Hybrid-attention-module-based Audio Spectrogram Transformer for Audio Classification

Lingqing Liu, Xiao Zhang and Shoji Makino

Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
E-mail: {lingqing_liu@moegi., zhang_x07@toki., s.makino@}waseda.jp

Abstract

Audio classification is a fundamental task in the audio processing field, aiming to categorize different kinds of audio. Audio Spectrogram Transformer (AST) achieves good performance with a fully attention-based convolution-free architecture. However, it has the weakness of extracting local information effectively due to its pure Transformer-based structure. To solve this problem, we propose a hybrid attention module (HAM) to assist the training process of the conventional AST model. The proposed HAM integrates convolution with self-attention mechanisms to capture both local and global features in a single training epoch. Additionally, the dual pooling strategy is involved in the proposed HAM, which makes the model more focused on important features for classification. The experimental results on the ESC-50 dataset and the Speech Commands V2 dataset showed that our proposed HAM-based AST model outperforms the conventional AST model in the classification task.

1. Introduction

Audio classification models have achieved large improvements based on neural networks. Early approaches to audio classification are mainly based on Convolutional Neural Networks (CNNs). These models, inspired by different classical CNN image classification architectures such as AlexNet, VGG and ResNet, have demonstrated a success in large-scale audio classification tasks [1]. However, CNN-based methods are often hard to comprehensively capture long-range information and global context, which are essential for training models for classification. To address this limitation, attention-based models have emerged as a promising method. AST [2] is a fully attention-based model for audio classification that does not rely on CNNs. By utilizing a self-attention mechanism, its simple architecture achieves strong performance in the audio classification task. Despite its success, the conventional AST model has the limitation of inadequate extraction of local features. More comprehensive audio features are helpful for model training, thereby improving classification accuracy.

The Conformer [3] resolves the limitation of Transformer in local feature extraction by integrating CNNs and Transformers. This method enables the model to capture both global and local features, achieving strong performance in the audio field. In addition to global and local feature capture, the Squeeze-and-Excitation Network weights the channels according to the importance of the features. At the same time, it introduces an attention mechanism that allows the network to focus on the target source, thus improving feature capture [4]. This approach proves to be beneficial for model training.

Besides, the channel attention mechanism (CAM) [5] allows the network to focus on the target in noisy observation, and the dual channel pooling strategy [6] shows the effectiveness of capturing different levels of contextual information for the Transformer-based methods. With these benefits, a network that assigns weights to features according to their importance has been proposed to further enhance the ability to capture features [7]. It has been proven that using the attention mechanism can enable models to capture more effective features for model training.

Taking these into account, in this paper, we propose a hybrid attention module (HAM) which integrates convolution with the multi-head self-attention mechanism. It can capture local and global features simultaneously, thereby assist training process of the model. Moreover, to refine the extracted features, the channel attention mechanism and the dual pooling strategy have also been introduced in this structure. By weighting feature channels based on their relative importance, the model can enhance sensitivity to different features and focus more on the target audio source. The proposed HAM-based AST model has been validated through simulations, and results demonstrate that the proposed method outperforms the conventional AST model.

2. Conventional Method

Audio Spectrogram Transformer (AST) is a purely Transformer-based model for audio classification. It directly processes audio spectrograms as input and uses a patch-based embedding approach similar to the Vision Transformer (ViT) [10].

The multi-head self-attention mechanism enables AST to focus on different regions of the spectrogram, therefore improving its ability to capture global information. Additionally, layer normalization and residual connection used in each Transformer block of AST ensure stable training and efficient gradient flow. Then, the output of the Transformer layer passes through the classification head to produce the classification label. It has demonstrated excellent performance on tasks such as environmental sound classification and speech recognition, showing its effectiveness in handling diverse audio classification challenges.

3. Proposed Method

In this paper, we propose a new AST model based on the hybrid attention module. The proposed HAM-based AST model captures local features and selects channels while ensuring its global information capture capability.

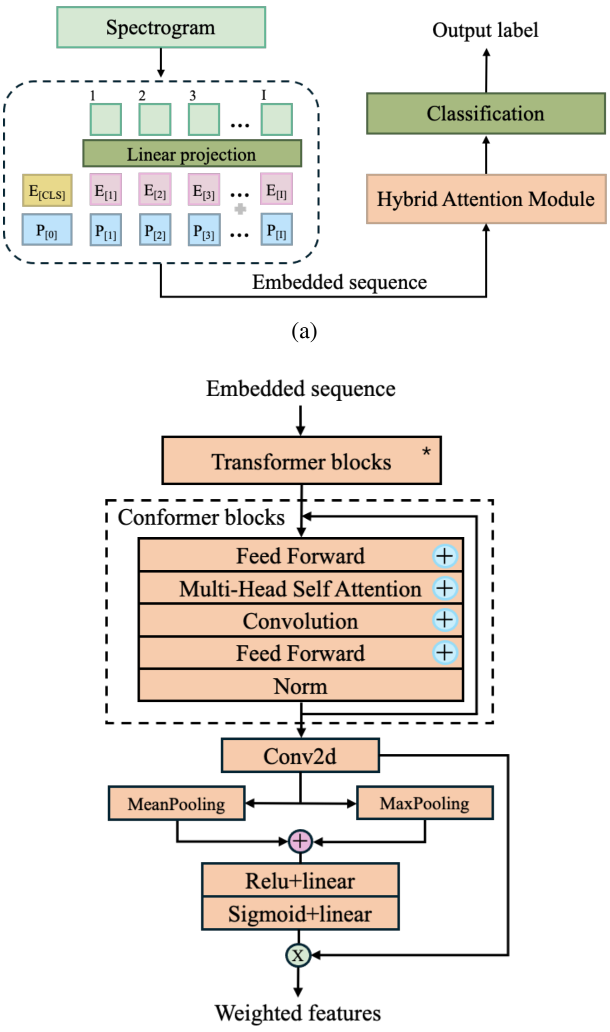
3.1 Overall Process Flow

Fig. 1a illustrates the overview process of the proposed HAM-based AST method. The input spectrogram is first divided into small, overlapping patches of shape 16×16, leading to I patches. These patches are flattened and then passed through a linear projection layer, therefore creating patch embeddings E_i for $i = 1, \ldots, I$. At the beginning of the embedding sequence, a CLS token is added. The final state of the CLS token is used as the aggregated sequence representation for the classification task. Then, these patch embeddings E_i are combined with positional embeddings P_i for $i = 0, \ldots, I$ to form the embedded sequence. This sequence captures the initial information of the input spectrogram. The embedded sequence is then input into the proposed HAM module, which integrates different attention mechanisms to obtain comprehensive features. Finally, these refined features will be used for the audio classification task.

3.2 Hybrid Attention Module

To better integrate global and local information and capture important features during the training process, we propose a hybrid attention module (HAM), which combines self-attention mechanism, convolution mechanism and channel attention mechanism. The structure of the proposed HAM module is shown in Fig. 1b. The HAM module takes the embedded sequence from the initial spectrogram patch processing as input and progressively refines the features through a series of computational blocks.

The module begins with Transformer blocks, which are utilized to initially extract global information from the input spectrograms. The multi-head self-attention mechanism of Transformer blocks enables the model to effectively capture information across distant temporal segments. The output of the Transformer blocks, referred to as $E_{\text{AST}} \in \mathbb{R}^{B \times N \times D}$, serves as an intermediate representation in the HAM module, where N represents the number of patches and D represents the embedded dimension.



(a)

(b)

Figure 1: (a) Overview of the proposed hybrid-attention-module-based AST model. (b) Architecture of the proposed hybrid attention module (HAM). Blue circle represents residual connection. Pink circle represents add. Green circle represents multiply.

Subsequently, E_{AST} passes through the Conformer blocks. These Conformer blocks are designed to supplement the global information obtained from Transformer blocks and the local information. For each Conformer block, it employs a combination of a multi-head self-attention module, a convolutional module, two feed-forward modules and a normalization layer within its architecture. These modules are connected using residuals. The multi-head self-attention module is used to refine the extraction of global features, while the convolution module enables the model to have the ability to extract local features for assisting in model training. We use $E_{\text{CON}} \in \mathbb{R}^{B \times N \times D}$ to represent the output sequence of Conformer.

The magnitude of the value of audio features can be mapped to the importance of the feature. We expect to improve the attention of the model to important features, and adjust the importance of individual feature channels dynamically. After preliminary extraction of audio features and getting E_{CON}, we use the dual channels design which is inspired by the design of conventional CAM++ [7]. This design utilizes max pooling and mean pooling in parallel to capture features to assist model training. Max pooling highlights the most significant features by selecting the maximum value within the feature map. This method effectively captures sharp information such as edges or corners, but might ignore less obvious but useful features. In contrast, mean pooling captures global information by computing the mean value in the feature map. Mean pooling is able to complement the global information that may be missed by max pooling.

The outputs of the dual pooling operations are combined to provide a comprehensive representation of feature importance. This combined representation utilizes normalize to adjust feature scales, ensuring stability in subsequent processing. Finally, we weight these features and take the weighted features as the final input features of the model training process for audio classification tasks.

4. Experiments

We evaluated the proposed HAM-based AST method on the ESC-50 [8] dataset and Speech Commands V2 [9] dataset. To verify the performance of our proposed HAM-based method, we compared the results with the conventional AST model [2]. For a better comparison with the conventional AST model, we used the same experimental settings as it in our experiments. To ensure the fairness and reliability of the experiment, we did experiments more than three times on each dataset and took the average result into account. Besides, we conducted a series of ablation experiments to further investigate the effect of different attention settings of our proposed method on the classification accuracy.

4.1 Datasets

ESC-50 dataset is the environmental sound dataset. It consists of 2,000 environmental audio clips that last five seconds each, and covers 50 semantical classes with 40 examples per class. The Speech Commands V2 (SC-V2) dataset is the speech dataset. It contains about 105,000 audio clips that last 1 second each. These clips represent simple spoken commands divided into 35 classes.

Table 1: Experimental Setup

Settings	**ESC-50**	**SpeechCommands-V2**
Num of classes	50	35
Initial learning rate	1e-5	2.5e-4
Epochs	25	30
Batch size	48	128
Pretraining	Audioset and ImageNet	
Loss	Cross Entropy	Binary Cross Entropy
Mixup	0	0.6

Table 2: Classification Accuracy of Conventional AST and Proposed HAM-based Method

Model	**ESC-50**	**SpeechCommands-V2**
AST	95.25%	96.32%
HAM (Proposed)	**96.41%**	**97.28%**

4.2 Training Settings

Table 1 shows the detailed training setup in our experiments. To achieve a more accurate comparison with the conventional AST method, we employed the same experimental settings as it. In particular, we directly used the pre-trained weights of the AST to reduce the computational resources and maintain the compatibility. The pre-trained AST models are trained on the Audioset [11] and ImageNet [12] datasets.

4.3 Results

Table 2 shows the comparison classification accuracy between our proposed HAM-based AST method and the conventional AST model on two datasets. On the ESC-50 dataset, the HAM-based model achieves an accuracy of 96.40%, which represents an improvement of 1.16% over the baseline AST model. On the SC-V2 dataset, the HAM model reaches an accuracy of 97.28%, outperforming AST by 0.96%.

Table 3 shows the results of different model configurations, where “# Conformer” represents the number of Conformer blocks. The data in the first two lines explore the effect of channel attention mechanism. By introducing the channel attention mechanism alone, the performance improves slightly to 95.33% on the ESC-50 dataset and 96.51% on the SC-V2 dataset, respectively. The data in the following two lines explore the effect of different number of Conformers on the classification accuracy when the channel attention mechanism has already existed. In cases where the channel attention mechanism has already existed, increasing the number

Table 3: Impact of the Number of Conformer Blocks and the Integration of Channel Attention Mechanism on Classification Accuracy at Different Settings

# Conformer	CAM	ESC-50	SpeechCommands-V2
0	–	95.25%	96.32%
0	✓	95.33%	96.51%
3	✓	95.73%	96.90%
6	✓	96.41%	97.28%

of Conformer blocks to six, the accuracy achieves 96.41% on ESC-50 and 97.28% on SC-V2, outperforming the model with three Conformers by 0.68% and 0.38% respectively, and outperforming the model without Conformers by 1.08% and 0.77% respectively.

4.4 Discussion

The experimental results show that the model based on HAM we proposed improved the classification accuracy on different types of audio datasets, validating our hypothesis and confirming that the hybrid attention module is effective to improve the audio classification performance.

While keeping the number of Conformer blocks fixed, the integration of the channel attention mechanism increased classification accuracy, proving the effectiveness of the channel attention mechanism. This may be because the channel attention mechanism supplements previously extracted information, therefore refining the feature space and improving classification. As the number of Conformer blocks increases in the case of the model with the existing channel attention mechanism, the accuracy of classification consistently increases. It indicated that Conformer blocks are useful to capture both local and global features by convolutional mechanism and self-attention mechanism respectively, making the model more effective in distinguishing between different classes of audio signals.

5. Conclusions

In this work, we propose a novel HAM-based model for audio classification tasks. The proposed HAM module can not only capture global and local information simultaneously, but also enhance the extraction of vital features for classification, solving the limitations of a single attention mechanism. We evaluated it on different types of audio datasets, outperforming the conventional AST model by 1.16% on ESC-50 and 0.96% on SC-V2, respectively. This improved performance shows the effectiveness of hybrid attention for feature extraction. Furthermore, ablation experiments confirm that the structure with hybrid attention outperform those relying on single attention mechanism, highlighting the importance of combining diverse attention strategies for audio classification tasks.

References

[1] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE ICASSP*, 2017, pp. 131–135.

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

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[5] Y.-Q. Yu, S. Zheng, H. Suo, Y. Lei, and W.-J. Li, "CAM: Context-aware masking for robust speaker verification," in *Proc. IEEE ICASSP*, 2021, pp. 6703–6707.

[6] C.-H. Tan, Q. Chen, W. Wang, Q. Zhang, S. Zheng, and Z.-H. Ling, "PoNet: Pooling network for efficient token mixing in long sequences," in *Proc. ICLR*, 2022.

[7] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," in *Proc. Interspeech*, 2023, pp. 5301–5305.

[8] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. ACM MM*, 2015, pp. 1015–1018.

[9] P. Warden, "Speech Commands: A dataset for limited-vocabulary speech recognition," arXiv preprint arXiv:1804.03209, 2018.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," CoRR, vol. abs/2010.11929, 2020.

[11] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017, pp. 776–780.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.