Designing Lightweight Deep Learning Models for Echocardiography View Classification

Hooman Vaseli*[1], Zhibin Liao*[1], Amir H. Abdi*[1], Hany Girgis*[1,2], Delaram Behnami[1], Christina Luong[2], Fatemeh Taheri Dezaki[1], Neeraj Dhungel[1], Robert Rohling[1], Ken Gin[1, 2], Purang Abolmaesumi†[1], and Teresa Tsang†[1,2]

[1]University of British Columbia
[2]Vancouver General Hospital, Vancouver, BC, Canada.

ABSTRACT

Transthoracic echocardiography (echo) is the most common imaging modality for diagnosis of cardiac conditions. Echo is acquired from a multitude of views, each of which distinctly highlights specific regions of the heart anatomy. In this paper, we present an approach based on knowledge distillation to obtain a highly accurate lightweight deep learning model for classification of 12 standard echocardiography views. The knowledge of several deep learning architectures based on the three common state-of-the-art architectures, VGG-16, DenseNet, and Resnet, are distilled to train a set of lightweight models. Networks were developed and evaluated using a dataset of 16,612 echo cines obtained from 3,151 unique patients across several ultrasound imaging machines. The best accuracy of 89.0% is achieved by an ensemble of the three very deep models while we show an ensemble of lightweight models has a comparable accuracy of 88.1%. The lightweight models have approximately 1% of the very deep model parameters and are six times faster in run-time. Such lightweight view classification models could be used to build fast mobile applications for real-time point-of-care ultrasound diagnosis.

Keywords: Echocardiography, View classification, View detection, Convolutional neural networks, Deep learning, Knowledge distillation.

1. INTRODUCTION

Transthoracic echocardiography (echo) is the most widely used imaging modality of cardiac assessment in which ultrasound data is acquired from standard cross sections (views) of the heart to study a variety of cardiac structures and functions. Echo acquisition is a manual procedure, where the clinician moves the imaging probe over several chest acoustic windows and fixates on the cross section of his or her choosing. Reliable interpretation of the data depends on its correct acquisition, requiring expertise and experience. Additionally, as the use of point-of-care ultrasound (POCUS) by non-cardiologists increases in the acute care specialties, such as anesthesiology,[7] automatic detection of echo views becomes even more important in speeding up echo acquisition and benefiting novice ultrasound users.

In the past decade, many efforts[1,3,8,10–12,15,16] have been made in automatic classification of echocardiography views, some of which have leveraged the deep learning models. However, most of these studies have not managed to tackle all the standard echo views. Moreover, feasibility of their approaches have not been tested over large-scale, diverse dataset. As a result, they likely do not meet the clinical needs of neither a cardiology department nor the POCUS use for acute care specialties. Furthermore, to facilitate adapting a deep learning approach in POCUS systems, it must be feasible to run view classification models in real time on mobile computation platforms (*e.g.*, Android phone) with limited computation power.Therefore, a more lightweight deep learning model is much preferred.

* Joint first authors
† Joint senior authors
Corresponding author: P. Abolmaesumi, purang@ece.ubc.ca, Tel: +1 604 827 4741

Table 1: Data distribution in the collected echocardiography dataset.

Window	Apical				Parasternal					Subcostal		Suprasternal
View	A2C	A3C	A4C	A5C	PLAX	RVIF	PSAX-A	PSAX-M	PSAX_PM	S4C	IVC	SUPRA
Cines	1,928	2,094	2,165	541	2,745	373	2,126	2,264	823	759	718	76
Frames	93,812	99,074	106,033	23,431	128,619	17,110	107,389	105,240	33,526	32,509	57,817	3,348

In this article, we present a lightweight deep learning model for echo view classification across 12 standard views from four major acoustic windows. Having comparable performance to best available deep learning architectures, our lightweight model has the size of less than 1% of the deep networks and runs six times faster, so it is feasible to be deployed on a mobile phone.

Our contributions in this paper are as follows: 1) we adapt the knowledge distillation method for training lightweight deep learning models in echocardiography view recognition domain; 2) we collect a large-scale echocardiography dataset of 3,151 unique patients to validate the training method; and 3) we show an ensemble of student networks, each with a different network structure, can maximize the knowledge distilled from a very deep network.

2. DATASET

The data used in this work was fetched from Vancouver General Hospital's echo database with approval from the Clinical Medical Research Ethics Board and consultation with the Information Privacy Office. The dataset was extracted randomly from the hospital server, including 3,151 unique patients who were diagnosed with various heart conditions and diseases during the time period of 2011 to 2015. In general, the dataset contains 16,612 echo cines (with a total of 807,908 frames) from 12 standard views taken from the four standard imaging windows, namely, *Apical*, *Parasternal*, *Subcostal*, and *Suprasternal*. The distribution of the data per class is shown in Table 1, and an example of dataset images (one per view class) can be found in Fig. 1. The data was originally acquired using various ultrasound machine models manufactured by *Philips*, *GE*, and *Siemens*. A senior cardiologist manually identified the view class associated with each cine series for the entire dataset.

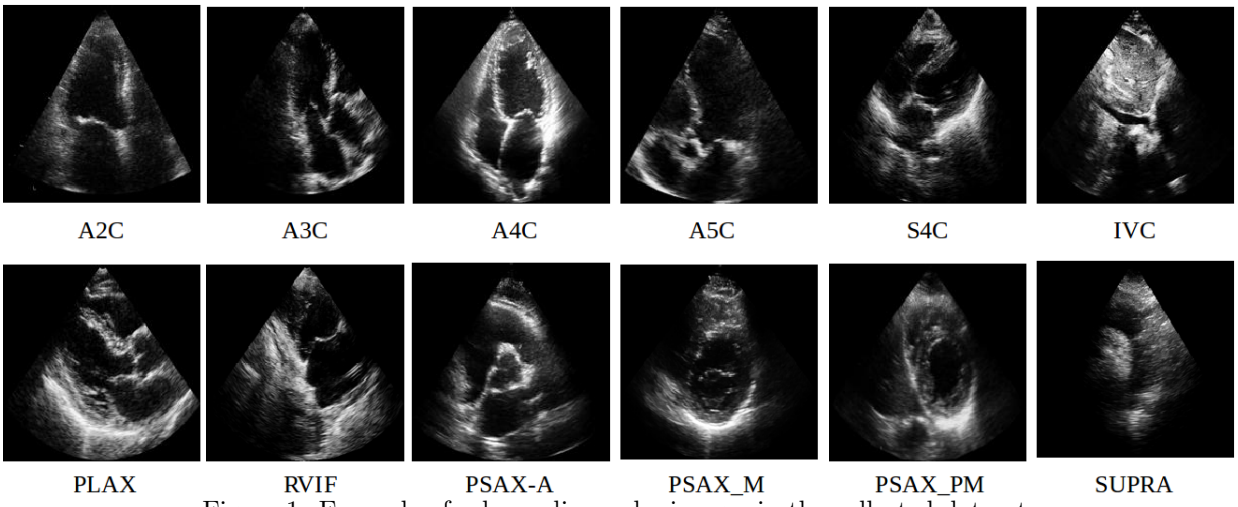
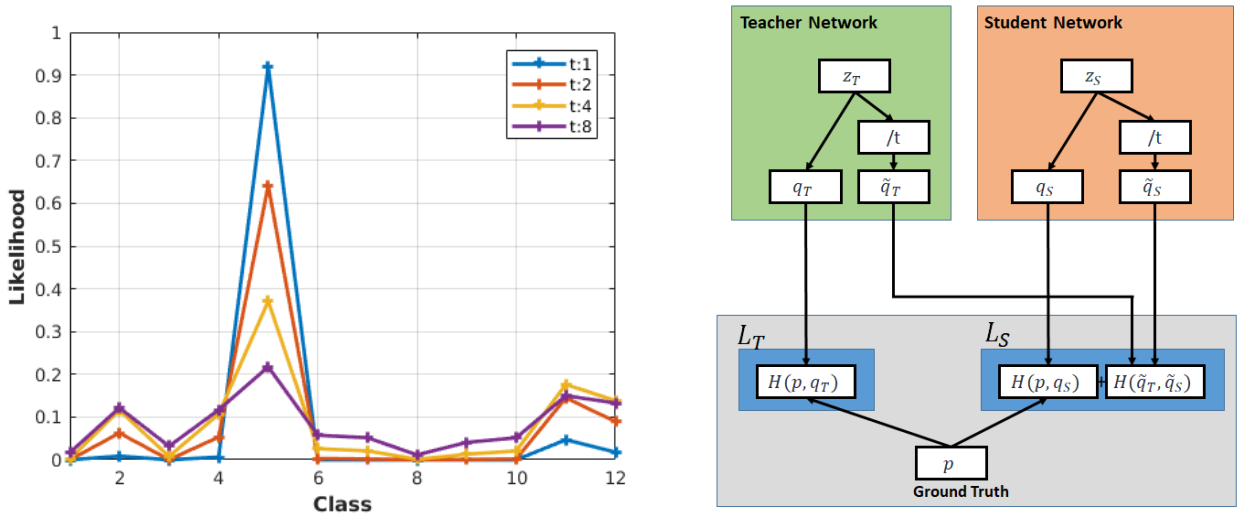


Figure 1: Example of echocardiography images in the collected dataset.

3. METHODOLOGY

3.1 Knowledge Distillation

Hinton *et al.*[5] introduced a compression method to transfer the learned knowledge of a *teacher network* to a *student network*. The main concept of this method is to train the student network with the output of the teacher



(a) (b)

Figure 2: (a) An example of the likelihood distribution softened with temperatures ($t = 1, 2, 4, 8$). (b) Tempered architecture and loss calculation for student network, where $\mathbf{q}_T$ and $\mathbf{q}_S$ are noted as the likelihood distributions when $t = 1$. The $[/t]$ block indicates the division-by-t operation.

network. In a classification task, let T and S be the teacher network and student network, respectively, for the same input $\mathbf{x}$. The pre-softmax logits $\mathbf{z}_T$ and $\mathbf{z}_S$ are tempered with the same temperature value t as

$$\tilde{\mathbf{q}}_T = \text{softmax}(\frac{\mathbf{z}_T}{t}), \tag{1}$$

$$\tilde{\mathbf{q}}_S = \text{softmax}(\frac{\mathbf{z}_S}{t}), \tag{2}$$

where $\text{softmax}(\mathbf{x}) = \frac{\exp(\mathbf{x}_j)}{\sum_i^{|\mathbf{x}|} \exp(\mathbf{x}_j)}$. The student network is trained to match its tempered output $\tilde{\mathbf{q}}_S$ with the tempered softmax prediction $\tilde{\mathbf{q}}_T$ of the teacher network and match the $\mathbf{q}_S$ (denote $\tilde{\mathbf{q}}_S$ for t=1) with the ground truth $\mathbf{p}$ for $\mathbf{x}$,

$$\mathcal{L}_S = \mathcal{H}(\tilde{\mathbf{q}}_T, \tilde{\mathbf{q}}_S) + \alpha \mathcal{H}(\mathbf{p}, \mathbf{q}_S), \tag{3}$$

where α is a trainable hyper-parameter to balance the two loss components, and $\mathcal{H}$ denotes the cross-entropy loss function.[13]

The reason of tempering $\mathbf{z}_T$ is to reduce the magnitude difference among the class likelihood values, allowing an easier transfer of the class similarity information captured in the prediction of the teacher network.[5] This is particularly important when the teacher network produces a very determined prediction (see Fig 2-(a)). Increasing the value of t allows the student network to map the output likelihood distribution of a sample much closer to the confused classes rather than the rest of the classes in the label space. This regularizes the training in the sense that these confused classes are mostly likely mis-classified by the teacher network.

3.2 Network Architectures

In this work, three different teacher networks are implemented, each of which consists of a Convolutional Neural Network (CNN) module and a Fully-Connected (FC) module. The CNN module for each teacher network is based on one of the three state-of-the-art deep learning architectures: VGG-16,[14] DenseNet,[6] and Resnet[4] (see Fig. 3-(a,c,e)). The FC module in all three teacher networks consists of the same two-layer structure with 1028 and 512 units, respectively, followed by a softmax classification layer. The student networks (see Fig. 3-(b,d,f)) are shallow versions of their respective teacher networks, with the use of less number of convolution filters and units in the FC layers. Particularly, the student networks contain only a single 256-unit FC layer, followed by a standard softmax layer and another tempered softmax layer to incorporate the knowledge distillation training (see

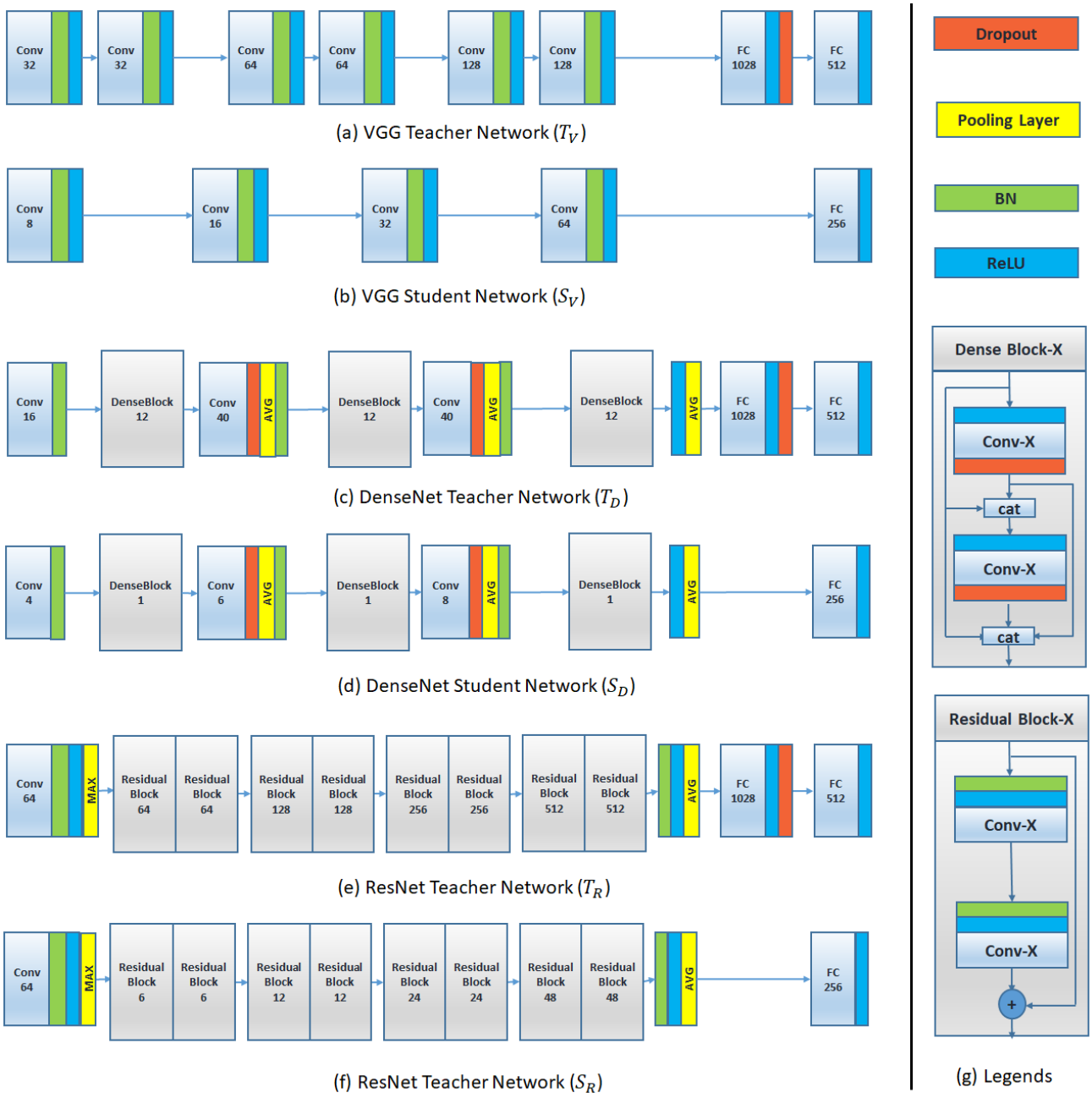


Figure 3: The respective teacher and student network architectures. The softmax classification layer in respective networks are omitted for clarity. [cat] denotes the concatenation operation.

Fig. 2-(b)). Note that the number of trainable parameters in the student networks are two orders of magnitude less than the teacher networks. Hence, they are sufficiently lightweight for execution on a mobile platform.

4. EXPERIMENT SETUP AND DATA AUGMENTATION

The data was split into mutually exclusive sets of training (60%), validation (20%), and test (20%) sets so that data from no patient appears in two of the sets. The patient information in the data was anonymized. We cross-validated the training of hyper-parameters based on the classification accuracy of the validation set networks, whereas we report the accuracy for the test set. Keras[2] deep learning library with TensorFlow backend was

Table 2: Performance comparison of the teacher and student models on the test set.

Network	No. Params	Acc_{avg}	Time/Frame [μs]
T_V	14 M	89.2	300
G_V	87 K	86.0	52
S_V	87 K	87.4	
T_D	10 M	87.5	176
G_D	260 K	83.7	68
S_D	260 K	84.6	
T_R	12 M	87.6	211
G_R	122 K	85.6	72
S_R	122 K	87.1	

(a)

Acc_{avg}	S_V	S_D	S_R		Combined
T_V(89.2)	87.4	84.3	86.6	→	87.5
T_D(87.5)	86.5	84.6	86.9	→	87.6
T_R(87.6)	86.7	84.6	87.1	→	**88.1**
↓	↓	↓	↓	↘	
Combined (89.0)	87.0	85.3	87.5		**88.1**

(b)

used to develop this work. The deep learning models were trained with the use of the Adam[9] optimizer. The ℓ_2 regularization weighting parameter was set to 1e-4.

In order to avoid bias towards more populated view classes, the stratified batch-making strategy was implemented. Each batch consists of 300 echo frames, equally taken from 12 echo view classes. Additionally, only the ultrasound beam of each echo frame was extracted and re-sized to create an 80×80 gray-scale image (see Fig. 1). Furthermore, to encourage generalization, on-the-fly data augmentation was implemented, including translation, rotation, and limited up-scaling of each training sample, independently. The rotation and translation ranges were validated by a cardiologist, to assert that the deployed data augmentation had not deform the image to the extent that it would be mis-classified by a human expert.

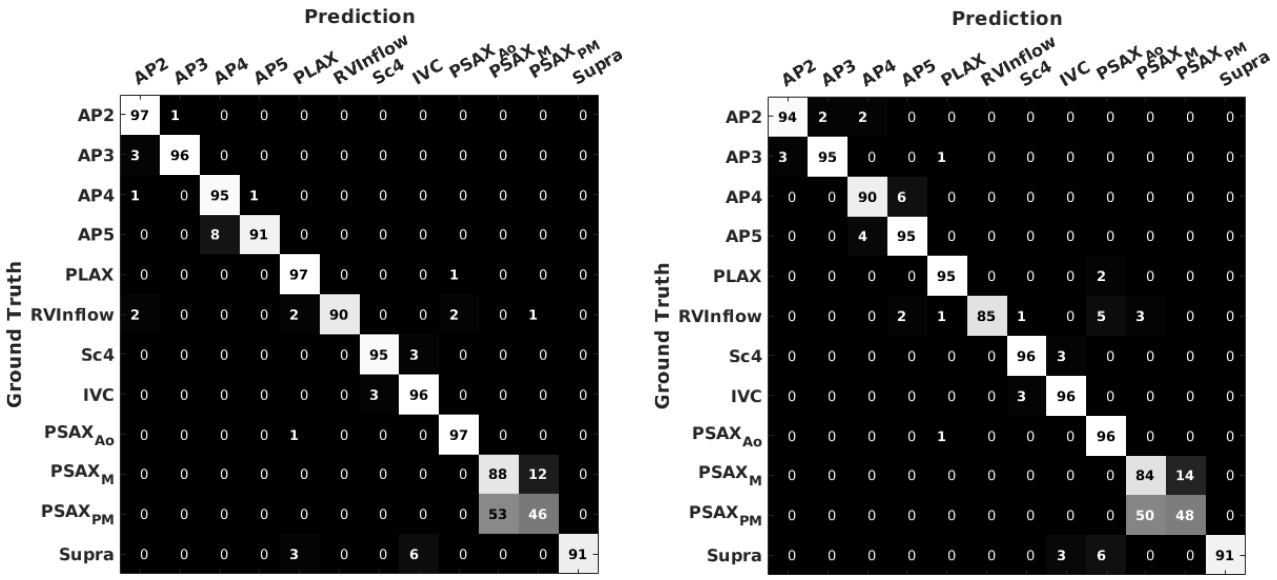
The cardiologist's labels for each echo cine series were used as the ground truth $\mathbf{p}$ for every individual frame of the cine. The teacher networks were trained using only $\mathbf{p}$ and generated $\tilde{\mathbf{q}}_T$. The $\mathbf{q}_S$ branch of the student networks were trained with $\mathbf{p}$, while the $\tilde{\mathbf{q}}_S$ branch of the student network were trained by $\tilde{\mathbf{q}}_T$ computed by the teacher network from the same on-the-fly-augmented input of the student network.

5. RESULTS

A total of three teacher and three student architectures were trained to classify the 12 standard echo views. Empirical evaluation shows that $t = 4$ (Fig. 2-(a)) produces the best results in training of the student models. In addition, we used $\alpha = 1$ (in Eq. 3) for all of our experiments. We did not observe improvement in the accuracy by reducing the α value. To classify each cine series, the likelihood predictions of all the containing frames are averaged to compute the prediction of the cine series. To evaluate the unbiased per view performance, the average class-wise classification accuracy across all 12 views is denoted as the *average accuracy* (Acc_{avg}). Moreover, the per-frame computation cost, calculated from a batch of size 300, is reported. For performance metrics to be comparable, all experiments were performed on the same system with GeForce GTX 980 Ti GPU, Intel(R) Core(TM) i7-6700 CPU, and 8 GB of RAM.

We show the experiment results in Table 2, where each number is an average of three identical model instances initialized by different random seeds. In Table 2-(a), we show the performance of all teacher networks ($T_{\{\text{V, D, R}\}}$), student networks ($S_{\{\text{V, D, R}\}}$), and the counterpart student networks trained from scratch (*i.e.*, training by only using the ground truth $\mathbf{p}$, indicating the performance of the student network without the knowledge from the teacher network, and denoted as $G_{\{\text{V, D, R}\}}$). The best performing student network is S_V, trained by T_V (which is also the best performing teacher network). Compared with G_V, the knowledge of T_V is able to contribute to 1.4% accuracy improvement; however, it is still behind T_V by 1.8%. Nevertheless, the inference time is 1/6 of teacher network. On the other hand, the student network S_D shows 0.9% improvement over G_V but it is 2.9% behind T_D. The student S_R is 1.5% better than G_R, but 0.5% behind T_R. These suggests, in general, the distillation method can extract useful knowledge to help the student networks to generalize better.

We further exam the cross-architectural knowledge distillation ability for all nine teacher and student network combinations, which is shown in Table 2-(b). The row names of Table 2-(b) indicate the teacher networks used in the training and the column names indicate the student network architecture. We also report the ensemble result of the row-wise, column-wise, and diagonal-wise three student networks in "combined" row and column



(a) (b)

Figure 4: Confusion matrices of (a) combined teacher networks $T_{\{V,D,R\}}$, and (b) combined student networks $S_{\{V,D,R\}}$ trained by teacher net T_V.

in Table 2-(b). It is interesting to see the combined performance for the same student networks, each trained by different teacher networks, are on average lower than the combined performance of the three student networks that trained by the same teacher network. This indicates that the student networks are able to converge to different features that can be complimentary, even with the guidance from the same teacher; on the other hand, the guidance from different teachers to the same student architecture does not encourage feature divergence. Finally, the top performing performance belongs to the ensemble of $S_{\{V,D,R\}}$ trained by T_R and the ensemble of $S_{\{V,D,R\}}$ trained by respective teacher network, achieving 88.1% accuracy, which is only 0.9% lower than the ensemble of the teacher networks. In Fig. 4, we show the respective confusion matrix of the teacher networks ensemble $T_{\{V,D,R\}}$ and the $S_{\{V,D,R\}}$ ensemble trained by the respective teach network.

6. CONCLUSION

In this work, we investigated the feasibility of knowledge distillation methodology in the context of echocardiography view classification. The method was extensively validated on a large echocardiography dataset. Compared to the state-of-the-art huge deep models, our models are lightweight, faster, and achieve comparable performance. Such lightweight models can be more efficiently deployed in mobile devices for use in POCUS systems in real-time recognition tasks. This can mitigate the overhead in image acquisitions for inexpert users.

REFERENCES

[1] Balaji, G.N., et al.: Automatic classification of cardiac views in echocardiogram using histogram and statistical features. Procedia Computer Science 46, 1569–1576 (2015)

[2] Chollet, F., et al.: Keras. https://github.com/keras-team/keras (2015)

[3] Gao, X., et al.: A fused deep learning architecture for viewpoint classification of echocardiography. Information Fusion 36, 103–113 (2017)

[4] He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE CVPR. pp. 770–778 (2016)

[5] Hinton, G., et al.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)

[6] Huang, G., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE CVPR (2017)

[7] Jorgensen, M., et al.: Point-of-care ultrasonography. OA Critical Care 1(1) (2013)
[8] Khamis, H., et al.: Automatic apical view classification of echocardiograms using a discriminative learning dictionary. Medical Image Analysis 36, 15–21 (2017)
[9] Kingma, D.P., Ba, J.L.: Adam: a Method for Stochastic Optimization. In: Proceedings of ICLR. pp. 1–15 (2015)
[10] Madani, A., et al.: Fast and accurate classification of echocardiograms using deep learning. npj Digital Medicine, Article number: 6 (2018)
[11] Narula, S., et al.: Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. Journal of the American College of Cardiology 68(21), 2287–2295 (2016)
[12] Park, J.H., et al.: Automatic cardiac view classification of echocardiogram. In: Proceedings of ICCV. pp. 0–7 (2007)
[13] Romero, A., et al.: Fitnets: Hints for thin deep nets. In: Proceedings of ICLR (2015)
[14] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Proceedings of ICLR. pp. 1–14 (2015)
[15] Wu, H., et al.: Echocardiogram view classification using low-level features. In: Proceedings of ISBI. pp. 752–755 (2013)
[16] Zhang, J., et al.: A Computer Vision Pipeline for Automated Determination of Cardiac Structure and Function and Detection of Disease by Two-Dimensional Echocardiography. ArXiv preprint (2017), https://arxiv.org/pdf/1706.07342.pdf