Efficient Ultrasound Image Analysis Models with Sonographer Gaze Assisted Distillation

Arijit Patra*, Yifan Cai*, Pierre Chatelain, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble

University of Oxford, Oxford OX3 7DQ, United Kingdom
arijit.patra@eng.ox.ac.uk; yifan.cai@eng.ox.ac.uk

Abstract. Recent automated medical image analysis methods have attained state-of-the-art performance but have relied on memory and compute-intensive deep learning models. Reducing model size without significant loss in performance metrics is crucial for time and memory-efficient automated image-based decision-making. Traditional deep learning based image analysis only uses expert knowledge in the form of manual annotations. Recently, there has been interest in introducing other forms of expert knowledge into deep learning architecture design. This is the approach considered in the paper where we propose to combine ultrasound video with point-of-gaze tracked for expert sonographers as they scan to train memory-efficient ultrasound image analysis models. Specifically we develop teacher-student knowledge transfer models for the exemplar task of frame classification for the fetal abdomen, head, and femur. The best performing memory-efficient models attain performance within 5% of conventional models that are 1000× larger in size.

Keywords: Model compression · Gaze tracking · Expert knowledge

1 Introduction

Current deep models for medical image analysis are recognized as having large memory footprints and inference costs, which are at odds with the increased focus on portability and low-resource usage [1]. While there have been studies on overparameterization of deep networks [2], efficient models have largely been defined empirically rather than using well-principled approaches. In this paper we explore efficient models using a combination of video and expert knowledge, defined by gaze tracking as a sonographer acquires an ultrasound (US) video. We propose a novel approach called Perception and Transfer for Reduced Architectures (PeTRA), a teacher-student knowledge transfer framework in which human expert knowledge is combined with ultrasound video frames as input to a large *teacher* model, whose output and intermediate feature maps are used to condition compact *student* models. We define a compact model as one that has a significantly reduced number of parameters and lower memory requirement compared to state-of-the-art models. Our objective is to achieve competitive accuracies with such compact models for our ultrasound image analysis task.

*Both authors contributed equally

Related Work. Model compression (or reduction) is a challenge in machine learning research due to both the interest in addressing over-parameterization [2] and for practical usage with reasonable computational resources. Model compression can be achieved through *pruning*, which consists in removing parameters based on feature importance [3]. However, pruning leads to compact models that are a sub-graph of the original model architecture, which unnecessarily constrains the architecture of the compact model. Knowledge transfer methods have been proposed that can transfer knowledge to an arbitrary compact model. Hinton et al. [4] introduce the concept of teacher-student knowledge distillation, which they define as a transfer of knowledge from the final layer of the large model to a compact model during the training of the latter. Romero et al. [5] extend the idea of knowledge transfer to include intermediate learnt feature maps in the training of the compact model as well. While model compression and teacher-student knowledge transfer have been studied in machine learning research, relatively few works deploy both concepts in ultrasound imaging settings despite research into ultrasound video understanding in terms of identification of standard fetal cardiac planes [6] and anatomy motion localisation [7] among others. Overcoming parameter redundancy is important to medical imaging as time required for diagnosis depends on model inference speeds, and memory footprint of algorithms come at the expense of storage space for other critical data. In a relevant study, [8] classify standard views in adult echocardiography by training traditional large deep learning models and use the method in [4] to train reduced versions of these models. In relation to using human knowledge in ultrasound video analysis, a related work concerns combining sonographer gaze and ultrasound video for fetal abdominal standard plane classification and gaze prediction [9]. Different from [8,9], we use a combination of distillation and intermediate feature adaptation along with human gaze priors for a fetal ultrasound anatomy classification task. Unlike [8], we do not use compact models derived from heavier models but those specifically proposed for low-compute situations.

Contributions. We propose a framework, Perception and Transfer for Reduced Architectures (PeTRA) which combines model knowledge transfer and expert knowledge cues. Our contributions are: 1) to train compact models using both final and intermediate knowledge distillation from large models for the exemplar task of anatomy classification of fetal abdomen, head, and femur frames from a free-hand fetal ultrasound sequence; 2) to incorporate sonographer knowledge in the form of gaze tracking data into a teacher model to enhance knowledge transfer. To our knowledge, this is the first attempt at model compression leveraging human visual attention with a teacher-student knowledge transfer approach.

2 Methods

Consider a K-class classification problem, which consists in finding the label $k \in [|1, K|]$ for an input $\mathbf{x}$. The output of a neural network can take the form $\mathbf{c} = \text{softmax}(\mathbf{z}) \in \mathbb{R}^K$, where $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^K$ is the raw output of the last

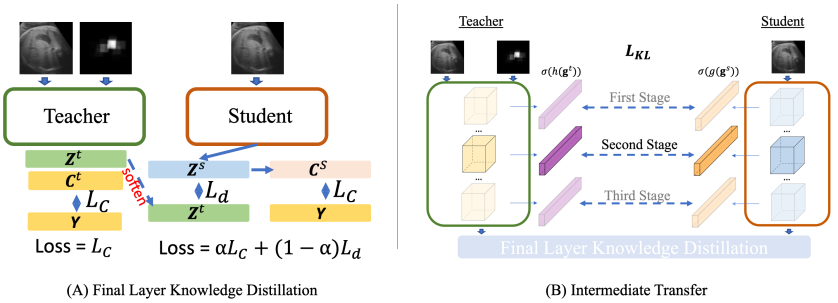


Fig. 1: Schematic of our proposed knowledge-distillation pipeline. (A) Final Layer knowledge distillation (B) Intermediate Transfer.

layer, or *logits*. We use a one-hot encoding for the classification target, such that, for a class $k \in [|1, K|]$, the corresponding target is $\mathbf{y} = (y_i)_1^K$ with $y_k = 1$ and $\forall i : i \neq k, y_i = 0$. The most commonly used loss function for multi-class classification is the categorical cross-entropy

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

2.1 Knowledge Transfer

Let $\mathcal{T}$ be a large *teacher* model and $\mathcal{S}$ a smaller *student* model. Model compression by knowledge transfer, first introduced in [10], consists in using the representations learnt by $\mathcal{T}$ to guide the training of $\mathcal{S}$ (Fig. 1). The key principle is that it is easier to learn using the representation of the teacher than it is to learn from the original input in the first place.

Final Layer Knowledge Distillation. Let $\mathbf{z}^t$ and $\mathbf{z}^s$ be the logits of the final layer of $\mathcal{T}$ and $\mathcal{S}$, respectively. Following [4], we first incorporate teacher knowledge by adding a distillation loss to the cross-entropy loss as:

$$L = \alpha L_c + \beta L_d \tag{2}$$

where L_c is the cross-entropy loss defined in Equation 1,

$$L_d = -\sum_{i=1}^{K} \text{softmax}\left(\frac{z_i^s}{E}\right) \log\left(\text{softmax}\left(\frac{z_i^t}{E}\right)\right) \tag{3}$$

is the distillation loss between teacher and student, and α, $\beta > 0$ are hyper-parameters controlling the relative influence of both terms. E is a *temperature* term introduced by [4] as a form of relaxation to *soften* $\mathbf{z}^t$ and $\mathbf{z}^s$. Indeed, having been obtained by a cross-entropy objective in $\mathcal{T}$, $\mathbf{z}^t$ may be too close to the one-hot target vector $\mathbf{y}$. Softening provides more information about the relative similarity of classes rather than absolute maxima.

Intermediate Transfer (IT). To leverage knowledge contained in intermediate representations of the teacher model, we consider intermediate knowledge transfer, or *hint learning* [5], in conjunction with final layer knowledge distillation. Let $\mathbf{g}^t$ be the output of an intermediate layer of the teacher model. It is used to produce a *hint* $\sigma(h(\mathbf{g}^t))$, where σ is a sigmoid activation function and h is a fully-connected (FC) layer. Similarly, an intermediate layer of the student model (a *guided* layer) is used to produce a regularization output $\sigma(g(\mathbf{g}^s))$, where g is a FC layer with the same output dimension as h. The hint is used to train the guided layer with a Kullback-Leibler (KL) loss

$$L_{\mathrm{KL}} = -\sum_j \sigma(g(\mathbf{g^s}))_j \log\left(\frac{\sigma(h(\mathbf{g^t}))_j}{\sigma(g(\mathbf{g^s}))_j}\right) \tag{4}$$

This creates a teacher model FC layer or *arm* (in purple, Fig. 1(B)) whose logits are associated with the student model FC *arm* (in orange, Fig. 1(B))in a KL divergence objective aimed at optimizing learned intermediate representations in the student model by supervising them with corresponding teacher model values (Fig. 1). Intermediate transfer essentially implements a regularization of the student learning using the most attentive intermediate features from the teacher. It is added to the optimization objective in Equation 2 in training:

$$L = \alpha L_c + \beta L_d + \gamma L_{\mathrm{KL}}, \tag{5}$$

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

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

2.2 Learning from Human Knowledge

We model the visual attention of a human expert looking at an image $\mathbf{I}$ through a gaze map $\mathbf{G}$. $\mathbf{G}$ is generated by recording the point-of-gaze of the human expert while looking at $\mathbf{I}$. To perform gaze-assisted knowledge distillation, we train the *teacher* model $\mathcal{T}$ to perform a classification task using both $\mathbf{I}$ and $\mathbf{G}$ as input. The student models still only "sees" the image $\mathbf{I}$. Thus, the teacher model can transfer not only the knowledge learned through its high number of parameters, but also knowledge extracted from the human visual attention. We test two different architectures for learning from image and gaze: $\mathcal{T}_{+\mathrm{gaze}}$ obtained by concatenation of extracted features of inputs (frame and gaze map) and $\mathcal{T}_{\times\mathrm{gaze}}$ by computing the element-wise product between resized gaze maps (28×28) and feature maps extracted from US frames (Fig. 2).

2.3 Data and Training Details

Data. Clinical fetal ultrasound videos with simultaneously recorded sonographer gaze tracking data was available from the PULSE study [11] . Ethics approval was

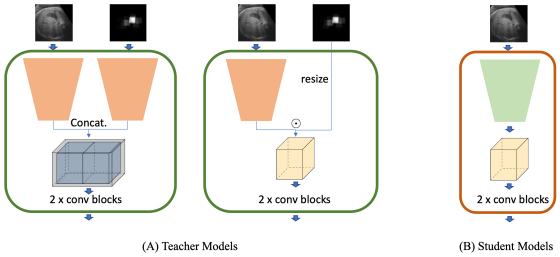


Fig. 2: Teacher and Student models used. (A) Teachers use concatenation or element-wise production to merge information from US image and visual attention map. (B) student only takes US image as input

obtained for data recording and data stored as per local data governance rules. From this dataset we extracted 23016 abdomen, 24508 head, 12839 femur frames. Gaze tracking data was recorded using a Tobii Eye Tracker 4C (Tobii, Sweden) that records the point-of-gaze (relative x and y coordinates with corresponding timestamp) at a rate of 90 Hz, effectively recording 3 gaze points per frame. Gaze points less than 0.5° apart were merged as a single fixation point. A sonographer visual attention map G was generated for each image by adding a truncated Gaussian with width corresponding to a visual angle of 0.5° at the point of fixation.

Training Details. We tested different student models to demonstrate the utility of the PeTRA approach: *SqueezeNet* [12] (**S**), *MobileNet* (0.25 width multiplier) [13] (**M**), and *MobileNet* v2 (0.35 width multiplier) [14] (**MV**) modified to accept single channel inputs and include a joint loss objective in Equation 5. These models are representative of the main types of compact architectures — squeeze-excite convolution blocks [12], group convolutions [13] and depthwise separable convolutions in groups [14]. Most other compact models proposed in computer vision literature derive from these basic architectures. For the teacher models we use a *VGG-16* feature extractor, modified to accept dual input of single-channel frames and gaze maps (with the depth of the first two fully connected layers changed from the original 4096 to 1024 and 512 to avoid overfitting). In a change to the standard VGG-16, for $\mathcal{T}_{\times\text{gaze}}$, the element-wise product after the fourth convolutional block is followed by the FC layers.In $\mathcal{T}_{+\text{gaze}}$, features are extracted by parallel convolutional blocks of the VGG16 and concatenated before FC layers. At inference, only one of the parallel blocks (processing single frame input, as gaze maps are not used at inference) and the following FC layers comprise the $\mathcal{T}_{\times\text{gaze}}$ model. This reflects in $\mathcal{T}_{\times\text{gaze}}$ having same number of parameters as $\mathcal{T}$ in Table 1. Data augmentation was performed using a 20 degrees rotational augmentation and horizontal flipping for both ultrasound and gaze map frames. Frames and corresponding gaze maps were resized to 224×224. All models were trained on 80 % (71 subjects) and tested on 20 % (18 subjects) of the dataset. Teacher models were trained for 100 epochs with learning rate

Table 1: Performance of MobileNetV2 (MV) with different configurations of knowledge distillation. IT indicates the level of intermediate transfer, if any.

Configuration			Validation accuracy				NetScore
Student	Teacher	IT	Abdomen	Head	Femur	Average	
MV	$\mathcal{T}$	–	0.63	0.67	0.69	0.66	60.16
$MV_{+\text{gaze}}$	$\mathcal{T}_{+\text{gaze}}$	–	0.71	0.73	0.68	0.71	61.43
$MV^1_{+\text{gaze}}$	$\mathcal{T}_{+\text{gaze}}$	1	0.73	0.74	0.70	0.72	61.67
$MV^2_{+\text{gaze}}$	$\mathcal{T}_{+\text{gaze}}$	2	**0.78**	**0.78**	0.76	0.77	62.84
$MV^3_{+\text{gaze}}$	$\mathcal{T}_{+\text{gaze}}$	3	**0.78**	0.77	**0.80**	**0.78**	63.06
$MV_{\times\text{gaze}}$	$\mathcal{T}_{\times\text{gaze}}$	–	0.84	0.84	0.79	0.82	63.93
$MV^1_{\times\text{gaze}}$	$\mathcal{T}_{\times\text{gaze}}$	1	0.80	0.83	0.79	0.81	63.72
$MV^2_{\times\text{gaze}}$	$\mathcal{T}_{\times\text{gaze}}$	2	0.86	**0.85**	0.83	**0.85**	**64.56**
$MV^3_{\times\text{gaze}}$	$\mathcal{T}_{\times\text{gaze}}$	3	**0.87**	**0.85**	**0.84**	**0.85**	**64.56**

of 0.005 and adaptive moment estimation (Adam) [15]. Students models were trained for 200 epochs over the *(N, image, label, logit)* set created for all N frames passed to the teacher model. The softening temperature value was set to 4.0 after a grid search for $E \in [|1, 10|]$. We investigated intermediate transfer at three different stages. First, second and third stage intermediate transfer was respectively applied from the 2nd, 4th, 5th maxpool layers in the teacher model to the FC arms after 2nd, 3rd, 5th maxpool layers of S and 3rd, 5th, 7th depthwise conv layer for M and MV. For experiments with intermediate transfer, such FC layer neurons were separately retained and appended to the set as *(N, image, label, logit,* IT_1*/../*IT_m*)*. α and β are set to 0.5 for equal influence of teacher knowledge and cross-entropy loss for the student model; γ is set at 1.

3 Results and Discussion

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

Performance. Final layer knowledge distillation improves the accuracy of the compact model compared to training without knowledge transfer for all students. Compact models trained using gaze-assisted knowledge distillation reach a higher accuracy than the same models trained with image-only knowledge distillation (+0.05 for $MV_{+\text{gaze}}$, +0.16 for $MV_{\times\text{gaze}}$, compared to MV). Intermediate transfer further improves knowledge transfer over final layer distillation only,

Table 2: Performance of teachers $\mathcal{T}_{+\text{gaze}}/\mathcal{T}_{\times\text{gaze}}$, student models (trained directly without teacher) and compared methods ($\mathcal{T}$ is teacher w/o gaze).No. of parameters are those in models used for inference.

Model						Validation accuracy			
Name	#parameters	Size(MB)	Time(ms)	MFLOP	NetScore	Abd.	Head	Femur	Avg.
$\mathcal{T}$	55,282,178	221.24	336.23	110.55	36.67	0.76	0.74	0.69	0.73
$\mathcal{T}_{+\text{gaze}}$	55,282,178	221.24	336.24	110.55	38.04	0.85	0.75	0.76	0.79
$\mathcal{T}_{\times\text{gaze}}$	213,320,002	884.96	637.13	464.31	28.21	**0.92**	**0.90**	**0.87**	**0.90**
S_{direct}	619,644	0.22	127.43	82.65	51.21	0.48	0.53	0.52	0.51
M_{direct}	738,658	0.27	159.28	98.71	52.50	0.56	0.61	0.64	0.60
MV_{direct}	**284,850**	**0.12**	**79.64**	**64.20**	**59.63**	0.57	0.67	0.68	0.64



Fig. 3: Performance-size trade-off. Left: $\mathcal{T}_{+\text{gaze}}$-trained students; Right: $\mathcal{T}_{\times\text{gaze}}$-trained students (enlarged in Appendix). Accuracy is averaged across classes.

with transfers at 3^{rd} level showing the best improvement in student model accuracy (+0.07 for $MV^3_{+\text{gaze}}$, +0.03 for $MV^3_{\times\text{gaze}}$, compared to $MV_{+\text{gaze}}$, $MV_{\times\text{gaze}}$). These trends are seen for all student models (S, M, MV).The baseline image-only knowledge distillation is analogous to prior work in [4] and [8].

Computational Complexity. We evaluated the computational complexity of the models by computing the number of floating point operations (FLOP) performed for inference (Table 2). We also report inference times for a batch of 100 frames from the test set using a 32GB/Intel core i7-4940MX/3.1Ghz laptop. The inference speed-up compared to $\mathcal{T}_{\times\text{gaze}}$ is 5× for S, 8× for MV and 4× for M (Table 2). For the same student models, using human gaze in teacher training does not change the computational complexity, but the performance metrics of student models when distilled from gaze-trained teachers are superior (Table 1).

Memory size. Student architectures (S, M, MV) show a 1000× to 7000× reduction of memory size compared to teachers (Table 2). The $MV^3_{+\text{gaze}}$ student with only 284,850 parameters achieves an average accuracy of 0.85, close to its teacher model $\mathcal{T}_{\times\text{gaze}}$ (0.90), and higher than $\mathcal{T}_{+\text{gaze}}$ (0.79) and $\mathcal{T}$ without gaze (0.73). Similar gains are seen for other students as well (Fig. 3). The MobileNet model M (738,658 parameters, 270 kB) trained with distillation from $\mathcal{T}_{\times\text{gaze}}$

attains accuracy (0.79) comparable to $\mathcal{T}_{+\text{gaze}}$ and higher than $\mathcal{T}$. Due to element-wise product operations, $\mathcal{T}_{\times\text{gaze}}$ has a higher number of parameters than $\mathcal{T}_{+\text{gaze}}$.

Model efficiency. To evaluate model efficiency as a trade-off between accuracy a, number of parameters p and computational cost c, we estimated the *NetScore* metric $\Omega = 20 \log\left(a^{\delta} p^{-\epsilon} c^{-\phi}\right)$ proposed in [16]. We provide other model data in Table 2 for completeness. Based on [16] we set $\delta = 2$, $\epsilon = 0.5$ and $\phi = 0.5$. For computational cost c, we use the number of FLOP instead of multiply-accumulate (MAC) operations in [16] because FLOP includes overheads such as pooling and activation beyond dot product and convolution operations. We report MFLOP (million FLOP) and NetScore values in Table 2. The units of a, p, c in Ω are percent, millions of parameters and MFLOP. The best NetScore is obtained by $MV^3_{\times\text{gaze}}$, the most compact model with the highest accuracy.

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

4 Conclusions

We proposed Perception and Transfer for Reduced Architectures as a general framework to train compact models with knowledge transfer from traditional large deep learning models using gaze tracking information to condition the solution without requiring such information at runtime. For the tasks of fetal abdomen, femur and head detection, compact model had an accuracy close to that of the large models, while having a much lower memory requirement. We found intermediate knowledge transfer to be more efficient when applied deeper in the networks. This is a proof-of-concept of human knowledge-assisted model compression for image analysis and the concept could be used for other modalities.

Acknowledgements We acknowledge the ERC (ERC-ADG-2015 694581, project PULSE), the EPSRC (EP/GO36861/1, EP/MO13774/1, EP/R013853/1), the Rhodes Trust, and the NIHR Biomedical Research Centre funding scheme.

References

1. Becker, D.M., et al.: The use of portable ultrasound devices in low-and middle-income countries: a systematic review of the literature. Tropical Medicine & International Health **21**(3) (2016) 294–311
2. Liu, B., et al.: Sparse convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 806–814
3. He, Y., et al.: Channel pruning for accelerating very deep neural networks. In: Proc. IEEE International Conference on Computer Vision. (2017) 1389–1397
4. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS 2014 Deep Learning Workshop. (2014)
5. Romero, A., et al.: FitNets: Hints for thin deep nets. arXiv:1412.6550 (2014)

6. Patra, A., et al.: Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer (2017) 276–284
7. Patra, A., et al.: Sequential anatomy localization in fetal echocardiography videos. arXiv preprint arXiv:1810.11868 (2018)
8. Vaseli, H., et al.: Designing lightweight deep learning models for echocardiography view classification. In: SPIE Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling. Volume 10951.
9. Cai, Y., et al.: SonoEyeNet: Standardized fetal ultrasound plane detection informed by eye tracking. In: 15th IEEE ISBI, IEEE (2018) 1475–1478
10. Buciluǎ, C., et al.: Model compression. In: Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2006) 535–541
11. PULSE: Perception ultrasound by learning sonographic experience, www.eng.ox.ac.uk/pulse. (2018)
12. Iandola, F.N., et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv preprint arXiv:1602.07360 (2016)
13. Howard, A.G., et al.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Sandler, M., et al.: MobileNetV2:inverted residuals and linear bottlenecks. In: CVPR. (2018)
15. Kingma, D.P., Adam, J.B.: A method for stochastic optimization. arXiv:1412.6980
16. Wong, A.: NetScore: Towards universal metrics for large-scale performance analysis of deep neural networks for practical usage. arXiv:1806.05512 (2018)