Meta-Learning with Pretrained Audio Representations Enables One-Shot Acoustic Signal Classification

Haoxiang Wu*, Zhengqiao Zhao*, Jingdong Chen*, and Jacob Benesty†
* Center of Intelligent Acoustics and Immersive Communications
Northwestern Polytechnical University, Xi'an, China
Email: zhengqiao.zhao@nwpu.edu.cn
† University of Quebec, Montreal, Canada

***Abstract*—Few-shot acoustic signal classification remains a challenging problem due to the high diversity and variability of acoustic data and limited availability of labeled samples. While pretrained audio classification models have proven effective for various acoustic signal classification tasks, fine-tuning them can still lead to overfitting in low-resource settings. In this work, we proposes an attention-based meta-learning framework that operates on the hidden states of a pretrained audio classification model. Specifically, we introduce a trainable hierarchical additive attention module to extract meaningful features from the hidden states of a large-scale pre-trained Audio Spectrogram Transformer (AST). The attention mechanism is trained with a simple meta-learning paradigm, enabling effective adaptation to one-shot learning tasks. We evaluates the proposed model on multiple acoustic signal classification tasks, including acoustic scene classification, sound event recognition and underwater vessel noise classification. Experimental results demonstrate that our proposed framework substantially outperforms the existing methods such as CNN-based prototypical networks in terms of one-shot classification accuracy. This research not only provides an efficient solution for low data resource acoustic pattern recognition tasks but also demonstrate the strong potential of pre-trained audio classification models when combined with meta-learning framework for few-shot learning.**

I. Introduction

Audio classification technology has broad applications in areas such as smart homes, environmental monitoring, and marine exploration. Conventional deep learning approaches typically rely on large-scale annotated datasets, making them prone to overfitting in few-shot learning scenarios. In practice, however, obtaining sufficient labeled data remains a major challenge. Metric learning frameworks such as Prototypical Networks [1] mitigate this issue by modeling inter-class relationships, but their feature extractors, usually Convolutional Neural Networks (CNNs), struggle to capture robust audio representations in complex acoustic environments.

Recently, pre-trained transformer models have shown remarkable feature extraction capabilities across domains including computer vision, natural language processing, and acoustic signal processing. Models such as Audio Spectrogram Transformer (AST) [2], wav2vec [3], BEATs [4], and HuBERT [5], trained on large-scale audio corpora, have achieved state-of-the-art performance in supervised audio classification. However, their potential in few-shot learning, where models must generalize from only a handful of labeled samples, remains underexplored. Leveraging pre-trained audio transformers within few-shot learning frameworks is therefore a key step toward advancing few-shot audio recognition.

In this work, we propose a meta-learning model that incorporates a pre-trained AST backbone for few-shot audio classification. Specifically, we introduce a meta-trained hierarchical attention module that effectively exploits the pretrained latent representations in one-shot settings, consistently outperforming competing approaches across diverse acoustic classification tasks.

The remainder of this paper is organized as follows. Section II reviews related work on few-shot audio classification. Section III introduces the proposed hierarchical attention meta-learning framework. Section IV describes the experimental setup, datasets, and results, and Section V concludes with a summary and discussion.

II. Related Work

Few-shot learning aims to train models to recognize novel categories using only a handful of labeled examples per class, typically between 1 and 5, within an N-way K-shot framework. This paradigm has attracted considerable attention in fields such as image processing, computer vision, and audio classification.

In image processing, significant progress has been achieved through diverse learning paradigms. Matching Networks introduce an attention-based mechanism that dynamically aggregates support samples via memory modules, enabling efficient one-shot adaptation [6]. MetaOptNet embeds a differentiable convex optimization layer (e.g., SVM) into the meta-learning loop, implicitly refining feature representations and achieving state-of-the-art performance on few-shot benchmarks [7]. Semantic Prompt advances multi-modal fusion by leveraging textual semantics to guide visual feature modulation through

979-8-3315-7206-8/25/$31.00

spatial attention, thereby enhancing focus on category-specific regions [8]. FSCE addresses few-shot object detection using a contrastive proposal encoding mechanism, which enforces tighter intra-class clustering and greater inter-class separation [9]. MAML introduces a model-agnostic meta-learning framework that allows rapid adaptation to new tasks through a few gradient updates [10].

In the audio domain, substantial progress has also been made. Heggan et al. systematically evaluated multiple few-shot learning algorithms across diverse everyday sound datasets [11]. Wang et al. constructed two specialized synthetic datasets for few-shot audio: FSD-MIX-CLIPS and FSD-MIX-SED, and conducted in-depth comparisons under complex auditory conditions [12]. Cheng et al. [13] adapted few-shot algorithms using a One-vs-Rest strategy, validating their effectiveness on the large-scale AudioSet dataset [14]. Liang et al. [15] introduced the Treff Adapter, a training-efficient module for CLAP [16], which employs a Cross-Attention Linear Model (CALM) initialized via cosine similarity. This approach enables effective training-free adaptation, surpassing metric-based methods while retaining zero-shot capability. Iqbal et al. [17] integrated Label Set Operation (LaSO) [18] features with prototypical networks for classroom audio classification, achieving superior accuracy by extracting LaSO features from scalograms using pretrained models.

Attention mechanisms have further enhanced few-shot learning by dynamically focusing on salient information, significantly improving feature discriminability in audio classification [19]–[24]. Building on this, Si et al. [25] explored attention-based methods in few-shot class-incremental audio classification. They proposed a multi-level embedding extractor (MEE) to derive meaningful representations from pretrained AST models. However, since the MEE module is trained and fine-tuned only during the base session and kept fixed in subsequent incremental sessions, its adaptability to new tasks remains limited.

III. PROPOSED METHOD

This paper proposes a meta-learning model for few-shot acoustic signal classification that leverages pretrained audio representations and a novel hierarchical latent attention mechanism . The detailed model structure and training strategy are described below.

A. Pretrained Audio Representations and Hierarchical Latent Attention Module

Audio Spectrogram Transformer (AST) is a transformer-based model for audio representation that adapts the Vision Transformer (ViT) architecture to audio by treating spectrograms as images, using patch-based processing to learn meaningful audio representations for downstream classification tasks. In AST, the input spectrogram are first divided into N patches of size $P_f \times P_t$, $N = \frac{F}{P_f} \times \frac{T}{P_t}$ where F is the total number of frequency bins, T is the total number of time frames. The patches are then processed by linear projection and positional encoding: $\mathbf{z}_i = \mathbf{E}\mathbf{p}_i + \mathbf{e}_i^{\text{pos}}$, where $\mathbf{E}$ denotes the learnable embedding matrix. Finally, the global audio representation $\mathbf{h}$ is given by the output embedding of the [CLS] token from the last layer, i.e., $\mathbf{h} = \text{Transformer}(\mathbf{Z})_{\text{[CLS]}}$ where $\mathbf{Z}$ is the input to the last transformer layer. In this work, to obtain robust audio features without requiring extensive task-specific data, our model utilizes a pretrained AST backbone (on ImageNet and AudioSet) with frozen weights. To further capture long-range temporal dependencies, we extract the hidden states from intermediate transformer layers (specifically, the outputs of Transformer layers 7 to 12) across 1,214 time steps. These hidden states are processed by a hierarchical latent attention module. First, a temporal attention mechanism [19] computes an adaptive, weighted summary for each layer, focusing on acoustically salient regions. Subsequently, a layer-wise attention mechanism aggregates these summaries into a single, final embedding. This two-stage hierarchy not only enhances focus on informative signal segments but also reduces the computational load and the number of trainable parameters. Figure 1 shows the overall structure of the proposed model.

B. Prototypical Network and meta-learning framework

We train our proposed model using an episodic training scheme [1]. For each episode, we construct a task with a query set and a support set with N classes and K samples per class (K shots). The model is optimized by minimizing the cross-entropy loss for query samples. Classification is performed by computing the Euclidean distance between a query embedding and each class prototype (derived from the support set), then applying a softmax over the negative distances to produce class probabilities. Specifically, for an N-way K-shot support set $\mathcal{S} = \{(x_1,\, y_1),\, \ldots,\, (x_{N\times K},\, y_{N\times K})\}$, the prototype for class k is calculated as the mean of embeddings from its support samples:

$$\mathbf{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(x_i, y_i)\in\mathcal{S}_k} f_\theta(x_i), \tag{1}$$

where $\mathcal{S}_k \subset \mathcal{S}$ denotes the subset of support samples belonging to class k, with $|\mathcal{S}_k| = K$. For a query sample x_q, the classification process involves:

1) Computing its embedding: $f_\theta(x_q)$.
2) Measuring distances to each prototype: $D(f_\theta(x_q), \mathbf{c}_k)$, e.g., the squared Euclidean distance:

$$D(f_\theta(x_q), \mathbf{c}_k) = \|f_\theta(x_q) - \mathbf{c}_k\|_2^2. \tag{2}$$

3) Generating probability distribution over classes using the softmax function, i.e.,

$$p_\theta(y = k|x_q) = \frac{\exp\big(-D(f_\theta(x_q), \mathbf{c}_k)\big)}{\sum_{k'=1}^{N} \exp\big(-D(f_\theta(x_q), \mathbf{c}_{k'})\big)}. \tag{3}$$

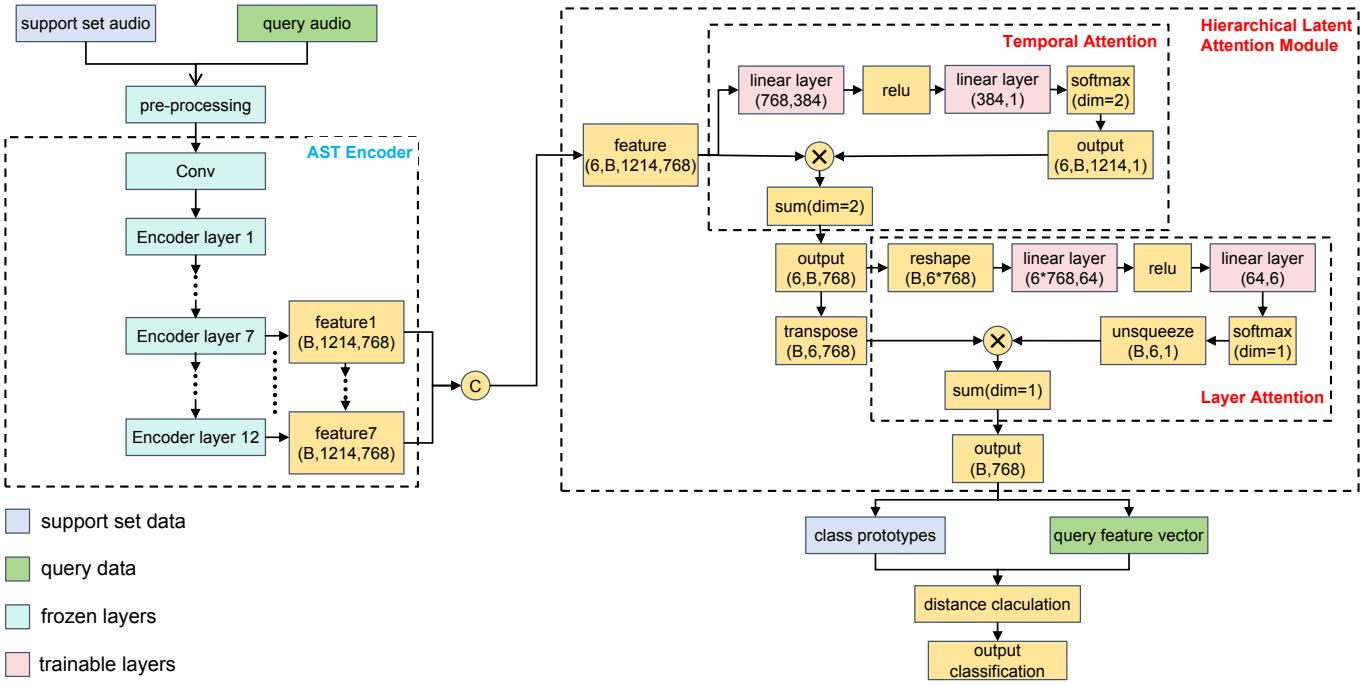

Fig. 1. The overview of the proposed few-shot learning model

IV. Experiment

A. Datasets

This study uses three datasets for experimental evaluation: the DCASE2018 Challenge Task 1 acoustic scene classification dataset ("Scenes") [26][1] , the DCASE2018 Challenge Task 5 monitoring of domestic activities dataset ("Events") [27][2] , and the Shipsear underwater vessel noise database ("Ships") [28]. All audio recordings are segmented into 10-second clips and resampled to 16 kHz for compatibility with the AST backbone.

Since the AST model is pretrained on the AudioSet dataset, some categories in our experimental datasets may overlap with those seen during pretraining. To fairly assess model performance, samples from overlapping classes are assigned to the training set, while samples from unseen classes are reserved for testing. Specifically, in the "Scenes" dataset, classes such as airport, metro_station, street_traffic, bus, metro, which appear in AudioSet, are used for training, while the remaining five (shopping_mall, street_pedestrian, public_square, park, and tram) are designated as test classes.

For the "Events" and "Ships" datasets, direct alignment with AudioSet ontology is less straightforward. Therefore, we construct the train–test split based on sample availability. In the "Events" dataset, four activities with sufficient samples (absence, watching_tv, working, cooking) are used for training, while five lower-frequency activities (social_activity, eating, other, dishwashing, vacuum_cleaner) are assigned to the test set. Similarly, in the "Ships" dataset, six vessel types with relatively large sample sizes (Motorboat, Mussel boat, Natural ambient noise, Ocean liner, Passengers, and RORO) comprise the training set, while the remaining six (Dredger, Fishboat, Pilot ship, Sailboat, Trawler, and Tugboat) form the test set.

For evaluation, a 5-way 1-shot learning setup is used for the "Scenes" and "Ships" datasets, while a 4-way 1-shot configuration is adopted for the "Events" dataset due to its smaller number of categories. In each few-shot task, the support set provides a single labeled example (one-shot) per class, and the query set contains ten samples per class for evaluation. Both sets are randomly sampled to simulate the data scarcity typically encountered in real-world applications.

B. Compared Models

We compare three main approaches: (1) the baseline Prototypical Network (PN); (2) a distance-based classification method that leverages global AST embeddings from the [CLS] token (AST+dist), assigning each query signal to the class of the nearest support example in the latent space; and (3) a meta-learning fine-tuning approach (AST+PN), where the pretrained AST model is frozen and two additional fully connected layers with 256 and 128 hidden units, respectively, are fine-tuned for classification within the meta-training framework.

C. Experimental Setup and Results

We employ a meta-training strategy with a learning rate of 10^{-4}. Each training episode uses a 1-shot support set and a 10-shot query set. We validate the model every 100 training

[1]https://dcase.community/challenge2018/task-acoustic-scene-classification.
[2]https://dcase.community/challenge2018/task-monitoring-domestic-activities.

TABLE I
N-WAY 1-SHOT CLASSIFICATION ACCURACY

Method	Scenes (N=5)	Events (N=4)	Ships (N=5)
PN	27.12% ± 0.72%	44.94% ± 1.21%	32.41% ± 0.81%
AST+dist	50.77% ± 0.97%	71.04% ± 1.12%	64.75% ± 0.82%
AST+PN	60.69% ± 0.80%	65.42% ± 1.30%	55.93% ± 0.91%
AST+ATT+PN	59.35% ± 0.69%	77.38% ± 1.09%	65.82% ± 1.05%

episodes on 100 validation tasks. The learning rate is halved if no improvement in accuracy is observed after 500 consecutive episodes and training is terminated if no improvement over 1000 consecutive episodes. The best-performing model is then evaluated on 10,000 testing episodes, with results reported as mean accuracy and standard deviation.

As shown in Table I, the proposed model consistently achieves performance that is comparable to, and often exceeds, competing methods across all three classification tasks. The pretrained AST model provides high-quality audio representations that support effective one-shot learning without additional training, as demonstrated by the strong results of the AST+dist method. In contrast, the baseline Prototypical Network (PN), which lacks access to pretrained representations, yields the weakest performance. Integrating the pretrained AST into the Prototypical Network (AST+PN) improves results, but the framework remains unable to fully exploit the rich latent features learned by the AST.

In comparison, our proposed approach (AST+ATT+PN), which combines the pretrained AST backbone, a hierarchical latent attention mechanism, and a meta-learning framework, achieves the highest performance. This demonstrates that the attention mechanism can be effectively trained within the meta-learning paradigm and leveraged to extract informative hidden states from the pretrained model for few-shot audio classification.

V. CONCLUSION

This paper presented a novel meta-learning framework that leverages pretrained audio representations. Instead of fine-tuning a pretrained model on extremely limited labeled data (e.g., a single sample per class), our approach introduces a hierarchical attention module trained with a simple yet effective meta-learning strategy to extract task-relevant latent representations tailored for one-shot learning. Specifically, the model directly utilizes hidden states from the intermediate layers of a pretrained AST and processes their temporal dependencies through the hierarchical attention module to preserve classification-relevant information. We evaluated the proposed framework on three widely used datasets and compared it against several competitive baselines. Experimental results demonstrate that our method consistently delivers substantial improvements in one-shot classification accuracy across diverse acoustic signal classification tasks. Although this study focused on one-shot settings, the results also suggest that the proposed training strategy effectively transfers general-purpose audio representations while simultaneously regularizing the model to mitigate overfitting—a key requirement in few-shot learning scenarios.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No. 62192713 and No. 62301442 and National Key Research and Development Program of China under Grant No. 2024YFF0505502.

REFERENCES

[1] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

[2] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575. DOI: 10.21437/Interspeech.2021-698.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[4] S. Chen, Y. Wu, C. Wang, *et al.*, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[5] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[6] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[7] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 657–10 665.

[8] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," *arXiv preprint arXiv:2303.14123*, 2023.

[9] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "Fsce: Few-shot object detection via contrastive proposal encoding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7352–7362.

[10] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, PMLR, 2017, pp. 1126–1135.

[11] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "Metaaudio: A few-shot audio classification benchmark," in *International Conference on Artificial Neural Networks*, Springer, 2022, pp. 219–230.

[12] Y. Wang, N. J. Bryan, J. Salamon, M. Cartwright, and J. P. Bello, "Who calls the shots? rethinking few-shot learning for audio," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2021, pp. 36–40.

[13] K.-H. Cheng, S.-Y. Chou, and Y.-H. Yang, "Multi-label few-shot learning for sound event recognition," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2019, pp. 1–5.

[14] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 776–780.

[15] J. Liang, X. Liu, H. Liu, *et al.*, "Adapting language-audio models as few-shot audio learners," *arXiv preprint arXiv:2305.17719*, 2023.

[16] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[17] R. Iqbal, C. Ritz, J. Yang, and S. Howard, "Few-shot audio classification model for detecting classroom interactions using laso features in prototypical networks," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2024, pp. 1–6.

[18] A. Alfassy, L. Karlinsky, A. Aides, *et al.*, "Laso: Label-set operations networks for multi-label few-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6548–6557.

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[20] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[21] Z. Wang and J. H. Hansen, "Audio anti-spoofing using a simple attention module and joint optimization based on additive angular margin loss and meta-learning," *arXiv preprint arXiv:2211.09898*, 2022.

[22] Z. Zhong, M. Hirano, K. Shimada, K. Tateishi, S. Takahashi, and Y. Mitsufuji, "An attention-based approach to hierarchical multi-label music instrument classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[23] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 316–320.

[24] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[25] Y. Si, Y. Li, J. Tan, Q. He, and I.-Y. Kwak, "Fully few-shot class-incremental audio classification using multi-level embedding extractor and ridge regression classifier," *arXiv preprint arXiv:2506.18406*, 2025.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Nov. 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840.

[27] G. Dekkers, S. Lauwereins, B. Thoen, *et al.*, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 32–36.

[28] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "Shipsear: An underwater vessel noise database," *Applied Acoustics*, vol. 113, pp. 64–69, 2016.