	25
2.2.3 Models: AlexNet and beyond	26
2.2.4 Transfer Learning	30
2.2.5 Video analysis	30
2.3 Applications in fetal ultrasound	32
2.3.1 Introduction	32
2.3.2 Applications in fetal cardiac ultrasound	34
2.3.3 Detection across video frames	35
2.4 Continual adaptation to novel data	36
2.4.1 Task Adaptation and Forgetting	36
2.4.2 Incremental Learning	37
2.4.3 Dynamic Architectures	38
2.4.4 Metric Learning	39
2.4.5 Related Ideas in Medical imaging	39

2.4.6	Interpretability in incremental learning contexts	40
2.5	Sonographer expertise meets deep learning	41
2.5.1	Understanding Sonographer perception	42
2.5.2	Model Compression	43
2.5.3	Teacher-Student Distillation	44
2.6	Concluding Remarks	45
3	Fetal echocardiography video analysis	47
3.1	Introduction	48
3.2	Proposed Contributions	50
3.3	Top-level video classification and cardiac localization	50
3.3.1	Problem definition	51
3.3.2	Spatial Baseline Models	53
3.3.3	Direct Temporal Encoding	55
3.3.4	Hierarchical Temporal Encoding	55
3.4	Datasets and Implementation	58
3.4.1	Training details	59
3.5	Experiments and Results	59
3.6	Multi-anatomy localization	65
3.6.1	Problem definition	66
3.6.2	Proposed Approach	67
3.6.3	Convolutional Architectures	69
3.6.4	Implementation	71
3.6.5	Results and Discussion	71
3.7	Conclusion	73
4	Continually learning from fetal ultrasound videos	75
4.1	Motivation	77
4.2	Lifelong, Continual or Incremental Learning	79
4.3	Proposed Contributions	81
4.4	Problem definition	83
4.5	Hierarchical Class Incremental Learning	84
4.5.1	Network Architecture Details	87
4.6	Experiments	94
4.6.1	Pre-processing	94
4.6.2	Baselines	96
4.6.3	Early training	97
4.6.4	Incremental training	98
4.6.5	Evaluation protocols	99
4.6.6	Incremental computational and memory overheads	100

4.7	Results and Discussions	101
4.7.1	Incremental Performance	101
4.7.2	Forgetting and retention.	103
4.7.3	Effects of Imbalance.	105
4.8	Conclusion	106
5	Multiple class incremental learning with attention	107
5.1	Motivation	109
5.1.1	Attention Map Quality	110
5.1.2	Attention driven continual curriculum	114
5.2	Model and Objective functions	116
5.2.1	A note on the distillation loss	117
5.3	Data preparation and Training	118
5.3.1	Training	118
5.3.2	Baselines	119
5.4	Results and Discussions	120
5.4.1	Attention map quality	120
5.4.2	Accuracy across stages	122
5.4.3	Limitations	123
5.5	Conclusion	125
6	Expert knowledge for reduced models	129
6.1	Introduction	129
6.2	Contributions.	133
6.3	Theory	133
6.3.1	Knowledge Transfer	135
6.3.2	Learning from Human Knowledge	139
6.4	Data and Training	140
6.4.1	Data	140
6.4.2	Training Details	144
6.5	Experiments, Results and Discussion	145
6.5.1	Performance	145
6.5.2	Computational Complexity	146
6.5.3	Memory size	147
6.5.4	Model efficiency	147
6.6	Concluding remarks	148
7	Conclusion	153
7.1	Contributions	153
7.2	Limitations	158
7.3	Future work	161
7.4	Concluding remarks	165

References

167

List of Figures

1.1	Scanning of the fetal heart	3
1.2	Fetal presentation	4
1.3	Four-chamber illustration of the fetal heart	6
1.4	Views and background frames	9
1.5	Creating the dataset of cardiac anatomies	12
2.1	The worldwide under-five mortality rate in 2015	23
2.2	The original AlexNet architecture, showing the different layers, window sizes and number of feature maps	26
2.3	Architecture Of VGG-16	27
2.4	The VGG series of deep convolutional networks with different depth configurations and convolutional block arrangements	28
2.5	The residual block of the ResNet architecture	29
2.6	Notable architectures validated on ImageNet	29
2.7	The three dimensional convolution kernel	32
2.8	The overview of the SonoNet architectures with different feature channel dimensions at the first layer.	33
2.9	The Temporal HeartNet approach to fetal echocardiography analysis	34
3.1	Three different view planes of the fetal heart.	53
3.2	Baseline CNN architectures for multi-task fetal heart prediction in ultrasound videos.	56
3.3	Direct Temporal Encoding (DTE) models modified from AlexNet. .	56
3.4	Direct Temporal Encoding (DTE) models modified from VGG-16. .	57
3.5	CNN architectures with HTE for multi-task fetal heart prediction in ultrasound videos.	57
3.6	CNN architectures with HTE with workflow.	60
3.7	Examples of predicted results by the Hierarchical Temporal Encoding architecture.	60
3.8	Confusion matrices for classification.	63
3.9	Different cardiac anatomical structures in the 4C view.	67
3.10	Detection and temporal progress pipeline for anatomical motion in fetal ultrasound videos.	69

4.1	The Incremental learning paradigm over a hierarchy of learnable features. Previously trained models need to adapt to new class data without compromising performance on previously seen class data. .	77
4.2	The ISUOG scanning protocols for LVOT, 3V and 4C view assessments and anatomical structures in the three standard viewing planes.	82
4.3	Overall workflow over multiple learning sessions.	85
4.4	Super-model for coarse classification. N1 and N2 modules.	86
4.5	Data preparation for hierarchical class incremental learning	89
4.6	Data distribution for incremental learning scenarios	90
4.7	Flow of control for hierarchical class incremental learning	90
4.8	Accuracy on new classes with values for 'p'	93
4.9	Baseline architecture for benchmarking in hierarchical class incremental learning	94
4.10	Diminishing old task performance for transfer learning baselines compared to near constant performance of Networks 1 and 2.	101
4.11	Ability to adapt to new information. Final accuracy over new classes is a proxy for New Task Adaptation. Network 2 generalizes best overall.	102
4.12	Catastrophic forgetting quantified by Overall Drop in class accuracies for old classes. Network 1 and 2 outperform baselines across classes.	103
4.13	Summary of model evaluation - Average overall drop in performance on old classes, and new task accuracies over new classes. Network 1 shows slightly less forgetting but Network 2 performs better on new tasks.	104
5.1	Schematic of saliency curriculum based exemplar retention - After training for a session, GradCAM modules compute attention maps from instances for OSQ calculations and selecting retention examples. Figures for representation purposes only	113
5.2	Map quality (OSQ) evolution over multiple task sessions for individual classes under the ARDS protocol. The three classes encountered as the new classes in Task 3 are not shown since the map quality for their validation sets is not evaluated beyond the first time they are trained for and hence forgetting is not a relevant issue for these classes.	126
5.3	Map quality (OSQ) evolution over multiple task sessions for individual classes under the DEDS protocol. The three classes encountered as the new classes in Task 3 are not shown since the map quality for their validation sets is not evaluated beyond the first time they are trained for and hence forgetting is not a relevant issue for these classes.	126
6.1	Schematic of our proposed knowledge-distillation pipeline	137

6.2	Teacher and Student models used. (A) Teachers use concatenation or element-wise production to merge information from US image and visual attention map. (B) student only takes US image as input . .	140
6.3	Data acquisition with gaze	141
6.4	Gaze maps derived from eye-tracking data superimposed on exemplar frames of the abdomen, head and femur scan.	141
6.5	Across student configurations (based on the teacher-student combination), validation accuracy upon the ultrasound frames for different states of gaze maps	142
6.6	Performance-size trade-off for student models	143
6.7	Architecture of student model S	150
6.8	Architecture of student model M	151
6.9	Student model MV architecture	151
6.10	Schematic for teacher configuration with gaze encoding fusion with extracted features	152
6.11	Schematic for teacher models with element-wise gaze multiplication	152

List of Abbreviations

1-D, 2-D		One- or two-dimensional, referring in this thesis to spatial dimensions in an image.
US		Ultrasound.
SIFT		Scale Invariant Feature Transform.
SVM		Support Vector Machine.
FV		Fisher Vector.
ISUOG		International Society of Ultrasound in Obstetrics and Gynecology Education Committee.
AC		Abdominal Circumference.
CHD		Congenital Heart Disease.
RV		Right Ventricle
LV		Left Ventricle.
RA		Right Atrium
LA		Left Atrium.
SVC		Superior Vena Cava.
FO		Foramen Ovale
PA		Pulmonary Artery.
PV		Pulmonary Vein.
LVOT		Left ventricular outflow Tract.
RVOT		Right ventricular outflow Tract.
4C		Four Chamber View.
3V		Three Vessel View.
CRL		Crown-rump Length.
BPD		Biparietal Diameter.
HC		Head Circumference.
FCN		Fully Convolutional Network.

ReLU	Rectified linear units.
PReLU	Parametric ReLU.
GMP	Global Max Pooling.
GAP	Global Average Pooling.
FC	Fully Connected.
KLD	Kullback-Leibler divergence.
RNN	Recurrent Neural Network.
LSTM	Long-Short Term Memory.
DTE	Direct Temporal Encoding.
HTE	Hierarchical Temporal Encoding.
CAM	Class Activation Mapping.
IOU	Intersection Over Union.
Adam	Adaptive Moment Estimation.
FLP	Femur Length Plane.
ACP	Abdominal Circumference Plane.
HCP	Head Circumference Plane.

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit...

There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain...

— Cicero's *de Finibus Bonorum et Malorum*

1

Introduction

Contents

1.1	Motivation	3
1.2	Datasets	8
1.2.1	Fetal echocardiography dataset	9
1.2.2	Dataset of cardiac anatomies	11
1.2.3	PULSE data	14
1.3	Thesis structure	16
1.4	Originality	18
1.5	Publications	19

The physiological development of the fetus through the gestation period is one of the most exhilarating events in the natural world. The process of maturation of organs and other physiological structures from an embryo is fundamental to mammalian life. In the process of this development during the gestation period, maternal and fetal physiological well-being is of interest. For most of human history, such monitoring of physiological health could not be carried out by direct observation of the growing fetus in-vivo. This changed over the last few decades with the advent of diagnostic ultrasound technology. Initial developments of ultrasound imaging can be traced back to inspirations from the Sound Detection and Ranging (SONAR) equipment used in naval warfare in the early 1940s. With the development of modern B-mode ultrasound after the introduction of sector scanning probes in

the 1970s, using ultrasound imaging for visualization and diagnostic assessment of the developing fetus has gained prominence.

Presently, most healthcare systems in the developed world include ultrasound imaging as a standard part of fetal screening requirements at different points of pregnancies [1]. The specific importance of ultrasound to the clinical problem of fetal health analysis stems from its non-ionizing nature, and non-interference with the health and physical well-being of the fetus. As such, fetal health and growth, which can be impacted by several factors like maternal nutrition, substance abuse, medication, or even environmental factors apart from congenital and genetic defects [1], is primarily assessed through imaging technologies reliant on ultrasound.

Ultrasound has multiple advantages over other imaging modalities in obstetrics. The costs incurred, both as capital investment in equipment and operational expenses, is lower than most imaging systems. A key advantage for ultrasound imaging systems stems from the scope for real-time visualization and interpretation as the sonographer scans a particular volume of interest with a hand-held probe [2]. A typical scanning session lasts 20-25 minutes compared to fetal MRI which usually lasts about 45-60 minutes. The relative safety is yet another advantage. Unlike X-ray or CT, ultrasound does not use any radiation which could be a risk to both the mother and the fetus.

It may be noted that MRI also does not use harmful radiation and can be used for fetal screening. However, it is still less preferable to ultrasound due to the potential magnetic field interaction effects with fetal tissue and the need for relatively complex machinery which result in higher cost in running and cause a slower scanning time. In MRI, the image has to be taken when the fetus is not moving for a while, but the fetus is more likely to move during the scanning. In addition, MRI has a heating effect on the fetus because of the magnetic field strength [3]. Additionally, the bulky machinery involved with MRI acquisition allows much lesser flexibility during scanning and can potentially be claustrophobic for patients involved. Thus, the relative safety and flexibility of ultrasound makes it the preferred modality for prenatal imaging, which accounts for roughly 40%

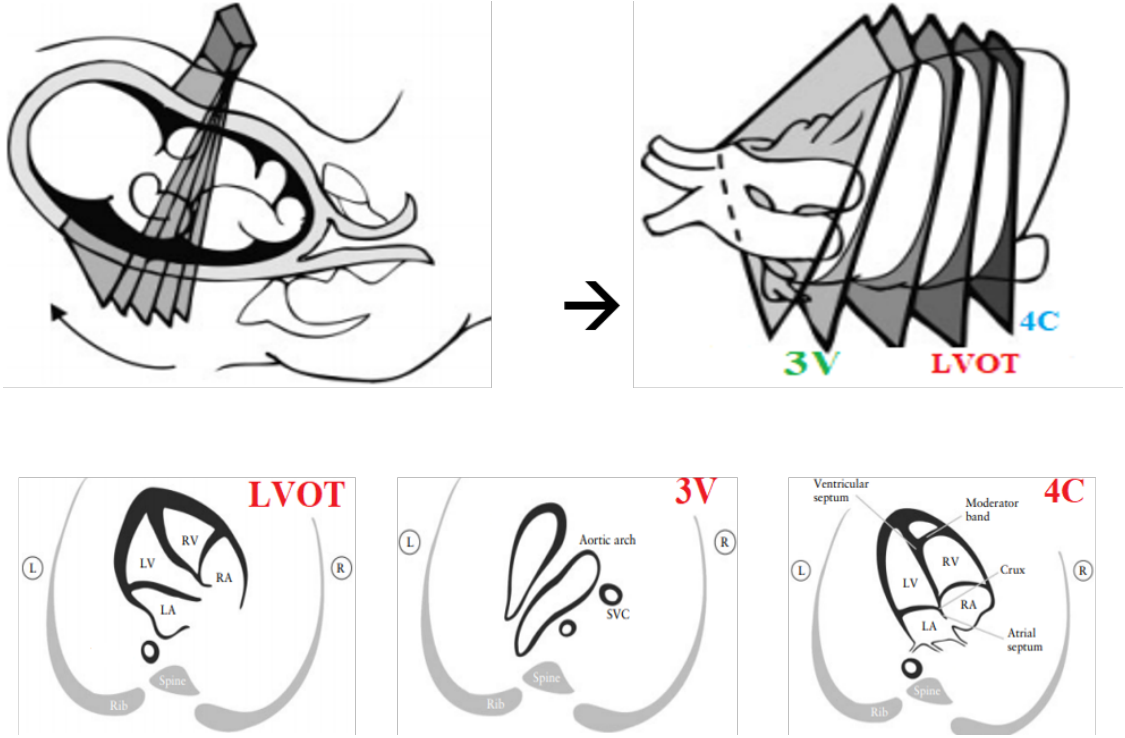


Figure 1.1: Locating fetal cardiac standard planes during the sweep protocol, with different angles of insonification resulting in different cross-sections of the fetal heart identified as fetal cardiac standard planes being visible (top). A schematic of the three fetal cardiac standard planes with the different cardiac structures used to characterize them (bottom)

of the total usage of ultrasound in medicine [4]. The applications of ultrasound in prenatal care ranges from fetal biometry tasks to detecting gestational ages and diagnosis of various fetal abnormalities [5].

1.1 Motivation

The gestation period for humans is typically about 40 weeks and this duration is divided into trimesters. Second trimester or mid-trimester ultrasound screening, usually performed between 18 and 22 weeks [6], is considered the optimal time window to obtain required views (good acoustic window) and to assess the structural and functional health of the developing fetus with a specific focus on diagnosing potential congenital abnormalities [7].

Congenital Heart Disease (CHD) is an umbrella term encompassing a set of abnormalities of the heart that are present at birth or when the fetus is developing.

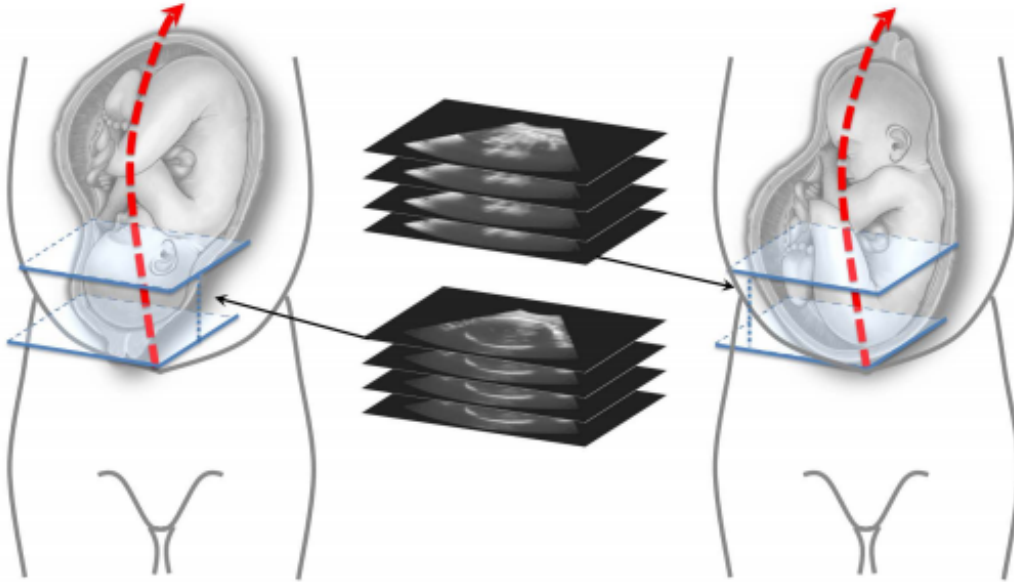


Figure 1.2: The fetus can be oriented in the cephalic and breech presentations with the former being most common (breech presentations account for about 1-3% of all deliveries). Breech presentation is defined as the fetus in a longitudinal presentation with the buttocks or the feet being closest to the cervix, instead of the head as would be the case with a cephalic presentation.

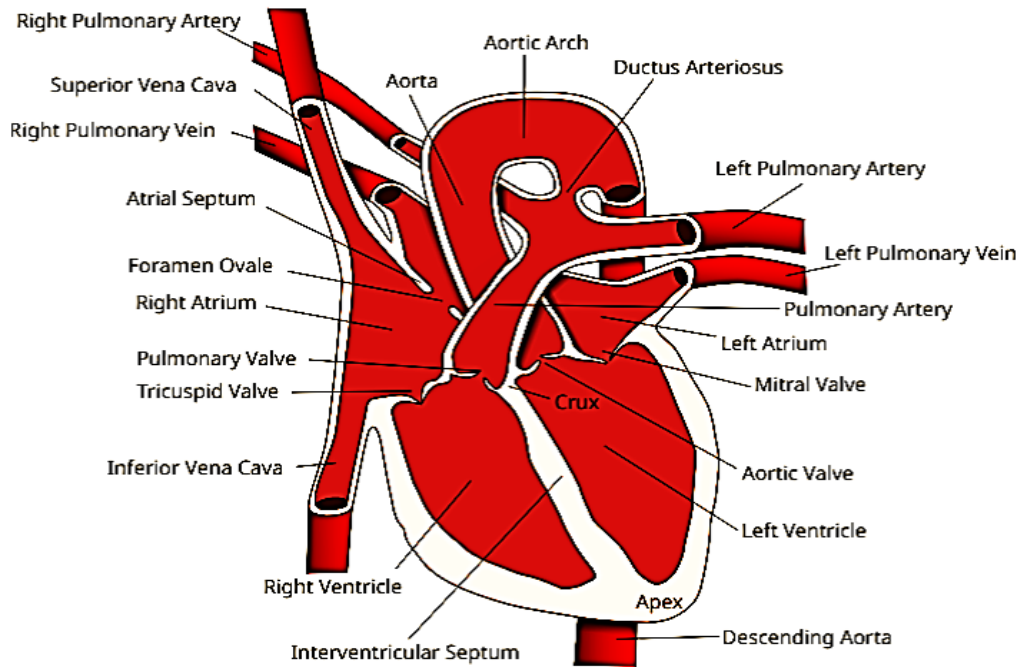
The list of possible conditions includes a variety of defects affecting different parts of the heart's anatomical regions. Multiple defects may co-occur and have complex interactions, along with a possible co-occurrence with other fetal anomalies or malformations in other organs. Worldwide incidence of CHD is estimated at approximately 8 in 1000 live births [8], with the possibility that antenatal populations would have a higher incidence due to miscarriages and terminations of pregnancy in cases with CHD. CHD is one of the most common types of birth defects and also is a significant cause of infant mortality, accounting for an estimated 42% of deaths according to statistics from the World Health Organization [1].

The diagnosis and interventions required for CHD depends on the nature of the defect and the stage it is detected in. Minor aberrations may require little to no intervention and may improve naturally as the infant grows. Some defects require either extensive medication or in some cases, post-birth surgery and life support. Additionally, CHD can lead to complications for the mother, particularly if she

has pre-existing heart conditions during pregnancy. Thus, a suitable detection of CHD during antenatal evaluations is essential for the condition and symptoms to be further studied and for possible interventions to be planned.

While ultrasound screening of the heart has gained prominence as a clinical tool for antenatal diagnosis of CHD and various researchers have concluded that antenatal diagnosis of heart defects improves treatment outcomes [9] [10], detection rates continue to be variable due to different scanning protocols, the inherent difficulties in the ultrasound modality, the nature of the defect [11] and sonographer expertise [12][13] which becomes difficult to build up without a significant experience. This is again a difficult prospect since anomalies are not seen frequently enough [13] [14] and diagnostic markers are observed in discrete sessions over extended time.

Visualisation of the four-chamber (4C) view - an axial view in which the left and right atria and the left and right ventricles are visible – forms the basis of fetal cardiac screening, and the analysis of the presence, size and function of the chambers and atrio-ventricular valves. It is to be mentioned that the four chamber view alone is insufficient to detect several defects like the transposition [15] or coarctation of great vessels [16]. As such, recent guidelines from the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) [6] suggest including the left ventricular outflow tract (LVOT) view and the three vessels (3V) view along with the four chamber (4C) view as shown in Figure 1.1. These different fetal cardiac standard planes or views are identified from different frames representing cross-sections of the fetal heart on the basis of the presence of the characteristic cardiac structures that are observed, with the angle of insonification of the probe and its relative orientation to the fetus determining the cross-section visible on the ultrasound scanner screen at a given point during the sonography scan. The four chamber view is obtained from an axial plane across the fetal thorax. The outflow tracts are typically obtained by angling the transducer toward the fetal head from a four-chamber view when the interventricular septum is aligned in the direction of the ultrasound beam. Examples of the different standard planes are shown in Fig. 1.4



(a) Anatomy of the fetal heart.



(b) Diastole.



(c) Systole.

Figure 1.3: Four-chamber illustration of the fetal heart and the end-systole and end-diastole states.

There are multiple challenges to the analysis of ultrasound videos as compared to natural images/videos. The variation in appearance of different tissues and anatomical structures (except bones, or fluid cavities) is subtle. The resolution is low, particularly for anatomical structures deeper in the body, because using higher frequencies to improve resolution leads to high signal attenuation. There are a number of characteristic features in ultrasound. Speckle is caused due to the interaction of the reflected sound waves with sub-resolution scatterers and manifests as additional texture patterns over and above the tissue structure at the scale of the imaging resolution [17], and is one of the leading causes of ill-defined edges in US images. Acoustic shadowing and enhancement occur when areas of the image lie behind tissue with particularly high or low echogenicity. This results in partial occlusion of regions of interest and intensity variations despite no underlying tissue changes. Along with the issues inherent in the modality, there occurs a significant variation in the anatomy of different subjects, with potentially different orientations and structural differences.

It is to be noted that the fetus does not remain stationary either, making analysis of fetal motion with respect to the probe movements a challenging task. In addition, there is the issue of intra- and inter- observer variability, with some dependence on the specific machine used as well. In fact, the American College of Radiology [18] suggested that over 300 examinations and at least 3 years of experience are required to interpret diagnostic ultrasound scans. Even when a sonographer is highly trained, there may be significant inter- and intra-observer variability in measurement [19]. Thus, there is a significant influence of the clinical expertise of the sonographer in arriving at diagnostically optimal interpretations of ultrasound video sequences.

The complexity associated with the diagnosis of ultrasound videos, and the pronounced scarcity of the corresponding expertise and equipment in several parts of the world, especially in the Low and Middle Income Countries (LMIC) has prompted the medical image analysis community to undertake research on automated algorithms for interpretation of fetal ultrasound. To that end, several applications of computer vision and machine learning techniques have been adapted in this setting.

Recently, deep learning based methods have shown a lot of promise in offsetting some of the challenges faced by traditional image analysis algorithms [20] and are increasingly being explored for different clinical application areas within the fetal ultrasound community. As such, fetal cardiac video analysis is explored through the lens of deep learning approaches with exemplar tasks in fetal echocardiography video analysis implemented using such techniques in this thesis.

As automated image analysis with deep learning has gained prominence within the medical image analysis community, it is pertinent to examine whether such systems can benefit from leveraging the knowledge and intuition of expert human sonographers. This is particularly critical given the aspect of trust and reliability, and on the potential of such incorporation of human knowledge towards improving the diagnostic efficiency of the automated pipelines. To that end, analyzing the scope for inclusion of human knowledge through gaze-tracking offers a novel approach to improve model performance [21] and an insight into the difference in the perception of the given ultrasound sequences between different expertise levels among sonographers. In this thesis, an attempt was also made to utilize such gaze fixation information in improving the performance of low-memory models in order to make efficient deployment on lower spec (and hence cheaper) equipment easier.

1.2 Datasets

For the experiments proposed in this thesis, data comes from two sources - 1) a fetal echocardiography dataset acquired for subjects between 20-35 weeks of pregnancy and 2) the PULSE study which aims to help build machine learning pipelines using a multimodal dataset covering different aspects of the ultrasound video acquisition workflow such as sonographer gaze, text reports, motion tracking data, audio and so on. A third dataset is derived from the first dataset and used for some experiments with anatomically cropped regions from cardiac frames. The data used in the thesis was collected from healthy subjects, and there was no evidence of pathological conditions such as congenital defects in any of the videos obtained. Subject level split was performed to prevent frames from a particular subject from being used

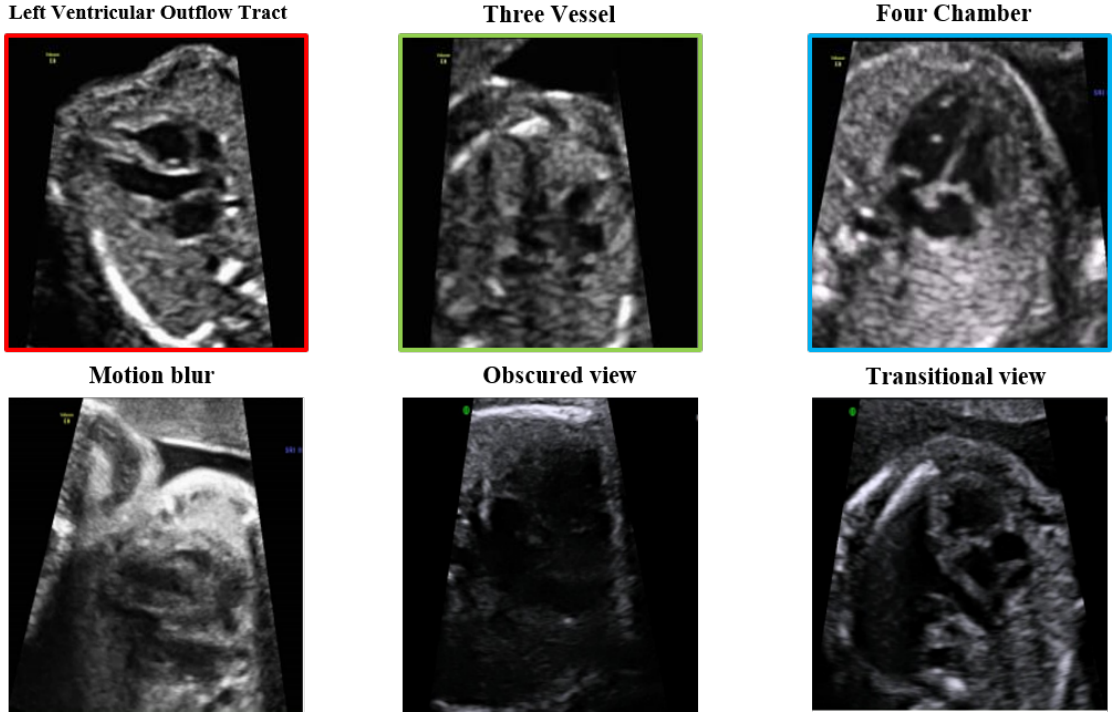


Figure 1.4: Examples of frames from the fetal echocardiography dataset showing the different standard planes and the other frames collectively categorized as background frames.

for both training and test phases. This subject level split was so designed as to approximately consider 20% of subjects for testing, along with cross validation where appropriate. Individual experiments implemented a 4-fold cross validation where the subject-wise split was used to have different sets of patients in the test cohorts. Annotations were performed with the aid of sonographers associated with the Oxford University Hospitals, including the John Radcliffe Hospital at Oxford.

The datasets are described as follows:

1.2.1 Fetal echocardiography dataset

This dataset of fetal echocardiography videos was acquired by Dr Christos Ioannou affiliated to the fetal medicine unit at the John Radcliffe Hospital in Oxford and has been used previously by Bridge et al in their work. This was a challenging task since long freehand sweep videos (beyond 2-3 seconds)(Fig. 1.1) covering multiple fetal cardiac standard planes are not retained as a part of normal clinical workflows. Our dataset was curated out of routine ultrasound scans performed on subjects using

a GE Voluson E8 machine, after securing applicable ethics approvals. The videos captured for the study were captured with the probe’s video capture tool, which records the frames as they appear to the sonographer during scanning time after the original equipment manufacturer’s on-device pre-processing stages and signal processing steps have been performed. Spatial cropping per frame was performed to remove extra information provided by the scanner and retain the ultrasound image alone. The frames so obtained had dimensions of 510 x 430 pixels (± 5 pixels). All the videos so captured in the process formed a part of this dataset without exclusion. The dataset collected in this setting consisted of 91 routine clinical cardiac screening videos from 12 subjects at gestational ages ranging from 20 and 35 weeks.

Each video had a duration between 2 and 10 s and a frame rate between 25 and 76 frames per second. This provided 39556 frames in total, across the different fetal cardiac standard planes and other frames consisting of sporadic occurrences anatomies like the abdomen, frames from the vicinity of the cardiac region not attributed to an anatomy of interest and frames where observations were blurred due to relative motion between the probe and the fetus. All of the frames that were not identified to be belonging to one of the three fetal cardiac standard planes (Four chamber view, three vessel view, left ventricular outflow tract view) are grouped together as ‘background’ frames for our purposes. The videos were annotated initially on a per-frame basis to provide ground truth information for training, validation and testing of proposed models and algorithms under guidance from clinical collaborators who were practicing fetal medicine specialists. The annotations comprised the label on the standard cardiac plane and as a background label in case a frame wasn’t deemed to belong on either of the three standard planes.

Additionally, the frames were individually annotated for each of the other global state variables including the position of the cardiac centre, the cardiac phase in radians, and the approximate value of the radius of the cardiac region. The last point about including the cardiac radius assumes that no magnification was employed while scanning. This is an acceptable assumption since any magnification employed may be tracked by the device software used for the construction of the

ultrasound images and it is straightforward to correctively establish the actual radius values. Assuming healthy growth and development of the fetus, the correlation of cardiac radius with the gestational age is established [22] and while subject-specific deviations may occur, it is expected that the algorithms proposed here would be robust towards minor variations in the same. In some experiments in this thesis, subsequent processing involved creating derivative datasets from subsets of this acquired dataset – primarily to obtain frames for individual cardiac structures. The manual annotations were performed using a randomly chosen subset of the videos from each of the subjects by Dr. Christopher Bridge and by Vaanathi Sundaresan independently over a course of two years from the completion of the first set of annotations. Annotations at the level of the sub-anatomies (cardiac structures in each of the cardiac standard planes) were performed by the author during the curation of these anatomical datasets from the global level dataset described earlier. The videos all consider healthy subjects and consist of scans of the fetal heart in a range of orientations. Following scanning guidelines prescribed by the ISUOG [6], the heart takes up approximately 30% or more of a frame in size.

1.2.2 Dataset of cardiac anatomies

Proceeding from the overall dataset of fetal echocardiography videos obtained from the 12 subjects, a dataset of cardiac sub-anatomies was created subsequently (Fig. 1.5). A sub-anatomy is defined as one of the cardiac anatomies that are relevant to the structure and function of the fetal heart and their selection for the dataset rests on their relevance in establishing diagnostic cues for congenital heart disease (CHD) diagnosis. Following these assumptions, the structures of interest were defined as the Left Ventricle wall (LV wall), the Right Ventricle wall (RV wall), the Tricuspid valve, the Mitral valve and the Foramen Ovale (from the 4C view); the Right Ventricle, the Left ventricle and Aorta (the 3V view); and the pulmonary artery, Descending Aorta, Superior Vena Cava (SVC) (from the LVOT view). The 91 subject level videos from the study described above were split into frames, and only the frames representative of one of the three fetal cardiac standard

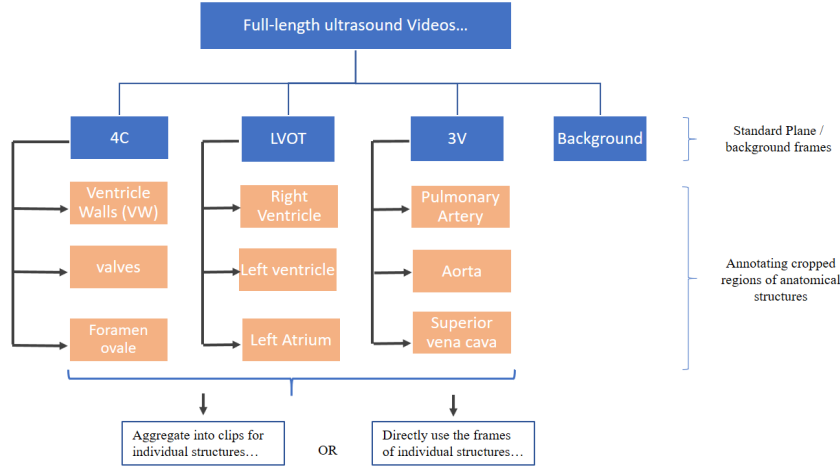


Figure 1.5: Annotating the cropped anatomical regions for creating the dataset of cardiac anatomies. Depending on the experiment, the cropped and annotated anatomical regions may be used directly or combined into clips if obtained from subsequent frames in the original source cardiac ultrasound video sequence.

planes were chosen for further processes. The frames were resized initially into 400 pixels x 400 pixels. The resizing operations were performed using the `cv2.resize()` functions available in the OpenCV package for Python. The resize function offers different interpolation options (`INTER_NEAREST`, `INTER_AREA`, `INTER_LINEAR`). In these experiments, bilinear interpolation (`INTER_LINEAR`) was used. Next, the relevant anatomical structures were manually identified for all the frames that identified as the end-systole, the end-diastole and also for the frame inferred as the midpoint between the end-systole and the end-diastole. While the fetal motion is unpredictable and there may be sudden changes in the fetal orientation owing to the relative motion between the fetus and the ultrasound probe, the annotation strategy is not unreasonable since such motions typically do not occur within a single fetal heartbeat and are not particularly abundant in this dataset.

In each of the frames chosen for manual annotation, the crop regions were identified in resolutions of 100 pixels x 100 pixels to cover the chosen anatomies. The regions to be cropped were identified visually, at the annotation stage, for the frames marked as representing the end-systole and end-diastole. In the intermediate frames between a pair of end-systole and end-diastole pairs, the location for the top-right and bottom-left corners for the bounding boxes were interpolated with

the starting and the ending corresponding points being the ones in the end systole and end-diastole frames. It is understood that the structures of these anatomies vary in their spatial extent and therefore the size of crop proposed may actually be superfluous for some anatomies, notably structures like the Foramen ovale that has a much lesser spatial visibility as compared to the ventricle walls. The size so chosen was to account for the diversity in spatial extent that may occur within the end-systole to the end-diastole, as there may be slight variations in the portion of the anatomies. In some experiments, a 50 x 50 and a 50 x 100 crop was implemented as well for these anatomies, as a 100 x 100 crop was found to be too large in spatial extent. This is described in the relevant sections (Chapter 3). The crops over these anatomies are automatically performed using a custom code written using MATLAB R2017a, based on the identified position of the nearest manually annotated frame in the chronological sequence of the frames obtained after dropping the background frames. The exercise allows us to obtain cropped regions for each of the chosen anatomies for each standard plane. Annotations were stored in separate folders identified by the label of the anatomy, the standard plane that it was derived from and whether the source frame was an end-systole/end-diastole or intermediate frame along with other metadata identifying the original subject labels that the videos were sourced from. After such a set of datasets has been created, subsequent processing is specific to the experiments proposed and are described in the chapters concerned. A common form of such processing involves a resizing of the anatomical crops and a reconstruction of video clips specific to anatomies in sets of a pre-defined number of frames (typically chosen as eight frames per clip). This clip so created would then be specific to a given anatomy within the cardiac region without transferring to another clip or artefact from some other standard viewing plane. This set of 8-frame clips so created for different anatomies is used for experiments in the Chapters 3 (in part), 4 and 5. Such a set of anatomical clips is created for proof-of-concept demonstrations of the algorithms or the technical pipelines proposed and are not necessarily in line with exact requirements of clinical practice, as such requirements for looking at specific regions of the heart may vary

from subject to subject particularly in the event of suspected anomalies, and also depends on the context of other physiological markers observed during the course of the scan. The anatomies so proposed for usage in these experiments are to provide a reasonable broad subset of fetal cardiac structures that may be useful for diagnosing a range of conditions or their absence, but given the limitations of the magnification in the data available and the indistinct appearance of several structures, only those structures in the heart region that are relatively clearly visible are considered.

The curation and annotation is largely semi-automatic as the manual annotation beyond the initial expert annotations is limited to identifying bounding box coordinates for the aub-anatomies at the end-systole and end-diastole frames. The bounding boxes for the intermediate frames are defined by their top-right and bottom-left coordinates which are estimated by interpolating linearly from the end-systole to end-diastole values. This linear interpolation can induce a slight error as the cardiac cycle itself is non-linear, but given that the bounding boxes account for a region in the vicinity of particular anatomical structures, the influence of such an offset caused due to the interpolation was not found to cause undesirable omissions or overlaps.

1.2.3 PULSE data

This dataset is curated as a result of studies conducted as part of a project entitled Perception Ultrasound by Learning Sonographic Experience (PULSE). The PULSE study protocol records full-length routine ultrasound scan video; probe movement; and the point-of-gaze of the sonographer on the monitor of the ultrasound scanner. All ultrasound scans included in this study were performed using a commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machines equipped with standard curvilinear (C2-9-D, C1-5-D), and 3D/4D (RAB6-D) probes. Synchronized eye tracking was undertaken using an eye tracker (Tobii Eye-tracking Eye Tracker 4C, Danderyd, Sweden) attached to the ultrasound machine; the validity of gaze-tracking was previously validated [23]. Probe movement was

recorded using an inertial measurement unit (Next Generation IMU (NGIMU), x-io Technologies Limited, Bristol, United Kingdom) rigidly attached to each probe.

This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051), and written informed consent was given by all participating pregnant women. Sonographers also consented to participate in the study at the outset, but do not have any visual or other signal to know that tracking devices are functioning. The PULSE project is funded by the European Research Council (grant ERC-ADG-2015 694581).

In the thesis, the specific subset of the PULSE data used include the ultrasound imaging frames relevant for the anatomies of the fetal head, fetal abdomen and fetal femur as obtained during the search the optimal fetal biometry planes for these anatomies, and the corresponding eye-tracking information which were used for generation of frame-level gaze maps during post-processing. This pre-processing is described in detail in chapter 6.

A four-fold cross validation approach is used with subject-level splits. In Chapter 3, the four fold cross validation works with a subject level split (10 subjects for training and 2 for testing). In Chapter 4, the initial and incremental training are both performed with the 4-fold cross validation, where the source of the specific anatomical classes is controlled for (the subject that cropped framed is derived from). In experiments where a retention of exemplars is important, such as in Chapter 5, the retained exemplars are primarily meant to reflect the model learning in previous optimization sessions. Thus, the retained exemplars themselves need to be subjected to the subject-level splits, but the new classes being trained for are made available under the same four-fold cross-validation settings as the previous training schedules. In other words, incremental training follows the same subject wise split configurations as the initial training, but with the new anatomical class data being sourced from the concerned subjects. In Chapter 6, all models were trained on 71 subjects and tested on 18 subjects of the dataset corresponding roughly to a 80:20 split over the total number of subjects in the dataset described

in Section 1.2.3, with the four folds traversing over the subject list and each fold having a different composition of the test subjects.

Overall, the objective is to automate different aspects of ultrasound image and video analysis for mid-trimester scans. As such, the primary objective of the video analysis experiments in Chapter 3 is to improve model performance through different methods of aggregating the spatial dynamics over time. Here, the question of computational efficiency is indirectly expressed as competing approaches can have differences in the model memory footprint, and this would also reflect on inference time, although this analysis was not an immediate objective of the study. In Chapter 4 and 5, the incremental learning properties of the models are investigated. While the primary indicator of performance is expressed in terms of the forgetting (decline in validation accuracy over incremental sessions) of learnt representations, the experiments address the question of memory usage due to storage of exemplars for replay, or a lack thereof. These aspects are touched upon in Sections 4.6.6 and 5.1.2. The specific focus on reduction of computational power and memory requirements in tandem with improvement in accuracy metrics is proposed in a sub-part of the thesis where lightweight deep learning architectures are studied, and their performance improved with the aid of cues from sonographer eye-tracking (Chapter 6). Here the requirement of lightweight deep learning architectures for efficient point of care deployment is introduced, and the improvement in performance metrics is brought about using a multimodal representation learning approach followed by distillation. This is validated through a new metric, NetScore, that captures the influence of accuracy, computational complexity (MFLOPS) and model size. Thus, the experiments in Chapter 6 attempt to directly contextualize the application of deep learning methods to an ultrasound image analysis tasks in terms of required computational power and memory, apart from accuracy.

1.3 Thesis structure

The aim of this research is to explore image analysis and deep learning algorithms for the automated analysis of fetal echocardiography videos that may form key

components of assistive tools for the overall diagnosis of fetal health. The potential tasks include view classification, anatomical structure detection, continually learning algorithms over variable data distributions and utilisation of human expert knowledge into automated ultrasound image analysis tasks. The thesis is structured into chapters related to the different facets of ultrasound video understanding and analysis, with reliable adaptations to novel information over extended time and with the examination of enhancements obtained by using expert knowledge. Concretely, the thesis chapters discuss the following:

Chapter 1 introduces the motivation behind the thesis and lays down the thesis structure. The descriptions of the datasets used for the rest of the chapters are also laid out here.

Chapter 2 deals with a survey of literature relevant to other chapters, and includes brief descriptions of key concepts relied upon in other chapters.

Chapter 3 builds a rationale for automated analysis of clinical fetal ultrasound video sequences and makes a demonstration of deep learning based ultrasound video analysis for fetal echocardiography understanding. The tasks addressed in this manner include the prediction of visibility; identification of standard viewing planes; localisation of the cardiac region by prediction of heart centre coordinates along with cardiac region radius and multi-anatomy localisation across frames for cardiac structure motion estimation.

Chapter 4 analyses the ability of ultrasound video and image analysis algorithms reliant on connectionist architectures to adapt to novel classes or data coming in incrementally over extended time in non-stationary distributions and dataset volumes. This is interpreted as a lifelong or continual learning problem in medical imaging settings, with extensions towards building such lifelong learning pipelines that can handle the paucity of labelled data in incremental regimes.

Chapter 5 builds on the ideas proposed in Chapter 4 on the need for continual learning regimes and integrates these ideas with the need for explainability in medical image analysis. This forms the basis of algorithms proposed in this chapter that can use the concepts of saliency and interpretability as a part the incremental

learning process, offering novel metrics for assessment of catastrophic forgetting in models using saliency map quality indices and also helping reduce such a decline in performance over sequential training sessions.

Chapter 6 appreciates the fact that recent automated medical image analysis methods have attained state-of-the-art performance but have relied on memory and compute intensive deep learning models. Reducing model size without significant loss in performance metrics is crucial for time and memory efficient automated image-based decision-making. Methods for the incorporation of expert sonographer knowledge are explored for improving reduced model performance in this chapter and it is proposed to combine ultrasound video with the information about eye fixations tracked for expert sonographers as they scan to train memory-efficient ultrasound image analysis models.

Chapter 7 summarizes the overarching contributions of the thesis and possible extensions towards future work and translational studies. The conclusion presented in this chapter is a recapitulation of the salient points presented in individual chapters along with their respective benefits and constraints. Aspects requiring further investigation and potential applications for the ideas proposed in the thesis are discussed as well.

1.4 Originality

I certify that the work presented is original and performed by my own efforts, except specific contributions made by collaborations where specified. I am the corresponding author of all publications mentioned, visualisations presented, and chapters written, along with relevant supplementary materials. Generally, my work has benefitted from inputs provided by my supervisor, collaborators and colleagues. In chapter 3, I was mentored by Dr Weilin Huang for the work on temporal encoding approaches towards video understanding, although all experimental contributions are personal. In chapter 4 and 5, all contributions are entirely mine, with the caveat that baseline results from methods proposed in other research papers were obtained with adapted versions of original implementations wherever possible and to the

extent practicable. Both the above chapters utilised the fetal echocardiography datasets curated by Dr. Christopher Bridge during his doctoral years, with some annotations being verified by me. The dataset of cardiac anatomies derived from this dataset and was prepared by me. In chapter 6, the contributions presented were results of a collaborative work under the PULSE study [24]. My contribution was centered around the ideas of using gaze fixation maps as supervisory signals for model compression via teacher student knowledge transfer. The datasets used here were acquired as a part of the PULSE project[24] and annotated by several authors. The gaze maps corresponding to ultrasound frames used in the work were obtained from the eye-tracking data by Yifan Cai and other team members involved with the PULSE project.

1.5 Publications

The following publications have resulted directly through the work described in this thesis:

1. Arijit Patra, Weilin Huang, J. Alison Noble, ‘Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video’, In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support 2017 Sep 14 (pp. 276-284). Springer, Cham. (oral presentation)(Chapter 3)
2. Arijit Patra and J. Alison Noble. "Multi-anatomy localization in fetal echocardiography videos." In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1761-1764. IEEE, 2019. (Chapter 3)
3. Arijit Patra, J Alison Noble, ‘Hierarchical Class Incremental Learning of Anatomical Structures in Echocardiography Videos’, under review at Journal of Biomedical Health Informatics. (Chapter 4)
4. Arijit Patra, J. Alison Noble, ‘Incremental learning of fetal heart anatomies using interpretable saliency maps’, The 23rd Conference on Medical Image Understanding and Analysis, 2019. (oral presentation) (Chapter 5)
5. Arijit Patra, Yifan Cai, Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris Papageorgiou, J Alison Noble, ‘Efficient Ultrasound Image Analysis Models with

Sonographer Gaze Assisted Distillation’, Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019 (Chapter 6)

*Alles Gescheite ist schon gedacht worden.
 Man muss nur versuchen, es noch einmal zu denken.*

*All intelligent thoughts have already been thought;
 what is necessary is only to try to think them again.*

— Johann Wolfgang von Goethe

2

Literature Review

Contents

2.1	Introduction	22
2.2	Advances in deep learning	24
2.2.1	Introduction	24
2.2.2	Pre-2012 models	25
2.2.3	Models: AlexNet and beyond	26
2.2.4	Transfer Learning	30
2.2.5	Video analysis	30
2.3	Applications in fetal ultrasound	32
2.3.1	Introduction	32
2.3.2	Applications in fetal cardiac ultrasound	34
2.3.3	Detection across video frames	35
2.4	Continual adaptation to novel data	36
2.4.1	Task Adaptation and Forgetting	36
2.4.2	Incremental Learning	37
2.4.3	Dynamic Architectures	38
2.4.4	Metric Learning	39
2.4.5	Related Ideas in Medical imaging	39
2.4.6	Interpretability in incremental learning contexts	40
2.5	Sonographer expertise meets deep learning	41
2.5.1	Understanding Sonographer perception	42
2.5.2	Model Compression	43
2.5.3	Teacher-Student Distillation	44
2.6	Concluding Remarks	45

2.1 Introduction

In this chapter, literature relevant to the concepts explored in this thesis is discussed. Medical imaging as a discipline has seen a rapid evolution over the last five decades and the rapid growth has reflected on the pace of research carried out in this field. Ultrasound imaging attained prominence for its suitability in obstetric imaging where its non-ionizing, non-invasive nature makes it the modality of choice, with widespread use gaining momentum in the 1970s. Over the years, there has been a reduction in under-5 mortality largely due to improvements in healthcare access in India, China and large parts of Africa and Latin America, but the reduction in neonatal mortality has not kept pace with the overall figures of under-five mortality. The number of deaths between 28 weeks of gestation and age five declined from 12.5 million in 2000 to 8.4 million in 2015, a decline of 2.6 percent per year [25]. Despite the proliferation of ultrasound imaging technologies, the drop in total neonatal mortality has been slower than that of the total under-5 mortality, and this reflects in a larger proportion of neonatal deaths in the overall figures for under-five mortality [25].

A key driver for such a high proportion of newborn deaths is the inadequate management of complications during pregnancies, and an ineffective monitoring of prenatal growth, apart from inefficient maternal healthcare. A non-trivial percentage of these deaths is reported from resource-poor settings in developing societies where both large scale capital equipment and critical expertise for performing diagnostically robust ultrasound imaging is in short supply. Additionally, there exists a global shortage of suitably trained sonographers, with the most pronounced deficit in the developing world [26]. The specific challenges associated with the ultrasound modality, such as the presence of speckle, enhancements, acoustic shadows and a dependence on specialized expertise motivated the medical image analysis research community towards deploying image understanding and pattern recognition methods towards automated diagnostic interpretation of ultrasound. Over three decades of intensive research has led to a large body of literature pertaining to different aspects of using ultrasound in clinical imaging, and in recent years, there has been

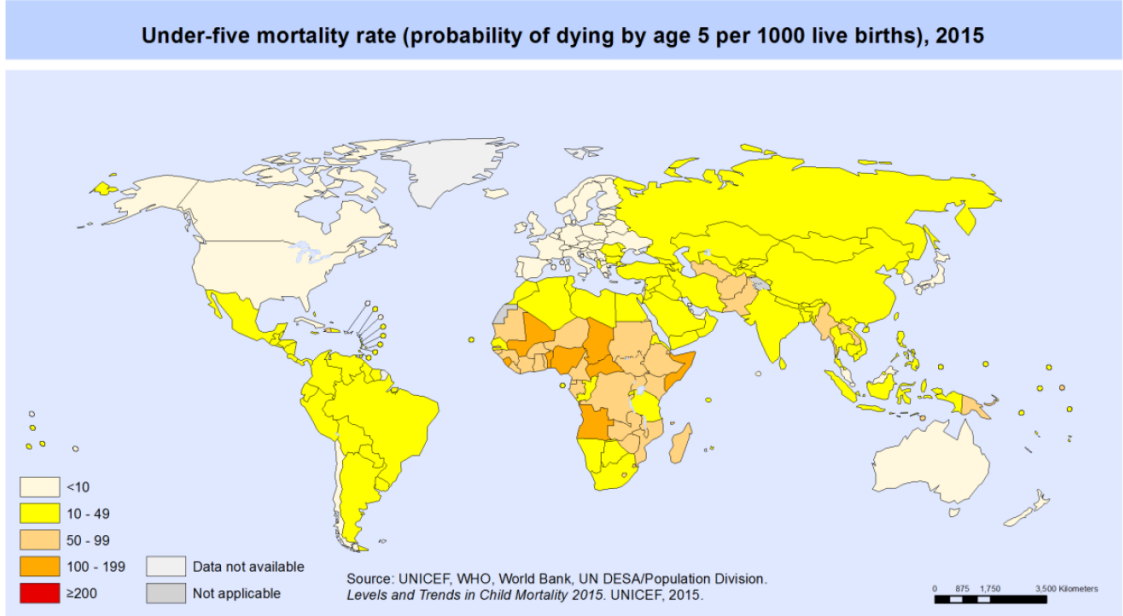


Figure 2.1: The worldwide under-five mortality rate in 2015. Around 42% of the under-five deaths are caused by neonatal mortality, a major cause of which derive from complications stemming from congenital defects. (source: World Health Organization).

a significant focus on combining computer vision and machine learning methods for robust analysis of diagnostic ultrasound images and video sequences. The thesis broadly concerns understanding fetal ultrasound dynamics through the lens of deep learning approaches, applied in clinically relevant situations of video analysis, incremental adaptation, interpretable continual learning and incorporation of sonographer expertise for improving compressed models' performance characteristics through teacher-student knowledge distillation. The literature reviewed in this chapter therefore begins with a discussion on the advances in the field of deep learning technologies and their relevance to computer vision and medical image analysis. This is followed by an introduction to research centered on ultrasound video analysis and interpretation. The next theme explored is the domain of sequential learning over extended time using deep networks in the ultrasound imaging regime along with variants that include interpretability measures such as saliency. Finally, the literature around model compression performance, teacher-student distillation and gaze-tracking in a fetal sonography context is surveyed.

2.2 Advances in deep learning

2.2.1 Introduction

Machine learning (ML) is defined as a set of methods that automatically detect patterns in data, and then utilize the uncovered patterns to predict future data or enable decision making. This is a broad definition encompassing a variety of possible routes of learning and prediction. Typically, two broad domains are supervised and unsupervised methods, with the catch that there is often some overlap between these and newer algorithms that seek to use principles of both sets of methods to obtain desired outcomes. In supervised learning, labelled training data is used to infer a functional representation underlying the dataset. Essentially, in the presence of training examples, optimization algorithms attempt to minimize the difference between the known outcomes and the predicted outcomes of a model given a set of parameters which may be adjusted or ‘learned’ over iterations towards attaining such a minimum. The model with these parameters can then be deployed for unseen examples to predict based on learned features. In contrast, unsupervised learning involves deriving inferences from available data without any labelled training data, thereby essentially learning a probability distribution based on prior data. Reinforcement learning involves a machine interacting with an ‘environment’ and predicting ‘actions’ that affect the ‘environment’ leading to the generation of a reward or punishment, and utilizes a reward maximization (or punishment minimization) approach to optimize the machine. This is analogous to feedback control systems, and to examples in decision theory.

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. Discovery of intricate structures in large data sets is enabled using backpropagation to indicate how a model should change its parameters used to compute the output representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data

such as text and speech[27]. Deep learning algorithms, in particular convolutional networks, have recently become a methodology of choice for analyzing medical images. Therefore, it is pertinent to look into the growth of deep learning as a field, with emphasis on computer vision applications and subsequent adaptations to medical imaging. Rudimentary artificial neural networks were formalized in 1950s, although usage picked up largely after the formalization of backpropagation as a potential training methodology [28][29] in the late 1960s. While the first innovations identifiable today as belonging to the class of deep learning algorithms is widely believed to be the convolutional network for handwriting recognition introduced in [30] followed by a comprehensive treatise on efficiently implementing the backpropagation algorithms for neural networks [30], the lack of computational resources led to the continued usage of traditional algorithms in computer vision, and by extension in medical image analysis. This changed with the advent of powerful parallel processing, GPUs and compatible algorithms.

2.2.2 Pre-2012 models

Unlike algorithms where features are hand crafted, deep learning learns image features implicitly. Several deep learning algorithms have emerged over the years, such as Deep Belief Networks [31][32], Restricted Boltzmann Machines (RBMs) [33], Auto-encoders (AEs) [34] and Convolutional Neural Networks (CNNs). However, CNNs have recently become a popular algorithm to solve computer vision problems. CNNs have achieved state-of-the-art performance in many problems, such as object detection, segmentation and classification [35][36]. CNNs are based on the concept of a multilayer neural network. Original images are input to a CNN and passed through several convolutional and max pooling layers, with possible additions like Local Response Normalization [35] and dropout as a regularization technique to prevent complex co-adaptations of training data by randomly introducing sparsity [37] emerging subsequently. A convolutional layer is simply the image convolution of the previous layer using a pre-assigned filter with learnable weights, and a max pooling layer is the layer that takes the most responsive unit as transition to the

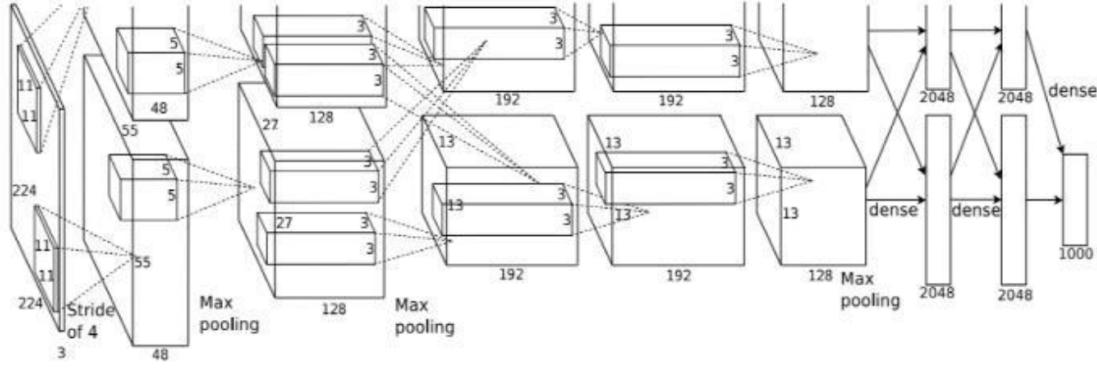


Figure 2.2: The original AlexNet architecture, showing the different layers, window sizes and number of feature maps with the division of processing between GPUs for different parts of the input.

next layer. After convolutional, max pooling and other layers, fully connected layers are attached. A fully connected layer connects all neurons in the previous layer to every single neuron it has. Fully connected layers then can be used to provide the input for a suitable loss function which can be optimized with respect to learnable parameters over multiple training iterations. Since the output from a previous layer is used as input to the next layer, hierarchical features can be learnt by the network with Stochastic Gradient Descent (SGD) [38].

2.2.3 Models: AlexNet and beyond

The watershed moment for the widespread adoption of deep learning was the contribution of Krizhevsky et al. [35] to the ImageNet Large Scale Visual Recognition Challenge [39] in December 2012. The proposed CNN, called AlexNet (Fig. 2.2), won that competition by a large margin. Subsequently, related but deeper architectures were developed in a bid to improve classification accuracies on ImageNet evidenced by the increase in the number of CNN models submitted to ILSVRC 2013. Alexnet was trained on ImageNet data which contained over 15 million annotated images from a total of over 22000 categories, with data augmentation implemented using image translations, horizontal reflections, and patch extractions. The training was found to be slow using the conventional hyperbolic tangent nonlinearity functions and was improved through a Rectified Linear Unit (ReLU). To avoid overfitting, dropout

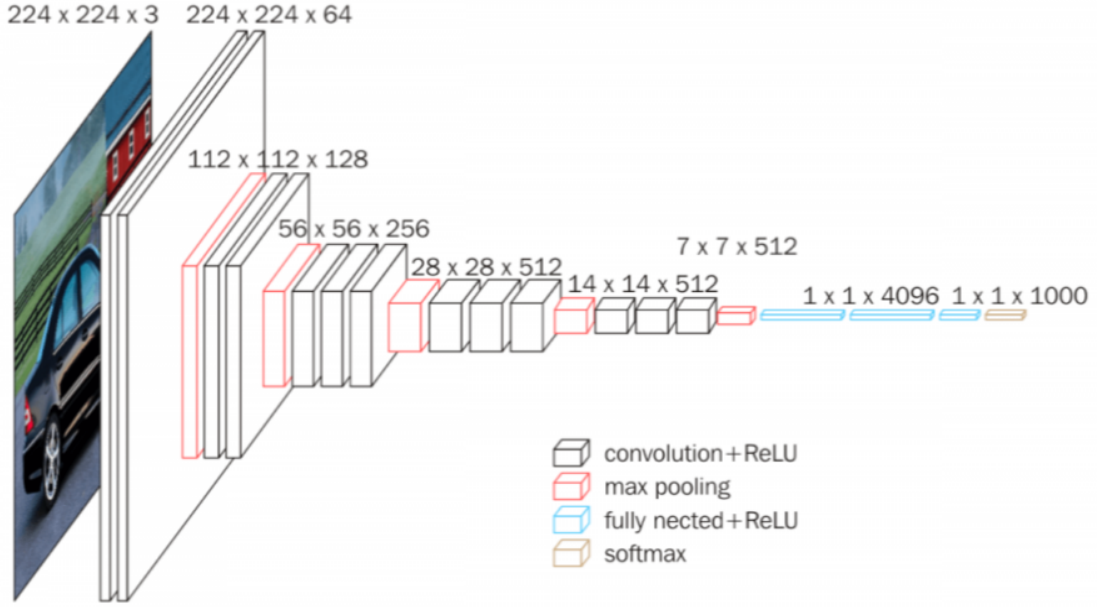


Figure 2.3: Architecture Of VGG-16. Reproduced from [36]

layers, which randomly reinitialized half of the weights from the previous layers were implemented. A feature of AlexNet, which made it amenable to computation using the two GTX 580 GPUs was the parallelized architecture, which essentially allowed the entire network to run in multiple threads and therefore not overwhelm the memory constraints in individual GPUs.

Following the success of AlexNet, Zeilus and Fergus [40] came up with ZFNet in 2013. This model achieved an 11.2% error rate. This architecture was more of a fine tuning to the AlexNet structure, but still developed some keys ideas about improving performance, and also attempted an intuitive explanation of the actual learning methodology for CNNs by introducing the Deconvolutional Network, a visualization which helps to examine different feature activations and their relation to the input space. Another modification was the size of the filters – 7x7 instead of the 11x11 in AlexNet, a trend that would continue in successive developments due to the ability of smaller filters to learn local features better.

Increasing network depths and smaller filter sizes continued to the VGG series of networks [41][42]. The 19-layer CNN with 3x3 convolutional filters all throughout with a stride and pad of 1, 2x2 max-pooling layers with stride 2, widely known as

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv1-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 2.4: The VGG series of deep convolutional networks with different depth configurations and convolutional block arrangements.

VGG-19, achieved a 7.3% error rate in ILSVRC 2014, establishing a new benchmark performance (Fig. 2.3). VGG reinforced the notion that convolutional neural networks have to have a deep network of layers in order for this hierarchical representation of visual data to work (Fig. 2.4). While the primary approach was a linear stacking of different layers, GoogleNet (22 layers, winner of ILSVRC 2014 with a top-5 error rate of 6.7%) [43], departed from the trend with its Inception module, where a parallelized operation was executed at the level of every convolutional step, essentially replacing the sequential convolutional and pooling workflow with a parallel one. A total of 9 Inception modules were used amounting to over 100 layers. The choice of average pooling to reduce dimensionality instead of fully connected layers reduced the requirement of a large number of parameters. Similar reductions through weight sharing in Inception modules meant that the overall number of parameters was 12 times lesser than AlexNet, with multiple crops of an image being used instead of full images to further economise memory requirements.

Following the notion of increasing capacity through depth, He et al. [44] came up with the idea of deep residual learning implemented through an architecture called ResNet, a 152 layer network that set new records for classification, detection and localization all through the same architecture (it won ILSVRC 2015 with an

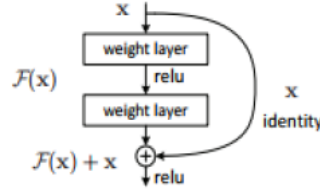


Figure 2.5: The residual block of the ResNet architecture. Input feature map to a layer gets concatenated to the output feature maps to be used as subsequent input.

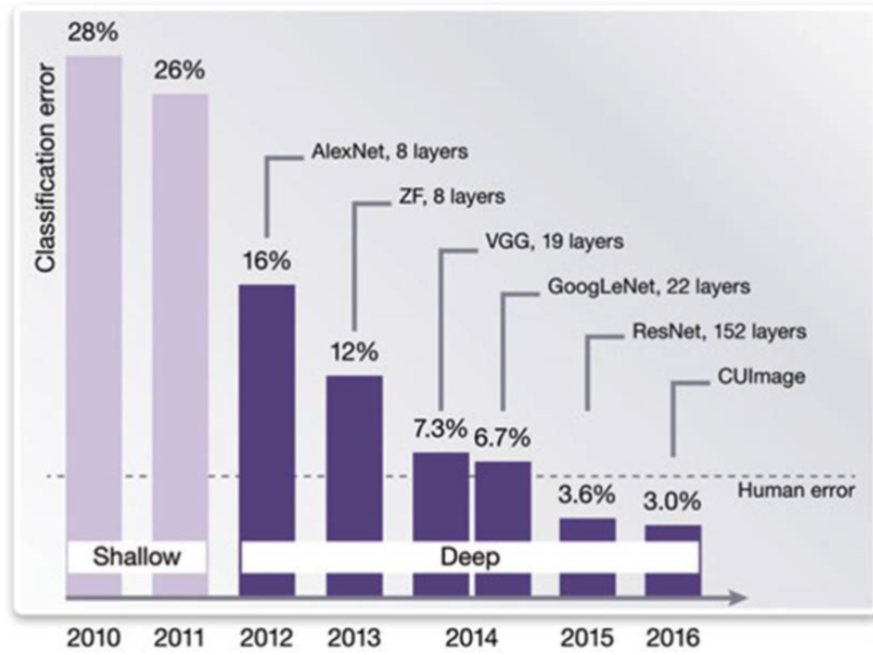


Figure 2.6: A summary of architectures that held state-of-the-art performances on the ImageNet dataset of natural images at different times [46]

error rate of 3.6%, which beats human performance). The ResNet introduced innovations like Residual Blocks where the input to a *conv-relu-conv* series is added to the result obtained from it, in order to obtain the feature maps to be used for further computation down successive layers (Fig. 2.5). This helps propagation of feature information in a more robust manner, besides enabling the gradients to spread out over multiple operations during back-propagation. This architecture has inspired several extensions and modifications, such as the idea of Multilevel Residual Networks being conceived by adding extra shortcut connections to the original ResNets level by level [45].

2.2.4 Transfer Learning

Apart from the steady growth in the complexity of networks in terms of depth, connectivity and nature of operations performed through different layers, other innovations have expanded the applicability and enhanced performance in different applications. A technique that has proven pertinent to medical image analysis, with successful adaptations to ultrasound is that of Transfer learning [47][48]. A common bottleneck in many deep learning applications is the availability of data, and particularly so in medical imaging where variability in observations, inadequate annotations, privacy regulations and storage constraints restrict the availability of large amount of data. Transfer learning is the process of using a model pre-trained on a ‘similar’ application with large datasets, and fine-tuning on the custom dataset, by restricting the gradient descent optimization process to adapt weights only for certain layers [49][50]. This tends to work because some of the features used for the overall learning process like edges, curves etc. are learned in lower layers in a variety of datasets for images acquired through different means and are fundamental to the overall learning across other datasets [51].

2.2.5 Video analysis

It should be mentioned that the networks over the years have gone beyond the two-dimensional space, which is important since there exist a non-trivial number of problems where images are three-dimensional volumes or videos with temporal information need to be dealt with. This is particularly true for medical imaging with modalities like MRI, or videos of ultrasound.

With the performance gains seen by the usage of CNNs in single image classification tasks, there were several papers that proposed an extension over multiple frames of images or full-sequence videos [52]. Early attempts towards using convolutional architectures for extracting spatio-temporal features saw strategies for fusing features at different CNN depths across multiple frames with the intuition that such aggregations over subsequent frames effectively aggregates information going forward in time and the time step so covered would be inversely proportional

to the frame rate of the video sequence. Karpathy et al. proposed multiple fusion strategies with an AlexNet feature extractor [53], with variants proposed including early fusion (the input layer is modified to accept a sequence of frames directly as input), slow fusion (progressively widening the video features receptive field by extending the fusion operation over the depth of the networks), and late fusion (which deploy individual convolutional networks as feature extractors for individual frames of the videos). The slow fusion approach was shown to perform better than the early fusion method as the gradual coalescing of video-level information was found to be more effective at temporal feature learning as compared to an outright accumulation of input frames. The late fusion variant was found to be inadequate in its ability to actually encode temporal information.

In line with the idea of fusion of learnt features, Fiechtenhofer et al. [54] proposed a two-stream 2D convolutional architecture for video recognition. The two streams were to learn to extract spatial features from frames and inter-frame motion features from the optical flow frames computed prior to the CNN training stage. The extracted features from the parallel CNNs were then used to independently arrive at classification confidence scores, with the softmax scores finally combined by averaging out. The proposed two-stream architecture has two downsides – a limited temporal resolution caused by a single frame operation for the spatial stream, and a two-frame aggregate information on the optical flow stream (since this stream accepts a single optical flow frame as input, it effectively captures motion information only for two consecutive video frames); and the inability to learn pixel-level associations between spatial and temporal features as the final classification scores are directly fused.

These issues led to the development of the 3D kernel [55] (Fig. 2.7) which directly operates on 3D data and works through with 3D feature maps which enable the end-to-end distillation of spatio-temporal information through trainable networks for tasks like human action recognition [56], real-time object recognition [57], medical applications like brain lesion segmentation in MRI images [58] and liver segmentation in CT scans [59].

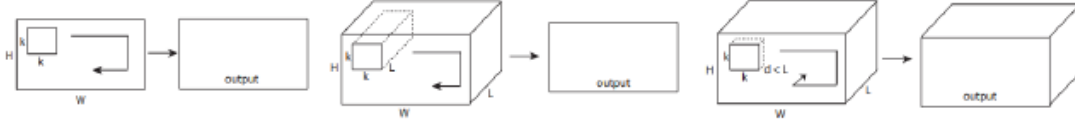


Figure 2.7: (Left to right) 2D convolution, 2D convolution on a video with multiple frames, 3D convolution on a volume generating volumetric outputs.

2.3 Applications in fetal ultrasound

2.3.1 Introduction

While there has been a significant amount of work pertaining to image and video level analysis for classification, detection, segmentation, action recognition, scene understanding and allied topics, both in the context of computer vision and in the subset of medical imaging, relatively few adaptations or innovations in these domains have been made for ultrasound imaging, especially pertaining to more complex domains like cardiology - both adult and fetal. This is partly due to the complexity of the modality of ultrasound that produces images with a significant proportion of artefacts and speckle, and due to the inherent difficulties of creating large and reliable medical datasets with high fidelity annotations. This seems to be changing with a recent spurt in academic and industrial interest in exploring the potential of using deep learning for ultrasound applications.

Automatic methods for ultrasound video analysis have also been developed both within traditional and deep learning systems. Kwitt et. al. [60] applied kernel dynamic texture models for labelling individual frames in an ultrasound video. Maraci et. al. [61] extended this concept to handle multiple object classes in real-world clinical ultrasound sequence. Bridge et al. [62] estimated a number of key parameters related to the fetal heart using handcrafted features developed with classification forests to distinguish different view planes of the fetal heart. Then a multi-variable prediction was formulated using a CRF-filter.

Previously, deep learning for anatomical description in ultrasound has been used for several applications like nerve segmentation [63], early-stage atherosclerosis detection from carotid ultrasound images [64], segmentation of Deep Brain Regions

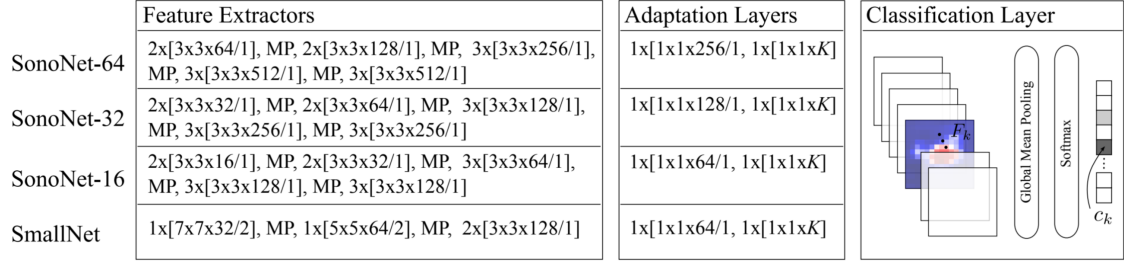


Figure 2.8: The overview of the SonoNet architectures with different feature channel dimensions at the first layer. A key difference from the VGG-16 model is the usage of global average pooling layers in the pre-softmax stage. Figure taken from [81]

using a combination of CNNs applied in patches identified with Hough voting [65] and robust segmentation in ultrasound volumes using Hough voting and 3D CNNs [66], Stacked deep polynomial networks based representation learning for tumor classification from small datasets of ultrasound images [67], anatomical structure detection using iterative multi-domain regularized deep learning [68], supervised segmentation of the left ventricle of adult hearts [69], the segmentation and estimation of carotid arterial diameters in ultrasound images [70], quantification of bull’s eye maps for myocardial strain estimation [71] [72] and learning based automated motion assessments as a determinant of normal functioning [73] [74]. On adult echocardiographic images supervised learning methods [75][76]] and probabilistic models [77] have been used for view classification. Apart from adult cardiac echocardiography, the work on fetal ultrasound includes the automated recognition and analysis of other fetal structures in images [78] [79] [80].

A popular application area where deep learning has been applied to fetal ultrasound pertains to view classification and standard plane identification from frames of short video sequences of 2D ultrasound sweeps. Chen et al. [68] employed a CNN architecture with five convolutional and two fully connected layers for detection of standard abdominal views, later extended to three scan planes with a recurrent architecture accounting for temporal information. Following this, Baumgartner et al. [81] implemented a weakly supervised localization for fetal structures using bounding boxes along with automated real-time detection of thirteen fetal standard views using a CNN architecture (Fig. 2.8). A CNN-based architecture was adopted

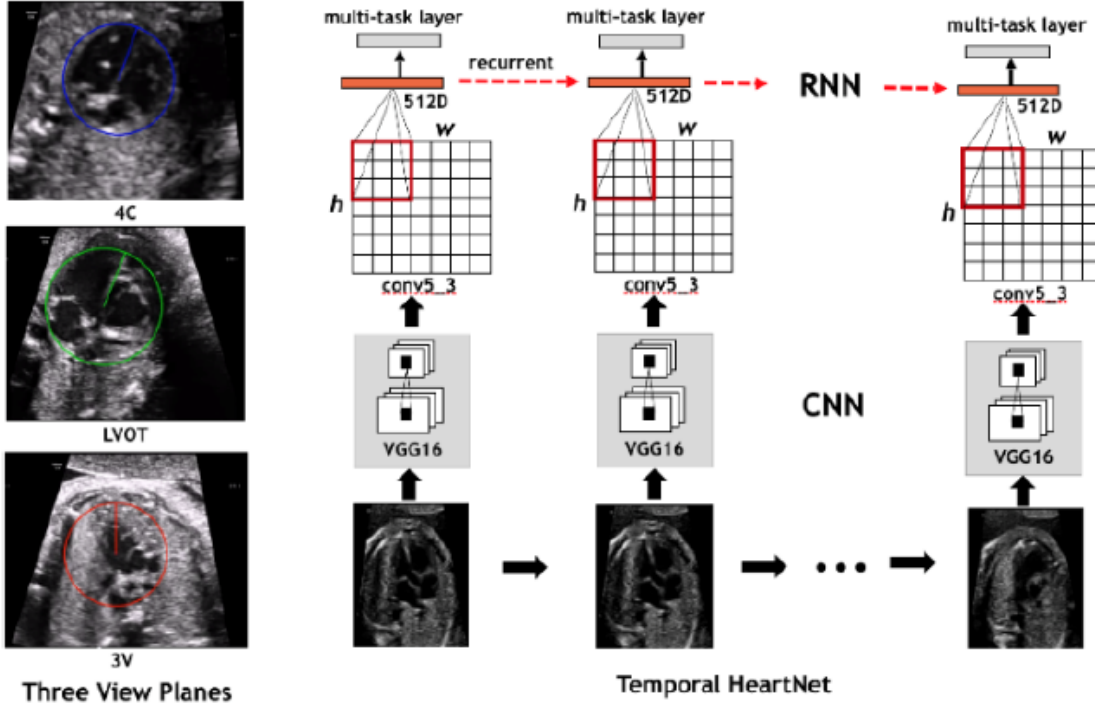


Figure 2.9: The Temporal HeartNet proposes aggregating spatio-temporal information through a RNN working on features extracted using frame-level CNNs for fetal echocardiography

by Yu et al. for fetal facial standard plane detection [82], which was attempted earlier using deep Fisher vectors in [83]. Recently, CNNs have been applied for image-level ultrasound video object classification with transfer learning for anatomical description [84] and recurrent models with knowledge transfer for prediction of fetal standard planes [85].

2.3.2 Applications in fetal cardiac ultrasound

Diagnosis of fetal cardiac ultrasound requires navigation to the standard cardiac viewing planes and observing the overall cardiac motion over a sequence of several frames, typically covering an entire heartbeat or more. Analysis of three standard fetal cardiac viewing planes, along with visibility and orientation was also presented using a segmentation inspired approach in [86] using fully convolutional networks. The Temporal HeartNet [87] proposed a different approach that developed a CNN architecture coupled with a RNN for describing jointly the fetal heart visibility, viewing plane, localization and orientation (Fig. 2.9). The Temporal HeartNet

architecture was the first deep learning based approach to the fetal cardiac view classification and localization tasks, and obtains an error of 21.6% calculated as an IoU metric. This is compared to Bridge et al.’s error rate of 34% on the same dataset and the inter-observer error of 26% in the tasks of viewing plane identification and prediction of cardiac radius.

2.3.3 Detection across video frames

Modelling of spatiotemporal dynamics in vision has been attempted in action recognition domains. While most video classification methods assume a single class per video sequence, action localization has been a challenging problem due to multiple detection, classification and localization tasks in variable spatial and temporal scales. Recently Hou et al. [88] presented a tube proposal network from 3D feature maps leveraging 3D feature map generations from 3D CNNs. Most localization work focused on integrating frame level object detection [89] with methods to build linkages across frames and defining the start and conclusion of disparate actions [90]. Action progress estimation was reported with global context and region proposals coupled with an RNN to regress to action end-points [91] with a single action per class.

In ultrasound imaging [87][92][93] [94] [**patra**] leveraged the spatiotemporal dynamics in videos, by characterizing the visibility, viewing planes and localization of the fetal heart using convolutional temporal fusion. To date, there have been no studies that performed anatomical detection with multiple regions of interest with temporal extensions for fetal ultrasound despite the importance of diagnostic assessment of physiological motion observed by medical imaging systems. Video understanding at the local level has largely been untouched for ultrasound where optical flow is difficult to estimate and generating unique motion tubes per structure is non-trivial. Unlike most problems in computer vision, relying on a presence of specific spatial cues is insufficient to isolate distinct motion tubes or to quantify start or end-points. This is because relevant anatomical structures are spatially present throughout but given the periodicity of anatomical motion, the characteristic

marker for motion completion involves isolating a single periodic cycle that may vary for different anatomical structures within the fetal heart and with overall heart rates for unhealthy instances.

Deep learning and its applications have grown by leaps and bounds in recent years in computer vision. It is evident that many of these methods have a great potential for automated analysis of medical images, and help ameliorate the characteristic issues in various modalities [95], eventually enabling the development of assistive technologies in the clinical setting.

2.4 Continual adaptation to novel data

Human learning can acquire knowledge and learn new examples building upon past learning, and this does not lead to a disappearance of past learning. In machine learning, Lifelong Learning is a computational approach that uses previously learned information to acquire new knowledge, using the idea that learning the n^{th} task should be easier than learning the $(n-1)^{th}$ task. The concept was first proposed by Thrun in a seminal work on continuous learning for robotic path planning [96], in response to critiques about the tendency of connectionist neural networks to demonstrate marked reductions in performance when trained over sequential data distributions without a joint training regime. Under assumptions of such new knowledge being incrementally added as new classes, the lifelong learning problem is essentially a class incremental learning problem.

2.4.1 Task Adaptation and Forgetting

Acquiring knowledge from examples is a continuous process and needs to evolve dynamically with new data. Most literature has considered the learning process as a function of static data (data known at the start). There have been relatively few attempts to place learning in a context where data is dynamically available [97]. Attempts to use generic features by transfer learning [49] rely on such features being common across a range of similar tasks [47]. A bottleneck to incremental learning in connectionist models is the tendency to converge to new tasks with

parameters unsuitable for modelling prior tasks. Goodfellow et al. [98] studied this phenomenon, called ‘catastrophic forgetting’, where features previously learned are lost on re-training for new tasks leading to a diminished performance on old tasks.

Techniques like fine-tuning [99] and feature extraction [100] are used to preserve learned features while adding new parameters to adapt to new tasks. Feature extraction is useful for cases where only a few new layers of parameters are initialized over and above learned layers borrowed from other similar tasks and a learning regime affects only these new layers [101]. In fine-tuning or transfer learning, the addition of new parameters is accomplished in a similar fashion but old parameters are selectively or wholly modified while training for new parameters [49].

A shortcoming of such techniques, apart from losing previous training results, is their requirement of similarly sized training datasets with a poor tolerance for class imbalance while re-training for new tasks. Using existing models for similar tasks pertinent to different modalities of datasets, a situation common in multi-modal medical imaging cases, can be tackled using domain adaptation[99]. Ideas like few-shot learning using Siamese networks [102][103] work with a few samples to improve generalization on trained models for tasks like image retrieval where variations exist among available data and comparative discrimination in input samples is essential.

2.4.2 Incremental Learning

Neural network based approaches towards avoiding catastrophic forgetting [97] have focused on 1) retraining with regularization to avoid drastic changes to learned weights; 2) multi-headed network expansion, either as modules for sub-tasks or fixed size modules for each new task [104], and 3) selective retraining with expansion [105]. Learning without Forgetting (LwF) [106][107] relied on an output level preservation of learned knowledge by using a distillation loss on prior learned logits. An initial dataset was used to train a network and the softmax logits for classes previously trained for were frozen and used as a supervision scheme along with new labeled data. A shortcoming here is the storage of past prediction probabilities during future task adaptations. In Elastic Weight Consolidation [108], parameter importance was

evaluated using a diagonal approximation of the Fisher information matrix, and a quadratic penalty imposed on the difference in parameters across old and new tasks. The key bottleneck of this approach is that the calculation of the Fisher information matrix is computationally intensive, and with the diagonal approximation used to resolve this issue there is an inevitable loss of parameter information. Zenke et al. [109] penalized alterations to relevant parameters on past tasks by a notion of synaptic relevance obtaining results similar to EWC. This synaptic relevance was computed by means of a perturbation to individual parameters and measuring the output space perturbation caused. Individual parameters were then controlled for future task adaptation by attaching a penalty proportional to the perturbation resulting from their modification.

2.4.3 Dynamic Architectures

Among dynamically evolving architectures, Xiao et al. [110] presented a node-branch formulation of an error-driven incremental learning system to include classes based on error thresholds. This strategy was used to address the diversity and size of the label space of the ImageNet dataset [111] and to allow for stage-wise classification of ImageNet classes instead of an end-to-end single classification. Rusu et al. [104] formulated fixed sized networks to be assigned for each new task, thus permanently locking all parameters used in a given task. Following concepts of rehearsal or mingling new class data with samples from past classes, iCaRL [112] was developed to incrementally learn from new classes by including exemplars from previous classes during new class training. Generative models [113] have been used in this paradigm as well, as explored in [114]. Drawing inspiration from memory consolidation in the brain, Kemker et al. [115] used a pseudo-rehearsal scheme with an auto-encoder reconstruction of past examples. A similar approach was used to design a dual generative replay process, based on inspirations from long term memory formation processes in the brain [116]. Approaches with regularization and with selective re-training require at least representative examples of past trained classes, which is a situation we seek to mitigate in line with potential clinical issues of data availability.

2.4.4 Metric Learning

Exploration of feature similarities in low dimensional spaces has been a mainstay in several areas like information retrieval, template matching, and increasingly few-shot learning [102]. In tasks that are inherently data sparse, achieving a hyper-plane separation between clusters of feature similarities becomes challenging and instead, a distance measure in the metric space has been suggested [103]. In incremental learning problems, an implicit step is novelty detection on the presented instance; namely in prior learning cycles has the instance been encountered or not. This is a stage where one can seek to exploit the few-shot nature of metric-based classification for class incremental learning, as a label different from what is already known by the system can be assigned to a novel exemplar.

2.4.5 Related Ideas in Medical imaging

Clinical data is incrementally and continuously acquired in clinical research studies and hospital settings, however incremental learning in medical imaging applications has not received much attention to-date. For instance, Giritharan et al. [117] used incremental learning to diagnose positive and negative cases in endoscopy images by adapting SVM boundaries to new data of diseased instances in subsequent clinical sessions.

Leveraging feature similarities for hierarchical classification of anatomical motion incrementally in modular architectures that are resilient to variable sized datasets remains an open problem. We explore such an incremental task without assuming any presence of past data, a scenario considered in a very limited number of works in the past, notably in FearNet and Progressive Networks. The former has a pseudo-rehearsal step on prior representations to generate exemplars at incremental run-time. The latter assigns different networks to different tasks rather than considering incremental class additions. Such an implementation leads to single networks for each of a multiple 1-class classification tasks, with a redundancy in the parameter space. Thus, primarily transfer learning approaches are the only knowledge extension schemes closest to our task conditions.

Classification of structures can be segregated into models of expert classifiers [118] specializing in different kinds of fine classification, as determined per instance by a coarse category (analogous to referrals from general practitioners to specialists in health practice). Then an incremental data distribution can be adapted to the overall pipeline by restricting fine-tuning for a new addition to the corresponding expert, leaving the integrity of learned representations in other expert models intact. This is different from standard fine-tuning and transfer-learning approaches in that the process of adapting to new data is conditionally segregated using coarse distinctions. This is a trade-off between incrementing fixed network capacity to a new additional task [104] and prior knowledge retention by conditioning access to the parameter space on prior decisions on input instance characteristics than impose a hard bound on model components used to incorporate new task data.

2.4.6 Interpretability in incremental learning contexts

While deep learning based methods have led to appreciable performance gains in several tasks, the issue of interpretability has increasingly received attention. The abstraction of the feature extraction and representation learning in the deep learning process inherently causes the model to be used as a ‘black box’ that takes in inputs and learns the associations over a series of steps not directly amenable to human-level interpretation. Therefore, there exists a need to rely on proxy measures towards ensuring deep model interpretability. A common method for making deep neural networks interpretable is to analyze the intermediate layers’ feature maps [119]. The idea is that if a certain region in the input space is particularly informative towards the final model decision, such regions will be amplified as the learnable parameters associated with those regions have steeper updates in the gradient-based optimization processes for learning input-output correspondences. This is expressed through the concept of saliency maps. A saliency map represents regions on the input image that were most ‘activated’ during the model prediction on it. Several methods have been proposed for obtaining such activation maps, and more generally for explaining the decisions of convolutional networks such as the class activation

mapping (CAM) [119] and subsequent variations like the gradient-modified class activation mapping or Grad-CAM approaches [120].

In the domain of interpretability of medical images, there has been a focus on utilizing attention mechanisms to understand decisions of machine learning models, such as attention mechanisms for interpretation in ultrasound images in fetal standard plane classification [121] and pancreas segmentation [122]. Utilizing the notions of interpretability in a continual setting or for enabling learning in incremental sessions is yet to be studied in literature and we introduce notions of class saliency and explainability for assessing and in influencing continual learning mechanisms.

While there have been efforts to automate the understanding of fetal ultrasound images and videos with machine learning approaches as described earlier, the possibility of catastrophic forgetting or adaptations to new classes over time has remained unaddressed and there are no prior publications on incrementally refining models with previously unseen classes as devised in this chapter.

2.5 Sonographer expertise meets deep learning

A third dimension of this thesis concerns possible methods of enhancing the performance, interpretability and utility of deep learning methods for fetal ultrasound image analysis by leveraging insights from expert sonographers who have made diagnostic interpretations on the same set of video sequences. Specifically, the motivating example chosen pertains to enabling higher performance characteristics for low-parameter, low memory footprint reduced models that can be efficiently deployed on lower spec devices such as portable machines, probes and similar hardware. Thus, the model compression efficiency is sought to be improved by incorporating cues from expert insights as a part of the training process for the pipeline. This is implemented as a teacher-student distillation method, where the gaze cues are additional signals at the teacher input.

With the proliferation of ultrasound in low-resource environments [123], the challenges towards improving maternal and neonatal health are partially addressed

in terms of reducing the barrier of high-cost capital equipment like complex, bulky scanners that were once the mainstay of clinical ultrasound. This disruption in the ultrasound hardware space still leaves open the question of expertise requirements for diagnostically robust ultrasound image and video analysis. As proposed by the overarching theme of this work, the latter is sought to be addressed by means of video understanding pipelines reliant on deep neural networks and associated techniques. The usage of deep networks often comes with the issues of heavy memory footprint, large inference times, high power consumption, and requirements of large volumes of data. Deployment on low spec hardware is particularly challenging due to such constraints, motivating the case of looking into model compression and knowledge transfer.

2.5.1 Understanding Sonographer perception

Literature on integrating actual sonographer experience with computational medical image analysis is sparse. In an early work pertaining to the usage of sonographer eye-tracking data for automated analysis, Ahmed et al. [124] proposed an eye-tracking based method for the identification of standardized image planes from 3D fetal abdominal ultrasound volumes and analyzing the visual localization strategies used by professional sonographers. A related work concerns combining sonographer gaze and ultrasound video for fetal abdominal standard plane classification using a set of fusion based architectures implementing different strategies for inclusion of sonographer gaze fixations [21]. The base architecture termed SonoEyeNet (SEN) was based on a SonoNet-16 [81] feature extractor and was primarily intended to demonstrate that the model classification performance can be improved by including cues related to sonographer perception. They show that with the usage of sonographer visual fixation information in the training process, the precision, recall and F1-score of abdominal circumference plane detection was increased by 22.9%, 24.9% and 24.0% respectively.

A drawback of the proposed strategy is the requirement of gaze maps corresponding to relevant frames during inference, a prospect that would be challenging to most

ultrasound scanning facilities. The authors sought to address this subsequently by proposing a pipeline called Multitask SonoEyeNet for generative gaze map prediction while simultaneously performing standardized abdominal circumference (AC) plane detection tasks [125]. The primary task was defined as the plane classification which was aided by the auxiliary task of saliency map generation. In this case, the inference stages would not require an explicit availability of gaze maps associated with input ultrasound frames as such gaze maps are simultaneously generated from a random distribution. These methods primarily concern improving performances of deep models on exemplar tasks in order to serve as a proof-of-concept for the utilization of expert gaze fixations to better inform the design of medical image analysis pipelines. Droste et al. [126] sought to demonstrate the feasibility of learning representations of ultrasound images without requiring manual annotations, and relying on modeling the sonographer visual attention in terms of their gaze fixations.

2.5.2 Model Compression

Model compression (or reduction) is a challenge in machine learning research due to both the interest in addressing over-parameterization [127] and for practical usage with reasonable computational resources. Model compression can be achieved through pruning, which consists of removing of parameters based on feature importance [128]. However, pruning leads to compact models that are a sub-graph of the original model architecture, which unnecessarily constrains the architecture of the compact model. Pruning out components of the network architectures can take several forms such as channel pruning where filter channels with activations below pre-set thresholds can be removed [129], or one could remove entire layers based on their redundancy towards the final task being performed [130]. Attempts have also been made towards computing parameter importance at the level of individual neurons [131] for removal of redundancies at a scale much more granular than channels or layers. Additionally, parameter importance can be computed with the Fisher information matrix for the parameters of the trained model. Those parameters with Fisher information values below set thresholds can be pruned out

in order to reduce model memory footprint and inference time [132]. This method has a drawback of being computationally expensive as the computation of the Fisher matrix over the whole parameter space can involve an intractably large number of operations. This is addressed by a diagonalization of the Fisher information matrix, with the downside that such a diagonalization can cause a significant information loss.

2.5.3 Teacher-Student Distillation

Over the years, knowledge transfer methods have been proposed that can transfer knowledge learnt by a large model to an arbitrary compact model. Hinton et al. [133] introduce the concept of teacher-student knowledge distillation, which they define as a transfer of knowledge from the final layer of the large model to a compact model during the training of the latter. The idea espoused by them seeks to leverage the fine representation learning capabilities of large teacher models afforded by their large parameter sets towards serving as auxiliary supervision during the optimization of reduced student models. Romero et al. [134] extend the idea of knowledge transfer to include intermediate learnt feature maps in the training of the compact model as well with the intuition that given the hierarchical learning capabilities of deep networks, the knowledge transfer process can be more effective in ‘guiding’ the student if applied as a regularization step over multiple intermediate stages rather than just an end-stage distillation objective. While model compression and teacher-student knowledge transfer have been studied in machine learning research, relatively few works deploy both concepts in ultrasound imaging settings.

Overcoming parameter redundancy is important to medical imaging as the time required for diagnostic evaluations depends on model inference speeds, and memory footprint of algorithms come at the expense of storage space for other critical data. In a relevant study concerning the transferability of ultrasound image classification across models of different capacity, Vaseli et al. [135] classify standard views in adult echocardiography by training traditional large deep learning models and use the method in [133] to train reduced versions of these models. The reduced models proposed in [135] are subsets of the heavier models constructed by removing layers

at different stages from the heavier teacher models, and therefore do not allow for exploiting the flexibility afforded by the teacher-student distillation process in terms of selecting the appropriate student models for efficient knowledge transfer. This thesis proposes the concept of Perception and Transfer for Reduced Architectures as a general framework to train compact models with knowledge transfer from traditional large deep learning models using gaze-tracking information to condition the solution without requiring such information at runtime. For the tasks of fetal abdomen, femur and head detection, compact reduced models had accuracies close to the large models, while having a much lower memory requirement.

2.6 Concluding Remarks

There has been an intense flurry of research activity in different facets of deep learning, scene understanding, video analysis, continual learning, model compression and human-experience centered machine learning. This has been matched by a correspondingly rich assortment of ideas proposed for fundamental research into algorithmic interventions in clinical imaging scenarios, addressing inefficiencies in medical image analysis platforms, exploration of questions pertinent to widespread deployment and associated public health and translational research questions relevant to a range of socio-economic and geographic situations. This thesis touches upon several of these themes in the quest towards establishing methods for fetal ultrasound video analysis with a clinically motivated set of tasks, under different data distribution situations and with the possibility of inclusion of sonographer perceptual cues. As such, the literature review aims to be as relevant as possible towards creating the context and background for subsequent chapters, and attempts to give a holistic overview of the current state-of-the-art relevant to the individual chapters proposed. Consequently, involved discussions around the technical details and motivations for different deep learning components are excluded, as are detailed expositions of translational research questions pertinent to ultrasound image analysis.

*Alles Gescheite ist schon gedacht worden.
 Man muss nur versuchen, es noch einmal zu denken.*

*All intelligent thoughts have already been thought;
 what is necessary is only to try to think them again.*

— Johann Wolfgang von Goethe

3

Fetal echocardiography video analysis

Contents

3.1	Introduction	48
3.2	Proposed Contributions	50
3.3	Top-level video classification and cardiac localization	50
3.3.1	Problem definition	51
3.3.2	Spatial Baseline Models	53
3.3.3	Direct Temporal Encoding	55
3.3.4	Hierarchical Temporal Encoding	55
3.4	Datasets and Implementation	58
3.4.1	Training details	59
3.5	Experiments and Results	59
3.6	Multi-anatomy localization	65
3.6.1	Problem definition	66
3.6.2	Proposed Approach	67
3.6.3	Convolutional Architectures	69
3.6.4	Implementation	71
3.6.5	Results and Discussion	71
3.7	Conclusion	73

Precisely analyzing fetal cardiac US videos is a challenging task even for human experts due to the complex appearance of different anatomical structures in a small region. Furthermore, there exist speckle, shadowing, enhancement and variations in contrast levels in US images. For the evaluation of the fetal heart, the sonographer or a fetal cardiologist would require to observe the heart over one or more cardiac

cycles. Thus, diagnosis of fetal ultrasound involves observing both spatial features and temporal cues making such diagnosis a spatio-temporal task. In this chapter, we analyze the potential for applying deep learning techniques to ultrasound video for describing key parameters of the fetal heart, with a focus on aggregating spatio-temporal features within deep convolutional architectures, and utilizing the same for understanding the visibility detection, viewing plane identification, cardiac localization and multi-anatomy motion analysis tasks in an automated fashion. Visibility detection for fetal cardiac ultrasound refers to the task of establishing whether or not the heart is observable on a frame. Viewing plane identification refers to the task of establishing whether or not a frame in which the heart is apparent belongs to one of the fetal cardiac standard viewing planes as recommended by the guidelines of the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) [6]. Localization refers to finding the center and radius of the heart. Multi-anatomy motion analysis refers to analyzing the structures such as valves, ventricle walls, foramen ovale and so on that enable the heart to perform its physiological role of haemodynamic circulation over a cardiac cycle.

The work described in this chapter has resulted in the following peer-reviewed publications:

1. Arijit Patra, Weilin Huang, and J. Alison Noble. "Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video." *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, Cham, 2017. 276-284.
2. Arijit Patra, J. Alison Noble, "Multi-anatomy localization in fetal echocardiography videos." *The 16th IEEE International Symposium on Biomedical Imaging (ISBI) 2019, Venice, Italy April 8-11, 2019*.

3.1 Introduction

The fetal cardiac screening guidelines of the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG) at 20 weeks of gestation recommend an initial analysis of the fetal cardiac standard views – the four chamber or 4C view,

the three Vessel or 3V view and the Left Ventricular Outflow Tract or LVOT view. The rationale for performing the identification of such standard cardiac viewing planes is to monitor the physiological status of vital cardiac anatomies and look for spatial and temporal cues of Congenital Heart Disease in the heart structures visible in these standard planes. Following a visibility assessment of whether a sequence actually includes the cardiac region and the heart or is a sequence of background frames and the subsequent viewing plane identification, it is important to accurately localize the heart region. The radial area included within the cardiac region is to be evaluated because the normal structural growth of the fetal heart through the gestation is primarily estimated by means of the heart radius.

Another diagnostic aspect of performing fetal echocardiography analysis is to monitor the structural and functional aspects of the sub-structures of the cardiac region – namely the valves, ventricle walls, blood vessels, septal wall, arteries and venal apparatus and the heart chambers. This is because the symptoms of congenital abnormalities often manifest as observable markers in one or more of these structures that comprise the fetal heart. Therefore in subjects where a possible incidence of a congenital heart condition is suspected during the routine second trimester scan, the follow-up specialist scanning process would closely examine these structures within the context of the cardiac cycle. In addition to assessments of structural abnormalities, a regionalized functional analysis is vital to overall diagnoses of fetal cardiac abnormalities. This includes an analysis of motion patterns of anatomical structures like the valves, the foramen ovale and the ventricle walls, with respect to global heart motion, as symptoms of several cardiac abnormalities may lie in anomalous motion patterns of these structures. Thus, a local analysis of such motions in a global context of the cardiac cycle is vital. Here, we propose a CNN-based pipeline to detect multiple anatomical structures in the fetal heart from fetal echocardiography videos and quantify their motion extents within an end-systole to end-diastole cycle.

3.2 Proposed Contributions

In this chapter, we consider the fetal heart video understanding problem through the lens of deep learning techniques, and focus on the assessment of visibility, viewing plane identification, localization and multi-anatomy detection and motion estimation from B-mode ultrasound video sequences. The chapter is structured to introduce the visibility and viewing plane joint-classification problem, followed by heart region localization (in terms of the heart radius and heart centre) and multi-anatomy localization and motion estimation tasks. The contributions of the chapter are:

1. We cast the problem of fetal heart analysis as a multi-task prediction in a hierarchical convolutional architecture that progressively encodes temporal information throughout the network. This is vital to retain both spatial detail and meaningful temporal patterns, which are key to accurate and automatic heart description in fetal ultrasound videos.
2. We investigate multiple temporal encoding architectures that learn a strong spatio-temporal representation tailored for ultrasound video representation. We study these approaches by analyzing the architecture for our specific ultrasound video analysis tasks, and the loss functions designed for joint classification and localization.
3. We propose tracking by detection methods for localization of multiple anatomical structures across frames of fetal echocardiography videos considering as exemplars of the approach, five diagnostically relevant sub-structures from the 4C view.

3.3 Top-level video classification and cardiac localization

The first step towards diagnostic evaluation of the fetal heart is to establish whether or not the heart is visible. Having determined the visibility of the heart in a frame, a classification task is performed to identify the standard viewing plane the given frame belongs to. Note that in frame-to-frame analysis, it is likely that some frames would involve transition from one viewing plane to the other, or to a background frame. Further, the unpredictable motion of the fetus relative to the probe means that some

frames could have blurred out or non-distinct anatomical cues. Such frames are grouped with the background frames in our experiments. The visibility and viewing plane identification together form a multi-class classification problem. The issues apparent in frame-to-frame analysis have led to most sonographers analyzing the fetal heart over multiple frames, or video sequences. Indeed, the process of ‘freezing’ of frames as applied to evaluations of several other parts of the fetal anatomy is often replaced by a retention of short video sequences of observed cardiac structures for later reference by many sonographers. Given the particular difficulty of acquiring an optimal standard viewing plane for the fetal heart, sonographers with relatively less experience in fetal cardiac analysis may prefer to save sequences extended over a few frames encompassing a particular cardiac standard plane. We adopt the same clinical motivation to perform video-level visibility and viewing plane classification followed by localization of the heart in terms of the centre (x,y) coordinates and the radius r together in a multi-objective setting.

3.3.1 Problem definition

The visibility relates to the presence or absence of the heart in each frame (with partial visibility of less than 50% being deemed as not visible). When the heart is present, the standard fetal cardiac viewing plane can be either the four chamber (4C) view, the left ventricular outflow tract (LVOT), or the three vessels (3V) view, as shown in Figure 3.1. By jointly considering the visibility and the standard fetal cardiac viewing identification, we include a background class to define a 4-class classification problem. The location of the heart is defined by its center and radius, which can be cast as a 3-parameter regression objective. From the model optimization perspective, the two tasks have two different objective functions and use separate targets or labels or ground truths. As such, we rely on the same input data for predicting multiple targets. One of the ways to enable such a joint prediction is to have an ensemble of models with individual models performing either the classification task with a cross-entropy objective or the localization task with regression loss terms such as an L2 loss. While an ensemble of models may

be reasonable considering the different nature of the objective functions, such an ensemble would lead to requirements of memory that scale with the number of distinct tasks pursued, a constraint untenable with the resource constrained environments that many automated medical image analysis algorithms are expected to be deployed in. Therefore, we attempt to have a single model that is versatile enough to perform a simultaneous classification and localization task by casting the learning problem in a multi-objective loss function between predictions of the model and the different labels for a given input sample. Our task is then to jointly predict the visibility, standard fetal cardiac viewing plane and location (x, y, r) of the fetal heart in ultrasound videos.

In this case, both classification and localization are formulated jointly as a multi-task learning problem within the convolutional architecture. The outputs of the two tasks are connected to the last fully-connected (FC) layer of a CNN, such as AlexNet [35] or VGGnet [36], as shown in Figure 3.3 and Figure 3.4. A cross-entropy loss with a softmax function is used for the classification task, while L2 loss is minimized for regressing the values of heart centre and radius. These configurations customize a regular CNN to our task, and form our baseline spatial architecture where the input is a single ultrasound image. Our temporal models are extended from this baseline architecture by allowing input of a sequence of continuous frames. To investigate the model capability for learning spatio-temporal details, we consider the aggregation of spatial features as they evolve temporally and implement a fusion of such features at different levels of the hierarchy of learnt convolutional features. The idea is that when a sequence of frames capture the cardiac motion, the different frames incorporate the spatial structure of the heart at those particular time points. Then these alterations in the visible spatial structure of the heart region can be understood as the consequence of the cardiac motion. This implies that the information introduced by the influence of temporal dynamics can be gathered from the spatial cues as long as one is able to accumulate such spatial cues over a certain length of time, as such temporal evolution would be apparent over several frames. Given that the knowledge representation that is learnt by convolutional

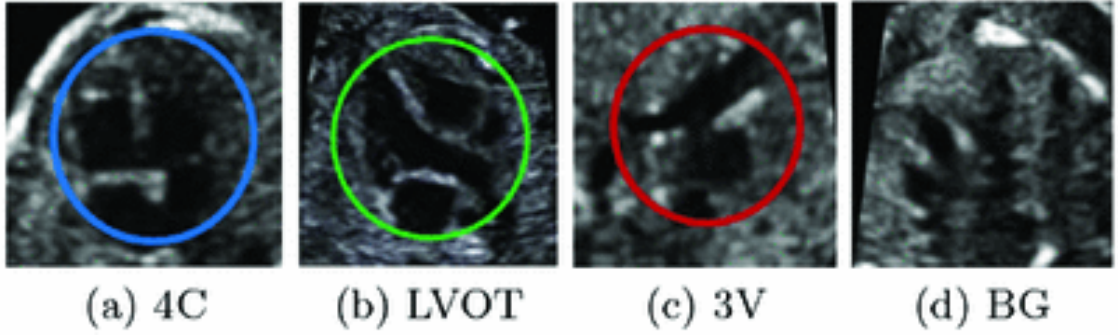


Figure 3.1: Three different view planes of fetal heart: (a) the four chamber (4C), (b) the left ventricular outflow tract (LVOT), (c) the three vessels (3V), and (d) the background (BG).

networks is a function of the patterns in the input space, the inclusion of temporal cues towards building more robust knowledge representations would require the input space to include a sequence of consecutive frames as a proxy for the temporal dynamics. Recently, temporal encoding methods have been introduced to enable action recognition using deep learning methods in [53], where the idea of aggregating spatial information was used. Differently from [53], the presented approach extends from the concept of temporal encoding of spatial information over a sequence of frames and introduces a hierarchical encoding strategy for ultrasound videos, within a multi-task setting of both standard plane classification and the localization of the cardiac region, unlike the single task setting of [53] (which does not address the localization problem). Additionally, in a departure from the ideas of [53], we use convolutional layers using small kernel size of 3×3 to capture finer details important to ultrasound video analysis tasks. Further, the influence of a progressively delayed fusion of spatial features is introduced, and benchmarked against other information fusion approaches for the analysis of fetal echocardiography videos.

3.3.2 Spatial Baseline Models

Our task is predicated on the question of whether the inclusion of temporal cues improve the classification and localization performance of deep neural networks in fetal echocardiography and to what extent is such an influence observed. Such a comparison is feasible only with baseline results that are obtained when using

only spatial features such as a single frame input. The network outputs in all the different models proposed and compared with are the class label (4C, 3V, LVOT, background); the coordinates of the center and the radius of the heart region (x, y, r) . For the first step of creating spatial baseline models, we leverage the classical AlexNet and the 16-layer VGGnet architectures, both of which are modified in our implementations to accept a single channel grayscale image. The two Spatial Baseline Models are referred to as SBM_{alex} and SBM_{vgg} respectively. The networks accept as inputs an ultrasound image with size of $224 \text{ pixels} \times 224 \text{ pixels}$. Details of the architectures are presented in Figure 3.2. Notice that the number of neurons in the fully-connected layers is changed from the VGG-16’s original 4096 to 1024 neurons in our adaptation in this experiment. This modification was made to avoid overfitting, as the number of predicted parameters in our task is significantly smaller than the 1000 categories of object classification in ImageNet [39] for which these architectures were originally designed. The last fully-connected layer is connected to the output layer which computes the softmax probabilities of four classes (3 view planes and the background frames), with the predicted heart center and radius for the localization circle.

Next, we address the question of aggregating spatial features across a specified temporal length. Convolutional operations modify the input into a set of feature maps using learnable filters that act as weights or parameters for that layer. A sequence of these convolutional operations enable a transformation of the input features into lower dimensional features that are ultimately used as model predictions (can be as interpreted as class confidence probabilities or directly as real numbers for a given target). Initial layers of the CNN learn basic features of the input space such as edges and deeper layers represent finer features [49]. The question then is what happens when the input space is not a single image but a temporal sequence of frames. The spatial features over multiple frames can be aggregated either at the outset or after transformations by convolutional operations. The former strategy is called Direct Temporal Encoding as explained below.

3.3.3 Direct Temporal Encoding

To compute temporal information, a straightforward approach is to modify the CNN architecture to allow for an input of multiple contiguous frames. Such a multi-frame input naturally includes temporal context which is important to produce a more reliable prediction. As the size of the CNN architecture is fixed by design, and the number of CNN parameters should be scalable, the input of the CNNs is configured and fixed to a short, fixed-length ultrasound video clip cropped from temporally contiguous frames. In experiments, we use an 8-frame video clip as input for all temporal models where the labels of the 4th frame is applied to the video clip.

For model configurations, the main idea is to extend network connectivities in the time domain, allowing these connectivities to automatically learn temporal dynamics between contiguous ultrasound images. Essentially, features learned over spatial levels need to be assessed for temporal developments. This can be understood as a multi-frame aggregation of learnable spatial features and the filters of the convolutional layers can be optimized with the availability of information from multiple temporal steps. To this end, we extend our spatial baseline models to Direct Temporal Encoding (DTE) by introducing direct temporal filters with kernel size of $k \times k \times T$ in the first convolutional layer, where T is the temporal dimension which is set to $T=8$ in our experiments. Here $k \times k$ is the spatial size of convolutional kernels which is identical to those of the spatial baseline models, e.g., 11×11 and 3×3 for the AlexNet and VGGnet respectively. These temporal filters are able to directly encode both spatial details and temporal connections of the continuous ultrasound frames. This enhances the low-level representation in the first convolutional layer, which in turn leads to stronger deep high-level representations by propagating them throughout the networks. The DTE architectures are described in Figure 3.3 and Figure 3.4.

3.3.4 Hierarchical Temporal Encoding

The DTE approach directly computes local temporal features at the pixel level, and is powerful enough to capture detailed low-level motion characteristics. However,



Figure 3.2: Spatial baseline models based on AlexNet (SBM_{alex}) and VGGnet (SBM_{vgg}).

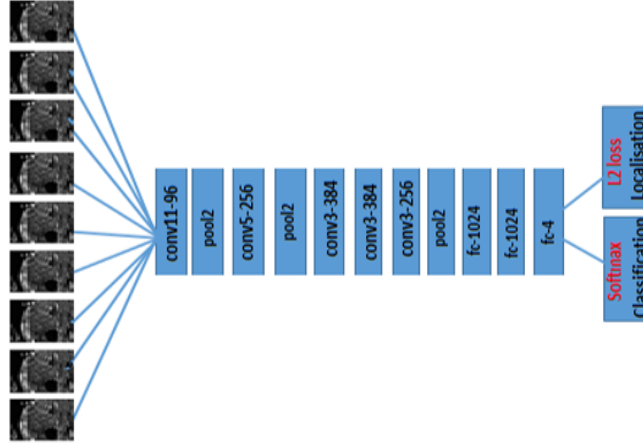


Figure 3.3: Direct Temporal Encoding (DTE) models modified from AlexNet.

CNN predictions are built on high-level deep representation (for example, the fully-connected layer features). It is important to investigate: (i) how this low-level temporal information is propagated efficiently and accurately to the final deep representation; and (ii) which level of the temporal information is crucial to the final prediction. Since ultrasound images often contain relatively weak visual information at the pixel level, temporal representation immediately computed at the first layer by the DTE may not be robust, and multi-layer propagation throughout the whole

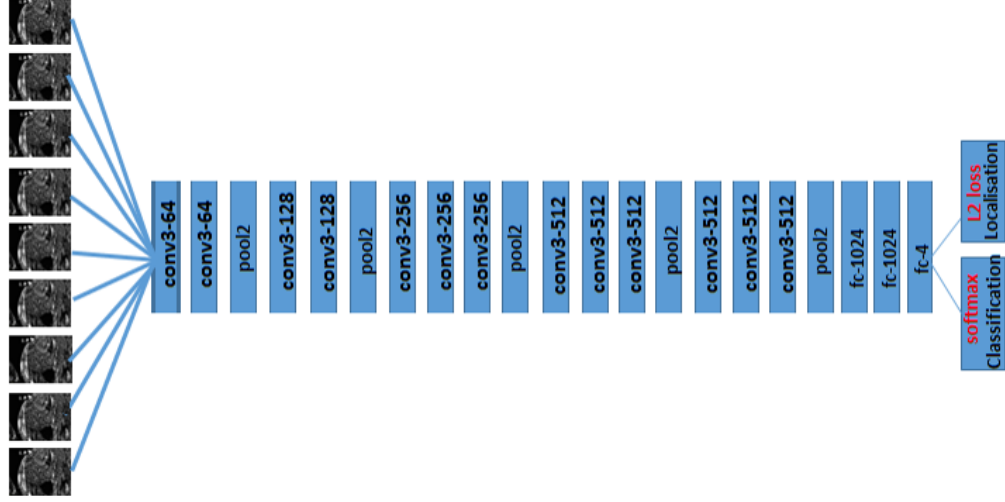


Figure 3.4: Direct Temporal Encoding (DTE) models modified from VGG-16.

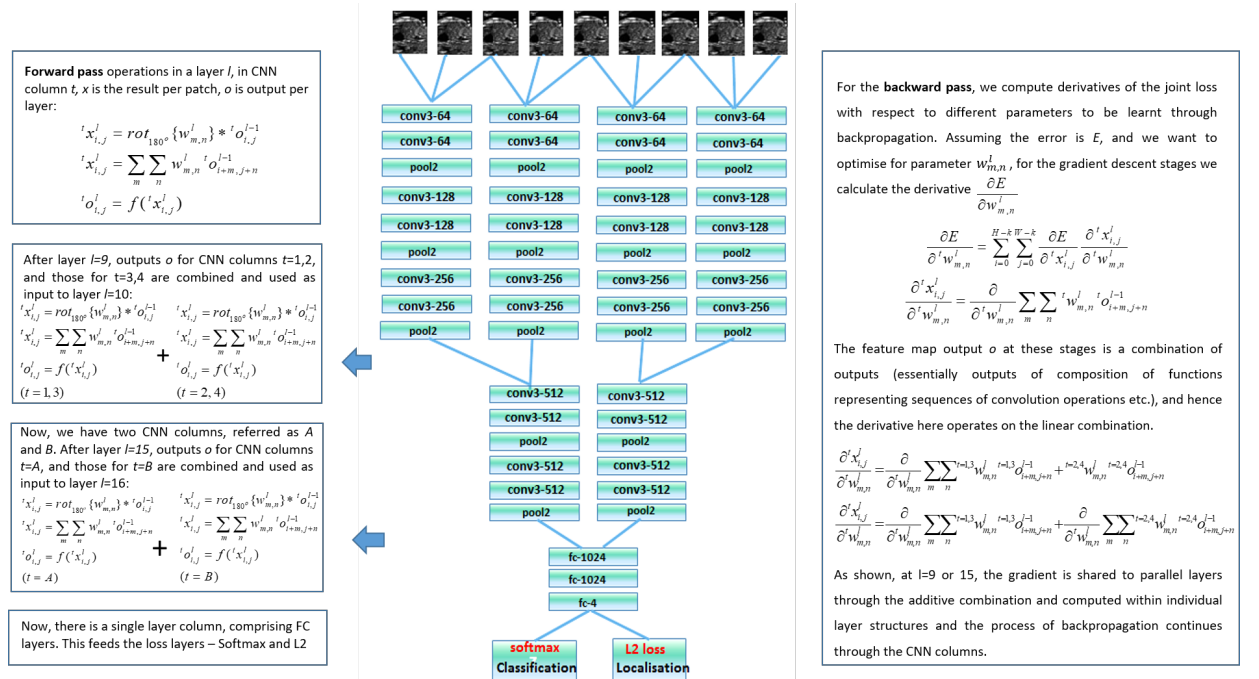


Figure 3.5: The proposed Hierarchical Temporal Encoding (HTE) that progressively incorporates spatiotemporal information throughout the network.

network may lead to a certain degree of information loss. The need is to develop a new temporal encoding approach that allows the temporal details of ultrasound videos to be computed more accurately and robustly, and to be propagated more efficiently throughout the network without significant information loss.

With these considerations, we have developed a new Hierarchical Temporal Encoding (HTE) approach. We present a number of technical improvements that elegantly tailor this slow fusion approach towards our task of fetal ultrasound description: (i) we incorporate multi-task prediction into this hierarchical architecture for jointly estimating the view plane and localization of the fetal heart. (ii) and, more convolutional layers using small kernel size of 3×3 are applied, resulting in a deeper model able to capture more spatio-temporal details which are particularly important to our task. These design features aim to capture key characteristics of ultrasound video interpretation to encode the temporal domain and the spatial domain, and a finer feature extraction from the previous layer is more desirable to propagate the features throughout the network.

Details of the HTE model are presented in Figure 3.5, where frames from the video clip are gathered in groups of three, and then fed into the first layers of multiple sets of CNNs which share the same parameters. These serve as quasi-independent CNNs until the 3rd pooling layer, where the outputs of the 3rd pooling layer are taken in groups of 2, and further fed into separate sets of two CNN structures also sharing the same parameters. This is maintained until the final encoding happens just prior to the first FC layer, and subsequent FC layers effectively process the overall information from the video. Therefore, a slow and progressive combination of the learned features enables the propagation of the diversity of information encoded in the temporal space.

3.4 Datasets and Implementation

Real-world clinical dataset

We start with the dataset described in Section 1.2.1. To recapitulate, this dataset consisted of 91 routine clinical cardiac screening videos from 12 subjects at gesta-

tional ages ranging from 20 and 35 weeks. Each video had a duration of between 2 and 10 s and a frame rate between 25 and 76 frames per second (39556 frames in total). It contained one or more of the three views of the fetal heart. Videos from 10 subjects were used for training, and the remaining 2 for test. Experiments implement a 4-fold cross validation for the train:test split. The particular subject-wise split is to roughly simulate an 80:20 train:test split, with the cross validation folds choosing different test subjects each time, so that the subjects available for testing is different across the folds.

Data preparation

For the training step, we split available videos into frames and applied data augmentation by performing an up-down and a left-right flipping. Then, we chose 8 frames at a time in the sequence of their original occurrence in the videos and combined them back into 8-frame sequences and repeated the same for the augmented data. This enabled ultrasound to obtain 9337 video clips with roughly equal samples from different view classes. The current frame is the 4th frame in a 8-frame video clip. All frames in the dataset of size 430 pixel $\times$ 510 pixel were cropped into 224 pixel $\times$ 224 pixel windows centered about the heart center, in order to be compatible with the input resolution chosen for our architectures.

3.4.1 Training details

All new layers across the networks used are initialized by using random weights. Our spatial baseline models are trained by using a 25-frame mini-batch with a learning rate of 0.01. The DTE models are trained with 8-frame chunks of videos using a batch size of 25 with a learning rate of 0.01. A batch size of 20 and a learning rate of 0.001 are used for the HTE.

3.5 Experiments and Results

We performed multiple sets of experiments for both the classification and regression tasks. We start with the generation of a strong feature-based baseline using single

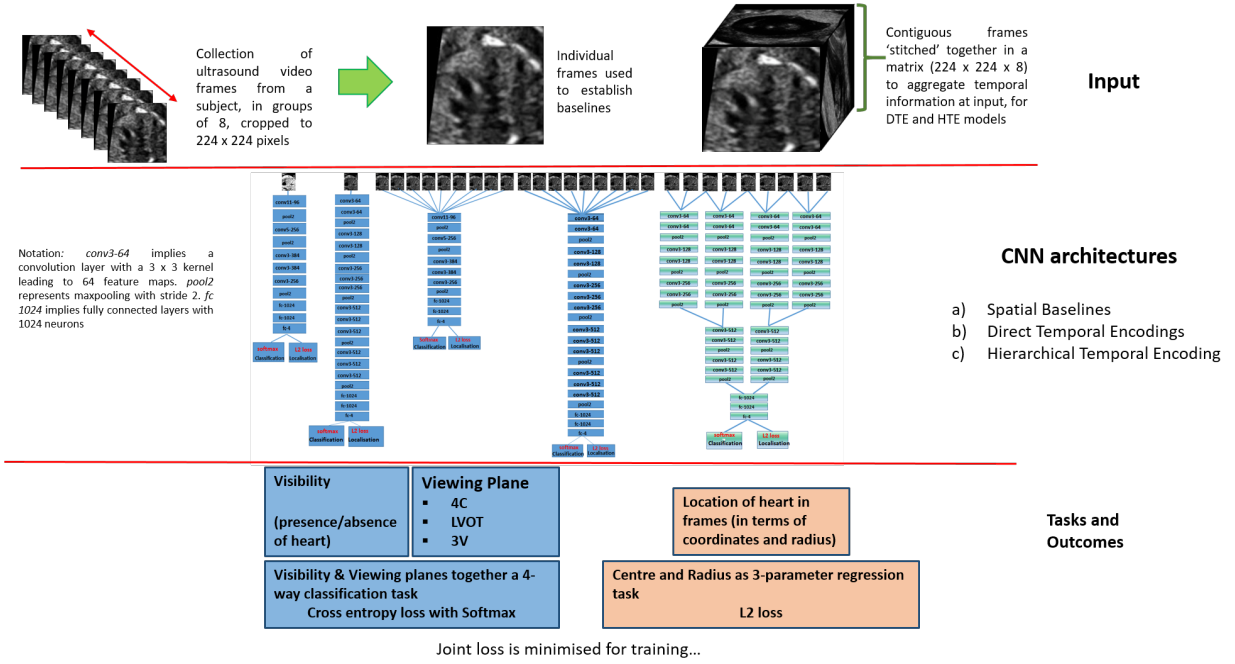


Figure 3.6: Workflow of experiments with baselines and different encoding strategies under multi-task objective functions.

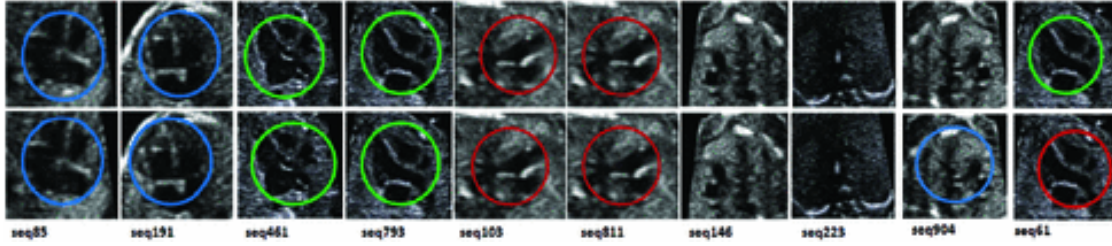


Figure 3.7: Examples of predicted results by the HTE (bottom) with Ground Truth (top). Colours (blue – 4C, green – LVOT, red – 3V) are used to demonstrate different viewing planes.

frames of video sequences, followed by temporal encoding models, which are then augmented by a hierarchical transfer of spatio-temporal details through the network. We analyze our models for multi-task learning from shared features, achieved by focussing on requisite regions of interest for each task.

Classification and localization performance

Our models are evaluated on real-world fetal cardiac screening videos. Results on several exemplar frames are shown in Figure Figure 3.7, and full results are compared in Table 1, where temporal models obtain better performance, particularly

on classification. Importantly, the improvements on the challenging classes are remarkable: 67.60% $\rightarrow$ 88.27% for the LVOT, and 59.11% $\rightarrow$ 79.80% for the 3 Vessels. The HTE outperforms DTE models by over 15% in both classes, which suggests that the HTE provides a more optimal method for encoding detailed and meaningful spatiotemporal features for fetal heart description. This can be further verified by the confusion matrices shown in Fig. 4. The advantage of the HTE is also affirmed by gains seen in the localization task. We compare our models with that in [62] which uses hand-crafted features and a CRF-filter. The IoU implementations of HTE achieve localization errors of 20.32% compared to the best result of 34% in [62]. Our models explore stronger deep spatiotemporal representations learned end-to-end with features shared over multiple tasks. Furthermore, our model is able to predict both the centre and radius for heart localization, while the radius is used as a strong prior information and only the centre is estimated in [62].

Table 1. Performances on real-world fetal cardiac screening videos, in accuracy rates (%). A correct localization is defined with an Intersection-over-Union ratio over 0.25 [62].

Method	Average Classification Accuracy %					Localization
	4 Chambers	LVOT	3 Vessels	Background	Overall	
SBM_{alex}	89.22	67.60	59.11	90.36	78.95	69.71
SBM_{vgg}	90.84	69.83	62.07	91.37	80.73	72.29
DTE_{alex}	88.68	73.18	57.63	88.83	79.16	72.21
DTE_{vgg}	91.11	70.95	67.00	89.34	81.78	77.26
HTE	81.67	88.27	79.80	86.80	83.58	79.68

Spatiotemporal encoding effects

In this subsection, we present further analysis of spatiotemporal aggregation of ultrasound representations throughout the CNN architecture. The performance of the direct encoding architecture, DTE_{alex} , can approach that of the deeper baseline model SBM_{vgg} with VGGnet. This is indicative of the dynamical interplay between competing factors governing the ability of deep networks to optimise learning of spatial and temporal information. While an information fusion at the level of extracted features at different time-steps or frames could capture significantly more contextual detail than single frame approaches, the aggregation may cause an overlap

between varying spatial features which tends to compromise the ability to learn distinct spatial features. This effect is pertinent to the ultrasound modality because of the non-trivial influence of imaging artefacts, shadows and speckle. Particularly, a pixel-level combination before the first convolution layer, as described in the DTE structures, tends to undermine the framewise variations in spatial information by directly fusing along the temporal direction. This trade-off is indirectly addressed in the HTE model, by feeding a relatively smaller temporal extent per convolutional network, and by aggregating multi-level spatiotemporal information progressively throughout the CNN architecture. This provides a more principled approach for spatiotemporal encoding that enables the final deep representation to essentially capture the entire temporal regime of the video segment. Thus, our HTE networks are able to achieve significant gains in accuracy for both the classification and localization tasks. It is to be noted that the optimal number of frames to achieve the best gains in this trade-off would exhibit significant variations as a function of the modality, the size of datasets and the complexity of features in spatial and temporal domains. While there has been considerable research into developing generalized convolutional architectures for ultrasound image analysis such as SonoNet [81], we perform baseline comparisons for the frame level and as the backbone in the DTE using the AlexNet and VGG-16 approaches. This is because we implement a multi-task learning objective along with multiple frames as input. SonoNet implements a global average pooling layer, while our approaches extend the multiple loss terms from the final fully-connected layer. This also follows from the fact that concurrently proposed approaches for echocardiography apply feature extraction using a pre-trained VGG block.

The proposed Hierarchical Temporal Encoding (HTE) is able to efficiently retain meaningful spatial and temporal details that are critical to implementing multi-task prediction in ultrasound video. We also investigated different temporal encoding models on a real-world clinical dataset, where the proposed HTE achieved significant performance gains. The performance of the different temporal encoding approaches varies across the views due to the inherent complexity of appearance of

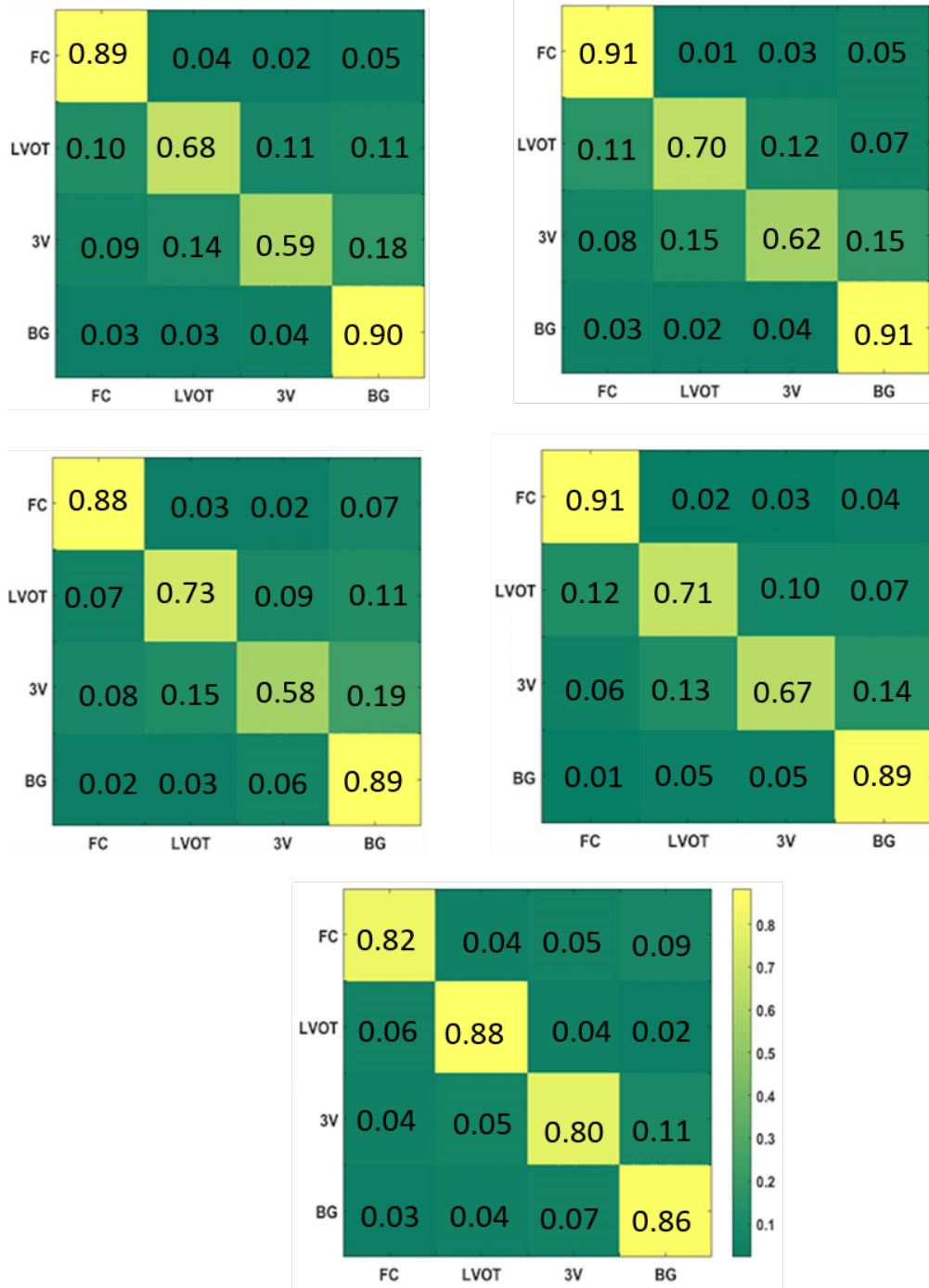


Figure 3.8: Confusion matrices for classification. L to R (top): SBM_{alex} , SBM_{vgg} ; $LtoR(middle)$: DTE_{alex} , DTE_{vgg} ; *Bottom* : HTE .

anatomical structures in different views, and the ambiguous motion patterns that cause susceptibility to be misclassified as background frames. As such, there is an added issue of starting with an imbalanced dataset (there are a larger number of 4C view frames originally) and despite data augmentation, the intra-class variability is not sufficiently represented for some of the views, considering the inherent complexity of ultrasound videos with the presence of artefacts, speckle, enhancements and acoustic shadows. While the HTE model attains substantial improvements in performance for challenging standard planes as the 3 Vessels and LVOT frames, the performance is better for the 4C frames using the DTE approach with a VGG backbone. This stands out compared to the expectation that the inclusion of temporal dynamics would improve the representation learning for the cardiac ultrasound videos. While overall performance improvements indicate that such temporal information can gainfully enable superior classification, the superior performance of the direct temporal fusion for the 4C frames can potentially be attributed to the fact that such frames have significant presence of the ventricular blood pools which appear as a blob region in the end-diastole and the corresponding end-systole appearance significantly diminishes the proportion of the frame that is occupied by the ventricular blood pools. This disparity in spatial features across the temporal window can potentially impair the ability of a late fusion strategy to sufficiently aggregate information pertinent to the class-level representation, as earlier layers likely overfit on specific feature regions. This is due to the relative consistency of the frames in a specific input set of three images out of the sequence that a particular convolutional branch has access to in the earlier stages. Thus, different strands of the HTE access sub-clips of three frames with the individual sub-clips processing frames that mutually differ significantly in spatial appearance, leading to the final representation learning being potentially adversely affected. This is not an issue when undertaking direct fusion as the whole temporal window undergoes the convolution operations with the same layers all through the network, enabling the relatively performance on the 4C frames by the DTE approach as compared to the HTE.

3.6 Multi-anatomy localization

Earlier in this chapter, we dealt with a top-level ultrasound video analysis for the fetal heart and focussed on overall spatial and temporal details. Now, the regionalized structural and functional analysis is cast as a motion localization problem for multiple anatomies in the cardiac region. Assuming the cardiac region to be an action space, the different sub-anatomies can be identified as localized sources of physiological motion. Such physiological motion analysis can be understood as analogous to multi-object action localization in computer vision literature, and spatial object detection pipelines can be temporally extended to adapt to our cardiac structure motion analysis problem.

Regionalized analysis of cardiac structures in fetal echocardiography sessions is vital towards detection of potential markers for several congenital heart disease classes. We cast the task of segregating anatomical motions in fetal ultrasound videos as an analogue of the multi-object action recognition problem in scene understanding contexts. The tracking-by-detection approach was combined with a regression to a motion progress property defined for individual structures within the video sequence. Fetal heart motion is an important diagnostic indicator for structural detection and functional assessment of congenital heart disease. In this section, we propose an approach towards integrating deep convolutional and recurrent architectures that utilize localized spatial and temporal features of different anatomical substructures within a global spatiotemporal context for interpretation of fetal echocardiography videos.

We formulate our task as a cardiac structure localization problem with convolutional architectures for aggregating global spatial context and detecting anatomical structures on spatial region proposals. This information is aggregated temporally by recurrent architectures to quantify the progressive motion patterns. We experimentally show that the resulting architecture combines anatomical landmark detection at the frame-level over multiple video sequences - with temporal progress of the associated anatomical motions to encode local spatiotemporal fetal heart dynamics and is validated on a real-world clinical dataset. Though motivated by

fetal echocardiography analysis, this method is transferable to motion analysis in other ultrasound tasks (or to other modalities with video data) and does not require optical flow information.

3.6.1 Problem definition

Multi-anatomy localization requires the ability to extract multiple region features from a given frame and classify those into the classes of anatomies under study or background. Aggregation of temporal context is then a replication of this process over sequential frames and combining similarly predicted region proposals as an anatomical class in the video. The structure detection is followed by motion quantification for which the anatomical deformation is treated as an action, or a sequence of actions in the physiological environment being visually interpreted. Thus, anatomical movement can be approximated as a constrained action recognition with the periodicity of physiological deformation serving as a marker of progress. Specific to our task of interpreting fetal echocardiography videos, we consider five key anatomical motions Figure 3.9, based on their clinical relevance towards diagnosis of 4C view Congenital Heart Disease [8] – the mitral and the tricuspid valves, the left and the right ventricle walls, and the foramen ovale.

It is pertinent that a few terms are defined here, in order to better contextualize the subsequent discussion:

1. **Motion progress** – This is used to describe the extent to which the motion of a particular anatomical structure completed the progression to the end-diastole state. Assuming a complete motion cycle as a progression from an end-systole to the next end-diastole state, the motion progress estimates the fraction of this path that has been traversed at a particular intermediate frame by each of the individual anatomies considered for tracking. On similar lines, a complete motion cycle for a particular anatomical structure can be understood as an interconnected set of local regions over a sequence of frames, essentially forming a 'tube' of a particular anatomical structure in motion. This is referred to as a motion tube or an echo tube in further discussion. This is mathematically defined by equation 3.1.

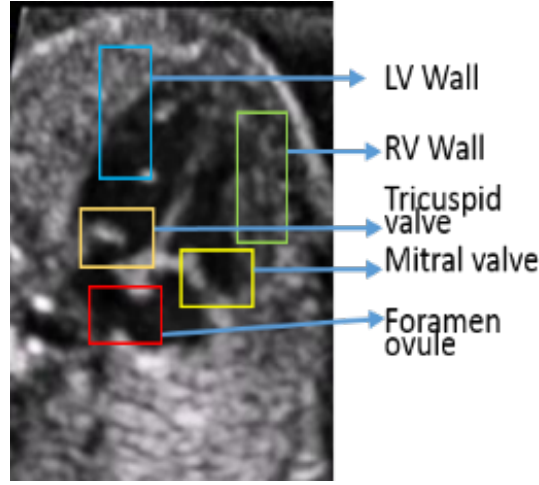


Figure 3.9: Different cardiac anatomical structures in the 4C view – the motions of these structures are assessed in this segment.

$$p = (N_s - N) / (N_e - N_s) \quad (3.1)$$

Here p represents motion progress, N is the frame being observed, N_s and N_e are start and end frames (taken as end-systole and end-diastole frames in an input clip), 0 and 1 are start and end of motion/action.

2. Motion localization – This term refers to the spatial level localization of a particular anatomical structure undergoing a progression from the end-systole to end-diastole state. This can be understood as successive localization of particular anatomical regions over the sequence of frames from the end-systole to the end-diastole frame.

3. Motion categorization – The idea here is to categorize the motion as being performed by a particular anatomical structure, essentially detecting anatomies on the same frame, successively across the sequence of frames from end-systole to end-diastole. This can be thought of as analogous to multiple object detection, but across a sequence of frames which themselves form a coherent video clip.

3.6.2 Proposed Approach

Initially, the presence of the heart in the four-chamber view is ensured using a pre-trained modified VGG-16 convolutional neural network, called Model-1, that

classifies the presence or absence of the 4C view. If the 4C view is visible in a given frame, we proceed with a detection routine that looks for spatial locations of the structures whose motion is of interest. To that end, we design a model inspired by [136] based on a VGG-16 architecture to train to classify between the five structures of interest and background. Next, extending to temporal localization and motion phase quantification require that structures of interest be recognized across multiple frames and their motion seen as successive detection of same structures over time.

We follow the region proposal algorithms proposed in [136] and generate multiple aspect ratio region proposals with the difference that we generate region proposals directly at input stage and constrain them to be in either 50 x 50 or 50 x 100 patches without overlap to keep computational requirements manageable. An unconstrained generation of bounding boxes with overlapping can potentially lead to a significant redundancy in terms of the probability of finding a structure of interest as each structure appears uniquely only once in a given frame. We compute their class confidence scores (representative of the anatomical structure) and refine the bounding boxes using Model-2, the convolutional architecture of which is a modified version of the approach in [136]. Model-2 also includes LSTM layers to aggregate temporal information over sequential frames. A difference from the methods in [136] is that we do away with ‘objectness’ calculations and treat backgrounds as a separate class. This is due to the difference of our task in the much lower number of probable classes per region and requirement to pass fully connected layers of patches of the same class to train a motion progress regression step later. The ground truth for the latter training is estimated from the start and end labels assigned to structural motions and normalized between 0 and 1. In a healthy heart, these are likely to be concurrent with the end-systole and end-diastole frames but there exist slight differences due to motion of the fetus and of the probe. We describe our pipeline in detail in subsequent sections [137].

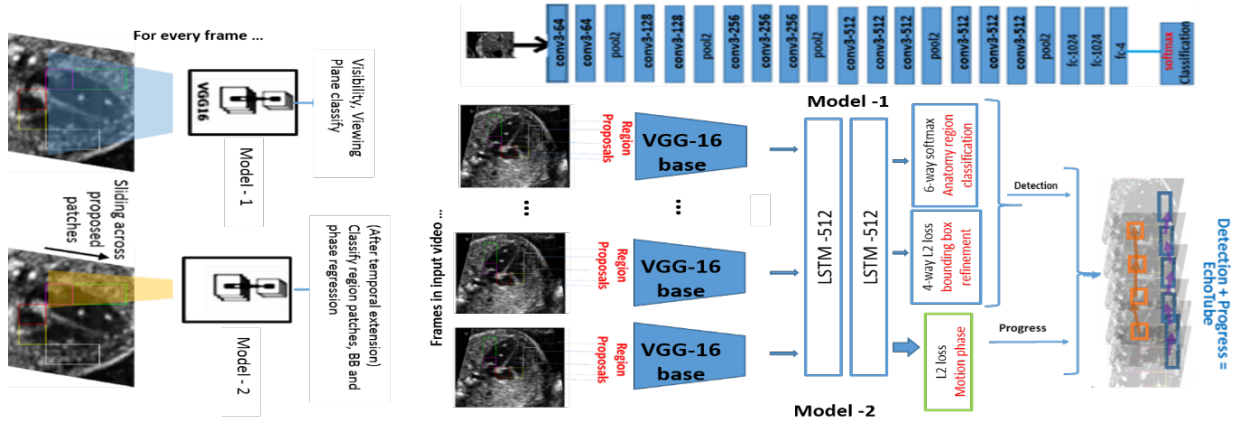


Figure 3.10: Detection and temporal progress pipeline for anatomical motion in fetal ultrasound videos.

3.6.3 Convolutional Architectures

First, it is important to ascertain the presence or absence of the heart. For this, Model-1 was trained on the dataset described in section 1.2.1 consisting of annotated viewing plane classes (4C, 3V, LVOT) and background frames. Next, Model-2 was trained to detect the five anatomical structures in the 4C view videos. Such a detection can be treated as a multi-class classification problem for different patches from the image. These patches are treated as individual labelled images that can be fed to a convolutional architecture capable of a 6-way classification (five anatomical structures and background). This training resulted in a model capable of resolving input regions into one of the structure classes. CNNs used have 13 convolutional layers with ReLU activation and 3 x 3 filters for fine structure aggregation, and 2 x 2 maxpooling after 2nd, 4th, 7th, 10th and 13th convolutional layers, followed by 2 fully-connected layers of 1024 and 512 units. The last fully-connected layer for Model-1 has 4 units linked to a 4-way softmax classification for visibility and viewing plane estimation. The penultimate fully-connected layer of Model-2 is used to feed spatial information per frame to LSTM modules for accounting temporal information across multiple frames in final detection. The LSTM stages give way to a 6-way softmax for structure classification, and two regression L2 losses, one for motion phase estimation and other for bounding box localization refinement. This is the multi-task convolutional-LSTM with classification and regression nodes

for temporal detection and local motion estimation. The latter task of estimating local anatomical motion progress is treated as a local phase detection problem. The phase regression is trained on a quantity defined to be the ratio of relative frame location from the nearest systolic frame to the number of frames between the end-systole to end diastole interval. This value is computed for every structure per frame. While it may appear trivial considering that it is analogous to the cardiac phase itself, such an equivalence holds true only for healthy hearts but digresses in case of wall-motion or structural abnormalities and the difference in local and global motion progress or phase in that case serves as a vital symptomatic marker for disease. Also, the Model-2 is pre-trained on anatomical structure videos on a pure classification stage to pre-condition it with parameters that would allow optimized label assignment for input region proposals. While this may seem redundant given that a multi-task training is performed anyway with both classification, detection, bounding box regression and local estimation phase, the pre-training is found to allow for better convergence and minimize overfitting due to the complexity of the detection task involved.

For the pipeline, region proposals (‘patches’ and ‘region proposals’ are used interchangeably for this task) defined by bounding boxes need to be efficiently computed and fed to pre-trained networks that classify them for every frame in every video sequence input. At training time, fully-connected outputs for patches classified as of a particular class across frames is passed to the motion progress LSTM modules, with a metadata file storing frame numbers of patch origin, patch label, sequence number and progress annotations computed with Equation 1. The latter is used as ground truth for motion progress training. To avoid multiple patches being identified with the same class from a given frame, we use a constraint that a single frame can have only one instance per class of anatomical structures and accept the patch detected with highest confidence per class. During test, the detection pipeline is run for a new input sequence and FC layer values are input to motion progress module for final prediction of each structure.

3.6.4 Implementation

Data pre-processing

This part utilises the dataset that was introduced in Section 1.2.2. For training, we split videos of the 4C view into frames and applied data augmentation by performing a top-bottom flip, and a rotational augmentation using angular steps of 10 degrees per frame, excluding for flips. This was applied to patches for five different anatomies, which are extracted in 50 x 50, and 50 x 100 pixel resolutions from the frames and used to pre-train CNN-2. The original frames are used for the visibility and view classification network training (Model-1). Next, we chose 32 frames at a time in the sequence of their original occurrence and concatenated them into 32-frame videos (the starting frame is an end-diastole frame) and repeated this for augmented data. This resulted in 4760 clips (clips are ignored if sourced from frames adjacent to background or a view change). In this, the start and end-frames were identified for each of the anatomical structures considered, with the end-diastole frame being the initial starting point. Both models were trained with mini-batches of 25 clips and a learning rate of 0.01. The objective function was optimized using the Adam optimizer with a momentum of 0.8.

3.6.5 Results and Discussion

Results for the motion localization task are summarized in Table 2. We briefly discuss the choice of metrics for the motion categorization task (identified as Echo Tube in Table 2). One aspect of such a task is the fidelity of temporal anatomy localization. The other aspect is an accurate identification of motion/action bounds. In our task, the former is a result of the efficiency of the object detection step, with a proxy being the average detection accuracy for each anatomical class represented in the echo tube over all the parsed frames, averaged over test instances (**Accuracy** in Table 2). The second is a measure of the correctness of the start and endpoints. This is presented as a ‘frame-difference’ metric, averaged over test instances and termed as **aFD** (average frame difference), i.e., we establish the frame number of the source of a patch classified as end of an echo tube and compare with ground truth.

We compared our method with a global context provided by the fully-connected embeddings from the CNN in Model-1 fused with the pre-LSTM fully-connect layers of Model 2. Anatomical similarities in some cases, notably valves and ventricle walls, prove to be difficult to accurately distinguish. This is due to a remarkable similarity of appearance in the ultrasound videos used for the task, among other factors like the crop-sizes of the regions used for training and the contextual information encoded, lack of information about heart orientation and the influence of speckle, shadows and enhancements. For motion localization tasks, the frame difference metric for motion patterns of the foramen ovale and the valves are much lower when averaged for test cases, but the same is not true for the ventricles. This is attributed to the relative inefficiency of using higher-level feature representations to track progress through a regression-based loss when the start and end morphologies are not particularly distinct over adjacent frames. A solution to this might be to use secondary sources of motion patterns, for instance, speckle-tracking. It is worth noting that the efficiency of motion/action detection as modelled using an overall average of per-video framewise average detection accuracies is not a particularly reliable localization performance indicator. This is because object detection is generally more accurate for well-defined patterns and localization relies on frame-to-frame distinction for similar spatial locations [91]. The two factors may not be necessarily correlated over long-interval video sequences despite the physiologically imposed periodicity. In Model-2, the spatial context helps with structure localization and a secondary temporal motion information is introduced through global motion patterns averaged through the recurrent LSTM steps. This is physiologically compatible as local anatomical motion is influenced by overall cardiac cycles. The method is found to be performing relatively well for foramen ovale motion categorization. This is an important fetal heart biomarker as it is the opening connecting the left and right atrium and plays a key role in channelizing oxygenated blood into the developing fetus. The motion of ventricle walls are localized and the extent over lengths relative to heartbeat intervals are vital clinical indicators for conditions of the heart and is an important measure of cardiac function.

Overall, in the second part of the chapter we presented a deep learning pipeline for quantifying motion patterns for multiple anatomical landmarks in the fetal heart and integrated the automated detection and motion progress which can form the basis for the analysis of diagnostically relevant functional motions in fetal echocardiography videos, as tested on real world clinical data. Though motivated by fetal echocardiography analysis, this method is transferrable to motion analysis in other ultrasound tasks (or to other modalities with video data) and does not require optical flow information.

Table 2. Motion/Action detection and localization performances over anatomies of interest in the 4 Chamber view.

Anatomy	Adapted two-stream Net [54]		EchoTube (ours)		
	Accuracy(%)	Bounds aFD	Accuracy(%)	Bounds aFD	
<i>Foramen ovale</i>	0.9375	2.5	0.9415	2.5	
<i>Mitral valve</i>	0.7185	2.3	0.7750	2.2	
<i>Tricuspid valve</i>	0.7300	3.2	0.7424	3.3	
<i>LV wall</i>	0.8232	5.8	0.8245	4.6	
<i>RV wall</i>	0.8016	4.2	0.8320	4.1	
<i>Overall</i>	0.8022	3.6	0.8231	3.34	

3.7 Conclusion

The chapter explored different facets of analysis of fetal cardiac ultrasound videos using a range of design choices and protocols derived from deep learning based methods by exploring the spatial and temporal dynamics in such videos. The video analysis pipelines so proposed seek to mimic steps of clinical analysis of ultrasound videos of the fetal heart, namely the initial steps of ascertaining the visibility of the cardiac region and locating the standard cardiac viewing planes, followed by localization of the heart by regressing to the centre of the heart and the radius and finally the motion analysis of individual structures in the four-chamber view. Several directions are possible with such an introductory treatment of automated analysis of fetal echocardiography videos, including extensions towards replicating the anatomical motion analysis in other standard viewing planes, locating non-standard planes, analyzing cases with congenital anomalies and so on. That said, fundamental questions on the video analysis methods, including their reliability over

the input data streaming characteristics need to be ascertained for the challenging clinical tasks described above. To that end, the next chapter will analyze the feasibility of the deep learning based architectures in incrementally learning from and adapting to variable data distributions in non-deterministic temporal windows.

*Alles Gescheite ist schon gedacht worden.
 Man muss nur versuchen, es noch einmal zu denken.*

*All intelligent thoughts have already been thought;
 what is necessary is only to try to think them again.*

— Johann Wolfgang von Goethe

4

Continually learning from fetal ultrasound videos

Contents

4.1	Motivation	77
4.2	Lifelong, Continual or Incremental Learning	79
4.3	Proposed Contributions	81
4.4	Problem definition	83
4.5	Hierarchical Class Incremental Learning	84
4.5.1	Network Architecture Details	87
4.6	Experiments	94
4.6.1	Pre-processing	94
4.6.2	Baselines	96
4.6.3	Early training	97
4.6.4	Incremental training	98
4.6.5	Evaluation protocols	99
4.6.6	Incremental computational and memory overheads . . .	100
4.7	Results and Discussions	101
4.7.1	Incremental Performance	101
4.7.2	Forgetting and retention.	103
4.7.3	Effects of Imbalance.	105
4.8	Conclusion	106

The previous chapter discussed the first steps towards automatically interpreting a fetal cardiac ultrasound video sequence by identifying the standard cardiac viewing planes before progressing on to localizing anatomical structures apparent on such

planes during the cardiac cycle. This chapter now introduces the notion of such automated understanding using neural networks being extended over multiple learning sessions, with a different set of input data distributions and without a necessary assumption of balanced datasets being available for all anatomical labels to be learnt. While ultrasound image and video analysis research has seen extensive use of deep neural networks, learning over multiple tasks is a challenge for connectionist networks due to tendencies of degradation in performance over old tasks while adapting to novel tasks [96] [98]. It is pertinent that adaptations to new data distributions over time are tractable with automated analysis methods as medical imaging data acquisition is typically not a static problem. So, one needs to ensure that a continual learning paradigm is implemented in machine learning methods deployed for medical imaging. Specifically, this chapter explores the idea of incremental learning for anatomical structures apparent in fetal echocardiography videos and proposes different strategies for continually learning new structures under different regimes of imbalance along with metrics for assessing the performance of such incremental learners. Approaches for such learning situations with a hierarchical classification strategy is explored in this chapter and with an exemplar rehearsal scheme are explored in the next. Note that the terms ‘incremental learning’, ‘continual learning’ and ‘lifelong learning’ are used interchangeably to refer to learning algorithms over multiple sessions with different data distributions, as there has not been a standardization of the definitions each of these terms may carry in machine learning literature or elsewhere. The research presented in this chapter has been accepted as a journal submission:

1. Arijit Patra, Alison Noble, ‘Hierarchical Class Incremental Learning of Anatomical Structures in Fetal Echocardiography Videos’, *Journal of Biomedical Health Informatics* (accepted).

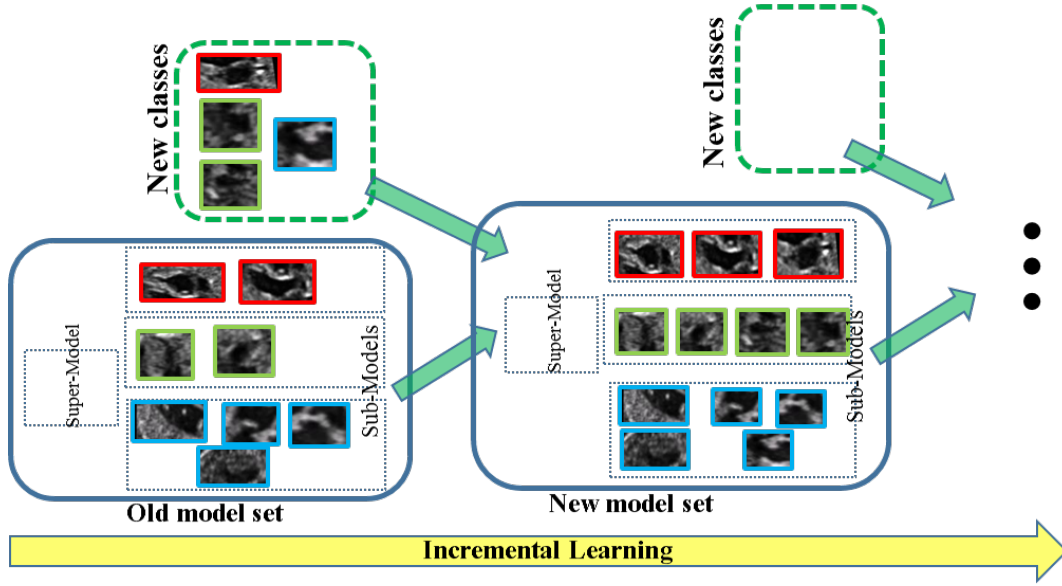


Figure 4.1: The Incremental learning paradigm over a hierarchy of learnable features. Previously trained models need to adapt to new class data without compromising performance on previously seen class data.

4.1 Motivation

As discussed in earlier sections, the scope for ultrasound video analysis with deep neural network approaches has shown significant promise. Adaptation of deep learning methods, such as convolutional and recurrent networks, have achieved state-of-the-art results for multiple medical image analysis tasks. Often a requirement for deep networks is the availability of large labeled datasets at the outset. However, in medical imaging applications data is often not available in large volumes or access to large patient datasets is restricted by hospital data governance policies. This poses a problem in medical imaging where most current deep-learning based models need large representative datasets to accommodate intra-patient variations, physiological differences due to disease, different acquisition methods and so on.

The storage of a range of medical imaging data is rendered difficult by legal complications brought on by privacy regulation, and with growing concerns around cybersecurity aspects of data centres. This is at odds with the fact that the diversity of physiological conditions on imaging data streams requires a continuous improvement in model performance for machine learning systems [138]. Under these

conditions, a spectre of decline in performance on previously learnt tasks is not acceptable from a reliability perspective. Currently, a limited number of machine learning based applications have been approved by regulators such as the US Food and Drug Administration [139], for a specific number of use cases such as assessment of diabetic retinopathy. These algorithms are locked in for specific applications. However, there is a case to be made for adaptive learning regimes for algorithms to be continuously upgraded to perform new tasks much like human clinicians. While an upper bound on performance could be maintained over incremental arrivals of data by simply retaining all the data indefinitely, such a retention of data itself may not be suitable in the light of privacy concerns and consent of use policies, apart from constraints on digital storage infrastructure. Although the latter can be addressed with cloud infrastructure unless constrained by local information risk management policies, the former concerns around informed consent at the subject-level and ensuring compliance to data protection laws may still prove to be a substantial effort. This lays the foundation for queries around learning over incremental streams of data, while not having to store past data and still retaining acceptable performance guarantees through the service life of the system. As such, the exact regulatory contours around model adaptation to data acquisition over extended time are yet to be formalised in many countries, to my knowledge, as practical systems still restrict themselves to single task static models for now. For most part, the legal standards for continual learning approaches are likely to derive from regulations such as Health Insurance Portability and Accountability Act (HIPAA) 1996 and their equivalents, which propose stringent compliance measures for information governance for clinical datasets [138]. Legislation around adaptation to continual arrival of data, either in a streaming or an offline fashion, is likely to evolve as the deployment of machine learning systems require increasing adaptation to data from different sites, over extended times and for compliance to emerging diagnostic requirements [139].

4.2 Lifelong, Continual or Incremental Learning

In machine learning, Lifelong Learning is a computational approach that uses previously learned information to acquire new knowledge, using the idea that learning the n^{th} task should be easier than learning the $(n-1)^{th}$ task [96]. Animal brains have evolved to retain significant amounts of information that are not compromised when learning new information in the form of novel tasks or experiences. While it is true that forgetting of some information happens naturally due to ageing, the process of memory loss in human brains occurs over extended periods of time rather than a sudden decline in knowledge retention of immediate previous learnt knowledge or acquired experiences. This compelling intuition has proved difficult to replicate in artificial learning algorithms reliant on connectionist models, where the representations optimized for with a certain input distribution stands susceptible to be affected in the event of further fine-tuning on a different data distribution over extended time. The issue of not retaining prior learnt representations to a reasonable degree has been studied for multiple neural network architectures, under an umbrella term of 'catastrophic forgetting' [98] in machine learning research. The previous and new knowledge can be in the form of old and new tasks or could be old classes and new classes within the same task. Under assumptions of such new knowledge being incrementally added as new classes, the lifelong learning problem is essentially a class incremental learning problem. In this situation, classes arrive and the trained model's knowledge is expanded in breadth by one class at a time.

How might lifelong learning be relevant to image-based diagnosis? In imaging-based diagnosis, available clinical datasets often contain a small number of samples of disease cases versus healthy cases (the baseline). Further, in the dynamic imaging case we consider in this chapter, information about a condition may appear over time as a sonographer scans multiple times over an organ. In both cases, lifelong learning needs to address the problem of class imbalance over all incremental learning stages without prior information on future data distributions. Achieving robustness to variations in data distribution over new unknown classes in undefined timescales is a non-trivial problem. Addressing such challenges for clinical diagnostic videos

requires deep learning based analysis to be amenable to successive learning from small datasets of future classes. This needs to happen without overfitting on new examples or losing the ability to generalize over past information. Availability of past datasets may not be necessarily assumed either. Additionally, the success of such continual learning approaches is related to overall model explainability over the multiple learning sessions the model is trained over. Therefore, it also makes sense to explore the possibility of including features such as saliency as a part of the design process for incremental learning algorithms.

To sum up, Incremental Learning here aims to enhance the representation space of a trained model by augmenting the number of classes recognized by it over time. A key bottleneck is the question of storage of data of current classes while training for new classes. Storage of old data indefinitely is impractical as it would increase the memory requirements at every incremental step. An ideal incremental learner for medical imaging should have the following characteristics – i) the model performance on new classes should be improved on account of prior training on old classes (forward transfer of knowledge), ii) accuracy on previous classes should not show significant decline upon training for new classes to be recognized, iii) the size of the models should not grow despite the number of classes seen, iv) the model should be able to handle incremental class imbalance where arriving new classes may have different dataset sizes compared to previously seen classes.

In classification tasks, incremental learning is supposed to increase the breadth of a classifier by training it to recognize new classes of objects, while retaining the learned knowledge of classes trained originally. The evaluation of the incremental learning pipeline is assessed either in terms of separate evaluation on the old and the new classes or tasks, or by a joint evaluation of the validation sets of both the old and new tasks. The former was termed as multi-headed evaluation and the latter is single-headed evaluation in computer vision applications [140]. Multi-headed evaluation tends to be easier since the model would not have to classify across old and new classes at the same time on a given input instance, whereas in single headed evaluation the decision on the input instance requires a classification over

all the labels seen across old and new tasks. Multi-headed evaluation is typically used in task incremental learning scenarios where it is not important that classes seen in both tasks are jointly encountered. This is inherently not the case for medical image analysis where classes seen in different incremental stages can be encountered together or simultaneously. Thus, the incremental learning pipelines designed for our class incremental ultrasound tasks should be subjected to a single-headed evaluation strategy, where the validation sets of the old and new classes need to be evaluated jointly. In the experiments in this chapter, the validation accuracies quoted are based on a single-headed strategy.

4.3 Proposed Contributions

In this chapter, the issue of catastrophic forgetting [98] in ultrasound image analysis is introduced and a set of solutions to ameliorate such loss of performance proposed. First, we propose to leverage hierarchical feature learning in convolutional networks in an incremental learning algorithm that aggregates a temporally-spaced inflow of new information and learns novel classes of data in a self-organizing fashion without significantly altering learned features from previous datasets. Second, we extend the hierarchical incremental learning pipeline to few-shot contexts where new class data arrives in sets with the number of exemplars much lesser than the initial classes the model is trained on. In both cases, we present a proof-of-concept on fetal echocardiography videos. Fetal echocardiography assessment in the 4C (four chamber) and non-4C view [8] requires the identification of a number of anatomical sub-structures. Such a multi-object classification task can be difficult in ultrasound videos and may require visualization over multiple scans. For instance, a sonographer may choose to visualize some sub-structures at higher resolution i.e. zoom-in, in a later part of a full-scan video or may find a better acoustic window in which to image the object at a later point of the full scan. While a comprehensive evaluation of the fetal heart during anomaly scanning requires a comprehensive view of the structures and physiological motion over a number of frames by the sonographer, a first step for a deep learning model to aid in the

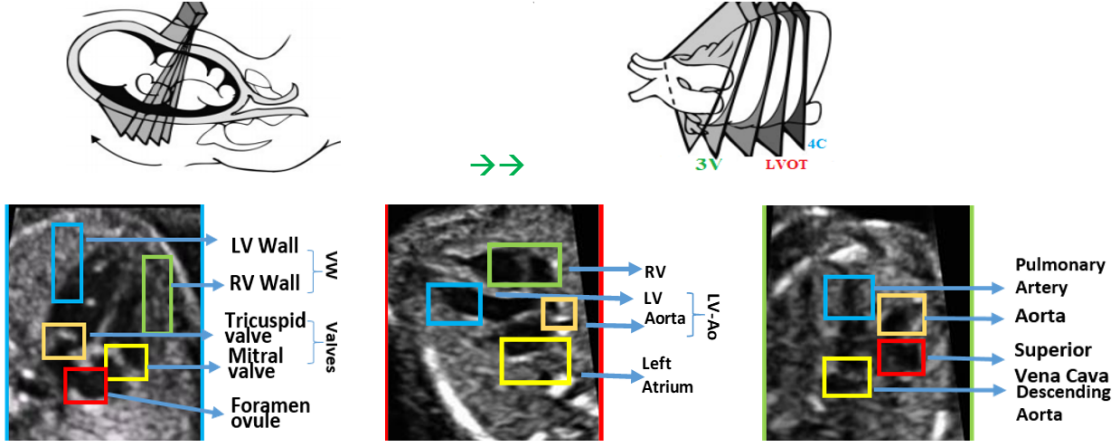


Figure 4.2: The ISUOG scanning protocols (top) for LVOT, 3V and 4C view assessments and anatomical structures in the three standard viewing planes. The anatomical structures visible in these standard planes (4C, LVOT, 3V standard planes, left to right) are considered for the class incremental learning problem.

process is to be able to identify individual structures of interest within the whole frame including the heart region both in still images and in video clips. In this case, the desire would be to fuse information from all video clips of the object to build a computational representation of the fetal heart.

Concretely, this chapter explores three directions in the attempt to propose continual learning for fetal ultrasound imaging:

- 1) Learning incrementally over multiple sessions with the introductions of new classes with dataset volumes in the vicinity of previously seen classes.
- 2) Learning incrementally over multiple sessions with new class dataset volumes being significantly reduced compared to prior classes.
- 3) Introduction of evaluation protocols and metrics for class incremental learning for ultrasound image analysis.

The chapter is structured in the order of the problem definition followed by the description of data preprocessing. This is followed by sections corresponding to the hierarchical approaches to incremental learning under the two extremes of data distributions. In these sections, the description of the underlying theoretical explanations are accompanied by implementation details and discussions on observed

results on the clinical fetal echocardiography dataset, annotated with the aid of certified sonographers associated with the Oxford University Hospitals.

4.4 Problem definition

We model fetal echocardiography dynamics at any instance based on currently available streams of data. The level of detail in the model is incrementally learned in a manner that does not significantly compromise prior model knowledge, in the two proposed experiments. Recall that in our fetal echocardiography videos, the fetal heart is visualized in three viewing planes – four chamber (4C), three-vessel (3V) and left-ventricular outflow tract (LVOT) [6]. Certain anatomical structures of the fetal heart can be visualized from one or more of these viewing planes. The 4C view includes structures such as the ventricles (their walls can be distinctly observed in the 4C view), the mitral and the tricuspid valves and the foramen ovale (Figure 4.2). The 3V view typically includes the pulmonary artery, the aorta, and the superior vena cava as the key informative structures carrying markers for physiological assessment [8]. The LVOT standard plane includes the left and right ventricles, the aorta (in our videos, the LV and the aorta are observed as a continuous cavity and labeled as LV-Ao for convenience) and the right atrium. Our problem deals with a class-incremental setting for detecting fetal cardiac structures: Ventricle Walls (VW), Foramen ovale (FO), mitral and tricuspid valves (4C view structures), left atrium (LA), right ventricles (RV), Left Ventricle and aorta (labeled as LV-Ao hereafter) and right atrium (LVOT view structures), Pulmonary Artery (PA), Aorta, Superior Vena Cava (SVC) (3V view structures). Out of these, structures are learnt in sets first as base categories in the initial task, followed by incremental task sessions where each of the new classes are introduced one by one. These are shown to the base-trained model in incremental stages in the class incremental learning experiments. The base training is performed on six classes across standard planes (VW, Valves, FO, RV, PA, Aorta) and the class incremental training is performed on three classes, one from each of the standard planes (LV-Ao, SVC, LA). Out of the latter class group, each class is sequentially introduced in

successive task sessions. This simulates a setting where the algorithm needs to adapt to new data distributions in the absence of a majority of exemplars from past distributions. The datasets used are the same as the ones described in Sections 1.2.1 and 1.2.2. The data sourced from 10 subjects is used for training and that from the remaining 2 subjects for testing.

4.5 Hierarchical Class Incremental Learning

In this section, we approach the problem of incrementally adapting to fetal cardiac structures by exploiting the hierarchy of features in the different standard viewing planes. Co-located structures in an anatomical region often have feature similarities suitable for hierarchical grouping and distinction from representations obtained from a different part of the fetus. We define two network models with different fine classification strategies representative of extreme situations of no class imbalance over new additions and highly imbalanced arrivals of new classes. We also define two experiments utilizing different network topologies conditioned for these two respective distributions handled (Figure 4.4).

Why does a hierarchy of features perspective into the incremental learning problem make sense? The convolutional neural networks used for most deep learning based methods are known to abstract information over a sequence of layers going from coarse features defining the input to more fine-grained descriptive features as one goes deeper and the network [49]. This is indeed the insight that informed the video analysis algorithms proposed in Chapter 3. Building on the same idea that was discussed there, we make the case that if multiple labels are possible for a set of images with one label being a coarse top-level description and the others being progressively more descriptive and going into sub-hierarchies, the hierarchical nature of learning in deep neural networks can provide predictions on these levels of classes at increasing depths. Therefore, a shallower CNN could be used to arrive at a coarse class description as early layers learn coarser descriptions such as edges and contours, which can be sufficient to differentiate between different coarse labels. Such a description can then be used to call a model that is an 'expert' on the finer

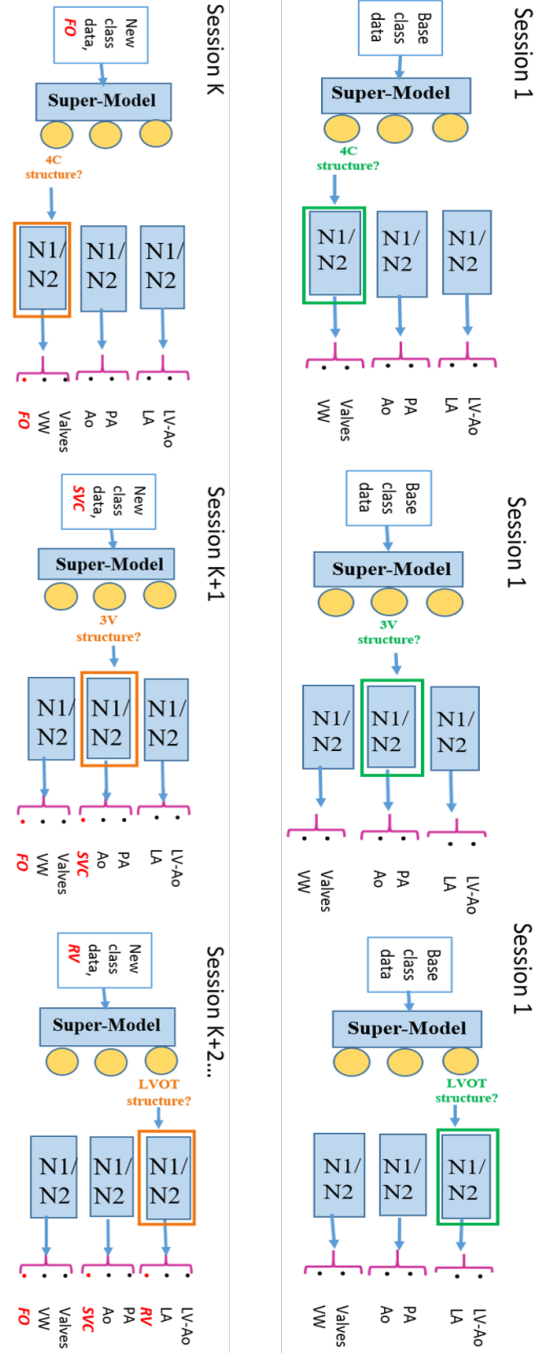


Figure 4.3: Overall workflow over multiple learning sessions (K is the number of sessions the model learns over). After initial training on base classes in Session 1, the model decides on the top-level class and directs to the corresponding sub-model for fine classification. Session K (and $K+1, 2, \dots$; $K=2$ in our tasks) is the learning session where the pipeline is shown previously unseen classes – FO, SVC and RV here. Top-level classification decision on previously unseen class data is taken by the super-model and the corresponding sub-model is adapted for the new data arrival. N1 is used as the sub-model configuration for Experiment 1 with balanced new class arrivals. N2 is used as the sub-model for few-shot new class arrivals in Experiment 2. Note the newly added classes are shown in red.

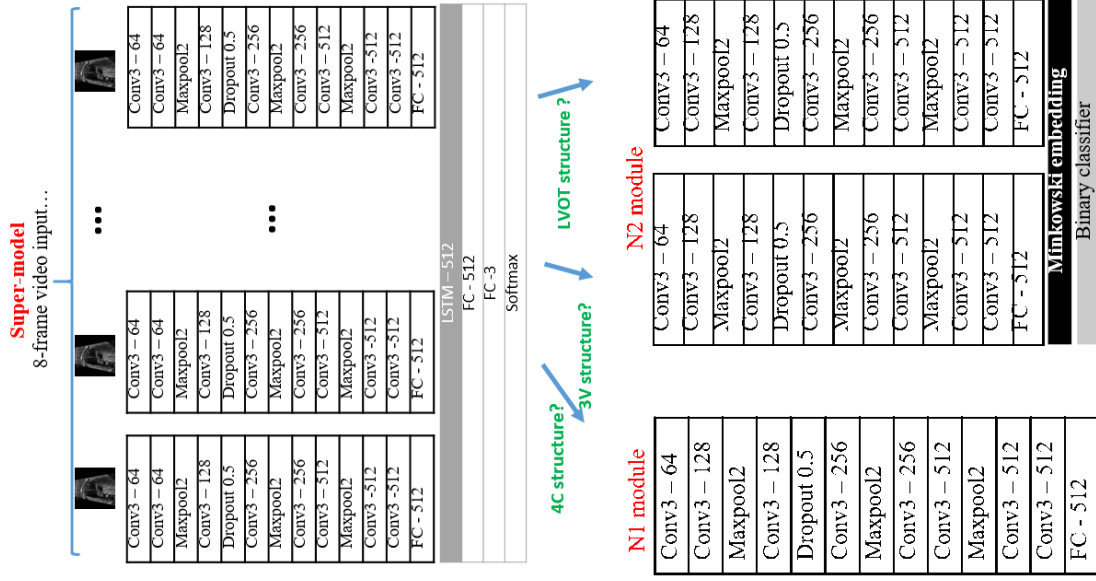


Figure 4.4: Super-model for coarse classification. N1 and N2 modules.

classes covered under a particular coarse label. This model, later termed here as a sub-model, is trained on identification of fine class categories [141].

Experiment 1. We consider the cases of incremental distributions with dataset sizes of novel classes approximately resembling those of the trained prior classes. Here we use a hierarchy of models with softmax probability-based self-organization throughout. This pipeline is termed Network 1.

Experiment 2. We consider the case of dataset sizes being imbalanced in novel classes as compared to prior learned distributions. We use a second hierarchy of models with the coarse classification stage employing a softmax classification approach to direct instances to sub-models. This adopts a similarity driven few-shot learning regime in a manner similar to metric learning approaches [103]. This pipeline is called Network 2.

Both Network 1 and Network 2 use the same pre-trained model for coarse-classifier stages as this is implemented as a common softmax probability based classifier. While actual distributions could lie in-between these extremes in incremental stages, such a distribution can be attempted in both strategies. Thus, in real world use-cases, the choice of architecture can be dictated by specific performances on available distributions.

4.5.1 Network Architecture Details

In this sub-section we define the architecture design components and discuss other aspects of design. First, we define two model types:

Super-model – A coarse classification stage, with a model implementing convolutional layers leading to a time-marching recurrent stage for taking video sequence input and classifying into pre-defined top-level classes. The softmax probabilities, averaged over a batch of inputs is used to determine the coarse or top-level class. This is different from standard classification pipelines where each instance is classified. While such information is available here, it is of secondary importance since the final outputs are the fine classes, and so the flow of control to the appropriate sub-model is prioritized).

Sub-model - Fine classification is implemented by a set of sub-models representing each coarse class the super-model can possibly generate. The models are pre-trained for classification (Network 1, with convolutional – fully connected – softmax architectures as N1 modules) or similar-class distance minimization (Network 2, with a Siamese architecture as N2 modules) on historically available categories.

We implement a conditional hierarchy of deep learning models, trained separately but working as a multi-stage pipeline. The flow of information is conditioned by top-level decision inputs based on the softmax probabilities or distance metrics depending on the stage of classification and specific implementation. The idea rests on a sequence of convolutional layers learning in terms of a hierarchy of features. This allows the learning routine firstly to learn super-classes (coarse classifier stage) based on general sets of features and to use that prior in a second learning step for detailed anatomical structure classification. The process essentially exploits the natural hierarchy of features apparent in fetal echocardiography video. In the limit of data distributions at the coarse classifier stage belonging to a single class and the original data being available in its entirety, the problem is reduced to a standard transfer learning problem where the addition of the incremental class assumes a new task similar in the same pattern to the old task.

Super-model: In both experiments, we first implement a convolutional network, inclusive of LSTM layers for temporal processing, with a 3-way softmax cross-entropy classifier to identify standard fetal heart views – *4C structures*, *3V structures*, and *LVOT structures*. This model essentially performs coarse classification. It consists of multiple convolutional layers, followed by LSTM layers and a final softmax classifier stage. The super-model places input data into one of the coarse classes and provides a confidence score per class based on which subsequent sub-models are called for fine classification.

The super-model is implemented as an architecture with eight convolutional layers with 3 x 3 filters accompanied by ReLU activation and max-pooling with 2 x 2 pooling windows and a dropout of 0.5 after the 3rd convolutional layer to avoid overfitting. The convolutional layers are followed by a single fully-connected layer of 512 units feeding into an LSTM layer of 512 recurrent units linking to two 512-way fully-connected layer linked to a N -way fully-connected layer feeding to a N -way softmax activation, where N is the number of super-classes ($N=3$ in our case). The model processes 8-frame video inputs frame by frame and passes the first fully-connected output to the LSTM layer to aggregate the video temporal dynamics. The model layers are initialized in a Xavier-improved scheme [142].

Sub-models: Sub-models are implemented per class of the super-model and initially trained independently of the super-model. There are two kinds of implementations for the two strategies explored, described in (Figure 4.4) as $N1$ and $N2$ modules. For *Network 1*, the sub-models deal with roughly similar dataset volumes as compared to the previous training data. Each $N1$ module is designed using five convolutional layers with alternate max-pooling followed by two fully-connected layers and a softmax classification layer comprises the fine classifier. The $N1$ module is replicated thrice into sub-models A, B and C corresponding to each of the three coarse classes. This process inherently assumes a condition of approximately balanced datasets and in situations of few-shot incremental training is sub-optimal as a symmetry driven augmentation is inadequate for complex ultrasound videos where overfitting results. This motivates *Experiment 2* and the design of *Network 2*.

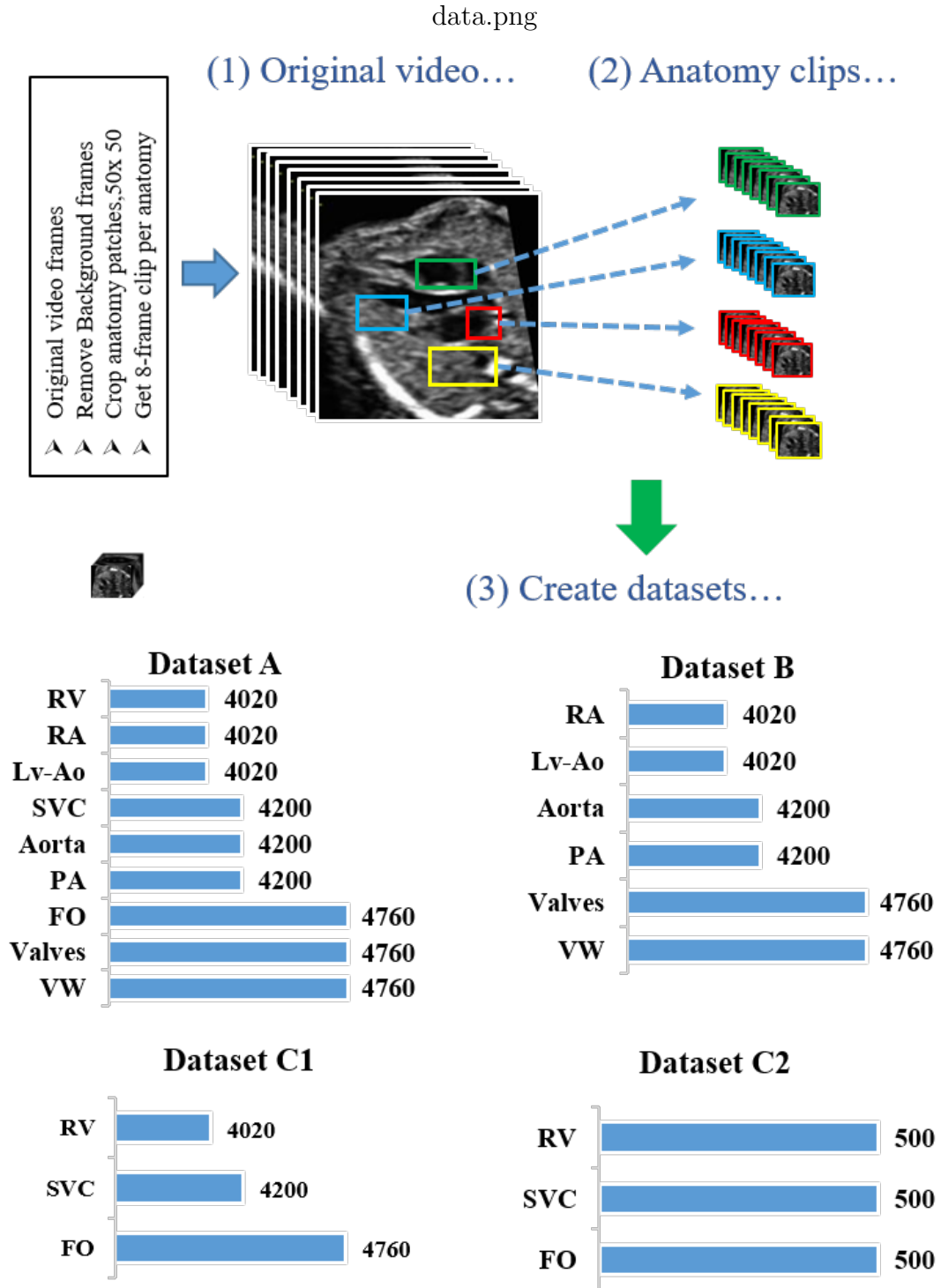


Figure 4.5: Data preparation from original video to clip generation and dataset aggregation.

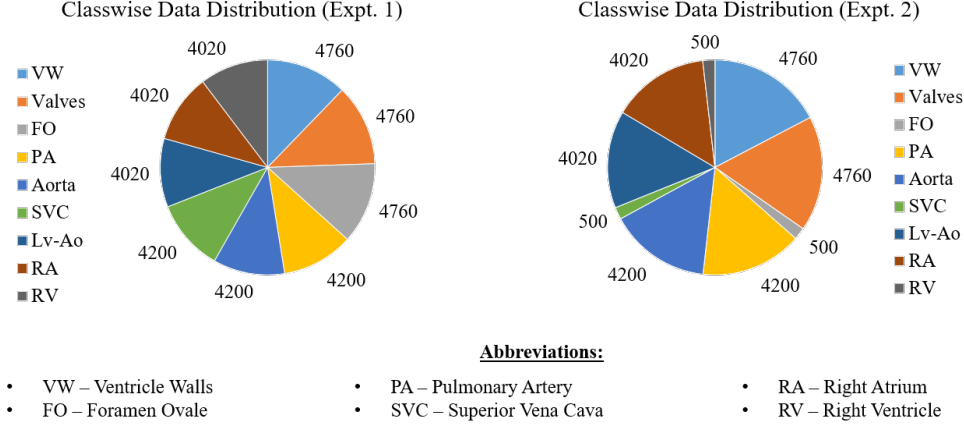


Figure 4.6: Data distributions simulated for base and incremental classes in the two experimental configurations described. Experiment 2 uses Network 2 with N2 modules for sub-models (old classes already trained on labeled in grey, new classes to be incrementally trained for labeled in red).

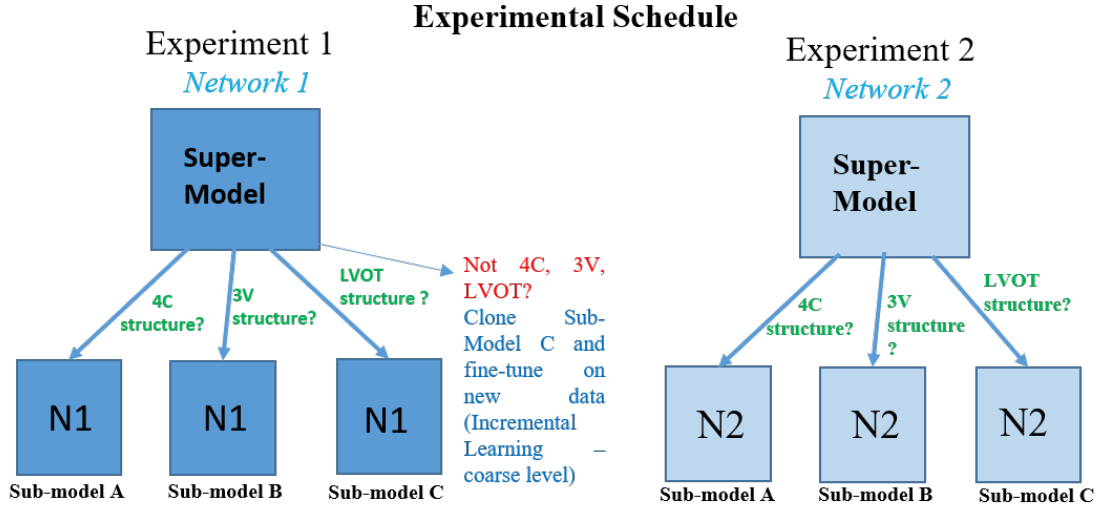


Figure 4.7: Experiments 1 and 2 with their data distribution situations. Experiment 1 uses Network 1 with N1 modules for sub-models.

In *Network 2*, the sub-models learn to discriminate between similar and different images using a Siamese Network [102] with a binary classifier. This is preferred over a direct softmax classification in order to optimize for a distance measure between images of similar and different structures. This is useful if we add to existing datasets a few examples of a novel class and want our classification models to be able to detect instances from this class without a need of retraining with augmentation or generative modelling. Using a training regime with a softmax

formulation in this case would suffer from the class imbalance problem of not being able to learn enough discriminative features with only a few examples of a new class, and not being able to predict optimally in the presence of such instances.

Using a distance metric that has been explicitly trained to be minimized for similar images and to be maximized otherwise is useful in this setting. This is because one only needs to compare an input image with representative examples of classes previously trained for and a threshold applied to distances computed can be used to categorize if a given example belongs to a pre-existing class. In models with a binary cross-entropy loss after distance layers, the same is possible from class confidence score outputs of the softmax function. This enables extension to novel classes of anatomies, as an image of such a structure is likely to be structurally different and encode different feature representations than those of classes already trained on. One can predict if an input instance belongs to a known set of classes or of a different unseen class based on the distance between the lower dimensional representations obtained for instances of these different images through the trained model. Over multiple such examples, it is possible to train a distance minimization for the same class exemplars, either by fine-tuning the full sub-model or the higher layers alone (feature extraction) while training for the intra-class distance minimization, without need of a large number of instances in the limit of small learning rates. Incremental learning can be boosted by fine-tuning the relevant sub-model to minimize distances for a novel class using its examples or by creating a new clone of a sub-model if softmax output from super-model is below a set threshold, in a case of a novel coarse class (this situation does not arise for the fetal echocardiography tasks described because the standard viewing planes where the cardiac structures are observed are predefined).

We train three $N2$ based sub-models (A, B, C) relevant to each coarse class. They are modelled as Siamese Networks [102] taking multiple-frame inputs. The strands process each input stream and optimize a distance metric by learning to classify ‘same’ and ‘different’ classes. The model has eight convolutional layers with 3×3 filters, ReLU activation and alternate max-pooling. The first layer is configured

to accept a sequence of eight images to aggregate temporal information directly similar to the Hierarchical Temporal Encoding design proposed in Chapter 3. After the convolutional layers is a 1024-way fully-connected layer that enables encoding of feature representations for distance computation. Each input is separately processed using shared weights, leading to two feature representations from each input pair. We encode similarity by a joint embedding function inspired from the Minkowski distance formulations expressed as:

$$\sum_i (|x_i - y_i|^p)^{1/p}, p \in R^+ \quad (4.1)$$

Following the generalized Minkowski formulation in equation 4.1, the embedding function used here is,

$$\sum_i (|f(x)_i - f(y)_i|^1) + (|f(x)_i - f(y)_i|^2) + (|f(x)_i - f(y)_i|^3) \quad (4.2)$$

where $f(x_i)$ and $f(y_i)$ are feature embeddings from instances compared. Instead of calling equation 4.2 a distance metric, we term it an embedding function because the function is not strictly a distance measure as the formula is an additive combination of different distance measures with different powers (here, a combination of a Manhattan, Euclidean and Minkowski distance measure). While most prior work in metric learning and related areas like few-shot learning has typically used $l1$ and $l2$ distances, we use a customized embedding function to leverage the numerical magnifications achieved with powers over absolute differences. This helps obtain a more reliable representation of similarity of features in our datasets of anatomical structures from ultrasound videos where structural anatomical definitions are influenced by the presence of speckle, enhancements and noise. Higher values of p enable a magnified representation of distances between feature representations and discriminative dissimilarities.

The value of p is a hyperparameter and we assume $p = 3$, by performing a grid search over integral values. The design choice of $p \in Z^+$ is assumed as non-integral values as number of metric distances added to form the Minkowski embedding is equal to p . We selected a value $p = 3$, as the improvement in performance was

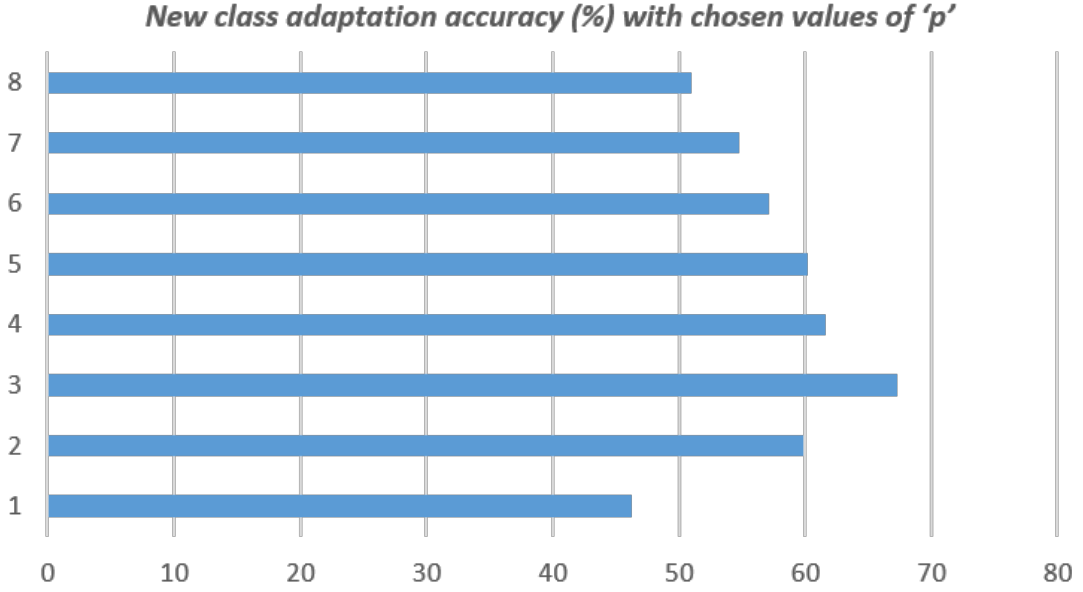


Figure 4.8: To choose the value for hyperparameter p , a grid search over integral values is performed and the metric monitored was the new class adaptation accuracy.

negligible. Distance between two representations from fully-connected layers is computed and constrained between 0 and 1 for classification to ‘same’ and ‘different’ classes optimized by logistic regression,

$$t \cdot \log(p(x_1 \circ x_2)) + (1 - t) \cdot \log(1 - p(x_1 \circ x_2)) \quad (4.3)$$

with $p(x_1 \circ x_2)$ (set equal to the normalized Minkowski embedding value obtained between feature representations of x_1 and x_2) as the probability that x_1 and x_2 are of the same class. Then, using equation 4.3 one obtains $t = 1$ if instances belong to the same class and $t = 0$ otherwise.

In both the variants of networks, the depths have been so chosen as to keep the overall numbers of parameters involved in a single forward pass in the same order or magnitude as would be in the baselines considered. This is to ensure that the order of parameterization has a minimal role in determining knowledge retention or forgetting. Parameters for temporal extension are delegated to the LSTM stages where feature representations across spatial frames are aggregated in a temporal dimension and final representations can be obtained thereafter. Again the choice

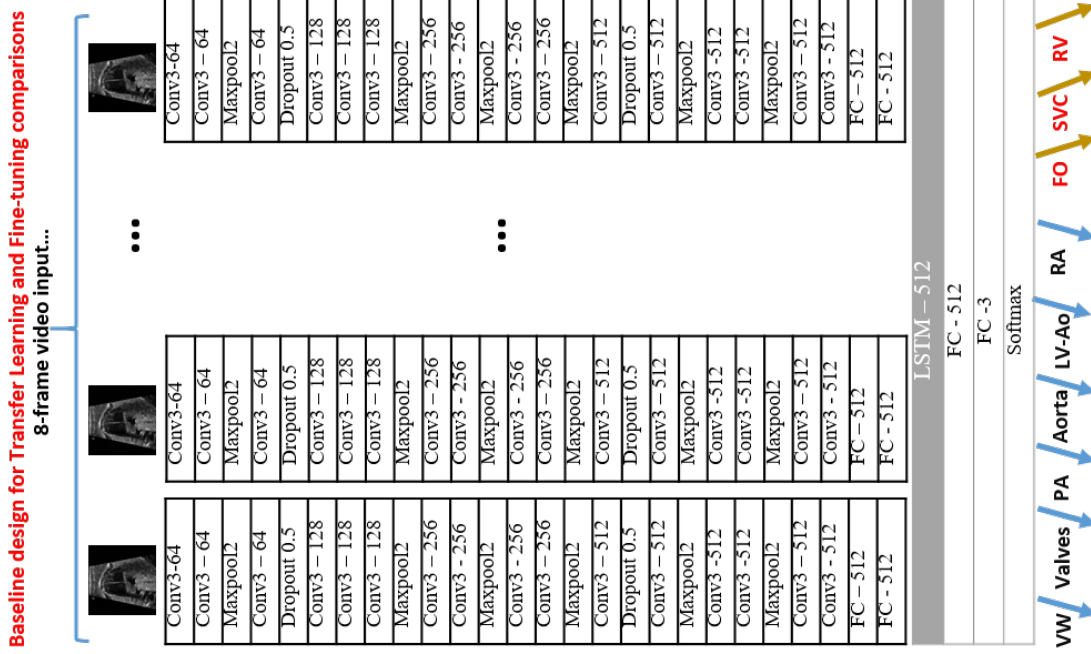


Figure 4.9: Baseline architecture for benchmarking in hierarchical class incremental learning

of the number of units is dictated by the requirement to constrain the order of parameterization to be within baseline parameter dimensionality.

4.6 Experiments

4.6.1 Pre-processing

We split the videos into frames and apply data augmentation. Resizing frames to 400 x 400, we extract relevant anatomical patches in resolutions of 100 x 100, 50 x 50 or 50 x 100 depending on the specific anatomical structure 4.2. These patches are later resized to 100 x 100 prior to model input stages in order to standardize input size between initial and incremental stages (note that the cropping and resizing are separately performed, with the latter becoming relevant only before model inputs). The patches are grouped into 8-frame clips in order of original occurrence in cardiac videos. This provides us 4760 8-frame clips of 4C view sub-anatomies (ventricle walls, valves, foramen ovale), 4200 clips of 3V view and its sub-anatomies and 4020

clips of LVOT view and sub-anatomies. Each video so obtained is labeled with the cardiac view it originated from and the structure it is. We group together videos of anatomical structures belonging to the 4C view and label it '*4C structures*', and repeat for clips from 3V and LVOT structures. The set of videos is accumulated with all the labeled sub-anatomies and called Dataset A. Next, we create a dataset with fine labels, called Dataset B. It comprises only ventricle walls and valves from the 4C structures set, pulmonary artery and aorta from 3V set and LV-Ao continuum and right atrium from LVOT set. The foramen ovale clips from 4C view, superior vena cava from 3V and right ventricles from LVOT are left for incremental learning. Dataset C has all incrementally added fine structures with their fine labels. C1 has novel fine structures in proportions similar to the old classes. C2 has 500 instances per new class. C1 and C2 are derived from C to simulate experimental conditions of standard incremental versus few-shot incremental learning and as such, dataset C is not the union of C1 and C2 but a parent dataset. Overall, Dataset A is the master dataset (used to derive test instances for assessing knowledge retention and overall drop in accuracies/ degree of forgetting), Dataset B is used for initial training, C1 for incremental training in Experiment 1 and C2 for the same in Experiment 2. C1 and C2 are used for incremental fine-tuning in baseline architectures. The data preparation strategy and respective distributions for the different sets of experiments are elaborated in Figure 4.7.

It is to be noted here that the incremental learning processes studied here work with the particular anatomical structures apparent in the standard cardiac viewing planes. In clinical workflows, an anatomical structure may be brought into the focus of the investigation but an explicit extraction of anatomical regions as patches is rarely performed. The step extending from the frame level view of the standard plane to the anatomical region is not included in the pipeline proposed and is replaced by a pre-extraction of the patches considered. In other words, in the experiments proposed here, the patches are extracted beforehand since the proof-of-concept focuses on incremental learning of anatomical structures. As such, the application of the models trained in this process can be applied in clinical

workflows under the assumption that a detection step for the anatomical structures (similar to the study in Section 3.6) is included prior to the incremental task.

4.6.2 Baselines

For comparisons with a direct coarse-to-fine classifier using incrementally growing data distributions, a baseline CNN-LSTM architecture (Figure 4.9) is used. The design of this model ensures the parameterization is in the same order as that of the super-model and a single sub-model combined. It is trained with all old fine classes (80:20 split on Dataset B, batch size of 50 and learning rate of 0.001) initially. Like data distributions used in the models proposed, in additions of new classes in *Experiment 1* and *Experiment 2*, we use datasets of new classes, C1 and C2, in a 80:20 split. Class prediction accuracies for old and new classes are used to benchmark *Experiment 1* and *Experiment 2*.

The baseline comparisons essentially replicate a transfer learning situation between the old and the new tasks, which in this case involve expanding model knowledge from a set of old classes towards including a set of newer incremental classes. Typically, the concept of transfer learning was not proposed to consider incremental class addition, but the fact that the new classes represent a different data distribution from the previously seen classes renders itself conceptually to the domain of class incremental learning. New tasks could be understood through the lens of either new analysis with existing data (such as expanding network capacity to perform a regression task, while having initially trained for a classification on the same dataset and same number of classes being considered) or, new tasks could involve optimization of an existing model over new data distributions such as new classes. The studies in the chapter are concerned with the second regime. As such, the ideas of transfer learning seek to explore whether model parameters trained for a certain optimization objective on certain data distributions can aid the learning over a new data distribution. In a broad interpretation, this ‘transfer’ can extend both ‘forward’ – where parameters in the trained model are better adapted to the new task training by virtue of having been subjected to an earlier schedule

of optimization as compared to learning the new task from scratch, or ‘backward’ – where the effects of the new task training upon performance on the old task validation sets is examined. Therefore, the idea that parameters are affected in incremental optimization steps for a pre-trained model, and that such effects can either positively or negatively influence both future learning and past validation performance allows its extension to continual learning settings. This is assuming the continual learning setting would have new sets of classes arriving over extended time in a manner that the data distributions are different in each incremental arrival.

4.6.3 Early training

We first train the super-model and the sub-models on the datasets assumed as old classes. Following this, incremental training is sequentially done on the datasets for the new classes. To begin, we use Dataset A for training and testing the super-model in an 80:20 split. The super-model needs to distinguish a structure from 4C, 3V and LVOT views respectively. This step is carried out with a mini-batch of 100 and a learning rate of 0.01 using Stochastic Gradient Descent in both cases. The next step is training sub-models for fine classifications into different fine structures of the initial classes. Dataset B is used with an 80:20 training-test data split and 3 sub-models are trained, one for a coarse class. For *Network 1*, inputs to the sub-models are of the same batch as that of the corresponding super-model as the average softmax probability of a batch is used to select right sub-models for fine classification rather than individual instance predictions. This is implemented as a conditional calling of pre-trained versions of relevant fine classifiers with the same batch for incremental training. As the entirety of a new set is assumed to be of the same class sharing coarse details, a single sub-model is adapted at any incremental stage.

For *Network 2*, inputs are randomly prepared in pairs of same or different classes with labels of 0 or 1, so as to use a binary classification (equal numbers of positive and negative pairs). Encodings are generated using shared weights of a Siamese network, and used to mutually optimize the embedding output normalized between 0 and 1. Here, the training creates models able to find similar or dissimilar images,

approximating a classification by similarity task analogous to the hard attention based discriminator in [103] with a distinction that a segregated fine-tuning is applied here to establish a footprint of novel data over time. A batch size of 50 is used with a learning rate of 0.0001. A low learning rate helps prevent drastic modifications in the parameter space during fine-tuning.

4.6.4 Incremental training

We perform incremental learning on the sub-classes left out in the early training stage. First, anatomical structures being incrementally learned are passed through the super-model to decide on coarse classification based on softmax probabilities obtained for three coarse classes. The case of incremental learning analyzed here is when the input instance belongs to a coarse class but not an existing fine-class. This is tested with foramen ovale (FO) examples from 4C classes, superior vena cava (SVC) from 3V cases and right ventricles (RV) from LVOT. Exemplar videos for each are fed to the super-model and coarse probabilities computed. Based on the coarse probability, relevant sub-models are adapted with instances from new fine classes in 100 epochs with a learning rate 0.0001. The number of samples is equal to the total dataset size in an 80:20 train-test split for the incremental learning part of *Experiment 1* and is restricted to 500 samples in *Experiment 2*. We use these 500 samples to generate randomized pairs with same new instances for positive pairs, and negative pairs using pair images from other classes to simulate few-shot increment steps for *Experiment 2*. In both these experiments, the train:test split, learning rate schedules and batch sizes are kept the same as in the early training stages.

It is noted that in case of *Experiment 2*, the sub-models are implemented as $N2$ modules that are essentially networks trained to compare feature representations derived from input video frames. This fundamentally implies that for novel classes, instances can be mutually compared for similarity as long as the classes themselves have an inherent association to the ones used for the first stage training. Thus novel class comparisons can theoretically be accomplished indefinitely without further incremental training given an oracle that collects and labels the data distribution

after its new instances are deemed mutually similar by the early-trained model. A caveat is that the reliance of discriminative power on initial training alone results in diminishing accuracy with diversity of incremental stages. To enable suitable retention of discriminative ability over diverse temporal data, we improve upon standard few-shot learning to fine-tune sub-models on incremental class data. The specific fine-tuning procedure applied for both classes of sub-models in both experiments is different to standard fine-tuning, as there is an enforced segregation on the sub-models actually fine-tuned and the refinement is governed by the coarse classification unlike in standard fine-tuning or transfer learning.

4.6.5 Evaluation protocols

For evaluation of incremental learning tasks defined in this section, we consider the retention of old knowledge and ability to generalize to new classes. For truly generalizable machine learning models, it is important to be able to retain both past knowledge and use the same for improve acquisition of future information (analogous to learning from experience) and potentially improve past performance itself due to acquisition of such future data. To that end, we define two evaluation metrics - Overall Drop and New Task Adaptation. Formally, the Overall Drop metric is supposed to detect the extent of drop in the chosen performance metric (in this case, the test or validation accuracy) on incremental training on new arriving class data. New Task Adaptation is the metric to assess the model performance on the new class data introduced. The former is observed by noting the drop in accuracies over validations on old classes during new data addition ('Overall Drop' in tables and in Figure 4.12). The second is evident from the validation accuracy on each new class added ('New Task Adaptation' in Figure 4.13). The priority of incremental learning pipelines then is to reduce the Overall Drop and improve the New Task Adaptations over different incremental stages.

At incremental stages, hold-out data for each class in Dataset A is used to test the pipeline for accuracies on every class in our experiments. The net decrease is the 'Overall drop', which is the difference in the validation data accuracy before new

class addition and that after all new data is trained for. Averaged over fine classes, the capacity of the architecture to mitigate forgetting is found as the Average Overall Drop (this is computed over the total number of classes, old and new, as new class accuracies are affected in subsequent incremental stages in baselines). New Task Adaptation has a reliable proxy in the validation accuracy on new classes after all of them are added. As then trained model is used to make predictions on unseen validation data of new classes, it shows forward transfer of knowledge without impact on model parameters during subsequent training.

4.6.6 Incremental computational and memory overheads

The sub-models are initialized for different coarse classes. Thus the memory requirements are conditional on the number of coarse groups studied. In clinical settings, top-level grouping can often be defined and memory budgets can be allocated at the beginning with a relative ease as compared to various finer classes as the latter can sometimes vary depending on scanning situations and physiological conditions. Arrival of new fine classes likely to share coarse characteristics with a pre-drawn set of top-level classes would thus not lead to increased memory usage during incremental training (though an average softmax probability cut-off at the coarse classification allows creating new sub-models in our pipeline, we do not have a use case relevant to that here). As we start with pre-trained models on old tasks, computational costs on new additions derive only from the fine-tuning on such data. It is a minimal addition on each increment as full training on new classes (or old plus new classes) is not performed. As for inference, the computation time depends on the number of floating point operations incurred in the matrix multiplications in the convolutional and recurrent stages along with the data loading operations. As such, the inference proceeds with the coarse label assignment and selection of the appropriate sub-model, which implies that the number of operations is the sum of those incurred in the super-model and one of the sub-models. Thus, the inference time stays the same regardless of the incremental stage since there is no

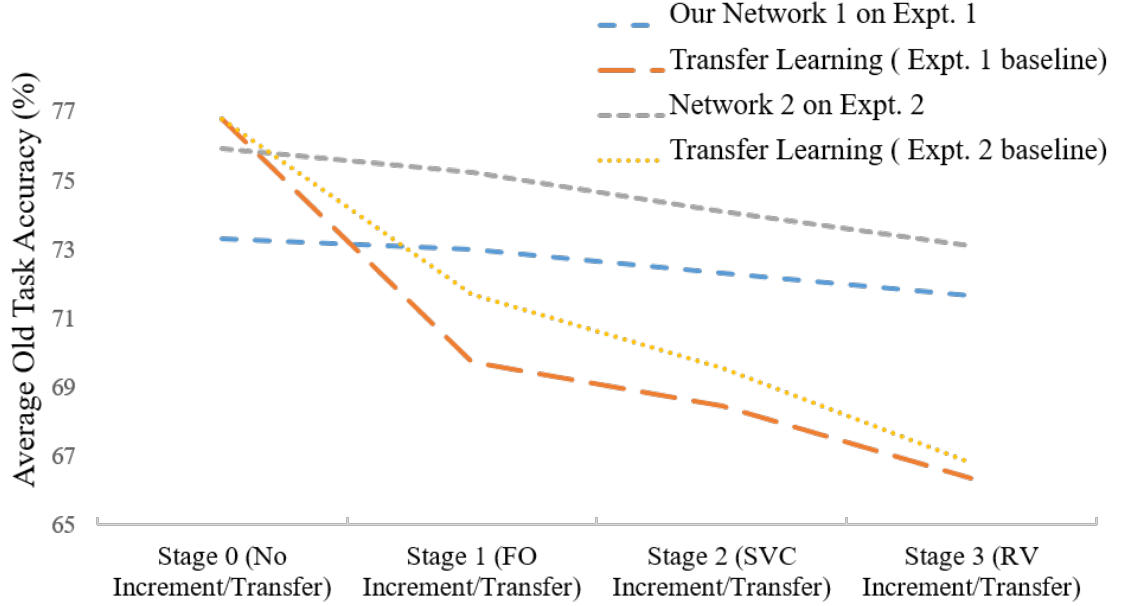


Figure 4.10: Diminishing old task performance for transfer learning baselines compared to near constant performance of Networks 1 and 2.

new addition of network stages unlike in continual learning approaches that rely on a dynamic expansion of networks such as [104].

4.7 Results and Discussions

Class specific validation accuracies are reported at every incremental stage. Therefore, the difference in the accuracy between the initial stage and the final stage reflects the class-specific drop in validation performance of the approach. This is the forgetting measure. Average model accuracies are obtained by averaging classwise accuracies for each stage, and offers a glimpse into the adaptation towards the new classes being added.

4.7.1 Incremental Performance

Using data distributions crafted for the *Experiments* 1 and 2, the incremental learning protocols defined earlier are studied. The progression of learning with the arrival of sparse data from new fine classes is evaluated for both *Networks* 1 and

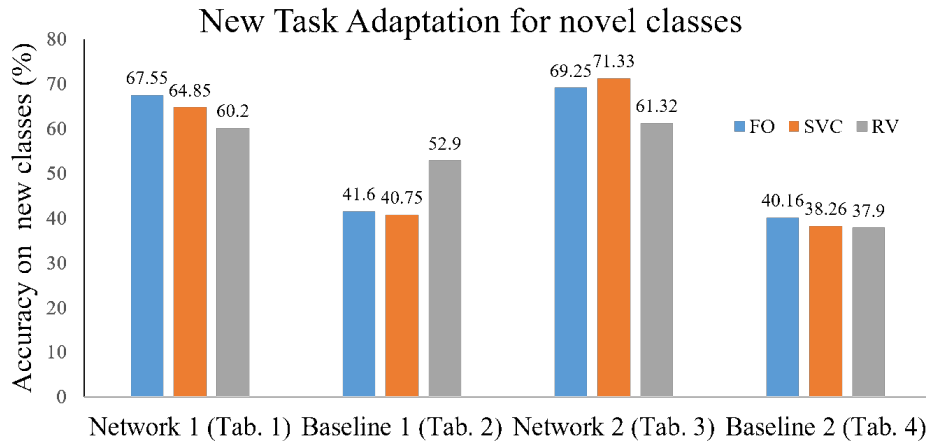


Figure 4.11: Ability to adapt to new information. Final accuracy over new classes is a proxy for New Task Adaptation. Network 2 generalizes best overall.

2 and presented in Table 1 and Table 3, with comparisons drawn with a baseline learning applied with and without the novel classes at each incremental stage. It is evident that in the hierarchical scheme, the effect of such inclusion is localized to the sub-models called with respect to the predicted coarse class and the other fine classification sub-models remain unaffected, as they do not participate in the adaptation to such new data. This can be seen as the accuracies stay constant with respect to the 'No Increment' stages unless new data is added within the coarse class (FO increment in 4C structures, *Sub-model A*; SVC increment in 3V structures, *Sub-model B*; RV increment in LVOT structures, *Sub-model C* in both modes of hierarchical incremental learning performed in *Network 1* and *Network 2*). Including examples from a novel class and training the relevant sub-model is seen to slightly reduce prediction accuracies for structures the sub-model was previously trained for. This is because the weights adapted for fine classification may not be best suited for predictions on previously learned structures but the values predicted are still close to original as compared to the corresponding steps in standard transfer learning baselines. Further, the incorporation of new sub-classes relevant to a single sub-model does not influence the fine-classification performance for the other two sub-models as shown in the prior fine accuracy figures remaining constant (shown in green in the tables) over these stages of class additions compared

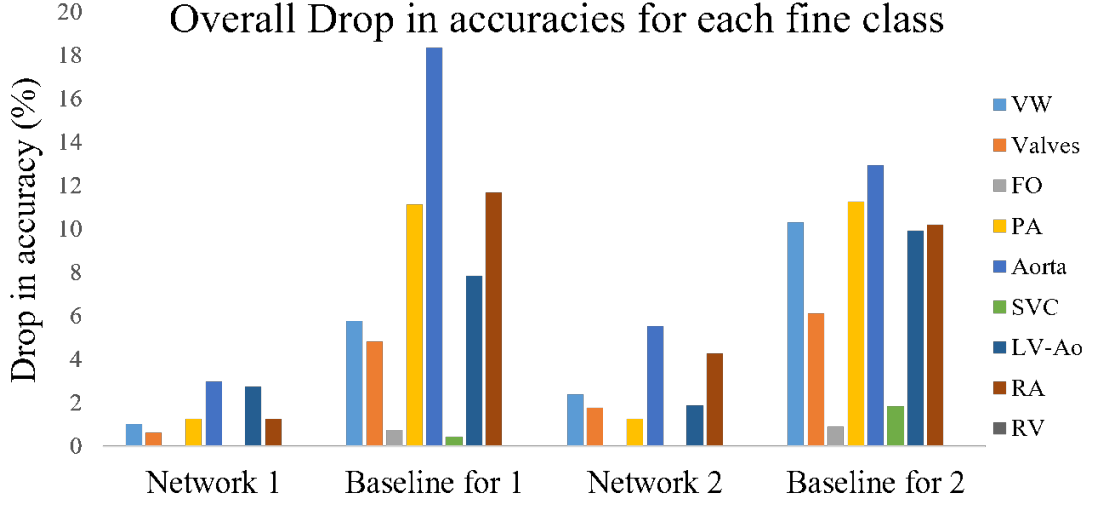


Figure 4.12: Catastrophic forgetting quantified by Overall Drop in class accuracies for old classes. Network 1 and 2 outperform baselines across classes.

to other sub-models. The progression of novel data addition in the proposed incremental few-shot scheme is compared to a standard coarse-to-fine model using a standard transfer learning/fine-tuning baseline, which has been a method of incorporating the ability to expand to new tasks in a variety of statistical learning approaches. It is seen that such a transfer learning model adapts poorly to an incorporation of novel classes (Tables 2, 4). This is reflected in the depreciation of the observed accuracies of old fine classes in each coarse bin, and a significantly low new task adaptation for novel data as well.

4.7.2 Forgetting and retention.

In *Experiment 1*, accuracies with new classes are diminished with direct transfer learning compared to *Network 1* – Foramen Ovale (67.55% vs 41.60%), Superior Vena Cava (64.85% vs 40.75%), Right Ventricle (60.20% vs 52.90%). In *Experiment 2*, baseline accuracies for the newly added classes of Formane Ovale (FO- 41.05%), Superior Vena Cava (SVC- 40.11%) and the Right Ventricle (RV- 37.90%) are much lower than the corresponding values obtained with our hierarchical models in *Network 2* (FO- 69.25%, SVC- 71.33%, RV- 61.32%). Also, in comparison to

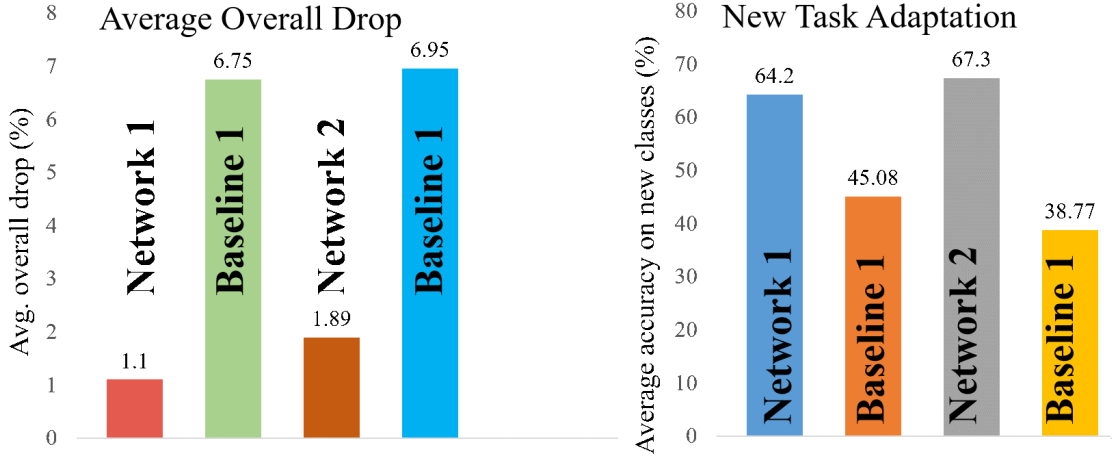


Figure 4.13: Summary of model evaluation - Average overall drop in performance on old classes, and new task accuracies over new classes. Network 1 shows slightly less forgetting but Network 2 performs better on new tasks.

our methods in both experiments, the accuracy for existing classes get modified with additions of new classes to a greater degree in direct transfer learning methods (Tables 2 and 4) where a single model is required to adapt to all the new classes after initial training on base classes. The metric 'Overall Drop' quantifies the presence of catastrophic forgetting in accuracy in both cases as averaged over all fine classes (4.12) for *Experiment 1*, 1.10 % vs 6.75 % in between using the hierarchical pipeline of Network 1 versus using the baseline; in *Experiment 2*, 1.89% vs 6.95% in between the modular *Network 2* vs the baselines. The improvement in knowledge retention for several difficult classes is starker – the overall drop in accuracy (4.12) reduces from 11.24% to 1.25% for the PA (Pulmonary artery) additions, from 12.94% to 5.52% for Aorta and from 10.30 to 2.38% for VW (Ventricle Walls) in *Experiment 2*. Such improvements are noted in *Experiment 1* too. This is due to drastic weight modifications in baseline architectures, explained using the well-studied limitations of neural networks and transfer learning [47][97] about trainable parameters being successively overwritten in these models. Our models adapt to novel classes better than baselines, quantified by the New Task Adaptation term for validation accuracy on new class data after all increments, both in average final accuracies (Figure (4.11)), and on the model averages (Figure (4.13)).

4.7.3 Effects of Imbalance.

To quantify effects of incremental imbalance, a comparison is also made by using *Network 1* on the data of *Experiment 2* (Table 5). The initial accuracies before new class additions are at par with values in Table 1 (which is obvious as at this stage the classification tasks are same). The subsequent divergence in accuracies is rapid for new classes (69.25% vs 40.10% for FO, 71.33% vs 43.15% for SVC, 61.32% vs 40.65% for RV), indicating certain misclassification despite modularization. The imbalance during incremental fine-tuning is thus seen to fare poorly in softmax classification settings. Contrarily, similarity methods are explicitly trained on distances among instances as the basis for categorization. This enables a direct extension to small numbers of novel instances, evident in better performance in comparisons on corresponding stages in Table 3 and Table 5. This is because when making a classification decision for an instance, we rely on proximity to a prototypical representation of the class in a low dimensional space, the quality of such a classification decision depends on the nature of the learnt prototypes for individual classes under study. It is possible to have a class prototype that is representative of a class even with a few examples of a given class as long as such examples accurately and sufficiently capture the diversity of features in a class. Such a prototype can be used to evaluate novel exemplars on how 'close' their representations are to the prototype. This allows a few-shot classification by detection of novelty of new exemplars, and the number of exemplars required for sequential training is significantly reduced. That said, the expansion in the set if initial exemplars used to train the model to obtain the class averaged prototypes can improve the learnt representation used as the prototype and lead to better outcomes on the classifications in incremental stages but at the cost of the computational and memory overheads of obtaining and storing a larger number of new class instances which is often not feasible in actual deployment situations.

4.8 Conclusion

We proposed experimental regimes with novel evaluation criteria for designing and assessing deep learning pipelines for continual learning on extremes of incremental data distributions in clinical settings using the task of fetal echocardiographic analysis as an exemplar of approach. Validations presented on clinical data with annotations from certified sonographers (as described in Section 1.2, which detailed the data acquisition and annotation strategies assisted by sonographers affiliated to the Oxford University hospitals) demonstrate the feasibility of the proposed approach for lifelong learning in clinical image analysis. Over methodological improvements on the proof-of-concept presented, future directions could include the adaptation of class incremental learning in cross-modal situations in line with clinical practice. The key hypothesis that observations may be missed in initial scans, with later examinations filling in diagnostic gaps, is extendable to multiple clinical imaging tasks. While the deployment of continual learning systems is still at an embryonic stage for most clinical systems, there is an emerging conversation in many regulatory organizations about the fact that machine learning algorithms, by nature, need to evolve into adaptive continual learners that can be corrected for emergent situations and a static approach is not sustainable for robust clinical decision support. This makes it pertinent that legal and regulatory frameworks enable the adoption of algorithms that allow for incrementally learning over novel information, without compromising patient privacy and confidentiality. Additionally, the ability to perform such continual learning on few-shot settings is of critical importance in rare diagnostic markers, and in cases with a large potential set of false positives. In such cases, achieving a learning model to perform robust diagnosis requires an incrementally resilient few-shot learner and our proposed method can be adopted with minor modifications.

5

Multiple class incremental learning with attention

Contents

5.1	Motivation	109
5.1.1	Attention Map Quality	110
5.1.2	Attention driven continual curriculum	114
5.2	Model and Objective functions	116
5.2.1	A note on the distillation loss	117
5.3	Data preparation and Training	118
5.3.1	Training	118
5.3.2	Baselines	119
5.4	Results and Discussions	120
5.4.1	Attention map quality	120
5.4.2	Accuracy across stages	122
5.4.3	Limitations	123
5.5	Conclusion	125

In the previous chapter, the idea of ensuring reliable machine learning models for fetal ultrasound video analysis was explored through the premise of creating algorithms that help learn over extended time across varying data distributions without compromising on the performance on previously introduced data. There, a class incremental setting was used to enable the model to increase its capacity by one class at each incremental stage. In practical image analysis situations, there may arise a requirement to adapt to multiple classes at once in a given stage, or there

may be a desire to learn more than one class in successive learning stages to save the training effort or power consumption associated with repeated model finetuning. This is a multi-class incremental learning setting, which is introduced in this chapter. This is different from the task incremental learning setting in the sense that these classes do not necessarily derive from different tasks in an ultrasound imaging context. Additionally, the reliability of such models over time in a clinical setting also rests on the ability to have an explanation for the decline in performance or otherwise, when adapted for new data distributions. A concern pertinent to deployments of machine learning models in healthcare settings is the ability to explain decisions made given certain input instances. In the event of learning over incremental new classes, such an ability to explain needs to be retained over all such sessions. In fact, the ability to obtain salient explanations for predictions for given input instances over learning sessions is indicative of the model’s tendency to retain salient past information while optimizing on novel class data. Therefore, such saliency may be considered as a part of the design process for continual learning algorithms. In that light, we propose a dual utilization of attention map based explanation strategies to implement the proposed continual learning setting. Additionally, we define novel quality metrics for class averaged attention maps that also quantify the ability of a model to learn continually. Note that the terms ‘incremental learning’, ‘continual learning’ and ‘lifelong learning’ are used interchangeably to refer to learning algorithms over multiple sessions with different data distributions, as in the previous chapter. The research presented in this chapter was presented in the following publication which was accepted for an oral presentation at:

1. Arijit Patra, Alison Noble, ‘Incremental learning of fetal heart anatomies using interpretable saliency maps’, In Medical Image Understanding and Analysis (MIUA) 2019.

5.1 Motivation

Convolutional Neural Networks learn hierarchical features that are aggregated towards a low-dimensional representation and the inability of the model to retain knowledge is manifested in this hierarchy as well. Since attention maps point out the most relevant pixels in an input image towards the classification decision made on it by a model [120], a digression of focus from these pixels is indicative of degradation of past knowledge. Thus, attention map quality can be used to quantify not only the overall decline in performance over old tasks, but offer detailed insights into relative decline at the level of individual classes in the task, and also for individual instances in a class (which can be used as a measure of some examples being especially difficult to retain). This attention based analysis is motivated by the fact that present measures of forgetting and knowledge transfer do not allow a granular class-specific assessment of learning processes or distinctions on the basis of difficulty of an instance, nor do they allow a scope for explainability of the continual learning process. Creating attention maps by finding class activations allows for feature level explainability of model decision on every learning cycle. A popular method in computer vision literature to address the issue of interpretability in convolutional neural networks is the concept of Class Activation Maps (CAM) [120]. A generalization of the CAM approach was proposed in the form of Grad-CAM [120] in order to enable compatibility with a variety of CNN based architectures without requiring any architectural modifications.

In this chapter, we focus on the question of whether methods designed to enable interpretability in neural networks can be used to gauge the efficacy of incremental or continual learning with deep neural network based models for ultrasound image analysis. As such, we prioritize the continual learning aspect rather than attempt to propose novel methods for visualizing and explaining deep network decisions. Thus, we choose the Grad-CAM approach for obtaining attention maps that allow a visualization of the features in the image pixels that are most relevant to a classifier decision and explore the proposed approach over a range of pre-existing algorithms for continual learning in machine learning literature and with the methods

introduced in this chapter. Grad-CAM is chosen as the method on account of the ability to use trained models with it without having to modify the architecture with the addition of extra layers such as global average pooling unlike in previous class activation mapping approaches, and its ability to be used without having to apply partial masking to the input image while evaluating attention. The former limitation is particularly untenable in the experiments here because of the reliance on using the final layers as the proxy for the learnt representations on the prior data distributions. The question of exploring a larger variety of methods to obtain the explanations themselves is not taken up since the relevance to the continual learning workflow heavily relies on a metric devised to gauge the quality of the attention maps made available and regardless of the method used to generate these, the rest of the workflow would rely on the map quality evaluation stages.

5.1.1 Attention Map Quality

At the conclusion of a given task of N classes, the validation set of each class is passed through the model which is then subjected to class activation mappings (CAM) to obtain attention maps using the GradCAM approach [120]. In the Grad-CAM approach, the input image is subjected to a forward pass with the trained model and a raw probability score for every possible class is obtained. Thereafter, the gradient of the score value y^c for a given class c is calculated with respect to each convolutional feature map A_k . A global average pooling on the set of gradients so obtained provides the neuron importance α_k of A_k . The set of A_k weighted by α_k are subjected to a ReLU activation to get a final attention map for class c .

Note that the specific method to obtain attention/saliency maps here is for demonstration only and any suitable method may replace it. We consider only the maps resulting from correct predictions because the explanations are generated even otherwise but is suboptimal for further inference as we require the class representations to emerge from correctly identified instances for future replay in incremental stages. Each instance’s map represents the understanding of the model for the decision taken on that instance. Averaged over the validation set of a

class in a task, this averaged attention map represents the average explanation of the model for the decisions of classification. In the absence of a ground truth, estimating a quality metric for the explanation based on an attention map is non-trivial. We try to assess the extent of forgetting by tracking the difference between the activations obtained just after the class or task has been learnt and that after a few other tasks are learnt subsequently in the next learning stage. This can be performed both for individual instances of classes and by considering classwise average attention maps. The latter is preferred in our experiments in line with our multi-class incremental setting.

Past literature has explored saliency map quality in terms of being able to mimic human gaze fixations or as a weak supervision for segmentation or detection tasks, in which ground truth signals were enforced even if weakly [143]. Metrics of evaluation for the predicted saliency maps in this case were mainly designed considering the availability of prior human gaze fixations as ground truth labels. That apart, saliency maps were evaluated by [143] in context of their attempts at designing explainable deep models. They interpret the efficiency of saliency maps as a reasonable self-sufficient unit for positive classification of the base image. Then, the smaller the region that could give a confident classification in the map the ‘better’ the saliency map for that instance. Mathematically, this was expressed as a logarithm of the ratio of the area and the class probability if that area was fed to the classifier alone. This quantity called the SSR (Smallest Sufficient Region) is expressed in [143] as $|\log(a) - \log(p)|$, where a is the minimum area that gives a class confidence score for the correct class as p . This method assumes that the concentration of informative pixels is a good indicator of attentive features, and is unsuitable for cases where features of interest are distributed spatially in a non-contiguous manner. As such, considering our dataset of fetal heart ultrasound, attentive regions may be distributed over a spatial region and the non-contiguous informative regions can be adequately quantified only through metrics designed for multiple salient region estimation. This cannot be adequately expressed by the original SSR metric since it requires the area a to be a single rectangular region.

We extend from the SSR concept and instead of thinking in terms of concentration of information in a single region, consider the extent of regions of informative content that may be spread over multiple non-contiguous rectangular regions [144]. To do so, we consider a grid of fixed uniform regions on the input image. Each grid region is taken as a small rectangular space whose information content is evaluated, the size of these small regions is fixed as a hyperparameter (we consider 224 pixels x 224 pixels image inputs, and 16 pixels x 16 pixels grid regions by optimizing for computational cost and accuracy). Each of these grid regions is evaluated by the trained model after a task for their prediction probabilities by themselves. Then for each grid region, a quantity $|-log(A_g) + log(p)|$ can be used, with $log(A_g)$ being constant for fixed size grid regions as A_g is the area of the 16 pixel x 16 pixel grid chosen to estimate the contribution of the region to the overall saliency map. The smaller this quantity, the more informative this region is. A threshold can be imposed and all n grid regions contributing can be summed up to express the Overall Saliency Quality (OSQ) as:

$$\sum_{k=1}^n (|\log(A_g^k) - \log(p)| / |\log(A_g)|) \quad (5.1)$$

This threshold depends on the desired class probability for the correct class (this expression would be valid even for incorrect classifications, but we choose regions only with correct predictions). Thus, a quality measure can be derived for each saliency map at all stages of tasks proceeding from 5.1. The expression above essentially gives an ‘absolute’ measure of attention map quality.

Retention of exemplars for efficient continual learning was done by random selection, by nearest class mean and so on. These methods do not consider the effect on actual performance on training or validation data while making the retention decision but derive from input distributions if not randomly selecting instances from the datasets. While in ideal cases, the class mean will reflect exemplars with high quality saliency maps (as one would trivially expect that the average class representation has the most volume of data and hence is more influential on the learning curve for these classes), it is not always true in case of classes with

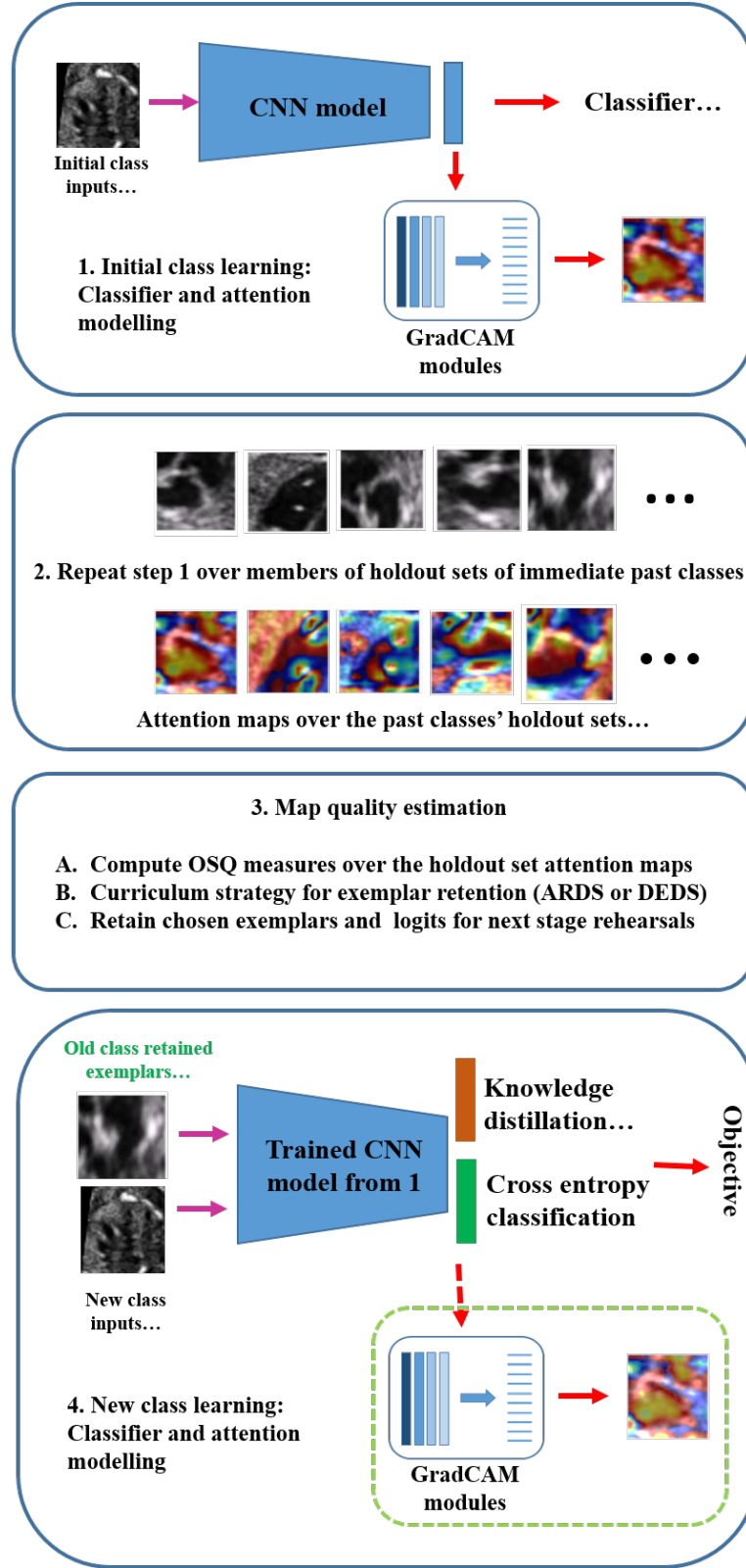


Figure 5.1: Schematic of saliency curriculum based exemplar retention - After training for a session, GradCAM modules compute attention maps from instances for OSQ calculations and selecting retention examples. Figures for representation purposes only

significant diversity and multiple sub-clusters of exemplars. In the latter, selecting exemplars by methods like herding [145], the need for retaining exemplars close to the class means is not optimal for retaining the diversity of class information, and the diversity of informative representations needs to be accounted for, which can be informed by attention based retention strategies [146].

5.1.2 Attention driven continual curriculum

We attempt to use the attention maps of past representations to help actively preserve knowledge and at the same time improve generalization to future classes. This is similar to transfer of knowledge from *Task N* to *Task (N+1)* through the attention maps of the former being used to condition the latter. After each task schedule we generate activation maps using the methodology presented by Selvaraju et al. [120] for the validation data per class and estimate the quality measure defined in the first part. Following our need for retaining prior knowledge while moving to the next task, we consider a selective retention strategy. In order to account for fixed memory allocations per class, a fixed number of exemplars may be retained. We try to establish the most informative of instances through a attention map quality based curriculum over available instances in a class.

We propose two ways of preparing an attention map based exemplar retention:

- 1) Assuming the presence of definitely identifiable salient cues in available instances, an average class attention map can be understood as an overall class decision explanation. A relative proximity at the pixel space of a given attention map of a given instance to this average attention map would indicate the suitability of such an instance for being stored as a representative exemplar for its class. This approach is termed the **Average Representative Distance Selection (ARDS)**.
- 2) Another alternative is to pre-select a set of most representative exemplars from a validation dataset of the class in terms of the class confidence probabilities, and use the normalized cumulative distance from their attention maps to every other instance's attention map. This is termed **Distributed Exemplar Distance Selection (DEDS)**, and is primarily useful for cases where the salient regions of interest have

a non-trivial spatial variation within the same class exemplar set. ARDS is suited when strong cues are localized and class prototype attention maps are useful. DEDS is useful in cases of a diversity of cues not similarly located on all exemplars or when a significant shift is caused by affine transformations. ARDS is computationally more efficient as a prototype-to-exemplar distance is computed in a single step. Actual choices between the two would depend on data characteristics. In both cases, the attention map is treated as a probability distribution and similarity is assessed by KL divergence between reference attention maps and instances' attention maps,

$$D_{KL}[e, d] = \int (e(x) - d(x)) \log(e(x)/d(x)) dx \quad (5.2)$$

considering that $e(x)$ and $d(x)$ are the respective maps being compared. As the computation for 5.2 over the pixel space is discretized, the integral form is replaced with a discrete summation,

$$D_{KL}[e, d] = \sum_{x=1}^{N_{pixels}} (e(x) - d(x)) \log(e(x)/d(x)) \quad (5.3)$$

As for choosing instances for the ARDS or DEDS calculations using equation 5.3, it would be computationally expensive to obtain attention maps for every single instance in the training sets. We choose 100 exemplars from each class's training set (based on their confidence probabilities in final epoch) for this process. This set will be referred to as the holdout set. The retention examples are chosen from training sets as they are replayed in future sessions and hence validation set examples cannot be used except as saliency benchmarks, i.e, while the benchmarks for average saliency or representative exemplar sets are derived from validation sets, these can't be retained as validation data can't be used in any part of training. In both cases, 30 instances are finally chosen after a grid search over integral number of samples between 10 and 50 (with the upper bound governed by memory constraints and lower bound on performance thresholds), to be retained in memory for future rehearsal. Choosing higher numbers of exemplars was found to lead to minimal performance gains (a detailed study of these trade-offs is left out for future work).

5.2 Model and Objective functions

For our architecture, we implement a convolutional network with 8 convolutional layers (64, 128, 128, 256, 256, 512, 512, 512 filters), interspersed alternately with maxpooling and a dropout of 0.5 after the third convolutional layer. This is followed by 1024-way fully-connected layers before a softmax classification stage. The focus of the work is not to achieve possible state-of-the-art classification accuracies on the tasks and datasets studied, but to investigate catastrophic forgetting in learning incrementally over multiple groups of novel classes. Thus, the base network used is significantly simplified to keep the order of magnitude of the number of parameters within that of other continual learning approaches reported in literature and enable a fair benchmarking. For a loss function, we implement a dual-objective of minimizing a shift on learnt representations in the form of prior logits in the final model layers using a knowledge distillation approach [133], and performing a cross-entropy classification on the current task classes. A quadratic regularization is additionally imposed with a correction factor μ that is set as a hyperparameter by grid search. The overall objective function is:

$$L = L_{cur} + L_{KD} + L_r \quad (5.4)$$

Equation 5.4 represents the overall loss function. Here L is the total objective composed of the softmax cross entropy as the current loss (equation 5.5), the knowledge distillation loss on past logits (equation 5.6), and the regularization term for the previously trained parameters:

$$L_{cur}(X_n, Y_n) = -\frac{1}{|N_n|} \sum_{i=1}^{N_n} \sum_{j=1}^{J_n} y_n^{ij} \cdot \log(p_n^{ij}) \quad (5.5)$$

where N_n is the number of examples in a batch, J_n is the number of classes, y_n^{ij} is the one-hot encoded label for an instance, and p_n^{ij} is the softmax prediction. For the distillation terms, the original labels not being available, y_o^{ij} is computed with new and retained examples and compared to stored logits for the old examples p_o^{ij} , giving a loss term:

$$L_{KD}(X_n, Y_o) = -\frac{1}{|N_n|} \sum_{i=1}^{N_n} \sum_{j=1}^{J_o} y_o^{ij'} \cdot \log(p_o^{ij'}) \quad (5.6)$$

, where $y_o^{ij'} = \frac{(y_o^{ij})^{1/\lambda}}{\sum_j (y_o^{ij})^{1/\lambda}}$ and $p_o^{ij'} = \frac{(p_o^{ij})}{\sum_j (p_o^{ij})}$

, where the distillation temperature for the logits λ is set at 2.0 through a hyperparameter search performed on integral values from 1 to 10.

The parameter regularization is imposed for already trained weights for the past classes and penalises the shift for current adaptation through the regularization term over the parameters as $L_r = \mu \sum_j ||w_o - w'||^2$. μ is set to 0.4 by grid search. The idea here is that while the retained exemplars from the past tasks are seen with the present data, the process of optimization should update parameters in a manner compatible with past prediction features. This ensures that parameters are adapted to present distributions without drastic shifts that adversely affect their ability to arrive at the optimal representations for previously seen examples. A distillation framework is implemented here as such a loss term in conjunction with a cross-entropy objective can enforce a regularization on representations from the past, which is not achieved by a simple parametric regularization.

5.2.1 A note on the distillation loss

The knowledge distillation concept is used in the objective function here for incremental learning, and will be used in the next chapter for enabling efficient model compression. The idea of using tempered final layer values for a supervision objective was originally proposed for model compression applications in Buciluă et al. [147] and later popularized by Hinton et al. [133] to enable transfer of learnt representations from deep neural networks with a large number of parameters to more lightweight deep learning models. They found it easier to train a classifier using the real-valued outputs of another classifier as target values than the ground-truth labels. The distillation loss was subsequently explored for a variety of applications in machine learning literature, including in lifelong representation learning [112]. Despite being empirically successful in transferring representations from one classifier to another in a variety of applications in recent years, the reasons behind efficiency

of knowledge distillation has not been established. Hinton et al. [133] postulated that the logit distribution of the incorrect outputs encodes information on the relative similarity between output categories for a given input instance. Such a knowledge of relative similarity between potential targets was demonstrated by them to lead to better convergence rates for the smaller models as compared to direct training with the one-hot encoding.

5.3 Data preparation and Training

The overall workflow used is schematically detailed in Figure 5.1. From the 39556 frames of different standard planes and background, we crop out relevant anatomical structures in patches of 100 x 100 from the frames, leading to a total of 13456 instances of 4C view structures (Ventricle Walls, Valves, Foramen ovale), 7644 instances of 3V view structures (Pulmonary Artery, Superior Vena Cava, Descending Aorta) and 6480 LVOT structures (Right Atrium, Left ventricle-Aorta continuum and Right Ventricle). A rotational augmentation scheme is applied with angular rotations of 10 degrees. Instances sourced from 10 subjects are used for training sets and the rest for validation. Note that in this case, we are working with frame-level images and not clips for the incremental learning experiments.

5.3.1 Training

For initial training, the number of base classes (N) are taken as 3. In the $(N+1)^{th}$ task following the creation of exemplar sets of the immediate past task, the training process is started for the $(N+1)^{th}$ task and so on (this goes on for 3 sessions in total in our case as we deal with a total of 9 classes of sub-anatomies). Batches are created between old and new data, and to further improve performance a distillation based regularization with representative logits of the past tasks is used along with the cross-entropy loss for the present task and exemplars. The training stage essentially involves a base training with the first set of classes, carried over 50 epochs with a learning rate of 0.001. This is followed by a class activation mapping stage with the recently trained model, and an averaged saliency map calculation per class.

This model is now fine-tuned for the group of classes for the next task with a joint distillation and cross-entropy loss over past logits and new labels for 50 epochs. The process continues over the remaining task sessions. GradCAM stages are carried out only after entire task session is completed and not in-between epochs. In those models where past exemplars are retained and rehearsed, these are interspersed with the batches during fine-tuning over the new class sets.

5.3.2 Baselines

For baseline comparisons, the network model described is adapted with the protocol in iCaRL [112], where the representation learning stages are followed by a storage of class-specific exemplars using a herding algorithm [145]. This is implemented in our datasets by computing average prototype representations through the penultimate fully-connected layers for the classes seen till the previous task. A multi-class adaptation of Learning without Forgetting (LwF.MC) [106] is attempted with our network and objective function without the weight regularization term, and logits for distillation are retained. For the end-to-end learning (E2EL) [148] method adapted to our network, a fixed memory version is followed to be a more accurate benchmark to our own fixed memory per class assumptions. The representative memory fine-tuning protocol in E2EL is implemented for baselines with the same training configurations as the initial training, except that the learning rate is reduced to 1/10th the initial value. A progressive distillation and retrospection scheme (PDR) [149] is implemented with replicated versions of our network serving as the teacher network for the distillation and the retrospection phases, with the exemplars generating retrospection logits for regularization while progressively learning on new data which are presented as a second set of logits which have been learned separately in another replication of the base network. In these implementations, storage of prior exemplars follows the same protocols as used in the original implementations.

5.4 Results and Discussions

We consider configurations of our approach in terms of the exemplar retention method used and adopt OSQ in tandem: 1) map quality with OSQ and ARDS, 2) map quality with OSQ and DEDS. Variants of our approach without storing rehearsal exemplars are also considered in terms of training a vanilla CNN with similar architecture as the base CNN used in other approaches. This is the same as using a simple CNN baseline with transfer learning over task sessions. Another version CNN(TL) also functions without retention strategies but with convolutional layers frozen while further task finetuning (this would reflect in a much higher difference between initial task performance and subsequent task values). The reported performances include the average accuracies at incremental levels, with and without using the salient retention scheme in step 2, benchmarked with our adaptations of methods in iCaRL, LwF.MC, E2EL and Progressive distillation (PDR). Using these benchmarks indirectly also allow comparison between different exemplar retention strategies, such as with naive and herding based methods explored in iCaRL and LwF. Also reported is the average saliency map quality in each stage. This OSQ metric is a proxy for the level of forgetting over multiple stages, and a difference between consecutive stages in the OSQ represents the decline in the models ability to seek out most salient image regions over classes. In addition, the OSQ is an indicator of stagewise model interpretability since accurate model explanations rely directly on the quality of attention maps for the ultrasound images used.

5.4.1 Attention map quality

The reported OSQ over learning sessions is the attention map quality averaged over all validation exemplars available for previously seen classes. We report the average value for tasks so far, since we want to look at broad trends in the overall saliency to assess the overall ability to retain knowledge. For future extensions, it is straight-forward to obtain these values at both class and instance level and only requires them to be input to the trained model and class activation maps

processed before the OSQ calculation. There is a difference in accuracies for baseline methods compared to original implementations due to our using the same base network for all baselines and models for uniform assessments (e.g. iCarL [112] originally used embeddings from 32 layer ResNet on CIFAR 100 but we use the iCaRL approach on our data and our base network). A trend is established where the inclusion of exemplars based on a saliency driven approach is seen to have a marked improvement on mitigation of forgetting, based on the OSQ metrics introduced here, and also on the validation accuracies averaged on previously seen classes for the task sessions considered (the past accuracy % in Table 1 for a task stage refers to validation accuracy of validation sets from previously seen classes, and the present accuracy % is the validation accuracy obtained on the present validation set). The attention map quality variation roughly corresponds to the past accuracy percentages demonstrating the efficacy of using map quality as a metric for evaluation of continual learning algorithms in medical image analysis.

In the presented demonstrations, a marked superiority of the methods incorporating attention map based rehearsal curricula is observed. In particular, the attention map quality on the holdout sets of the past tasks (up to one session backwards) shows a much reduced decline in both the ARDS and DEDS strategies for creating the set of exemplars for rehearsal in incremental task sessions. In between the ARDS and DEDS methods, the reduction in map quality on the Task 1 (T1) holdout sets is 0.018 (0.933 to 0.915) from the first to the second incremental learning sessions. The same alteration in the map quality on Task 2 (T2) is from 0.942 to 0.876 between the second and the third incremental learning steps. With the DEDS strategy, the respective declines are 0.023 (0.946 in the first session to 0.923 in the second session) and 0.095 (0.938 in the first session to 0.843 in the second session). The inter-task declines are slightly higher for the DEDS method in our experiments as compared to the ARDS, which is likely due to the relevant structures for diagnostic analysis being co-located within a relatively smaller region within the cardiac radius in our fetal echocardiography frames, resulting in a better assessment of attention map quality given the ARDS metric’s reliance on contiguous

regions. In both cases, the decline is much reduced as compared to either herding based exemplar selections as in the iCaRL and LwF-MC approaches, and also compared to standard baselines constructed out of transfer learning methods.

5.4.2 Accuracy across stages

The task accuracies capture overall accuracies over all test sets for all the tasks seen till that point. This captures the overall classifier performance on all classes and tasks it encountered over successive training sessions. The task accuracy in session 1, reflects the test data accuracy as 0.812, using the proposed approaches (the value stays the same for the first two methods as in the first training stage as the effect of the exemplar set for rehearsal is not relevant during initial training stages). In the second stage, the task accuracy is seen to be 0.845 for the ARDS and 0.863 for the DEDS strategy when comparing between first and second learning sessions. The task accuracies changes from 0.845 to 0.632 for ARDS and 0.863 to 0.636 for the DEDS method between the second and the third incremental learning stages. The trend here is one of slightly lower sequential overall accuracies for the ARDS method as compared to the DEDS method, signifying a better forward transfer performance of the DEDS strategy. The premise of this slight improvement in forward transfer is potentially on account of the distributed salient regions being able to capture a better intra-class diversity, and some of these features may be of importance towards learning similar features in the classes seen in future. The latter is plausible for the ultrasound videos under study since the anatomical regions learnt continually are often-located in near proximity within a small region overall. A key benefit proposed through the idea of using individual class and exemplar level attention maps is to be able to categorize class-specific forgetting characteristics, which is evaluated on the validation set for these classes. The results are shown in plots 5.2 and 5.3 which track the degree of forgetting through the OSQ metric which would decline over successive task sessions. Arresting this decline through the distillation based incremental learning strategy is observed over individual classes for the first two task sessions demonstrating the efficiency of our methods for individual classes. This

evaluation across individual classes for incremental learning performance is a novel method of interpreting incremental learning that has not been explored yet, to our knowledge, in continual learning literature where the overall accuracy is prioritized.

The methods that consider the retention of exemplars from the past overall show not only a better performance with respect to past task accuracies, but also demonstrate considerably higher current learning performances. This implies that a feature importance based identification of informative examples for classes of physiological markers not only improves the ability to better rehearse on past data during future learning stages but also transfers salient representative knowledge leading to better initialization of the parameters for improving performances on the present as well. Here, the diversity of the examples that can constitute a single class type representing a physiological region requires that multiple salient features can be used to explain the final optimization decisions and a diversity of informative exemplars need to be chosen for optimal forward propagation of knowledge while learning on future increments of tasks. The proposed pipeline ensures an inherent continual explainability of the decisions and how they shift over new data arrivals.

5.4.3 Limitations

It is pertinent here to discuss some potential issues with the approach presented since both the domains of explainability and incremental learning are under active research in machine learning communities and in application areas such as medical image analysis. An important bottleneck of the experiments proposed in this chapter lies in the rehearsal stages where class-specific exemplars need to be chosen for retention. Now, while using attention maps to select these exemplars works to ensure a good representative set for each class can be obtained, a pitfall here is that often the best possible exemplars may be skipped in the process since the attention maps are not invariant to translation. Therefore, an image which is somewhat translated from that of the majority of the instances in a given class may be left out of the rehearsal set even if it otherwise has a reasonable good classification confidence score for that class. This is less pronounced in our experiments because the samples finally

used for the holdout sets are of similar orientations as a majority of exemplars in the training and validation sets. This implies that even if the translation invariance is not accounted for while selecting representative exemplars, the rehearsal sets still have sufficiently representative set of exemplars for individual classes.

The methodology explored in this chapter relies on attention maps for quantification of knowledge retention efficiency and relies entirely on the generation of these maps for exemplars after the training processes in individual sessions. There have been ideas in medical imaging literature that propose the usage of attention gates for intermediate model layers [121], where the gate layers implement weights that boost the salient regions in intermediate stages. This enables the model to efficiently learn the intermediate representations by prioritizing the most significant features towards the output. For future studies, it may be worth exploring how catastrophic forgetting affects attention gate parameters, and the extent of the alterations to incremental attention maps when gating mechanisms are used concurrently. One could expect that the training process for the initial and incremental data would condition attention gate parameters to prioritize different subsets of the pixel space in the intermediate feature representations, and this would affect the model knowledge retention since the attention gates would be biased towards the most recently encountered data distributions. Another possible direction could be to explore the joint usage of attention gates along with the ideas of using attention maps for obtaining the retention curriculum, as in the feature maps for the most confident correct predictions could be used in addition to the attention maps in selecting what exemplars to retain. The intermediate attention gates would encode information about feature map importance through the model depth for the classes seen in a particular session and can therefore be used to impose an additional regularization on the calculation of ARDS and DEDS measures.

In this proposed method for continual learning paradigms exploiting attention maps, we presented metrics for assessment of continual learning in terms of attention map quality allowing for an understanding of the basis for predictions over multiple stages of learning and a shift in the learning of the same. We also utilized the

Table 5.1: Evolution of model performance over task sessions. Past accuracy refers to validation accuracy on past session classes (or past tasks). Map quality (MQ) is reported for present task session (leftmost column for each task session head) and for previous session classes.

Method	Task 1		Task 2				Task 3			
	<i>MQ</i>	<i>Task 1 acc</i> %	<i>MQ (T2)</i>	<i>MQ (T1)</i>	<i>Task 2 acc.</i> %	Past acc %	<i>MQ (T3)</i>	<i>MQ (T2)</i>	<i>Task 3 acc.</i> %	Past acc %
Ours (OSQ + ADRS)	0.933	0.812	0.942	0.915	0.845	0.704	0.913	0.876	0.632	0.568
Ours (OSQ + DEDS)	0.946	0.812	0.938	0.923	0.863	0.691	0.887	0.843	0.636	0.593
OSQ + std. CNN	0.871	0.811	0.840	0.631	0.778	0.592	0.852	0.802	0.661	0.455
OSQ + CNN (TL)	0.827	0.813	0.822	0.612	0.702	0.511	0.772	0.647	0.560	0.322
iCaRL	0.811	0.775	0.831	0.622	0.713	0.616	0.834	0.674	0.658	0.321
LwF - MC	0.773	0.762	0.767	0.668	0.732	0.529	0.822	0.731	0.621	0.301
E2EL	0.818	0.793	0.792	0.703	0.742	0.554	0.785	0.706	0.603	0.295
PDR	0.842	0.802	0.837	0.711	0.759	0.514	0.791	0.721	0.581	0.342

attention maps from the past task to select representative exemplars from classes seen during prior tasks during subsequent learning and developed a joint curriculum for creating such sets of exemplars. Our method makes the continual learning process interpretable to a degree, and thereby ensures that the forgetting and retention characteristics of models are explainable. Additionally, this happens to be the first exploration of how attention map quality evolves over incremental learning scenarios in medical image analysis, to our knowledge.

5.5 Conclusion

We proposed experimental regimes with novel evaluation criteria for designing and assessing deep learning pipelines for incremental data distributions in clinical settings using the task of fetal echocardiographic analysis as an exemplar of approach. In this

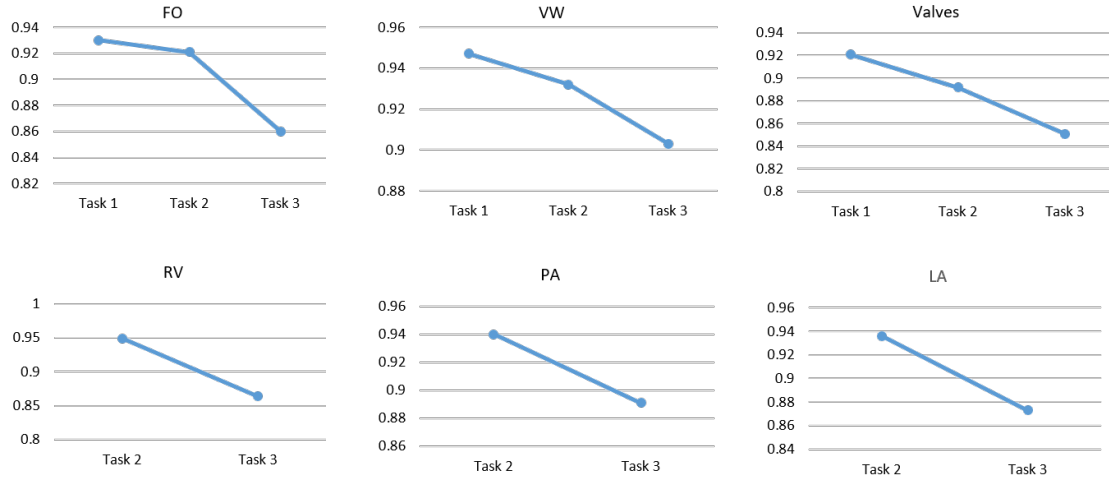


Figure 5.2: Map quality (OSQ) evolution over multiple task sessions for individual classes under the ARDS protocol. The three classes encountered as the new classes in Task 3 are not shown since the map quality for their validation sets is not evaluated beyond the first time they are trained for and hence forgetting is not a relevant issue for these classes.

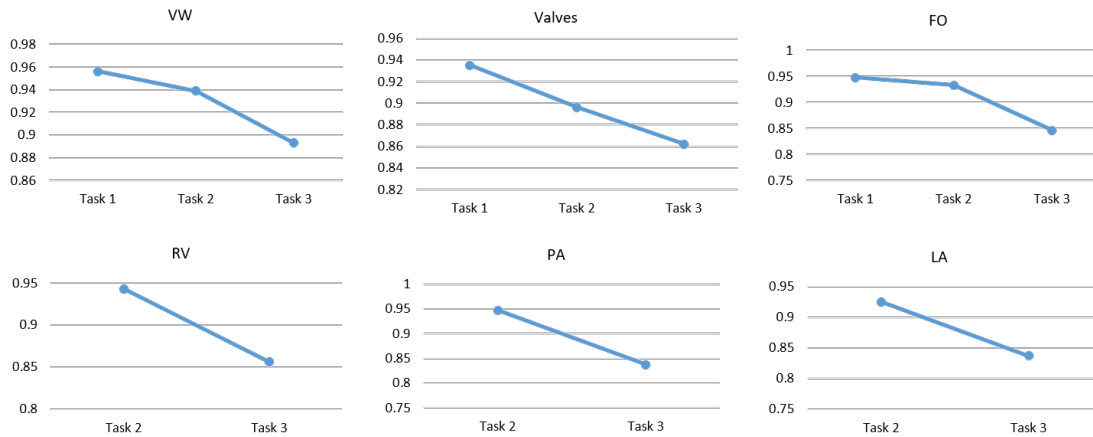


Figure 5.3: Map quality (OSQ) evolution over multiple task sessions for individual classes under the DEDS protocol. The three classes encountered as the new classes in Task 3 are not shown since the map quality for their validation sets is not evaluated beyond the first time they are trained for and hence forgetting is not a relevant issue for these classes.

chapter, a continual learning pipeline reliant on utilizing the evolving attention maps in multiple learning sessions was proposed to include a dimension of interpretability in the process. We also tied the explainability aspect to assessment metrics for incremental or continual learning performance in a departure from the convention in computer vision literature. Validations presented on clinical data with annotations from certified sonographers demonstrate the feasibility of the proposed approach for lifelong learning in clinical image analysis. Over methodological improvements on the proof-of-concept presented, future directions could include the adaptation of class incremental learning in cross-modal situations in line with clinical practice. The key hypothesis that observations may be missed in initial scans, with later examinations filling in diagnostic gaps, is extendable to multiple clinical imaging tasks. Thus, it is pertinent to attempt our lifelong learning method in application areas where developing abnormalities rely on imaging data. The ability to perform such continual learning stages with explainable attention maps is of importance when decision making needs to be validated by human oversight, and in cases with a large potential set of false positives as often may happen in cases of rare abnormalities. In such cases, achieving a learning model to perform robust diagnosis requires an incrementally resilient learner and our proposed method can be adopted with minor modifications.

*Alles Gescheite ist schon gedacht worden.
 Man muss nur versuchen, es noch einmal zu denken.*

*All intelligent thoughts have already been thought;
 what is necessary is only to try to think them again.*

— Johann Wolfgang von Goethe
 [von_goethe_wilhelm_1829]

6

Expert knowledge for reduced models

Contents

6.1	Introduction	129
6.2	Contributions.	133
6.3	Theory	133
6.3.1	Knowledge Transfer	135
6.3.2	Learning from Human Knowledge	139
6.4	Data and Training	140
6.4.1	Data	140
6.4.2	Training Details	144
6.5	Experiments, Results and Discussion	145
6.5.1	Performance	145
6.5.2	Computational Complexity	146
6.5.3	Memory size	147
6.5.4	Model efficiency	147
6.6	Concluding remarks	148

6.1 Introduction

Chapters 3, 4 and 5 looked into the analysis of fetal ultrasound videos with deep learning algorithms, which would attain predictive ability based on the the availability of input data and associated label information. It is pertinent to raise the point of comparison of such algorithms with the decades of experience possessed by several trained sonographers. Humans learn to acquire specialist knowledge through

an extended period of training, observation and practice. Models such as neural networks rely on a composition of parameterized functions applied sequentially to transform a given set of inputs into low dimensional feature representations that are then cast into an optimization problem to learn optimal parameters to model the given input-to-output relationships. Such a learned set of parameters then enables the model to approximate the function or a class of functions that relate the input distribution to the corresponding outputs. Indeed, one may ask whether a combination of both expert sonographer knowledge and the ability of deep networks to act as function approximators can open new dimensions towards automated ultrasound image analysis, or whether the inclusion of human expertise may help offset the heavy parameter and computational budget required by most current deep learning pipelines.

In that light, an attempt is made in this chapter to leverage expert sonographer knowledge for enhancing the performance of models with a lower parameterization and low memory footprint as compared to traditional models used to obtain state-of-the-art results on predefined tasks for standard datasets. Intuitively, such reduced models would have performance metrics such as accuracy well below heavy models with a large parameter set given an identical number of training epochs and conditions but these models come with the advantage of potentially higher inference speeds, lower memory footprint and a better suitability for implementation on lower cost hardware. Attempts towards improving the performance metrics of such reduced models is of significant importance towards enabling deep learning based models to be more versatile and usable in clinical settings, ranging from critical emergency care to on-the-field requirements in conflict zones and makeshift clinics. Current deep models for medical image analysis are recognized as typically having large memory footprints and inference costs, which are at odds with the increased focus on portability and desired consumer-grade quality even under low-resource use cases [123].

While there have been studies on overparameterization of deep networks [127], efficient models have largely been defined empirically rather than using well-

principled approaches. A question that needs to be asked is whether the performance gains in the desired tasks and datasets achieved by a successively staged increase in the number of model parameters optimized for are desirable under the constraints of on-device memory, power requirements and inference times after deployment. This is probably best answered in terms of a trade-off between model performance, memory footprint and inference time and energy costs.

Most of the research in medical image analysis has focused on improving standard performance metrics such as classification accuracy or Dice scores and similar measures. The constraints of memory and power requirements during real-time inference are not particularly pronounced for ultrasound imaging due to its applications in administering emergency care, maternal and prenatal care and similar interventions driven by public health concerns. This desired portability of ultrasound imaging and its versatile applications has led to the development of low spec, portable devices which need to economize on the memory and power requirements during operation. In this chapter we explore efficient models using a combination of video and expert knowledge, defined by gaze tracking as a sonographer acquires an ultrasound video. That said, the utility for reductions in inference time or memory footprint would still be of relevance in more capital intensive imaging modalities such as MRI or CT given the push towards remote reading and acquisition systems that are usually not equipped with GPUs or similar parallel processing hardware modifications. The requirement becomes particularly unique for ultrasound because of the need to couple acquisition and reading in near real-time in most clinical situations rather than primarily diagnosing separately from the acquisition as is often the case with many other imaging modalities. We propose a novel approach, Perception and Transfer for Reduced Architectures (PeTRA), wherein we use a teacher-student knowledge transfer framework in which human expert knowledge is combined with ultrasound video frames as input to a large *teacher* model, whose output and intermediate feature maps are used to condition compact *student* models. We define a compact model as one that has a significantly reduced number of parameters and lower memory requirement compared to state-of-the-art models.

Our objective is to achieve competitive accuracies with such compact models for our ultrasound image analysis tasks described in subsequent sections.

Note that there exists a rich literature on other strategies for model compression and acceleration of inference speeds in machine learning and computer vision literature, notably through methods like pruning and quantization. Pruning essentially involves a reduction of the model size by inducing sparsity on the basis of parameter or layer importance as judged by an appropriate metric such as Fisher information metrics or Fisher-Rao norms. The resulting model is a reduced version of the original model in terms of the number of connections or layers. Thus, it is a sub-graph of the original model and employs design strategies similar to the original architecture, i.e., a pruned version of a VGG-16 would still contain a sequence of convolutional blocks followed potentially by fully-connected layers, or a pruned version of an Inception-V3 would still use Inception blocks as the fundamental component of the model. Pruning as a process does not allow for situations where the heavier or deeper model employs a given design methodology, say an Inception block based design, and the lighter model has a different kind of a building block, say a squeeze-excite module such as in SqueezeNets. Such a situation may arise in order to make the model to be deployed better suited to the hardware requirements of the target device or due to limitations on the ability to adequately compute parameter importance for complex datasets such as medical imaging or video datasets since the Fisher information matrix calculations are computationally expensive for models with a large number of parameters and a diagonal approximation of the Fisher information matrix potentially ignores the importance of intermediate learnt representations of the model. This chapter is primarily motivated by the possibility of using human expert knowledge through the sonographer gaze information towards improving the performance of lighter or reduced models, instead of trying to reduce a pre-trained heavy or deep model. Therefore, while pruning and quantization methods can be directions for extending this research, they are not used as part of the experiments that seek to demonstrate

the usefulness of using human expert knowledge rather than achieve competitive compression properties for deep models.

6.2 Contributions.

We propose a framework, Perception and Transfer for Reduced Architectures (PeTRA) which combines model knowledge transfer and expert knowledge cues. Our contributions are: 1) to train compact models using both final and intermediate knowledge distillation from large models for the exemplar task of anatomy classification of fetal abdomen, head, and femur frames from a free-hand fetal ultrasound sequence; 2) to incorporate sonographer knowledge in the form of gaze tracking data into a teacher model to enhance knowledge transfer. The knowledge transfer process across models is also compared between situations of inclusion of expert eye fixation information versus otherwise. The section describes work that was first described in the following peer-reviewed conference publication:

- A. Patra, Y. Cai, P. Chatelain, H. Sharma, L. Drukker, A. Papageorgiou, J.A. Noble, ‘Efficient Ultrasound Image Analysis Models with Sonographer Gaze Assisted Distillation’, MICCAI 2019.

This chapter includes work that was a part of a collaboration. I developed the algorithms for gaze-assisted knowledge transfer in the knowledge distillation and intermediate transfer stages. The task of extracting sonographer gaze fixation maps was performed by Yifan Cai. The dataset used was obtained as a part of the PULSE study [24] and was curated and annotated by the other authors.

6.3 Theory

Early detection of Intra-uterine Growth Restriction (IUGR) and subsequent management of fetuses at high risk of low birth weight are crucial in the reduction of perinatal mortality, 98% of which occurs in the developing world [123]. Two dimensional (2-D) B-Mode ultrasound (US) is a common imaging modality for fetal growth monitoring and screening for potential anomalies during fetal development,

e.g. small for gestational age (SGA). The three common biometric measurements taken during second and third trimester antenatal scans are the Head Circumference (HC, Abdominal Circumference (AC) and Femur Length (FL); these are measured and compared against fetal growth standards to monitor fetal growth. The first step towards obtaining such biometry planes automatically is to be able to identify and distinguish between the femur, head and abdomen frames from a free-hand ultrasound sequence. This chapter begins with this exemplar task which is fundamentally a multi-label classification problem. The general theoretical formulation is described in subsequent sections, followed by a specific implementation for the anatomy classification task proposed above.

The difference in the number of parameters in the heavy teacher model and the reduced student model implies that in the gradient descent based algorithms reliant on a stochastic evaluation of parameters towards a minimum, the heavier model has a larger number of possible parameter configurations that can achieve the optimal value of the objective function leading to a quicker discovery of a set of such parameters. These parameters then encode the knowledge representation from a given data distribution as discovered in the gradient-based optimization process. The lighter student model is likely to have a lesser number of possible trajectories towards a similar minimization for the objective. In theory, given an infinite number of epochs, both models should be able to achieve the global minima and approximate the input data distribution without further intervention. In practice however, running the training routine involves computational resources which have a footprint in economic and environmental terms. In a finite number of epochs, the heavier model therefore has a better chance to achieving a set of parameter configurations that is better suited to approximate the input distribution as compared to a lighter student model.

Consider a K -class classification problem, which consists in finding the label $k \in [1, K]$ for an input $\mathbf{x}$. The mapping from the input space to the label space is assumed to be approximated by a deep neural network based model that relies on obtaining optimal parameters for such an approximation by optimizing the features

derived from the input instances towards their associated labels. The output of the neural network can then take the form $\mathbf{c} = \text{softmax}(\mathbf{z}) \in \mathbb{R}^K$, where $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^K$ is the raw output of the last layer, or *logits*. We use a one-hot encoding for the classification target, such that, for a class $k \in [1, K]$, the corresponding target is $\mathbf{y} = (y_i)_1^K$ with $y_k = 1$ and $\forall i : i \neq k, y_i = 0$. The most commonly used loss function for multi-class classification is the categorical cross-entropy formulation [35]:

$$L_c(\mathbf{y}, \mathbf{c}) = - \sum_{i=1}^K y_i \log c_i. \quad (6.1)$$

6.3.1 Knowledge Transfer

In neural networks with gradient-based training, network depth increases the nonlinearity of the function approximation process. While it may be possible to obtain multiple shallower networks with comparable performances, such a search is a trial-and-error exercise due to a lack of guarantees relating architecture and performance for given datasets. Thus, alternative approaches that can condition the optimization process itself in the direction of desired outcome are sought to reduce the number of shallow models to be examined for optimal performance-size trade-offs. Heavier models with a high parameter budget can attain a better generalization performance on a given task as compared to models with a lower number of parameters given an equal number of training epochs on a particular dataset. The latter evidently have a better inference time characteristics in deployment situations. There is a strong case for trying to obtain the best of both worlds - a good generalization performance in a reduced model. Here transfer of learnt logits from cumbersome models to the optimization process of smaller models becomes relevant. We consider such transfer as those of pre-softmax logits with distillation at this end-stage alone, and secondly, intermediate knowledge transfer along with end-stage distillation.

Let $\mathcal{T}$ be a large *teacher* model and $\mathcal{S}$ a smaller *student* model. Model compression by knowledge transfer, first introduced in [147], consists in using the representations learnt by $\mathcal{T}$ to guide the training of $\mathcal{S}$ (Figure 6.1). The key

principle is that it is easier to learn using the representation of the teacher than it is to learn from the original input in the first place. As the teacher model is configured to incorporate many times the number of trainable parameters as the student models, the optimization process on a given data distribution is likely to enable superior representation learning for the same number of training cycles or epochs. Therefore, the final pre-softmax neurons in such a teacher model is considered to be significantly more efficient in encoding the salient features of the overall dataset as relevant towards the final classification task (or any other relevant task in a general example).

As the student model possesses a parameter set dimension that is orders of magnitude smaller, the prospect of arriving at a pre-softmax embedding of a similar information content as that of the teacher is rarely feasible upon direct training from the original instances used to train the teacher model. In this situation, a possibility arises to actually direct the learnable neurons in the pre-softmax layers of the student model towards approximating the probability distributions learnt at a similar stage in the teacher. This effectively acts as an additional supervisory signal complementary to the one-hot encoding of the actual labels. This is interpreted as the knowledge of the heavier teacher model being leveraged to regularize the learnt representation of the reduced student model. Such a knowledge transfer is also possible at intermediate stages of the model, and the degree of transfer is representative of the influence of the learnt representations of the teacher on the learning process of the student. The chapter considers cases where final layer knowledge distillation is performed, and those where different configurations of intermediate transfer were investigated between different pairs of teacher and student models.

A. Final Layer Knowledge Distillation.

Let $\mathbf{z}^t$ and $\mathbf{z}^s$ be the logits of the final layer of $\mathcal{T}$ and $\mathcal{S}$, respectively. Following [133], we model the knowledge transfer stage as a simultaneous step with the cross-entropy objective minimization and incorporate the teacher knowledge by adding a distillation loss to the cross-entropy loss as:

$$L = \alpha L_c + \beta L_d \tag{6.2}$$

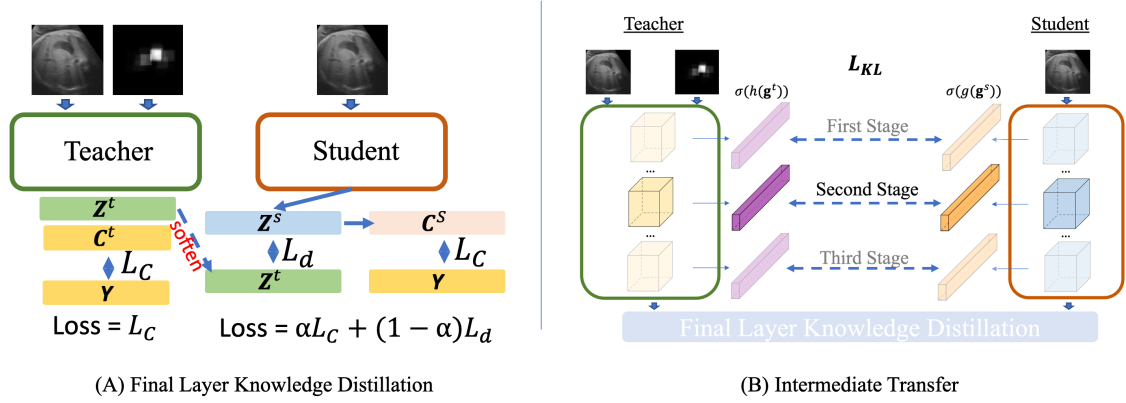


Figure 6.1: Schematic of our proposed knowledge-distillation pipeline. (A) Final Layer knowledge distillation (B) Intermediate Transfer.

where L_c is the cross-entropy loss defined in Equation 6.1,

$$L_d = - \sum_{i=1}^K \text{softmax} \left(\frac{z_i^s}{E} \right) \log \left(\text{softmax} \left(\frac{z_i^t}{E} \right) \right) \quad (6.3)$$

is the distillation loss between teacher and student, and $\alpha, \beta > 0$ are hyper-parameters controlling the relative influence of both terms.

E is a *temperature* term introduced by [133] as a form of relaxation to *soften* $\mathbf{z}^t$ and $\mathbf{z}^s$. Indeed, having been obtained by a cross-entropy objective in $\mathcal{T}$, $\mathbf{z}^t$ may be too close to the one-hot target vector $\mathbf{y}$. Softening provides more information about the relative similarity of classes rather than absolute maxima. The usage of a temperature term allows mitigating the steepness of the difference between probabilities of classes that share a relatively higher degree of feature similarity. The standard softmax expression devoid of additional temperature terms causes the probability of the correct class to approach a distribution of probabilities close to the hard maximum distribution of the one-hot encoding. Upon dividing the pre-softmax activations or logits by the temperature value, the post-softmax distribution is ‘smoothened’ out in comparison to the distribution one would obtain if the temperature value was set at 1.0. This means that the post-softmax distribution now has a reduced differential between the first and second most confident prediction and so on. Therefore, this set of activated neurons can encapsulate the teacher model’s knowledge in terms of relative difference between classes in the dataset and be used to transfer knowledge about such fine differences to the student training process.

B. Intermediate Transfer (IT).

To leverage knowledge contained in intermediate representations of the teacher model, we consider intermediate knowledge transfer inspired by the idea of *hint learning* [134], in conjunction with final layer knowledge distillation. The notion of learning representations from input data over a hierarchy of features across the depth of the network imply the possibility of directing the process of learning in the student model at different intermediate stages as well. This is akin to providing the student a *hint* as to arrive at parameter configurations during the training process that allows it to approximate the teacher’s learnt representations more closely over these different depths. Such a hint is extracted from the teacher model by deriving a fully-connected layer from an intermediate layer. These fully-connected layers and their analogous fully-connected layers from the student model (that can be understood as ‘receiving’ the hint by being conditioned by the teacher at intermediate stages) are termed as *arms*. Let $\mathbf{g}^t$ be the output of an intermediate layer of the teacher model. It is used to produce a *hint* $\sigma(h(\mathbf{g}^t))$, where σ is a sigmoid activation function and h is a fully-connected (FC) layer. Similarly, an intermediate layer of the student model (a *guided* layer) is used to produce a regularization output $\sigma(g(\mathbf{g}^s))$, where g is a fully-connected layer with the same output dimension as h . The hint is used to train the guided layer with a Kullback-Leibler (KL) loss:

$$L_{\text{KL}} = - \sum_j \sigma(g(\mathbf{g}^s))_j \log \left(\frac{\sigma(h(\mathbf{g}^t))_j}{\sigma(g(\mathbf{g}^s))_j} \right) \quad (6.4)$$

This creates a teacher model fully-connected layer or *arm* (in purple, Figure 6.1(B)) whose logits are associated with the student model fully-connected *arm* (in orange, Figure 6.1(B)) in a KL divergence objective aimed at optimizing learned intermediate representations in the student model by supervising them with corresponding teacher model values (Figure 6.1). The rationale for using KL divergence lies in its relationship with the empirical Fisher information matrix [150] that implicitly allows the estimation of parameter importance while prioritizing feature transfer. This intermediate KL divergence operation therefore implements a regularization of the student learning using the most attentive intermediate

features from the teacher, indirectly relying on the gaze information once again to enforce such regularization.

The KL divergence term is added to the optimization objective in Equation 6.2 in training:

$$L = \alpha L_c + \beta L_d + \gamma L_{\text{KL}}, \quad (6.5)$$

where $\gamma > 0$ is a hyperparameter controlling the influence of IT.

After training, the fully-connected (FC) arm is truncated. Resulting models have the same number of parameters as in the final layer knowledge distillation case, but with improved knowledge transfer from the teacher through intermediate layers.

6.3.2 Learning from Human Knowledge

The methods described in this chapter model the visual attention of a human expert looking at an image $\mathbf{I}$ through a gaze map $\mathbf{G}$. $\mathbf{G}$ is generated by recording the point-of-gaze of the expert sonographers while looking at $\mathbf{I}$. To perform gaze-assisted knowledge distillation, we train the *teacher* model $\mathcal{T}$ to perform a classification task using both $\mathbf{I}$ and $\mathbf{G}$ as input [151]. The student models still only 'see' the image $\mathbf{I}$ during its training cycle. The parameters of the teacher model then have undergone an optimization process that sought to establish a function between an output that was a clinically attached label and an input space that comprised of both the ultrasound image from the machine and a representation of human expert judgement about it. The student model is expected to be trained and evaluated without direct access to the expert judgement at any time. Thus, through the knowledge transfer protocols proposed above, the teacher model can transfer not only the knowledge learned through its high number of parameters but also knowledge extracted from the human visual attention. We implement two different teacher model architectures for learning from image and gaze: $\mathcal{T}_{+\text{gaze}}$ obtained by concatenation of extracted features of inputs (frame and gaze map) and $\mathcal{T}_{\times\text{gaze}}$ by computing the element-wise product between resized gaze maps (28×28) and feature maps extracted from ultrasound frames (Figure 6.11). These are two strategies for incorporating gaze

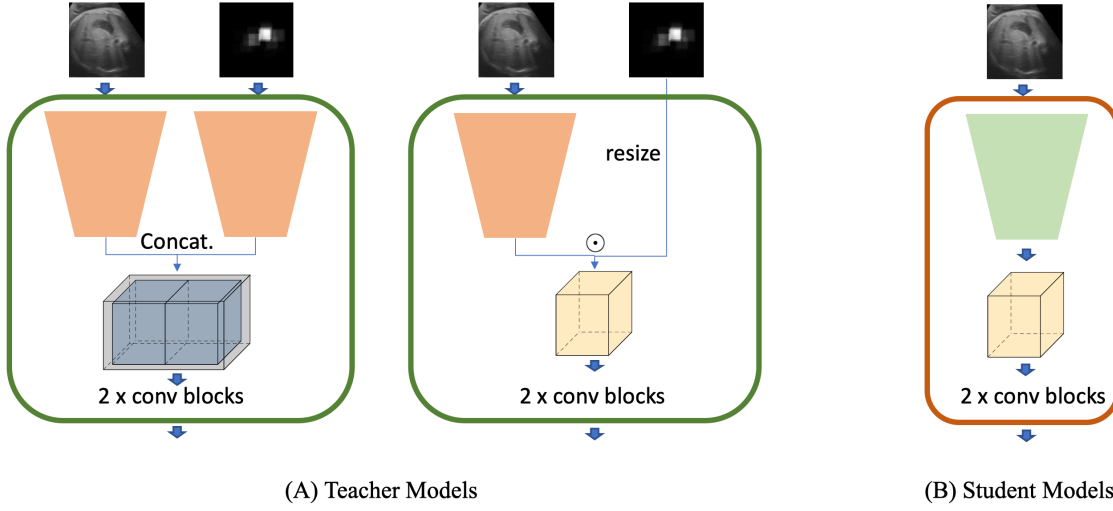


Figure 6.2: Teacher and Student models used. (A) Teachers use concatenation or element-wise production to merge information from US image and visual attention map. (B) student only takes US image as input.

information into the deep neural architecture, the former being a treatment of gaze maps as equivalent to the ultrasound frames from the standpoint of the neural network input layers and the latter implying that gaze information is akin to the attentive features learnt over the layers of the deep neural network and such gaze maps should therefore be introduced deeper into the networks.

6.4 Data and Training

6.4.1 Data

Clinical fetal ultrasound videos with simultaneously recorded sonographer gaze tracking data was available for 89 healthy subjects from the PULSE study, which is an interdisciplinary effort to understand how perceptual cues used by sonographers while performing diagnostic interpretations of clinical ultrasound videos and how such perceptual cues can be combined within machine learning algorithms to provide assistive technologies for sonography workflows [24] described in section 1.2.3 in Chapter 1.

Ethics approval was obtained for data recording and the datasets were stored in accordance with local data governance rules. From this dataset we extracted 23016 abdomen, 24508 head, 12839 femur frames. Gaze tracking data was recorded using a

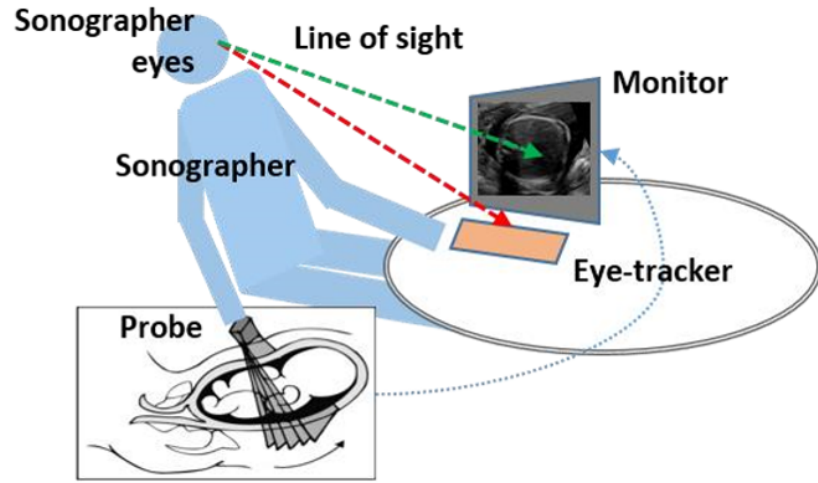


Figure 6.3: A schematic of an ultrasound session related to the PULSE study in progress with the sonographer gaze being tracked using the eye-tracker as they look over video frames while conducting a free-hand sweep. This schematic is to demonstrate the acquisition process for the data used in our experiments.

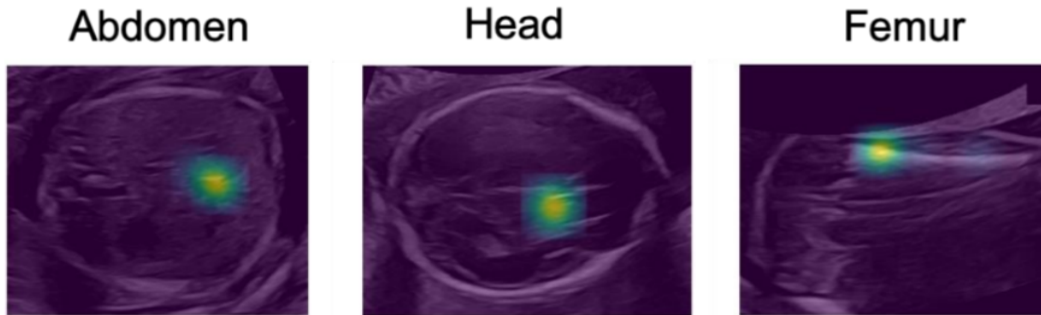


Figure 6.4: Gaze maps derived from eye-tracking data superimposed on exemplar frames of the abdomen, head and femur scan.

Tobii Eye Tracker 4C (Tobii, Sweden) that records the point-of-gaze (the relative x and y coordinates along with corresponding timestamps) at a rate of 90, effectively recording 3 gaze points per frame of the ultrasound sequence being observed. Often, the gaze points would be close to one another relative to the frame width as the structure of interest being observed in a given frame may have all the relevant features close together. Having multiple gaze maps then would cause an overlap and it is convenient to merge fixation points that are too close as a single fixation point.

Gaze points that were less than 0.5 degrees apart (as per the visual angle, which is the angle that would be formed between two lines of sight between the eyes of the

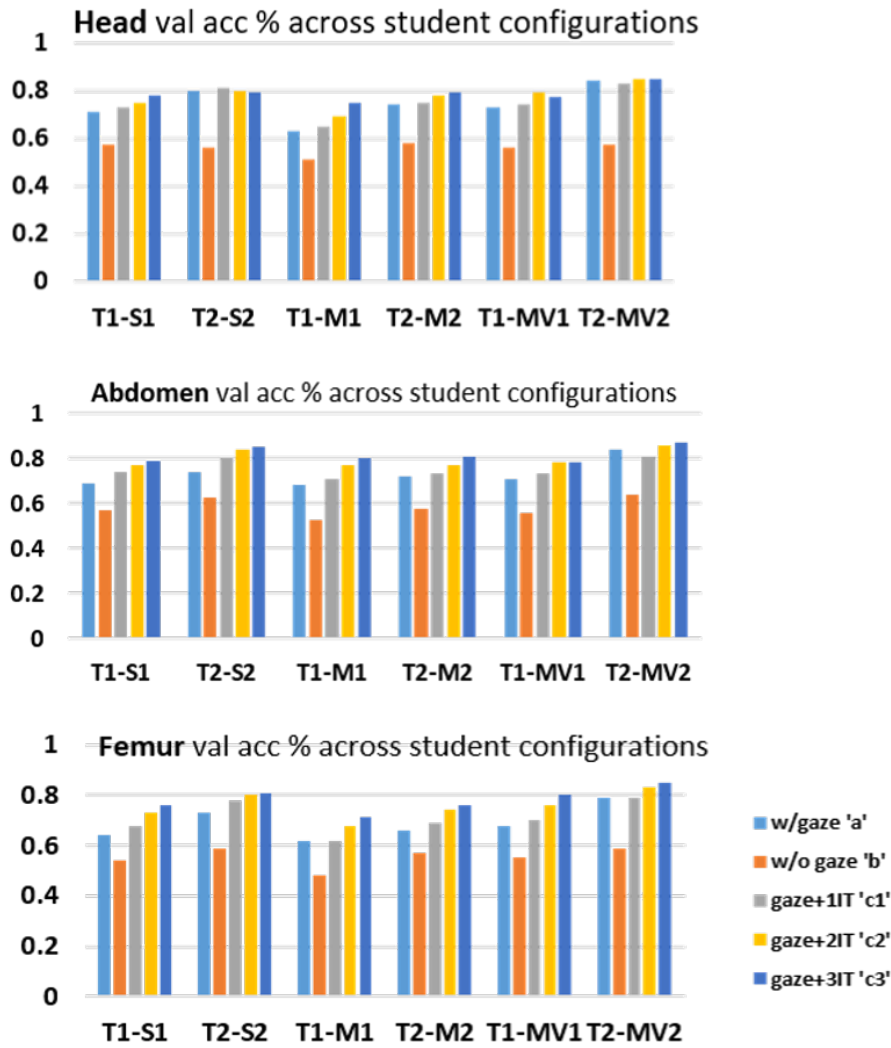


Figure 6.5: Across student configurations (based on the teacher-student combination), validation accuracy upon the ultrasound frames for different states of gaze maps being shown to the teacher and teacher-student intermediate transfer are considered (light blue – teacher trained with gaze maps (w/gaze), orange – teacher trained only on frames (w/o gaze), gray – teacher trained with gaze and intermediate transfer at first stage (gaze+1IT), yellow - teacher trained with gaze and intermediate transfer at second stage (gaze+2IT), dark blue- teacher trained with gaze and intermediate transfer at third stage (gaze+3IT))

sonographer and the point observed on the screen) were merged as a single fixation point in the study presented here. A sonographer visual attention map G was generated for each frame by adding a truncated Gaussian with width corresponding to a visual angle of 0.5 degrees at the point of fixation. The gaze map was obtained on the same frame dimensions as the actual image frame.

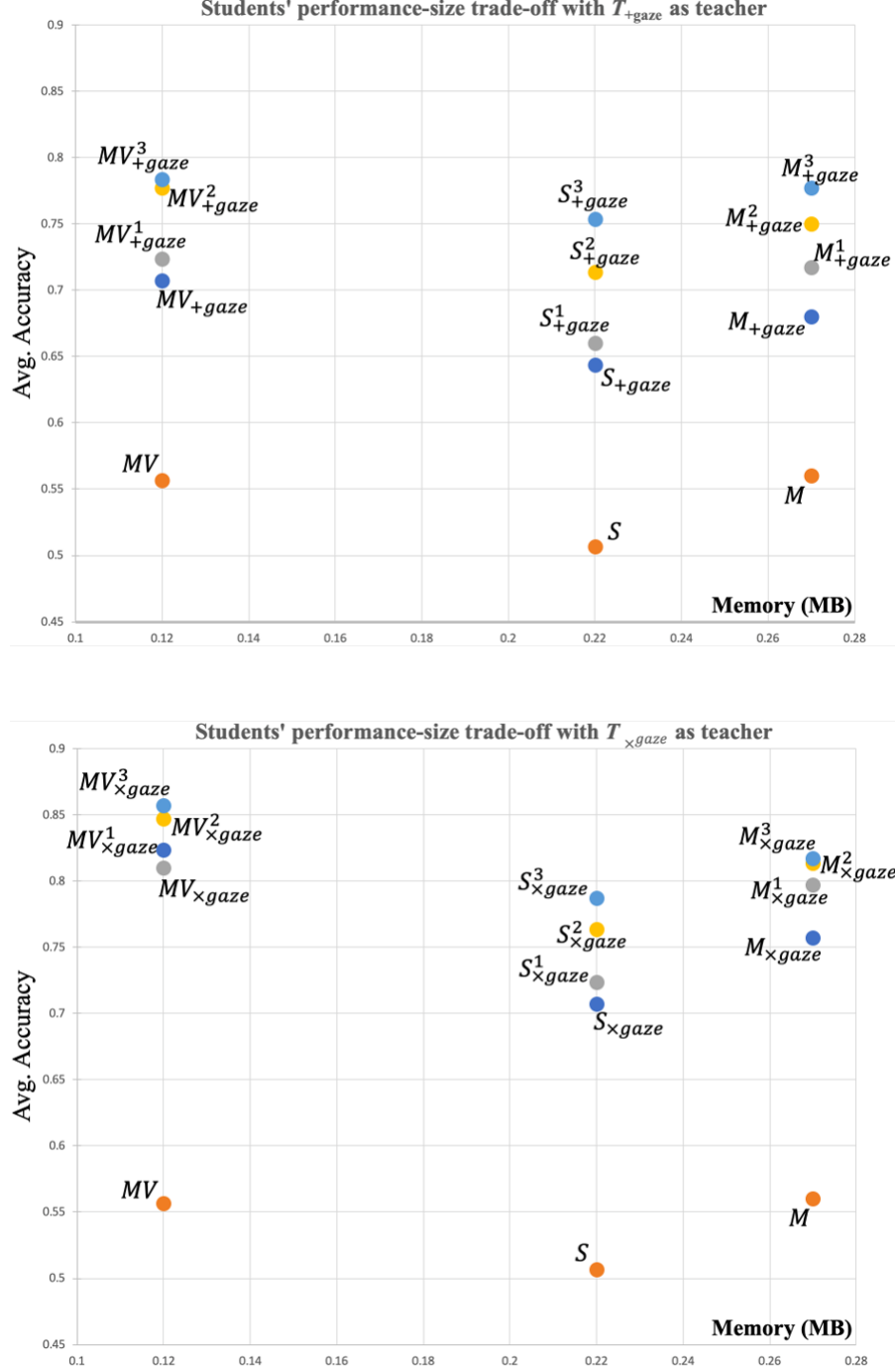


Figure 6.6: Performance-size trade-off. Top: $T_{\times gaze}$ -trained students; Bottom: T_{+gaze} -trained students. Accuracy is averaged across classes. Teacher trained with Gaze + 3rd IT achieves best results overall.

6.4.2 Training Details

We tested different student models to demonstrate the utility of the PeTRA approach: *SqueezeNet* [152] (**S**), *MobileNet* (0.25 width multiplier) [153] (**M**), and *MobileNet* v2 (0.35 width multiplier) [154] (**MV**) modified to accept single channel inputs and include a joint loss objective in Equation 6.5. These models are representative of the main types of compact architectures — squeeze-excite convolution blocks [152], group convolutions [153] and depthwise separable convolutions in groups [154]. Most other compact models proposed in computer vision literature derive from these basic architectures, apart from those models that are pruned versions of bigger models. We describe the student model architectures in figures Figure 6.7, Figure 6.8 and Figure 6.9.

For the teacher models we use a *VGG-16* feature extractor, modified to accept dual input of single-channel frames and gaze maps (with the number of neurons in the first two fully connected layers changed from the original 4096 to 1024 and 512 respectively to avoid overfitting as the number of classes considered here is 3, compared to 1000 in the ImageNet dataset). In a change to the standard VGG-16, for $\mathcal{T}_{\times\text{gaze}}$, the element-wise product after the fourth convolutional block is followed by the fully-connected layers. In $\mathcal{T}_{+\text{gaze}}$, features are extracted by parallel convolutional blocks of the VGG-16 and concatenated before the fully-connected layers.

At inference, only one of the parallel blocks (processing single frame input, as gaze maps are not used at inference) and the following fully-connected layers comprise the $\mathcal{T}_{\times\text{gaze}}$ model. This reflects in $\mathcal{T}_{\times\text{gaze}}$ having same number of parameters as $\mathcal{T}$ in Table 1. Data augmentation was performed using a 20 degrees rotational augmentation and horizontal flipping for both ultrasound and gaze map frames. Frames and corresponding gaze maps were resized to 224×224 . The training and testing split is decided on a subject-wise basis in line with the norm in medical imaging tasks so as to ensure that data from a single subject do not end up in both the training and test datasets. All models were trained on 71 subjects and tested on 18 subjects of the dataset corresponding roughly to a 80:20 split over the total number of subjects in our dataset. Teacher models were trained for 100 epochs with

learning rate of 0.005 and optimized using an adaptive moment estimation optimizer (Adam). Student models were trained for 200 epochs over the $(N, image, label, logit)$ set created for all N frames passed to the teacher model. The softening temperature value was set to 4.0 after a grid search for $E \in [1, 10]$. We investigated intermediate transfer at three different stages. First, second and third stage intermediate transfer was respectively applied from the 2nd, 4th, 5th maxpool layers in the teacher model to the fully-connected arms after 2nd, 3rd, 5th maxpooling layers of S and 3rd, 5th, 7th depthwise conv layer for M and MV . For experiments with intermediate transfer, such fully-connected layer neurons were separately retained and appended to the set as $(N, image, label, logit, IT_1/./IT_m)$. α and β are set to 0.5 for equal influence of teacher knowledge and cross-entropy loss for the student model; γ is set at 1.

6.5 Experiments, Results and Discussion

We report the classification accuracy for MobileNetV2 (MV) in Table 1, and the accuracy of teacher models and compact models trained without knowledge transfer in Table 2. We also report the number of parameters, memory requirement and inference time of models in Table 2. Complete overall results for the variants of students are shown in Figure 6.6. Student models are named X^l , X^l_{+gaze} and $X^l_{\times gaze}$ when trained using knowledge from $\mathcal{T}$, $\mathcal{T}_{+gaze}$ and $\mathcal{T}_{\times gaze}$, respectively. $X \in \{S, M, MV\}$ is the student architecture and $l \in \{1, 2, 3\}$ is the stage used for intermediate transfer.

6.5.1 Performance

Final layer knowledge distillation improves the accuracy of the compact model compared to training without knowledge transfer for all the student models. Compact models trained using gaze-assisted knowledge distillation reach a higher accuracy than the same models trained with image-only knowledge distillation (+5% for MV_{+gaze} , +16% for $MV_{\times gaze}$, compared to MV). Intermediate transfer further improves knowledge transfer over final layer distillation only, with transfers at the 3rd level showing the best improvement in student model accuracy (+7% for MV^3_{+gaze} ,

+3% for $MV_{\times\text{gaze}}^3$, compared to $MV_{+\text{gaze}}$, $MV_{\times\text{gaze}}$). These trends are seen for all student models (S , M , MV). The image-only knowledge distillation is analogous to prior work in [133] and [135] which studied the efficacy of knowledge distillation between teacher and student models in improving the knowledge representation in the latter. Thus, the baseline results for the efficiency gains for the student models in the knowledge distillation framework used here are those that do not involve the usage of gaze maps when training the heavier teacher model. The difference in the performance in the student model when trained with the incorporation of the sonographer gaze information is then attributed as the model performance gain through the inclusion of human expert knowledge [155].

Across student configurations (based on the teacher-student combination) validation accuracy for the abdomen, head and femur frames in different states of gaze maps being shown to the teacher and teacher-student intermediate transfer are considered. It is evident in Figure 6.5 that across classes, the best performances for the student models are obtained upon inclusion of gaze maps and the higher intermediate transfer stages. Additionally, for equivalent intermediate transfer stages and gaze map inclusion strategies, the student model based on the MobileNet V2 implementation is observed to attain the best classification performances for all the classes considered.

6.5.2 Computational Complexity

We evaluate the computational complexity of the models by computing the number of floating point operations (FLOPs) performed for inference as reported in Table 2. We also report inference times for a batch of 100 frames from the test set using a 32GB/Intel core i7-4940MX/3.1Ghz laptop. The inference speed-up compared to $\mathcal{T}_{\times\text{gaze}}$ is $5\times$ for S , $8\times$ for MV and $4\times$ for M . For the same student models, using human gaze in teacher training does not change the computational complexity as the number of parameters involved in the computational graph stays the same, but the performance metrics of student models when distilled from gaze-trained teachers are superior.

6.5.3 Memory size

Student architectures (S , M , MV) show a $1000\times$ to $7000\times$ reduction of memory size compared to teachers (Table 2). The MV_{+gaze}^3 student with only 284850 parameters achieves an average accuracy of 0.85, close to its teacher model $\mathcal{T}_{\times gaze}$ (0.90), and higher than $\mathcal{T}_{+gaze}$ (0.79) and $\mathcal{T}$ without gaze (0.73). Similar gains are seen for other students as well (Figure 6.6). The MobileNet model M (738658 parameters, 270) trained with distillation from $\mathcal{T}_{\times gaze}$ attains accuracy (0.79) comparable to $\mathcal{T}_{+gaze}$ and higher than $\mathcal{T}$. Due to element-wise product operations, $\mathcal{T}_{\times gaze}$ has a higher number of parameters than $\mathcal{T}_{+gaze}$.

6.5.4 Model efficiency

To evaluate model efficiency as a trade-off between accuracy a , number of parameters p and computational cost c , we estimated the *NetScore* metric $\Omega = 20 \log(a^\delta p^{-\epsilon} c^{-\phi})$ proposed in [156]. The NetScore metric was chosen for this purpose because it is directly proportional to the performance measures such as accuracy, and inversely proportional to the model size and number of computations required. Such an evaluation metric is suited in this case because empirical characteristics used to determine the efficacy of reduced models revolves around their ability to offer superior performance at a reduced memory footprint and inference time. We provide other model data in Table 2 for completeness, and in-depth comparison over different aspects such as actual inference times, number of parameters used and actual size on disk.

Similar to [156] we set $\delta = 2$, $\epsilon = 0.5$ and $\phi = 0.5$, with these values obtained through experiments on a range of models proposed for mobile deep learning applications [156] and the importance of accuracy being somewhat prioritized over that of the model size and computational power requirements but with the caveat that the latter two criteria are still significantly represented. For computational cost c , we use the number of FLOP instead of multiply-accumulate (MAC) operations in [156] because FLOP includes overheads such as pooling and activation beyond dot product and convolution operations. We report MFLOP (million FLOP) and

NetScore values in Table 2. The units of a , p , c in Ω are percent, millions of parameters and MFLOPs. The best NetScore value is seen to be obtained by $MV_{\times \text{gaze}}^3$, the most compact model with the highest accuracy. Thus, one aspect to note is that not only does the ability to include human knowledge help improve performance even beyond the original models that were directly trained on it, the compressed or reduced model itself has certain characteristics that influence its ability to efficiently adapt to such teacher knowledge. In other words, how good a 'learner' a student model is would also depend on the nature of the student model architecture as evident from the success of the inverted residual approach in MobileNet v2 designs evident here. The additional insight here is that not only do such strategies to make models lightweight help with the size-performance tradeoffs when directly learning from input data, but also the ability to learn by proxy as in the case of teacher-student distillation shown here.

The best performing reduced models achieve within 5% of the accuracy of full models with 1000x fewer parameters. The reduction of memory footprint and inference times make them potentially attractive for deployment in a clinical setting on equipment with lower computational power and memory storage.

6.6 Concluding remarks

The chapter proposed Perception and Transfer for Reduced Architectures as a general framework to train compact models with knowledge transfer from traditional large deep learning models using gaze tracking information to condition the solution without requiring such information at runtime. For the tasks of fetal abdomen, femur and head detection, compact models had accuracies close to the large models, while having a much lower memory requirement. The best performing reduced models achieve within 5% of the accuracy of full models with 1000x fewer parameters. The reduction of memory footprint and inference times make the reduced models very attractive for deployment in a clinical setting on lower spec (and hence cheaper) equipment. The significance of a 5% accuracy reduction for an assistive technology given other pertinent constraints depends on the final end-use application, as many

healthcare settings may prioritize flexibility of operating conditions and accept a slight decline in accuracy of the diagnostic software. Usability studies on this trade off and other aspects in a given clinical setting forms a part of future translational studies. We chose our ultrasound task as the classification of these three biometry planes deliberately as biometry assessment is a fundamental diagnostic test in obstetric settings around the world and is performed at a range of healthcare centers and hospitals with a wide diversity of equipment in use.

While ultrasound image analysis with deep models help reduce need for specialized expertise, they need to be lightweight for use on such low cost devices and on portable handheld machines for low-resource and emergency settings. So memory footprint is as relevant as accuracy metrics for such models to be clinically worthwhile in systems where access to medical imaging is financially or operationally constrained. This is a proof-of-concept of human knowledge-assisted model compression for image analysis and the concept could be used for other modalities.

Conv1-96
Maxpool2
fire1_squeeze = Conv1-16
fire1_expand1 = Conv1-16
fire1_expand2 = Conv3-32
merge1 = concat([fire1_expand1, fire1_expand2])
fire2_squeeze = Conv1-16
fire2_expand1 = Conv1-16
fire2_expand2 = Conv3-64
Maxpool2
merge2 = concat([fire2_expand1, fire2_expand2])
fire3_squeeze = Conv1-32
fire3_expand1 = Conv1-32
fire3_expand2 = Conv3-64
merge3 = concat([fire3_expand1, fire3_expand2])
Maxpool2
fire4_squeeze = Conv1-32
fire4_expand1 = Conv1-32
fire4_expand2 = Conv3-128
merge4 = concat([fire4_expand1, fire4_expand2])
fire5_squeeze = Conv1-48
fire5_expand1 = Conv1-48
fire5_expand2 = Conv3-192
merge5 = concat([fire5_expand1, fire5_expand2])
Maxpool2
fire6_squeeze = Conv1-48
fire6_expand1 = Conv1-48
fire6_expand2 = Conv3-192
merge6 = concat([fire6_expand1, fire6_expand2])
fire7_squeeze = Conv1-64
fire7_expand1 = Conv1-64
fire7_expand2 = Conv3-256
merge7 = concat([fire7_expand1, fire7_expand2])
Maxpool2
fire8_squeeze = Conv1-64
fire8_expand1 = Conv1-64
fire8_expand2 = Conv3-256
merge8 = concat([fire8_expand1, fire8_expand2])
Dropout 0.5
Conv1-10
Global avg pool 13 x 13
FC – N
Softmax

Figure 6.7: Architecture for student S. Broadly inspired by the squeeze-excite method in SqueezeNet [152], a similar computational graph is used with ‘fire’ modules of 1 x 1 convolutional filters (conv1) followed by 3 x 3 convolutional (conv3) filters. ‘merge’ refers to a concatenation of the layers listed inside the concat operation following it. The actual number of feature maps output at individual layers is different from the original SqueezeNet models and was designed according to our tasks.

Conv3 - 32 α s2
Conv d/w 3 - 32 α s1
Conv1 - 64 α s1
Conv d/w 3 - 64 α s2
Conv1 - 128 α s1
Conv d/w 3 - 128 α s1
Conv1 - 128 α s1
Conv d/w 3 - 128 α s2
Conv1 - 256 α s1
Conv d/w 3 - 256 α s1
Conv1 - 256 α s1
Conv d/w 3 - 256 α s2
Conv1 - 512 α s1
{{Conv d/w 3 - 512 α s1 Conv1 - 512 α s1}} x 5
Conv d/w 3 - 512 α s2
Conv1 - 1024 α s1
Conv d/w 3 - 1024 α s2
Conv1 - 1024 α s1
avgpool 7 x 7
FC - 512
FC - 512
FC - 256
Dropout 0.5
FC-N
Softmax

Figure 6.8: Architecture of student model M. The width multiplier, α , is the factor of reduction in the number of output channels (for example in Conv3 - 32, the layer involves 3x3 convolutional kernels with 32 channels as output – if $\alpha=0.25$, that is 8 feature channels as output). ‘s1’ implies a stride of 1 and ‘s2’ a stride of 2. The green marking implies that this block is repeated sequentially – in this case 5 times indicated by the x5 in white. The basic structure follows the MobileNet design with a difference in the number of fully-connected layers (3 FC layers before the Dropout layer unlike the original designs). Conv d/w refers to depthwise convolution.

Conv3 - 32 α s2
{conv1 - 16 α Conv d/w 3 - 16 α Conv1 - 16 α } x 1 (1) s1
{conv1 - 24 α Conv d/w 3 - 24 α Conv1 - 24 α } x 2 (6) s2
{conv1 - 32 α Conv d/w 3 - 32 α Conv1 - 32 α } x 3 (6) s2
{conv1 - 64 α Conv d/w 3 - 64 α Conv1 - 64 α } x 4 (6) s2
{conv1 - 96 α Conv d/w 3 - 96 α Conv1 - 96 α } x 3 (6) s1
{conv1 - 160 α Conv d/w 3 - 160 α Conv1 - 160 α } x 3 (6) s2
{conv1 - 320 α Conv d/w 3 - 320 α Conv1 - 320 α } x 1 (6) s1
Conv1 - 320 s1
Avgpool 7x7
Dropout 0.5
FC - N
Softmax

Figure 6.9: Student model MV architecture. The width multiplier is α . Blocks in green are sequentially repeated with identical configurations with the number of such repetitions indicated in white. The original MobileNet v2 design included the concept of expansion or the selective increase in the number of feature channels within certain stages in the network, which we adapt with the factor of expansion shown here in red. The strides are marked in blue (s1 is a stride of 1 and s2 is a stride of 2). As in the other student models, the MV student uses modules and concepts inspired by the original MobileNet v2 design, but is not meant to replicate any of networks as such in the original design.

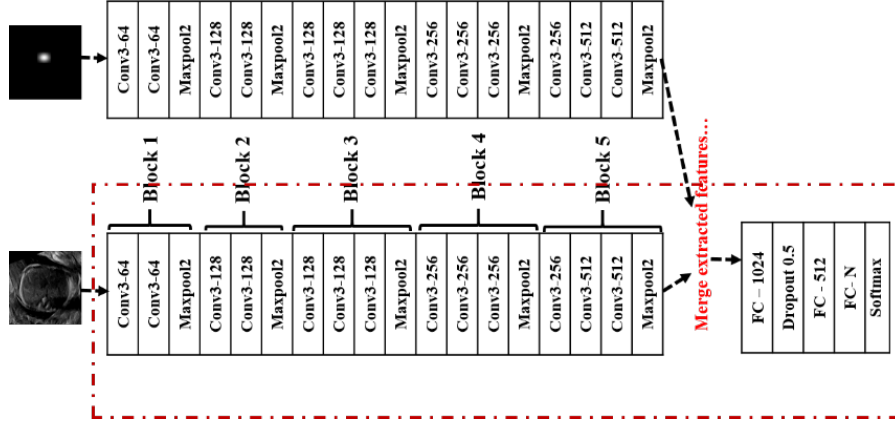


Figure 6.10: The schematic for teacher configuration T+gaze : Convolutional sets derived from the VGG-16 architecture are used to extract features from the input frame ultrasound image and the corresponding gaze map (derived from the gaze tracking data as explained in Sec 2.3). The fully-connected layers are set at 1024 and 512 units. Intermediate convolutional layers with 128 channel outputs in Block 3 and 256 channel outputs in block 4 represent other deviations from the original VGG-16 architecture. These design changes were made specific to our data to reduce the number of parameters involved and control overfitting. While feature extraction was done using convolutional sets with shared parameters, one of the sets is retained with the fully-connected layers as the model used during inference (red dotted rectangle). N (in FC – N) refers to the number of classes. Note: conv3-64 means a layer with 3x3 convolutional kernels and 64 output filters. Maxpool2 refers to a maxpooling operation with a 2x2 pooling window. FC – 1024 implies a fully connected layers with 1024 units. A similar naming convention is used for the rest of the models.

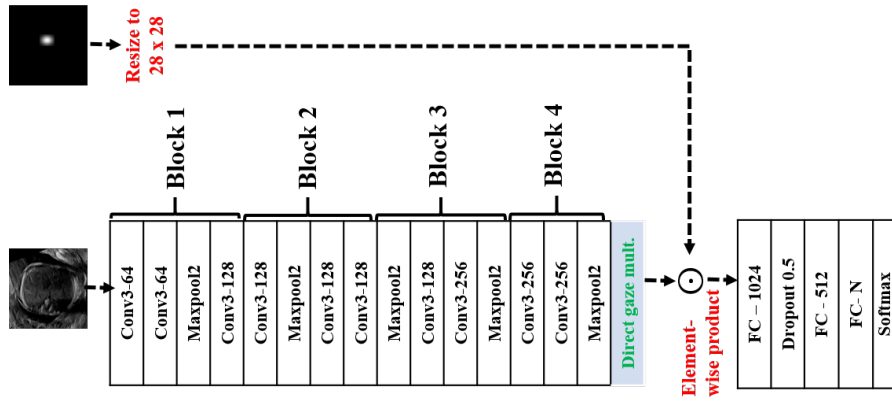


Figure 6.11: The schematic for teacher configuration Tx gaze :Element wise multiplication of the resized (to 28 x 28 pixels). Features are extracted from the input ultrasound frame. After the fifth maxpooling layer, the feature maps are 28 x 28 in dimensions. Element-wise multiplication between each of these feature maps and the resized gaze map is now carried out here. This resultant tensor is passed for further computation into the fully-connected layers. Note the block numbering is kept different from that in the first teacher schematic to account for the different positions of the maxpooling layers in the two modes. N (in FC – N) refers to the number of classes.

Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit...

There is no one who loves pain itself, who seeks after it and wants to have it, simply because it is pain...

— Cicero's *de Finibus Bonorum et Malorum*

7

Conclusion

Contents

7.1	Contributions	153
7.2	Limitations	158
7.3	Future work	161
7.4	Concluding remarks	165

Attempts to explore and address the many dimensions of the fetal ultrasound video analysis problem were made in the course of the preceding chapters, with specific tasks chosen for their clinical relevance, algorithmic challenges or novel human-machine interaction approaches. This chapter summarizes the contributions, their shortcomings and limitations, and discussed avenues for future work around the ideas proposed in the thesis.

7.1 Contributions

The key themes explored in this thesis include:

1. A temporal encoding algorithm for spatiotemporal aggregation of fetal ultrasound video sequences towards visibility assessments, standard fetal cardiac viewing plane identification and cardiac region localization

We proposed an automated video understanding algorithm for fetal ultrasound videos using a purely convolutional architecture design. The proposed variants of the temporal encoding design are used to absorb spatial features information from sequential frames, thereby building a temporally consistent feature representation to be passed on for final classification and localization objectives. The best performing temporal encoding variant, the Hierarchical Temporal Encoding architecture, accepts sets of frames as input in one of the four parallel convolutional stacks and performs a sequence of convolutional operations independently till deeper in the network. At an intermediate depth, the computed feature maps are fused pairwise and passed into another parallel set of convolutional stacks, which again perform a convolutional operation sequence to arrive at sets of high-level feature maps to be fused into an overall feature representation.

This representation is the overall spatiotemporal feature representation for the given video sequence, which is then optimized with respect to the class labels and the localization coordinates (centre coordinates and radius) in a multi-task fashion. Results demonstrated competitive performance with respect to previously explored methods using particle-filtering approaches, and this was achieved without using recurrent architectures, with validations presented on a real-world clinical fetal echocardiography dataset.

2. Multi-anatomy localization from fetal echocardiography video streams

Regionalized analysis of cardiac structures in fetal echocardiography sessions is vital towards detection of potential markers for several congenital heart disease classes. We cast the task of segregating anatomical motions in fetal ultrasound videos as an analogue of the multi-object action recognition problem in scene understanding contexts. The tracking-by-detection approach was combined with a regression to a motion progress property defined for individual structures within the video sequence. Fetal heart motion is an important diagnostic indicator for structural detection and functional assessment of congenital heart disease. We propose an approach towards

integrating deep convolutional and recurrent architectures that utilize localized spatial and temporal features of different anatomical substructures within a global spatiotemporal context for interpretation of fetal echocardiography videos.

3. Hierarchical Class Incremental Learning for fetal echocardiography videos

Chapter 4 proposes ultrasound video interpretation algorithms that enable novel classes or instances to be added over time without significantly compromising prediction abilities on prior representations learned on earlier datasets or classes with a different data distribution as compared to the currently available classes or datasets. The motivating application is diagnostic fetal echocardiography video and frame-level analysis in mid-trimester scans. In this application, a sonographer may acquire multiple video clips of the heart at different points in the full scan. The goal is to make a complete inference of the health of the fetal heart from those multiple clips. Recording full diagnostic fetal echocardiography video is currently not widely done in clinical practice. Such diagnostic videos can more typically be available in varying lengths and summarize a number of diagnostic sub-tasks of varying difficulty. Thus in this application, although large clinical datasets may be available at onset to build well-performing ultrasound image analysis solutions, data may also become available in variable volumes over extended time to assist in algorithm refinement. This suggests a need for ultrasound video interpretation algorithms that enable novel classes or instances to be added over time, without significantly compromising prediction abilities on prior representations.

To address this need, we propose to use incremental learning to build a hierarchical network model that allows for a parallel inclusion of previously unseen anatomical classes without requiring prior data distributions. Super classes are obtained from coarse classification followed by fine classification to allow the model to self-organize anatomical structures in a sequence of categories through a modular architecture.

We show that this approach can be adapted with new variable data distributions without significantly affecting previously learned representations. Two extreme

situations of new data addition are considered; (1) new class data is made available over time with volume and distribution similar to prior available classes, and (2) imbalanced datasets arrive over future time to be learned in a few-shot setting. In either case, availability of data from prior classes is not assumed. The evolution of the learning process is validated using incremental accuracies of fine classification over novel classes and compared to results from an end-to-end transfer learning-derived model fine-tuned on a clinical dataset annotated by experienced sonographers. The modularization of subsequent learning reduces the depreciation in future accuracies over old tasks from 6.75% to 1.10% using balanced increments. The depreciation is reduced from 6.95% to 1.89% with imbalanced data distributions in future increments, while retaining competitive classification accuracies in new additions of fine classes with parameter operations in the same order of magnitude in all stages in both cases.

4. Saliency aware continual learning for ultrasound image analysis

Chapter 4 builds up on the case it made for enabling efficient continual learning pipelines for medical image analysis by extending concepts of interpretability and saliency to the class incremental learning problem. To explore interpretable lifelong learning for deep neural networks in medical imaging, we introduce a perspective of understanding forgetting in neural networks used in ultrasound image analysis through the notions of attention and saliency. Concretely, we propose quantification of forgetting as a decline in the quality of class specific saliency maps after each subsequent task schedule.

Another challenge in the design of continual learning algorithms in machine learning research is to decide upon methodologies for retention of suitable exemplars from prior data distributions in rehearsal and regularization driven approaches. In medical imaging applications, such an optimal retention becomes even more important as a suitable diversity of physiologically relevant image-level information needs to be represented in selective replays during incremental learning sessions. To that end, we introduce a novel method for knowledge transfer from past tasks to

present by a saliency-guided retention strategy for past exemplars, which improves the ability to retain past knowledge while optimizing parameters for current tasks. Additionally, the redefinition of catastrophic forgetting measures as an equivalent of saliency map quality makes the incremental learning pipeline interpretable to a significant degree. Experiments on a clinical fetal echocardiography dataset demonstrate a state-of-the-art performance for our protocols. The techniques proposed in this chapter are fundamentally extendable to other medical imaging modalities as well.

Given the foundations laid here, multiple future directions are possible starting with an exploration of classwise characteristics in terms of forgetting and retention performances and the trends of decline in map quality over intra-class variations. This is likely to be a natural follow-up of the ideas proposed here. Another immediate direction would be to study other strategies for saliency-curriculum driven exemplar retention. While we used a fixed number of exemplars for retention and rehearsal over future tasks, other approaches like a variable retention based on class difficulty are possible. Future directions can also include expanding to different tasks and datasets, using the saliency based exemplar scheme with other lifelong learning methods, using generative replay for estimating past saliency maps and images without need to retain exemplars, and so on. While we have demonstrated on approaches on distillation-based preservation, and using class activation derived saliency maps, the concept is generally applicable with any other continual learning pipeline, and can use other methods of estimating saliency maps, with different base architectures or objectives.

5. Improving the efficiency of ultrasound image analysis models with sonographer gaze assisted distillation

Chapter 5 looks into the possibility of combining deep learning based methods for understanding fetal ultrasound imaging with the expertise of sonographers who have spent a significant part of their professional careers in performing clinically robust interpretations of fetal ultrasound. This symbiosis of human and machine

learnt knowledge is deployed towards the application of improving performance metrics of reduced models with a low memory budget, which is a relevant problem given that ultrasound hardware has seen significant innovations towards portability and on-field usage in recent years. Recent automated medical image analysis methods have attained state-of-the-art performance but have relied on memory and compute intensive deep learning models. Reducing model size without significant loss in performance metrics is crucial for time and memory-efficient automated image-based decision-making. Traditional deep learning based image analysis only uses expert knowledge in the form of manual annotations. Recently, there has been interest in introducing other forms of expert knowledge into deep learning architecture design. This is the approach considered in the paper where we propose to combine ultrasound video with eye fixations tracked for expert sonographers as they scan to train memory-efficient ultrasound image analysis models. Specifically we develop teacher-student knowledge transfer models for the exemplar task of frame classification for the fetal abdomen, head, and femur. The best performing memory-efficient models attain performance within 5% of conventional models that are 1000 times larger in size.

7.2 Limitations

1. Video understanding and analysis under incorrectly assigned labels or with weak supervision, and handling variable input lengths for videos

The video understanding pipelines proposed in Chapter 3 assume that the annotations for all the videos and their frames are correct and complete for both the standard viewing planes, and the cardiac region radii. This is not necessarily the case in multiple ultrasound image analysis datasets, due to the difficulty in parsing these videos even for the human observer and the high level of intra-class variability. It is common that a standard viewing plane gets mistaken as a background frame on account of a high degree of acoustic shadowing, or enhancements or even due to physiological variations such as a crowded rib area or maternal body fat. There is

also the issue of how to label those frames where there is only a part of the relevant anatomical structures visible, i.e., the anatomical structures that are characteristic of a given standard plane or other frame level aspects do exist but are obfuscated by a high level of shadowing, or are partly outside of the acoustic fan region.

In these cases, the question is whether to treat such frames as background frames as they rarely serve clinically important diagnostic frames, or to label them as the standard plane to make our models more robust. The models explored thus far are trained on frames and video clips where the relevant anatomical features are visible to a relatively high degree of clarity. Future studies would need to look at expanding the intra-class diversity in order to account for the different cases of partially visible anatomies, high shadow regions and those frames with an unclear feature definition. This also boils down to the human annotators being able to accurately decide which frames are to be labelled in what classes in such cases. Methods that can offer an uncertainty estimate, such as Bayesian networks, are also options that may be explored in these cases.

Another shortcoming with current implementations of our video analysis is that the input layers of the convolutional architectures take a fixed number of frames as input. This is a limiting aspect of our models as often the long-range dynamics of fetal cardiac motions encode important markers of physiological abnormality. This can be resolved in future work by using a dynamic computational graph (supported by packages like PyTorch and recent implementations of Tensorflow) instead of a static computational graph, as the latter would require initializing separate models for different input sequences. It may also be noted that evaluating the fairness [157] [158] [159] of algorithms used in healthcare is gaining significant interest, and these directions may be incorporated into video analysis algorithm development as well in the years to come.

2. Accounting for intra-class variations in class incremental learning

While class incremental learning scenarios proposed in Chapter 4 look into mechanisms for incorporating previously unseen classes into the model-training

pipeline over discrete arrivals of data streams, it may be noted that in many medical imaging situations the intra-class diversity is significant on account of possible physiological variations and intra-operator biases. This is particularly pronounced for ultrasound as the variations can also result from different degrees of speckle, enhancements and acoustic shadowing.

As such, the catastrophic forgetting that we talk of as a phenomenon resulting from the shift in the input data distributions, may not always be a function of the different classes being added over extended time. That is, if individual datasets from the same class were to be shown to the model over different sessions without all of them being simultaneously present, a forgetting effect can still be observed and therefore class-incremental learning alone may not be sufficient for building truly lifelong learners for ultrasound image analysis. Rather, there exists a need for considering the forgetting and intransigence effects of deep neural networks at the level of individual batches of instances even within a given class when such batches are introduced over different training sessions. This also implies that improving the model performance for a given class by introducing a bigger subset of exemplars for that class may not necessarily be free of a decline in performance on another subset of the same class.

These effects are not considered in the proof-of-concept class incremental learning situations proposed in Chapter 4, and may form the basis for future studies. The saliency based continual learning pipeline does have a method for considering exemplar level effects of forgetting and to track the incremental learning efficiency at the granular level of individual instances , through the ability to assess saliency map quality over successive learning sessions and a corresponding correlation with model forgetting and forward transfer. This may serve as a good starting point for further studies into individual effects on overall model resilience to catastrophic forgetting.

3. Dependence of the sonographer gaze prior on actual sonographer experience

The gaze fixation maps are computed using the eye-tracking data of sonographers who were viewing the ultrasound video during its clinical evaluation. This implies that sonographers were actively looking for diagnostic markers of interest in the process of evaluating an ultrasound scan over multiple frames. The knowledge of which markers are relevant and at what points in the scan they might appear is acquired through multiple practice sessions and on the job experience. It takes years of experience to build the intuition of when to free a given frame while performing a free-hand sweep and tentatively in which areas of the screen a specific diagnostic marker may appear given the state of the scan a few frames prior. As such, tracked gaze points for different sonographers may have a significant difference over individuals with different expertise levels, as the strategies used to locate relevant structures on the ultrasound video would vary, and only get refined over several years. This can cause a non-trivial diversity in the quality of gaze fixation maps generated for different individual sonographers for the same frames. While using such gaze maps for conditioning our ultrasound image analysis tasks, this aspect needs to be considered along with intra-class variations at the frame level. Inclusion of expert knowledge comes with the intra-expert variability and this can cause a difference in the performance of the algorithms used to learn from both streams of input information.

7.3 Future work

Over the course of the themes explored in this thesis, there exist several avenues for future extensions based on the ideas developed here. There is always a scope to improve upon the algorithms developed in the chapters, and addressing some of their shortcomings as elaborated above. There is also a strong element of research into translation of the techniques proposed into clinical workflows. In particular, the model compression enhancements proposed in chapter 5 need to be evaluated in a real-world setting, as it is only through such translational studies that we can gauge whether or not the trade-offs explored are worthwhile in actual clinical

settings. Overall, a few avenues of future work that seem to be particularly in line with the motivation of this thesis are:

1. Enabling continual learning over online streaming of videos

The ideas around continual learning proposed in chapter 4 mainly deal with class incremental situations over extended time. Often, real-world clinicians have to make decisions on the fly as they see a new structural defect on the screen while performing a free-hand sweep. Such a difference in the image-level feature space, as it arrives in a sequence of frames of a video, is difficult to adapt to in an online fashion. Algorithms in ultrasound imaging that are expected to be truly continual learners needs to be able to account for these unforeseen variations over the course of a single video sequence. This can potentially be implemented as an online learning problem, although to our knowledge, there has been little effort in this direction. In this situation, the idea of having different modules for different tasks or classes is a challenging prospect since there is a limited memory available and frame-to-frame diversity can be significant. The question is how one builds a suitable incremental learning pipeline that absorbs features and their differences across frames, without having to dynamically expand or store a large number of prototype exemplars. One possible solution is to explore the online variants of the Elastic Weight Consolidation (EWC) approach [160], and to combine them with regularization strategies such as pseudo-rehearsal based knowledge distillation. There is also a possibility of using a dynamic computational graph to enable taking in an arbitrary numbers of frames as input.

2. Inclusion of other forms of sonographer-video interaction such as probe motions, audio and textual information

We have demonstrated that the inclusion of expert knowledge into deep learning paradigms for fetal sonography can make models more efficient with a lower memory footprint. It is an interesting aspect to evaluate how the incorporation of other cues evident in the overall sonographer experience while performing diagnostic ultrasound. Specifically, the inclusion of probe motion information can be of

significant utility in terms of providing real-time guidance towards standard planes, evaluating experience levels, providing corrective assistance to clinicians and so on. Tracking of probe motions can be done using an Inertial Measurement Unit (IMU). An information about probe motion in the six degrees of freedom, i.e., in terms of the motion and velocity over each time step for both translation and rotation, can provide information about how individual sonographers respond to the video dynamics visible to them during scanning time. This, along with gaze patterns, can serve to define the diagnostic patterns of individual sonographers and in due course can be used for training robotic sonography units via imitation learning [161]. Even in the algorithms presently explored, particularly the ones that involve modelling of spatiotemporal cardiac motions, incorporation of probe motion is useful because the fetus under scanning moves around non-deterministically and the motion apparent in the ultrasound video is a result of the relative motion between the fetus and the probe (assuming that the influence of subject motion is negligible). Another point to note is that the ultrasound diagnosis process happens in a fairly dynamic environment with the range of observational cues being verbally communicated by the sonographer, and a subset of such information may also be mentioned in case reports. Such an audio based information stream can be coupled with the video sequence inputs, and potentially with corresponding gaze maps as well. Such a fusion of audio information with corresponding frames can help establish correspondences between sonographer observations with possible landmarks of interest at the frame level. Such audio level cues and transcribed text reports may also indicate sonographer sentiment as apparent during scan acquisition and a sentiment analysis based approach on such data may yield useful cues about the relative interest in a particular section of a video sequence which may be particularly useful in the event of anomalous or pathologically significant markers appearing in a subset of frames being acquired.

3. Extension of the motion understanding studies using a reinforcement learning approach

The studies presented on understanding the motion of cardiac structures and those used for spatiotemporal modelling for applications like plane finding can be understood in terms of a scenario where an agent or a swarm of agents are tasked with accomplishing a target in space (say, accurately finding a suitable viewing plane) and depending on the proximity of their chosen action to what a trained expert would do, are assigned rewards or penalties. This is a suitable application of Reinforcement Learning (RL) where an agent is tasked with a certain objective and every stage where it gets closer to the objective, a ‘reward’ is assigned to it. Such a reward is often modelled through the Bellman equations [162] and can be used as an optimization objective to be maximized over the course of the training on available data. Typically, a RL agent does not have a prior knowledge of the action space and tries to seek the reward by a strategy of exploration of possible actions and exploitation of a subset of those actions. As such, the agent needs to maximize overall rewards and arrive at the objective. This maximization of overall reward can come at the cost of sacrificing short-term gains. Additionally, the agent may be assigned a reward based on the time taken to arrive at an optimal solution. Therefore, the agent needs to learn over available training examples how to interact with the environment towards attaining a suitable combination of exploration and exploitation steps to obtain the final objective. With a rich variety of multimodal information available in the sonography room, it is an exciting domain to explore the utility of RL agents similar to the directions explored in other domains pertaining to computer aided interventions such as robotic surgical planning [163]. One can define objectives in terms of finding standard planes, localizing anatomies of interest, figuring out optimal sweep patterns, finding correct probe motions for different situations and so on. Additionally information such as gaze patterns, sonographer speech and text reports, motion patterns and expert feedback can all be used as reward components for the RL agent. Possible applications of successful RL based agents in fetal ultrasound are significant, ranging from robotic sonography setups to automated plane finding under challenging situations of high shadows

and artefacts. Therefore, this can be a promising avenue of future work building upon the fundamentals presented in this thesis.

7.4 Concluding remarks

Advances in fetal sonography has revolutionized the course of prenatal healthcare over the last few decades and may be credited with ensuring a better quality of life for millions of people worldwide. It is high time that advances in machine learning based research into medical image computing and computer aided interventions give ultrasound diagnostics the propulsion it needs to accelerate into the areas still not adequately served and technological problems not fully resolved. This thesis attempts to make a contribution in some of these areas, and address a few key challenges with respect to automated understanding of fetal ultrasound videos. All validations of the solutions proposed are performed on real-world clinical fetal ultrasound datasets and it is hoped that these algorithms are able to inspire further research into relevant areas in the years to come.

*The first kind of intellectual and artistic personality
belongs to the hedgehogs, the second to the foxes ...*

— Sir Isaiah Berlin

References

- [1] World Health Organization, World Health Organization, et al. “Report on the regional consultation towards the development of a strategy for optimizing fetal growth and development”. In: *WHO Regional Office for the Eastern Mediterranean: Cairo* (2005).
- [2] Margaret B McNay and John E Fleming. “Forty years of obstetric ultrasound 1957-1997: from A-scope to three dimensions.” In: *Ultrasound in medicine & biology* 25.1 (1999), pp. 3–56.
- [3] JP De Wilde, AW Rivers, and DL Price. “A review of the current use of magnetic resonance imaging in pregnancy and safety implications for the fetus”. In: *Progress in biophysics and molecular biology* 87.2-3 (2005), pp. 335–353.
- [4] Francis A Duck. “Is it safe to use diagnostic ultrasound during the first trimester?” In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 13.6 (1999), pp. 385–388.
- [5] Alfred Abuhamad et al. “Standardized six-step approach to the performance of the focused basic obstetric ultrasound examination”. In: *American journal of perinatology* 2.01 (2016), pp. 090–098.
- [6] JS Carvalho et al. “ISUOG Practice Guidelines (updated): sonographic screening examination of the fetal heart”. In: *Ultrasound in Obstetrics & Gynecology* 41.3 (2013), pp. 348–359.
- [7] Sangeeta Pathak and C Lees. “Ultrasound structural fetal anomaly screening: an update”. In: *Archives of Disease in Childhood-Fetal and Neonatal Edition* 94.5 (2009), F384–F390.
- [8] Nick Archer and Nicky Manning. *Fetal cardiology*. Oxford University Press, 2009.
- [9] Mark K Friedberg et al. “Prenatal detection of congenital heart disease”. In: *The Journal of pediatrics* 155.1 (2009), pp. 26–31.
- [10] R Andrews et al. “Outcome of staged reconstructive surgery for hypoplastic left heart syndrome following antenatal diagnosis”. In: *Archives of disease in childhood* 85.6 (2001), pp. 474–477.
- [11] GD Hill et al. “Disparities in the prenatal detection of critical congenital heart disease”. In: *Prenatal diagnosis* 35.9 (2015), pp. 859–863.
- [12] Ph Pézard et al. “Influence of ultrasonographers training on prenatal diagnosis of congenital heart diseases: a 12-year population-based study”. In: *Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis* 28.11 (2008), pp. 1016–1022.

- [13] Lindsey Allan. “Antenatal diagnosis of heart disease”. In: *Heart* 83.3 (2000), pp. 367–367.
- [14] R Chaoui. “The four-chamber view: four reasons why it seems to fail in screening for cardiac abnormalities and suggestions to improve detection rate”. In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 22.1 (2003), pp. 3–10.
- [15] Damien Bonnet et al. “Detection of transposition of the great arteries in fetuses reduces neonatal morbidity and mortality”. In: *Circulation* 99.7 (1999), pp. 916–918.
- [16] O Franklin et al. “Prenatal diagnosis of coarctation of the aorta improves survival and reduces morbidity”. In: *Heart* 87.1 (2002), pp. 67–69.
- [17] Yuichi Notomi et al. “Measurement of ventricular torsion by two-dimensional ultrasound speckle tracking imaging”. In: *Journal of the American College of Cardiology* 45.12 (2005), pp. 2034–2041.
- [18] American College of Radiology et al. “ACR-SPR-SRU practice guideline for performing and interpreting diagnostic ultrasound examinations”. In: *American College of Radiology website* (2011).
- [19] I Sarris et al. “Intra-and interobserver variability in fetal ultrasound measurements”. In: *Ultrasound in Obstetrics & Gynecology* 39.3 (2012), pp. 266–273.
- [20] Riccardo Miotto et al. “Deep learning for healthcare: review, opportunities and challenges”. In: *Briefings in bioinformatics* 19.6 (2017), pp. 1236–1246.
- [21] Y Cai et al. “SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking”. In: *15th IEEE ISBI*. IEEE. 2018, pp. 1475–1478.
- [22] Kuntharee Traisrisilp et al. “Reference Ranges for the Fetal Cardiac Circumference Derived by Cardio-Spatiotemporal Image Correlation From 14 to 40 Weeks’ Gestation”. In: *Journal of Ultrasound in Medicine* 30.9 (2011), pp. 1191–1196.
- [23] Pierre Chatelain et al. “Evaluation of gaze tracking calibration for longitudinal biomedical imaging studies”. In: *IEEE transactions on cybernetics* 50.1 (2018), pp. 153–163.
- [24] PULSE. “Perception Ultrasound by Learning Sonographic Experience, www.eng.ox.ac.uk/pulse”. In: (2018).
- [25] Li Liu et al. “Levels and causes of mortality under age five years”. In: *International Bank for Reconstruction and Development, World Bank. Reproductive, maternal, newborn, and child health: disease control priorities 2* (2015), pp. 71–83.
- [26] Marveen Craig. “Sonographer shortages: a day late and a dollar short?” In: *Journal of Diagnostic Medical Sonography* 19.4 (2003), pp. 261–271.
- [27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), p. 436.
- [28] Seppo Linnainmaa. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors”. In: *Master’s Thesis (in Finnish), Univ. Helsinki* (1970), pp. 6–7.

- [29] Paul Werbos. “Beyond Regression:” New Tools for Prediction and Analysis in the Behavioral Sciences”. In: *Ph. D. dissertation, Harvard University* (1974).
- [30] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [31] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets”. In: *Neural computation* 18.7 (2006), pp. 1527–1554.
- [32] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.
- [33] Asja Fischer and Christian Igel. “An introduction to restricted Boltzmann machines”. In: *Iberoamerican Congress on Pattern Recognition*. Springer. 2012, pp. 14–36.
- [34] Yoshua Bengio et al. “Learning deep architectures for AI”. In: *Foundations and trends® in Machine Learning* 2.1 (2009), pp. 1–127.
- [35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.
- [36] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [37] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.
- [38] Léon Bottou. “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of COMPSTAT’2010*. Springer, 2010, pp. 177–186.
- [39] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.
- [40] Matthew D Zeiler and Rob Fergus. “Visualizing and understanding convolutional networks”. In: *European conference on computer vision*. Springer. 2014, pp. 818–833.
- [41] Karen Simonyan and Andrew Zisserman. “Two-stream convolutional networks for action recognition in videos”. In: *Advances in neural information processing systems*. 2014, pp. 568–576.
- [42] Ken Chatfield et al. “Return of the devil in the details: Delving deep into convolutional nets”. In: *arXiv preprint arXiv:1405.3531* (2014).
- [43] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [44] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [45] Sasha Targ, Diogo Almeida, and Kevin Lyman. “Resnet in resnet: Generalizing residual architectures”. In: *arXiv preprint arXiv:1603.08029* (2016).

- [46] Gordon Cooper. “Software Framework Requirements For Embedded Vision”. In: (2017). URL: <https://semiengineering.com/software-frameworkrequirements-for-embedded-vision/>.
- [47] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.
- [48] Hariharan Ravishankar et al. “Understanding the mechanisms of deep transfer learning for medical images”. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 188–196.
- [49] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.
- [50] Ariadna Quattoni, Michael Collins, and Trevor Darrell. “Transfer learning for image classification with sparse prototype representations”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.
- [51] Ali Sharif Razavian et al. “CNN features off-the-shelf: an astounding baseline for recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 806–813.
- [52] Di Wu, Nabin Sharma, and Michael Blumenstein. “Recent advances in video-based human action recognition using deep learning: A review”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 2865–2872.
- [53] Andrej Karpathy et al. “Large-scale video classification with convolutional neural networks”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732.
- [54] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. “Convolutional two-stream network fusion for video action recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1933–1941.
- [55] Du Tran et al. “Learning Spatiotemporal Features with 3D Convolutional Networks”. In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015).
- [56] Shuiwang Ji et al. “3D Convolutional Neural Networks for Human Action Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231.
- [57] Daniel Maturana and Sebastian Scherer. “VoxNet: A 3D Convolutional Neural Network for real-time object recognition”. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2015).
- [58] Konstantinos Kamnitsas et al. “Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation”. In: *CoRR* abs/1603.05959 (2016). arXiv: 1603.05959. URL: <http://arxiv.org/abs/1603.05959>.
- [59] Qi Dou et al. “3D Deeply Supervised Network for Automatic Liver Segmentation from CT Volumes”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 Lecture Notes in Computer Science* (2016), pp. 149–157.
- [60] Roland Kwitt et al. “Localizing target structures in ultrasound video—a phantom study”. In: *Medical image analysis* 17.7 (2013), pp. 712–722.

- [61] Mohammad Ali Maraci et al. “Searching for structures of interest in an ultrasound video sequence”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pp. 133–140.
- [62] Christopher P Bridge, Christos Ioannou, and J Alison Noble. “Automated annotation and quantitative description of ultrasound videos of the fetal heart”. In: *Medical image analysis* 36 (2017), pp. 147–161.
- [63] Adel Hafiane, Pierre Vieyres, and Alain Delbos. “Deep learning with spatiotemporal consistency for nerve segmentation in ultrasound images”. In: *arXiv preprint arXiv:1706.05870* (2017).
- [64] Rosa-Maria Menchón-Lara, José-Luis Sancho-Gómez, and Andrés Bueno-Crespo. “Early-stage atherosclerosis detection using deep learning over carotid ultrasound images”. In: *Applied Soft Computing* 49 (2016), pp. 616–628.
- [65] Fausto Milletari et al. “Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound”. In: *Computer Vision and Image Understanding* 164 (2017), pp. 92–102.
- [66] Christine Kroll et al. “Coupling convolutional neural networks and hough voting for robust segmentation of ultrasound volumes”. In: *German Conference on Pattern Recognition*. Springer. 2016, pp. 439–450.
- [67] Jun Shi et al. “Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset”. In: *Neurocomputing* 194 (2016), pp. 87–94.
- [68] Hao Chen et al. “Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 487–495.
- [69] Gustavo Carneiro, Jacinto C Nascimento, and António Freitas. “The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods”. In: *IEEE Transactions on Image Processing* 21.3 (2011), pp. 968–982.
- [70] Rosa-Maria Menchón-Lara, Andrés Bueno-Crespo, and José Luis Sancho-Gómez. “Estimation of the Arterial Diameter in Ultrasound Images of the Common Carotid Artery”. In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2015, pp. 358–367.
- [71] Hasmila A Omar et al. “Quantification of cardiac bull’s-eye map based on principal strain analysis for myocardial wall motion assessment in stress echocardiography”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1195–1198.
- [72] H Omar et al. “Myocardial wall motion assessment in stress echocardiography by quantification of principal strain bulls eye maps: P299”. In: *European Heart Journal Cardiovascular Imaging* 18 (2017).
- [73] Hasmila A Omar et al. “Automated myocardial wall motion classification using handcrafted features vs a deep cnn-based mapping”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 3140–3143.

- [74] Hasmila A Omar et al. “Improving visual detection of wall motion abnormality with echocardiographic image enhancing methods”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 1128–1131.
- [75] Dhruv Agarwal, KS Shriram, and Navneeth Subramanian. “Automatic view classification of echocardiograms using histogram of oriented gradients”. In: *2013 IEEE 10th International Symposium on Biomedical Imaging*. IEEE. 2013, pp. 1368–1371.
- [76] Ritwik Kumar et al. “Echocardiogram view classification using edge filtered scale-invariant motion features”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 723–730.
- [77] Shahram Ebadollahi, Shih-Fu Chang, and Henry Wu. “Automatic view recognition in echocardiogram videos using parts-based representation”. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. Vol. 2. IEEE. 2004, pp. II–II.
- [78] Ana IL Namburete, Bahbib Rahmatullah, and J Alison Noble. “Nakagami-based Adaboost learning framework for detection of anatomical landmarks in 2D fetal neurosonograms”. In: *Annals of the BMVA 2* (2013), pp. 1–16.
- [79] Bahbib Rahmatullah, Aris T Papageorgiou, and J Alison Noble. “Integration of local and global features for anatomical object detection in ultrasound”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 402–409.
- [80] Mohammad Yaqub et al. “Automatic detection of local fetal brain structures in ultrasound images”. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, pp. 1555–1558.
- [81] Christian F Baumgartner et al. “Real-time detection and localisation of fetal standard scan planes in 2d freehand ultrasound”. In: *CoRR abs/1612.05601* (2016).
- [82] Zhen Yu et al. “A deep convolutional neural network-based framework for automatic fetal facial standard plane recognition”. In: *IEEE journal of biomedical and health informatics 22.3* (2017), pp. 874–885.
- [83] Baiying Lei et al. “Automatic recognition of fetal facial standard plane in ultrasound image via fisher vector”. In: *PloS one 10.5* (2015), e0121838.
- [84] Y Gao, Mohammad Ali Maraci, and J Alison Noble. “Describing ultrasound video content using deep convolutional neural networks”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 787–790.
- [85] Hao Chen et al. “Standard plane localization in fetal ultrasound via domain transferred deep neural networks”. In: *IEEE journal of biomedical and health informatics 19.5* (2015), pp. 1627–1636.
- [86] Vaanathi Sundaresan et al. “Automated characterization of the fetal heart in ultrasound images using fully convolutional neural networks”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 671–674.

- [87] Weilin Huang et al. “Temporal HeartNet: towards human-level automatic analysis of fetal cardiac screening video”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 341–349.
- [88] Rui Hou, Chen Chen, and Mubarak Shah. “Tube convolutional neural network (T-CNN) for action detection in videos”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5822–5831.
- [89] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.
- [90] Xiaojiang Peng and Cordelia Schmid. “Multi-region two-stream R-CNN for action detection”. In: *European conference on computer vision*. Springer. 2016, pp. 744–759.
- [91] Federico Becattini et al. “Am I done? Predicting action progress in videos”. In: *arXiv preprint arXiv:1705.01781* (2017).
- [92] Arijit Patra, Weilin Huang, and J Alison Noble. “Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 276–284.
- [93] Arijit Patra and JA Noble. “Sequential anatomy localization in fetal echocardiography videos”. In: *arXiv preprint arXiv:1810.11868* (2018).
- [94] Arijit Patra and J Alison Noble. “Multi-anatomy localization in fetal echocardiography videos”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 1761–1764.
- [95] Juan P Viguera-Guillen et al. “Parallel Capsule Networks for Classification of White Blood Cells”. In: *arXiv:2108.02644* (2021).
- [96] Sebastian Thrun and Tom M Mitchell. “Lifelong robot learning”. In: *Robotics and autonomous systems* 15.1-2 (1995), pp. 25–46.
- [97] German I Parisi et al. “Continual lifelong learning with neural networks: A review”. In: *Neural Networks* (2019).
- [98] Ian J Goodfellow et al. “An empirical investigation of catastrophic forgetting in gradient-based neural networks”. In: *arXiv preprint arXiv:1312.6211* (2013).
- [99] Nima Tajbakhsh et al. “Convolutional neural networks for medical image analysis: Full training or fine tuning?” In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1299–1312.
- [100] Jeff Donahue et al. “DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition”. In: *CoRR* abs/1310.1531 (2013). arXiv: 1310.1531. URL: <http://arxiv.org/abs/1310.1531>.
- [101] Jeffrey Donahue et al. “Long-term recurrent convolutional networks for visual recognition and description”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634.
- [102] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. “Siamese neural networks for one-shot image recognition”. In: *ICML deep learning workshop*. Vol. 2. 2015.

- [103] Oriol Vinyals et al. “Matching networks for one shot learning”. In: *Advances in neural information processing systems*. 2016, pp. 3630–3638.
- [104] Andrei A Rusu et al. “Progressive neural networks”. In: *arXiv preprint arXiv:1606.04671* (2016).
- [105] Jaehong Yoon et al. “Lifelong learning with dynamically expandable networks”. In: *arXiv preprint arXiv:1708.01547* (2017).
- [106] Zhizhong Li and Derek Hoiem. “Learning without forgetting”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017), pp. 2935–2947.
- [107] Arijit Patra et al. “Multimodal continual learning with sonographer eye-tracking in fetal ultrasound”. In: *ASMUS 2021: 2nd International Workshop of Advances in Simplifying Medical UltraSound*. Springer, 2021.
- [108] James Kirkpatrick et al. “Overcoming catastrophic forgetting in neural networks”. In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526.
- [109] Friedemann Zenke, Ben Poole, and Surya Ganguli. “Continual learning through synaptic intelligence”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 3987–3995.
- [110] Tianjun Xiao et al. “Error-driven incremental learning in deep convolutional neural network for large-scale image classification”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 177–186.
- [111] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [112] Sylvestre-Alvise Rebuffi et al. “icarl: Incremental classifier and representation learning”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, pp. 2001–2010.
- [113] Vinayak Abrol, Pulkit Sharma, and Arijit Patra. “Improving generative modelling in VAEs using multimodal prior”. In: *IEEE Transactions on Multimedia* (2020).
- [114] Timothée Lesort et al. “Generative models from the perspective of continual learning”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.
- [115] Ronald Kemker and Christopher Kanan. “Fearnnet: Brain-inspired model for incremental learning”. In: *arXiv preprint arXiv:1711.10563* (2017).
- [116] Arijit Patra and Tapabrata Chakraborti. “Learn more, forget less: Cues from human brain”. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Nov. 2020.
- [117] Balathasan Giritharan, Xiaohui Yuan, and Jianguo Liu. “Incremental classification learning for anomaly detection in medical images”. In: *Medical Imaging 2009: Computer-Aided Diagnosis*. Vol. 7260. International Society for Optics and Photonics. 2009, 72603W.
- [118] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. “Expert gate: Lifelong learning with a network of experts”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3366–3375.

- [119] Bolei Zhou et al. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.
- [120] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.
- [121] Jo Schlemper et al. “Attention gated networks: Learning to leverage salient regions in medical images”. In: *Medical image analysis* 53 (2019), pp. 197–207.
- [122] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).
- [123] Dawn M Becker et al. “The use of portable ultrasound devices in low-and middle-income countries: a systematic review of the literature”. In: *Tropical Medicine & International Health* 21.3 (2016), pp. 294–311.
- [124] Maryam Ahmed and J Alison Noble. “An eye-tracking inspired method for standardised plane extraction from fetal abdominal ultrasound volumes”. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2016, pp. 1084–1087.
- [125] Yifan Cai et al. “Multi-task SonoEyeNet: detection of fetal standardized planes assisted by generated sonographer attention maps”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 871–879.
- [126] Richard Droste et al. “Ultrasound Image Representation Learning by Modeling Sonographer Visual Attention”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 592–604.
- [127] Baoyuan Liu et al. “Sparse convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 806–814.
- [128] Hao Li et al. “Pruning filters for efficient convnets”. In: *arXiv preprint arXiv:1608.08710* (2016).
- [129] Yihui He et al. “Channel pruning for accelerating very deep neural networks”. In: *Proc. IEEE International Conference on Computer Vision*. 2017, pp. 1389–1397.
- [130] Sajid Anwar and Wonyong Sung. “Compact deep convolutional neural networks with coarse pruning”. In: *arXiv preprint arXiv:1610.09639* (2016).
- [131] Ruichi Yu et al. “Nisp: Pruning networks using neuron importance score propagation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9194–9203.
- [132] Ming Tu et al. “Reducing the model order of deep neural networks using information theory”. In: *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE. 2016, pp. 93–98.
- [133] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *NIPS 2014 Deep Learning Workshop*. 2014.
- [134] Adriana Romero et al. “FitNets: Hints for thin deep nets”. In: *arXiv:1412.6550* (2014).

- [135] Hooman Vaseli et al. “Designing lightweight deep learning models for echocardiography view classification”. In: *SPIE Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 10951.
- [136] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [137] JA Noble and A Patra. “Multi-anatomy localization in fetal echocardiography videos”. In: (2019).
- [138] Cecilia S Lee and Aaron Y Lee. “Clinical applications of continual learning machine learning”. In: *The Lancet Digital Health* 2.6 (2020), e279–e281.
- [139] Food, Drug Administration, et al. *Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)-discussion paper*. 2019.
- [140] Prithviraj Dhar et al. “Learning without memorizing”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5138–5146.
- [141] Arijit Patra and Julia Alison Noble. “Hierarchical Class Incremental Learning of Anatomical Structures in Fetal Echocardiography Videos”. In: *IEEE Journal of Biomedical and Health Informatics* 24.4 (2020), pp. 1046–1058.
- [142] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, pp. 249–256.
- [143] Piotr Dabkowski and Yarın Gal. “Real time image saliency for black box classifiers”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 6967–6976.
- [144] JA Noble and A Patra. “Incremental learning of fetal heart anatomies using interpretable saliency maps”. In: *Communications in Computer and Information Science* 1065 (2020).
- [145] Max Welling. “Herding dynamical weights to learn”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 2009, pp. 1121–1128.
- [146] Arijit Patra and J Alison Noble. “Incremental learning of fetal heart anatomies using interpretable saliency maps”. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2019, pp. 129–141.
- [147] Cristian Buciluă et al. “Model compression”. In: *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006, pp. 535–541.
- [148] Francisco M Castro et al. “End-to-end incremental learning”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 233–248.
- [149] Saihui Hou et al. “Lifelong learning via progressive distillation and retrospection”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 437–452.
- [150] Shun-Ichi Amari. “Natural gradient works efficiently in learning”. In: *Neural computation* 10.2 (1998), pp. 251–276.

- [151] Arijit Patra et al. “Efficient ultrasound image analysis models with sonographer gaze assisted distillation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 394–402.
- [152] Forrest N Iandola et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size”. In: *arXiv preprint arXiv:1602.07360* (2016).
- [153] Andrew G Howard et al. “MobileNets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).
- [154] Mark Sandler et al. “MobileNetV2: Inverted residuals and linear bottlenecks”. In: *CVPR*. 2018.
- [155] AT Papageorgiou et al. “Efficient ultrasound image analysis models with sonographer gaze assisted distillation”. In: *Lecture Notes in Computer Science* 11767. Part IV (2018).
- [156] Alexander Wong. “NetScore: Towards universal metrics for large-scale performance analysis of deep neural networks for practical usage”. In: *arXiv:1806.05512* (2018).
- [157] Mengnan Du et al. “Fairness in deep learning: A computational perspective”. In: *IEEE Intelligent Systems* (2020).
- [158] Tapabrata Chakraborti, Arijit Patra, and J Alison Noble. “Contrastive fairness in machine learning”. In: *IEEE Letters of the Computer Society* 3.2 (2020), pp. 38–41.
- [159] Alvin Rajkomar et al. “Ensuring fairness in machine learning to advance health equity”. In: *Annals of internal medicine* 169.12 (2018), pp. 866–872.
- [160] Hippolyt Ritter, Aleksandar Botev, and David Barber. “Online structured laplace approximations for overcoming catastrophic forgetting”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 3738–3748.
- [161] Yan Duan et al. “One-shot imitation learning”. In: *Advances in neural information processing systems*. 2017, pp. 1087–1098.
- [162] Leemon Baird. “Residual algorithms: Reinforcement learning with function approximation”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37.
- [163] Brijen Thananjeyan et al. “Multilateral surgical pattern cutting in 2d orthotropic gauze with deep reinforcement learning policies for tensioning”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 2371–2378.