HIERARCHICAL KNOWLEDGE DISTILLATION: A MULTI-STAGE LEARNING APPROACH

Technical Report

Quoc Thinh Vo, David K. Han

Drexel University, College of Engineering
Electrical and Computer Engineering Department
3100 Market St, Philadelphia, PA 19104, USA
qv23, dkh42@drexel.edu

ABSTRACT

This technical report details our approach to Task 1 of the 2023 Detection and Classification of Acoustic Scenes and Event (DCASE2023) [1], which focuses on the classification of recorded audios for acoustic scene recognition. The task calls for a quantized model of no more than 128KB in memory allowance for model parameters and a maximum of 30 millions of multiply-accumulate operations (MMACS) per inference. Our solution exploits log-mel sprectrogram features and leverages multiple data augmentations. Our proposed methodology utilizes an audio spectrogram transformer (AST) [2] as the teacher model and multiple Convolutional Neural Network (CNN) models as students in a hierarchical knowledge distillation (KD) framework. This approach aids in bridging the substantial parameter disparity between the teacher model, which has over 86 million parameters, and our compact CNN-based model limited to just 119,526 parameters. Upon network training completion, the variable type of the weight data is converted into type INT8 to meet the size constraints. Our INT8 model achieves a log-loss of 1.59 and an accuracy of 46.01% on the TAU Urban Acoustic Scenes 2022 Mobile Development [3] dataset's standard test set, signifying the efficacy of our framework. Our proposed method demonstrates the potential of distillation strategies in optimizing smaller models without compromising their learning ability in a hierarchical approach.

Index Terms— audio spectrogram transformer, log-mel spectrogram, acoustic signal classification, knowledge distillation

1. INTRODUCTION

Acoustic scene classification is a domain dedicated to identifying the soundscapes of recorded audios, an area that has seen significant advancements in applications. Numerous classification methods have emerged in recent years such as CNN-based models [4] and transformer models [2]. The first task of DCASE 2023 proposes a captivating challenge: participants are tasked to develop a method capable of identifying various auditory scenes, such as airports, parks, streets, and more, within a one-second audio clip. This must be achieved while complying the limitations of less than 30 MMACS and 128KB in maximum memory allowance for model parameters. This presents an interesting challenge that combines machine learning, audio processing, and resource optimization for edge devices.

Thanks to the Office of Naval Research for funding.

To thoroughly exploit the time-frequency characteristics of audio data, we developed an approach that incorporated diverse data processing techniques in data augmentation and feature extraction. We design and train a compact CNN-based model by applying numerous data augmentation techniques in the time-frequency domain to enhance training data variety and model adaptability. To further improve the model's performance, we apply KD concept that had previously exploited by Schmid et al.[5], and Gong et al.[6]

In summary, the main contributions of our proposed method are:

1. We designed a compact CNN model tailored to satisfy the specified constraints of the challenge.

2. We applied the KD technique to enhance the performance of our compact model, thus optimizing its learning from the teacher model.

3. We developed two supplementary CNN-based models that act as bridges between the teacher model and the compact student model to facilitate the knowledge transfer.

2. PROPOSED METHODOLOGY

2.1. Hierarchical Knowledge Distillation Framework

The proposed approach is a hierarchical knowledge distillation (H-KD) framework with a pre-trained AST model serves as the teacher, guiding a cascade of CNN-based student models. Initially, the teacher model, having been pre-trained, imparts its knowledge to the successive student models. The first student model, Student 1, is a Resnet-18 based model consisting of over 11M parameters. The second student model, Student 2, is a CNN-based model with over 746K parameters. Lastly, Student 3, the final lightweight model, designed similarly to the architecture of Student 2 but is more compact, consisting of merely 119K parameters.

Our proposed CNN-Transformer knowledge distillation framework utilizes the strengths of both these architectures, facilitating the absorption of knowledge by our lightweight CNN-based model (with 119,526 parameters) from the teacher model (with over 86 million parameters), despite their huge difference in size.

During the training process, the teacher model remains in evaluation mode - effectively frozen and unaltered. Student models 1 and 2 are initially pre-trained with the challenge development data and are subsequently fine-tuned throughout the training of Student 3. As described in Figure 1, the first distillation learning layer incorporates a pre-trained AST as a teacher model. The loss for Student 1 is computed based on both its label loss and the distillation loss

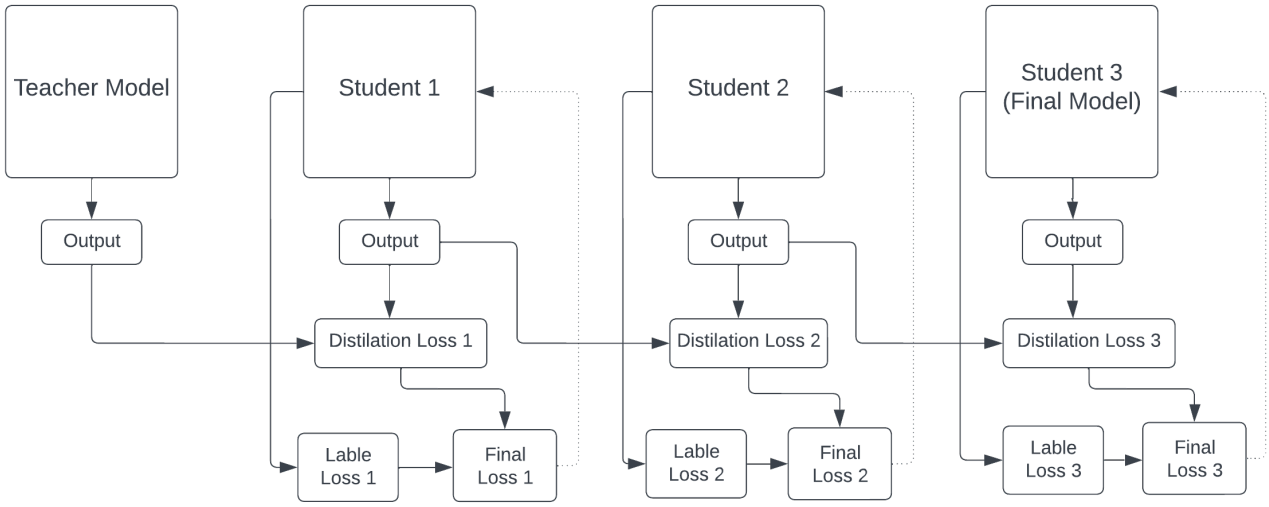

Figure 1: Proposed methodology on effective training of Hierarchical Knowledge Distillation

Class	Baseline Log-Loss	Baseline Accuracy (%)	Proposed Model Log-Loss	Proposed Model Accuracy (%)
Overall	**1.575**	**42.90**	**1.591**	**46.01**
Airport	1.534	39.4	1.430	50.41
Bus	1.758	29.3	1.839	41.48
Metro	1.382	47.9	1.958	40.17
Metro Station	1.672	36.0	1.728	40.51
Park	1.448	58.9	0.898	74.55
Public Square	2.265	20.8	2.277	22.29
Shopping Mall	1.385	51.4	1.736	40.71
Street Pedestrian	1.822	30.1	1.896	32.42
Street Traffic	1.025	70.6	0.981	72.69
Tram	1.462	44.6	1.579	44.80

Table 1: Class-wise Log-loss and Accuracy on the DCASE 2022 Development Test Data for Baseline and the Proposed Model.

- a measure of the difference between the teacher's output and Student 1's output. A similar process is used in the second distillation learning layer, which involves Student 1 and Student 2. Student 2's loss is also computed based on its label loss and the distillation loss between Student 1's output and Student 2's output. Finally, in the third distillation learning layer, Student 3's loss is computed, considering both its label loss and the distillation loss between Student 2's output and Student 3's output.

The loss function that is used throughout the whole framework as follows:

$$\mathrm{L} = \alpha \cdot \text{student\_loss} + (1 - \alpha) \cdot \text{distillation\_loss} \quad (1)$$

The balancing coefficient is denoted by α, while the student label loss and Kullback–Leibler divergence are represented by student loss and distillation loss, respectively. The activation function is softmax and the label loss is calculated using cross entropy. As described, we use the Kullback–Leibler divergence as distillation loss. In the context of cross-model KD, there may be a discrepancy in the softness of the logit distributions between the teacher and student models. To mitigate this, we selectively apply a temperature factor (T) solely to the logits of the teacher model, thereby enabling explicit control over this difference [6].

Based on experimental results, our final submission model has the following configurations: α = 0.45 and T = 1.5.

2.2. Data Augmentation

To enhance the model's performance and generalization ability, we implement data augmentation in the time and frequency domains. Time-domain data augmentation techniques include mix up [7, 8] and adding white noise to signal data. Frequency-domain data augmentation is achieved using SpecAugment [9]. These methods has been exploited by various experiments in the domain.

2.3. Backbone Architecture

We design a CNN-based model as the backbone of our solution, consisting of a mix of standard convolution (Conv2d) layers and depthwise separable convolution layers.

The architecture begins with a standard convolution layer that takes a single-channel input and outputs 16 channels. This is followed by alternating depthwise separable convolution layers, and

another standard convolution layer. The depthwise separable convolution layers increase the receptive field while maintaining computational efficiency. A series of depthwise separable convolution layers follows, each of which is designed to further process the features and increase the model's ability to capture complex patterns in the data.

Each convolutional layer is followed by a batch normalization operation to stabilize the learning process and reduce internal covariate shift. The output from these layers goes through a ReLU activation function to introduce non-linearity into the model.

To handle overfitting, dropout and max pooling layers are used. The max pooling layers are strategically placed to progressively reduce the spatial dimensions of the data, allowing the model to focus on the most salient features.

The output from the final convolution layer is flattened and passed through two fully connected (Linear) layers, which serve to aggregate the learned features and map the output vector to the final 10-classification classes.

Quantization stub layers (QuantStub and DeQuantStub) are included in the model for quantization-aware training. These layers do not change the model's behavior during training but allow for the conversion of the model's parameters to INT8 during the post-training quantization process.

The CNN-based model contains mix of standard and depthwise separable convolution layers, batch normalization, and dropout, has been meticulously designed to handle the challenges of acoustic scene classification under the constraints of model size and computational efficiency.

2.4. Training

We trained the model using back-propagation and RAdam optimizer with a batch size of 64 and cross-entropy loss function. A cosine learning rate scheduler was used to reduce the learning rate during training when validation accuracy ceased to increase.

2.5. Quantization and Inference

Post-training quantization is applied to convert the weights in the model to INT8, reducing the model size but slightly decreasing the accuracy. Results in this report were obtained using the quantized model.

3. EXPERIMENTAL RESULTS

For testing log-loss and accuracy observations, Table 2 shows the performance improvement achieved by incorporating KD into the training of our proposed CNN model. For consistency and fairness, all experiments were conducted over the same number of training epochs.

The baseline model achieved a log-loss of 1.575 and an accuracy of 42.90%. Our proposed CNN model, before KD was implemented, demonstrated a slightly higher log-loss of 2.05 and a slightly lower accuracy of 39.89%. The initial application of KD was done by calculating the distilation loss directly between the AST - teacher model and the compact CNN model. The experiment improved the CNN model, as reflected by the reduced log-loss of 1.89 and increased accuracy of 42.61%.

Our final and most effective model, the proposed CNN model trained within a comprehensive KD framework, exhibited a further reduced log-loss of 1.59 and much higher accuracy of 46.01%. This marks a significant improvement over the baseline model, validating the efficacy of our proposed model and the applied KD technique. Additionally, when being trained with the H-KD framework, the proposed CNN model converges faster in term of effective training.

For a more detailed analysis, Table 1 presents class-wise log-loss and accuracy for both the Baseline model and our proposed model on the TAU Urban Acoustic Scenes 2022 Mobile Development test data. This allows us to further understand the individual class performance and the overall impact of our proposed CNN model with H-KD framework.

Model	**Log-Loss**	**Accuracy (%)**
Baseline	1.575	42.90
Proposed CNN	2.05	39.89
Proposed CNN + KD	1.89	42.61
Proposed CNN + KD framework	**1.59**	**46.01**

Table 2: Log-loss and Accuracy on the TAU Urban Acoustic Scenes 2022 Mobile Development - Test Set.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a hierarchical knowledge distillation framework that utilized audio spectrogram transformer and CNN-based models in a multi-stage learning approach. The method outperformed the Baseline system accuracy and demonstrated that it could bring up the performance of a light-weighted model, in addition to helping the student model converge faster.

We aim to improve the model's capabilities by exploring new method of incorporating self-supervised learning [10] for addressing data sparsity in the teacher model and further improve the student model robustness.

5. ACKNOWLEDGMENT

We would like to offer our thanks to the Office of Naval Research for funding. Grant No. N00014-21-1-279.

6. REFERENCES

[1] I. Martín-Morató, F. Paissan, A. Ancilotto, T. Heittola, A. Mesaros, E. Farella, A. Brutti, and T. Virtanen, "Low-complexity acoustic scene classification in dcase 2022 challenge," 2022. [Online]. Available: https://arxiv.org/abs/2206.03835

[2] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[3] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, pp. 56–60. [Online]. Available: https://arxiv.org/abs/2005.14623

[4] D. Kim, S. Park, D. Han, and H. ko, "Multi-band cnn architecture using adaptive frequency filter for acoustic event classification," *Applied Acoustics*, vol. 172, p. 107579, 01 2021.

[5] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "Cpjku submission to dcase22: Distilling knowledge for lowcomplexity convolutional neural networks from a patchout audio transformer," DCASE2022 Challenge, Tech. Rep, Tech. Rep., 2022.

[6] Y. Gong, S. Khurana, A. Rouditchenko, and J. Glass, "Cmkd: Cnn/transformer-based cross-model knowledge distillation for audio classification," *arXiv preprint arXiv:2203.06760*, 2022.

[7] H. Yu, H. Wang, and J. Wu, "Mixup without hesitation," in *Image and Graphics: 11th International Conference, ICIG 2021, Haikou, China, August 6–8, 2021, Proceedings, Part II 11*. Springer, 2021, pp. 143–154.

[8] R. Mars and R. K. Das, "A device classification-aided multitask framework for low-complexity acoustic scene classification."

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.