CS230

Teacher-Student Distillation for Heart Sound Classification

Pranav Sriram
prsriram@stanford.edu

Abstract

Automated classification of heart sound data has been an active area of research in recent years [1]. While recent research has involved the utilization of contrastive learning methods in order to aid with heart sound classification, there remains room for improvement. Furthermore, while distillation with a contrastive teacher model has been shown to improve performance on ImageNet classification, it has not yet been applied to heart sound classification tasks [2]. This paper explores the effectiveness of a distilled teacher-student CNN on binary heart sound classification. We find that our teacher-student network is able to roughly match performance of the teacher network while being half its size. Furthermore, our teacher-student network was able to achieve a test AUROC of .915.

1 Introduction

While heart sound classification has been an area of interest for many years [3], progress has improved recently as digital stethoscopes have gained popularity, contributing to an expanding set of cardiac sound data [4]. With that said, a significant portion of this sound data is unlabeled.

This problem is not unique to heart sound data. Due to the lack of reliable labeled data in the medical AI sector, contrastive learning methods have recently gained widespread prominence in a variety of medical classification tasks [5]. Contrastive learning is a self-supervised deep learning method that is generally effective in paradigms with low amounts of labeled data [6].

Through Stanford's AI for Healthcare Bootcamp, our research lab has developed a contrastive approach to heart sound classification. However, model performance with this contrastive approach still has room for improvement.

This project seeks to improve model performance by applying one-phase knowledge distillation. Knowledge distillation is a training technique where we train a teacher-student model using the output values (soft labels) from a teacher network. This was inspired by recent research that has shown distillation with a contrastive teacher model improved performance on ImageNet classification [2].

We constructed a teacher-student CNN model and utilized the same training regime as [2] (further explained in Appendix 8.1). Afterward, we evaluated the performance of the teacher-student model against that of the original contrastive teacher model. The input to our teacher-student model was a spectrogram corresponding to a heart phonocardiogram (PCG) recording. The output of this model was a binary prediction on whether the PCG recording was normal or abnormal.

This project is important and interesting because it may help facilitate advances in the medical audio classification space. Increasing the accuracy of automated cardiac sound algorithms could help doctors make better medical diagnoses. Additionally, if distillation ends up being promising in the context of cardiac sound data, it may also be beneficial in other contexts (such as lung sound classification).

,

2 Related work

2.1 Heart Sound Classification

Currently, some common approaches to heart sound classification include support vector machines (SVMs) [7] as well as some classifiers that are heavily dependent on independent feature extraction systems [8]. With that said, there also exists research on the use of CNNs for heart sound classification [1]. In [1], researchers were able to achieve an accuracy of 79.5% on binary heart sound classification via use of a CNN and the Windowed-sinc Hamming filter algorithm. The promising results reported in [1] were part of what motivated our decision to use a CNN model architecture for this project.

2.2 Contrastive Learning

Contrastive learning is a machine learning framework where learning occurs by encouraging representations from similar examples to be close to each other. This framework was introduced in [6]. It is especially effective for tasks lacking labeled data. In a contrastive setup, an input and an augmented version of that same input are considered to be a positive pair. Any other combination of inputs is considered to be a negative pair. Learning proceeds by attempting to optimize similarity in representations between positive pairs. In doing so, contrastive models develop an understanding of a given domain.

Contrastive learning has shown strong results in some medical domains. For example, researchers at Stanford found that pre-training with the contrastive framework Moco-CXR improves performance on chest X-ray interpretation [9]. Based on ongoing research via the AI for Healthcare Bootcamp, my lab has also found promising results when applying contrastive learning to the heart sound domain.

2.3 Distillation

Distillation involves the use of a teacher-student model that is trained on soft labels from a teacher model. As we've discussed, [2] shows that one-phase distillation can have promising results on ImageNet classification. In addition, many other papers have found training with distillation can improve classification performance [10].

There has also been some existing research on distillation in audio domains. For example, the paper "Multi-Representation Knowledge Distillation For Audio Classification" shows that distillation can improve performance across a variety of different audio classification tasks [11]. Despite these promising results, there has not been much research into the effectiveness of distillation in medical audio classification. This paper will attempt to fill this hole in existing research.

3 Dataset and Features

For our dataset, we utilized the PhysioNet/CinC Challenge 2016 dataset [4]. The PhysioNet/CinC Challenge 2016 dataset consists of 3,240 phonocardiogram (PCG) recordings varying from 5 to 120 seconds in length and classified as "normal" and "abnormal." 2,575 of these recordings had the label "normal" and 665 had the label "abnormal." Normal recordings came from healthy patients and abnormal recordings came from patients with confirmed cardiac conditions (such as coronary artery disease) [4]. We split up the Physionet dataset into pre-train, fine-tune, validation (dev), and test sets according to patient ID with each patient only appearing in one set.

The pre-train dataset consisted of 2,040 examples. The fine-tune, validation, and test sets each consisted of 400 examples each. The dev and test sets were class-balanced with 200 normal and abnormal examples each. Thus, the pre-train and fine-tune datasets collectively contained 2,175 normal and 225 abnormal recordings. This class imbalance made classification more challenging.

Recordings were padded and converted into Mel Spectrograms. Spectrogram size was set to 50 pixels by 2805 pixels. An example heart spectrogram is provided in Appendix 8.2. While no data augmentation was used for the teacher-student model, Google's SpecAugment was used to train the contrastive teacher model [12].

4 Methods

Three models were trained for this project: a distilled teacher-student model, a contrastive teacher model, and a student baseline model.

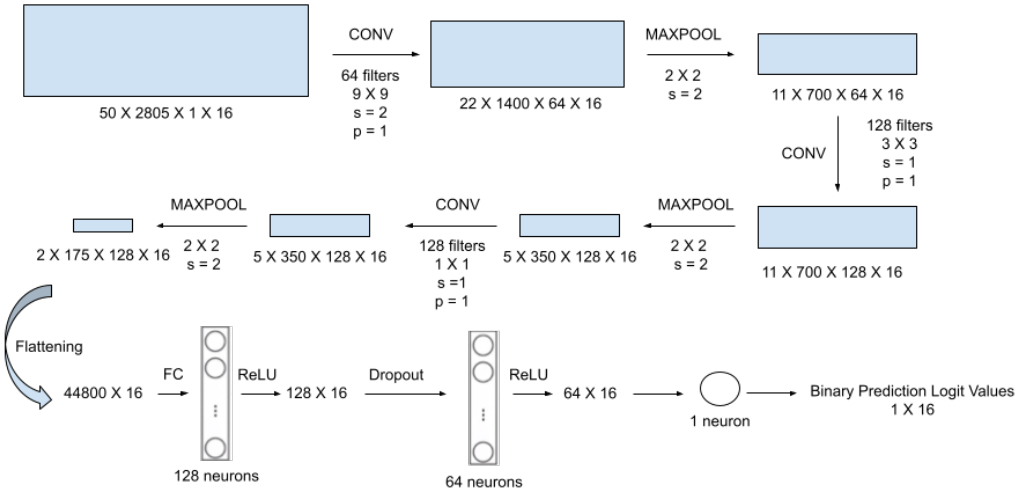

Figure 1: Teacher-Student / Student Baseline CNN architecture: The spectrogram input was 50 by 2805 and a batch size of 16 was used. LeakyReLU was used as an activation function following the CONV layers. The model consisted of 5,838,465 trainable parameters.

The teacher-student architecture (shown in Figure 1) was inspired by classic convolutional networks, especially AlexNet and VGG-16. LeakyReLU was used as an activation function following CONV layers and batch size was 16. The output value corresponds to the binary prediction logit value. A negative output value means the model predicts the spectrogram corresponds to a normal heart sound and a positive value means the model predicts the spectrogram corresponds to an abnormal heart sound.

The teacher-student model was trained under a supervised approach with the psuedo-labels coming from the output of the contrastive teacher model. A Binary Cross Entropy with Logits loss function was chosen for the teacher-student network.

The teacher model was pre-trained under a contrastive approach. Its architecture consisted of a ResNet-18 encoder followed by a multilayer perceptron evaluator. The number of trainable parameters in the teacher model was 11,827,905, over double that of the teacher-student / student baseline models.

The teacher model used a Normalized Temperature-scaled Cross Entropy (NT-Xent) loss function for pre-training as specified in [6]. This loss function can be expressed for a positive pair of examples as:

$$\mathbb{l}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

Figure 2: NT-Xent Loss Function

Positive pairs for our teacher model were defined as an input spectrogram and its corresponding augmented spectrogram (generated via SpecAugment). z_i and z_j correspond to the vectors generated by passing our contrastive encodings for i and j through a projection head. Finally, the similarity function ($sim(z_i, z_j)$) corresponds to cosine similarity. The teacher model used a temperature of .07. Finally, this loss was summed across all positive pairs in our dataset.

Lastly, the student baseline model used the same architecture specified in Figure 1. However, unlike the teacher-student model, it was trained with the true label values for each spectrogram. It also used a Binary Cross Entropy with Logits loss function.

5 Experiments/Results/Discussion

5.1 Hyperparameters

For our experiments, the main hyperparameter tuned was the learning rate. We experimented with three different learning rates: $\alpha = 10^{-3}, 10^{-4}, 10^{-5}$. When the teacher-student model was run with a learning rate of $\alpha = 10^{-3}$, predictions were inconsistent and the model appeared somewhat instable. When run with a learning rate of $\alpha = 10^{-5}$, this problem was resolved but training was quite slow. A learning rate of $\alpha = 10^{-4}$ was a good balance, with the model's performance remaining relatively stable and training taking a reasonable amount of time.

The teacher model used a batch size of 16. Thus, to ensure consistency, we ran the teacher-student model with a batch size of 16 as well.

5.2 Experimental Procedure

As mentioned, we trained three separate models. Our experimental procedure was inspired by the methodology used in [2] (high-level diagram provided in Appendix 8.1).

The contrastive teacher model was pre-trained for 3 epochs on our pre-train dataset using a NT-Xent loss function. Importantly, it was not provided with any of the true labels for the pre-train dataset. Afterwards, it was supervised fine-tuned on our fine-tune dataset for 30 epochs. For fine-tuning, the teacher used a Binary Cross Entropy with Logits loss function. Checkpoints were saved at the end of each epoch, and the checkpoint with the highest dev set AUROC was stored. This helped avoid overfitting to the fine-tune set; the checkpoint with the highest dev set AUROC ended up being that after the second epoch of fine-tuning.

The teacher-student model was trained for 15 epochs on the pre-train dataset. It was given the output of the aforementioned teacher model as soft labels. Notably, it was not provided with any of the true labels for the pre-train dataset. In line with the methodology followed in [2], the teacher-student model was not trained at all on the fine-tune set. Checkpoints were saved the same way as the teacher model. The best checkpoint ended up being that after the last epoch.

Finally, the student baseline model was also trained for 15 epochs on the pre-train dataset. It was given the true labels for the pre-train dataset. Checkpoints were saved the same way as above; the best checkpoint was after epoch 11.

5.3 Results & Discussion

The primary evaluation metric on our dev set was AUROC. This metric is justified since our dev set was class-balanced. These results are detailed in the table below:

Model	Dev Set AUROC
Teacher	**.926**
Teacher - Student	.907
Student Baseline	.820

Table 1: Dev AUROC Performances. Teacher-Student Dev AUROC plot included in Appendix 8.3.

As shown in the above table, the teacher model had the highest dev set AUROC. Surprisingly, despite having access to the true labels on the pretrain dataset, the student baseline model performed the worst on the dev set. This is likely due to class imbalance. Since about 90% of our pre-train dataset consisted of normal examples, our student model is incentivized to decrease pre-train loss by guessing "normal" overly often. However, this behavior did not translate as well on the class-balanced dev set.

Using a weighted Binary Cross Entropy Loss function for just the student baseline was considered; however, it was ultimately determined that this would serve as an unfair baseline (since by definition, the teacher-student model cannot possibly have access to the same weighted loss function).

On our test set, we calculated AUROC, generated a confusion matrix, and determined precision, recall, and F_1 scores for both of our classes:

		Predicted Label					
		Teacher		Teacher - Student		Student Baseline	
		Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Actual Label	Normal	154	46	150	50	198	2
	Abnormal	21	179	8	192	146	54

Table 2: Confusion matrices for teacher, teacher-student, and student baseline models.

	Teacher **(AUROC: .927)**			Teacher - Student (AUROC: .915)			Student Baseline (AUROC: .864)		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Normal	.88	.77	.82	.95	.75	.84	.58	.99	.73
Abnormal	.80	.90	.84	.79	.96	.87	.96	.27	.42

Table 3: AUROC, precision, recall, and F1 scores for the aforementioned models.

These results closely align with our dev set results. Teacher and Teacher-Student performance is very good, both resulting in test AUROC's above .9, which is typically considered outstanding on medical diagnostic tasks [13]. While the teacher model has a better AUROC, if we consider test accuracy, our teacher-student model (accuracy of 85.5%) actually outperforms our teacher model (accuracy of 83.25%) slightly. This is especially impressive given the fact the teacher-student model only contains 5,838,465 trainable parameters while the teacher model contains over double that (11,827,905 trainable parameters). Both models have F1 scores for the normal and abnormal classes between .8 and .9.

Additionally, the test results for the student baseline model lends credence to our class imbalance theory. We can see that the student baseline has an extremely high recall on the normal class at .99 (correctly labeling 198 out of 200 normal examples as normal). However, the recall on the abnormal class is extremely low at .27. Looking at the confusion matrix for our student baseline, we see that it has an extremely high amount of false negatives (146). However, it's important to note that despite the class imbalance issues plaguing the student baseline model, it still achieves a test AUROC of .864.

In terms of qualitative evaluation, we examined some of the raw audio files and spectrogram inputs our teacher-student model predicted incorrectly. As untrained human listeners, we were unable to hear any obvious patterns in the audio or identify patterns in the spectrogram. In the future, we plan to show a trained medical professional our incorrectly predicted heart sounds files and spectrograms to see if they can help with pattern identification.

6 Conclusion/Future Work

We were able to train a teacher-student model that roughly matched the performance of our teacher model in terms of test accuracy and AUROC. This is significant given our teacher-student model was less than half the size of the teacher model.

For future work, we'd like to explore using a focal loss function across all three of our models. This may help solve the student baseline model's class imbalance issue, in turn providing a more useful baseline for student performance on this task. Looking at the dev AUROC plot in Appendix 8.3, we see that it's possible teacher-student performance would slightly benefit from longer training. Thus, we'd also like to experiment with using early stopping (train each model until dev AUROC decreases on three consecutive epochs) instead of a fixed epoch regime. Additionally, we'd like to explore changing the architecture of the teacher-student model and test whether two-phase distillation yields better results than our current one-phase approach.

7 Contributions

All contributions on the distillation algorithm were made my myself. The entirety of my research lab (Lung Heart Classification Team in the AI for Healthcare Bootcamp) contributed to the contrastive teacher model utilized. Thanks to Pranav Rajpurkar for advising this work.

References

[1] Ryu, Heechang, Jinkyoo Park, and Hayong Shin. "Classification of heart sound recordings using convolution neural network." *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016.

[2] Chen, Ting, et al. "Big self-supervised models are strong semi-supervised learners." *arXiv preprint arXiv:2006.10029* (2020).

[3] Redlarski, Grzegorz, Dawid Gradolewski, and Aleksander Palkowski. "A system for heart sounds classification." *PloS one* 9.11 (2014): e112673.

[4] Liu, Chengyu, et al. "An open access database for the evaluation of heart sound algorithms." *Physiological Measurement* 37.12 (2016): 2181.

[5] Zhang, Yuhao, et al. "Contrastive learning of medical visual representations from paired images and text." *arXiv preprint arXiv:2010.00747* (2020).

[6] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

[7] Whitaker, Bradley M., et al. "Combining sparse coding and time-domain features for heart sound classification." *Physiological measurement* 38.8 (2017): 1701.

[8] Safara, Fatemeh, et al. "Multi-level basis selection of wavelet packet decomposition tree for heart sound classification." *Computers in biology and medicine* 43.10 (2013): 1407-1414.

[9] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. "Moco pretraining improves representation and transferability of chest x-ray models." *arXiv preprint arXiv:2010.05352* (2020).

[10] Ruffy, Fabian, and Karanbir Chahal. "The state of knowledge distillation for classification." *arXiv preprint arXiv:1912.10850* (2019).

[11] Gao, Liang, et al. "Multi-representation knowledge distillation for audio classification." *arXiv preprint arXiv:2002.09607* (2020).

[12] Park, Daniel S., et al. "Specaugment: A simple data augmentation method for automatic speech recognition." *arXiv preprint arXiv:1904.08779* (2019).

[13] Mandrekar, Jayawant N. "Receiver operating characteristic curve in diagnostic test assessment." *Journal of Thoracic Oncology* 5.9 (2010): 1315-1316.

8 Appendix

8.1 High-Level Overview of Training Regime

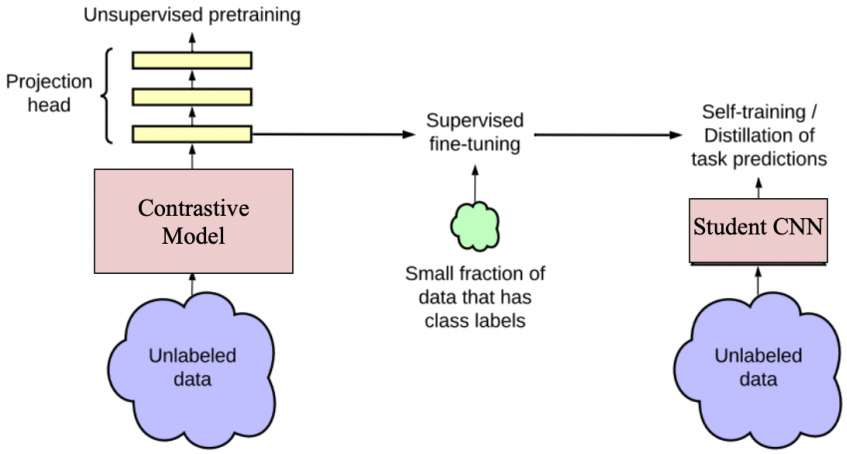

Figure 3: High-level overview of our overall training regime. First, we pre-train our contrastive teacher model on a large unlabeled dataset. Then, we fine-tune it on a small fine-tune dataset with labels (normal supervised learning). Next, we pre-train a teacher-student CNN on the same unlabeled dataset using the logit values outputted by the teacher model. Notably, the teacher-student network is not fine-tuned. Both the teacher and teacher-student networks are then evaluated. This image was edited from [2].

8.2 Heart Spectrogram Example

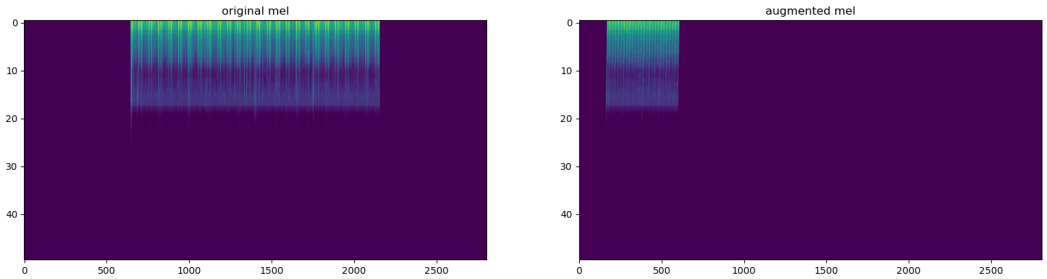

Figure 4: On the left, is an example heart spectrogram from our dataset. On the right, is the result after passing the mel spectrogram through SpecAugment. Both the original and augmented spectrograms were used for training our contrastive teacher model. In contrast, the teacher-student model only used the original mel spectrogram.

8.3 Teacher-Student Dev AUROC Plot

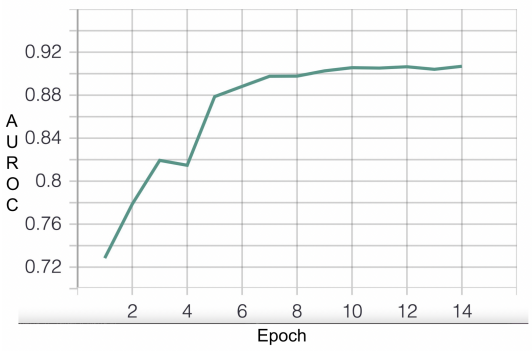

Figure 5: This graph shows the teacher-student model's Dev AUROC performance over time. While Dev AUROC, appears to increase roughly linearly across the first 8 epochs (except for the drop around epoch 4), the rate of increase does slow down following epoch 8. With that said, it's still possible the teacher-student model may benefit from extended training.