TENCENT SUBMISSION TO DCASE23 TASK1: LOW-COMPLEXITY DEEP LEARNING SOLUTION FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

Weicheng Cai, Mingyuan Zhang, Xiang Zhang

Tencent Inc. Beigjing, China
wilsoncai@tencent.com

ABSTRACT

In this technical report, we present the Tencent team's entry for Task 1 Low-Complexity Acoustic Scene Classification in the DCASE 2023 challenge. We mainly follow the DCASE 2022 1st place solution from the CP-JKU team and have made some adjustments to meet the requirement of this year. Our approach involves employing knowledge distillation to train low-complexity CNN student models using Patchout Spectrogram Transformer (PaSST) models as teachers. We initially train the PaSST models on Audioset and then fine-tune them using the TAU Urban Acoustic Scenes 2022 Mobile development dataset. Lastly, we quantize the student models to enable 8-bit integer-based inference computations to meet the low-complexity constraints in edge devices.

Index Terms— acoustic scene classification, knowledge distillation, Vision transformer, PaSST, CNN

1. INTRODUCTION

Task 1 in the DCASE 2023 Challenge is to identify one of the ten predefined scenes given one-second audio clips [1, 2]. This task is challenging because the information contained in a 1-second audio signal is quite limited. Besides, the goal is to ensure generalization across multiple devices, and some devices appear only in the test subset. This year, the Acoustic Scene Classification (ASC) situation focuses on achieving classification using devices with limited computational power and memory capacity, which in turn imposes constraints on model complexity, including the number of parameters and the count of multiply-accumulate (MACs) operations. Specifically, the maximum memory allowance for model parameters is 128KB (Kilobyte), and the limit of the maximum number of MACS per inference is 30 million. Submissions will be ranked by weighted average rank of classification accuracy, memory usage, and MACs. Hence, the systems being submitted should place significant emphasis on effectively balancing the trade-off between classification accuracy and model size.

In the last decade, as neural networks gained prominence, Convolutional Neural Networks (CNNs) have emerged as the de-facto standard for end-to-end audio classification models. These models strive to establish a direct mapping between audio waveforms or spectrograms and their corresponding labels. [3, 4]. More recently, neural networks based purely on self-attention, such as the Audio Spectrogram Transformer (AST) [5], have been shown to further outperform deep learning models constructed with CNNs on various audio classification tasks, thus extending the success of Transformers from natural language processing, and computer vision to the audio domain. Based on the AST model, many of its variants, such as Patchout Spectrogram Transformer (PaSST) [6], Hierarchical Token Semantic Audio Transformer (HTS-AT) [7] have achieved remarkable results in audio tagging and sound event detection tasks.

The outstanding performance of these AST-based variants in audio classification and detection tasks demonstrates the potential of self-attention mechanisms and transformer architectures in the field of audio processing. Conversely, although Transformer models exhibit superior performance, their high complexity makes them less computationally efficient than CNN models, especially for resource-limited devices.

In the research conducted by Gong et al. [8], the authors uncover an intriguing relationship between CNN and Transformer models, advocating for the application of cross-model knowledge distillation (KD) in audio classification tasks. By using either a CNN or an AST model as the teacher and training a distinct model as the student through knowledge distillation, the student model exhibits substantial enhancement and exceeds the teacher's performance. Consequently, the knowledge-distilled CNN model, which has only 8M parameters, surpasses the original AST with 88M parameters on the FSD50K dataset.

At the same time, the investigators at CP-JKU propose a training technique for effective CNNs that employ offline KD from high-performing, intricate transformers [9]. By merging this training strategy with an efficient CNN architecture inspired by MobileNetV3, the resulting models surpass earlier solutions in both efficiency and predictive performance.

Drawing from their previous research, the CP-JKU team submitted their solutions for DCASE 2022 Task 1 and won 1st place [10]. Our approach primarily follows the CP-JKU team's winning solution for DCASE 2022, with some modifications to fulfill this year's requirements. Our method involves using knowledge distillation to train low-complexity CNN student models with PaSST models serving as teachers. The PaSST models are trained on Audioset and subsequently fine-tuned using the TAU Urban Acoustic Scenes 2022 Mobile development dataset. Finally, the student models are quantized to facilitate 8-bit integer-based inference computations, adhering to the low-complexity constraints necessary for edge devices. In our submissions, PaSST and Audioset are the only external data sources used.

2. KNOWLEGE DISTILLATION

Although deep neural networks have achieved significant success in various tasks, including image classification, and speech recognition, there is a growing need to develop resource-efficient deep neural networks (e.g., with fewer parameters) without compromising accuracy. The challenge of deploying large deep neural network models is especially pertinent for edge devices with limited memory and computational capacity.

To tackle this challenge, a model compression method was proposed to transfer the knowledge from a large model into training a smaller model without any significant loss in performance [11]. KD transferring knowledge from a cumbersome teacher model to a lightweight student model has been investigated to design efficient neural architectures with high accuracy with a few parameters. The KD process captures and "distills" the knowledge in an ensemble of large models into a smaller single model that is much easier to deploy without significant loss in performance.

Knowledge is transferred from the teacher model to the student by minimizing a loss function, aimed at matching softened teacher logits as well as ground-truth labels. The logits are softened by applying a "temperature" scaling function in the softmax, effectively smoothing out the probability distribution and revealing inter-class relationships learned by the teacher. We use KD in its original form, as introduced in [11], and set the "temperature" parameter to 3 as in [10].

3. TEACHER MODEL

For the training of teacher models, our goal is to achieve the highest possible accuracy using large models or even multiple ensemble large models without considering the model size.

3.1. AST

We use an Audio Spectrogram Transformer (AST) [5] pretrained on AudioSet full dataset here. First, the raw audio waveform is downsampled using a sampling rate of 16kHz. Each waveform is then converted to a sequence of 128-dimensional log Mel filterbank (Fbank) features computed with a 25ms Hanning window every 10ms. We then split the spectrogram into a sequence of 16×16 patches with an overlap of 6 in both time and frequency dimensions as described in [5]. We use the ImageNet-pretrained model as our initial weights and use an initial learning rate of 1e-5 and train the model for 5 epochs, the learning rate is cut into half every epoch after the 2nd epoch.

After the model is trained, we fine-tune the AST model on the TAU Urban Acoustic Scenes 2022 Mobile development dataset. We use an initial learning rate of 1e-5 and decrease the learning rate with a factor of 0.85 for every epoch after the 5th epoch.

3.2. EfficientNet

EfficientNet [12] is a recently proposed CNN architecture that has shown an advantage in both accuracy and efficiency over previous architectures. The original EfficientNet-B2 model for image classification has 9.11M parameters. Our training configuration here is the same as that of AST, except that we have replaced the model from AST to EfficientNet and increased the number of training epochs to 30.

3.3. ConvNeXt

The ConvNeXt model was proposed in [12], and it is a pure convolutional model, inspired by the design of Vision Transformers, that claims to outperform them. Our training configuration here is the same as that of the EfficientNet setup except that we have replaced the model from EfficientNet with ConvNext.

Teacher Model	**Accuracy(%)**
AST	56.7
EfficientNet	55.2
ConvNext	53.4
SwinTrans	56.9
PaSST-Ensemble	**62.5**

Table 1: Accuracy results of different teacher models on the validation set

3.4. SwinTrans

The Swin Transformer is a type of Vision Transformer. It builds hierarchical feature maps by merging image patches in deeper layers and has linear computation complexity to input image size due to the computation of self-attention only within each local window. It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. The HTS-AT approach introduces the Swin Transformer block with a shifted window attention for sound classification and detection and shows good performance [7]. Our training configuration here is the same as that of the AST setup, but replace AST with SwinTrans.

3.5. PaSST

PaSST extends the AST model further by introducing a technique called patchout to tackle the quadratic scaling of attention layers concerning the sequence length and to improve the generalization of trained transformers. PaSST models are well suited for training on downstream tasks in a short amount of time resulting,

To train the PaSST model, the raw audio signal is downsampled using a sampling rate of 32kHz, and the input features are extracted from the raw audio signals using a Short Time Fourier Transformation (STFT) with a window size of 800 with 40% overlap. We apply a Mel-scaled filter bank to 128 frequency bins. We use the pre-trained multiple PaSST models provided by [10], and average the logits of all four different PaSST models to further improve the results.

3.6. Experiemnt results

As shown in Table 1, the PaSST ensembled model achieves the best performance with 62.5% accuracy on the development set. Due to its excellent performance, we adopt the PaSST-Ensemble as our teacher model in the following experiments and use it to produce the "soft labels" for knowledge distillation.

4. STUDENT MODEL

To meet the competition requirements, the student model needs to pursue the smallest possible model size while ensuring accuracy. The study in [13] shows that a Receptive Field Regularized Convolutional Neural Network (RFR-CNN) and its variants CP-ResNet performs well in previous editions of the DCASE challenge.

The raw audio signal is down-sampled using a sampling rate of 32 kHz and the input features are extracted from the raw audio signals using a Short Time Fourier Transformation (STFT) with a

	ACC(%)	Logloss	MEM	MACs
system 1	57.5	1.147	127.6K	28.8M
system 2	**58.1**	1.178	79.9K	21.4M
system 3	57.4	1.190	79.9K	21.4M
system 4	57.0	1.198	**63.5K**	**19.5M**

Table 2: Results of the quantized models on the provided development set split in terms of accuracy and validation loss

window size of 2048 and overlap of 36%. We apply a Mel-scaled filter bank to end up with 256 frequency bins. We experiment with mixing features and label information using Mixup [14] and mixing the style of the recordings using MixStyle [15].

For system 1, we strictly follow the CP-ResNet model structure described in [10]. The CNN's initial width is set to 32 channels. The number of channels for the next three residual blocks is 32, 64, and 92, respectively. The input feature size of the neural network is 1 × 256 × 44. After the convolution operation, a feature map of size 95 × 15 × 10 is obtained. Then, global pooling is performed to obtain a 92-dimensional feature and finally connected to a fully connected layer with 10 output nodes.

For system 2 and system 3, we modify the kernel size for the first convolutional layer in Stage 3 from 3 × 3 to 1 × 1. This adjustment can significantly reduce the number of model parameters and reduce the number of model calculations. The distinction between System 2 and System 3 lies in their training methodologies; System 2 is exclusively trained using Mixup, while System 3 does incorporate both Mixup and Mixstyle in its training process.

In System 4, we implement an additional modification by altering the kernel size of the initial convolutional layer in Stage 2 from from 3 × 3 to 1 × 1. This change contributes to a further reduction in the model's complexity.

5. QUANTIZATION

We use Post-Training Static Quantization as implemented in PyTorch to quantize all model parameters and perform all inference computations with 8-bit integers. We use a subset of the training data for calibration. We quantize all model parameters and the input data using the quantization stub inserted into the model's forward pass.

6. SUBMISSIONS AND RESULTS

The final results on the development set split are reported in Table 2. We use the official NeSsi toolkit to compute MEM and MACs parameters after the model quantization step.

7. CONCLUSION

In this technical report, we described the Tencent submission to Task 1 of the DCASE 2023 challenge. We tried several different teacher models and choose an ensemble of the audio spectrogram transformer PaSST as our teacher model. We then tried to compress the knowledge into a low-complexity CP-ResNet student model, while maintaining as much of the predictive performance as possible. Finally, our quantized 8-bit light student model achieves 57.0% accuracy with lower model complexity compared with the official baseline.

8. REFERENCES

[1] 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

[2] 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

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] Y. Gong, Y.-A. Chung, and J. Glass, "PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.

[5] ——, "AST: Audio Spectrogram Transformer," July 2021.

[6] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," Mar. 2022.

[7] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 646–650.

[8] Y. Gong, S. Khurana, A. Rouditchenko, and J. Glass, "CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification," Mar. 2022.

[9] F. Schmid, K. Koutini, and G. Widmer, "Efficient Large-scale Audio Tagging via Transformer-to-CNN Knowledge Distillation," Nov. 2022.

[10] F. Schmid, S. Masoudian, K. Koutini, and G. Widmer, "CP-JKU submission to dcase22: Distilling knowledge for low-complexity convolutional neural networks from a patchout audio transformer," DCASE2022 Challenge, Tech. Rep., June 2022.

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," Mar. 2015.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Sept. 2020.

[13] K. Koutini, H. Eghbal-zadeh, and G. Widmer, "Receptive Field Regularization Techniques for Audio Classification and Tagging with Deep Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1987–2000, 2021.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond Empirical Risk Minimization," Apr. 2018.

[15] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain Generalization with MixStyle," Apr. 2021.