Generating Textual Captions for Ultrasound Visuals in an Automated Fashion

Mohammad Alsharid

Jesus College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2021

Abstract

Generating captions for ultrasound images and videos is an area that is yet to be fully studied and explored. The aim of the work in this thesis is to learn joint image-text representations to describe ultrasound images with rich vocabulary consisting of nouns, verbs, and adjectives. Preparing medical image captioning benchmarks is challenging for two reasons: (a) describing medical images with specific terminology requires expert knowledge of medical professionals; and (b) the sensitive nature of medical images prevents wide-scale annotation, for instance, using crowd-sourcing services (e.g. Amazon Mechanical Turk) and similar methods. Therefore, automatic image captioning has not been widely studied on ultrasound images before, the challenge being enhanced by the lack of readily available large datasets of ultrasound images with captions.

First, the thesis explores different combinations of recurrent neural networks, concatenation techniques, word embedding vectors in different model architecture configurations. We identify in this process the configuration most suitable for the fetal ultrasound image captioning task and the dataset at hand. We show that a configuration incorporating an LSTM-RNN and word2vec embeddings and using a merge-by-concatenation operation performed best. The thesis then explores three solutions to the challenge of working with real world datasets. We introduce a curriculum learning based strategy that incorporates the complexities of the image and text information to prepare the data for training. We show that by training captioning models with the order of data samples determined by the curriculum, we can achieve higher scores on the evaluation metrics with the same amount of data. We also look into augmenting the data through the creation of pseudocaptions to pair up with caption-less images. Finally, we explore leveraging other available data from a different modality, specifically eye gaze points, to supplement available image-text data. We find that using eye gaze data can help in training models that score relatively higher on the evaluation metrics; however since the improvements are small and the pre-training steps involved are considerable, this leads us to the recommendation that improving base models should take precedence over relying on data from other modalities to improve the performance of captioning models.

To the best of our knowledge, the work in this thesis is the first attempt to perform automatic image captioning on fetal ultrasound images (video frames), using sonographer spoken words to describe their scanning experience. The thesis can help serve as a blue print for future endeavours in fetal ultrasound captioning by providing guidelines to follow and pitfalls to avoid and as an aid for those attempting medical image captioning, more generally.

Generating Textual Captions for Ultrasound Visuals in an Automated Fashion

Mohammad Alsharid

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2021

Acknowledgements

I would like to thank my supervisor, Prof. Alison Noble, colleagues, family, and friends for their support. I am grateful for the helpful discussions I have had with Harshita Sharma, Daniela Massiceti, Farah Shamout, Rasheed El-Bouri, Yifan Cai, Richard Droste, Lokhin Lee, Denise Dempsey, Yipei Wang, Peter Liu, Clare Teng, Alexander Gleed, Yangdi Xu, Felipe Moser, Ping Lu, Jielai Zhang, Jianbo Jao, Jaan Toots, Lior Drukker, Pierre Chatelain, Yuan Gao, Ruobing Huang, Spencer Dunleavy, Hosuk Ryou, Hasmila Omar, Arijit Patra, Weidi Xie, and many others for the advice they have given me and for all they have done that helped support me throughout this journey at Oxford. I would also like to thank the Rhodes Trust for their institutional support.

Abstract

Generating captions for ultrasound images and videos is an area that is yet to be fully studied and explored. The aim of the work in this thesis is to learn joint image-text representations to describe ultrasound images with rich vocabulary consisting of nouns, verbs, and adjectives. Preparing medical image captioning benchmarks is challenging for two reasons: (a) describing medical images with specific terminology requires expert knowledge of medical professionals; and (b) the sensitive nature of medical images prevents wide-scale annotation, for instance, using crowd-sourcing services (e.g. Amazon Mechanical Turk) and similar methods. Therefore, automatic image captioning has not been widely studied on ultrasound images before, the challenge being enhanced by the lack of readily available large datasets of ultrasound images with captions.

First, the thesis explores different combinations of recurrent neural networks, concatenation techniques, word embedding vectors in different model architecture configurations. We identify in this process the configuration most suitable for the fetal ultrasound image captioning task and the dataset at hand. We show that a configuration incorporating an LSTM-RNN and word2vec embeddings and using a merge-by-concatenation operation performed best. The thesis then explores three solutions to the challenge of working with real world datasets. We introduce a curriculum learning based strategy that incorporates the complexities of the image and text information to prepare the data for training. We show that by training captioning models with the order of data samples determined by the curriculum, we can achieve higher scores on the evaluation metrics with the same amount of data. We also look into augmenting the data through the creation of pseudocaptions to pair up with caption-less images. Finally, we explore leveraging other available data from a different modality, specifically eye gaze points, to supplement available image-text data. We find that using eye gaze data can help in training models that score relatively higher on the evaluation metrics; however since the improvements are small and the pre-training steps involved are considerable, this leads us to the recommendation that improving base models should take precedence over relying on data from other modalities to improve the performance of captioning models.

To the best of our knowledge, the work in this thesis is the first attempt to perform automatic image captioning on fetal ultrasound images (video frames),

using sonographer spoken words to describe their scanning experience. The thesis can help serve as a blue print for future endeavours in fetal ultrasound captioning by providing guidelines to follow and pitfalls to avoid and as an aid for those attempting medical image captioning, more generally.

Contents

List of Figures **xi**

List of Abbreviations **xix**

1 Introduction **1**
1.1 Clinical Motivation 1
1.2 Contributions and Publications 3
1.3 Thesis Structure 4

2 Literature Review **7**
2.1 Introduction 7
2.2 Speech: Augmenting and Identifying Audible Human Expressions . 8
2.2.1 Reasons for Conducting a Literature Review on Speech . . . 8
2.2.2 Audio Data Augmentation 9
2.2.3 Sound eXchange (SoX) 12
2.3 Transcribed Speech is Text: Transcribing Spoken Words 14
2.3.1 Speech Transcription in a Medical Context 14
2.4 Text and Images: Multimodal Data for the Captioning Task 16
2.4.1 Image Captioning 16
2.4.2 Ultrasound Image Captioning 32
2.4.3 Video Captioning 33
2.4.4 Gaze Tracking Information in Captioning 33
2.4.5 Medical Image Captioning and Curriculum Learning 34
2.5 Conclusion 34

3 Datasets **37**
3.1 Introduction 38
3.1.1 Speech Transcription Tools 39
3.2 Descriptions of Datasets 40
3.2.1 Dataset Class. 1 41
3.2.2 Dataset Class. 2 42
3.2.3 Dataset Cap. 1 42
3.2.4 Dataset Cap. 2a 50

3.2.5 Dataset Cap. 2b . . . 50
3.2.6 Dataset Cap. 3 . . . 51
3.2.7 Analysis of Sonographer Vocabulary (Freeze Frame Vocabulary vs. Probe Motion Vocabulary) . . . 53
3.3 Microphone Matters . . . 56
3.3.1 Microphone Comparisons . . . 56
3.3.2 Chosen Microphone . . . 62
3.3.3 Audio Challenges . . . 62
3.3.4 Justification for Purchasing and Using Two Microphones . . . 64
3.3.5 Source Separation . . . 64
3.3.6 Choice of USB Audio Interface . . . 65
3.3.7 Placement of Microphones and Audio Interface . . . 65
3.4 Evaluation Metrics . . . 69
3.4.1 Evaluation Metrics for Experiments Conducted in Chapter 5 69
3.4.2 Evaluation Metrics for Experiments Conducted in Chapter 6 70
3.5 Conclusion . . . 71

4 Building Captioning Models for the Fetal Ultrasound Context 73
4.1 Introduction . . . 73
4.2 Image Captioning Model . . . 75
4.2.1 Model Architecture . . . 75
4.2.2 Results and Discussion . . . 80
4.3 Summary . . . 89

5 Improving the Performance of Captioning Models through Curriculum Learning and Pseudo-Caption Creation 91
5.1 Overview . . . 92
5.2 Originality and Individual Role . . . 94
5.3 The Dual Curriculum . . . 95
5.3.1 Introduction . . . 95
5.3.2 Model and Training Details . . . 97
5.3.3 Results and Discussion . . . 101
5.4 The Course-Focused Dual Curriculum . . . 104
5.4.1 Introduction . . . 105
5.4.2 Model and Training Details . . . 105
5.4.3 Image Captioning Model Architecture . . . 106
5.4.4 Experiments . . . 107
5.4.5 Results and Discussion . . . 107
5.5 Pseudo-Caption Preparation Pipeline . . . 110
5.5.1 Introduction . . . 110

5.5.2 Model and Training Details . . . 112
5.5.3 Experiments . . . 116
5.5.4 Results and Discussion . . . 117
5.6 Summary . . . 120

6 Leveraging Data from Other Modalities in Fetal Image Captioning 123
6.1 Introduction . . . 123
6.1.1 Chapter Outline . . . 125
6.1.2 Changes from Previous Chapters . . . 125
6.1.3 Purpose of Chapter . . . 126
6.2 Originality and Individual Role . . . 126
6.3 Gaze-Assisted Captioning . . . 127
6.3.1 Introduction . . . 127
6.3.2 Model and Training Details . . . 128
6.3.3 Results and Discussion . . . 139
6.4 Conclusions . . . 145

7 Conclusions and Future Work 147
7.1 Conclusions . . . 147
7.2 Future Work . . . 149
7.2.1 Video Navigation . . . 149
7.2.2 Using Curriculum Learning with Simpler Models and Video Captioning and Paving the Way Towards Clinical Translation 150
7.2.3 Using Probe Motion Data in Captioning . . . 150
7.2.4 Using Audio Directly in the Captioning Process . . . 151
7.2.5 Using a Transformer-based Captioning Model . . . 152

Appendices

A Demo Scripts 157
A.1 "*_fetal" file . . . 157
A.2 "*_terms" file . . . 158

B Extra Figures 161

References 163

List of Figures

2.1 Central Concept of the methodology of Karpathy *et al.* [30] which relies first on aligning images with text that make up the training data (multimodal embedding) before using the inferred alignments in text generation. 18
2.2 An illustrative example comparing top-down and bottom-up captioning. In top-down captioning, the feature information representing the entire image is used in generating the caption. With bottom-up captioning, objects in the image are first identified. The words associated with the identified objects would be used by the language model by generating a caption that includes all the words. 21
2.3 The template-based captioning method of Farhadi *et al.* [52] where every image-caption pair has a meaning space representation in the form of <object, action, scene>. At inference, once the meaning space of an image is determined, it becomes straightforward to generate a caption. The same weight settings for the <object, action, scene> triplet will consistently lead to the same caption being generated. We acknowledge the work of Farhadi *et al.* [52] as being the source of this figure. © Springer 2010. 22
2.4 The neural image captioning (NIC) model introduced by Vinyals *et al.* [24] that later inspired many deep learning based captioning models. We acknowledge the work of Vinyals *et al.* [24] as being the source of the figure. © IEEE 2015. 25
2.5 A Region-based Convolutional Neural Network (R-CNN) is used with a Bidirectional Recurrent Neural Network (BRNN) to generate a similarity score between an image region and a descriptive caption [30]. We acknowledge the work of Karpathy *et al.* [30] as being the source of the figure. © IEEE 2017. 27
2.6 The multimodal RNN generative model of Karpathy *et al.* [69]. We acknowledge the work of Karpathy *et al.* [69] as being the source of the figure. © IEEE 2015. 28
2.7 The captioning framework of Lu *et al.* [70] utilises entity discovery and linking methods. We acknowledge the work of Lu *et al.* [70] as being the source of the figure. 29

3.1 A barchart shows GCS accuracy evaluations with WER is shown. It also shows how phrase hints can improve GCS transcription when transcribing a demo script. Some errors in transcription are to be expected (up to 25% are reported in [93]). However, we noticed a pattern where non-native English speakers who spoke with a lower pitch have more errors in their transcribed speech than their counterparts. 45
3.2 Evaluating GCS accuracy with WER. It also shows how phrase hints can improve GCS transcription when transcribing a list of terms. . 46
3.3 A piechart showing the distribution of parts of speech in Dataset Cap. 1. 48
3.4 The data acquisition and processing pipeline for restrospectively acquired audio is shown from recording sonographer audio to the preparation of image caption pairs. A sonographer watches a scan video while speaking about the contents of the video in the same manner they would during a scan. Their speech is recorded and then passed through the GCS API to Google Cloud Speech in order to get a transcription of the audio recording. From Google Cloud Speech, we obtain the transcription along with the timestamps of when each word in the transcription was uttered. The ELAN annotation tool is used to confirm that the transcriptions match the ultrasound content and to fix any errors that might have arisen during the transcription process. We also combine through the ELAN annotation tool adjacent words to create a full sentence if it is meaningful to do so. From the ELAN annotation tool, we then get the captions with their start and end times. Using these start and end times along with the timestamps of each video frame, it becomes a straightforward process to match captions to the video frames they apply to. 49
3.5 A described structure timeline plot shows us what structure a sonographer is talking about throughout a scan. The numbers on the horizontal axis represent the frame number in an ultrasound scan video. Multiple structures can be present on the screen at the same time; however, the structure that the sonographer speaks about is the one shown in the described structure timeline plot. If the sonographer speaks about the multiple present structures, then we will see thin slices alternating along the timeline plot. An example of this can be seen between frames 3000 and 4000 where there is a sudden change on what structure is being described from abdomen to kidney and then back to abdomen before finally again returning to kidney. 49

3.6 A 2D t-SNE visualisation of Dataset Cap. 2a. 0, 1, 2, and 3 represent the different anatomical structures (abdomen, head, heart, and spine respectively). 54
3.7 Word clouds showing the words spoken by a sonographer during a scan. The larger the word in a word cloud, the more prevalent it is in the dataset. Nouns are shown in red. Adjectives are shown in green. Verbs are shown in blue. Adverbs are shown in yellow. . . . 55
3.8 Video clip of a beating fetal heart centred around a freeze frame (outlined in blue) with the corresponding caption: "*you can see the heart beating very nicely and this is a four chamber view and three vessel trachea view*" is shown. 55
3.9 Some of the microphones that were compared are shown. 56
3.10 The old audio acquisition setup with the SHURE MX184 microphone is shown. In the lower image, when the sonographer is looking at the patient, they would not be speaking in the direction of where the microphone had been placed. 58
3.11 How the Audio Technica AT2035 microphone could be installed in the scan room is shown. 59
3.12 How the JABRA Conference Speakerphone could be installed in the scan room is shown. 60
3.13 The two ways a Tonor Wireless Headset could potentially be worn by a sonographer during a scan are shown. 61
3.14 How a Portable Audio Recorder (such as Roland R26) could be installed in the scanroom. 66
3.15 How a boundary microphone (such as the PCC160) could be installed in the scan room. 67
3.16 How a boundary microphone (such as the MXL AC-404) could be installed in the scan room. 68
3.17 (a) Example of a fetal ultrasound image with sonographer description. (b) Word cloud of most frequently occurring words in sonographer vocabulary. Red, green, and blue represent nouns, adjectives, and verbs, respectively. The size of a word in the word cloud is proportional to its frequency of use. 72

4.1 A high level abstraction of the system architecture initially envisioned at the beginning of the doctoral research. 74

4.2 The image captioning model (concatenation configuration) is shown. A VGG16 that has been fine-tuned on US images is used as a feature extractor for image information. In the text branch, an LSTM-RNN is provided with a sequence of previously generated words as input in the form of Word2vec embedding vectors. Its final hidden state is used as a representation of the sequential text information. The two vectors are concatenated before a prediction for the next word to generate in the sequence is made. At inference, that generated word is added to the partial caption (words generated so far). During training, teacher forcing is used whereby the next word to be added to the partial caption comes from the ground truth caption. In the experiments of this chapter, the vocabulary has a size of 117 unique words. The size of the vocabulary is directly related the size of the dataset, since with more captions, a greater variety of words could be encountered. 76

4.3 An image showing the poorly generated captions using natural image captioning models, showing why fetal image captioning need to be trained separately from natural image captioning models [30]. . . . 77

4.4 A confusion matrix of anatomical labels for the best performing configuration is shown. Based on this matrix, we can conclude that most generated captions describe the right anatomical structure depicted in the image; however, there are cases (11%) when heart-related captions are generated for images depicting abdomens. . . 84

4.5 Ground truth and good generated captions for a couple of images in the test set. The higher the softmax probability associated with a generated word, the darker the green color of that generated word. . 85

4.6 Ground truth and good generated captions for another couple of images in the test set. The higher the softmax probability associated with a generated word, the darker the green color of that generated word. 86

4.7 Examples where the ground truth and generated captions do not match exactly, but the latter may describe the image contents with relevant terminology. This mismatch is reflected in the low objective scores in the Results section. Also, the confusion between heart views is evident in Fig. 4.7a. Please note that the words 'aortic' and 'valve' in the ground truth caption of Fig. 4.7b are not in the training set. The higher the softmax probability associated with a generated word, the darker the green color of that generated word. 87

4.8 Additional examples where the ground truth and generated captions do not match. Note that the stomach is visible in Fig. 4.8b, but the sonographer happened to be talking about a rib in this instance. The higher the softmax probability associated with a generated word, the darker the green color of that generated word. 88

5.1 The image captioning model is shown. The input ultrasound image is that of a heart. 'n' represents the maximum sequence length. A VGG16 that has been fine-tuned on US images is used as a feature extractor for image information. In the lower branch, an LSTM-RNN is provided with a sequence of previously generated words as input in the form of Word2vec embedding vectors. Its final hidden state is used as a representation of the sequential text information. The two vectors are concatenated before a prediction for the next word to generate in the sequence is made. 99

5.2 The distribution of anatomical structures in an MD-IC (upper-left), a CD-IC (upper-right), and a WD-IC (bottom-middle) with non-balanced batches (with respect to anatomy class) is shown. CD-IC is more likely than others to have batches contain samples of the same anatomical class. 103

5.3 Qualitative results for an abdomen image and a heart image are shown. GT is for Ground Truth as spoken by a sonographer. NC is for a model that was trained with 'No Curriculum'. WD-DC is for a model that was trained with a Wasserstein Distance-Dual Curriculum. 103

5.4 The image captioning model architecture is shown. A feature vector from the image branch and the text branch of the model are merged together before predicting the next word in the sequence. The text input is the partial caption. At inference, the partial caption includes the <start> token and the previously predicted words. 106

5.5 Qualitative results for an abdomen image. GT is for Ground Truth. This is the caption that was provided by the sonographer directly. DC is for Dual Curriculum. This is the caption that was generated by a model that was trained with a dual curriculum. I1-CF-DC is for Image-First Course-Focused Dual Curriculum. This is the caption that was generated by a model that was trained with an I1-CF-DC. T1-CF-DC is for Text-First Course-Focused Dual Curriculum. This is the caption that was generated by a model that was trained with a T1-CF-DC. 109

5.6 Qualitative results for a heart image. This is the caption that was provided by the sonographer directly. The full forms of GT, DC, I1-CF-DC, T1-CF-DC can be found in the caption of Fig. 5.5. . . . 109

5.7 A visualisation on how the different similarity measures group data samples. We can see that cosine similarity's subsets are relatively more uniform. With cosine similarity, a subset is more likely to contain data samples of the same anatomical structure. In other words, with cosine similarity, the most similar image to an image is more likely to be of the same anatomical structure, and hence their captions are relatively more likely to be applicable to both. This is important because we rely on image similarity to retrieve captions from which we then extract nouns that along with the anatomical label are used to create pseudo-captions for images that lack captions. 114

5.8 The sequence-to-sequence model architecture that 'translates' the sequence consisting of the anatomical label and the extracted nouns into a pseudo-caption is shown. In the input sequence, the first 'spine' is the anatomical label, and 'end' and the second 'spine' are the extracted nouns. The words in the sequence are represented with word embedding vectors before being passed as input to the encoder and decoder RNNs. 115

5.9 The late merge image captioning model used is shown. The model consists of two branches, an image branch and a text branch. A feature vector associated with the image information from the image branch and a feature vector associated with the text information from the text branch are concatenated together before a prediction of the next word to generate in the sequence is made. Max length represents the maximum number of words a caption of this anatomical structure could consist of. '?' represents the vocabulary size associated with this specific anatomical structure. 116

5.10 Qualitative results for a heart image image and an abdomen image. GT stands for Ground Truth as spoken by a sonographer. NP stands for model trained with No Pseudo-captions. WD stands for model regularized with Word Dropout. WP stands for model trained With Pseudo-captions (our proposed method). 118

6.1 An example of where the ground truth and the generated captions do not match. Note that the stomach is visible in Fig. 6.1a, but the sonographer happened to be talking about a rib in this instance. The higher the softmax probability associated with a generated word, the darker the green color of that generated word. This figure was originally shown in Chapter 4. 124

6.2 Word clouds showing the diversity in the dataset. A larger word occurs more often. Red represents a noun, green represents an adjective, blue represents a verb, and yellow represents an adverb. Words that make up the other parts-of-speech have been dropped for the sake of clarity and to emphasise more meaningful words that come in the form of nouns, adjectives, verbs, and adverbs. 130

6.3 Video clip of a beating fetal heart centred around a freeze frame (outlined in cyan) with the corresponding caption: "*you can see the heart beating very nicely and this is a four chamber view and three vessel trachea view*". 131

6.4 The architecture of the multi-modal model is shown in the this figure. The specifics of the residual operation that is performed on the outputs of Blocks A and C are shown on the right side of this figure. The gazeless model configuration only includes Blocks A and B. Block A represents the branch of the model where the spatial feature information is extracted from the video clip for each of its sampled frames by a VGG16 CNN. Block B represents the branch handling text information. The sequence of words generated so far are tokenized and embedded with a Word2vec embedding vector before being passed as input to an LSTM-RNN. The last hidden state of this LSTM-RNN is concatenated with the linearized feature vector from the convolutional LSTM. The real gaze and predicted gaze model configurations also include Block C. In Block C, either the ground truth visual attention maps are used (in the case of the real gaze configuration) or predicted gaze saliency maps are predicted and then used for each sampled frame in a video clip by a previously trained saliency prediction model. The extracted feature blocks from Block A and the gaze maps from Block C are combined together through the residual operation shown in detail on the right side of the figure. In the gazeless configuration, there is no residual operation to be performed on the sequence of feature blocks. They are passed directly to the convolutional LSTM. 132

6.5 Two fetal ultrasound frames, their real gaze attention maps, and their corresponding predicted gaze attention maps using the method of Cai et al (2018). This figure shows information do the different model configurations (gazeless, real gaze, predicted gaze) learns from. **(a)** A fetal ultrasound frame of a head (left), that frame's real gaze attention map (center), and the corresponding predicted gaze attention map (right). **(b)** A fetal ultrasound frame of a spine (left), that frame's real gaze attention map (center), and the corresponding predicted gaze saliency map (right). The attention in the real gaze attention map is instantaneous in the sense that it reflects where exactly one sonographer looked at on that frame. The predicted gaze saliency map represents an averaged attention that partially mirrors the shape of the anatomical structure on the screen. The real gaze attention map and the predicted gaze saliency map are correlated but are not the same. In both maps, regions in the same vicinity are activated. 136
6.6 Confusion matrices comparing a model trained with cross entropy loss (left) and a model trained with focal loss (right) are shown. . . 137
6.7 A histogram of the most commonly occurring words in the training dataset. Red represents nouns. Green represents adjectives. Blue represents verbs (excluding forms of the verb 'to be'). Yellow represents adverbs. Black represents all other parts-of-speech. . . . 140
6.8 Qualitative results from a random fold are shown. GT is for Ground Truth as spoken by a sonographer. GL is for the gaze-less model configuration. RG is for the real gaze model configuration. PG is for the predicted gaze model configuration. 146

7.1 The image captioning model to be developed in the future. It follows the same skeleton as previously introduced captioning models, but with the the key difference being using transformer blocks in lieu of an LSTM-RNN. 153
7.2 Some captions that have been generated by the transformer-based captioning model are shown. TrF stands for Transformer-based captioning model. 153

B.1 A model trained with MD-IC or WD-IC is able to achieve comparable results within one single epoch of training to a model fully trained through the stochastic mini-batch training process. 162

List of Abbreviations

API . . . . . Application Programming Interface

ARS . . . . . Anatomical Relevance Score

B1 . . . . . BLEU-1

BLEU . . . . . Bilingual Evaluation Understudy

BLEU-1 . . . . Bilingual Evaluation Understudy-1-gram

BLEU-2 . . . . Bilingual Evaluation Understudy-2-gram

BLEU-3 . . . . Bilingual Evaluation Understudy-3-gram

BLEU-4 . . . . Bilingual Evaluation Understudy-4-gram

BLEU-n . . . . Bilingual Evaluation Understudy-n-gram

BRNN . . . . . Bidirectional Recurrent Neural Network

CD . . . . . Cosine Distance

CD-IC . . . . . Cosine Distance-based Image Curriculum

CD-DC . . . . Cosine Distance-based Dual Curriculum

CIDEr . . . . . Consensus-based Image Description Evaluation

CNN . . . . . Convolutional Neural Network

CTC . . . . . Connectionist Temporal Classification

DC . . . . . Dual Curriculum

DNN . . . . . Deep Neural Network

EDL . . . . . Entity Discovery and Linking

EMBR . . . . Edit-based Minimum Bayes Risk

GCS . . . . . Google Cloud Speech

GT . . . . . Ground Truth

IC . . . . . Image Curriculum

LAS . . . . . Listen Attend and Spell

LD . . . . . Levenshtein Distance

LSTM . . . . . Long Short-Term Memory

LSTM-RNN . Long Short-Term Memory-Recurrent Neural Network

MD . . . . . Mahalanobis Distance

MD-IC . . . . Mahalanobis Distance-based Image Curriculum

MD-DC . . . . Mahalanobis Distance-based Dual Curriculum

NC . . . . . No Curriculum

NN . . . . . Nearest Neighbour

NP . . . . . No Pseudo-captions

R-CNN . . . . Region-based Convolutional Neural Network

RNN . . . . . Recurrent Neural Network

ROUGE-L . . Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence

RL . . . . . ROUGE-L

TrF . . . . . Transformer

US . . . . . Ultrasound

VGG16 . . . . Neural network with 16 convolutional layers introduced by the University of Oxford's Visual Geometry Group

WD . . . . . Wasserstein Distance

WD-DC . . . . Wasserstein Distance-based Dual Curriculum

WER . . . . . Word Error Rate

WP . . . . . With Pseudo-captions

Give a man a fish and you feed him for a day. Teach a man how to fish and you feed him for a lifetime.

— Lao Tzu [1]

1 Introduction

Contents

1.1 Clinical Motivation 1
1.2 Contributions and Publications 3
1.3 Thesis Structure 4

1.1 Clinical Motivation

As part of routine care, pregnant women are offered a detailed fetal anomaly ultrasound (US) scan at approximately 20 weeks of gestation to identify any fetal malformations. Medical images can be challenging to describe for the layperson who does not possess the sufficient expert knowledge needed to do so. A tool that can be of assistance in such scenarios can potentially have a wide appeal and prove to be useful to a number of different users. In this thesis, we build machine learning models from audio-video data that can potentially serve as the foundational building blocks of a prototype of such a tool. However, we must keep in consideration the fact that datasets used in medical image analysis come from the real world, and such datasets are often small-sized, in terms of number of data samples, and imbalanced. For that reason, it is essential that we carefully consider

how to prepare the data and how to train a model in such a way that we mitigate the difficulties associated with using real world data.

The technical idea explored in this thesis is to investigate the development of a framework for interpreting ultrasound (US) video automatically and conveying this interpretation in (English) text form. Such computational methodology, fully developed and validated, might in the future provide an educational tool for trainees, support occasional users of US, or assist in conveying information to a patient about a scan in an intuitive way.

The most obvious possible application for text generation in a medical imaging context would be report generation or diagnostics, and there is recent work that has been done for that in the case of diagnostics [2, 3]. In our work, we aim to generate correct and meaningful text that represents how a sonographer may describe the scan that they are conducting. The description must be plausible but is not necessarily diagnostically meaningful and is generated from the real world training audio. Effectively, the models we build would allow play-by-play commentary of fetal ultrasound sonography but not necessarily accurate clinical diagnosis of fetal ultrasound. The raw and temporal nature of our collected data allows us instead to look at more interesting aspects, such as describing the scanning activity. We are also able to explore more nuanced applications that are only possible when in possession of the sonographers' audio recordings that accompany the fetal ultrasound scans and the transcribed textual form of those audio recordings combined with the corresponding visual content.

That is the reason why we have primarily opted to look into the use of deep learning based text generation models in an effort to make it possible to build prototype models that could aid in the dissemination of knowledge from experts to beginners in sonography. There is a need to be met when it comes to making it easier to learn sonography. We intend to do our part in addressing this need through our text generating models.

Text generation could be used in the development of a communication tool that would provide subjects with information from ultrasound visual content in the

absence of a professional sonologist. If embedded within an engineering solution, the research output that will come out of working on this thesis might be used in the future as a part of the building block of an interesting potential product, an ultrasound video player that is able to generate descriptive text on the fly. The text would be describing the fetus to the expecting parents. This allows parents who purchase photos of their fetus to also be given a video of their ultrasound scan. While the scan video is viewed through this media player software built on top of our prototype models, text would be generated that describes to them what is shown on the screen. This application would be a consumer-focused product that might appeal to parents that purchase printed photos of their unborn children. These parents would drive the early demand for a tool of this nature. It is important to note that such a tool is currently an aspiration, but this thesis makes the first step towards this interpretive goal.

Essentially, this research aims to provide a deeper understanding of how to build algorithms with capabilities of this kind by combining and applying machine learning, computer vision, and natural language processing techniques on real world fetal ultrasound video data and their accompanying audio recordings. The aim of our work is to learn joint image-text representations to describe ultrasound images with rich vocabulary consisting of nouns, verbs, and adjectives. The resulting deep learning models may be useful, in the future, if embedded within systems developed to aid in interpreting the contents of frames and clips from ultrasound scan videos.

1.2 Contributions and Publications

- A fetal ultrasound image captioning model. Content associated with this contribution have been published in [4] (Early Acceptance, Oral Presentation).

- Approaches to order and augment data for training through a multi-modal curriculum learning based approaches for captioning and pseudocaption creation. Content associated with this contribution have been published

in [5] (Best Presentation Award, Full Presentation) as well as in [6]. There is a third paper currently being drafted.

- An approach to leverage eye gaze data in the captioning process and a model architecture for fetal ultrasound video captioning. Content associated with this contribution is currently being formatted into a journal paper with the aim of submitting the paper to the Medical Image Analysis journal.

1.3 Thesis Structure

Here in **Chapter 1**, we discuss the motivation for the doctoral research described in this thesis. We also summarise our contributions which are discussed individually in more detail in later chapters. Finally, this section specifically shows the structure of the thesis.

Chapter 2 is the literature review chapter. We summarise literature covering image captioning, US-specific image captioning, video captioning, and the use of eye gaze tracking data in image captioning. A particular focus is given to deep learning based captioning techniques. These are the main topics that are relevant to the work discussed in this thesis.

Chapter 3 describes the data used in this thesis. The details of the dataset are summarised as is the data acquisition process, in particular how the audio data was acquired, the different ways we looked into acquiring audio data, and the strategy we followed in exploring the audio-to-text transcription process.

Chapters 4, 5, and 6 describe the main, novel algorithmic contributions of the thesis. All three chapters share a common theme. They all primarily revolve around how to train fetal ultrasound image (and video) captioning models when in possession of real data which is often of a smaller number than those that make up established benchmark datasets used in natural image captioning. Each of the three chapters proposes a different way to tackle this challenge that is innate to real world medical data; by using (1) specific model architectures, (2) preparing the data for training in a particular way and augmenting it through the creation of

pseudocaptions, or (3) leveraging other available data, such as eye gaze tracking data, to compensate for the data constraints (or supplement available data).

Chapter 4 concerns building fetal image captioning models specifically for our real world data. An automatic natural language processing (NLP)-based image captioning method to describe fetal ultrasound video content by modelling the vocabulary commonly used by sonographers and sonologists is introduced. We present a recurrent neural network (RNN)-based-model that we have built to caption the ultrasound images. We compare RNN, word embedding, and feature vector merging options. This chapter discusses the specific models we have built, training them with the aforementioned real world data. Results show that the proposed models can learn joint representations of images and text to generate relevant and descriptive captions for anatomies, such as the spine, the abdomen, the heart, and the head, in clinical fetal ultrasound scans.

Chapter 5 introduces techniques to better perform fetal image captioning without changing the base model introduced in the previous chapter. First, we explore curriculum learning approaches tailored to the captioning problem; though, the methods introduced are likely to be suitable for other tasks incorporating computer vision and natural language processing. These curriculum learning approaches prepare the data samples for training based on the inherent characteristics of the dataset by re-ordering from 'easy' examples to 'hard' or 'complex' ones. Second, we introduce a data augmentation strategy to create pseudocaptions for images that are in our possession but lack corresponding textual captions. This chapter discusses the advantages of adopting the aforementioned pre-training processes and techniques to overcome the challenge posed by real world data.

Chapter 6 describes how to leverage eye gaze tracking data in video clip captioning to overcome the challenge posed by real world data. We have also built a model that can use human gaze in generating text and a model that can rely on predicted gaze coming from a saliency prediction model to generate text.

Chapter 7 presents a brief summary of the thesis contributions and how the doctoral thesis opens up some avenues for future research.

There is nothing impossible to him who will try.

— (traditionally attributed to) Alexander the Great of Macedonia [7]

2 Literature Review

Contents

2.1 Introduction 7
2.2 Speech: Augmenting and Identifying Audible Human Expressions 8
2.2.1 Reasons for Conducting a Literature Review on Speech 8
2.2.2 Audio Data Augmentation 9
2.2.3 Sound eXchange (SoX) 12
2.3 Transcribed Speech is Text: Transcribing Spoken Words 14
2.3.1 Speech Transcription in a Medical Context 14
2.4 Text and Images: Multimodal Data for the Captioning Task 16
2.4.1 Image Captioning 16
2.4.2 Ultrasound Image Captioning 32
2.4.3 Video Captioning 33
2.4.4 Gaze Tracking Information in Captioning 33
2.4.5 Medical Image Captioning and Curriculum Learning . . 34
2.5 Conclusion 34

2.1 Introduction

In the first part of this literature review, we review relevant literature on speech recognition and transcription. Speech has a relevant role in the project as it is one of our sources of raw data (recorded audio). We are interested in the speech of sonographers, as they provide their own personal commentaries on ultrasound video.

Another objective of this review is to provide supporting evidence to determine ways to handle audio datasets of intelligible speech. Is there a way to augment audio data in the same way we would augment image data, and if so, is it effective? How have other researchers created their own audio datasets? The first part of the literature review attempts to address these questions. Reviewing material on speech was essential, as we investigated how to build the dataset required to conduct research.

The second part of the literature review is on image captioning. The thesis is primarily concerned with correspondence mapping between speech-transcribed text and ultrasound visuals to generate descriptive captions, and hence the second section of the literature review is dedicated to reviewing work on image captioning in particular, but also includes a section on video captioning as well as one on curriculum learning and how it has been used in medical imaging.

2.2 Speech: Augmenting and Identifying Audible Human Expressions

This section is split into two subsections. Subsection 2.2.2 concerns handling datasets of audio and augmenting them. Subsection 2.3.1 discusses speech transcription in a medical context. While there are all kinds of interesting work being done with deep learning on audio, the review primarily focuses on speech rather than environmental sounds (such as sirens for example) or the like; although theoretically, the techniques in the papers discussed could be equally plausibly appropriate for non-comprehensible non-human sounds if the models are trained to detect them.

2.2.1 Reasons for Conducting a Literature Review on Speech

Our goal is to transcribe audio into text which is then mapped to video frame segments with emphasis on ensuring correct synchronization between that which is said and that which is seen. Chapter 3 is dedicated to how we prepared our data including details on how the audio was transcribed. For that reason, we felt that there is a need to discuss speech and audio at length in the literature review chapter.

If one were to consider developing their own audio transcription model, then cleaning, pre-processing, augmenting, and managing noisy data are important steps that need to be considered in the audio-to-text transformation. The Google Cloud Speech (GCS) application programming interface (API) is a readily available commercial solution for speech recognition and transcription [8] that we have used due to its ease of use, relative accuracy, and low cost, but it was still important at the early stages of the work to look at what other researchers are using and creating themselves and considering whether there is merit in building our own speech transcription approach or picking up an off-the-shelf tool and integrating it into our framework.

2.2.2 Audio Data Augmentation

Data augmentation is used to combat model overfitting by increasing the amount of available training data [9]. In order to train models to be more robust when deployed in real world situations and environments and be able to perform well even under a variety of different conditions, Chiu *et al.* [10] add "artificial noise", including background human noise such as office chatter, to speech recordings. Augmenting audio also helps prevent overfitting that is observed when only using the original data in the models being developed. Four different ways to augment audio data are introduced below:

1. **Multi-Style Training (MTR)** is an audio augmentation technique used to make a speech recognition model more resilient when under "stress", which have been described as conditions of high noise that are not encountered during training [11]. MTR is used to prevent overfitting [10]. It requires the same speakers to speak in a number of different styles, such as the way they speak after returning from work or while driving a car, that are different from how they usually speak in normal conditions.

   It is not difficult for speakers to produce speech of different styles, such as the two mentioned above [11]. As a form of MTR, Chiu *et al.* [10] added twenty

kinds of noise including café sounds and background music. This addition of noise resulted in a signal to noise ratio (SNR) that ranged between 5 and 25 decibels. However, Ko *et al.* [9] make the point that while adding noise to the audio will improve the resilience of the model when it attempts to recognize and transcribe speech in everyday conditions, there is no clear evidence that it would be helpful when recognizing audio in a controlled environment where there is a single speaker whose speech we are interested in. This opinion stands against that of Lippmann *et al.* [11] who say that even in a stress-free environment, MTR can still improve performance because it also makes a model more adaptable to the variations that exist in 'normal' speech from one human to another.

2. **Vocal Tract Length Perturbation.** (VTLP) is about applying the reverse of Vocal Tract Length Normalisation (VTLN), a speech normalisation technique where a random warp factor is applied in each time-step of training [12, 13]. VTLP is another audio data augmentation technique, but Ko *et al.* [9] does not consider it to be as effective as augmenting audio data by changing the speed of the audio signal. Audio speed perturbation has the same effect as VTLP and pitch perturbation, but it performs better than both techniques separately and if the two techniques are combined, by applying one after the other. The authors test out audio speed perturbation on four different tasks with initial hours of audio data ranging from 100 to 960. Doing so is an attempt to showcase the wide applicability of their approach [9]. although, it is possible to keep the warping factors constant throughout as Ko *et al.* [9] have done. The warp factor α is chosen randomly [12]. The range it could be chosen from can be specified as well in the same way that Ko *et al.* [9] have done where they fixed the range to be [0.9, 1.1]. Equation 2.1 shows the transformation used by Ko *et al.* [9] which is based on the speech normalisation method of Lee and Rose [14].

$$f' = \begin{Bmatrix} f\alpha & f \leq F_{hi}\frac{min(\alpha,1)}{\alpha} \\ S/2 - \frac{S/2-F_{hi}min(\alpha,1)}{S/2-F_{hi}\frac{min(\alpha,1)}{\alpha}}(S/2-f) & otherwise \end{Bmatrix} \quad (2.1)$$

f is the frequency. f' is the warped frequency. S is the sampling frequency. F_{hi} is a boundary frequency [12]. Applying VTLP leads to an augmented dataset which contains variances of what originally was the same input data. What happens in this kind of perturbation is that some parameters and values are changed in order to artificially produce sounds that resemble those that would be produced by vocal tracts of different lengths. There are a number of ways to perform data augmentation through VTLP, but the approach taken and favoured by Ko *et al.* [9] is to be consistent in applying the same warping factors on each sample of audio data. Two audio datasets would be created by applying two sets of warping factors. The first consists of 0.9, 1.0, 1.1, and the second consists of 0.9, 0.95, 1.0, 1.05, 1.1. Each original feature vector of the audio will be warped by the warping factors. Each of the two datasets is then used to train two different deep neural networks [9].

3. **Speed Perturbation** results in a time signal warped by some warping factor. Speed perturbation has effects similar to those produced by VLTP but with the added effect of a change in the number of frames for a single utterance of speech.

   A number of audio augmentation techniques are compared by Ko *et al.* [9], but they find producing copies of an audio signal with different speeds to be the most favourable. For each audio signal, two new audio signals are produced which are simply the original modified by two different speed factors, 0.9 and 1.1. It is about altering the speed of speaking. The authors recommend this approach because of how easy it is to implement and because it can produce an average improvement of 4.3% in word error rate (WER) (6.7% improvement in the Switchboard (SWB) benchmark). They obtained that score by testing the method with four different Large Vocabulary Continuous Speech Recognition (LVCSR) tasks. The Switchboard evaluation contains

data from the CallHome English dataset which contains speech uttered with accents that are foreign, and therefore, it represents a challenging speech recognition task [9].

Audio speed perturbation has the same effect as VTLP and tempo perturbation, but it performs better than both techniques separately and if the two techniques are combined. The authors test audio speed perturbation on four different tasks with raw audio data ranging from 100 to 960 hours, showcasing the wide applicability of their approach that is capable of dealing with audio data of different lengths [9].

Speed perturbation results in a time signal warped by some warping factor. Speed perturbation has effects similar to those produced by VLTP but with the added effect of a change in the number of frames for a single utterance of speech. Using three-fold (90%, 100%, and 110%) speed perturbation data augmentation gave the best results for the LVCSR tasks in the form of improvement in WER [9]. Three-fold in this context refers to to the fact that a system was used where three modified feature vectors were made from the original feature vector [9].

4. **Tempo Perturbation.** Tempo is also referred to as speech rate [9]. The difference between tempo perturbation and the previously mentioned speech perturbation, as suggested by Ko *et al.* [9], is that in tempo perturbation, although the tempo would change by a factor that is randomly chosen [15], it would not cause a change in the pitch of the audio.

2.2.3 Sound eXchange (SoX)

Sound eXchange (SoX) is a command prompt based utility used to manipulate audio. It can be used to apply speed perturbation. It is interesting to note that Google used SoX as well when demonstrating the power of their GCS API [16], which is what we then decided to use for audio transcription. SoX is primarily used to change an audio's sampling rate [17]. It is also used as an audio recorder

Table 2.1: A table showing performance of different configuration of different perturbation techniques including VTLP, tempo-perturbation, and speed perturbation through word error rate [9]. We acknowledge Ko *et al.* [9] as being the source of this table.

System	Fold	Epochs	SWB	CHE	Total
Baseline	1	6	13.7	27.7	20.7
VTLP	3	2	13.1	26.5	19.9
VTLP	5	2	13.2	26.7	20
VTLP+time-warped	3	2	13.3	26.8	20.1
Tempo-perturbed	3	2	13.5	27	20.3
Speed-perturbed	3	2	13.1	26.1	19.7
Speed-perturbed	3	6	12.9	25.7	19.3

which is how Google used SoX during a demonstration of the Google Cloud Speech Application Programming Interface (API) [16]. Tempo perturbation can also be done using the SoX utility, which is built on a Waveform Similarity based OverLap-Add (WSOLA) technique [18]. In addition to the original audio, a copy of it with tempo of 90% and one with a tempo of 110% are produced. The SoX utility can also be used for speed perturbation by creating copies with modified speeds of 90% and 110%, increasing the audio dataset by a factor of three. Using three-fold (90%, 100%, and 110%) speed perturbation data augmentation gave the best results. After speed perturbation, VTLP came second in effectiveness. Tempo perturbation came in last place; however, it did lead to better results than in the case where no data augmentation was performed [9]. Table 2.1 compares the results of the different techniques.

Ragni *et al.* [19] and Kanda *et al.* [20] attempt data augmentation for languages that have little data to work with. This was initially seen as promising for our project because while there is no shortage of audio data in the English language, there is not an abundant of openly available sonographic medical English speech, so we did think that perhaps the same techniques used in those two publications could be applicable for our use case.

2.3 Transcribed Speech is Text: Transcribing Spoken Words

2.3.1 Speech Transcription in a Medical Context

The system developed by Chiu *et al.* [10] transcribes verbal communication in a medical context between a doctor and a patient. Approximately 14,000 hours' worth of data was available. The average length of a conversation was 10 minutes, but there were conversations up to 120 minutes. Some of the speech that was transcribed was of a medical nature, since the audio came from a natural form of communication, a conversation between a patient and the doctor treating the patient. The data was anonymized, and any text that could contribute to identifying individual persons in the transcripts of the conversations was also anonymized. In our case, patient information did not need to be mentioned much, if at all. This paper is related to our work because in both projects the speech intended for transcription is of a medical nature [10]. There are two key differences however:

1. Our project involves a single speaker, the sonographer. The raw unedited audio might include more voices, but effectively, for training the models only the text from the sonographer is relevant.

2. Chiu *et al.* [10] transcribe a natural form of communication, a conversation between a patient and the doctor treating the patient. The speech to be transcribed in our project, on the other hand, are the thoughts of the sonographer spoken out loud for our analysis and research.

Chiu *et al.* built two models:

1. A Connectionist Temporal Classification (CTC) phoneme based RNN model. The Oxford dictionary defines a phoneme as being "perceptually distinct units of sound in a specified language that distinguish one word from another" [21]. In simple terms, it is the sound an English consonant letter makes. Both models built by Chiu *et al.* [10] would be appropriate for medical transcription

although audio data cleanup would be necessary, especially for the CTC model. The CTC model performance was measured by word error rate (WER). It achieved a WER score of 20.1% [10]. The CTC model was evaluated based on its ability to detect medical phrases. Two CTC models were trained, one unidirectional and the other bidirectional. The bidirectional model had a recall rate of 86% and a precision of 92%, while the unidirectional model had a recall rate of 84% and a precision of 88%.

2. A Listen Attend and Spell (LAS) grapheme based end-to-end model. A grapheme is defined as being "the smallest meaningful contrastive unit in a writing system" [22], so an alphabetic letter is a grapheme. The LAS model proved to be relatively more resilient to noise in the audio data [10] and not reliant on language models. A language model is a model with a probability distribution over the vocabulary that makes it possible to predict what word usually comes after a sequence of other words in a phrase or sentence. This is useful in determining which of two or more words is said when they sound the same but are written and mean different things [23]. In general, having a language model that is adapted for the medical domain improves speech recognition accuracy. Accuracy is more important in such cases than in others because a bad transcription could lead to a bad prescription, for example, and eventually clinically harming the patient [10]. LAS models achieved a WER of 18.1% [10]. The LAS model was evaluated based on its ability to pick up the names of treatment drugs. During their experiments, both types of models had the same training, validation, and test datasets [10].

The attention model performs well with single word utterances even if they are not very audible and can handle sudden truncations in an audio segment better. CTC, on the other hand, has most of its errors in transcribing the beginnings and ends of audible utterances. However, LAS does make noticeable mistakes in transcribing the non-medical parts of the conversation. Chiu *et al.* [10] suggest this is so because with the LAS model, there is no "external language model" to

make use of. To improve its performance, the unidirectional CTC model needed to be trained with an output delay. The model's output for the first f frames is dropped and the last input frame is re-fed into the model f times [10].

The CTC model performance was measured by word error rate (WER). It achieved a score of 20.1%, while the LAS model achieved 18.1% [10]. It is important to point out though that the two were not compared the same way. The CTC model was evaluated on its ability to detect all medical phrases, while the LAS model only had to pick up the names of treatment drugs; although, the authors did at least ensure that when training, both types of models would have the same training and test data sets. There were two kinds of CTC models made, one unidirectional and the other bidirectional. The bidirectional had a higher recall rate and precision (86% vs. 84%, 92% vs. 88%).

2.4 Text and Images: Multimodal Data for the Captioning Task

2.4.1 Image Captioning

To generate a caption that is able to describe the visual information within an image has been a challenge that brought together the fields of computer vision, deep learning, and natural language processing [24]. Vinyals *et al.* [24] makes use of knowledge in all three fields in order to produce captions that are adequate enough to convey in writing the content of provided images. Vinyals *et al.* [24] were able to outperform on metrics, such as BLEU [25], the state-of-the-art models at the time (back in 2015) including work done by Ordonez *et. al* [26] and Kulkarni *et al.* [27]. Starting in 2015, there has been more attention given to image captioning as new challenge datasets, such as the COCO Image Captioning Task Dataset [28], became available and new approaches to use them were introduced [29].

Image-Text Alignment

Alignment between visual material and text is considered by Karpathy *et al.* [30] with the general concept represented in Figure 2.1. The overall process that

Karpathy *et al.* [30] discusses has two main parts. The input to the whole system has images accompanied by sentences that describe them. The output includes generated sentences for entire images as well as regions within an image.

In the approach, the training data include images and their ground truth captions, and the goal is to find some possible alignment between an entry's image and its caption. The alignment is represented with a latent vector. With this vector that captures the alignment, they investigate the possibility of generating text from image or retrieving image from text.

Their alignment model consisted of CNNs and BRNNs that operate in parallel in different pipelines. A multimodal embedding layer is used to align the results of both pipelines despite the modalities being different by having them represented in the same vector space [30].

Once they are embedded in the same vector space, the alignment objective "learns (this common) representation" in such a way that makes it possible for concepts across the two modalities that refer to the same idea or hold closely related meanings to be closely placed in the d-dimensional vector space, where d could be a value from 1000 to 1600, as the work of Karpathy *et al.* [30] shows. The way the alignment is measured is explained below. The alignment model is dependent on the image-caption score S_{ic}, which itself is dependent on scores between image regions and words, since if words apply to regions of the image, one can come to the conclusion that the caption the words formulate apply to the entire image. From that conclusion, arises Karpathy *et al.* [31]'s definition of S_{ic} as being dependent on the dot product between the r-th region in an image and the w-th word in a caption. From this definition of Karpathy *et al.* [31], arises Karpathy *et al.* [30]'s definition of S_{ic} where a word of the caption aligns solely to the image region that it fits the most and is most applicable to.

$$S_{ic} = \sum_{w \in g_c} max_{r \in g_i} v_r^T s_w \tag{2.2}$$

In Eqn. 2.2, w represents a word, g_c represents a collection of parts of a caption c, r represents a region in an image, g_i represents a collection of parts of an image

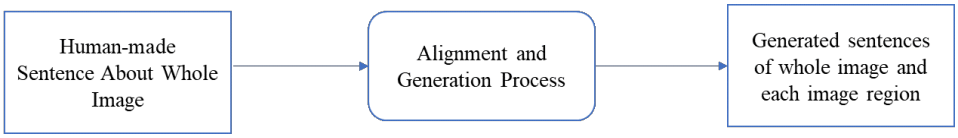


Figure 2.1: Central Concept of the methodology of Karpathy *et al.* [30] which relies first on aligning images with text that make up the training data (multimodal embedding) before using the inferred alignments in text generation.

i, and $v_r^T s_w$ is the dot product between r and w. With S_{ic}, the alignment objective results in having a greater score for image-caption pairs that are related to one another. Pairs that are not related will have that reflected in their low score. The alignment model is useful when using the actual caption generator, which is based on a multimodal RNN that generates new captions for images and their constituent regions by making use of the "inferred alignments" [30].

Evaluation Metrics

Before discussing different image captioning techniques, it is worthwhile to introduce the evaluation metrics used when determining how well an image captioning system or model performs. There are a number of metrics used to evaluate generated image captions including the Bilingual Evaluation Understudy (BLEU) score [25, 32], the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score [33], the F-measure [34], and the Consensus-based Image Description Evaluation (CIDEr) score [35]. BLEU and ROUGE are defined in Chapter 3. Below, we include an explanation of METEOR which will also define the F-measure which METEOR depends on.

METEOR is a common metric used primarily for evaluating machine translations by comparing the unigrams (such as words, for example) in the generated text with the ones in the original human-made ground truth [36]. It is often used to compare different captioning models. The METEOR score is calculated in the following way. The unigram or, specifically in the case of image captioning, word precision and recall need to first be calculated [36].

$$WordPrecision = \frac{words_{mapped}}{words_{generated}} \tag{2.3}$$

$$WordRecall = \frac{words_{mapped}}{words_{groundtruth}}, \tag{2.4}$$

where $words_{mapped}$, represents the number of words that exist in both the generated sentence and the ground truth. The words have been "mapped" from one to the other. The number of words in the generated sentence is represented by $words_{generated}$, while $words_{groundtruth}$ represents the number of words in the original sentence that is being used to evaluate the generated sentence [36].

$WordPrecision$ and $WordRecall$ are combined into a single value, a harmonic mean between the two where greater weighting is placed on one of the two values. [34, 36]. This single value is the F-measure [34]. Here, the F-measure, represented by the F_{mean}, has different weighting for the recall and the precision, with the recall's weight being nine times more than that of precision [36].

$$F_{mean} = \frac{10 * WordPrecision * WordRecall}{WordRecall + 9 * WordPrecision}[36] \tag{2.5}$$

In this context, we define chunks to be sequences of at least two words. Equation 2.5 does not take into account whether or not entire chunks in the generated caption and the ground truth match. If the generated caption and the ground truth are very similar in the sense that they have similar words in a similar order, then it is expected that there will be fewer chunks but ones that are very long. Considering the length of these chunks is a good indicator of how much a generated caption looks like the ground truth one. If the generated caption and the ground truth are completely identical, then there is one long chunk that covers the entire caption. The way that METEOR takes chunks into account is by introducing a penalty that decreases the value of the harmonic mean proportional to the number of chunks by up-to 50%.

$$penalty = 0.5(\frac{chunks}{words_{mapped}})^3 \tag{2.6}$$

$$METEOR\_Score = (1 - penalty) \cdot F_{mean} \tag{2.7}$$

For information on evaluation metrics that we have used throughout the thesis, we refer the reader to Chapter 3. The metrics discussed are useful in many NLP-based tasks; however, we noticed the absence of a domain specific metric, and so in Chapter 4, we introduce our own, the Anatomical Relevance Score.

Image Captioning Techniques

There are currently two high-level established ways to perform image captioning [37, 38]: (a) text retrieval where descriptions are stored beforehand and retrieved using scores between stored and queried images [26]; and (b) text generation where novel text descriptions are generated. The latter is achieved using top-down or bottom-up approaches [39]. In the top-down approach, an image is described by translating visual representations to text, and in the bottom-up approach, constituent objects and concepts in an image are described with words that are then combined into sentences using language models [40]. In both cases, CNNs and RNNs (Recurrent Neural Networks) are built from the images and text, respectively [24, 38]. An example is shown in Fig. 2.2.

Another way to divide image captioning approaches is as follows. Image captioning methods can be broadly divided into three categories [4, 38, 41, 42]; template-based [27], text-retrieval-based [26] and text-generation-based [39, 40]. The different possible established ways to perform image captioning are discussed in detail below:

Template-Based

In template-based captioning, computer vision techniques are used in order to detect objects within an image and obtain its features which are fed into natural language generation models [27, 40, 43]. These models then generate words that can be used to assemble a caption. Attempts to caption images which rely on similarity scores, sentence templates, rule-based text generation, or ontologies can be put in this category. These techniques may operate in combination with extractors of image features [44, 45]. According to Lyndon *et al.* [46] however, these approaches lack the flexibility needed to transcend into a more generalized form of image

(a) A top-down caption.

(b) A bottom-up caption.

Figure 2.2: An illustrative example comparing top-down and bottom-up captioning. In top-down captioning, the feature information representing the entire image is used in generating the caption. With bottom-up captioning, objects in the image are first identified. The words associated with the identified objects would be used by the language model by generating a caption that includes all the words.

captioning. These approaches do not go beyond the highly-specific areas that they have been trained for and the exact captions that they have been trained with, especially when compared to end-to-end deep learning based approaches. In other words, Lyndon *et al.* [46] is suggesting that such approaches do not generalize as well as end-to-end deep learning based methods. Template-based methods typically have a fixed template sentence with blanks which the model fills in with the appropriate words, which can cause a lack of flexibility in the potential output as produced captions end up being very similar to one another structurally and grammatically, following the fixed template.

Some work that attempt to describe action videos make use of templates with

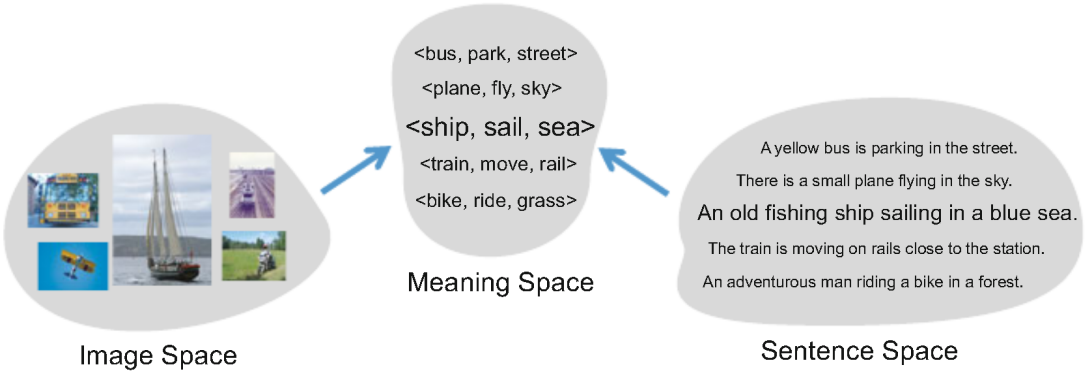


Figure 2.3: The template-based captioning method of Farhadi *et al.* [52] where every image-caption pair has a meaning space representation in the form of <object, action, scene>. At inference, once the meaning space of an image is determined, it becomes straightforward to generate a caption. The same weight settings for the <object, action, scene> triplet will consistently lead to the same caption being generated. We acknowledge the work of Farhadi *et al.* [52] as being the source of this figure. © Springer 2010.

hand-crafted grammatical rules [47–51]. The templates have special positions for different parts of a sentence, such as the sentence's subject, verb, and object. Some of these work then rely on detecting visual concepts in the video to then fill in the template caption with the appropriate words.

Text Retrieval-Based

To do image captioning using a **text retrieval system**, descriptions are retrieved by the system which calculates a score between the model's vocabulary and a test image. The higher this score, the higher the association a word in the vocabulary has to the image the developers are interested in generating a caption of [26, 53–56]. Nearest Neighbour techniques have been used to "generate" a caption of words when provided with an image [29, 52, 55]. Text-retrieval involving Nearest Neighbour (NN) based techniques that perform image captioning are investigated by Devlin *et al.* [29]. The techniques differ in how they determine an image's "nearest neighbours" [29]. They include using simple features like the "gist" of an image [57], using features that are extracted by deep convolutional networks such as VGGNet16 [58], and using features that are extracted by deep convolutional networks created specifically with the intended goal of generating captions. Therefore, those deep convolutional networks were fine-tuned for that specific task accordingly, as was done by Fang *et*

al. [59]. The dataset used in the experiments done by Devlin *et al.* [29] was the MS-COCO dataset [28]. In order to generate a caption for a certain image, the techniques used by Devlin *et al.* [29] attempt to identify a group of "NN images" that come from the training set. A caption of one of the images in the group that is most suitable for the test image would then be returned as the caption of the test image. The most suitable caption is called the "consensus caption" [29]. The most suitable caption is determined by calculating BLEU [25] or CIDEr [35] scores for each couple of captions. Every caption is paired up with every other caption in the group, and the BLEU or CIDEr scores is calculated for every combination. The caption that has the highest average score when compared to the other captions is chosen [29]. Fig. 1 in [29] shows a visualisation of the selection process. Algorithms that retrieve captions can at times outperform methods that generate new descriptions according to Devlin *et al.* [29]. It is suggested by Devlin *et al.* [29] that image captioning systems that rely on text retrieval can have good performance if the dataset consists of images that can be separated into groups of visually similar images. Such datasets make NN approaches to image captioning worthwhile. Fig. 1 in [29] shows a plot where individual points are captions of images in the MS-COCO dataset. One can clearly see that the points are distributed in clusters with points that are within or around the same cluster have captions that are similar. By extension, this similarity suggests that their associated images are also similar. In summary, text retrieval is where descriptions are stored beforehand and retrieved using scores between stored and queried images [26], and these retrieval-based methods rely on finding visually similar images from the training set to the image in question and then associating it with the caption(s) of the most similar image(s). Retrieval-based captioning is when a pre-existing caption (or a combination of captions) is retrieved from the training dataset and associated with a new unseen image [26]. The caption chosen is decided by first determining which of the images in the training set have the highest similarity scores with the unseen image in question.

Generation-Based

In summary, generation, as its name suggests, is the process whereby captions are generated. These generated descriptions might not exist in the training set, but the vocabulary used will come from words encountered in the training set. To use an **RNN** to generate new captions that potentially do not exist in the text of the training set. Image features obtained from a **CNN** would usually influence the recurrent neural network into selecting words from the training set's vocabulary that when assembled and put one after the other make a coherent phrase or sentence that is closely connected to what is visible in the image [38, 60–64]. Depending on the architecture, the RNN may or may not directly come into contact with the image feature information. Text generation, where novel text descriptions are generated, is achieved using top-down or bottom-up approaches [39]. In the top-down approach, an image is described by translating visual representations to text, and in the bottom-up approach, constituent objects and concepts in an image are described with words that are then combined into sentences using language models [40]. In both cases, CNNs and RNNs are built from the images and text, respectively [38, 62]. Generation-based methods typically rely on deep learning models consisting of an encoding CNN to describe an image, and a textual caption is generated by learning joint image-text embeddings, with RNNs [4, 38, 62] or transformers [65, 66], which are neural networks that rely entirely on attention mechanisms in lieu of recurrence or convolutions [67], serving as language models.

Deep learning based techniques that are composed of a text encoding RNN and a text generating RNN had achieved success in the task of machine translation. Inspired by this success, attempts were made to adapt this approach for image captioning by replacing the text encoding RNN with a feature extracting CNN. From 2015 on-wards, it became common to see attempts at methods that are deep learning based and make it possible to generate captions in an automated fashion using models that are end-to-end, as first introduced by Vinyals *et al.* [24]. The form of image captioning inspired by the work of Vinyals *et al.* [24], which they have hence named "neural image captioning", often involves an RNN in the form of

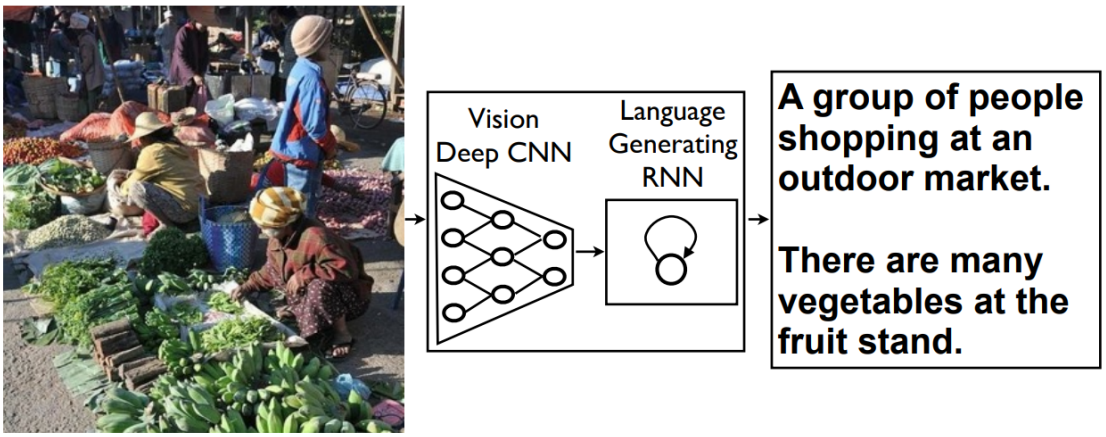


Figure 2.4: The neural image captioning (NIC) model introduced by Vinyals *et al.* [24] that later inspired many deep learning based captioning models. We acknowledge the work of Vinyals *et al.* [24] as being the source of the figure. © IEEE 2015.

a decoder designed to generate text. It is shown in Fig. 2.4. The image information extracted from the CNN would be inserted into the RNN. There is a different way, however, to use an RNN that does not involve such an insertion, in which its sole role would be as a feature extractor of textual information. The image information would instead be merged with the last obtained feature vector of the textual information [38, 68]. This approach is called the "merge model" by Tanti *et al.* [38] who say that, typically, results obtained using the merge model are better than the more common inject model, in which CNN-extracted image information are inserted into an RNN.

This category is where the models we build and introduce in this thesis fall under, but a related sub-category, discussed below, has helped inspire us to think of ways to incorporate more information (not only the traditional image-caption pairs) in the pre-training and training stages of the captioning models. It is for this reason that we have selected these papers to discuss, and that is what makes them relevant to our work.

Hybrid Approaches. There are also intermediate, hybrid ways to caption images which use some of the earlier approaches in addition to deep learning to generate new descriptions. For example, as shown in Fig. 2.5, Karpathy *et al.* [30] use a Region-based CNN (R-CNN) in conjunction with a BRNN in order to

generate a similarity score between an image region and a descriptive caption. The similarity scores play a part when training a caption-generating multimodal RNN.

In the work of Karpathy *et al.* [30], natural language descriptions of images and their regions are generated. Karpathy *et al.* [30] uses natural images and their sentence descriptions to train an alignment model that can create a correspondence between the two types of data. The correspondence is represented by an image sentence score S_{kl}. The alignment model consists of a combination of CNNs over image regions, bidirectional RNNs over sentences, and a final structure that aligns the two modalities through a multimodal embedding.

The alignments are fed into multimodal RNN to learn to generate new descriptions of image regions [30]. In Fig. 2.5, the process with which Karpathy *et al.* [30] evaluates the image-sentence score S_{kl}. Through an R-CNN, regions encompassing an object, will be represented in an image feature vector. Words will also be embedded in with a vector through a BRNN. These two vectors occupy the same multimodal space. The inner product between them is calculated. This is repeated for every object-word combination. The inner product represents the pairwise similarity. From the resulting matrix, the image-sentence score can be obtained.

Fig. 2.6 shows multimodal recurrent neural network of Karpathy *et al.* [30] that is used to generate text. In the first step, the RNN is conditioned on the image information. In later steps, it takes in the context from the previous time step. At every step, there is also another word that goes into the RNN as input. This model uses 'START' and 'END' as special tokens.

A model composed of a CNN followed by a Long Short-Term Memory (LSTM) is used by Lu *et. al* [70] to generate caption templates. The fillable slots in the caption template would then be filled by entity discovery and linking (EDL) methods [71]. Doing it so allows Lu *et. al* [70] to generate captions with their approach that contain a high level of specific detail, a person's name for example instead of just referring to them as 'person', that would otherwise not be straightforward for a pure deep learning-based image captioning system without having many examples of that person and their name in the training data. Most generated captions, as

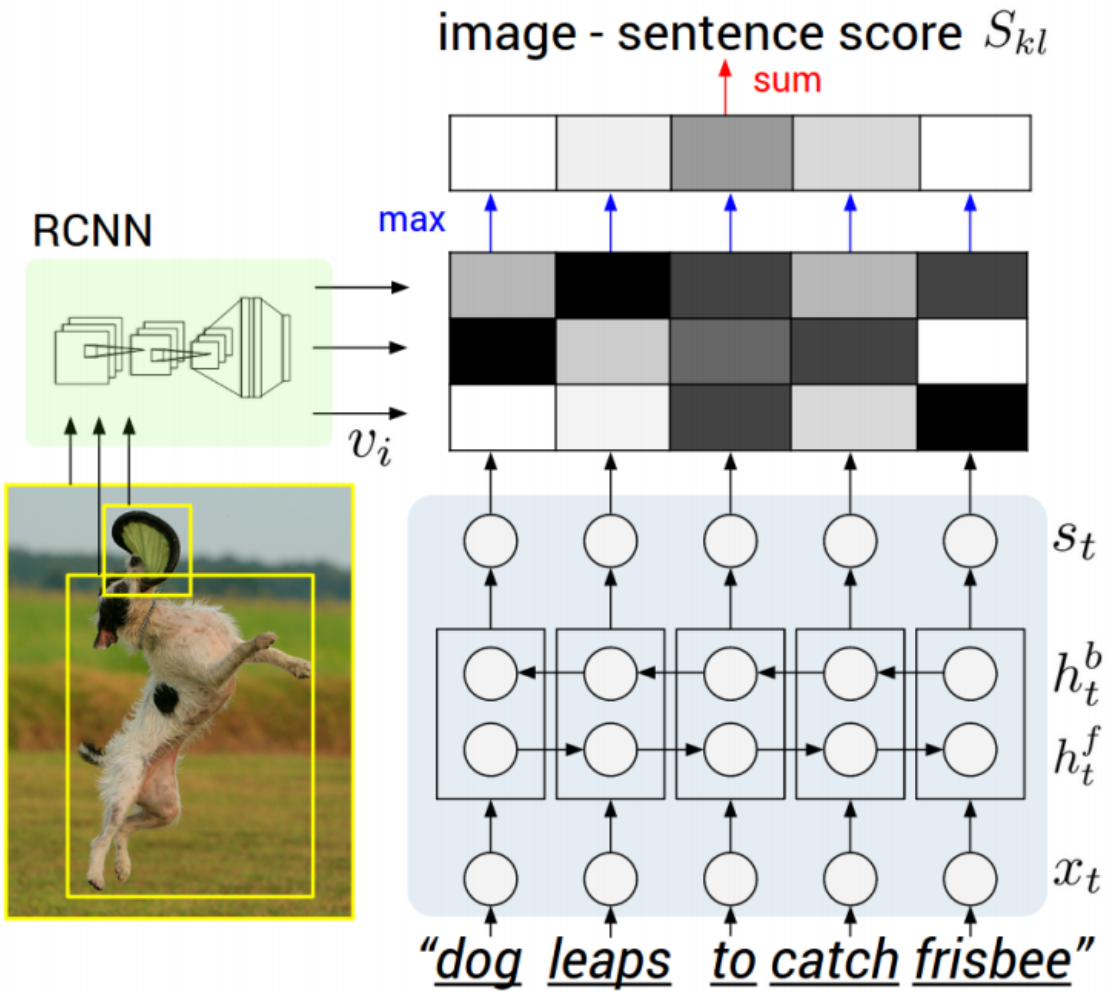


Figure 2.5: A Region-based Convolutional Neural Network (R-CNN) is used with a Bidirectional Recurrent Neural Network (BRNN) to generate a similarity score between an image region and a descriptive caption [30]. We acknowledge the work of Karpathy *et al.* [30] as being the source of the figure. © IEEE 2017.

Lu *et al.* [70] explain, lack the kind of details one would find in a human-made caption. Their adopted framework is illustrated in Fig. 2.7. These details could include named entities, many of which are proper nouns. Their trained system would take in images and hashtags. The hashtags would serve as the source for the named entities, the proper nouns [70]. Their system consisted of a CNN-LSTM model that produces a template in the form of "<team> player celebrates with the trophy in <place> <date>" [70]. The produced templates become proper captions when the "blanks" (the words betweeen < >) in the templates are filled with the specific information, the named entities, obtained from the hashtags that

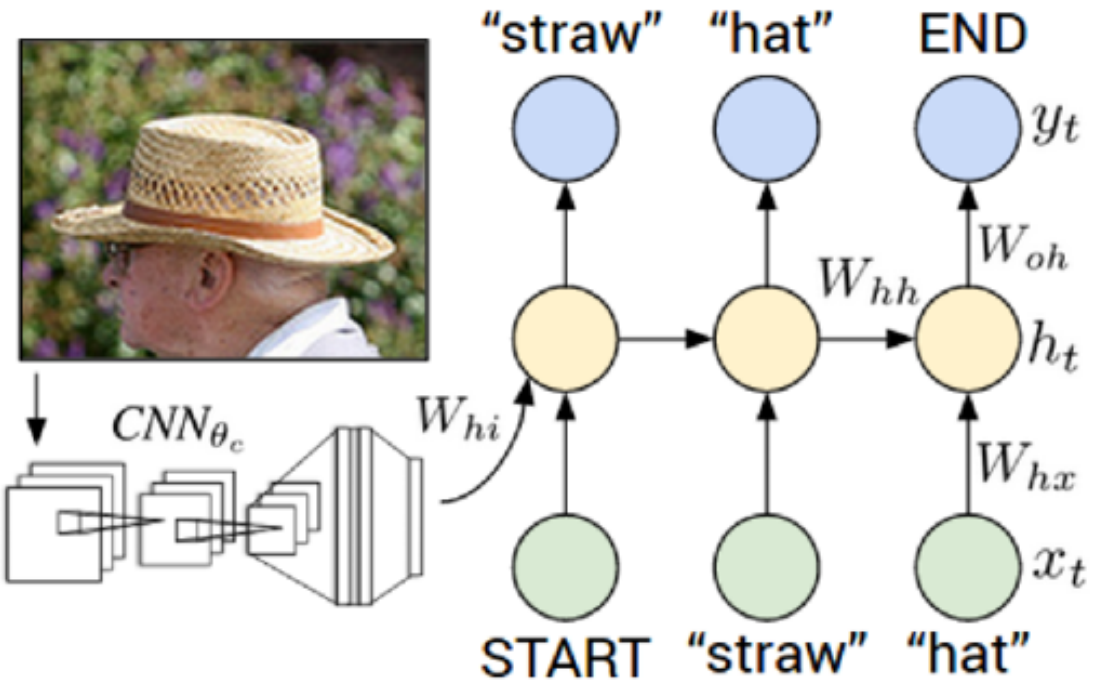


Figure 2.6: The multimodal RNN generative model of Karpathy *et al.* [69]. We acknowledge the work of Karpathy *et al.* [69] as being the source of the figure. © IEEE 2015.

were provided as input with the images. With the named entities and a template, they can get the following caption, "Red Sox player celebrates with the trophy in Boston November 2, 2013" [70]. Experiments on a new dataset consisting of images from Flickr that Lu *et al.* [70] collected themselves demonstrated that their entity-aware captioning system [70] is able to generate captions that are detailed to the extent that they would resemble the kind of writing that a news reporter would prepare for the target image. Their proposed system has a METEOR score on their chosen benchmark dataset that is three times higher than the baseline system raising the score from 4.8 to 13.6 [70].

Another kind of image captioning technique that falls under this sub-category is proposed by Cohn-Gordon *et al.* [72]. A system that incorporates a Rational Speech Acts model [73, 74] that generates captions for images that are unique enough such that they would not be applicable to other images that visually resemble the original image is discussed in the work of Cohn-Gordon *et al.* [72]. A good caption needs to describe the object in an image and perhaps what the object is doing or what is

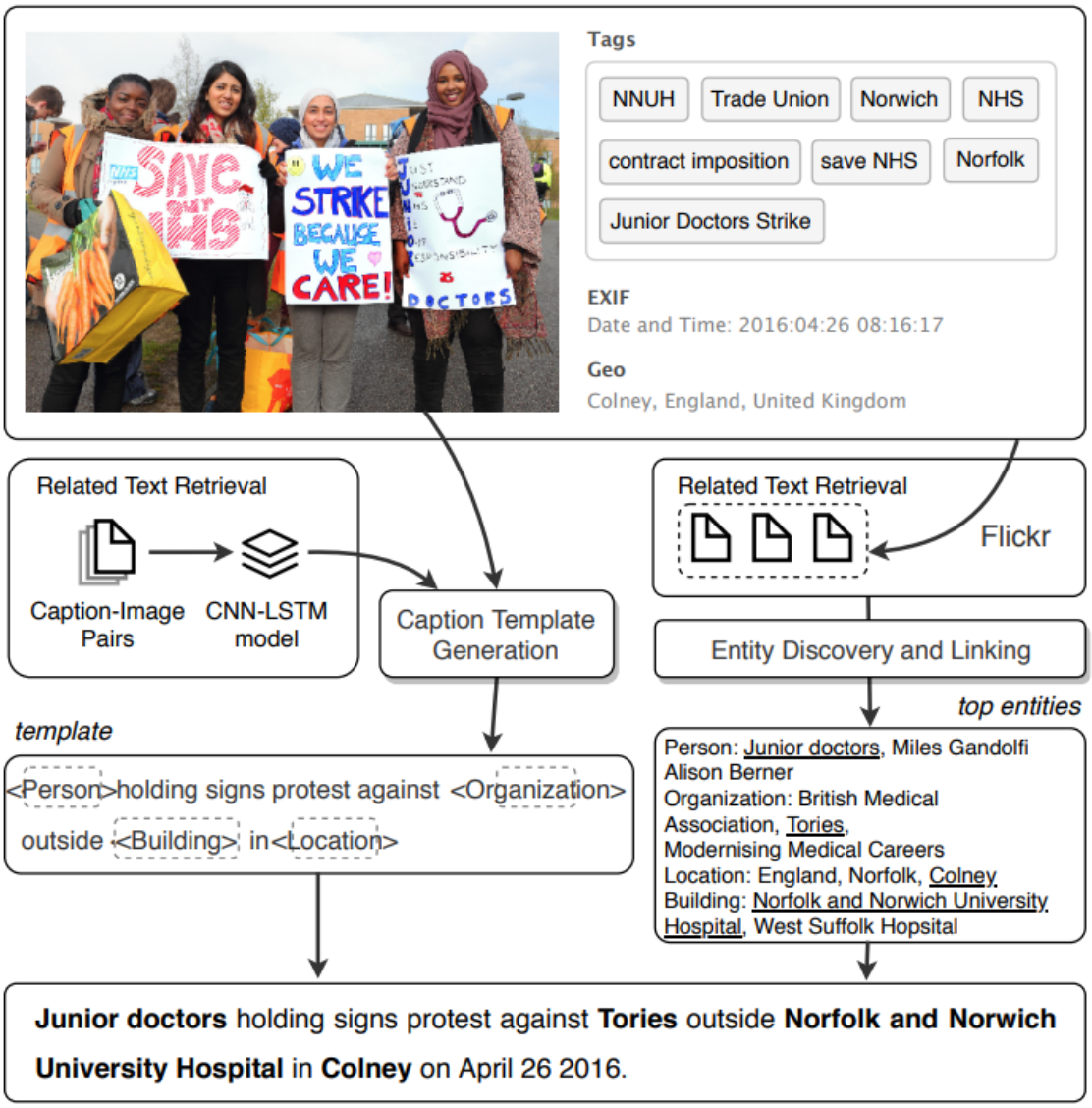


Figure 2.7: The captioning framework of Lu *et al.* [70] utilises entity discovery and linking methods. We acknowledge the work of Lu *et al.* [70] as being the source of the figure.

being done to the object. A good caption is defined as a caption that describes what objects are visible, what actions are being taken by or on the visible objects, and how the objects relate to each other. A visual description is a description of some visual content and the scene that they are part of. In the context of this thesis, the actions (measuring, scanning, and observing), the off-screen object performing these actions (the sonographer), the on-screen objects that these actions are being performed on (the fetus and the mother's abdomen) are fixed. Nonetheless, in a given segment

of time, there is some variation in which of the three actions are taking place and which parts of the fetus are visible on screen during that time segment.

According to Cohn-Gordon *et al.* [72]. An excellent caption needs to go a step further than by making the generated caption unique enough (by using more adjectives for example) such that another image of the same kind of object involved in the same kind of activity would require a different caption to be relevant to it. It is about using specific words to make a description more unique to the target image, as was done in the work of Cohn-Gordon *et al.* [72] where the inclusion of the adjective 'red' resulted in a more unique caption to the target image. What sets this paper apart from others that discuss systems that are able to align images with human intelligible text is that they describe their caption-generating model as being "pragmatically informative" [72] because it takes in the target image and a general caption before generating a more relevant caption. According to Cohn-Gordon *et al.* [72], a pragmatically informative model is one that is able to generate a caption that accurately describes the image in such a way that is unique to it with the generated description being not applicable to different images of the same category or that visually resemble the original image. The authors explain that it is not enough for captions to be correct. They also need to emphasise certain features of the images. This emphasis allows the captions to be appropriately specific, making it possible to set aside a unique image from a group of images that look like it visually. The group of images would be similar and would be part of the same domain or context, but one of them would be different in a specific way. So, for example, all the images could be images of dogs, but the target image is unique because it is the only one that is a close-up of the dog's face, and so the caption, in order to be discriminative itself, would need to include either 'head' or 'face' in addition to 'dog' [72].

Training such a system that is described as being a "pragmatic speaker" requires training related images together at a time. The authors mention that simpler image captioning methods generate captions that are too nonspecific or excessively unnecessarily detailed because such systems do not take into account how similar or different one target image is from a group of visually similar images. They lack

the ability to be pragmatic in their generated descriptions. A Rational Speech Acts (RSA) pragmatic speaker has one goal; to produce an intelligible description for a target image that is part of a group of images with that description being unique to the target in that specific context. The caption would need to be true but also rich enough for the target to be identifiable in this context [72].

Other RSA-based models achieve distinctive captions by inferring and reasoning over all possible utterances or words; however, because of the number of possibilities, that approach is hard to deal with. Cohn-Gordon *et al.* [72] went with their implementation that worked with individual characters, and hence, limited the number of possibilities. The neural language model part of the system produces one character after another while maintaining a balance between the pragmatic informative nature of RSAs and a necessary abidance to the laws of language. The authors say that this approach is better at identifying a target from within a set of similar images [72]. A system that is only able to generate a caption that is true but not distinctive is referred to as a literal speaker [72]. The pragmatic speaker, in order to generate a distinctive caption, would need to be aware of the process undergone by the literal speaker as well as the caption that it would generate. It needs to know how a typical image captioning model would describe the image in order to be able to generate a description that would be distinctive. Fig. 2 in Cohn-Gordon *et al.* [72] shows the difference in the generated captions between a literal speaker and a pragmatic speaker.

Before working on the pragmatic speaker, Cohn-Gordon *et al.* [72] had to first train a literal speaker which would be in the form of a CNN+LSTM model like in the work of Vinyals *et al.* [24]. The trained model would have a probability distribution P(caption|image) [72]. For every image, there would be a set of captions with an accompanying probability value. Captions that are sensible and suitable for the image would have a high value. Ones that are irrelevant would have a low value but not necessarily zero [72]. The novelty in the work of Cohn-Gordon *et al.* [72] is its use of RSA models with character level inference rather than at the level of entire words and utterances. Doing so primarily improves efficiency,

but the authors also demonstrate that it outperforms RSA-based models that with word level inference [72].

2.4.2 Ultrasound Image Captioning

There have only been a few studies investigating the captioning of ultrasound images. We are aware of only two previous ultrasound image captioning works [2, 46]. In the work of Lyndon *et al.* [46], captions are generated for the ImageCLEF dataset including other radiology images, which also uses a top-down deep-learning based text generation approach. In our work, a reduced complexity is achieved using a merged configuration in which image feature vectors are not included as part of the input sequence to the recurrent network. Lyndon *et al.* [46] developed an image captioning model, capable of generating descriptions for a wide variety of medical images by having an RNN come after a CNN that is responsible for extracting image information. The RNN uses the image feature vector from the CNN as it generates text that is descriptive to the image. This is an approach based on the work of Vinyals *et al.* [24], and it has been a popular choice because it discusses the first fully-differentiable deep learning model for caption generation that works end-to-end and produces good results [24]. In the work of Zeng *et al.* [2], the image captioning task is performed on adult abdominal ultrasound with a focus on diseases of the kidney and the gallbladder, where a structure and an associated disease are classified before generating a description with an RNN trained specifically on words of that structure. Both Zeng *et al.* [2] and Lyndon *et al.* [46] use text reports as a raw source of textual data. We use sonographer voice-over recordings to describe the videos in real-time, thereby providing a richer description of the spatio-temporal video content. Our contribution to medical image captioning comes in the form of a convolutional neural network (CNN) based captioning model for second trimester fetal ultrasound images that fuses text and image information for the next word generation. It is described in our work [4] and Chapter 4.

An image captioning framework to caption ultrasound images of the abdomen is also built in the work of Zhu *et al.* [65]. That framework begins with a classifier

to identify the structure of interest in the given abdominal ultrasound image, before passing the ultrasound image to an encoding convolutional neural network (CNN) of the captioning model of that structure. In contrast, in Chapter 5, the classifier is effectively also responsible for encoding the image information for the captioning model.

2.4.3 Video Captioning

In our work, we consider text generation-based methods for video, describing the spatio-temporal visual content with words from the expert vocabulary of medical professionals. We are interested in video captioning as US is a dynamic imaging modality. Sonographers talk about what they see on their screens while they scan, so for our work, the models we train do not read still images for the purpose of creating a radiology report.

For video description and activity recognition, to date, there have been primarily two deep learning approaches to learn spatio-temporal visual information [75]. The first approach employs 3D CNNs to accommodate temporal information in the third dimension. However, a significant amount of data is required to train a 3D CNN [75]. The second approach employs standard 2D CNNs and learns temporal dependencies via RNNs. Due to the flexibility of the latter approach in transfer learning tasks where limited training data is available, we have selected this approach for work in the thesis on video captioning in which we use a convolutional LSTM based RNN architecture.

2.4.4 Gaze Tracking Information in Captioning

The first work to incorporate gaze tracking information in automatic captioning used the ground truth gaze in the form of a histogram of fixations [76]. The image was separated into regions with each region having a corresponding fixation from the histogram. Incorporating the gaze information involved the value of a fixation being multiplied with a feature representation of its corresponding patch [76].

To encourage thorough exploration of the visual scene, Sugano and Bulling [76] used both the human visual attention and computed soft attention [64] to filter the visual features extracted by the encoding CNN. In contrast, in our case, the ground truth visual attention is computed as a 2D attention map around Cartesian coordinates of gaze points [77]. Different from the work of Sugano and Bulling [76], the work done in Chapter 6 modifies the attention filtering mechanism [77] by implementing a residual operation [78], which is computationally more efficient than calculating soft attention.

2.4.5 Medical Image Captioning and Curriculum Learning

In the work of Matiisen *et al.* [79], an algorithm, referred to as the teacher, is used to specify to a student network what sub-task to work on first in a curriculum learning framework that consists of teacher-student networks. A student network is the main machine learning model that is being trained. The teacher algorithm(s) aims to monitor the progress of the student's learning and from that determine what subtasks should the student train for at each training step. Examples of using curriculum learning exist in computer vision [80] and natural language processing [81–83] as well as a few in biomedical imaging applications, such as classification of lesions in chest X-rays [84] to address weakly labeled data and the issues caused by it and detection of cardiac MR motion artefacts [85] to deal with class imbalance and overcome its associated limitations.

2.5 Conclusion

Before discussing fetal image captioning in later chapters and further elaborating our motivations behind it, we briefly summarise image captioning here. Automatic image captioning combines computer vision with natural language processing to generate a textual statement, called a caption, to represent image content. Image captioning has been widely explored in other recent external work for natural images with benchmark datasets [37], however, most established image-captioning datasets do not include medical images, and the majority of image captioning

competitions, such as the COCO 2015 Image Captioning Task challenge [28], reflect this fact. We see relatively few attempts at image captioning when the source of data comes from the medical domain, even less from fetal ultrasound in particular, aside from the ImageCLEF Caption Challenge which focuses on multiple radiology image categories [86].

The man who moves a mountain begins by carrying away small stones.

— Confucius [1]

3 Datasets

Contents

3.1 Introduction 38
3.1.1 Speech Transcription Tools 39
3.2 Descriptions of Datasets 40
3.2.1 Dataset Class. 1 41
3.2.2 Dataset Class. 2 42
3.2.3 Dataset Cap. 1 42
3.2.4 Dataset Cap. 2a 50
3.2.5 Dataset Cap. 2b 50
3.2.6 Dataset Cap. 3 51
3.2.7 Analysis of Sonographer Vocabulary (Freeze Frame Vocabulary vs. Probe Motion Vocabulary) 53
3.3 Microphone Matters 56
3.3.1 Microphone Comparisons 56
3.3.2 Chosen Microphone 62
3.3.3 Audio Challenges 62
3.3.4 Justification for Purchasing and Using Two Microphones 64
3.3.5 Source Separation 64
3.3.6 Choice of USB Audio Interface 65
3.3.7 Placement of Microphones and Audio Interface 65
3.4 Evaluation Metrics 69
3.4.1 Evaluation Metrics for Experiments Conducted in Chapter 5 69
3.4.2 Evaluation Metrics for Experiments Conducted in Chapter 6 70
3.5 Conclusion 71

Table 3.1: A summary of each dataset discussed in this thesis with a brief mention of some its key points is shown with a focus on the data samples used in the experiments.

Dataset	Original Task Associated with Dataset	Number of Samples in Train Set	Number of Samples in Test Set
Dataset Class. 1	Image Classification	990	-
Dataset Class. 2	Image Classification	41029	2721
Dataset Cap. 1	Image Captioning	2800	560
Dataset Cap. 2a	Image Captioning	12808	9979
Dataset Cap. 2b	Image Captioning	41029	14601
Dataset Cap. 3	Video Captioning	198	36

Table 3.2: A summary of each dataset discussed in this thesis with a brief mention of some its key points is shown with a focus on the videos the samples come from.

Dataset	Total Number of Scan Videos	Mean Video Length in Minutes
Dataset Class. 1	71	-
Dataset Class. 2	8	-
Dataset Cap. 1	5	37 (20-56)
Dataset Cap. 2a	10	32 (8-56)
Dataset Cap. 2b	10	32 (8-56)
Dataset Cap. 3	10	32 (8-56)

3.1 Introduction

This chapter describes the datasets that we have used in this thesis and how it was preprocessed and prepared for machine learning modelling. A high level summary of the datasets is provided in Tables 3.1 and 3.2. Subsequent analysis chapters will use the resulting annotated datasets. This chapter also discusses the way that the raw audio data was acquired. For the live audio recording sessions, audio is not recorded for the first 90 seconds after commencing a scan, allowing the sonographer and those present to exchange personal information not pertinent to our research.

A commercial Voluson E8 version BT18 (General Electric Healthcare, Zipf, Austria) ultrasound machines* equipped with standard curvilinear (C2-9-D, C1-5-

*More information on this machine can be found here: `https://www.gehealthcare.co.uk/products/ultrasound/voluson-e8`

D), and 3D/4D (RAB6-D) probes were used to perform all the ultrasound scans used in our work. This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051). Written informed consent was given by all participating pregnant women [87, 88].

3.1.1 Speech Transcription Tools

Dragon Medical [89] and Google Cloud Speech [16] are the two main automatic transcription tools that we investigated. One advantage of Google Cloud Speech is that it provides timestamps with its transcriptions. Timestamps allow better calibration of the text with the video. Dragon Medical uses a subscription model, costing a single user $99 per month in an annually renewable license, while Google Cloud Speech, after the first complementary 60 minutes, charges $0.006 every 15 seconds for up to one million minutes. Google Cloud Speech is a cloud-based tool. Dragon Medical can be used offline but does require online connectivity to authenticate users. Cloud-based tools can raise ethical concerns when it comes to the privacy of the sonographers in retrospective recordings. Therefore, permission from these sonographers was first obtained. One of the benefits of Dragon Medical, however, is that it possesses a large medical vocabulary and is a tool specifically designed for use in a clinical setting.

We evaluated the accuracy of transcribing speech recordings with Google Cloud Speech (version 1.3.2). The accuracy evaluated through the word error rate (WER) calculated using a modified Levenshtein Distance algorithm, which is defined later in this paragraph, focused on word-to-word comparisons rather than character-to-character comparisons. Overall word error rate is important for proper context and understanding; however, WER of specific words are more important. Levenshtein Distance (LD) is effectively a measure of two strings' similarity [90]. It is the number of deletions, insertions, or substitutions needed to make one string identical to the other. The higher the value, the more different the strings are. For example, "Heart" and "Heart" have LD = 0, and "Heart" and "Part" have LD = 2.

Table 3.3: Google Cloud Speech accuracy is evaluated with word error rate. We have a number of different speakers reading the same document. We compared the inclusion of phrase hints in the transcription process. Phrase hints are a feature of Google Cloud Speech where terms and phrases that are likely to be spoken are provided to API hence why they are referred to as phrase hints.

File (Speaker)	Without Hints	With Hints
fetal_A1	20.13%	19.81%
fetal_A2	12.62%	12.94%
fetal_A3	11.66%	11.98%
fetal_A4	15.18%	15.34%
fetal_A5	29.55%	29.87%
fetal_H	18.37%	19.17%
fetal_L	41.69%	40.42%
fetal_P	47.28%	45.53%
fetal_R	24.12%	24.76%
fetal_Y	20.13%	19.81%

Dragon Medical (Practice Edition 4), unlike GCS, does not offer an API that a Python script can rely on to automatically perform large scale transcription of multiple files, making it straightforward to place transcription into the entire pre-processing pipeline. Dragon Medical, instead, offers users an interface where they can have transcriptions generated live during a recording or processing one retrospectively recorded file at a time by importing it into the software. In addition, early qualitative results did not suggest that Dragon Medical would provide more accurate transcriptions for our audio data. Finally, we decided to use Google Cloud Speech for our transcription needs in this thesis.

3.2 Descriptions of Datasets

The PULSE project data includes scans of the second trimester of pregnancy from 18 to 22 weeks of gestational age. Second trimester scans are commonly performed to identify if any anomalies exist in the growing fetus in the UK in accordance with the Fetal Anomaly Screening Programme (FASP) [91]. The sonographers performing these scans look at a number of structures, but in this thesis, we are interested in the abdomen, the head, the heart, and the spine. These are the top four most represented anatomical structures in the PULSE dataset by number of image-caption pairs.

The disease distribution of the cohort in our datasets is not relevant to the objectives of this thesis. In this thesis, we aim to investigate the potential for building ultrasound captioning models reliant on the vocabulary of sonographers as we aim to build and establish a core capability of sonographer-like description generation. We are interested in normal screening examples; therefore, we reiterate that the datasets we have collected are appropriate for our application, since clinical diagnosis and disease detection is not the focus of our work.

All of the video and gaze data has been captured in the PULSE project's ultrasound scanning room and was acquired using the instrumented ultrasound machine (PULSE system) described in section 3.1. An initial set of five audio recordings was acquired retrospectively for five videos (see Section 3.2.3). Subsequently, a further set of five audio-video datasets were acquired "live".

3.2.1 Dataset Class. 1

Dataset Class. 1 consists of images and labels (rather than textual captions). It can potentially be used for classification like tasks. Manual labelling of video clips in the US scan videos based on the viewed anatomies was performed as described in Sharma *et al.* [92]. 990 data samples were used in the finetuning of pre-trained VGG16 that was later used in experiments throughout the chapters of the thesis. When first starting the DPhil program in 2017, it was sensible to use VGG-16 in the start of experimentation. Other work that fell under PULSE had achieved considerable success using VGG-16. For the sake of consistency, we have kept using VGG-16 throughout, for example, when moving on to explore how curriculum learning can be used to improve the performance of the captioning models discussed in the thesis. Using a different CNN, such as Inception-ResNet would not change the messaging of the thesis. Also, work is being done as part of the PULSE project to build PULSE-specific CNN models. The first of such models is based on VGG-16. By consistently using VGG-16 throughout this thesis, we have made it very straightforward for us in the future to replace the model weights with those of the PULSE specific models.

3.2.2 Dataset Class. 2

Dataset Class. 2 is a dataset related to Dataset Class. 1, for they come from the same source. 41,029 images and their labels were used in the process of creating 2,721 pseudo-captions as explained in Chapter 5. These 2,721 pseudo-captions were then used along with Dataset Cap. 2b in the experiments of Section 5.5.

3.2.3 Dataset Cap. 1

This section talks about the acquisition and processing of data that was used specifically for the experiments of Chapter 4 (Image Captioning). Full-length routine fetal second-trimester ultrasound scan videos acquired by an expert sonographer were available for research from the PULSE study [87]. We had the sonographer retrospectively record voice-overs in English for five anonymised videos with a mean duration of 37 minutes (range: 20-56 minutes). A total of 160 minutes of audiovisual content was recorded. From the full-length videos, freeze frames were automatically detected through project specific tools developed by Richard Droste that detects freeze frames through optical character recognition. The sonographers freeze a frame when they find a suitable view of interest for diagnostic examination, which are the anatomical standard planes. The display frame is automatically cropped to include only the ultrasound image without the user interface. The speech recordings were pre-processed for anonymisation and then transcribed using the Google Cloud Speech (GCS) API (version 1.3.2) [8]. At first, we considered one of two speech transcription tools: Dragon Medical [89] or Google Cloud Speech? Dragon Medical's use of a large medical vocabulary reflects its usefulness as a transcription tool for medical purposes. It appears to have a better medical vocabulary, but the question we asked ourselves then was: do we need a transcription tool with a vocabulary for sonography that is that sophisticated? After going through the first few recordings, we were sure we would not. Sonographers' language when speaking about or during a scan turns out to be simpler than how they would describe it when writing up a report. For example, when talking about the healthiness of the heart, sonographers would say something along the lines of 'we can see the baby's heart beating nicely'.

Table 3.4: Examples of misdetected words are shown. 'Occurence' refers to the number of times this word appears in the same documents that the compared speakers read from. 'Errors' refers to the number of times a word was misdetected and transcribed erronenously as a result. '% of time identified wrongly' shows the percentage of time this word was identified wrongly and it is calculated by $Errors/(Occurrence * Number\_of\_Speakers)$.

Word	Occurrence	Errors	% of time identified wrongly
Coiled	1	10	100.00%
foetus'/foetus	8	34	42.50%
belly	2	6	30.00%
abdomen	3	8	26.67%
anatomy	1	7	70.00%

Google Cloud Speech provides timestamps with the transcribed words allowing better calibration of the text with the video along words to appear at the same time as the relevant content in the ultrasound scan videos. It also costs less. The accuracy of the Google Cloud Speech transcription was evaluated through the word error rate (WER). The WER was calculated using a modified Levenshtein Distance (LD) algorithm focused on word-word comparisons rather than character-character comparisons. Overall word error rate is important for proper context and understanding; however, WER of specific words more important. We can do this comparison of specific words by giving a higher penalty for misdetecting an important word or simply not counting words we are not interested in when calculating the word error rate. What determines whether we are interested in a word or not is if the word is relevant to the context associated with fetal ultrasound scanning sessions. One suggestion on how to implement this modified Levenshtein Distance algorithm is to remove filler words when calculating the Distance. While filler words play a role for us in daily conversation, they may hamper the transcription process leading to wrong results. Their presence in sentences would not necessarily take away from the semantic meaning that the sentence would try to convey. LD and WER become a matter of finding out how much of the meaningful relevant words does the transcriber get right. The calculated global, total similarity shows that relevant words are more easily detectable, as Table 3.4 shows for some examples. Occurrence is how many times a word appears in our demo script that we had people read. Such tables allow us to see which words are identified wrongly and how often.

In a way, this modified Levenshtein Distance is a precursor, an inspiration that lead to the evaluation metric (Anatomical Relevance Score) we developed that is described later in this chapter. Levenshtein Distance (LD) is a measure of two strings' similarity. It is the number of deletions, insertions, or substitutions needed to make one string identical to the other. The higher the value, the more different the strings are. For example, "Heart" and "Heart" have $LD = 0$, but "Heart" and "Part" have $LD = 2$. Some words do get misdetected by the transcriber. However, the Google Cloud Speech API allows us to phrase hints to their dictionary to aid in the transcription process [16]. Words that we have seen the transcriber got wrong had been re-added into its dictionary as phrase hints. Adding phrase hints biases the transcriber to properly detect words we need and expect to be said often. In addition, before comparing the original and transcribed documents, all words in the documents were lowercase and the punctuation had been removed because the chosen transcriber does not quite follow proper capitalisation and punctuation rules, and so, I did not want such differences to be classified as errors. In Tables 3.3 and 3.5 and Figs. 3.1 and 3.2, we evaluate GCS's accuracy with WER by having different speakers read a demo script (titled "fetal_*" where '*' is the name of the speaker) and having them read a list of terms (titled "terms_*" where '*' is the name of the speaker). These files can be found in Appendix A. Each recording of each speaker (both demo script and list of terms) has two transcriptions, one produced with phrase hints that are additions to GCS's dictionary and one without. GCS is fine-tuned with the contents of a document. This document contains our phrase hints. This is the same document that speakers read from when recording "terms_" where '*' is the name of the speaker. The WER (word error rate) is reported in percentages after WER is calculated through a script based on the Levenshtein Distance algorithm. What can be noticed is that native English speakers (speakers named A2, A3, and A4) score considerably better than those that are not.

In the future, we may want to consider transcription options that are specifically suited to serve for use cases that handle small data. An interesting thing to explore in the future with the WER is to look and plot how performance changes with

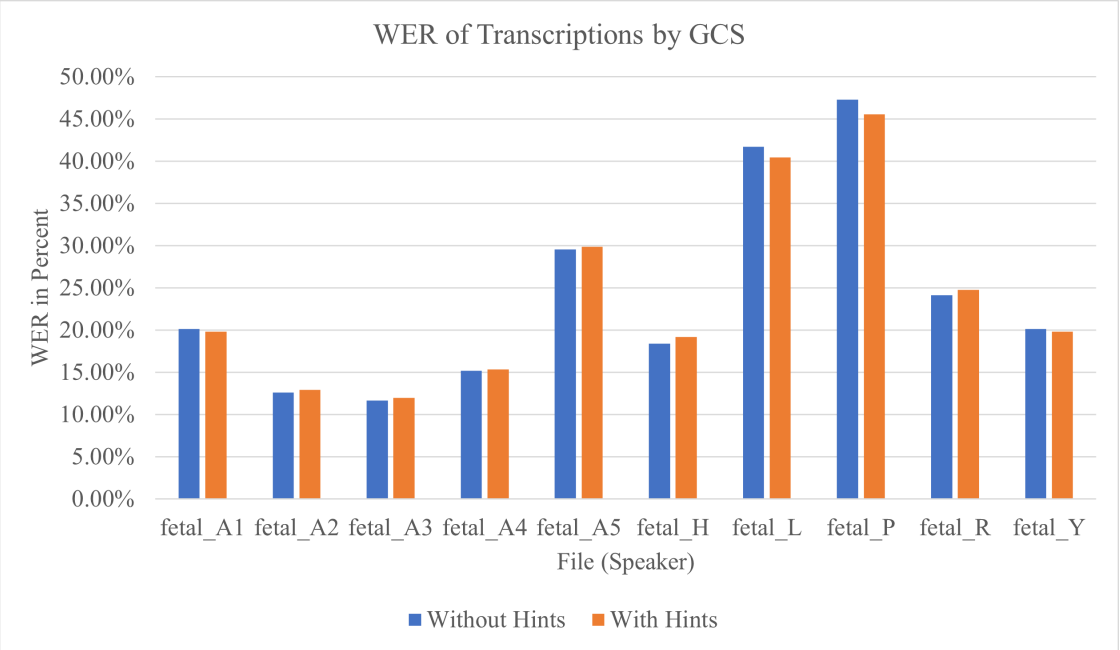


Figure 3.1: A barchart shows GCS accuracy evaluations with WER is shown. It also shows how phrase hints can improve GCS transcription when transcribing a demo script. Some errors in transcription are to be expected (up to 25% are reported in [93]). However, we noticed a pattern where non-native English speakers who spoke with a lower pitch have more errors in their transcribed speech than their counterparts.

Table 3.5: Evaluating GCS accuracy with WER when transcribing a list of terms.

File (Speaker)	Without Hints	With Hints
terms_A1	34.65%	14.85%
terms_A2	15.84%	14.85%
terms_A3	10.89%	5.94%
terms_A4	15.84%	15.84%
terms_A5	23.76%	13.86%
terms_H	39.60%	26.73%
terms_L	26.73%	14.85%
terms_P	64.36%	57.43%
terms_R	45.54%	35.64%
terms_Y	18.81%	10.89%

changes in the amplitude of the audio recordings. It is about seeing how WER changes. It also allows us to answer the question: Is changing the amplitude more important than ensuring proper native pronunciation in audio recordings?

WER matters most for relevant words which are relatively few, but what are the relevant words? They include: (1) medical terms, (2) adjectives and nouns

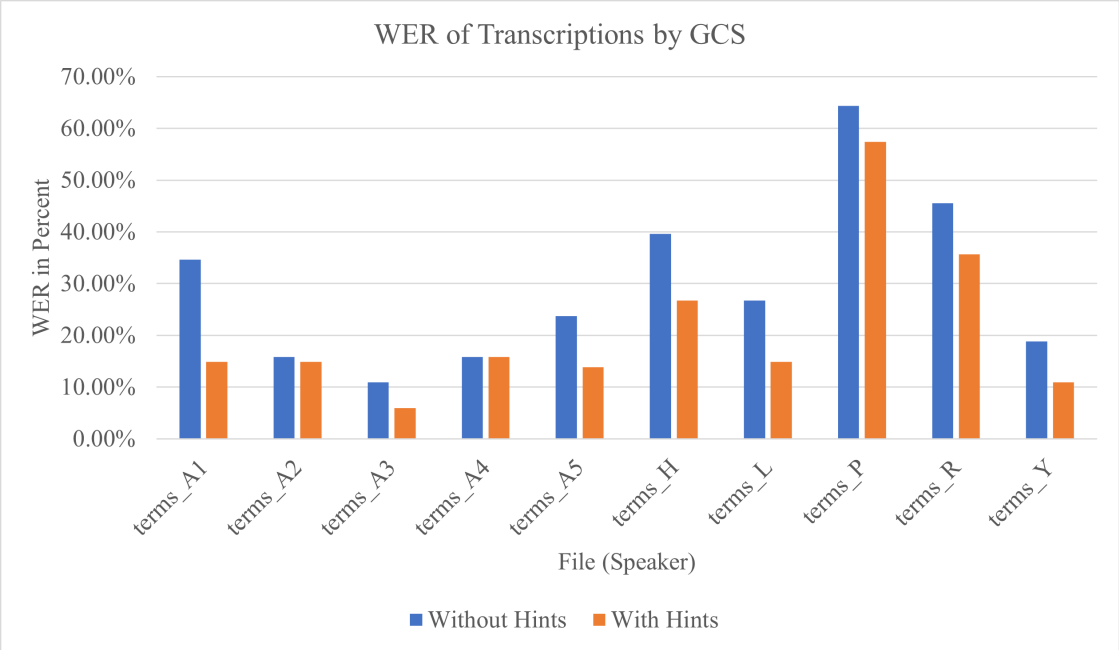


Figure 3.2: Evaluating GCS accuracy with WER. It also shows how phrase hints can improve GCS transcription when transcribing a list of terms.

instrumental in description, and (3) articles and prepositions that are meaning changers (e.g.: using 'the' for specification or distinction). While articles and prepositions do not really carry descriptive weight, their absence or presence changes the meaning of a sentence. Articles such as 'the' indicate form specification or distinction, so can they truly be discarded? While individual words carry or lack of sentiment that can be added up or nullified when words are put in a sentence. Descriptions are more reliant on the choice of words and the order with which the words are written. Nonetheless, overall word error rate does matter. We must keep in mind however that some words are certainly more important than others, but if the switching of articles "the" and "a" for example is enough to change the meaning of our description, then we must take heed of what we drop from our calculations. This thought process on the WER helped in later formulating the Anatomical Relevance Score, particularly in considering the fact that some words are more important and relevant than others.

The speech transcription tool, Google Cloud Speech, is not without its limitations. Detecting speech in audio requires clear pronunciation and using the words in a

sensible context. Some words are only detected if said in a sentence that made sense contextually, as in these words could only be detected if they were put in a sentence that made sense for them to be detected in. For example, 'ears' can be detected as 'years' unless put in a sentence like 'I pierced my ears', and 'sole', as in 'the sole of the foot', can be detected as 'soul' unless used in a sentence talking about a part of the anatomy. GCS processes a recorded audio file in approximately 50 second intervals. Sometimes, the transcriptions are returned by GCS with a couple of words randomly grouped together with no space between them.

GCS is designed for natural language, but the recordings contain additional medical vocabulary. The transcriptions contained a few errors which were corrected by manual post-processing. ELAN, a multimedia annotator for audiovisual content, was used to synchronize video contents with generated transcriptions and to correct erroneous text [94]. After the transcribed words were manually checked, grouped, and synced, a file containing the captions with start and end times was produced to automatically align video frames with captions. Fig. 3.4 shows the process of creating image-caption pairs.

The raw text was cleaned by removing punctuation, replacing numeric characters with their word equivalents, and removing stall words (e.g. 'so yeah', 'well'). Special tokens denoted a caption's start and end. The resulting caption length varied between 1–22 words, with a vocabulary of 158 unique words and distribution of adjectives, determiners, nouns, and verbs is 12.7%, 22.2%, 28.0%, and 16.0%, respectively. The remaining 21.1% are prepositions, pronouns, adverbs, and other parts-of-speech. Hence, the combined dataset was composed of real-world fetal anomaly ultrasound video freeze frames and their associated captions.

For training the deep learning models, we excluded captions that do not describe one of the four anatomical structures of highest interest, *i.e.*, head, heart, spine, and abdomen. These anatomical classes were selected due to having the most representation in the collected dataset compared to other anatomical classes, and they form 40%, 22%, 20%, and 18% of the data, respectively.

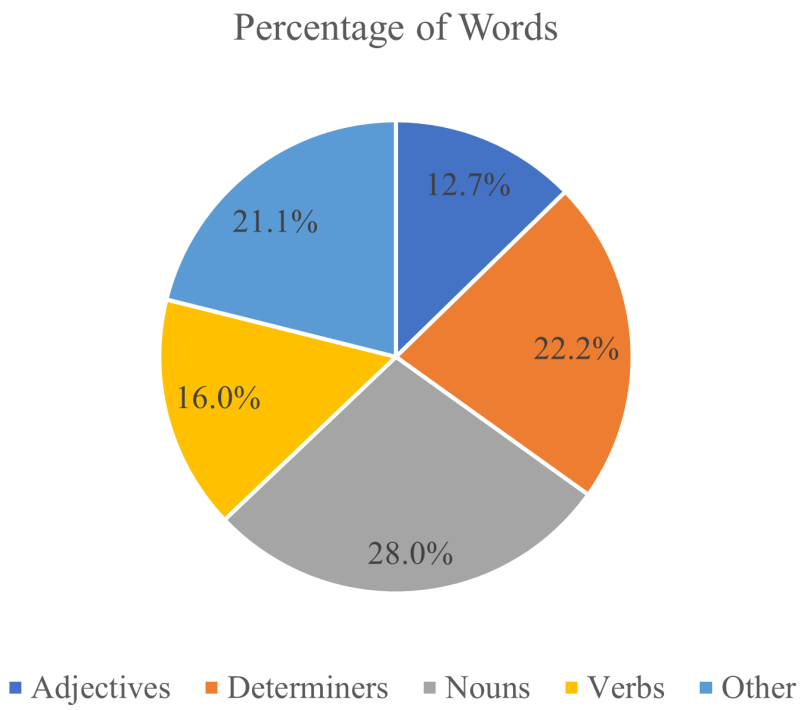


Figure 3.3: A piechart showing the distribution of parts of speech in Dataset Cap. 1.

From caption pre-processing, vocabularies of each of the four anatomical classes of interest were obtained. Lexical diversity scores, specifically MTLD [95], to measure word variety in each vocabulary were obtained as 21.2, 20.3, 17.9, and 21.8, respectively. To address class imbalance, an equal number of unique captions were included for each anatomical class. In addition to class imbalance, caption imbalance is observed as some captions correspond to more than one video frame because sonographers spend a different amount of time looking at different fetal structures. The training set consisted of 2,800 image-caption pairs, and the validation set consisted of 560 image-caption pairs. The images were resized to 224×224 pixels.

The described structure timeline plot in Fig. 3.5 shows what structure a sonographer is talking about for the duration of a scan. The x-axis represents the frame id, with '0' being the first frame in an ultrasound scan video. The described structure timeline also helps to identify where the key frames lie for the duration of an ultrasound scan.

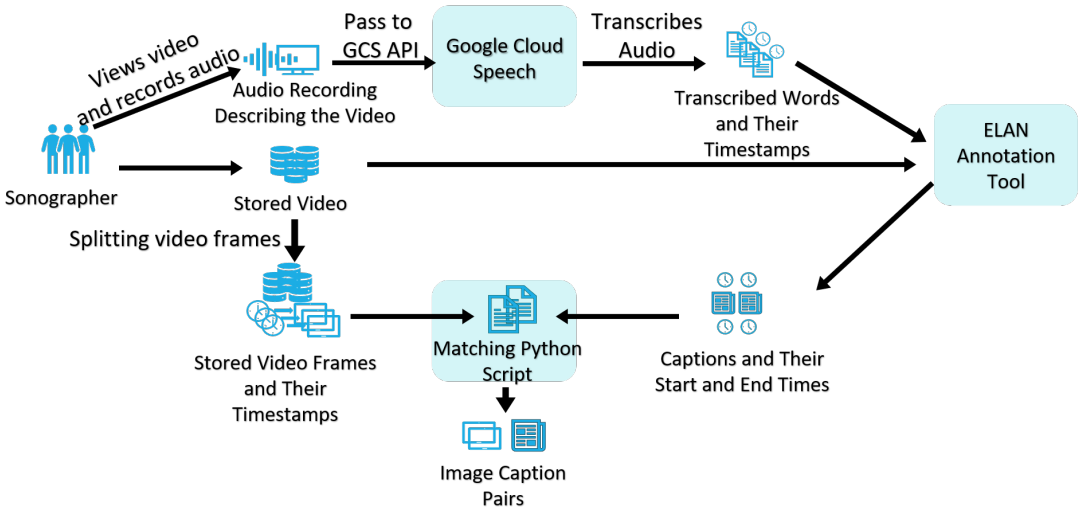


Figure 3.4: The data acquisition and processing pipeline for restrospectively acquired audio is shown from recording sonographer audio to the preparation of image caption pairs. A sonographer watches a scan video while speaking about the contents of the video in the same manner they would during a scan. Their speech is recorded and then passed through the GCS API to Google Cloud Speech in order to get a transcription of the audio recording. From Google Cloud Speech, we obtain the transcription along with the timestamps of when each word in the transcription was uttered. The ELAN annotation tool is used to confirm that the transcriptions match the ultrasound content and to fix any errors that might have arisen during the transcription process. We also combine through the ELAN annotation tool adjacent words to create a full sentence if it is meaningful to do so. From the ELAN annotation tool, we then get the captions with their start and end times. Using these start and end times along with the timestamps of each video frame, it becomes a straightforward process to match captions to the video frames they apply to.

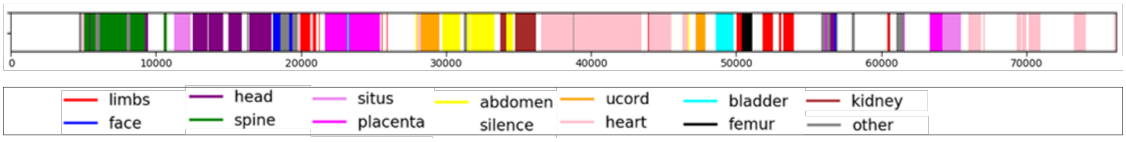


Figure 3.5: A described structure timeline plot shows us what structure a sonographer is talking about throughout a scan. The numbers on the horizontal axis represent the frame number in an ultrasound scan video. Multiple structures can be present on the screen at the same time; however, the structure that the sonographer speaks about is the one shown in the described structure timeline plot. If the sonographer speaks about the multiple present structures, then we will see thin slices alternating along the timeline plot. An example of this can be seen between frames 3000 and 4000 where there is a sudden change on what structure is being described from abdomen to kidney and then back to abdomen before finally again returning to kidney.

Table 3.6: The breakdown of the datasets used in the experiments of Chapters 5. The scan video with the most image-caption pairs was held out as the test set.

Data Information	Training Set	Validation Set	Test Set
Number of US Scans (Videos)	7	2	1
Number of Image-Captioning Pairs	12,808		9,979

3.2.4 Dataset Cap. 2a

This section discusses the data used in Chapter 5 and how it was prepared. We collected a dataset consisting of the timestamps of consecutive frames of ultrasound video scans and the aligned captions that describe their content. What distinguishes this dataset from other medical text-visual datasets is its transcribed speech component. The annotated (transcribed) data comes from ten different scans, each that of a different subject. The scans in this dataset have been performed by two different sonographers. In Fig. 3.6, we visualise a subset of this dataset in 2D using t-SNE (applied on the image feature vectors), while mentioning captions of some of the sampled datapoints.

The breakdown of the data used in Chapter 5 is shown in Table 3.6. Data from seven of the ten videos were used for training. Two of the ten served as a validation set, and the tenth video was the test set. This arrangement gave us 12,808 image-caption pairs to train and validate models on and 9,979 image-caption pairs for model testing.

Pre-processing of the data samples consisted of cropping the images to remove the US system's GUI, resizing the US images to 224×224 pixels, removing punctuation from the captions, and converting uppercase letters to lowercase. Images were randomly augmented by rotations and horizontal flipping.

3.2.5 Dataset Cap. 2b

As part of the PULSE study, videos of full-length fetal ultrasound scans were acquired as well as their accompanying audio recordings. The real image-caption pairs where prepared by transcribing the audio data from 10 scan videos as described in previous sections. The mean length of those ten videos is 32 minutes. When training the

image captioning model, real image-caption pairs from seven videos were used at the training set. Image-caption pairs from two videos were included in the validation set, and the remaining image-caption pairs from the final video made up the test set.

There were a total of 23,558 real image-caption pairs in the training dataset; however, this included successive frames of the same segment that look similar if not the same. In the validation set, there are 17,471 real image-caption pairs. The image-pseudo-caption pairs created from Dataset Class. 2 were used with the training dataset. Although, they numbered only 2,721 image-pseudo-caption pairs, the images are relatively more distinct having included a step size of 16 between each sampled frame. Note that the frames that constitute the image-pseudo-caption pairs come from eight videos that are different from the 10 videos mentioned in the previous paragraph. As part of standard pre-processing practice, the images were cropped to remove the user interface and then resized to 224x224 pixels. The text had its punctuation dropped, and all of its letters were made lowercase.

3.2.6 Dataset Cap. 3

This subsection discusses pre-processing multi-modal data for gaze-assisted video captioning (Chapter 6).

Video Pre-processing

The image-label paired data in Dataset Class. 1 were used in finetuning the VGG16 [58] that was then used to extract frame features.

Clip-Caption Pair Generation

Speech recording of sonographers is acquired for ten routine full-length second-trimester US scan videos from ten subjects, where they described the live contents of the scan including the anatomical knowledge and scanning actions performed. Video captions were obtained after transcribing sonographer speech recordings. Five of the audio recordings were transcribed through the Google Cloud Speech API [96], while the rest containing significant conversation were transcribed manually in order to avoid the accidental transcription of non-sonographer speech consisting of

Anatomy	Distribution
Abdomen	11.7%
Head	31.4%
Heart	40.5%
Spine	16.4%

Table 3.7: The distribution of the anatomical structures in the dataset used in the experiments of Chapter 5.

the speech of the pregnant women or accompanying persons [4]. During a fetal US scan acquisition, the sonographer performs search, fine-tuning and interpretation of different fetal anatomies. The sonographer can freeze at a frame where they view the required anatomical structures, for further interpretation. Based on the freezing action, a video clip corresponding to every unique caption with the centre frame as a freeze frame was extracted. Then, **12 frames** were sampled from the segment spanning the length of the caption. Theoretically, with more sampled frames, we can better capture the temporal changes in information. Early experimentation allowed us to empirically determine this theory. Fig. 6.3 shows a clip-caption pair. The average length of a clip-caption pair in our dataset before concatenation **is 7.25 seconds**. In general, in routine fetal US scans, sonographers spend unequal amounts of time looking at different anatomies of interest, which was also observed in Datasets Class. 1 and Class. 2 [92]. This led to a class imbalance for the four common anatomical classes considered in this work, namely, abdomen, head, heart, and spine. The distribution of the four anatomies is 11.7%, 31.4%, 40.5%, and 16.4% as shown in Table 3.7.

Gaze-Tracking

The use of eye gaze tracking data in this thesis was made possible through the work of Chatelain *et al.* [97] which is also part of the PULSE project. A Tobii Eye Tracker 4C (Tobii Sweden) was used in the work of Chatelain *et al.* [97] and other studies conducted as part of the PULSE project that utilised eye gaze tracking data. This work helped to calibrate the eye gaze tracking, making it possible and

providing guidance for work and studies that make use of eye gaze tracking data, such as gaze points, as done in Chapter 6, for example.

Gaze tracking data (x-y coordinates and time stamps) of sonographers were acquired **at 90*Hz*** using an eye tracker (Tobii, Sweden) as part of the PULSE study, simultaneously with the voice recording and filtered following the protocol of Cai *et al.* [98]. Binary maps are created using gaze data, with gaze-points labeled as 1 and others 0. Sonographer visual attention maps are subsequently generated by convolving the binary map with a **Gaussian kernel of size 40 pixels**, given an observer-to-screen distance of 0.5, human field of view of 1.5°visual angle, and screen dimensions of 33.2 cm $\times$ 20.7 cm. The visual attention map is further normalised so that each pixel value ranges in $[0, 1]$. A *spatio-temporal saliency prediction model* [99] was pre-trained using a large simultaneous gaze-tracking dataset of second-trimester scans as part of the PULSE study. The predicted saliency maps were used as input for one of the model configurations explored in Chapter 6. For that reason, we bring up this saliency prediction model when discussing data preparation.

3.2.7 Analysis of Sonographer Vocabulary (Freeze Frame Vocabulary vs. Probe Motion Vocabulary)

Preparing datasets for the experiments has allowed us to explore the available vocabulary in depth. Implementing and pre-processing data for the captioning models puts us in an ideal position to look into and analyse the vocabulary of sonographers and how they speak. We have noticed that sonographers speak slightly differently depending on whether they have frozen the frame to focus on structures appearing in the frame or if they are moving the probe. The word clouds in Fig. 3.7 show the different words in these two different scenarios.

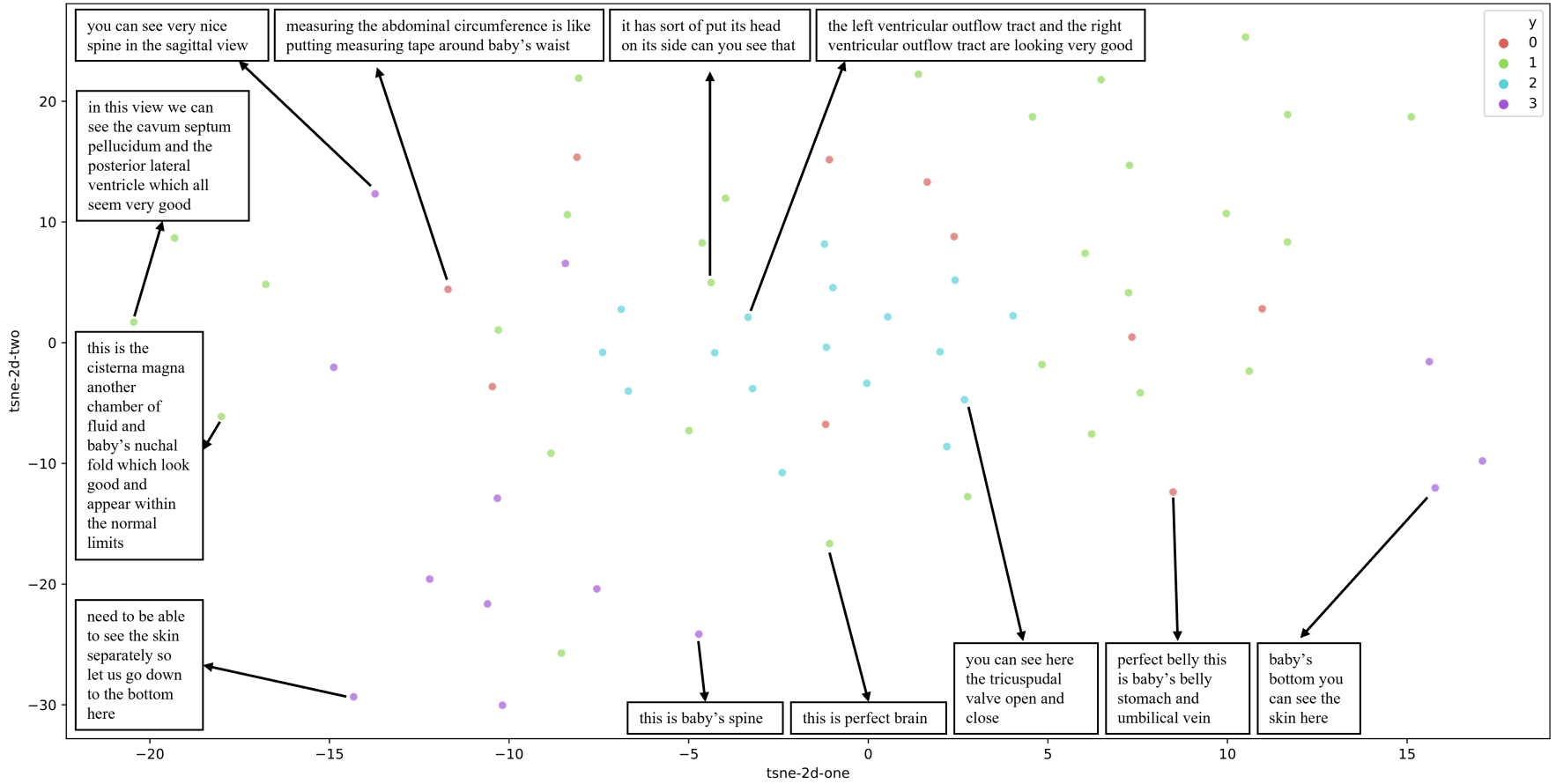


Figure 3.6: A 2D t-SNE visualisation of Dataset Cap. 2a. 0, 1, 2, and 3 represent the different anatomical structures (abdomen, head, heart, and spine respectively).

(a) Words spoken when image is frozen. **(b)** Words spoken when image is not frozen.

Figure 3.7: Word clouds showing the words spoken by a sonographer during a scan. The larger the word in a word cloud, the more prevalent it is in the dataset. Nouns are shown in red. Adjectives are shown in green. Verbs are shown in blue. Adverbs are shown in yellow.

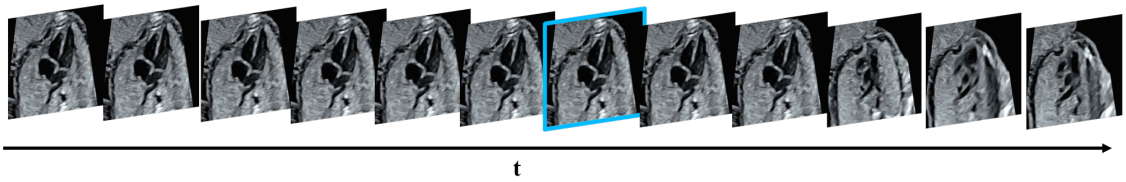


Figure 3.8: Video clip of a beating fetal heart centred around a freeze frame (outlined in blue) with the corresponding caption: "*you can see the heart beating very nicely and this is a four chamber view and three vessel trachea view*" is shown.

3.3 Microphone Matters

3.3.1 Microphone Comparisons

A high level summary of investigated microphones is shown in Table 3.8. The chosen microphone setup is shown in Fig. 3.15 (top). In this subsection, we compare different microphones including a lapel microphone (SHURE MX184[†]), the Audio Technica AT2035[‡], the JABRA Conference Speakerphone[§], the TONOR Wireless Headset Microphone, and the Portable Audio Recorder Roland R-26[¶] as well as two boundary microphones, PCC160[‖] and MXL AC-404[**].

Figure 3.9: Some of the microphones that were compared are shown.

SHURE MX184

The audio quality of this microphone was found to be decent if used as intended ie. if worn by the speaker, for it is a lapel microphone; however, after a couple of weeks of recordings, sonographers were found to prefer to leave it on the desk rather than wearing it on themselves. Doing so made them too far from where the microphone was placed. The kit was also shared by multiple sonographers. This sharing of kit

[†]https://www.shure.com/en-MEA/products/microphones/mx184
[‡]https://www.audio-technica.com/en-us/at2035
[§]https://www.mea.jabra.com/business/speakerphones/jabra-speak-series
[¶]https://www.roland.com/global/products/r-26/
[‖]https://www.akg.com/Microphones/Boundary%20Layer%20Microphones/PCC160.html
[**]https://mxlmics.com/product/ac-404/

happened before the start of the COVID-19 pandemic. Now, this would be more problematic. In order to decide on the quality of the microphone, I would listen to a recording that is at least 5 minutes long and try to see if I can successfully pick up all the words by the sonographer (or the person playing the role of the sonographer) in the audio recording. In order to pick up the voice of a sonographer, the gain had to be boosted which at times led to background noises, such as the air conditioning, to be accentuated. This microphone follows a super-cardioid polar pattern. The microphone was not an impedance to sonographers when they did not wear it. It was easily placed behind the US machine's screen almost hidden from view. There were instances where it could pick up undesired audio though. One of its benefits was the fact that it did not require manual support unless settings on the microphone and the accompanying adaptor were accidentally changed. We define manual support as being me or another member of the team with an engineering background needing to re-set up the microphone system regularly or when recording needs to commence. No manual support indicates that the microphone works seamlessly without requiring sonographers to personally do anything to commence recording. The audio acquired live during scanning with this microphone went on to be transcribed to be used in the experiments conducted for this thesis. It is relatively expensive at £252 without VAT.

Audio Technica AT2035

The Audio Technica AT2035 produced excellent quality, but background noises were still present and could overwhelm speech. The quality was determined in the same way as was done for the SHURE MX184. The microphone is quite big, and it definitely could have impeded sonographers during the scan. It still picked up non-desired audio, but it did not require much manual support unless the microphone and the adaptor were accidentally moved. It is priced at £133.5.

JABRA Conference Speakerphone

The appeal of using the JABRA Conference Speakerphone was its size and portability ensuring its presence would not be a hindrance to the sonographers, but the quality

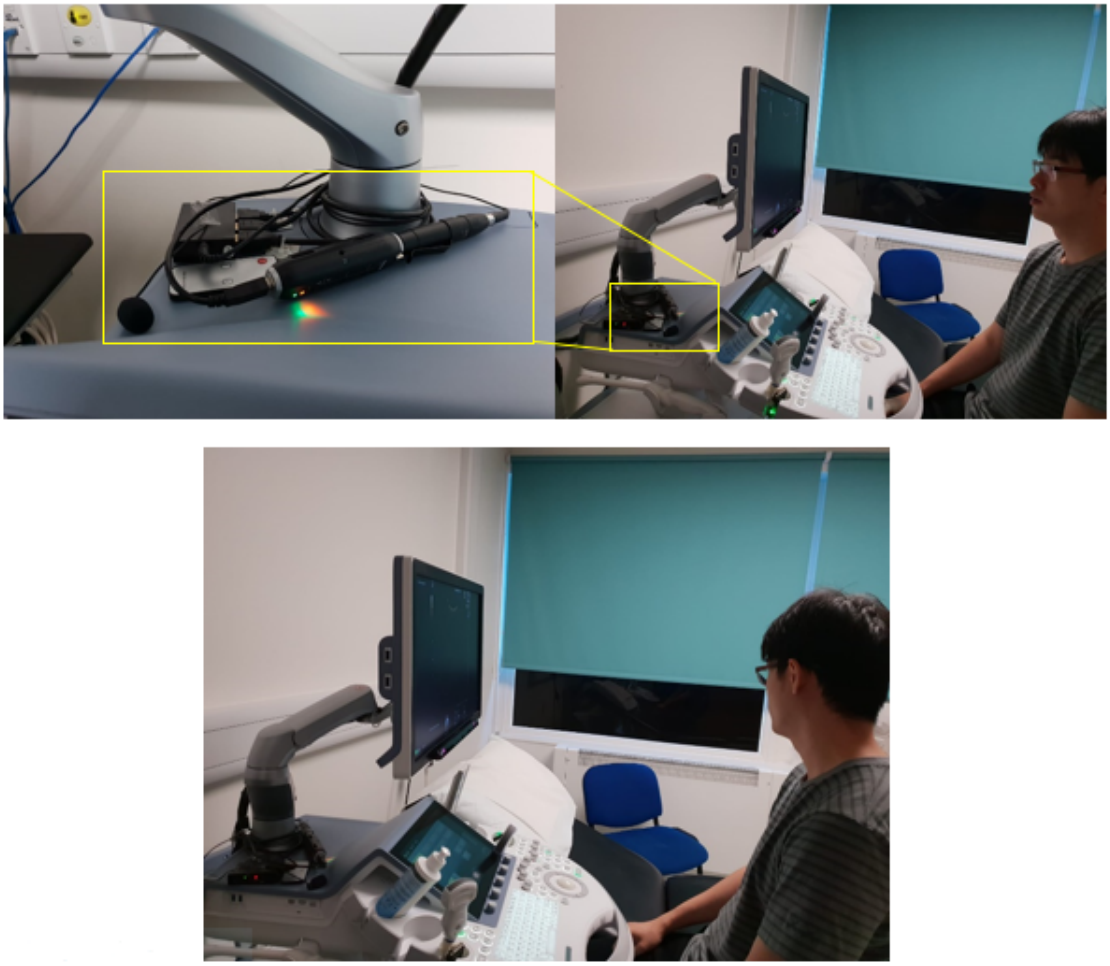


Figure 3.10: The old audio acquisition setup with the SHURE MX184 microphone is shown. In the lower image, when the sonographer is looking at the patient, they would not be speaking in the direction of where the microphone had been placed.

is not at an acceptable standard for audio recording. There was a noticeable 'graininess' in the recorded audio. In addition to managing to pick up non-desired audio, it needed to be recharged if it was to be used wirelessly. It is priced at £79.

Tonor Wireless Headset Microphone

The Tonor Wireless Headset Microphone was investigated as a potentially more comfortable wearable microphone than a lapel one, but it become evident when having sonographers try it that they would rather not have to deal with remembering to wear a microphone. The audio quality was good, but the microphone needed to be worn in order to attain good quality recordings. It picked up undesired audio

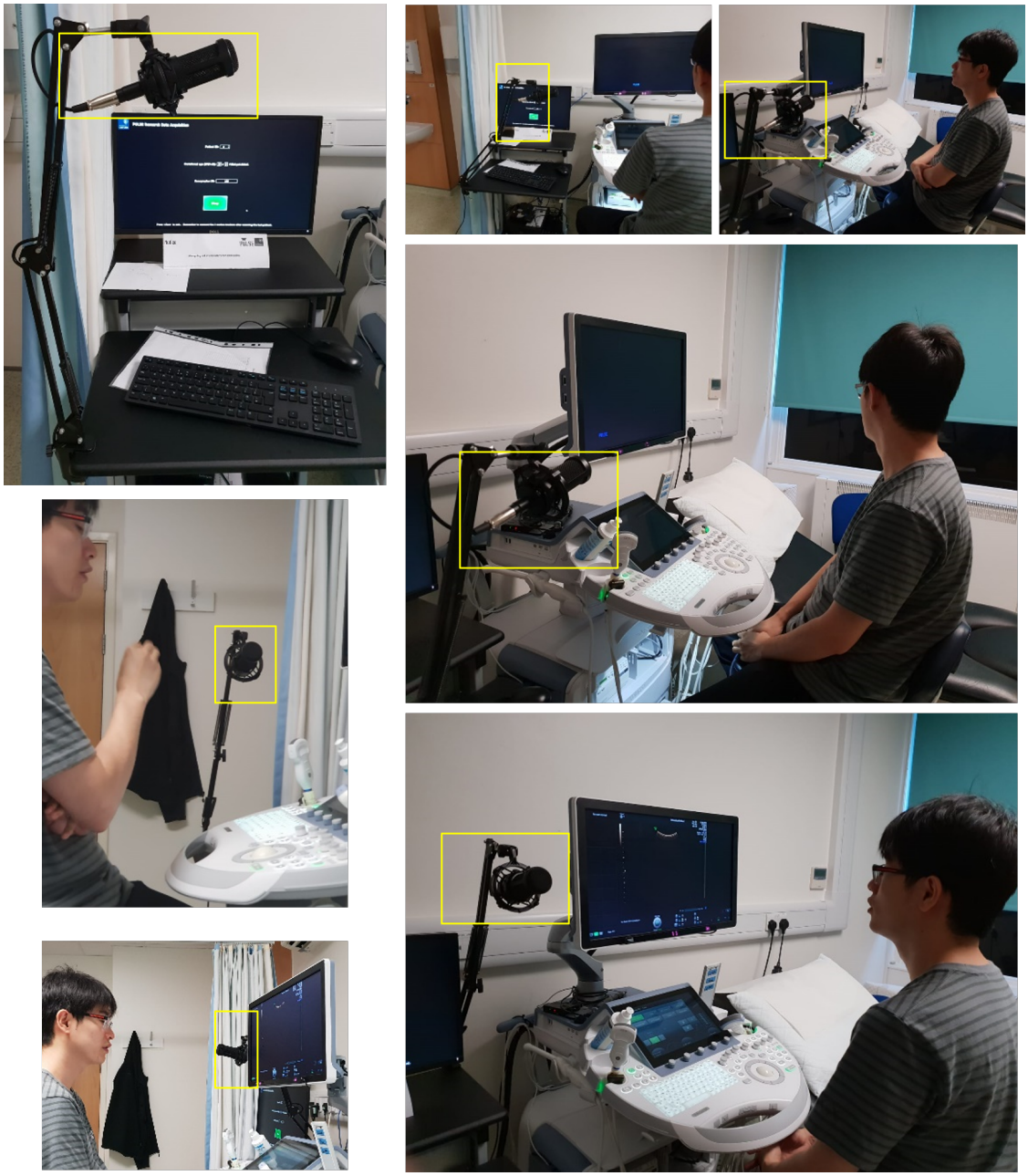


Figure 3.11: How the Audio Technica AT2035 microphone could be installed in the scan room is shown.

though, but it also required significant manual support because it would need to be be recharged daily. It is the cheapest of all the microphones investigated costing £20.

Portable Audio Recorder: Roland R26

With the portable audio recorder, the main problem is the fact that the transferring of data has to be done manually, so it was not the option for us going forward as

Figure 3.12: How the JABRA Conference Speakerphone could be installed in the scan room is shown.

integrating it with the rest of acquisition system would have been neither convenient nor straightforward. The data is stored locally on the audio recorder itself rather than on the local machine in the scan room. The audio quality is superb when compared to the other microphones; however, it would require significant manual support from the sonographers as they will need to turn it on and off manually themselves. As with the other microphones, non-desired audio continued to be picked up. The exact model we performed our tests with is now discontinued, but the price of similar products range from £167 to £350.

Boundary Microphone: PCC160

The PCC160's audio was quite good relative to some of the previously discussed options. It did not impede the sonographers in performing their duties, but it did have to be placed close enough to the sonographer, roughly within an arm's reach.

Figure 3.13: The two ways a Tonor Wireless Headset could potentially be worn by a sonographer during a scan are shown.

Although it did pick up some non-desired audio, the intelligibility of said audio with this microphone was significantly lower when compared to the previously described microphones. The PCC160 does not require any manual support whatsoever unless it has been moved from its location. It is priced at around £355.

Boundary Microphone: MXL AC-404

Like the PCC160, the MXL ACL-404 audio was quite good relative to some of the previously discussed options. We determined how good it was, following the same process discussed earlier; a 5 minute audio recording was made, and the easier it was to identify the spoken words in the recording, the better the audio quality of the microphone was considered to be. It did not impede the sonographers in the performing of their duties, but it did have to be placed close enough to a sonographer, roughly within an arm's reach. Although it did pick up some non-desired audio, the intelligibility of said audio with this microphone was significantly lower when compared to previous microphones. The MXL AC-404 does not require

any manual support whatsoever unless it has been moved from its location. It is also reasonably priced at approximately £100.

3.3.2 Chosen Microphone

Initially we used a SHURE MX 184 Lapel Microphone; however, sonographers found it cumbersome to put on. We spent around 3 months investigating alternatives. These investigations have been described throughout this section. In summary, we ended up selecting two PCC160 microphones which are boundary microphones meant to be placed on a surface. One was placed close to the the sonographer, and the other was placed close to the pregnant mother. The use of two rather than one microphone is motivated by wanting to perform source separation. This new microphone setup was more to the liking of the sonographers, since it did not require them to wear kit or turn it on when commencing a scan. Also, the audio was found to be of good quality.

In general, all three categories of microphones discussed in this section had good quality if they were used as they are intended to be used. The exception being speaker phones meant for conference calls rather than audio recordings. Portable audio recorders were slightly better when it came to picking up low whispering sounds than the others. Sonographers may at times speak quietly when they are deeply focused on the object on the screen in front of them. All the microphones do pick up undesired speech (i.e. the voice of the pregnant women and others). At times, it can be somewhat unintelligible as with the SHURE MX184, and at times it can be very clear as with the AT2035. The boundary microphones were the best compromise in terms of these conflicting requirements where the quality of the sonographer speech needs to be good while having non-sonographer speech dampened.

3.3.3 Audio Challenges

Low Whispering Voice of Speaker

One way to handle the low whispering voice a sonographer may have is to boost the microphone gain. That does help; however, it will also increase the signal

Microphone	Audio Quality	Sonographer Likeability	Non-Desired Audio	Manual Support	Polar Pattern	Price	Overall (R: Recommended/ NR: Not Recommended)
SHURE MX184 (Lapel)	✓	X	✓	X	Super-Cardioid	-	NR
Audio Technica AT2035	✓	X	X	✓	Cardioid	£133.5	NR
JABRA Conference Speakerphone	X	✓	X	X	Omni-Directional	£81.49	NR
Tonor Wireless Headset Microphone	✓	X	✓	X	-	£20	NR
Portable Audio Recorder: Roland R26	✓	X	X	X	3 patterns (2 Cardioid and 1 Omni -Directional)	£167 to £350	NR
Boundary Mic: PCC160	✓	✓	X	✓	Super-Cardioid	∼£355	R
Boundary Mic: MXL AC-404	✓	✓	X	✓	-	∼£100	R

Table 3.8: The microphones investigated and key characteristics of the different types and models of microphones are summarised.

amplitude of all sources of sounds in the scan room not only the sonographer's speech. We requested that sonographers try and speak loudly if possible while bringing the microphone closer to them.

Air Conditioner Noise

Air conditioning was a source of noise in the recording. The ideal solution would have been to turn the air conditioner off, and for scan sessions that were on cold winter days, that was the case. A surface above the microphone could have shielded it from air conditioner noise, but there were concerns with the surface interfering with the sonographer's work and there was no guarantee that the sound from the air conditioner would not reach the microphones, as the sound waves could have easily bounced off the walls and other objects in the small sized US scan room.

We also worked on ensuring that the boundary microphone was placed as close as possible to the speaker so that their voice would prevail over the noise coming from the air conditioner. Any remaining noise was removed computationally through the application of various noise filters.

3.3.4 Justification for Purchasing and Using Two Microphones

Recording the audio during a scanning session means that the non-desired speech of the subject and those that may be accompanying her will inevitably be captured along with the desired sonographer's speech. We have done tests to prove that only a few wearable microphones will not manage to pick up the non-desired speech in the small ultrasound scan room. The size of the room makes it quite likely that the sound waves will reflect off the walls and come into contact with the polar pattern of non-wearable microphones even when directed away from the source of the non-desired speech. For this reason, we have concluded that source separation is needed.

3.3.5 Source Separation

While blind source separation from one microphone is theoretically possible, source separation is made easier having multiple microphones in the room [100, 101]. Some

established source separation techniques that we have tried, ended up splitting audio recordings into human sounds and non-human sounds. This splitting does not solve our problem. A related concept to source separation, called speaker diarisation, is focused specifically on splitting an audio recording into that of multiple human speakers, answering the question: who spoke when? [102]. One tool we have found that works quite well in that regard is VoxSort [103]. In many cases, VoxSort can successfully split audio into speaker-specific segments given a recording from one single microphone. In order to ensure a higher success rate however, we introduced two microphones, one closer to the sonographer and the other closer to the subject. Audio signals from some of the recordings need to be normalised with Audacity [104] before using VoxSort on the recording.

3.3.6 Choice of USB Audio Interface

PCs are not built to allow multiple microphones to be used at the same time. For that reason, we also installed a USB audio interface. We have done our tests on the UR22mkII (approximately £136) which we then acquired.

3.3.7 Placement of Microphones and Audio Interface

During our tests, the microphone closer to the sonographer was placed on the desk of the PC to the left of the ultrasound machine within no more than five metres from the sonographer's mouth. Ideally, we might need to look into the possibility to clamp the microphone on the ultrasound machine itself in order to ensure that the placement does not hinder the sonographer and to ensure that the within-five-metres requirement is always satisfied, since some sonographers prefer to stand rather than sit during the scans, adding more distance between them and the microphone. The USB Audio Interface can fit neatly on top of the computer case and has been placed there since our early microphone comparisons.

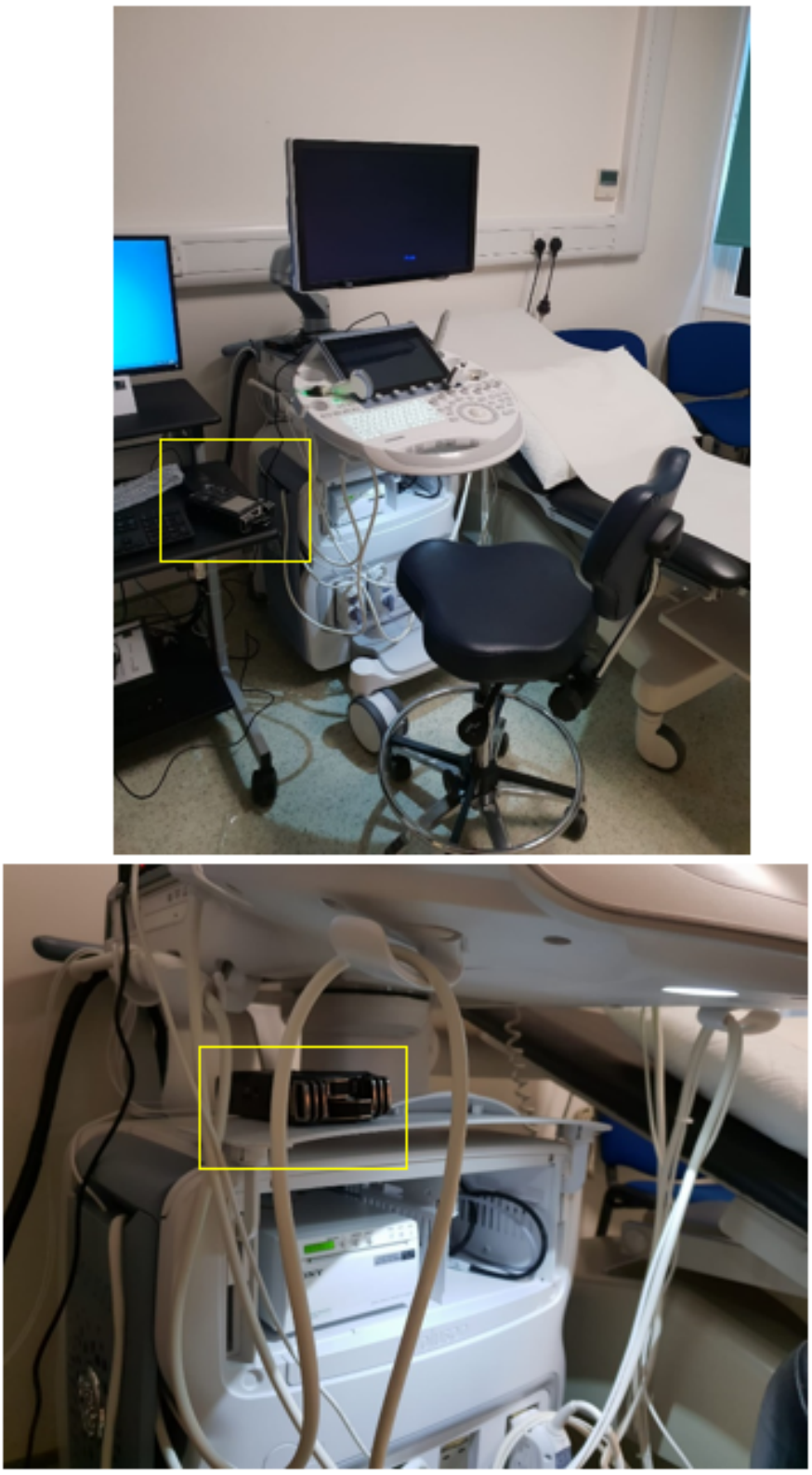


Figure 3.14: How a Portable Audio Recorder (such as Roland R26) could be installed in the scanroom.

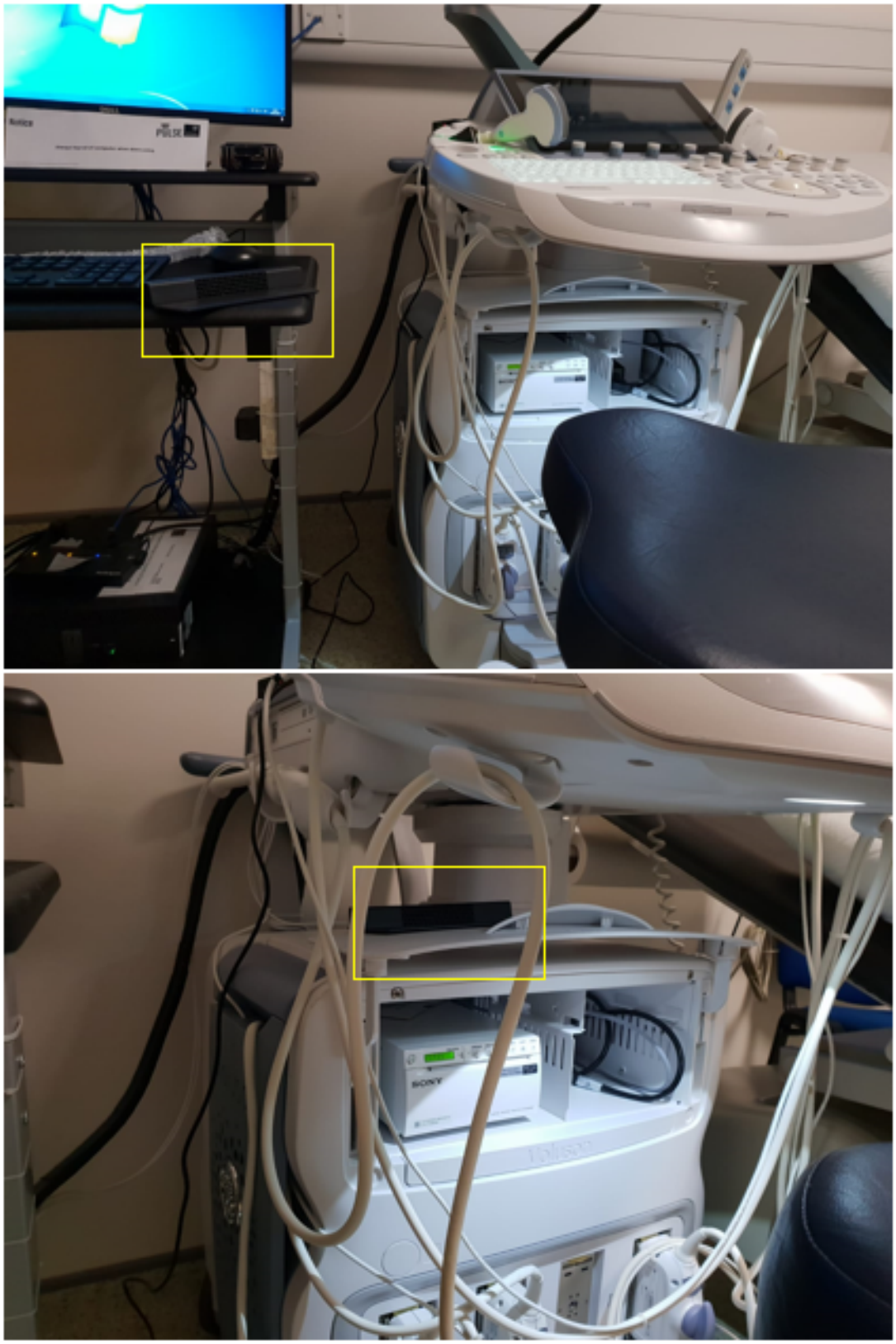


Figure 3.15: How a boundary microphone (such as the PCC160) could be installed in the scan room.

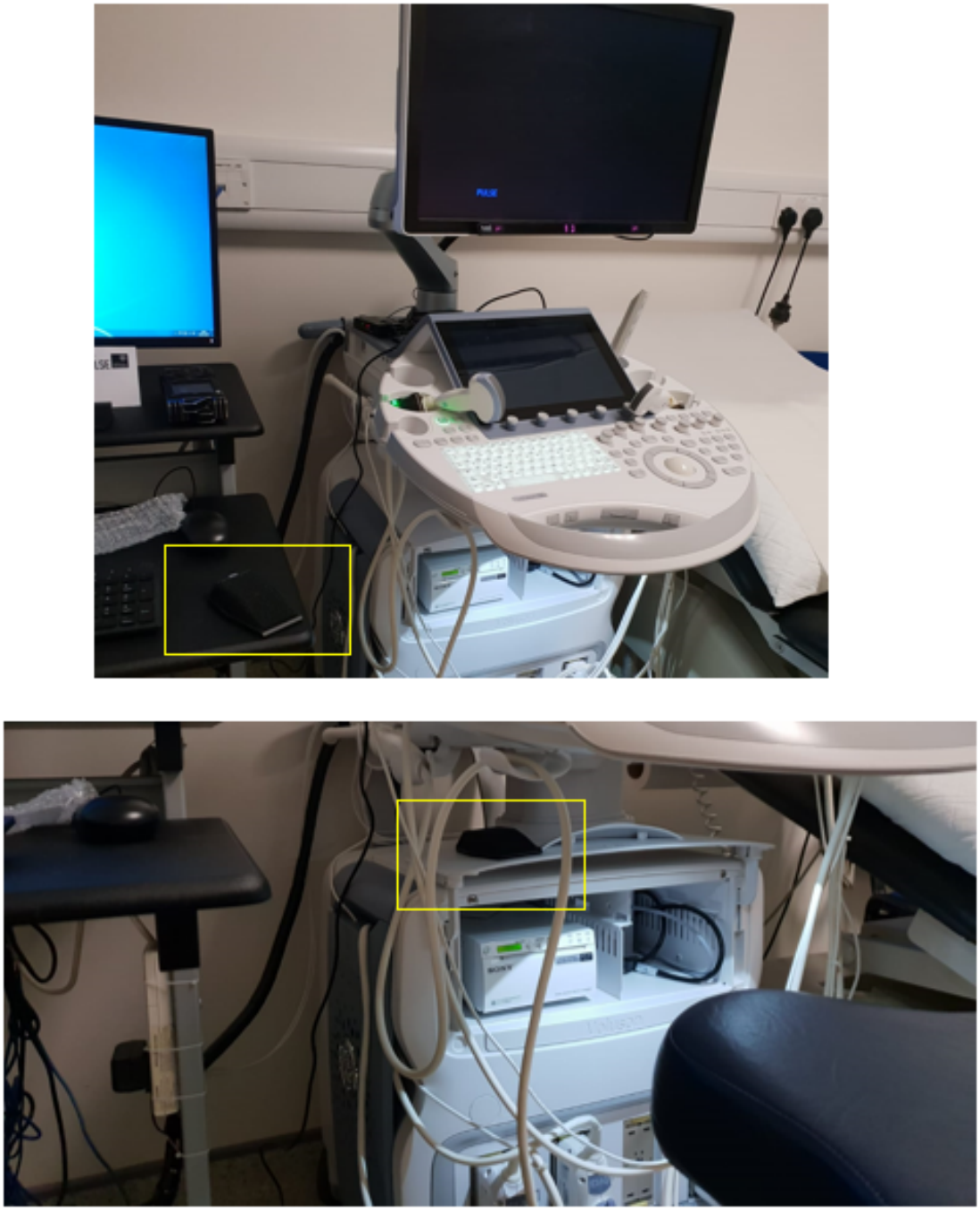


Figure 3.16: How a boundary microphone (such as the MXL AC-404) could be installed in the scan room.

3.4 Evaluation Metrics

Different model configurations were compared using the established general metrics *BLEU* (Bilingual Evaluation Understudy) and *ROUGE-L* (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence), a grammar score *GB* (GrammarBot) [105], a classification metric *Class. F1* (F1 score), and an anatomical description metric *ARS* (Anatomical Relevance Score). Objective metrics *BLEU* [25] and *ROUGE-L* [33] are calculated between the ground truth captions and the generated captions.

These two metrics (*BLEU* and *ROUGE-L*) are commonly used when evaluating image captioning models but may lead to lower values when the pair of captions do not show exact matches. $B1$ and RL look at the degree of overlap between the words of the ground truth and the words of the generated caption. $BLEU$ will look for overlap at the level of one word (*BLEU-1*), two words (*BLEU-2*), three words (*BLEU-3*), and four words (*BLEU-4*). However, both $BLEU$ and $ROUGE$ do not take into consideration whether the overlapping words hold any semantic value. Hence, for our caption generation task, grammar-based, classification-based, and subjective metrics were additionally explored. To evaluate captions grammatically, the average number of grammatical mistakes in a generated caption was calculated. Classification $F1$ scores were calculated by determining the caption class as the class of the vocabulary that has the highest overlap with the predicted caption.

We split the evaluation metrics into two categories: syntax-focused metrics and semantics-focused metrics. The syntax-focused metrics include *BLEU-1* (B1) [25], *ROUGE-L (RL)* [33] and $GrammarBot(GB)$ [105]. The semantics-focused metrics include $F1$ and ARS. GB simply measures the quality of the grammar in the generated captions.

3.4.1 Evaluation Metrics for Experiments Conducted in Chapter 5

To evaluate the model performance, we use the $F1$ score, *BLEU-1 (B1)* [25], and *ROUGE-L (RL)* [33]. *BLEU-1* and *ROUGE-L* have been slightly modified for

the task at hand as follows. Traditionally, both these metrics check that two captions are similar based on their constituent words with no regard to the actual semantic relevance of the generated words to those in the ground truth caption. In order to prevent relatively high scores being obtained when no relevant anatomical terminology is generated, we calculate *B1* and *RL* by modifying them according to the anatomical classification performance. Each image has a ground truth label (one of the four anatomical structures), hence, when evaluating a captioning model, the $B1$ and RL scores for the corpus (i.e the image-captioning test set of that structure) are multiplied by the percentage of images that are correctly classified by the image classification model of the same curriculum. If the percentage of correctly classified images is high (e.g. 90%) for that structure, the $B1$ and RL scores for that structure will only be slightly reduced. If the percentage of images that are misclassified is high, the $B1$ and RL scores will drop. This will penalise models for generating captions describing the wrong structure. We then take the mean of the scores for each structure.

3.4.2 Evaluation Metrics for Experiments Conducted in Chapter 6

The $F1$ score as described in this work uses scikit-learn's implementation of the $F1$ score were it is the weighted average of recall and precision. The ground target values represent the actual classes (anatomical structures) of the ground truth captions. The predicted target values represent the predicted classes (anatomical structure). To determine the anatomical structure of a caption, we had prepared a python script that that determines the structure given a caption. It returns the most likely anatomical class for that caption. We refer to this script as the structure determiner. The structure determiner works by finding which of the vocabularies of the anatomical structures has the most word overlap with the caption of interest. These vocabularies have been handcrafted and prepared with assistance from clinical experts. You can find the equation of the $F1$ score below.

$$F1 = 2 * \frac{recall * precision}{recall + precision} \tag{3.1}$$

The highest possible value is one, and the lowest possible value is zero. The Anatomical Relevance Score (ARS) is also reliant on the structure determiner. ARS is explained in more depth in Chapter 4.

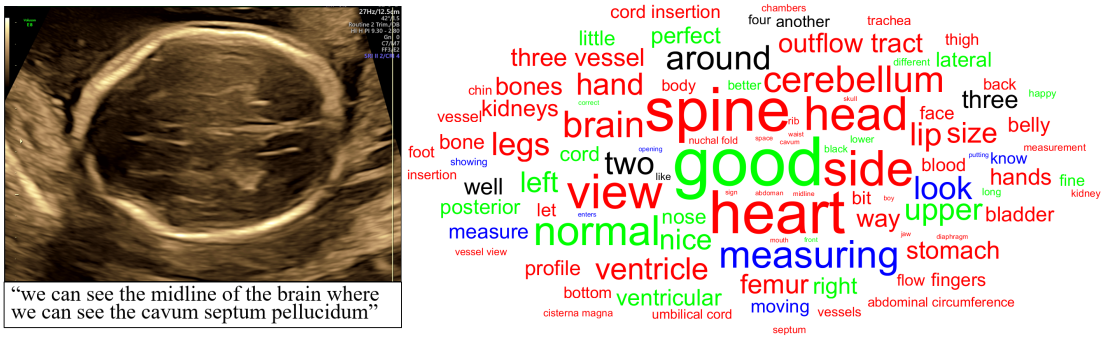
3.5 Conclusion

The thesis uses a number of datasets. A summary of these datasets were shown in Tables 3.2 and 3.1. Depending on the dataset, the data could include samples of combinations of image frames, text (captions), class labels, and gaze points. These datasets make it possible to fine-tune and train networks for the tasks of fetal ultrasound image classification and captioning.

The best microphone in terms of audio quality are either big in size or need to be worn by the sonographers. The best compromise for all the conflicting requirements was the boundary microphone with the caveat that it needs to be placed no more than one metre from the speaker.

Dealing with audio capture and transcribing it was an important early step before performing the experiments described in this thesis. Options for audio recording were explored and challenging requirements that needed to be met were handled. The word cloud in Fig. 3.17b shows the most common spoken sonographer words used to describe fetal ultrasound scans in our work. These commonly spoken words are textual outputs of the speech transcription performed on the retrospectively obtained recordings. The retrospectively recorded speech was transcribed with Google Cloud Speech [16]. The accuracy of the Google Cloud Speech transcription was evaluated through the word error rate calculated using a modified Levenshtein algorithm [90] focused on word-word comparisons rather than character-character comparisons.

Building the framework to transcribe the audio was done, making it possible to then build the fetal ultrasound image captioning models introduced in later chapters by training them on data samples such as the one shown in Fig. 3.17a.



(a) **(b)**

Figure 3.17: (a) Example of a fetal ultrasound image with sonographer description. (b) Word cloud of most frequently occurring words in sonographer vocabulary. Red, green, and blue represent nouns, adjectives, and verbs, respectively. The size of a word in the word cloud is proportional to its frequency of use.

Measure what is measurable, and make measurable what is not so.

— Galileo [106]

4

Building Captioning Models for the Fetal Ultrasound Context

Contents

4.1 Introduction 73
4.2 Image Captioning Model 75
4.2.1 Model Architecture 75
4.2.2 Results and Discussion 80
4.3 Summary 89

4.1 Introduction

Our initial vision for this piece of work differs from the contents of this chapter. At that time, we were interested in creating a correspondence mapping between speech-transcribed text and ultrasound visuals to generate descriptive captions, where ultrasound visuals refer to both ultrasound images and video clips. The main idea is that the latent space that represents this correspondence mapping could be used for two bidirectional tasks, generating captions given an image representation and retrieving a relevant image given a textual sequence. This text-video association given enough textual description of what is seen in an ultrasound scanning session

makes it possible to develop an automated way to describe ultrasound video with rich sonographic medical vocabulary as one potential downstream task.

To acquire rich textual descriptions of the ultrasound video, we asked sonographers to describe visual scenes (image contents) while performing an ultrasound scan or watching a previously recorded ultrasound scan video.

A successful captioning model would be able to adequately emulate the medical descriptions of professionals. A perfect emulation would not be sensible, as that would imply that the model may be overfitting to the speakers we have in our data. The captioning models need to learn the vocabulary of sonographers and use them to automatically generate medical descriptions of ultrasound video. In future work, we may want to investigate the latent space that makes this correspondence mapping possible and if it can be used for other tasks, such as image retrieval or video indexing.

In Fig. 4.1, we show a high-level abstraction of the analysis architecture that was considered at the beginning of the DPhil. It can provide generated captions and correspondence mapping as well as the similarity score given image and speech.

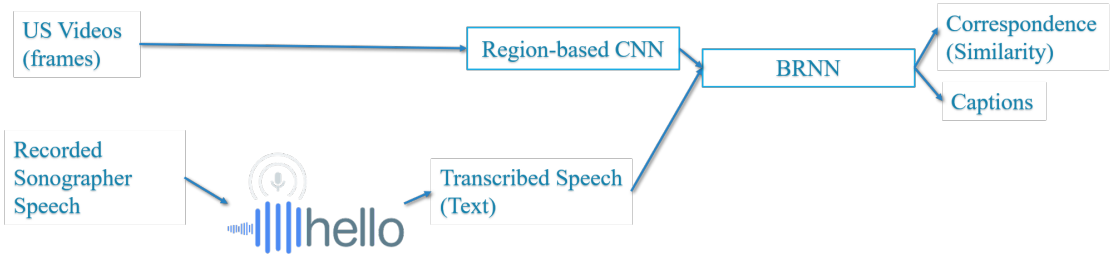


Figure 4.1: A high level abstraction of the system architecture initially envisioned at the beginning of the doctoral research.

The motivation for exploring this mapping in the future is the fact that if we do create this mapping between the images and the text as a basis for a system that can return one in the absence of the other, then we can generate descriptions if given images and index a video database if given a caption.

Using text as an input source to train a deep learning model to automatically caption ultrasound (US) images was new and did not exist when I first started my DPhil in 2017. The first couple of papers on this subject came out only in late 2017 [46] and 2018 [2]. Our first publication on this subject was in 2019 [4].

In addition to our objective of wanting to understand how sonographers describe ultrasound content, our work is set apart because of the nature of our data, which has driven the models we investigated and produced.

We use sonographers' spoken vocabulary to describe ultrasound videos with captions in an automated fashion. As described in Chapter 3, we asked sonographers to speak over ultrasound videos retrospectively to acquire the required rich descriptions of the contents of the ultrasound videos. Our goal was to generate rich captions from ultrasound frames as illustrated in Fig. 3.17b.

The generated captions are similar to the words spoken by a sonographer when describing the scan experience in terms of visual content and performed scanning actions. We train deep learning models consisting of convolutional neural networks and recurrent neural networks in merged configurations to generate captions for ultrasound video frames. We evaluate different model architectures using established general metrics (*BLEU*, *ROUGE-L*) and application-specific metrics.

4.2 Image Captioning Model

4.2.1 Model Architecture

The image captioning architecture is shown in Fig. 4.2. The reason why we have chosen this kind of architecture is as follows. It became apparent early on that we would be exploring with an architecture that combines a convolutional neural network (CNN) with a recurrent neural network (RNN). Initially, we had considered using a recurrent neural network in an insert configuration [38] as was done by Karpathy *et al.* [69]. However, the merits exhibited by a merge configuration [38, 107] where the RNN is not aware of the image information (and therefore require the RNN to learn less) was an important design choice we made early on especially since we are dealing with small sized datasets, and any architectural tweaks that can limit the debilitating effect of this challenge had to be taken. An RNN was used solely as the textual feature extractor and the encoded image information from a CNN was combined with the textual features in merged configurations [38, 107]. One branch of the model is a CNN based on the VGGNet16 [58] architecture,

Figure 4.2: The image captioning model (concatenation configuration) is shown. A VGG16 that has been fine-tuned on US images is used as a feature extractor for image information. In the text branch, an LSTM-RNN is provided with a sequence of previously generated words as input in the form of Word2vec embedding vectors. Its final hidden state is used as a representation of the sequential text information. The two vectors are concatenated before a prediction for the next word to generate in the sequence is made. At inference, that generated word is added to the partial caption (words generated so far). During training, teacher forcing is used whereby the next word to be added to the partial caption comes from the ground truth caption. In the experiments of this chapter, the vocabulary has a size of 117 unique words. The size of the vocabulary is directly related the size of the dataset, since with more captions, a greater variety of words could be encountered.

pre-trained on the ImageNet dataset and fine-tuned on fetal ultrasound standard planes of the abdomen, face, brain, femur, heart, spine, and placenta. The other branch represents the text encoding part of the model, including an embedding layer, which embeds the words in the sequence into a vector, followed by an RNN. Fig. 4.3 shows the nonsensical captions we would get for ultrasound images if we had used an image captioning model for natural image captioning on ultrasound images without any training, fine-tuning, hyperparameter tuning, or architectural modifications.

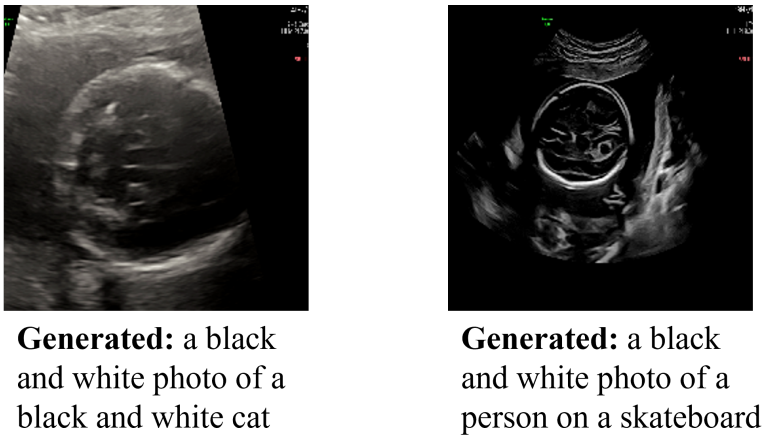


Figure 4.3: An image showing the poorly generated captions using natural image captioning models, showing why fetal image captioning need to be trained separately from natural image captioning models [30].

Features are extracted from the ultrasound video frames using the fine-tuned CNN. A textual caption is encoded by an embedding layer followed by a recurrent layer. The branches are merged, followed by a fully connected and decision-making layer. The model configurations generate the next word in a caption at every step as the probability distribution over the words in the vocabulary which, for this chapter, consists of 117 words, since the size of the vocabulary is dependent on the size of the textual dataset and the variety of captions that comprise it.

We comparatively evaluated different embeddings, namely, **word2vec** embedding trained on the Google News corpus [108], **GloVe** embedding trained on Wikipedia-2014 corpus [109], and a plain random initialization.

Word2vec is a shallow neural network trained to predict the context around a given word in a skip-gram model [110]. **GloVe** incorporates word co-occurrence

probabilities with the idea that words occurring together often enough are likely to hold underlying semantic meaning. The embedded word vectors are learnt by an RNN consisting of a Long Short-Term Memory (LSTM) unit [111] or a Gated Recurrent Unit (GRU) [112]. GRUs have less trainable parameters than LSTMs and require fewer operations, which makes GRUs more efficient to train, scaling down well to smaller datasets.

The two branches produce tensors of different lengths (200 and 300, respectively) that are joined together by merging. However, output vectors of equal length are used in addition merging configurations to perform element-wise addition. We compare two merge methods, concatenation and addition. In concatenation configurations, text and image feature vectors of unequal length are combined to deliberately force the model towards relying more on the text branch when generating the next word in a sequence to have a textually well-structured generated caption. In addition configurations, however, output vectors of an equal length of 300 are used.

Training Process

Dataset Cap. 1 was used for the experiments described in this chapter. Please refer to Chapter 3 for details on Dataset Cap. 1. Each image in the dataset was augmented twice; first by rotating by an angle between -30°and 30°, and second by horizontally flipping the image. Pre-trained VGGNet16 was first fine-tuned on ultrasound images. During training of image captioning models, 'teacher forcing' was applied where the ground truth sequences of increasing length were used at every step rather than the sequence of the words generated by the model at previous steps [113].

The model was called in a recursive fashion with the sequence of generated words so far being iteratively fed into the model at every time-step, along with the corresponding image. This process continued until the model generated a special end token or the maximum caption length was reached. Adam optimization [114] and categorical cross-entropy loss were applied during training. Early stopping was used to stop training when validation loss did not improve for more than five epochs. Early preliminary experimentation showed that for our data and architecture the

validation loss rarely improved after more than five epochs of continuous decrease. Dropout (rate between 0.4 and 0.5) was used to reduce overfitting. During inference, the model relied on its previously generated words to generate the next word.

Details of the evaluation metrics that were used can be found in Chapter 3, but we also devised a new measure called an Anatomical Relevance Score (ARS). The ARS is determined by matching words in a generated caption with the terminology of the anatomical class of interest. For example, an image of an abdomen may have a ground truth caption about the ribs, but if the generated caption describes the stomach, it is not an erroneous caption. ARS is calculated using Equations 4.1, 4.2 and 4.3

$$CS_k = \left(\sum_{i=1}^{L(W^c)} \mathbf{1}_{V_k}(w_i^c) \right)^{-1} \sum_{i=1}^{L(W^c)} \mathbf{1}_{V_k}(w_i^c) p_i \tag{4.1}$$

where CS_k is a score that a caption has in relation to the anatomical class k, K is the set of four anatomical classes, V_k is the vocabulary set of class k, L is the length of a caption W^c consisting of words w_i^c with softmax probabilities p_i, $\mathbf{1}_V(\cdot)$ is an indicator function which returns 1 if w_i is in V and 0 otherwise.

$$SS_c = \begin{cases} \max\limits_{k \in K} CS_k & \text{if} \arg\max\limits_{k \in K}(CS_k) = GT_c \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

where SS_c is a score that only considers CS_k if it has the ground truth anatomical class GT_c.

$$ARS = \frac{1}{C} \sum_{i=1}^{C} \tag{4.3}$$

where ARS is the Anatomical Relevance Score and C is the total number of captions in the set.

We provide a simple example to illustrate how to calculate the Anatomical Relevance Score. Assume a test set of size $C = 1$. The ground truth (GT) caption for the single test image is 'we are looking at the abdomen now'. In the GT caption, there is one word of anatomical relevance, "abdomen". The generated caption for the test image is 'this is the baby's belly'. In the generated caption,

the word 'belly' is a synonym of 'abdomen', and both are anatomically relevant to the abdomen as an anatomical structure. If a word is anatomically relevant to some structure, its synonyms are also anatomically relevant to the same structure. CS_{head}, CS_{heart}, and CS_{spine} are equal to zero as there are no words in the generated caption relevant to the head, heart, or spine. $CS_{abdomen}$ is dependent on $p_{stomach}$. Suppose $p_{stomach} = 0.7$, then according to Equation 4.1, $CS_{abdomen} = 0.7$ since there is only one word in the generated caption of anatomical relevance to the abdomen. By definition, $SS_c = \arg\max_{k \in K}(CS_k)$ in the case where $\arg\max_{k \in K}(CS_k)$ is also the anatomical class being described in the ground truth caption. Hence for this example, $SS_c = CS_{abdomen} = 0.7$. As $C = 1$, then $ARS = 0.7$.

4.2.2 Results and Discussion

Quantitative Evaluation

Table 4.1 presents a quantitative evaluation for different model configurations where the bold font indicates the best score for that metric. An overall score is obtained by calculating the mean of the scores (*GB* was normalised and inverted). The overall best performing model was the Fine-tunable-Word2vec-LSTM-Concatenation configuration, which is used to demonstrate an anatomy-focused evaluation in Table 4.2 and Fig. 4.4.

For a subjective measure, Likert Scale based evaluations are performed where a medical professional was asked to give a score of 0 ('No'), 1 ('Neutral'), or 2 ('Yes') in response to the following statements about a generated caption, namely it: (1) accurately describes the image; (2) has no incorrect information; (3) is grammatically correct; (4) is relevant for this image. For each caption, the responses were averaged. These scores are reported in Table 4.2 as *LSS* (Likert Scale Scores). Knowing the original image class and resulting caption classes, we plot the confusion matrix for the best performing configuration in Fig. 4.4.

Table 4.1: Evaluation results of model configurations are shown. The best model configuration is highlighted.

Word Embedding	RNN	Merge Mode	*BLEU-4*	*ROUGE-L*	*GB*↓	*Class. F1*	*ARS*	Overall
Fine-Tunable GloVe	LSTM	Concatenation	0.066	0.536	1.091	0.809	0.680	0.385
		Addition	0.081	0.580	0.9	0.948	0.686	0.397
	GRU	Concatenation	0.081	0.585	**0.889**	0.502	0.455	0.261
		Addition	0.094	0.561	0.923	0.529	0.449	0.268
Fine-Tunable Word2vec	LSTM	Concatenation	**0.105**	**0.594**	1.214	0.970	0.536	**0.427**
		Addition	0.045	0.546	0.929	0.679	0.506	0.297
	GRU	Concatenation	0.080	0.523	1.200	0.764	0.594	0.376
		Addition	0.086	0.539	1.077	0.609	0.476	0.307
Pretrained Word2vec	LSTM	Concatenation	0.085	0.574	1.200	0.921	0.567	0.413
		Addition	0.063	0.529	1.267	0.641	0.537	0.348
	GRU	Concatenation	0.066	0.530	1.100	0.768	**0.718**	0.385
		Addition	0.062	0.545	0.917	0.714	0.648	0.334
Random Initialisation	LSTM	Concatenation	0.075	0.560	1.222	**0.975**	0.564	0.422
		Addition	0.091	0.536	1.188	0.805	0.539	0.362
	GRU	Concatenation	0.067	0.507	1.308	0.763	0.632	0.394
		Addition	0.084	0.525	0.857	0.625	0.547	0.287

Discussion

Table 4.1 shows that there is no clear superior configuration across the different metrics, but overall, the Fine-tunable-Word2vec-LSTM-Concatenation configuration performs the best across the different metrics. Its generated captions are shown in Figs. 4.5, 4.6, 4.7, and 4.8. It is marginally outperformed in anatomical classification scores by the Random-Initialisation-LSTM-Concatenation configuration but scores higher in *BLEU-4* and *ROUGE-L*, implying the usefulness of pre-trained embeddings to ensure superior caption structuring compared to randomly initialised vectors. Word2vec embeddings were found to be more effective than GloVe embeddings for the fetal ultrasound datasets.

It is interesting to note that, in most cases, concatenation performed better than addition, and LSTM units outperformed GRUs, even for our limited datasets. It is to be expected that concatenation outperforms addition in our case, since the element-wise addition of feature vectors is not intuitive nor meaningful when two different modalities are involved. Addition would be a logical merging operation to perform if the feature vectors had come from the same data modality. However, since this is not the case, we explain the relatively worse performance of addition compared to concatenation to be due to the possible loss of information when merging feature vectors of different modalities through addition. Among the anatomical classes, from Table 4.2, abdomen and head show low scores in *BLEU-4* and *ROUGE-L* due to having the highest lexical diversity. Spine does well in *ROUGE-L* and *GB* because of its lower lexical diversity, however, *BLEU-4* is zero due to the absence of 4-gram overlaps but *BLEU-3*=0.319 is achieved. We attribute this to the fact that for spines the generated captions are often shorter than the ground truth. The generated captions do not learn and include the anatomically irrelevant stop words and filler words that are part of the ground truth (for example, 'ok that's good so yeah spine' because of the irregularity of their use. By irregularity of their use, we refer to the fact that the phrasing and choice of stop words and filler words by the sonographer is not necessarily dependent on observable characteristics of the visual content. This is different from what anatomies they mention which are

directly dependent on what is visible on the screen. However, not generating those stop words can penalize the models as has happened with the BLEU-4 score. Also, some spine-related vocabulary (such as 'cervical' that were part of captions in the test data did not appear in the captions of training data. The resulting *BLEU* and *ROUGE-L* scores can be explained by the fact that the generated captions do not use the exact same words that are in their ground truth equivalents. Hence, the models are penalized by BLEU and ROUGE-L even when generated captions correctly describe the image. At first glance, the GB scores appear high, but it is important to note that it penalizes the model for lack of capitalisations in the generated captions for names of some of the anatomical structures.

From *LSS*, we can see that the heart class is more challenging. In clinical practice, a fetal heart is typically identified by its beating motion (a video clip) rather than a still image. Further, the current captioning model is not trained to distinguish between the different heart views, but the textual description can be heart view specific. For example, the system generated a caption for the right ventricular outflow tract when it was given an image of a left one and this error is penalized in the current system. Adding more image-caption pairs of distinct heart views may solve this problem. Fig. 4.4 shows that all classes are accurately identified; however, the model struggles with 11% of abdomen images, misclassifying them as hearts. On investigation, it was found that for these specific images the stomach bubble has an elongated appearance, which has some resemblance to a heart view or heart chamber. Some qualitative examples are shown in Figs. 4.5, 4.6, 4.7, and 4.8. The metric scores for these qualitative examples are shown in Table 4.3.

Table 4.2: Evaluation results for the different anatomical structures are shown.

Structure	*BLEU-3*	*BLEU-4*	*ROUGE-L*	*GB↓*	*Class. F1*	*ARS*	*LSS*
Abdomen	0.000	0.000	0.533	0.667	0.886	0.316	0.625
Head	0.122	0.058	0.479	2.000	**1.000**	0.213	0.625
Heart	0.252	**0.140**	0.581	0.857	0.993	**0.843**	0.500
Spine	**0.319**	0.000	**0.789**	**0.000**	**1.000**	0.771	**1.000**

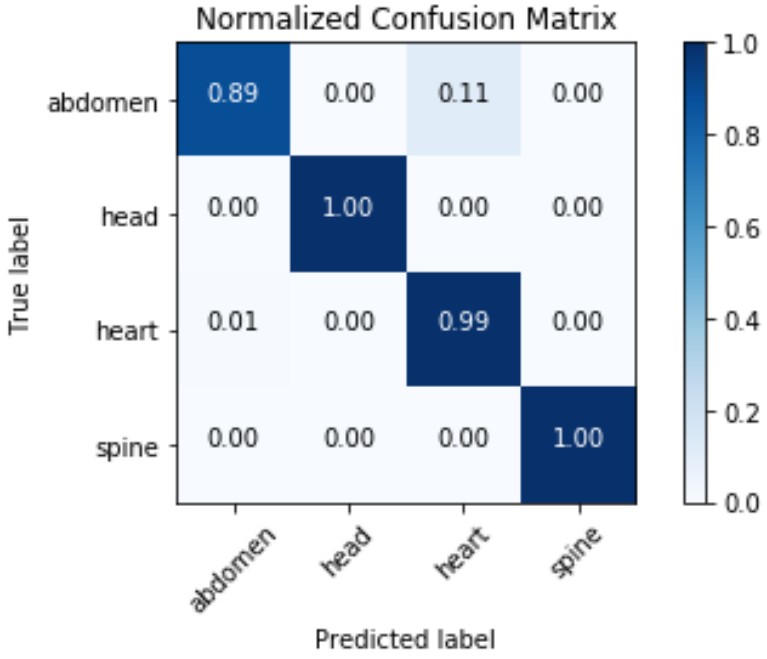
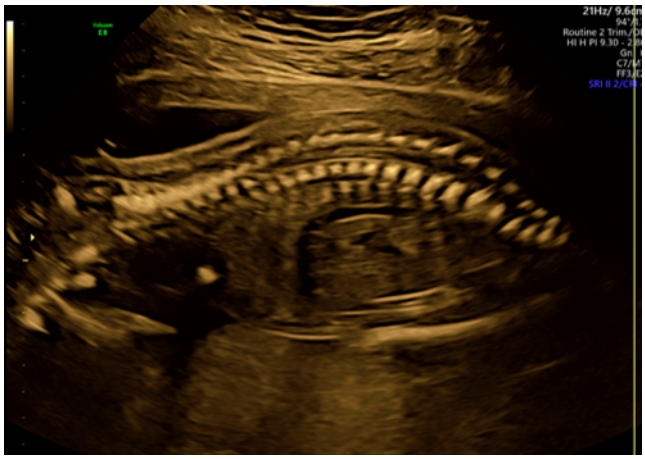


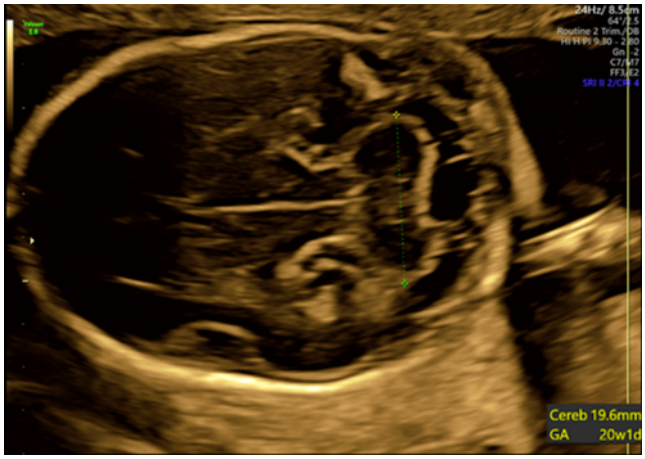
Figure 4.4: A confusion matrix of anatomical labels for the best performing configuration is shown. Based on this matrix, we can conclude that most generated captions describe the right anatomical structure depicted in the image; however, there are cases (11%) when heart-related captions are generated for images depicting abdomens.

Table 4.3: Individual metric results for the examples in Figs 1.10 to 1.13. The scores were only calculated between the ground truth and the generated caption individually for each single example rather than an overall score for the entire test set.

	B1	B2	B3	B4	R	F1	GB
Fig 1.10a	0.37	0.21	0.00	0.00	0.72	1.00	0
Fig 1.10b	0.50	0.27	0.00	0.00	0.67	1.00	0
Fig 1.11a	0.67	0.58	0.56	0.47	0.83	1.00	0
Fig 1.11b	0.50	0.24	0.00	0.00	0.56	1.00	0
Fig 1.12a	0.64	0.49	0.44	0.34	0.79	1.00	0
Fig 1.12b	0.43	0.27	0.00	0.00	0.56	1.00	0
Fig 1.13a	0.30	0.00	0.00	0.00	0.42	1.00	0
Fig 1.13b	0.29	0.00	0.00	0.00	0.47	1.00	0



(a) Ground truth: **'we can see the cervical spine'**
Generated: **'the spine'**



(b) Ground truth: **'measuring the cerebellum'**
Generated: **'the cerebellum has a good measurement'**

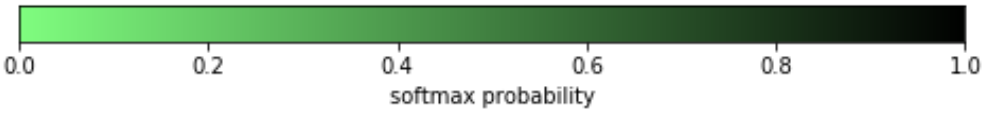
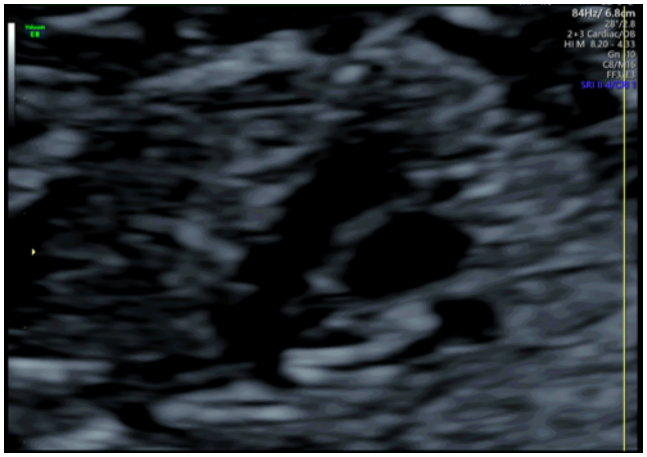
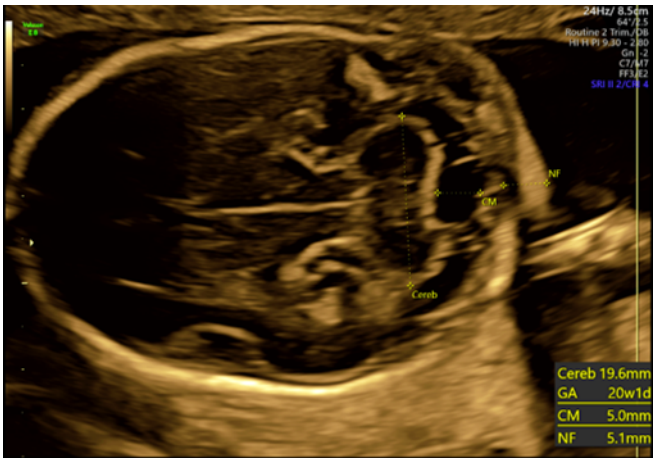


Figure 4.5: Ground truth and good generated captions for a couple of images in the test set. The higher the softmax probability associated with a generated word, the darker the green color of that generated word.



(a) Ground truth: **'three vessel trachea view'**
Generated: **'this is the three vessel trachea view'**



(b) Ground truth: **'the measurements of the nuchal fold are good'**
Generated: **'this is the cisterna magna and nuchal fold'**

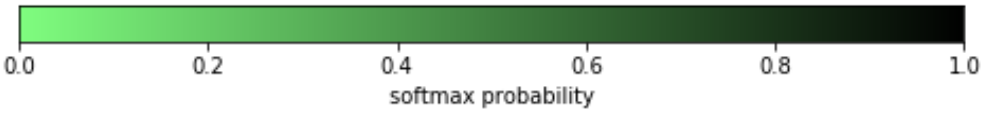
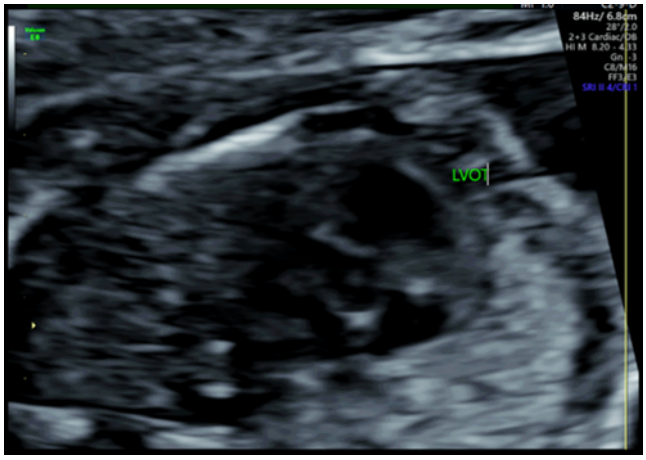


Figure 4.6: Ground truth and good generated captions for another couple of images in the test set. The higher the softmax probability associated with a generated word, the darker the green color of that generated word.



(a) Ground truth: **'this is the left ventricular outflow tract'**
Generated: **'the right ventricular outflow tract'**

(b) Ground truth: **'the aortic valve'**
Generated: **'the right ventricular outflow tract'**

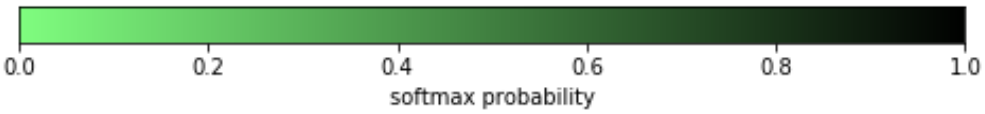
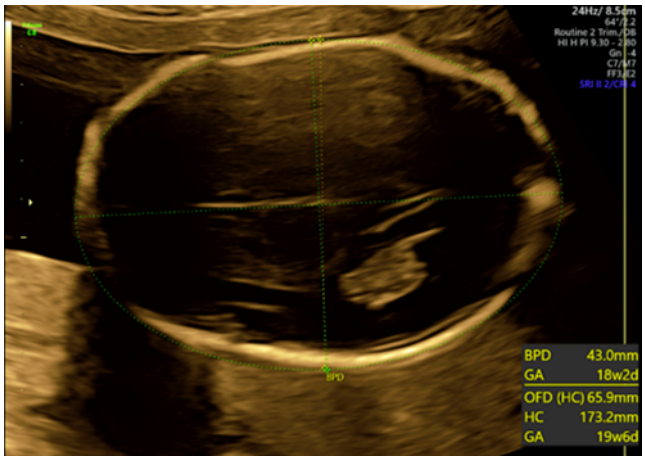


Figure 4.7: Examples where the ground truth and generated captions do not match exactly, but the latter may describe the image contents with relevant terminology. This mismatch is reflected in the low objective scores in the Results section. Also, the confusion between heart views is evident in Fig. 4.7a. Please note that the words 'aortic' and 'valve' in the ground truth caption of Fig. 4.7b are not in the training set. The higher the softmax probability associated with a generated word, the darker the green color of that generated word.



(a) Ground truth: **'now measuring the head side to side and around the head'**
Generated: **'this is the cisterna magna and nuchal fold'**

(b) Ground truth: **'rib'**
Generated: **'we can see the stomach'**

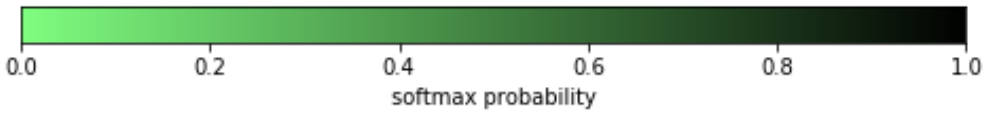


Figure 4.8: Additional examples where the ground truth and generated captions do not match. Note that the stomach is visible in Fig. 4.8b, but the sonographer happened to be talking about a rib in this instance. The higher the softmax probability associated with a generated word, the darker the green color of that generated word.

4.3 Summary

We describe an automatic natural language processing (NLP)-based image captioning method to describe fetal ultrasound video content by modelling the vocabulary commonly used by sonographers. The generated captions are similar to the words spoken by a sonographer when describing the scan experience in terms of visual content and performed scanning actions.

Using full-length second-trimester fetal ultrasound videos and text derived from accompanying expert voice-over audio recordings, we train deep learning models consisting of convolutional neural networks and recurrent neural networks in merged configurations to generate captions for ultrasound video frames. We evaluate different model architectures using established general metrics (*BLEU*, *ROUGE-L*) and application-specific metrics. Results show that the proposed models can learn joint representations of image and text to generate relevant and descriptive captions for anatomies, such as the spine, the abdomen, the heart, and the head, in clinical fetal ultrasound scans.

We proposed an automatic image captioning method to describe fetal ultrasound video content from four types of anatomical structures using real-world sonographer vocabularies. The Fine-tunable-Word2vec-LSTM-Concatenation performed best among the different evaluated model configurations. Richer vocabularies and extensions to spatio-temporal models will be considered in future work.

In the next chapter, we discuss and introduce important methods that can be used to improve the quality of the captioning results before the commencement of model training. These techniques have to do with: (1) either using statistics of the distribution of the training data in determining the ordering of data samples for training, and (2) using existing annotated data to pseudo-caption images that have no textual annotations.

Instrumental or mechanical science is the noblest and, above all others, the most useful.

— Leonardo da Vinci [106]

5

Improving the Performance of Captioning Models through Curriculum Learning and Pseudo-Caption Creation

Contents

5.1 Overview 92
5.2 Originality and Individual Role 94
5.3 The Dual Curriculum 95
5.3.1 Introduction 95
5.3.2 Model and Training Details 97
5.3.3 Results and Discussion 101
5.4 The Course-Focused Dual Curriculum 104
5.4.1 Introduction 105
5.4.2 Model and Training Details 105
5.4.3 Image Captioning Model Architecture 106
5.4.4 Experiments 107
5.4.5 Results and Discussion 107
5.5 Pseudo-Caption Preparation Pipeline 110
5.5.1 Introduction 110
5.5.2 Model and Training Details 112
5.5.3 Experiments 116
5.5.4 Results and Discussion 117
5.6 Summary 120

5.1 Overview

In Chapter 4, we introduced the main model that we perform the experiments in this chapter on. The theme of the previous chapter was overcoming challenges with data limitations by introducing an appropriate captioning model and framework that can generate captions despite inherent challenges. In the current chapter, the theme shifts to overcoming the challenge present in realistically sized medical datasets by considering the data, specifically the order with which the model interacts with data samples at training and the augmentation of said data.

This leads to the second contribution in the thesis which is **a novel curriculum learning based training algorithm for natural language processing (NLP) based fetal ultrasound image captioning**. Datasets containing medical images and corresponding textual descriptions are relatively rare and hence, smaller-sized when compared to the datasets of natural images and their captions. This fact inspired us to develop an approach to train a captioning model suitable for real world medical data. Our datasets consist of real-world ultrasound video along with synchronised and transcribed sonographer speech recordings. We propose a "dual-curriculum" method for deriving an ultrasound image captioning model. The "dual-curriculum" method prepares training batches in consideration of the two (dual) modalities involved in image captioning, text and images, by ranking data samples based on the inherent complexity associated with both the image information and the text information. The evaluation results (presented in section 5.3.3) show an improvement in all performance metrics when using curriculum learning over stochastic mini-batch training for the individual task of image classification as well as using a dual curriculum for image captioning.

The aim of the dual-curriculum method is to have the model train on 'easier' image-caption pairs in the earlier stages of training, leaving more complex data samples to later on. The method relies on building and learning from curricula of image and text information for the ultrasound image captioning problem. We compare several distance measures for creating the dual curriculum and observe

the best performance using the Wasserstein distance for image information and tf-idf metric for text information.

We also introduce a curriculum learning captioning method to caption fetal ultrasound images by training a model to dynamically transition between two different modalities (image and text) as training progresses. Specifically, **we propose a course-focused dual curriculum method**, where a course is training with a curriculum based on only one of the two modalities involved in image captioning. The evaluation results show that transitioning between the text complexity and the image complexity over epochs of training improves results when compared to the scenario where the complexities of both modalities are considered in equal measure in every epoch.

We compare two configurations of the course-focused dual curriculum; an image-first course-focused dual curriculum which prepares the early training batches primarily on the complexity of the image information before slowly introducing an order of batches for training with increasing focus based on the complexity of the text information, and a text-first course-focused dual curriculum which operates in reverse.

Finally, as another pre-training step, **we propose an approach to augment and increase the number of data available to train a fetal ultrasound image captioning model when in possession of only a small number of images that have corresponding textual descriptions by leveraging an existing larger dataset annotated for image classification-like tasks**. This allows a seamless automatic annotation process for the images that lack human-prepared captions to then use them in training models for the image captioning task.

The process first requires us to identify, for the target image, an image from the smaller captioning dataset that it is most similar to it according to cosine similarity. Along with the corresponding classification label that the target image already possesses, the nouns from the caption of that similar image are identified

and fed into an encoder-decoder sequence-to-sequence model to produce a pseudo-caption for the target image. These nouns represent the semantically meaningful content in the image.

5.2 Originality and Individual Role

Upon learning about curriculum learning, I began thinking how potentially useful curriculum learning could be in training image captioning models. Very early on, I was convinced that there needs to be a mechanism in the curriculum learning approach that considers both modalities involved in image captioning. Using the open source Python library frameworks, Keras (version 2.2.4) and Tensorflow (version 1.9.0), work was done on creating a curriculum learning approach specific to image captioning as well as later building the pseudo-caption creation pipeline for the third section of this chapter. The work resulted in two papers [5, 6]. I conceived the initial presented idea. Rasheed El-Bouri and I were involved in algorithmic design and coding. I was also responsible for integrating the different modules and finally training and evaluating the models as well as tabulating results. Harshita Sharma gave advice on performing an ablation study and provided me some relevant literature to review. Lior Drukker served as the primary expert annotator. Aris T. Papageorghiou and J. Alison Noble oversaw the work. This chapter has resulted in a third piece of work (described in section 5.5) on using pseudo-captions in training image captioning models. It was inspired by non-purely supervised NLP tasks. In that work, I conceived the presented idea and was responsible for algorithmic design, coding, training and evaluating models, as well as tabulating results. Harshita Sharma gave comments on the work conducted. Lior Drukker reviewed the medical content in the papers and served as primary expert annotator. Aris T. Papageorghiou and J. Alison Noble oversaw the work.

5.3 The Dual Curriculum

5.3.1 Introduction

In this section, we consider a training regimen well-suited and tailored for small-sized multi-modal medical data of images and text. The contribution of this section is the development of algorithms to prepare suitable curricula using curriculum learning techniques from a relatively small-sized training dataset and the demonstration of how this leads to improved deep learning-based image classification and captioning models for fetal ultrasound images. To the best of our knowledge, this is the first work to propose a curriculum learning technique to prepare curricula based on both medical image and text data concurrently for the task of medical image captioning.

Randomly shuffling the training dataset throughout the epochs of training is standard practice in deep learning. This randomness is meant to help regularise model training. The goal of regularisation is to prevent the values of model parameters from getting stuck in local minima during optimization.

In curriculum learning, the defining concept is the order in which data samples are introduced. Thus, curriculum learning also attempts to perform regularisation but in a more organised, less random way. Curriculum learning was introduced by Bengio et. al [115] to mimic the way that learning occurs in humans. Humans are trained with a curriculum that begins with easy concepts and progresses to harder, more 'complex' topics [116]. In the literature, a number of ways to define curriculum learning have been proposed [117]. In this section, curriculum learning is independent of any specifics of the model architecture or its hyperparameters that determine how it is trained. The only aspect under consideration is how the training batches are prepared and presented to the network. Curriculum learning techniques address the fact that optimising the parameters of a neural network is a non-convex problem. Therefore, a function's global minimum is not necessarily reachable, and the lowest possible error cannot be obtained during normal stochastic mini-batch gradient descent training [115, 117].

Related Works.

There have only been a few studies investigating the captioning of ultrasound images. A convolutional neural network (CNN) based captioning model for second trimester fetal ultrasound images that fuses text and image information for the next word generation was described in Chapter 4. In that method, random shuffling of training data was used without any optimization based on sample complexity. We improve on the findings in the previous chapter, using a curriculum learning approach for small-sized datasets; however, a direct comparison with that chapter is not possible due to the use of different datasets (details in Chapter 3).

Related work that explores image captioning has been reviewed in Chapter 2. In particular, we remind the reader of the image captioning framework to caption ultrasound images of the abdomen that is built in [65]. That framework begins with a classifier to identify the structure of interest in the given abdominal ultrasound image, before passing the ultrasound image to the encoding convolutional neural network (CNN) of the captioning model of that structure. In contrast, in this thesis, the classifier is effectively also responsible for encoding the image information for the captioning model.

Examples of using curriculum learning exist in computer vision [80] and natural language processing [81–83] as well as a few in biomedical imaging applications, such as classification of lesions in chest X-rays [84] to address weakly labeled data and detection of cardiac MR motion artefacts [85] to deal with class imbalance. In this chapter, we consider techniques to build curricula to aid in the captioning of second trimester fetal ultrasound images. To the best of our knowledge, this chapter (as well as its associated published papers) is the first work towards optimizing results in medical image captioning, a multi-modal task, for small-sized real-world datasets, by proposing the training of the deep neural network in a structured manner through a dual curriculum.

5.3.2 Model and Training Details

Curriculum Learning and Distance Measures.

To understand curriculum learning, we first define a metric, H, that encodes some notion of complexity (examples of H are defined in the paragraphs below). We then split the training dataset into batches according to this complexity such that $H[B_i] < H[B_j]$, where B is a batch of data from our training set, i is an index of an earlier batch, and j is an index of a later batch. These batches are ordered for presentation to the network and training according to the choice of curriculum. There is no current consensus on what a general curriculum looks like as a curriculum is usually tailored to the problem at hand. For example, for some problems, models achieve better performance when trained with 'anti-curricula' (presentation of data from high 'complexity' or entropy to low) [115, 118]. In this section, we experiment with various configurations and curriculum strategies.

It has been hypothesized that curriculum learning performs a similar function to numerical continuation methods [115], whereby a complex surface is approximated by a smooth version before increasing the complexity of that surface to become more similar to the original [119]. Entropy is a measure of randomness and uncertainty. Data samples with low entropy with respect to the mean of the samples are 'easier' to learn from. In terms of the curriculum, by optimising with a low entropy, 'easy' to learn batch, we provide a smoother version of the prediction error surface. By training sequentially with data in higher 'complexity' batches, we progressively increase the complexity of the surface while already having some parameters within the domain of a minimum which takes the solution closer to the global minimum, thus giving model weights that can provide better results. In this section, when building an image-based curriculum, we experiment with the Mahalanobis distance [120], the cosine similarity and the Wasserstein distance as H metrics. We perform experiments using a forward curriculum (low H to high H). The data samples for image focused tasks are image feature vectors extracted through a finetuned VGG16 [58].

The **Mahalanobis distance** d_{ml} measures how far, or how different, an image feature vector $\boldsymbol{x_n}$ is from the mean of the samples $\boldsymbol{u}$. It is the multi-dimensional generalisation of standard deviation from the mean and given in Eqn. 5.1,

$$d_{ml} = ((\boldsymbol{x_n} - \boldsymbol{u})^T \mathbf{S}^{-1} (\boldsymbol{x_n} - \boldsymbol{u}))^{(1/2)} \tag{5.1}$$

where $\mathbf{S}$ is the covariance matrix.

The **cosine similarity** d_{cos} is defined in Eqn. 5.2. We use a fixed reference (mean) vector to compare the dot products of model inputs with.

$$d_{cos} = \boldsymbol{x_n}\boldsymbol{u}/|\boldsymbol{x_n}||\boldsymbol{u}| \tag{5.2}$$

The **Wasserstein distance** d_{ws} is defined as the distance between two distributions and is mathematically defined in Eqn. 5.3,

$$d_{ws} = \inf_{\pi \in \Gamma(\hat{\boldsymbol{x}}_{\boldsymbol{n}}, \hat{\boldsymbol{u}})} \int_{\mathbb{R}\times\mathbb{R}} |m - n| d\pi(m, n) \tag{5.3}$$

where Γ denotes the collection of all measures on a space, $\mathbb{R} \times \mathbb{R}$, over all joint distributions of the random variables m and n with marginals $\hat{\boldsymbol{x}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{u}}$ respectively. In this case, $\hat{\boldsymbol{x}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{u}}$ are the softmaxed distributions of our features $\boldsymbol{x_n}$ and the mean of the features $\boldsymbol{u}$ respectively. The use of the Wasserstein distance arose to determine the optimal transport of piles of sand, represented by the distributions $\hat{\boldsymbol{x}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{u}}$ [121]. To use a Wasserstein distance-based curriculum, we first apply a softmax function on the image feature vectors, representing them as distributions. Then, the distance between the distribution of every image vector and the distribution of the mean feature vector is computed. The SciPy python library version 1.2 [122] has a number of readily-available functions that can compute distance measures. We have used their `wasserstein_distance()` function to calculate the Wasserstein distance between two distributions.

For preparing text-focused curricula, we used term frequency–inverse document frequency (tf-idf) score [123], as calculated by Eqn. 5.4:

$$d_{ti} = \frac{\#(w \in c)}{\Sigma_{w_i}^{W} \#(w_i \in c)} \ln \frac{\#(c \in C)}{\#(c\ if\ w \in c)} \tag{5.4}$$

where d_{ti} is a word's tf-idf score, $\#()$ is a count function, w is a word, c is a caption, W are all the words in caption c, and C are all the captions in the dataset. It provides a measure of the meaningfulness that a word possesses. For each caption, the tf-idf scores of its constituent words are summed, allowing us to incorporate the length of a caption into its complexity.

The dual curriculum, as used in the captioning task, uses two metrics of complexity, one for each data modality, namely, image H_1 and text H_2. Each metric ranks the data samples s in the dataset S differently. With dual curriculum, we combine these rankings. The scales used by the different metrics are not the same, so it was necessary to first normalise each metric before obtaining a combined measure of 'difficulty'. This allows us to take into account the difficulty associated with both images and text as shown in Eqn. 5.5:

$$d_{dc} = \frac{H_{1s} - min(H_1)}{max(H_1) - min(H_1)} + \frac{H_{2s} - min(H_2)}{max(H_2) - min(H_2)} \tag{5.5}$$

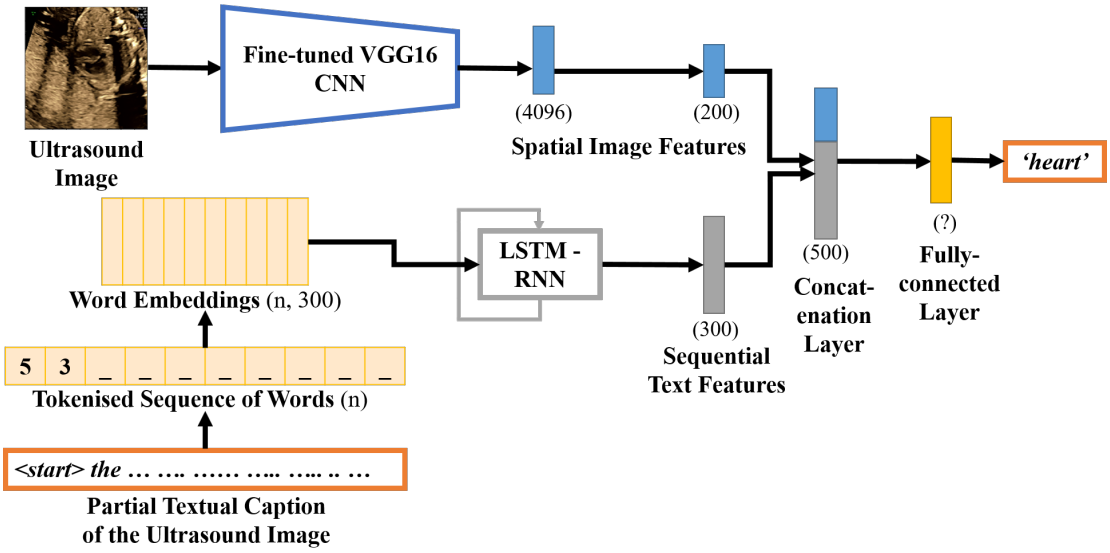


Figure 5.1: The image captioning model is shown. The input ultrasound image is that of a heart. 'n' represents the maximum sequence length. A VGG16 that has been fine-tuned on US images is used as a feature extractor for image information. In the lower branch, an LSTM-RNN is provided with a sequence of previously generated words as input in the form of Word2vec embedding vectors. Its final hidden state is used as a representation of the sequential text information. The two vectors are concatenated before a prediction for the next word to generate in the sequence is made.

Image Classification Model. Before considering the image captioning task, we first consider which distance metric works best for the simpler task of fetal ultrasound image classification. This image classification model is later used as the image information branch of the image captioning model shown in Fig. 5.1. Specifically, the problem is to classify an ultrasound image given four classes; the abdomen, the head, the heart, and the spine. We consider a pre-trained VGG16 [58] that has been fine-tuned on fetal ultrasound images. The images and labels used in fine-tuning the VGG16 are obtained from video clips in the US scan videos that were manually labelled based on the viewed anatomies [92]. The VGG16 architecture provides image feature vectors that can serve as input for a shallow 3-layer image classifier. We compare four variants of the model for evaluation of distance measures trained using: 1) mini-batches with stochastic gradient descent and random sampling, 2) a Mahalanobis distance based image curriculum (MD-IC), 3) a cosine similarity based image curriculum (CD-IC), and 4) a Wasserstein distance based image curriculum (WD-IC).

Image Captioning Model.

Fig. 5.1 shows the image captioning model in the process of generating a caption, one word at a time, for a heart image. The image captioning model consists of two separate branches for image and text modalities, combined in a late merge configuration using a concatenation layer followed by a fully-connected classification layer. A late merge configuration is used where the image information and text information (shown in the upper and lower branches in Fig. 5.1) are only combined towards the later (top) layer of the model. In the lower branch, an LSTM-RNN is provided with a sequence of previously generated words as input in the form of Word2vec embedding vectors. Its final hidden state is used as a representation of the sequential text information. Vectors from the two branches are concatenated before a prediction for the next word to generate in the sequence is made. The upper branch followed by a fully-connected layer with softmax activation represents the image feature classifier model. '?' represents a varying number of words depending

on the vocabulary of the anatomical structure that the specific captioning model would correspond to. For the image captioning task, we explored three different dual-curriculum options (MD-DC, CD-DC, WD-DC) that combine metrics discussed in the previous section with the tf-idf score. Four captioning models were trained, one for each anatomical structure. During training the partial caption comes from the ground truth caption, whereas at inference, the model would rely on its previously generated words. At inference, the US image is first classified using the trained US image classifier. Once the associated anatomical structure of the image is determined, the corresponding captioning model is deployed to caption the image. Posing the captioning problem this way places a high importance on first correctly identifying the structure.

Training Process.

In all experiments, the loss function was categorical cross entropy, and Adam optimisation was used. When training, the number of batches was 32 for the classification task and eight for the captioning task. In the captioning tasks, the words were embedded with fine-tunable Word2vec embedding vectors that were pre-trained on the Google News corpus [124]. Models were trained for 150 epochs with early stopping. During training, teacher forcing was used where partial captions consisted of the ground truth words rather than the words the model generated at previous steps, but at inference, the captioning model had to rely on its previously generated words to generate the next word in the caption. Dropout was applied throughout with a rate between 0.4 and 0.5.

Experiments.

The dataset used for the image classification task Dataset Class. 1 as described in Chapter 3. The dataset used for the image captioning task was Dataset Cap. 1.

5.3.3 Results and Discussion

The results for the image classification task are shown in Table 5.1 (P: Precision, R: Recall). We observe that the WD-IC trained model outperforms the traditionally

Table 5.1: Image Classification

Metric	P	R	$F1$
No Curriculum	0.83	0.76	0.73
CD-IC	0.63	0.73	0.67
MD-IC	0.94	0.95	0.95
WD-IC	**0.96**	**0.96**	**0.96**

Table 5.2: Image Captioning

Metric	$B1$	RL	ARS	$F1$
No Curriculum	0.18	0.37	0.23	0.77
CD-DC	0.22	0.34	0.22	0.73
MD-DC	0.25	**0.47**	0.40	0.96
WD-DC	**0.27**	0.42	**0.43**	**0.97**

trained (no curriculum) model and the CD-IC trained model. The latter two obtained relatively lower scores. The MD-IC trained model performed well and was only slightly inferior in terms of accuracy to the WD-IC model. For the classification task, the WD-IC and MD-IC models performed comparably well, and we also observed the same behavior for captioning.

Fig. 5.2 shows the distribution of structures in ten batches prepared by an MD-IC, a CD-IC, and a WD-IC. For visualisation purposes, we assume that the data is split into ten batches. A higher batch number means 'harder' data samples in the batch. From this visualisation, we can observe class imbalance in the batches. Preliminary experiments in image classification showed that the curricula struggled when batches had a disproportionate number of samples of the different structures as in Fig. 5.2. Future work will investigate this observation. To alleviate this problem, for the rest of the experiments reported in this section, batch preparation was done through the curricula as follows. Rather than having a single mean of the whole dataset from which a distance measure is calculated for each data sample, each structure had its own mean. In each of these balanced batches, for each structure, we had data samples with the lowest entropy relative to that structure's mean instead of the data samples with the lowest entropy relative to the mean of the whole dataset. Early experimentation also suggests that using the image curriculum leads to comparable results to a model fully trained with stochastic gradient descent within one epoch of training, implying potential faster convergence as Fig. B.1 in the appendix shows. This experiment used the same dataset but also included femur and placenta images.

Qualitative results of the image captioning task are shown in Fig. 5.3. Quantitative results are summarized in Table 5.2. *B1 (BLEU-1)*, *RL (ROUGE-L)*,

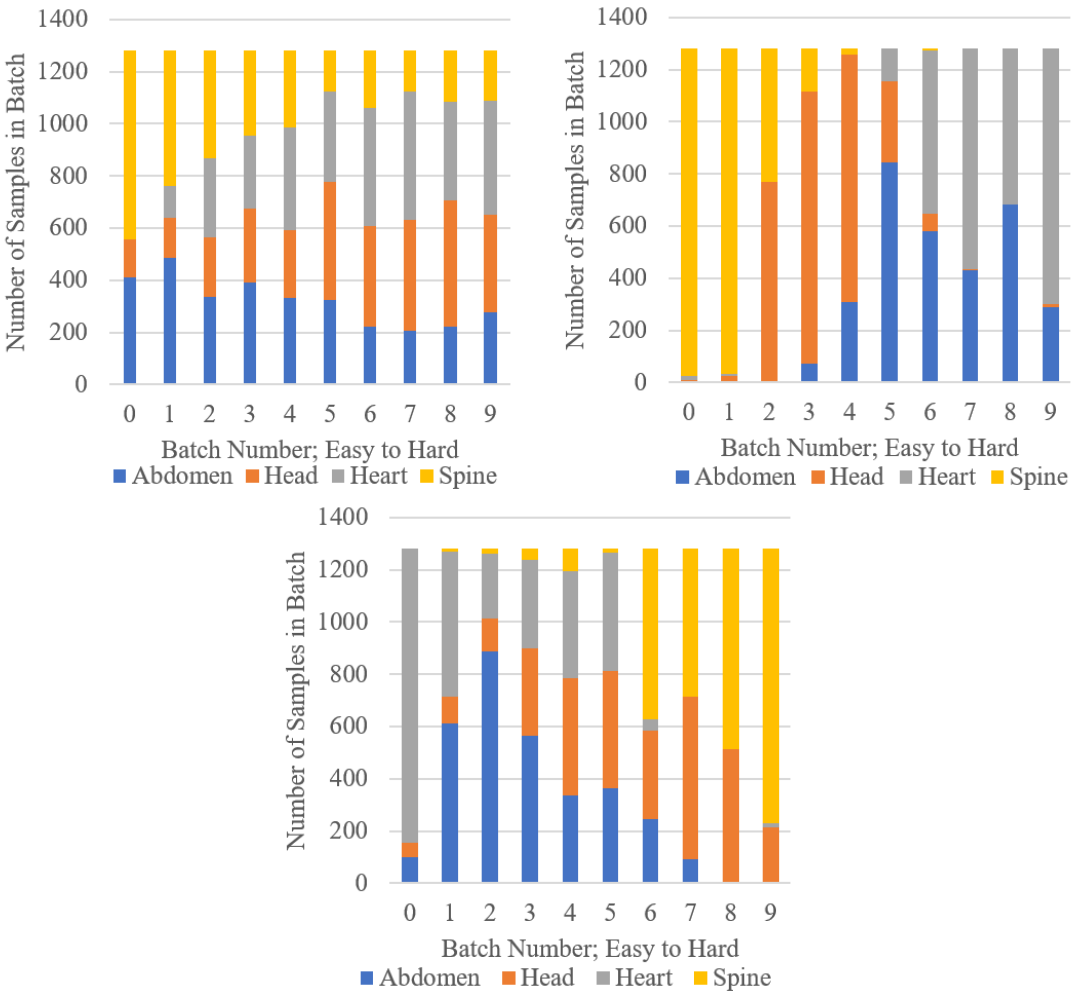
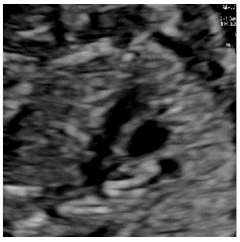


Figure 5.2: The distribution of anatomical structures in an MD-IC (upper-left), a CD-IC (upper-right), and a WD-IC (bottom-middle) with non-balanced batches (with respect to anatomy class) is shown. CD-IC is more likely than others to have batches contain samples of the same anatomical class.

GT: 'belly'
NC: 'good heart'
WD-DC: 'this is baby's belly the black space the black circle is the baby's stomach'



GT: 'the three vessel view'
NC: 'is the the ventricular outflow tract'
WD-DC:'three vessel view'

Figure 5.3: Qualitative results for an abdomen image and a heart image are shown. GT is for Ground Truth as spoken by a sonographer. NC is for a model that was trained with 'No Curriculum'. WD-DC is for a model that was trained with a Wasserstein Distance-Dual Curriculum.

F1, and *ARS* are evaluation metrics that have been discussed in Section 3.4 and Subsection 4.2.1. We notice an improvement when compared to the traditionally trained model (NC). WD-DC is superior to other curricula for three out of four metrics. The reason *RL* is lower for WD-DC than for MD-DC is discussed as follows. *RL* looks for the longest common subsequence (series of words) between the ground truth and generated captions, whereas $B1$ looks for single words in common. There might be a certain multiple-word-long way to phrase captions that exists in the training set that the model trained with MD-DC seems to be better at emulating; however, not phrasing the captions in this way is not indicative of loss of relevant information because otherwise we would have noticed *ARS* and $F1$ to be lower for WD-DC than for MD-DC.

Despite the empirical evidence otherwise, forgetting could theoretically be a potential problem when using the same order of data samples every epoch. Nonetheless, we alleviate this concern by using the course-focused dual curricula introduced in Section 5.4 where there still is a strategy to be followed in the ordering of data samples but one that leads to a different order each epoch of training.

5.4 The Course-Focused Dual Curriculum

In the experiments of Chapter 4, random shuffling of training data was used without any optimization based on sample complexity. In Chapter 5, we improve on the findings of Chapter 4 using a curriculum learning approach for small-sized datasets; however, a direct comparison with that paper is not possible due to the use of different datasets.

Earlier in this chapter, we have introduced a curriculum learning method that can be used in training fetal ultrasound image captioning models. In this section of Chapter 5, we take this further, by having unequal influence on the two modality-specific metrics that together constitute the previously introduced dual curriculum. Each metric represents the inherent complexity of one of the two modalities present in image captioning.

5.4.1 Introduction

In this section, we explore a natural extension of using a linear combination of the complexity metrics of a single multi-modal data sample. This means that rather than assuming that both metrics contribute equally to the arrangement and ordering of batches in every epoch, one of the complexity metrics is more influential than the other in a given epoch.

The modalities are clearly different and complementary, and one may be more challenging (e.g. data samples deviate greatly from the mean) to learn by the network than the other. The aim of this section is to quantitatively determine the best weighting combination and explore the advantages, or otherwise, of doing this. It answers the question: Would unequal weighting be beneficial?

Earlier in Chapter 5, we considered techniques to build curricula to aid in the captioning of second trimester fetal ultrasound images. The previous section discussed improving results in medical image captioning, a multi-modal task, by proposing the training of the deep neural network in an organised manner through a dual curriculum.

5.4.2 Model and Training Details

This section investigates unequal weighting of the two 'courses' (image and text) in the curriculum. Specifically, we explore two variants of the course-focused dual curriculum (CF-DC); an image-first course-focused dual curriculum (I1-CF-DC) and a text-first course-focused dual curriculum (T1-CF-DC). The equation for the CF-DC is shown in Eqn. 5.6:

$$d_{cf} = w_1 \frac{H_{1s} - min(H_1)}{max(H_1) - min(H_1)} + w_2 \frac{H_{2s} - min(H_2)}{max(H_2) - min(H_2)} \tag{5.6}$$

where w_2 and w_1 linearly interchange ($w_2 = 1 - w_1$). In I1-CF-DC, $w_1 = 1.0$ and decreases in linear decrements to 0.0. With T1-CF-DC, $w_1 = 0.0$ and increases in linear increments to 1.0.

Unlike the vanilla dual curriculum, a course-focused dual curriculum is a hierarchical curriculum that prepares and orders different curricula for model

training first based on the weighting associated with a modality (w_1 and w_2) and then based on the complexity of the data samples according to H_1 and H_2.

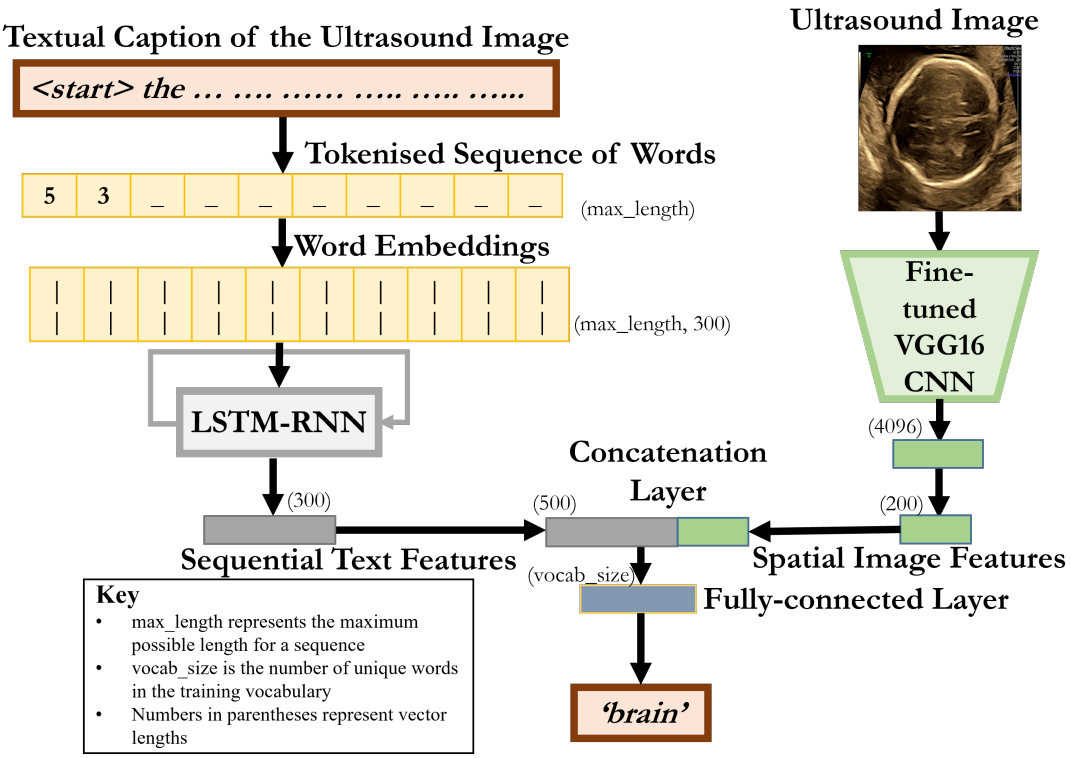


Figure 5.4: The image captioning model architecture is shown. A feature vector from the image branch and the text branch of the model are merged together before predicting the next word in the sequence. The text input is the partial caption. At inference, the partial caption includes the <start> token and the previously predicted words.

Another way to update w_1 and w_2 is to evaluate the model with the validation set after every curriculum iteration and to observe whether the model has scored better with syntax-focused metrics ($B1$ [25], RL [125], GB[126]) or semantics-focused metrics (ARS [4], F1). w_1 is decreased and w_2 is increased if semantics-focused metrics are higher than syntax-focused metrics or vice-versa. This is similar to the approach followed in [117] where performance on the validation set is used to guide the automatic creation of a curriculum.

5.4.3 Image Captioning Model Architecture

Fig. 5.4 shows the image captioning model architecture that is used in this section which is the late merge captioning model used in our papers [4, 5] and discussed

Table 5.3: Quantitative results are being shown comparing a model trained without a curriculum learning-based approach only using random sampling (NC), a model trained with DC, and models trained with CF-DCs. The evaluation metrics are split into two categories syntax or semantics focused.

	Syntax-Focused			Semantics-Focused	
Metric	*B1*	*RL*	*GB↓*	*ARS*	*F1*
NC	0.18	0.37	—	0.23	0.77
DC [5]	0.12	42.39	1.60	**0.43**	**0.97**
I1-CF-DC	**0.34**	**48.51**	**0.89**	0.39	0.96
T1-CF-DC	0.30	45.78	1.01	**0.43**	**0.97**

in the previous chapter. It consists of two branches. The right branch handles the image information and consists of a fine-tuned VGG-16 and fully-connected layers. The left branch embeds each tokenised word with a Word2vec embedding vector before passing it through a recurrent neural network. A flat feature vector from each branch is concatenated together and a prediction for the next word in the sequence is made based on the concatenated output.

5.4.4 Experiments

The dataset used for the image captioning task in this section was Dataset Cap. 2a. Information on this dataset can be found in Chapter 3.

5.4.5 Results and Discussion

Table 5.3 compares the results of CF-DC trained models introduced in this section with the results of a DC trained model and a traditionally trained model (NC for no curriculum) [5]. Table 5.3 shows that one approach (I1-CF-DC) performs better with respect to metrics focused on evaluating sentence quality or the syntax (B1, RL, GB), and another approach (DC) is better with respect to metrics that deal with semantics and anatomical relevance (ARS, F1). However, T1-CF-DC has identical performance metric scores to DC on the semantics-focused metrics while improving scores on the syntax-focused metrics.

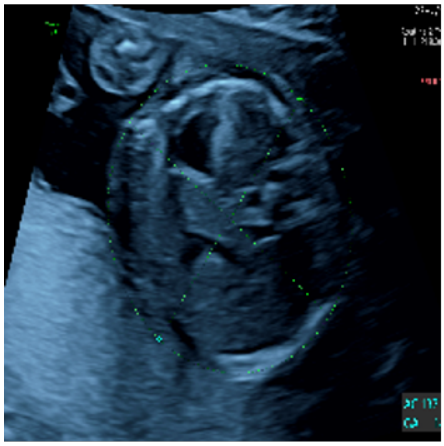
The question that arises is why are CF-DCs better performing when considering all the metrics together overall. In the vanilla dual curriculum (DC) where w_1 and

w_2 both have fixed values of 1.0 throughout training, we are effectively trying to merge two different curricula together. There is nothing to guarantee that one curriculum is not interfering with another's progress. There might be instances of forgetting as described in [79].

I1-CF-DC's lower score on semantics-focused metrics can be explained by the fact that the earlier focus on the image-associated complexity metric when assigning a dual curriculum score to a data sample loses its effect. The more complex and widely varied information in the text makes it more likely to forget what the model has learned from first primarily learning from the image-focused early part of the curriculum.

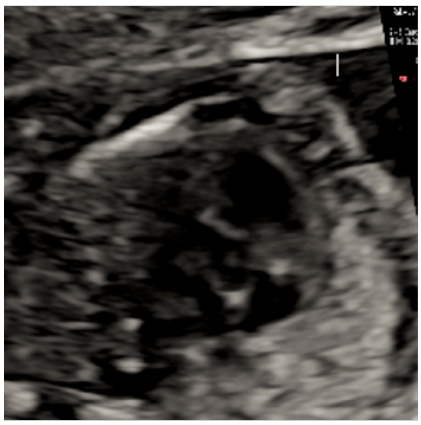
However, for T1-CF-DC, the data samples that are first encountered are primarily determined based on how 'easy' or 'hard' they are on the text-associated complexity metric. There is less forgetting that had to happen when later dealing with the image-focused part of the curriculum, and so, T1-CF-DC still outperformed DC and NC on syntax-focused metrics. The reasons for this are twofold. There is a greater variety and therefore complexity in the text information. We can imagine that a captioning model is effectively an image classifier with four possible classes connected with a language model that has hundreds of possible 'classes' that it needs to learn, each associated with a single word in the training vocabulary. The second reason is the fact that even when dealing with easy examples according to the text complexity metric, the model is still learning the image-based semantics through the anatomical vocabulary. The words of the vocabulary carry semantic meaning relevant to the image information while the opposite (learning the image information doesn't lead to learning the words) does not necessarily hold. These reasons explain why CF-DC trained models have relatively most balanced scores (ie. scoring the highest or second-highest on all metrics) across the different metrics.

Table 5.4 compares the performance of course-focused dual curricula on complex data. A complex data sample in an image captioning problem is an image-caption pair with an image feature vector that has a high Wasserstein distance from the mean of the image feature vectors and a caption that is long (in terms of



GT: this is where we measure the abdominal circumference
DC: us measure around the baby's belly
I1-CF-DC: let us measure around the baby's belly
T1-CF-DC: measuring the abdomen is like putting measuring tape around the baby's waist

Figure 5.5: Qualitative results for an abdomen image. GT is for Ground Truth. This is the caption that was provided by the sonographer directly. DC is for Dual Curriculum. This is the caption that was generated by a model that was trained with a dual curriculum. I1-CF-DC is for Image-First Course-Focused Dual Curriculum. This is the caption that was generated by a model that was trained with an I1-CF-DC. T1-CF-DC is for Text-First Course-Focused Dual Curriculum. This is the caption that was generated by a model that was trained with a T1-CF-DC.



GT: this is the left ventricular outflow tract
DC: ventricular outflow tract and right ventricular outflow tract are looking very very good
I1-CF-DC: ventricular outflow tract
T1-CF-DC: this is the left ventricular outflow tract

Figure 5.6: Qualitative results for a heart image. This is the caption that was provided by the sonographer directly. The full forms of GT, DC, I1-CF-DC, T1-CF-DC can be found in the caption of Fig. 5.5.

number of words) and consists of more infrequent words as determined by the tf-idf scores. To prepare this table, we took the test set, ranked each data sample by its complexity, and took the more complex half of the test set to evaluate on. Effectively, we found the same observation as in the previous experiment with I1-CF-DC performing better with syntax-focused metrics while T1-CF-DC

performing better with semantics-focused metrics.

Table 5.4: Comparing CF-DCs on Complex Data

	Syntax-Focused			Semantics-Focused	
Metric	*B1*	*RL*	*GB↓*	*ARS*	*F1*
I1-CF-DC	**0.38**	**46.68**	**0.56**	0.40	0.96
T1-CF-DC	0.33	42.66	0.88	**0.42**	**0.97**

5.5 Pseudo-Caption Preparation Pipeline

5.5.1 Introduction

In this section, we investigate the potential of using the larger existing image label (anatomical class) classification datasets to increase the image-caption pairs to train captioning models on. We aim to make use of the small number of captions available to automatically generate and associate a pseudo-caption to an image that lacks any corresponding caption that describes its content.

We also compare our approach with other existing text-based regularization techniques used in natural language processing (NLP) such as word dropout [127] and SwitchOut [128].

Related Works

An image captioning approach that does not require image-caption pairs was introduced in [129]. That method relies on the existence of a dataset of images, a corpus of text that may be unrelated to the images, and a previously trained visual concept detector that can detect visually meaningful concepts in an image. The corpus is used to train the model [129] to generate captions structurally well while the concepts detected through the visual concept detector is used to train the captioning model to associate certain keywords with objects in an image. Both the visual and textual information are represented in the same latent space. From that space, [129] demonstrates that it is possible to reconstruct either the visual or textual information. [129] also introduces the con2sen model which generates

a pseudo-caption from the objects detected in an image of interest. This uses the output of the con2sen model to pretrain their image captioning model.

Another approach used in [130] identifies concepts in images without captions that come from image recognition datasets and tries to find captions from an existing image-captioning dataset that are most semantically similar to those identified concepts. It allows for training the captioning models for objects that may not exist in the captioning dataset by identifying objects that do exist in the captioning dataset that it may be most semantically similar to.

Other work attempts to first train models on image-captions pairs in one domain, and then transfer the learned knowledge into another domain that is lacking in paired data [131, 132]. Other work first identifies the concepts exhibited in an object, say by object identification and uses that to build sentence templates around the identified objects in the image [27, 50, 133–135]. However, this kind of approach would require annotated data to train the concept identifier or object detector. Our approach attempts to circumvent this requirement by relying on acquiring these concepts that are exhibited by the nouns of the captions of similar images.

Identifying images that are similar is a cornerstone for work that relies on text retrieval and has been used in image captioning. In [136], the authors rely on using K-nearest neighbours to identify which image in the dataset is the target image most similar to. However, the downside of using text retrieval is the fact that assigning novel captions to the target images cannot happen as all possible captions come from the training dataset. However, our work is different from these studies in that we circumvent the requirement of having a visual concept or object detector and the annotated data to train one by proposing an undemanding method to create useful pseudo-captions from an existing image captioning dataset.

Data augmentation can be considered to be a model regularization technique [137]. Some regularization techniques used in NLP include word dropout, where several words in a sequence are randomly chosen to be dropped [127]. Another technique is SwitchOut, which randomly swaps some words in a sequence with

other words that exist in the training vocabulary [128]. We compare the proposed augmentation approach with these regularization techniques.

The step we introduce in this section is that nouns are identified in the retrieved caption and then along with the anatomical label of the target image are fed into a con2sen-inspired [129] model, which is a model that generates a sentence given concepts. This approach makes it possible to introduce an aspect of potential novel caption generation that would be missing otherwise.

5.5.2 Model and Training Details

We automatically annotate images with pseudo-captions for use in fetal US image captioning by leveraging existing image-captions pairs. The captions of images that are most similar to the captionless images will have their nouns used in creating pseudo-captions for the captionless images. This form of augmentation allows us to increase the number of image-caption pairs the model can be trained on. Our approach involves four main steps: (1) text retrieval, (2) noun extraction, (3) pseudo-caption creation from anatomical labels and extracted nouns, and (4) caption generation through an image captioning model architecture and framework.

Before we discuss the methods further, we briefly describe the available image data that we start with and what annotations already come with them. We can split the image data into two categories based on whether or not they come with captions that describe their content. The images come from the video frames of anomaly fetal ultrasound scans. The captions come from the transcribed text of the audio recordings of sonographers describing the content of the ultrasound scan videos. The images of the second category that lack any captions come from video frames but of different anomaly scans that lack accompanying audio recordings. All images in both categories possess a high-level automatically assigned anatomical label that is associated with the main visible anatomical structure in the image.

The whole process consists of four steps: (1) text retrieval, (2) noun extraction, (3) pseudo-caption creation, and (4) caption generation.

Step 1: Text Retrieval.

First, we perform text retrieval. We first calculate the cosine similarity between the image feature vector of the captionless image and the image feature vector of every single image in the training dataset. We retrieve the caption of the most similar image to the target image. Our captionless images came from the freeze frames of eight different fetal ultrasound scans with a step size of 16 between sampled frames.

We have also initially considered using the Wasserstein distance [121] or Mahalanobis distance [120] instead of the cosine similarity, but we have determined cosine similarity to be more likely to ensure that similar images will be of the same anatomical class or structure, and therefore, their captions would be more likely to still be applicable to that of the target image. This conclusion is evidenced in Figure 5.7.

Specifically, in Figure 5.7, we have the data samples of the real image-caption pairs split into subsets for visualisation purposes; with lower numbered subsets being ones that contain feature vectors that are most similar to the mean feature vector in the dataset. We observe that cosine similarity is the one most likely to ensure that data samples of the same structure (anatomical class) end up being grouped close together as part of the same subset based on their image feature vectors. For this reason, we have opted to work with the cosine similarity rather than Wasserstein distance or Mahalanobis distance. The cosine similarity $similarity_{cos}$ was defined in Eqn. 5.2.

Step 2: Noun Extraction.

After retrieving the caption of the most similar image to the captionless image, we extract its nouns through the TextBlob python library framework which can be used for parts-of-speech tagging as well as other NLP tasks [138]. When provided with a string, TextBlob returns a list of tuples. Each tuple consists of a word in the original string and its parts-of-speech tag. From that returned list, we create a sublist

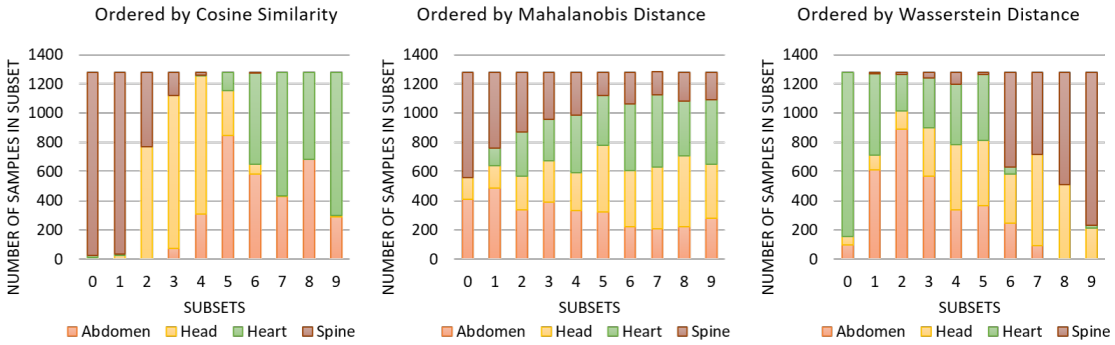


Figure 5.7: A visualisation on how the different similarity measures group data samples. We can see that cosine similarity's subsets are relatively more uniform. With cosine similarity, a subset is more likely to contain data samples of the same anatomical structure. In other words, with cosine similarity, the most similar image to an image is more likely to be of the same anatomical structure, and hence their captions are relatively more likely to be applicable to both. This is important because we rely on image similarity to retrieve captions from which we then extract nouns that along with the anatomical label are used to create pseudo-captions for images that lack captions.

consisting solely of the nouns that exist in the original string, and we use the parts-of-speech tags to identify those nouns. We hypothesise that these nouns adequately represent the inherent concepts associated with the image feature vector in question.

Step 3: Pseudo-Caption Creation.

Pseudo-captions are created from the anatomical labels and extracted nouns. We train an encoder-decoder sequence-to-sequence model often traditionally used for translation to transform a sequence of extracted nouns and the anatomical label ('abdomen', 'head', 'heart', or 'spine') of the captionless image to a pseudo-caption. This model is illustrated in Fig 5.8. This is similar in spirit to the con2sen model [129] which generates a pseudo-caption from the objects detected in an image of interest.

To train this sequence-to-sequence model, we perform the noun extraction on our real captions. The extracted nouns and the anatomical labels serve as the input while the real captions are the target outputs.

Step 4: Caption Generation.

Captions are generated through an image captioning model architecture and framework shown in Fig. 5.9. For training, we proceed with the real image-

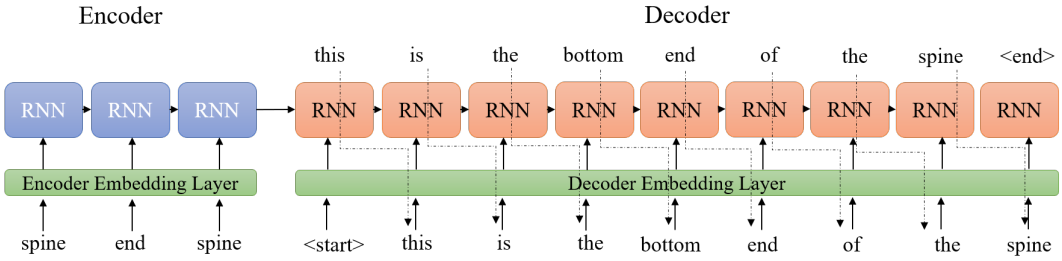


Figure 5.8: The sequence-to-sequence model architecture that 'translates' the sequence consisting of the anatomical label and the extracted nouns into a pseudo-caption is shown. In the input sequence, the first 'spine' is the anatomical label, and 'end' and the second 'spine' are the extracted nouns. The words in the sequence are represented with word embedding vectors before being passed as input to the encoder and decoder RNNs.

caption pairs augmented by image-pseudo-caption pairs with the pseudo-captions being generated in the previous step. The framework, which follows our paper [5], also includes an image feature vector classifier that classifies an image to one of four possible classes, 'abdomen', 'head', 'heart', and 'spine'. There are four variants of the model shown in Fig. 5.9, one for each of the four anatomical structures. So, there is a separate captioning model for each anatomical structure, and once the image feature vector classifier classifies an image, the appropriate captioning model is initiated to then caption the said image.

The actual image captioning model is the same late merge captioning model discussed in the previous chapter. The model is made up of two parts, one focused on the text and the other focused on the image. The text-focused branch in this figure is in the lower half of the figure, and it shows each word in the input sequence is given a token before going through an embedding layer. The embedding layers uses weights from a Word2vec network that has been pretrained on the GoogleNews corpus. The sequence is then passed through a recurrent neural network. The upper half of the figure shows the image-focused branch which consists of a VGG16 convolutional neural network (CNN) that has been fine-tuned on fetal ultrasound images of the same gestational age. The CNN is followed by a couple of fully-connected layers. This image-focused branch on its own serves as the aforementioned image feature vector classifier. From both branches, a flat feature vector is obtained, one containing the inherent image information and the other containing the equivalent

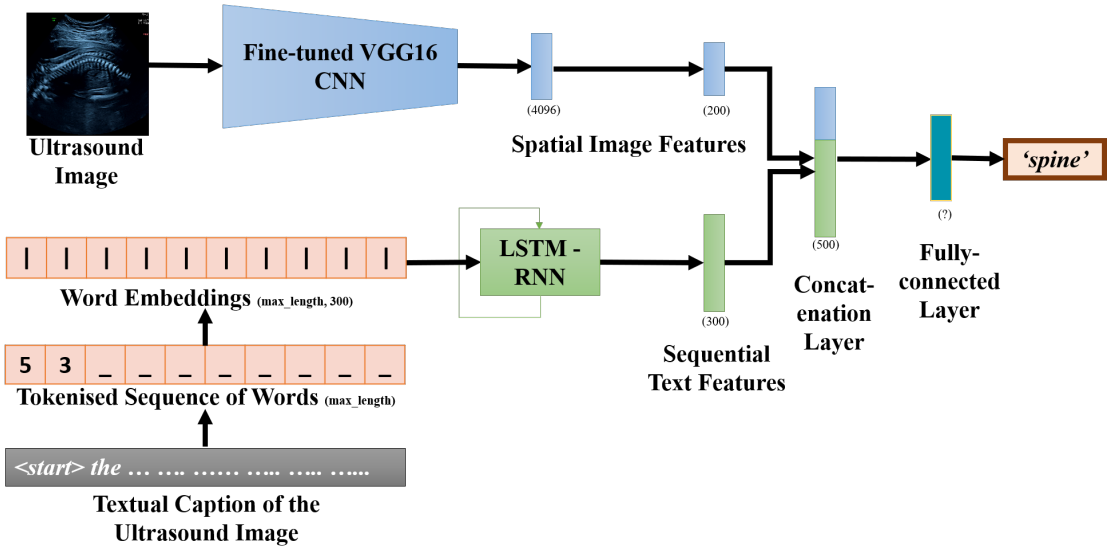


Figure 5.9: The late merge image captioning model used is shown. The model consists of two branches, an image branch and a text branch. A feature vector associated with the image information from the image branch and a feature vector associated with the text information from the text branch are concatenated together before a prediction of the next word to generate in the sequence is made. Max length represents the maximum number of words a caption of this anatomical structure could consist of. '?' represents the vocabulary size associated with this specific anatomical structure.

but for the text. The vectors are concatenated together. Finally, a prediction for the next word to be generated is made. The process is repeated until the maximum possible length is reached or a special end token ('<end>') is generated.

5.5.3 Experiments

We trained multiple models to compare our approach (including image-pseudo-caption pairs in the training dataset) with a baseline (no text augmentation or regularization) and other regularization techniques (word dropout [127] and SwitchOut [128]). The number of epochs was set at 150, but early stopping was applied if the validation loss did not improve for seven consecutive epochs. We applied a batch size consisting of 32 image-caption pairs. The Adam optimization algorithm was used with a learning rate of 0.001.

Datasets and Data Preparation

The dataset that was used for the image captioning task in this section is Dataset Cap. 2b. Information on this dataset can be found in Chapter 3.

5.5.4 Results and Discussion

The quantitative results are shown in Table 5.5. Qualitative results for two typical examples are shown in Fig. 5.10. When compared to only image-caption pairs, we can immediately see an increase in almost all of the evaluation metrics when incorporating the image-pseudo-caption pairs with the exception of $F1$ where both obtained the high score 0.97/1.00. This simply implies that generated captions from both will describe the appropriate structure. However, from the *ARS* score, we notice that the models that trained with pseudo-captions score higher. The *ARS* incorporates softmax probabilities into its calculation. A higher *ARS* score then translates to the model producing higher softmax probabilities for the relevant terms. Roughly, we can say the model is more sure that these generated words are more relevant. All of the syntax-focused metrics are better. With regards to the *BLEU-1* score, specifically, the set of models trained with pseudo-captions receives a score 0.30 which falls within the range that [139] would describe as being 'understandable and good'. On the other hand, the set of models trained without pseudo-captions obtains a *BLEU-1* score of 0.13. A score of 0.13 falls within the range that [139] would describe as 'hard to get the gist of'. We can see this behavior exemplified in the generated captions in Fig. 5.10. This result is unsurprising. With more data, a model is more likely to learn proper sentence structure as it has more data to learn from, and it is pleasing to see that the pseudo-captions although not provided directly by a sonographer can still help a model to score higher on the evaluation metrics.

All the syntax-focused metrics are better with our proposed approach except *GB*; however, that can be explained by the fact that word dropout drops words from the sequence, and so the model ends up learning to generate shorter sentences. With shorter sentences, the number of grammatical mistakes decreases. Word dropout

and SwitchOut have notably lower scores on the semantics-focused metrics. This phenomenon can be explained by the fact that, with word dropout and Switchout, anatomically relevant words could be dropped or switched out during training. Even with a word dropout rate of only 0.2, a caption of, say, five words will lose one word, and that could be the word that refers to the anatomical content. There is a high chance of losing semantically meaningful words in shorter captions when using word dropout and similar techniques.

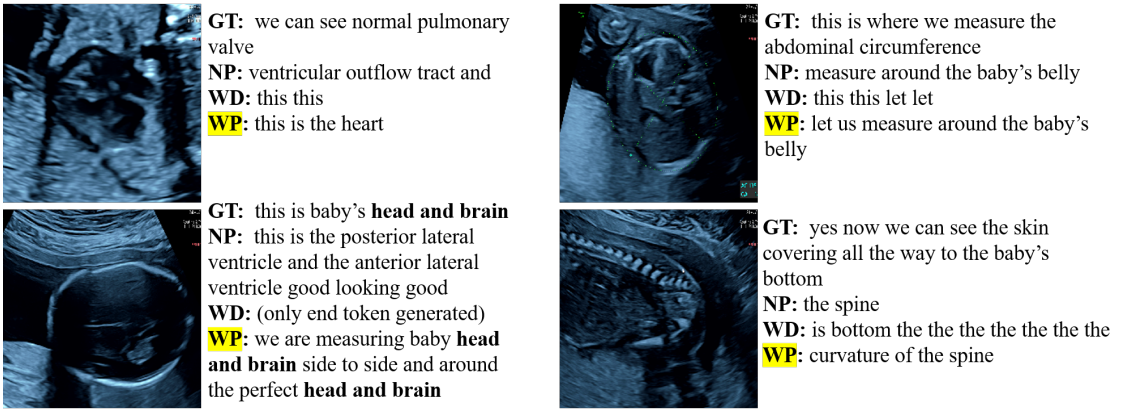


Figure 5.10: Qualitative results for a heart image image and an abdomen image. GT stands for Ground Truth as spoken by a sonographer. NP stands for model trained with No Pseudo-captions. WD stands for model regularized with Word Dropout. WP stands for model trained With Pseudo-captions (our proposed method).

Table 5.5: Quantitative results comparing our proposed augmentation approach with other regularization techniques

Methods	Syntax-Focused			Semantics-Focused	
	BLEU-1	*ROUGE-L*	*GB↓*	*ARS*	*F1*
No pseudo-captions (baseline)	0.17	0.24	1.26	0.39	**0.97**
With pseudo-captions (ours)	**0.30**	**0.42**	1.14	**0.77**	**0.97**
Word Dropout token level	0.01	0.03	**0.52**	0.05	0.34
Word Dropout vector level [127]	0.07	0.17	0.84	0.03	0.03
SwitchOut [128]	0.11	0.20	2.55	0.08	0.85
Synonym swapping	0.11	0.20	2.62	0.09	0.73

5.6 Summary

This chapter has made specific contributions:

1. We first present a novel curriculum learning approach to train a natural language processing (NLP) based fetal ultrasound image captioning model. Datasets containing medical images and corresponding textual descriptions are relatively rare and hence, smaller-sized when compared to the datasets of natural images and their captions. This fact inspired us to develop an approach to train a captioning model suitable for small-sized medical data. Our datasets are prepared using real-world ultrasound video along with synchronised and transcribed sonographer speech recordings. We call this approach the “dual-curriculum" method. The method relies on building and learning from curricula of image and text information for the ultrasound image captioning problem. We compare several distance measures for creating the dual curriculum and observe the best performance using the Wasserstein distance for image information and tf-idf metric for text information. The evaluation results show an improvement in all performance metrics when using curriculum learning over stochastic mini-batch training for the individual task of image classification as well as using a dual curriculum for image captioning.

   We determined that curriculum learning helps to train more accurate models than using a traditional stochastic mini-batch preparation process does for image captioning. We built curricula for image classification and image captioning tasks suitable for small-sized medical datasets. In particular, we found the Wasserstein-distance-based curriculum and the Mahalanobis-distance-based curriculum to be good options for automatic fetal ultrasound image captioning.

2. The second contribution of this chapter is to propose a curriculum learning captioning method to caption fetal ultrasound images by training a model to dynamically transition between two different modalities (image and text) as training progresses. Specifically, we propose a course-focused dual curriculum

method, where a course is training with a curriculum based on only one of the two modalities involved in image captioning. We compare two configurations of the course-focused dual curriculum; an image-first course-focused dual curriculum which prepares the early training batches primarily on the complexity of the image information before slowly introducing an order of batches for training based on the complexity of the text information, and a text-first course-focused dual curriculum which operates in reverse. The evaluation results show that dynamically transitioning between text and images over epochs of training improves results when compared to the scenario where both modalities are considered in equal measure in every epoch.

3. The third contribution of this chapter is to propose an approach to augment and increase the number of data available to train a fetal ultrasound image captioning model when in possession of only a small number of images that have corresponding textual descriptions by leveraging an existing larger dataset annotated for image classification like tasks. The process first requires identifying, for the target image, an image from the smaller captioning dataset that it is most similar to. Along with the known corresponding classification label that the target image already possesses, the nouns from the caption of that similar image are identified and fed into an encoder-decoder sequence-to-sequence model to produce a pseudo-caption for the target image. These nouns represent the semantically meaningful content in the image. This allows a seamless automatic annotation process for images that lack human-prepared captions to then use them in training models for the image captioning task.

   In Section 5.5 of this chapter, we demonstrate a process to create pseudo-captions for fetal ultrasound images that lack any captions by making use of existing image-caption pairs. We proceed to show that these image-pseudo-caption pairs can improve results in the fetal ultrasound image captioning task.

This chapter explored what can be done to improve model performance before starting to train the models. It explored ordering data through a dual curriculum and by augmenting available text data through the pseudo-caption creation pipeline. The next chapter looks into improving model performance for video, rather than image, captioning by leveraging an additional data modality, eye gaze tracking data.

Education is the best provision for old age.

— Aristotle [106]

6

Leveraging Data from Other Modalities in Fetal Image Captioning

Contents

6.1 Introduction 123
6.1.1 Chapter Outline 125
6.1.2 Changes from Previous Chapters 125
6.1.3 Purpose of Chapter 126
6.2 Originality and Individual Role 126
6.3 Gaze-Assisted Captioning 127
6.3.1 Introduction 127
6.3.2 Model and Training Details 128
6.3.3 Results and Discussion 139
6.4 Conclusions 145

6.1 Introduction

When we were finishing up working on Chapter 4, we realised that the captioning model we built, could very well generate appropriate captions, but the sonographer could be talking about another object that is also present in the image. An example of this is shown in Fig. 6.1a.

This led us to consider the use of eye gaze data in captioning which was readily available as part of the data of the PULSE project. There is great novelty in the

(a) Ground truth: **'rib'**
Generated: **'we can see the stomach'**

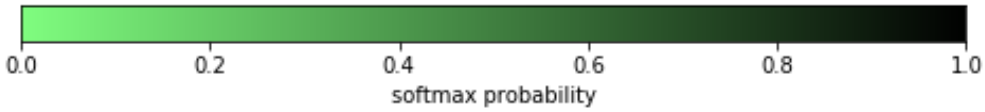


Figure 6.1: An example of where the ground truth and the generated captions do not match. Note that the stomach is visible in Fig. 6.1a, but the sonographer happened to be talking about a rib in this instance. The higher the softmax probability associated with a generated word, the darker the green color of that generated word. This figure was originally shown in Chapter 4.

leveraging of eye gaze data to aid in the captioning process in medical imaging. The model generates text aided by the use of eye gaze tracking information of the sonographers that reflects their visual attention on parts of the video clip they consider relevant when describing the content.

Our third contribution in this thesis is a novel gaze-assisted natural language processing (NLP)-based video captioning method to describe routine second-trimester fetal ultrasound scan videos with words from the vocabulary of spoken sonography. As will be shown later in this chapter, results show that the proposed gaze-assisted models can generate richer captions for clinical fetal ultrasound scan videos at the expense of the perceived sentence structure and that the generated captions are similar to sonographer speech in terms of discussing the visual content and the scanning actions performed.

The novelty of our multi-modal method is that the textual captions that describe the spatio-temporal scan video content are learnt from sonographer speech recordings

and the generation of captions is assisted by sonographer gaze-tracking information reflecting their visual attention while performing live-imaging and interpreting a frozen image. As part of an ablation study, we compare spatio-temporal deep networks trained using three multi-modal configurations, namely; a gaze-less neural network, a neural network using real sonographer gaze, and a neural network using automatically predicted gaze. We evaluate the machine learning model architectures through established general text-based metrics (BLEU, ROUGE-L, F1 score) and domain-specific metrics that consider the richness and efficiency of the generated captions with respect to the scan video.

6.1.1 Chapter Outline

This chapter covers how human gaze can assist in automatic captioning of ultrasound video. In the beginning, we first talk about what makes this challenge different from what has been encountered earlier, particularly in Chapter 4. The methods used to process the data and prepare the clip-captions for the experiments conducted for this specific chapter are also discussed. The architecture of the model and the process whereby it has been trained is then presented.

We then present results and discuss the conclusions that we can draw from them. We conclude by addressing where this work can be further explored going forward in the future.

6.1.2 Changes from Previous Chapters

What is new and different in this chapter when compared to the previous chapters is first and foremost the fact that we moved from image captioning to video captioning. We swap cross-entropy loss for focal loss when training the captioning models. We also incorporate the use of eye gaze tracking data into the captioning process, to leverage data from a third modality to aid in fetal visual captioning.

6.1.3 Purpose of Chapter

What is it that we do in this chapter? We propose an original multi-modal deep learning method to describe ultrasound videos using sonographer spoken words. We are interested in being able to adequately caption visual information with a temporal component that would be lost when only attempting to caption single images.

An example of where this is applicable is in a sentence, such as 'nice contraction of the heart four chamber view we can see the right and left ventricles'. A single still image would not show the beating of the heart. Since our data is spatio-temporal in its original raw form, it is not surprising to find an example such as this.

Our model learns to associate sonographer-specific vocabulary with fetal ultrasound clips. The contribution is to propose an original multi-modal deep learning method to describe ultrasound videos automatically in the same way that sonographers do using the spoken word while incorporating eye gaze data in the captioning process.

6.2 Originality and Individual Role

Training the different captioning model configurations with varying levels of gaze involvement fell under my purview. Tabulating results and the experiments mentioned in detail in this chapter were done by me. Comparing word embedding vectors for fetal video captioning was also a task that I handled. The saliency prediction model was work done by Yifan Cai who had used predicted gaze saliency maps in an image classification related task [140]. He prepared the predicted gaze saliency maps used in this chapter. Generating real gaze attention maps from real gaze points was made possible through Python tools prepared by Richard Droste. The gaze data acquisition system had been developed as part of the PULSE project.

6.3 Gaze-Assisted Captioning

6.3.1 Introduction

Video captioning, the subject of this chapter, extends the concept of image captioning by exploiting rich, spatio-temporal information in the video clip of interest [141]. Our interest in this chapter is to investigate automatic ultrasound video captioning. In particular, in this chapter, the objective is to build a video captioning method to describe routine second-trimester fetal ultrasound scans in the sonographer spoken vocabulary. A novelty of our work is that we pose this as a *multi-modal* problem, using ultrasound video, text, and gaze as input. We propose an original multi-modal deep learning framework to describe ultrasound (US) scan videos assisted by sonographer gaze information. To the best of our knowledge, this is also the first attempt to perform automatic *video* captioning using sonographer descriptions derived from their speech recordings, that is, a captioning method developed for live scan videos rather than images in fetal ultrasound scans.

In this chapter, we enrich our models, making them more accurate, by leveraging data from another modality that is collected as part of the PULSE project, eye-gaze tracking data. There is additional novelty in leveraging sonographer gaze-tracking information to aid the video captioning process, as the model generates text that reflects the sonographer visual attention during the acquisition and spoken description of the scan video. We hypothesise that the additional information prevalent in the eye gaze tracking data will help build better performing captioning models.

Visual attention, as described in [140], refers to the actual human attention in the form of their gaze behaviour when viewing video clips or images. The motivation behind using gaze for visual attention is the hypothesis that the gaze information could help in training better deep learning models. This hypothesis was investigated and proven by [140] for the task of standard plane detection.

Contributions

This chapter proposes an original multi-modal deep learning method to caption routine fetal US scan videos using sonographer spoken words derived from their

speech recordings, and gaze-tracking data recorded during the scan acquisition process. This work is novel because: (1) it is the first attempt to perform automatic video captioning on fetal ultrasound scans using sonographer spoken vocabulary; (2) it proposes a novel mechanism to use expert gaze information. In terms of a quantitative evaluation metrics, the video caption generating model with gaze is found to perform better than one without. Furthermore, we are not aware of prior work that utilizes gaze-tracking information to generate text for a medical imaging modality or, more generally, any work that attempts medical video captioning.

6.3.2 Model and Training Details

Methods

In this subsection and the following ones, we discuss the methods used in the experiments conducted for this chapter.

Multi-Modal Data Preprocessing

In the subsubsections below, we describe the multi-modal data preprocessing steps.

Video, Gaze, and Audio Capture

Clip-caption pairs were defined as follows. During a fetal US scan acquisition, the sonographer performs search, fine-tuning and interpretation of different fetal anatomies. The sonographer can freeze at a frame where they view the anatomy. Based on the freezing action, a video clip corresponding to every unique caption with the centre frame as a freeze frame was extracted. Then, 12 frames were sampled from the segment of the clip spanning the length of the caption. Fig. 6.3 shows a clip-caption pair.

The nature of the words in the dataset can be seen in the word clouds in Fig. 6.2. In image captioning tasks developed for spatio-temporal data [4], such as videos and speech recordings, each caption can correspond to more than one image, as spoken text naturally covers a temporal segment of visual content. In the current chapter, we are not so constrained and have a unique caption for each video clip. The average length of a clip-caption pair in our dataset before concatenation is

7.254 seconds. In general, in routine fetal US scans, sonographers spend unequal amounts of time looking at different anatomies of interest, which was also observed in our manually labelled video clip dataset [92]. This leads to a natural class imbalance for the four common anatomical classes considered in this work, namely, abdomen, head, heart, and spine. The distribution (% of clip-caption pairs) of the four anatomies is 11.7%, 31.4%, 40.5%, and 16.4% respectively.

Gaze-Tracking

The gaze data was filtered following the protocol described in [98]. Binary maps were created from gaze data, with gaze-points labeled as 1 and others 0. Sonographer visual attention maps were subsequently generated by convolving the binary map with a Gaussian kernel with $\sigma = 40$ pixels, assuming an observer-to-screen distance of 0.5 m, human field of view of 1.5° visual angle, and screen dimensions of 33.2 × 20.7. The visual attention map was further normalised so that each pixel value is in the $[0, 1]$ range. One of the three captioning model configurations that we explore later in this chapter is a model that incorporates **predicted gaze saliency maps**. These predicted gaze saliency maps are obtained from a *spatio-temporal saliency prediction model* [99] and attempt to highlight the relevant parts of the anatomy, spatially localizing them in the image. In [99], 10 frames are sampled from a video clip on which to predict the saliency of. For this work, this spatio-temporal saliency prediction model was pre-trained using the same large simultaneous gaze-tracking dataset of second-trimester scans [99] with the aim, however, of predicting saliency maps of 12 sampled frames, rather than ten, from a video clip. Video clips from the PULSE data as well as the accompanying real gaze data was used in training the spatio-temporal saliency prediction to generate saliency maps. These generated saliency maps are what we refer to in this chapter as predicted gaze.

Video Captioning Model Architecture

We summarize the video captioning neural network architecture in Fig 6.4. The architecture is inspired by the split-based image captioning method [4, 38, 68]. We consider three configurations of this model; the gaze-less (GL) configuration, the

(a) Word cloud when image is frozen or probe is stationary.

(b) Word cloud during the search or fine-tuning process.

Figure 6.2: Word clouds showing the diversity in the dataset. A larger word occurs more often. Red represents a noun, green represents an adjective, blue represents a verb, and yellow represents an adverb. Words that make up the other parts-of-speech have been dropped for the sake of clarity and to emphasise more meaningful words that come in the form of nouns, adjectives, verbs, and adverbs.

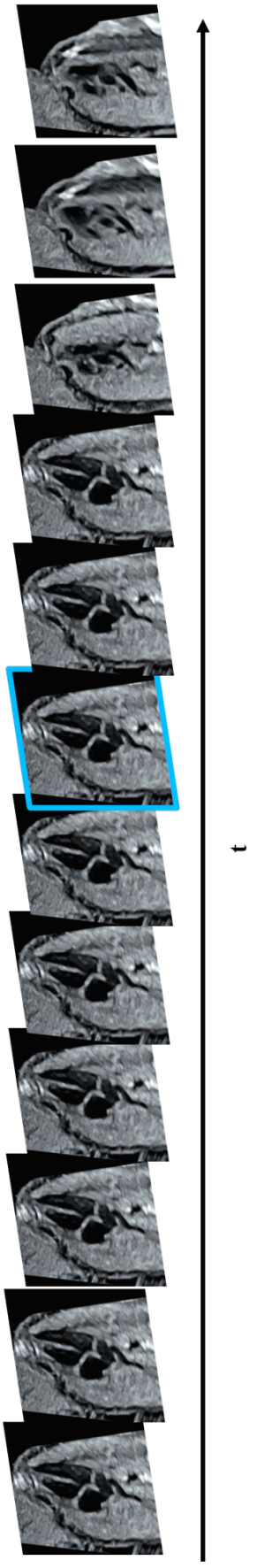


Figure 6.3: Video clip of a beating fetal heart centred around a freeze frame (outlined in cyan) with the corresponding caption: "*you can see the heart beating very nicely and this is a four chamber view and three vessel trachea view*".

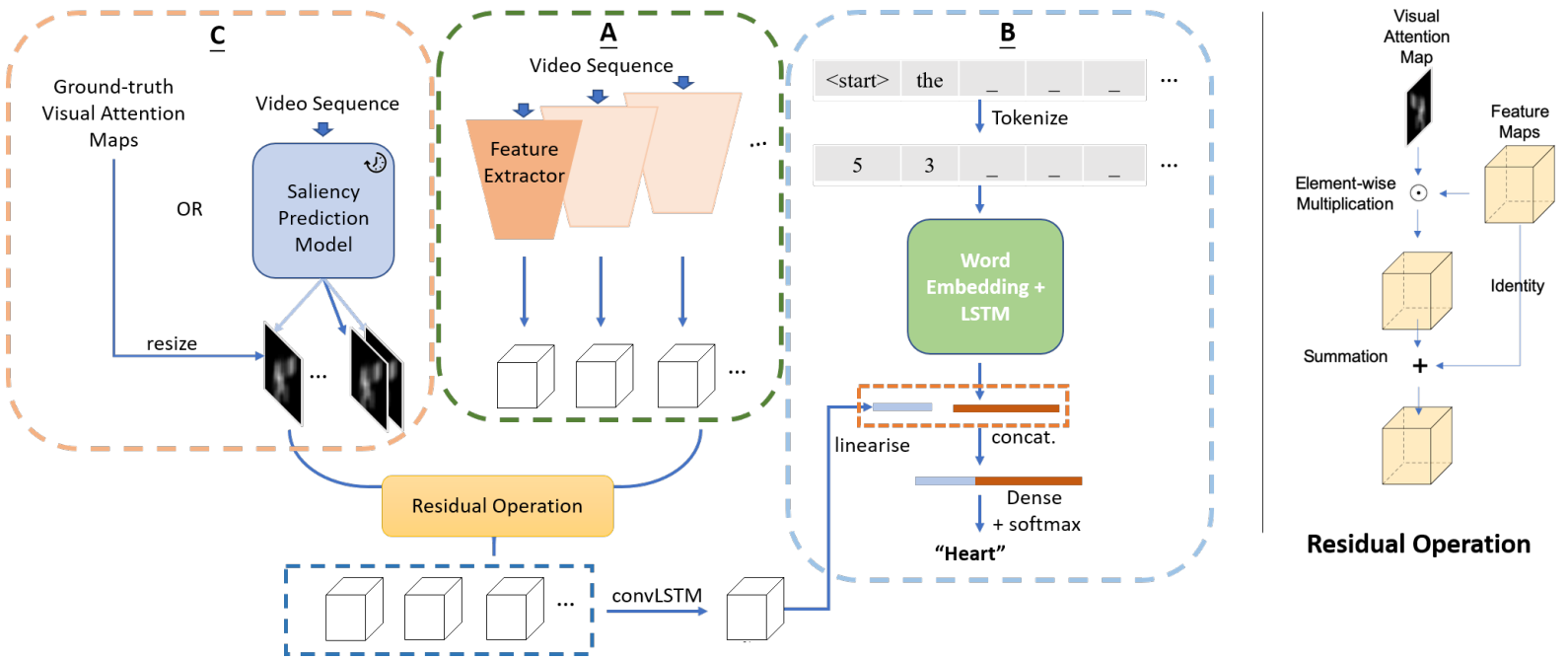


Figure 6.4: The architecture of the multi-modal model is shown in the this figure. The specifics of the residual operation that is performed on the outputs of Blocks A and C are shown on the right side of this figure. The gazeless model configuration only includes Blocks A and B. Block A represents the branch of the model where the spatial feature information is extracted from the video clip for each of its sampled frames by a VGG16 CNN. Block B represents the branch handling text information. The sequence of words generated so far are tokenized and embedded with a Word2vec embedding vector before being passed as input to an LSTM-RNN. The last hidden state of this LSTM-RNN is concatenated with the linearized feature vector from the convolutional LSTM. The real gaze and predicted gaze model configurations also include Block C. In Block C, either the ground truth visual attention maps are used (in the case of the real gaze configuration) or predicted gaze saliency maps are predicted and then used for each sampled frame in a video clip by a previously trained saliency prediction model. The extracted feature blocks from Block A and the gaze maps from Block C are combined together through the residual operation shown in detail on the right side of the figure. In the gazeless configuration, there is no residual operation to be performed on the sequence of feature blocks. They are passed directly to the convolutional LSTM.

real gaze (RG) configuration, and the predicted gaze (PG) configuration. The GL configuration, as its name suggests, has not been trained with gaze data, and the part of the framework that handles gaze information is not relevant to it. The RG configuration uses attention maps that come directly from the real gaze points of the performing sonographers as explained in more detail in this section. The PG configuration uses saliency maps that have been generated by the previously trained saliency prediction model.

We explain the key features of Fig. 6.4 next.

Block A depicts the *visual spatio-temporal branch*, where each video clip is encoded by first extracting successive features of each constituent frame from the last convolutional block of a fine-tuned VGG16 [58], and then feeding these as a sequence into a convolutional-LSTM layer [142] to model the spatio-temporal context in the video clip. The last hidden state of the convolutional LSTM layer is flattened before going through a fully-connected layer.

Block B presents the *captioning branch* for learning joint video and text embeddings. Here, the tokenized sequence of a partial caption is passed through an embedding layer that represents each word with a 300-dimensional embedding vector. The sequence can be up to 83 elements long, since that is the longest caption in our training dataset. The sequence of embedding vectors is the input to an LSTM layer, the final hidden state of which is concatenated with the vector from the visual branch. Finally, a softmax operation is applied to the concatenated vector to output a probability distribution over the training vocabulary from which the next generated word is determined. The gaze-less (GL) model configuration is effectively the model architecture with Blocks A and B (but without Block C). The real gaze model configuration and the predicted gaze model configuration all include Blocks A, B, C. They differ in that in the predicted gaze model, in Block C, the gaze maps come from the saliency prediction model. Whereas, in the real gaze model configuration, the gaze maps are the ground truth visual attention maps obtained through real sonographer eye gaze points.

The justification for using this architecture, illustrated in Fig. 6.4 as **Block A** and **Block B**, is primarily motivated by the fact that we have a relatively small sized dataset to work with. The captioning branch in **Block B** shows that that text encoding LSTM will not encounter the image information, making it easier to train the LSTM-RNN as it would only need to learn the text information. There is benefit to doing this for captioning when working with a relatively small sized dataset [38, 68]. **Block A** was built to make it relatively easy to transfer learn from a VGG16 pretrained on ImageNet and fine-tuned on fetal ultrasound images. We wanted to be able to extend our image captioning model in [4] to video captioning while making the most of what is available to be trained on. We used that same CNN only duplicated twelve times, once for every sampled frame, in the sequence before being fed into a convolutional LSTM to process and handle changes in temporal information between those three frames.

The gaze-assisted model configurations that include real gaze (RG) or predicted gaze (PG) have the gaze information included through a *gaze-encoding branch* shown in **Block C**. In the RG configuration, ground-truth real gaze visual attention maps are directly resized to match the dimension of the features extracted from each frame in **Block A**. In the PG configuration, input US images in **Block A** are fed through a spatio-temporal variant of a pre-trained saliency prediction model [99], which we describe below, to predict visual attention maps. In both cases, the predicted visual attention maps filter the features extracted from **Block A** using a residual operation, *i.e.* element-wise multiplication followed by identity summation to highlight visually-salient regions in image features.

The spatio-temporal saliency prediction model first models "static visual attention" using the method of [98] and predicts a visual attention map on each input video frame. Then, the features extracted are fed into a bi-directional convolutional LSTM to model "dynamic visual attention", accounting for the temporal variation of sonographer visual attention. In addition to the *mean squared error (MSE)* and *Kullback-Leibler Divergence (KLD)* losses, the model employs a soft-Dynamic Time Warping (sDTW) loss [143] to align predicted visual attention maps and

ground-truth maps. Fig. 6.5 shows a couple of fetal ultrasound frames and their corresponding real gaze attention maps and predicted visual attention maps.

Model Training

Cross-entropy loss is among the losses commonly used in NLP tasks [144] and captioning specifically [4, 38, 68]. Focal loss is a modified version of cross-entropy loss that gives a greater importance to misclassifications by down-weighting correct classifications when training a deep learning model, hence, it is useful for classification problems that have imbalanced input datasets [145]. In our training vocabulary, word imbalance is observed. For example, 'the', being an essential article used in the English language, is unsurprisingly one of the most commonly occurring words in our dataset, having over 700 instances. On the other hand, 'iliac' is a word that is significantly more relevant anatomically but exists in only one caption. However, the more common words, despite their grammatical importance, are not as essential for the anatomical description as some of the less-represented words. Figure 6.7 shows the count of the most commonly occurring words in the training dataset. In practice, without addressing natural imbalance, a model is more likely to generate the more common words due to their prevalence. As captioning can be considered as a classification problem with words as classes, the vocabulary imbalance justifies choosing focal loss in this work.

Equation 6.1 defines the focal loss function FL

$$FL = \sum_{w=1}^{v} -\alpha \left(1 - p_w\right)^{\gamma} y_w \log(p_w) \tag{6.1}$$

where α=0.25 and γ=2 and where w is an index in a word list, v is the total number of words in the vocabulary, y_w is the associated ground truth value of that word, and p_w is the associated softmax probability of that word. α is described as a weighting factor and γ is the focusing parameter. We use the same values for α and γ (α=0.25 and γ=2) as in [145].

Figure 6.5: Two fetal ultrasound frames, their real gaze attention maps, and their corresponding predicted gaze attention maps using the method of Cai et al (2018). This figure shows information do the different model configurations (gazeless, real gaze, predicted gaze) learns from. **(a)** A fetal ultrasound frame of a head (left), that frame's real gaze attention map (center), and the corresponding predicted gaze attention map (right). **(b)** A fetal ultrasound frame of a spine (left), that frame's real gaze attention map (center), and the corresponding predicted gaze saliency map (right). The attention in the real gaze attention map is instantaneous in the sense that it reflects where exactly one sonographer looked at on that frame. The predicted gaze saliency map represents an averaged attention that partially mirrors the shape of the anatomical structure on the screen. The real gaze attention map and the predicted gaze saliency map are correlated but are not the same. In both maps, regions in the same vicinity are activated.

Equation 6.2 defines the cross-entropy loss function CE

$$CE(p, y) = \begin{cases} -log(p) & \text{if } y = 1 \\ -log(1-p) & \text{otherwise} \end{cases}. \tag{6.2}$$

We performed some experiments to compare focal loss and cross entropy loss for captioning. The experiments used the image captioning model architecture introduced in Chapter 4. The confusion matrices in Fig. 6.6 show how the models trained with the two different losses compare in terms of performance. The experiment involved training two image captioning models with the same dataset and settings used in the experiments of Chapter 4. The difference is one model was trained with a cross-entropy loss while the other model was trained with a focal loss. The confusion matrices show the performance of the two models by determining whether the anatomical class of the generated caption matches that of the the ground truth.

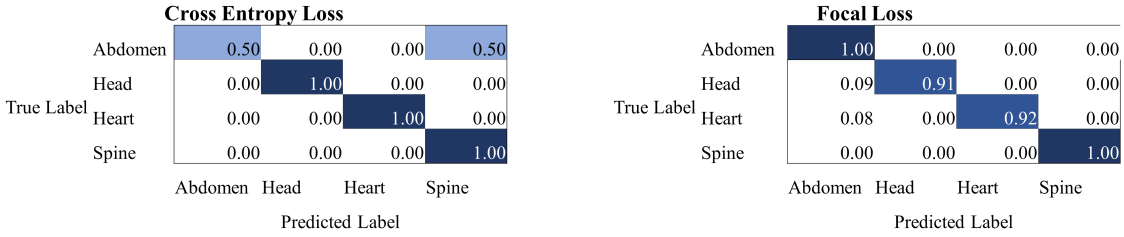


Figure 6.6: Confusion matrices comparing a model trained with cross entropy loss (left) and a model trained with focal loss (right) are shown.

To augment the original training dataset, the sequence of frames was randomly augmented on the fly during training by either rotating by an angle between -30° and 30° around the vertical y-axis or by horizontally reflecting the image. Captions were augmented by synonym swapping through WordNet [146]. For every caption, a word in it was randomly chosen to be replaced by its synonym but on the condition that the synonym was already a part of the training vocabulary. We used Adam optimization [147] when training deep learning models with stochastic gradient descent. Early stopping was done when the validation loss ceased to improve after 28 epochs. To reduce overfitting, dropout with a rate between 0.4 and 0.5 was used. The different model configurations (see Section 6.3.2) were trained with

teacher forcing [113]. Teacher forcing is a commonly used practice to train recurrent neural networks with sequential information. It involves providing a sub-sequence of ground truth words of appropriate length as input to the network at every time step, regardless of what output (sequence of words) has been generated by the model in the previous time steps [113]. So, if a ground truth caption is 'the healthy spine', and we are at timestep $t = 2$, the partial caption that will be provided as input to the model at that timestep is ['start-token', 'the', 'healthy'] which come from the ground truth caption. The model being trained, though, may have generated different predicted words in the previous timesteps. For example, it may have generated 'a' at $t = 0$, and 'curved' at $t = 1$, but ['start-token', 'a', 'curved'] will not be used as input to the model at $t = 2$. This training behaviour continues at all other timesteps. At inference, the model relies on its previously generated words and the corresponding sequence of frames to generate the next word. We use a framework where there is a captioning model for each of the four anatomical structures. A video clip classifier based on **Block A** classifies the video clip and then starts the appropriate captioning model based on the classification output.

Word Embeddings

As initially mentioned in Chapter 4, word2vec is an established embedding method by which words are represented by embedding vectors [124]. It is effectively a neural network with a single hidden layer trained to predict either the next word given previous words or the context, as the target, around the word given as input [110, 124]. BioWordVec is a vector representation for words that are specifically relevant to the biomedical domain [148]. It is based on the fastText embedding [149] but goes beyond by incorporating Medical Subject Headings (MeSH) terms, which is effectively in the form of an ontology [148]. On the other hand, word2vec is not specifically optimised to make use of medical text and information. The key improvement of fastText [149] over word2vec is that fastText takes the subunits of words into consideration whereas word2vec does not. One could see how that could be useful in the biomedical domain, where even the word 'biomedical' itself

could be divided into the sub-units, 'bio' and 'medical'; however, truly noticing this advantage depends on having a dataset consisting of enough words that can be divided into these sub-units [148]. In this work, we compare the word2vec and BioWordVec embeddings within our modelling framework and dataset in a experiment reported and discussed in Section 6.3.3.

Evaluation Metrics

As mentioned previously in Chapter 3, we use two sets of metrics to evaluate modelling performance; one focused on the quality of video clip classification and one focused on the quality of the generated caption. With regards to the visual content classification-related metrics, we calculate the classification F1-Score and a specifically designed anatomical relevance score (ARS) that was introduced in [4]. Caption quality is reported using $BLEU$ [25], and $ROUGE\text{-}L$ [33]. More information on these metrics can be found in Chapter 3. We also report metrics that measure caption efficiency and richness. $Efficiency$ should score correct captions with as few words as possible higher. $Richness$ should reflect how complete a generated caption is. To calculate efficiency and richness, we drew inspiration from precision and recall within the context of information retrieval [150]. Equations 6.3 and 6.4.

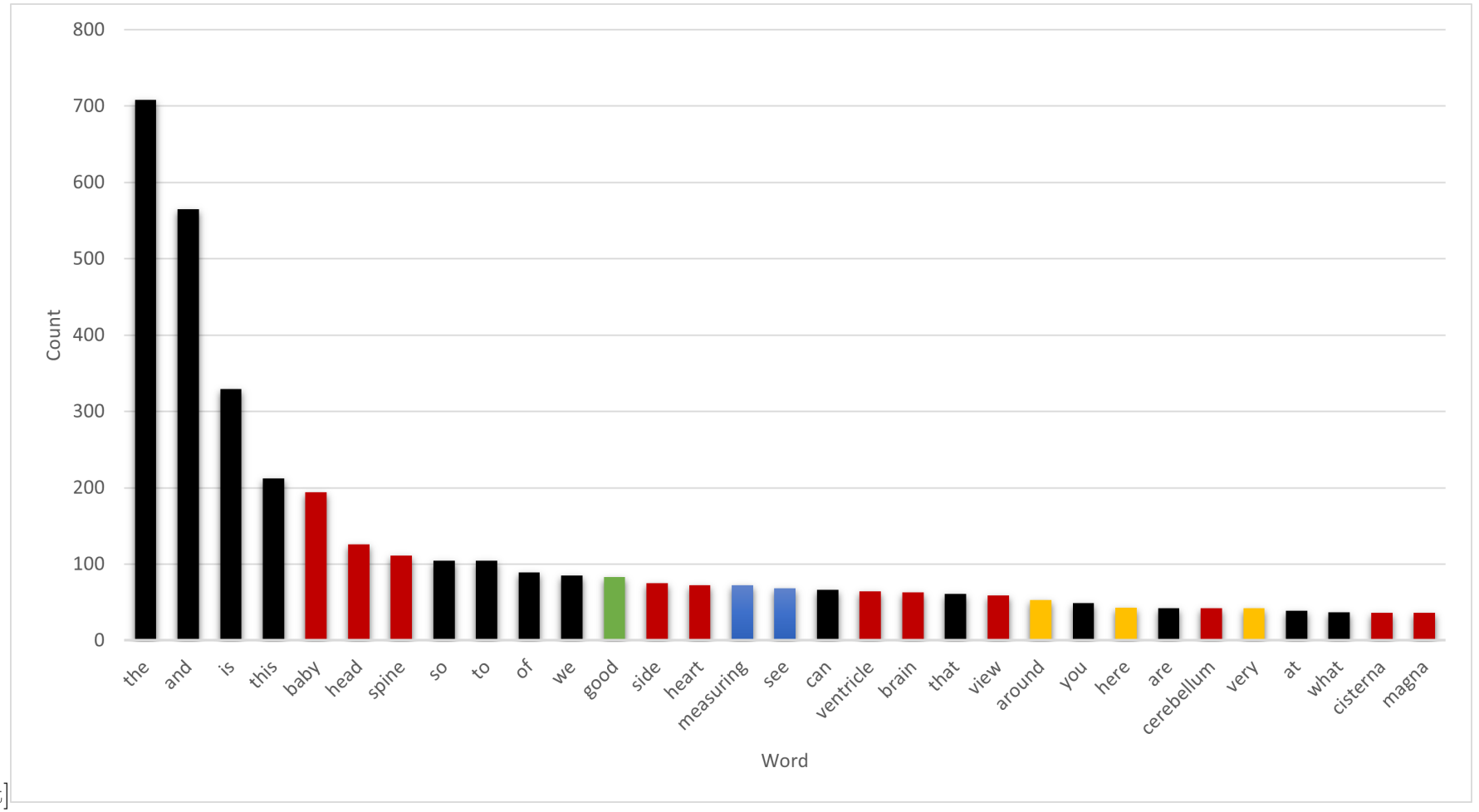
$$Richness = \frac{Number\ of\ Generated\ Relevant\ Words}{Number\ of\ Possible\ Relevant\ Words} \tag{6.3}$$

$$Efficiency = \frac{Number\ of\ Generated\ Relevant\ Words}{Total\ Number\ of\ Words} \tag{6.4}$$

6.3.3 Results and Discussion

Quantitative Evaluation

Leave-one-out cross-validation has been performed. Specifically, there are nine different runs, where video clips from eight different scans go into the training set and video clips from a ninth scan make up the validation set. A tenth completely different video that is not part of the cross-validation experiment is set aside to later be used to evaluate the trained models. Table 6.1 shows the scores obtained



[t]

Figure 6.7: A histogram of the most commonly occurring words in the training dataset. Red represents nouns. Green represents adjectives. Blue represents verbs (excluding forms of the verb 'to be'). Yellow represents adverbs. Black represents all other parts-of-speech.

with the three multi-modal configurations. GL, RG, PG are the gaze-less, real gaze, and the predicted gaze configurations respectively. $B1$ to $B4$, $R\text{-}L$, $F1$, ARS, and *Rich.* represent *BLEU-1* to *BLEU-4*, *ROUGE-L*, the *F1 score*, the *Antomical Relevance Score*, *Richness*, respectively, with standard deviations in brackets. For more information on the evaluation metrics, please check Chapter 3. The results shown in Table 6.1 are those after leave-one-out cross validation has been performed. The scores of every fold were then averaged before being shown here.

Table 6.3 compares results when using a more general word embedding model trained on a general corpus, word2vec, and when using a more biomedical-focused word embedding model trained on a corpus of a medical nature, BioWordVec (BioWV), with the RG configuration.

Discussion.

We observe that video clip caption generation assisted by sonographer gaze-tracking information improves results as Tables 6.1 and 6.2 show. The gaze-assisted configurations outperform the gaze-less configurations on efficiency, implying more inclusion of relevant terminology in generated text when incorporating gaze information in the captioning models. It is interesting to observe the high scores for the predicted gaze configuration obtained when compared to the real gaze configuration. We reason that this is so since with predicted gaze, all frames would have attention maps that they could leverage from. With real gaze, on the other, only frames with real sonographer gaze points will have an attention map generated for them. Some frames simply do not have any gaze points because the sonographer may have been looking at the subject or they were staring at a part of the screen that was not showing the anatomical content. With the predicted gaze assisted configuration, every single frame would have attention maps predicted for it, allowing it to leverage predicted gaze information for every frame when captioning. The predicted gaze map is a prediction of what the experts look at in a given image. The $BLEU$ scores are lower than what might be expected in other natural language processing tasks including image captioning [41, 62]. These results can be explained by the fact that only one

Table 6.1: Results of model configurations with varying levels of gaze involvement. The three rows show the results of: the gazeless (GL) model, the model that uses real gaze (RG) attention maps, and the model that uses predicted gaze (PG) saliency maps. The columns show the different evaluation metrics with which we compare the different model configurations. B1-B4, RL, and Rich. are more caption-focused evaluation metrics, while F1 and ARS are class-focused. More information on the different evaluation metrics can be found in Section 6.3.2 and the Appendix. The scores in bold in each column are the ones that are highest for that evaluation metric.

	B1	B2	B3	B4	R	F1	ARS	Eff	Rch
GL	0.27 (0.02)	0.12 (0.01)	0.07 (0.01)	0.03 (0.00)	0.43 (0.02)	0.77 (0.12)	**0.24** (0.05)	0.17 (0.01)	0.05 (0.00)
RG	0.28 (0.01)	0.10 (0.01)	0.05 (0.01)	0.02 (0.00)	0.44 (0.01)	0.81 (0.08)	0.18 (0.03)	0.19 (0.01)	0.05 (0.00)
PG	**0.33** (0.01)	**0.15** (0.01)	**0.08** (0.01)	**0.04** (0.00)	**0.48** (0.01)	**0.91** (0.03)	0.23 (0.03)	**0.25** (0.01)	**0.08** (0.00)

Table 6.2: Results of model configurations with varying levels of gaze involvement being shown for each anatomical structure specifically. The best score for the combination of a particular structure and a particular metric is in bold.

	Structure	B1	B2	B3	B4	R	F1	ARS	Eff	Rch
GL	Abdomen	0.27	0.08	**0.03**	**0.00**	0.41	0.68	**0.26**	0.12	0.05
	Head	0.27	0.15	0.11	0.06	0.40	0.77	0.30	0.17	0.06
	Heart	0.34	0.15	**0.09**	**0.06**	0.45	0.87	**0.21**	0.21	0.03
	Spine	0.21	0.09	0.04	0.02	0.48	0.76	0.19	0.20	0.07
RG	Abdomen	0.30	0.07	0.02	**0.00**	0.42	0.75	0.20	0.14	**0.06**
	Head	0.30	0.16	0.13	0.07	**0.46**	0.98	0.27	0.22	0.07
	Heart	0.33	0.12	0.05	0.02	0.46	0.80	0.12	0.23	0.03
	Spine	0.20	0.06	0.01	0.00	0.42	0.70	0.11	0.17	0.06
PG	Abdomen	**0.35**	**0.11**	0.00	**0.00**	**0.47**	**0.79**	0.15	**0.15**	**0.06**
	Head	**0.35**	**0.21**	**0.17**	**0.10**	0.44	**1.00**	**0.40**	**0.32**	**0.14**
	Heart	**0.36**	**0.16**	**0.09**	0.04	**0.51**	**0.92**	**0.21**	**0.29**	**0.05**
	Spine	**0.27**	**0.12**	**0.05**	**0.03**	**0.51**	**0.92**	**0.17**	**0.22**	**0.08**

Table 6.3: Results of Model Configurations with Different Word Embeddings

Config.	B1	B2	B3	B4	RL	Rich.	F1	ARS
Word2vec	**0.257**	**0.188**	**0.160**	**0.093**	35.320	**0.109**	**0.778**	**0.359**
BioWV	0.133	0.094	0.077	0.044	**44.20**	0.089	0.750	0.295

sonographer provided 'ground truth' captions to compare generated captions with. Many natural image captioning datasets, such as MS-COCO [28], Pascal VOC 2008 [52], Flickr8k [151], and Flickr30k [152], could have up to five different 'ground truth' captions for the same image [62]. We, on the other hand, only have one set of ground truth labels as a reference. This fact means that a relatively mediocre *BLEU* score (when compared to those obtained in other NLP tasks) which directly compares the degree of overlap between a generated caption and the reference is not enough to dismiss generated captions as erroneous since the *BLEU* scores only have a single reference to compare them with. We would like to highlight, however, how the incorporation of gaze improves the scores and qualitative performance.

The PG-trained model scores highest, but the RG-trained model does not. We attribute this result to the fact that with PG-trained models, the attention is focused on objects (or parts of objects), so the saliency in a head image would be depicted as around a head for example. This is in contrast to the input of the RG-trained models where the attention is localized to pixel regions based on gaze points. Another possible reason at play, as mentioned earlier, is that not all frames have real gaze points (from which to create real gaze attention maps), but all frames can have predicted gaze saliency maps generated for them. In Table 6.3, we compare word embeddings. It is interesting to note that BioWordVec underperformed when compared to word2vec in all but two metrics. This result is explainable by the fact BioWordVec might be too focused on biomedical text which does not constitute the majority of our our data. Our data, coming primarily from the natural speech of a medical professional, might not need embeddings of words and terms that would be used in medical reporting as spoken vocabulary can be distinct from written vocabulary. Word2vec might be more suitable for transcribed speech than BioWordVec. [148] also determined that using MeSH terms

with BioWordVec did not significantly help in NLP tasks involving clinical notes, but they might be effective with NLP tasks involving text from PubMed articles. We consider transcribed sonographer speech to be more similar in structure and content to clinical notes than to PubMed articles hence why we see our results as being in line with that of [148].

6.4 Conclusions

We presented a novel gaze-assisted natural language processing (NLP)-based video captioning method to describe routine second-trimester fetal ultrasound scan videos in a vocabulary of spoken sonography. The novelty of our multi-modal method is that the textual captions that describe the spatio-temporal scan video content are learnt from sonographer speech recordings and the generation of captions is assisted by sonographer gaze-tracking information reflecting their visual attention while performing live-imaging and interpreting a frozen image. As part of an ablation study, we compare spatio-temporal deep networks trained using three multi-modal configurations, namely; a gaze-less neural network, a neural network using real sonographer gaze, and a neural network using automatically predicted gaze. We evaluated the machine learning model architectures through established general text-based metrics (*BLEU, ROUGE-L, F1 score*) and domain-specific metrics that consider the richness of the generated captions with respect to the scan video. Results show that the proposed gaze-assisted models can generate richer and more diverse captions for clinical fetal ultrasound scan videos at the expense of the perceived sentence structure and that the generated captions are similar to sonographer speech in terms of discussing the visual content and the scanning actions performed. Qualitative examples are shown in Fig. 6.8

We have proposed an automatic video captioning method to describe spatio-temporal video content using words from the spoken vocabulary of professional sonographers in routine second-trimester fetal ultrasound scans. Utilizing gaze in captioning helps achieve higher scores on evaluation metrics, specifically *BLEU-1* to *BLEU-4*, *F1*, *ARS*. Predicted gaze has the added benefit of allowing all frames

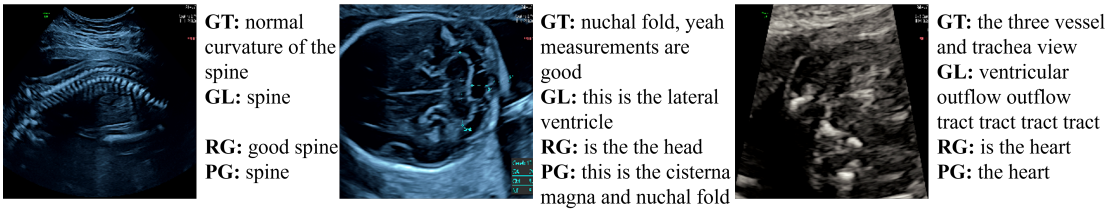


Figure 6.8: Qualitative results from a random fold are shown. GT is for Ground Truth as spoken by a sonographer. GL is for the gaze-less model configuration. RG is for the real gaze model configuration. PG is for the predicted gaze model configuration.

to have accompanying gaze information that could help in the captioning process, while real gaze, although it better reflects the important anatomical content, has the downside of being only available for those frames where the sonographers happened to be looking at the screen. Word2vec embedding outperformed BioWordVec in gaze-assisted captioning.

There is a single light of science and to brighten it anywhere is to brighten it everywhere.

— Isaac Asimov [1]

7 Conclusions and Future Work

Contents

7.1 Conclusions 147
7.2 Future Work 149
7.2.1 Video Navigation 149
7.2.2 Using Curriculum Learning with Simpler Models and Video Captioning and Paving the Way Towards Clinical Translation 150
7.2.3 Using Probe Motion Data in Captioning 150
7.2.4 Using Audio Directly in the Captioning Process 151
7.2.5 Using a Transformer-based Captioning Model 152

7.1 Conclusions

The research described in this thesis concerns using text as an input source along with ultrasound (US) visuals to train deep learning models in order to automatically caption ultrasound (US) images and video clips. We briefly summarise the main contributions of the thesis.

- **Chapter 3: Preparing a fetal ultrasound image captioning dataset from full-length second-trimester fetal ultrasound videos and text derived from accompanying expert voice-over audio recordings.** We began, in Chapter 3, by presenting the way we acquired audio in the real-world

clinical setting and the processing steps that converted this to text. This has been of great use to my work and the work of colleagues in the lab who are currently conducting their own experiments using this dataset which is the first of its kind. Chapter 4 used a smaller version of the annotated dataset. Later chapters included the same data samples as well as newly annotated ones, effectively doubling the size of the used dataset. In total, over 55,000 captions, covering four different anatomical structures (abdomen, head, heart, and spine) from ten of the audio recordings (of two different sonographers) were used in the experiments described in this thesis.

- **Chapter 4: Training fetal image captioning models by using the aforementioned dataset.** A paper on this contribution has been published [4]. In this chapter, we developed an approach that used a CNN as an image feature extractor and an RNN as either a text generator or a text feature extractor, allowing us to look at the fetal image captioning problem. The model grew in complexity and changed to improve its performance on scores in automatic captioning.

- **Chapter 5: Introducing the dual curriculum, and using it to train fetal ultrasound image captioning models. A variant of the dual curriculum, called the course-focused dual curriculum was also introduced where focus is shifted linearly from one complexity metric to the other throughout model training.** In Chapter 5, we investigated training image captioning models with different curricula built during each epoch with a changing weighting for each modality, allowing us to deal with the relatively small-sized nature of real datasets. We show that, in the fetal image captioning context, following a particular training protocol can improve performance when compared to training with randomly ordered training samples.

- **Chapter 5 Section 5.4: We proposed an augmentation method where pseudo-captions are created for caption-less images. These pseudo-captions can be used in training captioning models.**

- **Chapter 6: The fetal ultrasound image captioning models have been extended to video captioning and to make use of eye gaze information.** In this chapter, a convolutional LSTM-based model is used to caption ultrasound video clips that is assisted with either a real gaze attention map or a predicted gaze saliency map. In this chapter, we looked into captioning fetal ultrasound video clips and using gaze information in the captioning process. We have also looked into using embedding vectors that are built on biomedical vocabulary rather than general-use word embeddings in this chapter.

7.2 Future Work

There are a number of different ways to take the work in the thesis forward. Below, we list and discuss some of them.

7.2.1 Video Navigation

The usefulness of generating text has been described at length in Chapter 1 and throughout this thesis. The image captioning model could be further developed towards providing a handy tool to aid minimally trained clinical professionals and users of ultrasound, helping them learn what it is they see on an ultrasound system screen. On the other hand, being able to retrieve images or clips given some kind of text prompt would make for an efficient, convenient tool that can navigate and traverse hours of video to find video clips that are relevant to the text prompt. Although we do not explore this aspect in this thesis towards achieving the above, we are keen in the future to investigate how the weights of the models we trained could be transferred to this other task, allowing us to discover the full potential of the correspondence mapping first mentioned in Chapter 1.

Our total dataset consists of the speech of three sonographers; two are female native English speakers, and the other is a male non-native English speaker. Knowing this information, we can explore how different sonographers speak differently based on these parameters (level of English fluency, gender), and in what way there is consistency in sonographer speech, transcending the differences between sonographers.

7.2.2 Using Curriculum Learning with Simpler Models and Video Captioning and Paving the Way Towards Clinical Translation

In Chapter 5, we have explored variants of the dual curriculum for fetal image captioning. A text-first course-focused dual curriculum was found to perform best. In the future, we will investigate whether this conclusion holds true for the temporal equivalent of video clip captioning. In the future, we will also investigate if curriculum learning could help obtain comparable results while using a simpler model with less trainable parameters for real-time on-the-fly video captioning applications.

The current work is not quite ready for clinical translation; however, the use of simpler models would also be important to ensure a successful **clinical translation** in the future. Simpler models (with less parameters) make faster, real-time captioning possible, allowing for descriptions to be generated and shown while a scan is being performed or a video is being watched. As might be expected, the captioning models would need to be trained on more anatomical structures, not only the four ones that have been the focus of this thesis. The data for such training would need to be annotated and labelled as needed. These aforementioned steps would need to occur before clinical translation can happen.

7.2.3 Using Probe Motion Data in Captioning

The generated text could either be a sentence that describes the prevailing anatomical structure (such as 'we can see the heart beating nicely') or a caption that gives navigational guidance advice (such as 'calculating the abdominal circumference is like putting measuring tape around the baby's waist') to either acquire a particular kind of diagnostic image or or to advise on what to immediately do next with

the probe. This objective could be met by using a model akin to that of a Multi-Domain Network where earlier layers are common to different strands or domains but the later layers could be specific to certain domains [153]. For us, we could build a network inspired by the aforementioned one [153] but where the domains have to do with image navigation and anatomy recognition and understanding. Motion data could potentially be used in conjunction with text to build a model capable of describing the content of a video and how a professional sonographer performs an ultrasound scan.

7.2.4 Using Audio Directly in the Captioning Process

We have explored using another modality (gaze) in the captioning process in Chapter 6. Another modality that could be incorporated is the audio data that the text was transcribed from. Audio signals could be used as input along with the visual information. Audio augmentation techniques (such as altering the speed of speaking) discussed in Chapter 2 could potentially be important in this future task.

The speech that was transcribed for the work done in this thesis are the thoughts of the sonographer spoken out loud for our analysis and research. One reason we wanted to consider working on sonographer speech recognition was because of the great amount of work that has been done in speech recognition, which is only increasing in accuracy and use; however, we decided years ago to use Google Cloud Speech [8], but if we were to consider creating our own speech recognition model down the line, the following steps would include looking more thoroughly into the Listen-Attend-and-Spell (LAS) grapheme-based model to see how it can be adapted for our case [16]. Since our project will not involve words of everyday conversation between multiple people, the LAS model appears to be appropriate for our use despite its limitations. We would also need to compare it in depth with the readily available off-the-shelf commercial option of Google Cloud Speech [16]. Our earlier obtained audio contained only the voice of a single person, so the problem in [10] of multiple speakers was not a concern of ours in the beginning. Only a single sonographer talked over a pre-recorded video; and therefore, less pre-processing of

audio data was necessary when recording audio retrospectively. [10] also had the challenge of dealing with background noises, which we worked on minimizing by having the then retrospectively acquired speech sessions in controlled environments; although, we expected to still have the same difficulties associated with recognising speaker accents and technical niche terminology which would persist. In addition to the challenges posed in personal quirks in speech, some utterances are difficult to transcribe, and therefore, context plays a role to the extent that transcribers even add words that were not said at times in order to make a caption proper and a phrase complete [10, 16]. With regards to data augmentation, speed perturbation seems straightforward and does lead to an improvement as [9] shows; therefore, I would recommend applying it to increase the size of available data should we seek to develop and train our own audio speech recognition models; however, we did not do this in the thesis, as we were not looking to train audio recognition or transcription models. Live acquired audio is more important and relevant for this work. It has its challenges as the data chapter discusses, but it better reflects the state of mind of the sonographer during the scan and their thoughts on the objects they see on the screen.

7.2.5 Using a Transformer-based Captioning Model

Deep learning is a rapidly evolving area with new architectures appearing all the time, some becoming base architectures which last the test of time. One such approach which has appeared during the latter part of my thesis is the Transformer [67]. It would be interesting to consider how Transformers could be used in place of the architectures employed in this thesis. By way of early illustration, a transformer-based captioning model is shown in Fig. 7.1. Some preliminary qualitative results of the transformer-based captioning model are shown in Fig 7.2.

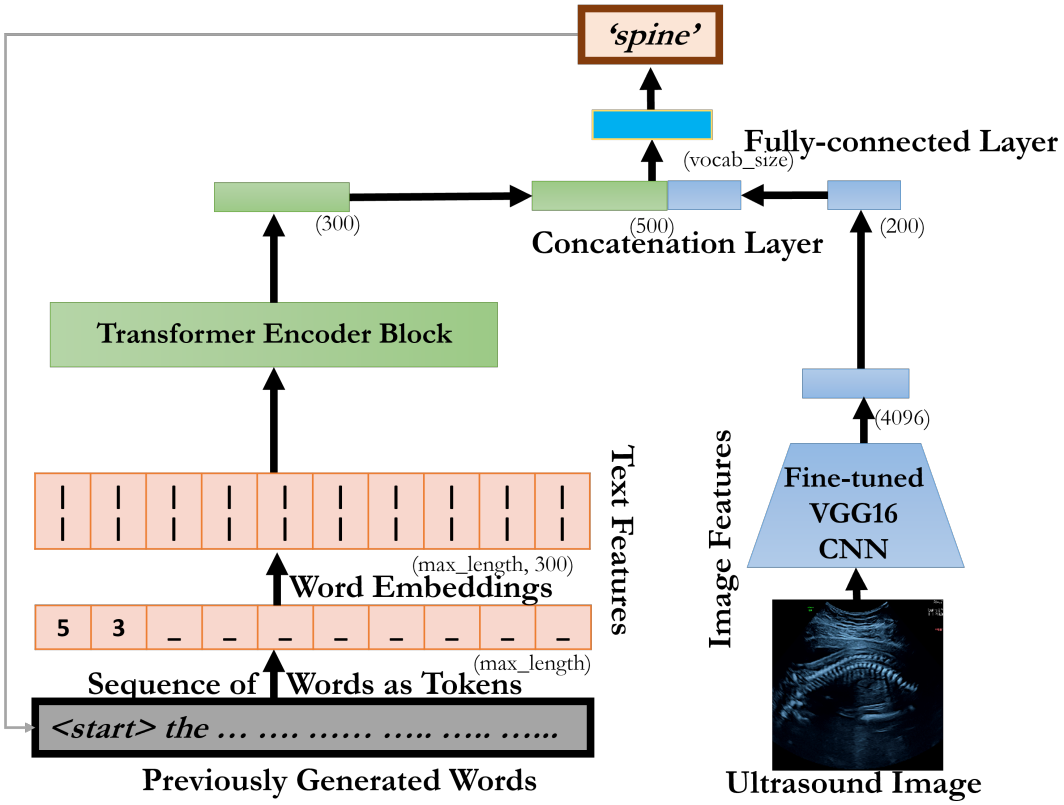


Figure 7.1: The image captioning model to be developed in the future. It follows the same skeleton as previously introduced captioning models, but with the the key difference being using transformer blocks in lieu of an LSTM-RNN.

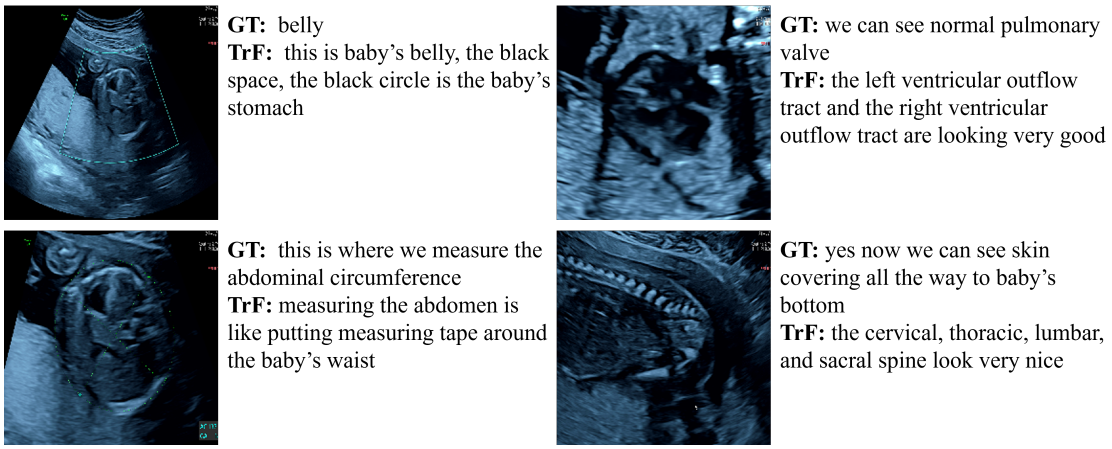


Figure 7.2: Some captions that have been generated by the transformer-based captioning model are shown. TrF stands for Transformer-based captioning model.

Appendices

True glory consists in doing what deserves to be written, in writing what deserves to be read.

— Pliny the Elder [1]

A Demo Scripts

Contents

A.1 "*_fetal" file 157
A.2 "*_terms" file 158

A.1 "*_fetal" file

The contents of the _terms file is as follows:

"Could you tell me your first and last names please? Great. Do you know your estimated date of delivery? This is great. Thank you. So I will start by making a quick survey of your foetus and I will show you in detail all the parts that we could see of your foetus' anatomy. So here we can see the baby's heart. You can hear it as well. This is the baby's belly, stomach, spine, this is the umbilical cord. Then now we are going to take a measurement of your foetus' belly, that's the foetus' abdomen. Ok great. This is the foetus' head, you can see really inside his brain. These would be called the ventricles of the brain, coiled Plexus, fluid within the brain, the skull. I am going to take a measurement of that as well. So the baby's facing down now, so his eyes are to your bottom. Ok. This is the baby's abdomen again. You can see now the umbilical vein, both kidneys, the bowel. Here is one

kidney and the other kidney, intestines, stomach again. This is the baby's spine and pelvis. Now he is changing positions. So it might be easier for us to see some other parts now. This is what we call the cerebellum, that's the small part of the brain, the small brain. You can see both ears. This is the lung. You can see there trachea, we can see the diaphragm which looks great. This is a spinal cord. You can see the baby's face, the nose, the chin, you know he is opening his mouth right now. Now you can see one hand and the second hand. Clearly we can count five fingers on this side and on the other side as well. Now I will show you, this is the umbilical cord and the insertion to the abdomen. Let's have a look at both the legs. So this is one leg, this is the left leg. We can see that the position of the leg is good. And this is the second leg, the right leg. We can see that the position of the sole is also good. I'm going to take a measurement of the foetus' long bone, it's the Femur. Great. Now I'm going to take measurements of all other bones. So I'm going to take their hands bones, their legs bones. I'm going to take all the measurements so as to make sure they are first of all, all formed and all well, and secondly that they have grown well till now. This all looks very good. So I'm very happy with the measurements. Now clearly you can say that this is a girl. The labia, yeah, you are going to have a girl. This is wonderful. Ok let's go to the foetus' heart. So, that the heart is looking right towards us right now. So we can see the heart very clearly. You can see the four chambers of the heart. You can see the blood flow. The blue represents blood flow which is going one way and the red the other way. So, red this is something, blood flow is coming towards me and red, the blue sorry, is going away from the —-You could see that both ventricles and both atria formed very well. I can see the out tract flow which is good. Ok, and since he is looking at us, take a look, we can see both eyes, we can even see the lenses and they are formed nicely and well. You can see the foetus' profile again. We can see both lips very good now. They have formed well as also. This is a part of the tongue bulging outside for a second now."

A.2 "*_terms" file

The contents of the _terms file is as follows:

"Head shape Cavum septum pellucidum Ventricular atrium Cerebellum Nuccal fold Coronal view Lips Nasal tip Situs Heart Four chambered Aorta Left ventricle Pulmonary artery Right ventricle Three vessel view Three vessel — Stomach Intrahepatic Umbilical vein Abdominal wall Cord insertion Diaphragm Kidneys Renal pelvis Bladder Vertebra Skin Cervical Thoracic Lumbar Sacral Limbs Femur Tibia Fibula Radius Ulna Humerus Metacarpal Placenta Amniotic fluid Uterine Head circumference Trans cerebellar diameter Abdominal circumference Rib Umbilical vein Sagittal plane Femur length Spine Lower spine Iliac bone Lateral Inferior vena cava Right atrium Left atrium Outflow tract Aortic valve Pulmonary valve Duct Superior vena cava That's it"

B
Extra Figures

Normal Stochastic Mini-Batch Training

		Abdomen	Head	Heart	Spine	Femur	Placenta
	Abdomen	0.84	0.04	0.00	0.07	0.00	0.05
	Head	0.00	0.99	0.00	0.00	0.00	0.00
	Heart	0.02	0.00	0.98	0.00	0.00	0.00
	Spine	0.01	0.03	0.00	0.96	0.00	0.00
	Femur	0.00	0.00	0.01	0.00	0.76	0.23
True Label	Placenta	0.52	0.31	0.00	0.00	0.00	0.18
		Abdomen	Head	Heart	Spine	Femur	Placenta
		Predicted Label					

Mahlanobis Distance Based Image Curriculum

		Abdomen	Head	Heart	Spine	Femur	Placenta
	Abdomen	0.78	0.09	0.00	0.07	0.00	0.05
	Head	0.00	0.99	0.00	0.01	0.00	0.00
	Heart	0.03	0.00	0.98	0.00	0.00	0.00
	Spine	0.00	0.04	0.00	0.96	0.00	0.00
	Femur	0.00	0.00	0.01	0.00	0.77	0.23
True Label	Placenta	0.53	0.25	0.00	0.00	0.00	0.19
		Abdomen	Head	Heart	Spine	Femur	Placenta
		Predicted Label					

Wasserstein Distance Based Image Curriculum

		Abdomen	Head	Heart	Spine	Femur	Placenta
	Abdomen	0.79	0.05	0.00	0.07	0.00	0.05
	Head	0.00	1.00	0.00	0.00	0.00	0.00
	Heart	0.03	0.00	0.97	0.00	0.00	0.00
	Spine	0.00	0.04	0.00	0.96	0.00	0.00
	Femur	0.00	0.00	0.01	0.00	0.77	0.23
True Label	Placenta	0.33	0.60	0.00	0.00	0.00	0.07
		Abdomen	Head	Heart	Spine	Femur	Placenta
		Predicted Label					

Figure B.1: A model trained with MD-IC or WD-IC is able to achieve comparable results within one single epoch of training to a model fully trained through the stochastic mini-batch training process.

There is no wealth like knowledge,
no poverty like ignorance.

— Ali ibn Abi Talib [1]

References

[1] 2K Games. *Sid Meier's Civilization IV*. [PC CD-ROM]. 2005.

[2] Xian-Hua Zeng, Bang-Gui Liu, and Meng Zhou. "Understanding and generating ultrasound image description". In: *Journal of Computer Science and Technology* 33.5 (2018), pp. 1086–1100.

[3] Xianhua Zeng et al. "Deep learning for ultrasound image caption generation based on object detection". In: *Neurocomputing* (2019).

[4] Mohammad Alsharid et al. "Captioning ultrasound images automatically". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 338–346.

[5] Mohammad Alsharid et al. "A curriculum learning based approach to captioning ultrasound images". In: *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*. Springer, 2020, pp. 75–84.

[6] Mohammad Alsharid et al. "A course-focused dual curriculum for image captioning". In: *International Symposium on Biomedical Imaging*. Springer, 2021, pp. 75–84.

[7] Adam Hart-Davis. *History: From the Ancient to the Modern World*. London: Dorling Kindersley Limited, 2016.

[8] Google Cloud Platform. *Speech API - Speech Recognition*. URL: `https://cloud.google.com/speech/` (visited on 12/11/2017).

[9] Tom Ko et al. "Audio augmentation for speech recognition". In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[10] Chung-Cheng Chiu et al. "Speech recognition for medical conversations". In: (2017). arXiv: `1711.07274`. URL: `http://arxiv.org/abs/1711.07274`.

[11] R Lippmann, Edward Martin, and D Paul. "Multi-style training for robust isolated-word speech recognition". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*. Vol. 12. IEEE. 1987, pp. 705–708.

[12] Navdeep Jaitly and Geoffrey E Hinton. "Vocal tract length perturbation (VTLP) improves speech recognition". In: *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*. Vol. 117. 2013.

[13] Anton Ragni et al. "Data augmentation for low resource languages". In: (2014).

[14] Li Lee and Richard Rose. "A frequency warping approach to speaker normalization". In: *IEEE Transactions on Speech and Audio Processing* 6.1 (1998), pp. 49–60.

[15] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. "Elastic spectral distortion for low resource speech recognition with deep neural networks". In: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE. 2013, pp. 309–314.

[16] Google Cloud. *Cloud Speech API Demo*. 2017. URL: https://www.youtube.com/watch?v=z8g3XM16eRM.

[17] SoX. *Sound eXchange*. URL: http://sox.sourceforge.net/.

[18] Werner Verhelst and Marc Roelands. "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech". In: *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. IEEE. 1993, pp. 554–557.

[19] Anton Ragni et al. "Data augmentation for low resource languages". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* September (2014), pp. 810–814.

[20] Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. "Elastic spectral distortion for low resource speech recognition with deep neural networks". In: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE. 2013, pp. 309–314.

[21] Oxford Dictionaries. *Definition of Phoneme in English*. URL: https://en.oxforddictionaries.com/definition/phoneme.

[22] Oxford Dictionaries. *Definition of Grapheme in English*. URL: https://en.oxforddictionaries.com/definition/grapheme.

[23] Jason Brownlee. *Gentle Introduction to Statistical Language Modeling and Neural Language Models*. 2017. URL: https://machinelearningmastery.com/statistical-language-modeling-and-neural-language-models/.

[24] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 07-12-June-2015 (2015), pp. 3156–3164. arXiv: 1411.4555.

[25] Kishore Papineni et al. "BLEU: A method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[26] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. "Im2text: Describing images using 1 million captioned photographs". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1143–1151.

[27] Girish Kulkarni et al. "Babytalk: Understanding and generating simple image descriptions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2891–2903.

[28] Tsung-Yi Lin et al. "Microsoft COCO: Common objects in context". In: *European Conference on Computer Vision*. Springer. 2014, pp. 740–755.

[29] Jacob Devlin et al. "Exploring nearest neighbor approaches for image captioning". In: *arXiv preprint arXiv:1505.04467* (2015).

[30] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2017), pp. 664–676. arXiv: 1412.2306.

[31] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. "Deep fragment embeddings for bidirectional image sentence mapping". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1889–1897.

[32] KantanMT. *Cloud-based Machine Translation Platform*. URL: https://www.kantanmt.com/whatisbleuscore.php.

[33] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text Summarization Branches Out*. 2004, pp. 74–81.

[34] Yutaka Sasaki. "The truth of the F–measure". In: *Manchester: School of Computer Science, University of Manchester* (2007).

[35] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4566–4575.

[36] Satanjeev Banerjee and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 2005, pp. 65–72.

[37] Raffaella Bernardi et al. "Automatic description generation from images: a survey of models, datasets, and evaluation measures". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2017, pp. 4970–4974.

[38] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. "What is the role of recurrent neural networks (RNNs) in an image caption generator?" In: *arXiv preprint arXiv:1708.02043* (2017).

[39] Quanzeng You et al. "Image captioning with semantic attention". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4651–4659.

[40] Desmond Elliott and Frank Keller. "Image description using visual dependency representations". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1292–1302.

[41] Raffaella Bernardi et al. "Automatic description generation from images: A survey of models, datasets, and evaluation measures". In: *Journal of Artificial Intelligence Research* 55 (2016), pp. 409–442.

[42] Imane Allaouzi et al. "Automatic caption generation for medical images". In: *Proceedings of the 3rd International Conference on Smart City Applications*. 2018, pp. 1–6.

[43] Margaret Mitchell et al. "Midge: Generating image descriptions from computer vision detections". In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 747–756.

[44] Siming Li et al. "Composing simple image descriptions using web-scale n-grams". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. 2011, pp. 220–228.

[45] Benjamin Z. Yao et al. “I2T: Image parsing to text description”. In: *Proceedings of the IEEE* 98.8 (2010), pp. 1485–1508.

[46] David Lyndon, Ashnil Kumar, and Jinman Kim. “Neural captioning for the ImageCLEF 2017 medical image challenges.” In: *CLEF (Working Notes)*. 2017.

[47] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. “Natural language description of human activities from video images based on concept hierarchy of actions”. In: *International Journal of Computer Vision* 50.2 (2002), pp. 171–184.

[48] Marcus Rohrbach et al. “Translating video content to natural language descriptions”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 433–440.

[49] Anna Rohrbach et al. “Coherent multi-sentence video description with variable level of detail”. In: *German conference on pattern recognition*. Springer. 2014, pp. 184–195.

[50] Sergio Guadarrama et al. “Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2712–2719.

[51] Ran Xu et al. “Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework.” In: *AAAI*. Vol. 5. Citeseer. 2015, p. 6.

[52] Ali Farhadi et al. “Every picture tells a story: Generating sentences from images”. In: *European Conference on Computer Vision*. Springer. 2010, pp. 15–29.

[53] Ankush Gupta, Yashaswi Verma, and CV Jawahar. “Choosing linguistics over vision to describe images”. In: *26th AAAI Conference on Artificial Intelligence*. 2012.

[54] Micah Hodosh, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899.

[55] Rebecca Mason and Eugene Charniak. “Domain-specific image captioning”. In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. 2014, pp. 11–20.

[56] Richard Socher et al. “Grounded compositional semantics for finding and describing images with sentences”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 207–218.

[57] Aude Oliva and Antonio Torralba. “Building the gist of a scene: The role of global image features in recognition”. In: *Progress in Brain Research* 155 (2006), pp. 23–36.

[58] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[59] Hao Fang et al. “From captions to visual concepts and back”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1473–1482.

[60] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. “Multimodal neural language models”. In: *International Conference on Machine Learning*. 2014, pp. 595–603.

[61] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. "Unifying visual-semantic embeddings with multimodal neural language models". In: *arXiv preprint arXiv:1411.2539* (2014).

[62] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3156–3164.

[63] Junhua Mao et al. "Deep captioning with multimodal recurrent neural networks (M-RNN)". In: *arXiv preprint arXiv:1412.6632* (2014).

[64] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International Conference on Machine Learning*. 2015, pp. 2048–2057.

[65] Xinxin Zhu et al. "Captioning transformer with stacked attention modules". In: *Applied Sciences* 8.5 (2018), p. 739.

[66] Yuxuan Xiong, Bo Du, and Pingkun Yan. "Reinforced transformer for medical image captioning". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2019, pp. 673–680.

[67] Ashish Vaswani et al. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017).

[68] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. "Where to put the image in an image caption generator". In: *Natural Language Engineering* 24.3 (2018), pp. 467–489.

[69] Andrej Karpathy and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3128–3137.

[70] Di Lu et al. "Entity-aware image caption generation". In: *arXiv preprint arXiv:1804.07889* (2018).

[71] Xiaoman Pan et al. "Unsupervised entity linking with abstract meaning representation". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2015, pp. 1130–1139.

[72] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. "Pragmatically Informative Image Captioning with Character-Level Inference". In: (2018). arXiv: 1804.05417. URL: http://arxiv.org/abs/1804.05417.

[73] Michael C Frank and Noah D Goodman. "Predicting pragmatic reasoning in language games". In: *Science* 336.6084 (2012), pp. 998–998.

[74] Noah D Goodman and Andreas Stuhlmüller. "Knowledge and implicature: Modeling language understanding as social cognition". In: *Topics in Cognitive Science* 5.1 (2013), pp. 173–184.

[75] Zuxuan Wu et al. "Deep learning for video classification and captioning". In: *Frontiers of Multimedia Research*. 2017, pp. 3–29.

[76] Yusuke Sugano and Andreas Bulling. "Seeing with humans: Gaze-assisted neural image captioning". In: *arXiv preprint arXiv:1608.05203* (2016).

[77] Yifan Cai et al. “SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1475–78.

[78] 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.

[79] T. Matiisen et al. “Teacher-student curriculum learning”. In: *IEEE transactions on Neural Networks and Learning Systems* (2019).

[80] Zhou Ren et al. “Deep reinforcement learning-based image captioning with embedding reward”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 290–298.

[81] Volkan Cirik, Eduard Hovy, and Louis-Philippe Morency. “Visualizing and understanding curriculum learning for long short-term memory networks”. In: *arXiv preprint arXiv:1611.06204* (2016).

[82] Yulia Tsvetkov et al. “Learning the curriculum with Bayesian optimization for task-specific word representation learning”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 130–139.

[83] Cao Liu et al. “Curriculum learning for natural answer generation”. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 2018, pp. 4223–4229.

[84] B. Park et al. “A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-PA X-ray screening for pulmonary abnormalities”. In: *Scientific Reports* 9.1 (2019), pp. 1–9.

[85] Ilkay Oksuz et al. “Automatic CNN-based detection of cardiac MR motion artefacts using k-space data augmentation and curriculum learning”. In: *Medical image analysis* 55 (2019), pp. 136–147.

[86] David Lyndon, Ashnil Kumar, and Jinman Kim. “Neural captioning for the ImageCLEF 2017 medical image challenges.” In: *CLEF (Working Notes)*. 2017.

[87] Department of Engineering Science. *PULSE*. 2019. URL: https://www.eng.ox.ac.uk/pulse/.

[88] Lior Drukker et al. “Transforming obstetric ultrasound into data science using eye tracking, voice recording, transducer motion and ultrasound video”. In: *Scientific Reports* 11.1 (2021), pp. 1–12.

[89] Nuance Communications. *Medical Speech Recognition*. 2021. URL: https://www.nuance.com/en-gb/healthcare/physician-and-clinical-speech/dragon-medical.html.

[90] Nish Parikh, Gyanit Singh, and Neel Sundaresan. “Chapter 20 - Query Suggestion with Large Scale Data”. In: *Handbook of Statistics*. Ed. by C.R. Rao and Venu Govindaraju. Vol. 31. Handbook of Statistics. Elsevier, 2013, pp. 493–518.

[91] NHS. *FASP Ultrasound Handbook*. 2015. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/443865/FASP_ultrasound_handbook_July_2015_090715.pdf.

[92] Harshita Sharma et al. "Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 987–990.

[93] Foteini Filippidou and Lefteris Moussiades. "A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer. 2020, pp. 73–82.

[94] Han Sloetjes and Peter Wittenburg. "Annotation by category-ELAN and ISO DCR". In: *6th International Conference on Language Resources and Evaluation (LREC 2008)*. 2008.

[95] Philip M McCarthy and Scott Jarvis. "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment". In: *Behavior Research Methods* 42.2 (2010), pp. 381–392.

[96] Google. *Google Cloud Speech-to-Text - Speech Recognition*. URL: `https://cloud.google.com/speech-to-text/`.

[97] 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.

[98] 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. 871–9.

[99] Yifan Cai et al. "Spatio-temporal visual attention modelling of standard biometry plane-finding navigation". In: *Medical Image Analysis* 65 (2020), p. 101762.

[100] Sam T Roweis. "One microphone source separation". In: *Advances in Neural Information Processing Systems*. 2001, pp. 793–799.

[101] Siouar Bensaid, Antony Schutz, and Dirk TM Slock. "Single microphone blind audio source separation using EM-Kalman filter and short+ long term AR modeling". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer. 2010, pp. 106–113.

[102] Laurent Girin, Sharon Gannot, and Xiaofei Li. "Audio source separation into the wild". In: *Multimodal Behavior Analysis in the Wild*. Elsevier, 2019, pp. 53–78.

[103] Integrated Wave Technologies, Inc. *Voxsort diarization: Who spoke when?* 2020. URL: `http://www.voice-sort.com/`.

[104] James Crook. *Audacity*. May 2021. URL: `https://www.audacityteam.org/`.

[105] *GrammarBot*. 2020. URL: `https://www.grammarbot.io/`.

[106] 2K Games. *Sid Meier's Civilization V*. [Online] Steam edition. 2010.

[107] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. "Where to put the image in an image caption generator". In: *Natural Language Engineering* 24.3 (2018), pp. 467–489.

[108] Google. *Google Code Archive - Long-term storage for Google Code Project Hosting*. URL: `https://code.google.com/archive/p/word2vec/`.

[109] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing*. 2014, pp. 1532–1543.

[110] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119.

[111] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[112] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[113] Ian Goodfellow et al. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.

[114] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[115] Yoshua Bengio et al. "Curriculum learning". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 41–48.

[116] Rasheed El-Bouri et al. "Hospital admission location prediction via deep interpretable networks for the year-round improvement of emergency patient care". In: *IEEE Journal of Biomedical and Health Informatics* (2020).

[117] R. El-Bouri et al. "Student-teacher curriculum learning via reinforcement learning: predicting hospital inpatient admission location". In: *arXiv preprint arXiv:2007.01135* (2020).

[118] Pietro Morerio et al. "Curriculum dropout". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3544–3552.

[119] Eugene L. Allgower and Kurt Georg. *Numerical Continuation Methods: An Introduction*. Vol. 13. Springer Science & Business Media, 2012.

[120] Prasanta Chandra Mahalanobis. "On the generalized distance in statistics". In: National Institute of Science of India. 1936.

[121] Victor M Panaretos and Yoav Zemel. "Statistical Aspects of Wasserstein Distances". In: *Annual Review Of Statistics And Its Application, Vol 6* 6.ARTICLE (2019), pp. 405–431.

[122] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272.

[123] "TF–IDF". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G.I. Webb. Boston, MA: Springer US, 2010, pp. 986–987.

[124] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[125] Lin Chin-Yew. "ROUGE: A package for automatic evaluation of summaries". In: *Proceedings of the Workshop on Text Summarization Branches Out*. 2004, pp. 56–60.

[126] *GrammarBot*. 2020. URL: `https://www.grammarbot.io/`.

[127] Yarin Gal and Zoubin Ghahramani. "A theoretically grounded application of dropout in recurrent neural networks". In: *arXiv preprint arXiv:1512.05287* (2015).

[128] Xinyi Wang et al. "Switchout: an efficient data augmentation algorithm for neural machine translation". In: *arXiv preprint arXiv:1808.07512* (2018).

[129] Yang Feng et al. "Unsupervised image captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4125–4134.

[130] Lisa Anne Hendricks et al. "Deep compositional captioning: Describing novel object categories without paired training data". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1–10.

[131] Tseng-Hung Chen et al. "Show, adapt and tell: Adversarial training of cross-domain image captioner". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 521–530.

[132] Wei Zhao et al. "Dual learning for cross-domain image captioning". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, pp. 29–38.

[133] Abhinav Gupta et al. "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 2012–2019.

[134] Niveda Krishnamoorthy et al. "Generating natural-language video descriptions using text-mined knowledge". In: *Proceedings of the Workshop on Vision and Natural Language Processing*. 2013, pp. 10–19.

[135] Jesse Thomason et al. "Integrating language and vision to generate natural language descriptions of videos in the wild". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 1218–1227.

[136] Jacob Devlin et al. "Language models for image captioning: The quirks and what works". In: *arXiv preprint arXiv:1505.01809* (2015).

[137] Andriy Burkov. "The hundred-page machine learning book". In: *The Hundred-Page Machine Learning Book*. 2019, pp. 100–101.

[138] *Textblob: Simplified Text Processing*. 2020. URL: `https://textblob.readthedocs.io/en/dev/`.

[139] Google Cloud. *Evaluating Models | AutoML Translation Documentation*. 2020. URL: `https://cloud.google.com/translate/automl/docs/evaluate`.

[140] Yifan Cai. "Deep learning sonographer visual attention". PhD thesis. Oxford, UK: University of Oxford, 2019.

[141] Pingbo Pan et al. "Hierarchical recurrent neural encoder for video representation with application to captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1029–1038.

[142] Shi Xingjian et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in Neural Information Processing Systems*. 2015, pp. 802–810.

[143] Marco Cuturi and Mathieu Blondel. "Soft-DTW: a differentiable loss function for time-series". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 894–903.

[144] Yoav Goldberg. "A primer on neural network models for natural language processing". In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420.

[145] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2980–2988.

[146] George A. Miller. "WordNet: A lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[147] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *The international conference on learning representations*. 2015.

[148] Yijia Zhang et al. "BioWordVec, improving biomedical word embeddings with subword information and MeSH". In: *Scientific Data* 6.1 (2019), pp. 1–9.

[149] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *arXiv preprint arXiv:1607.04606* (2016).

[150] Allen Kent et al. "Machine literature searching VIII. Operational criteria for designing information retrieval systems". In: *American Documentation* 6.2 (1955), pp. 93–101.

[151] Cyrus Rashtchian et al. "Collecting image annotations using Amazon's Mechanical Turk". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 2010, pp. 139–147.

[152] Peter Young et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 67–78.

[153] Hakan Bilen and Andrea Vedaldi. "Universal representations: The missing link between faces, text, planktons, and cat breeds". In: *arXiv preprint arXiv:1701.07275* (2017).