Adapting Language-Audio Models as Few-Shot Audio Learners

Jinhua Liang[⋆]*, Xubo Liu*[*]*, Haohe Liu*[*]*, Huy Phan*[†]*, Emmanouil Benetos*[⋆‡]*,*
Mark D. Plumbley[*]*, Wenwu Wang*[*]

[⋆] Centre for Digital Music, Queen Mary University of London, UK
[*] Centre for Vision, Speech and Signal Processing (CVSSP)
[†] Amazon Alexa [‡] The Alan Turing Institute, UK

{jinhua.liang, emmanouil.benetos}@qmul.ac.uk, huypq@amazon.co.uk,
{xubo.liu, haohe.liu, m.plumbley, w.wang}@surrey.ac.uk

Abstract

Contrastive language-audio pretraining (CLAP) has become a new paradigm to learn audio concepts with audio-text pairs. CLAP models have shown unprecedented performance as zero-shot classifiers on downstream tasks. To further adapt CLAP with domain-specific knowledge, a popular method is to finetune its audio encoder with available labelled examples. However, this is challenging in low-shot scenarios, as the amount of annotations is limited compared to the model size. In this work, we introduce a **Tr**aining-**eff**icient (Treff) adapter to rapidly learn with a small set of examples while maintaining the capacity for zero-shot classification. First, we propose a cross-attention linear model (CALM) to map a set of labelled examples and test audio to test labels. Second, we find initialising CALM as a cosine measurement improves our Treff adapter even without training. The Treff adapter outperforms metric-based methods in few-shot settings and yields competitive results to fully-supervised methods.

Index Terms: Contrastive language-audio pretraining, few-shot learning, domain adaptation, audio classification

1. Introduction

Learning new concepts from a small set of examples is challenging in machine learning. It is a de-facto issue in audio domains where high-quality labels are more labor-intensive to obtain. While existing few-shot algorithms exploit available annotations by directly learning in the few-shot setting (i.e., n-way k-shot problem) [1, 2], contrastive language-audio pretraining (CLAP) provides a new paradigm to learn audio concepts using large-scale audio-text pairs. CLAP has shown an impressive capacity for zero-shot knowledge transfer in audio classification [3, 4, 5], largely due to the use of a large-scale datasets and its different training objective.

To further adapt CLAP to downstream datasets, existing methods directly finetune the audio encoder with examples in target domains [4, 5]. However, finetuning is ill-suited for few-shot settings because the number of available examples is extremely small with respect to the number of model parameters. In addition, updating the parameters of CLAP's audio encoder with discriminative learning would break the connection to its language encoder. A natural question is: *can we bootstrap a CLAP in few-shot scenarios while maintaining its capacity as a zero-shot classifier?*

In this paper, we propose the Training-efficient (Treff) adapter to bootstrap CLAP models by bridging zero-shot knowledge transfer and few-shot finetuning. Our Treff adapter contains two core designs: a cross-attention linear model (CALM) and cosine initialisation. First, we devise CALM where a set of support audio clips and test audio clips are mapped to their probability distributions. Second, we find that with a proper initialisation, using CALM as a cosine similarity measurement yields a better performance even without training. Figure 1 summarises performance of various adaptation approaches on the ESC-50 dataset v.s. the number of examples (i.e., shots) per class. Our Treff adapters outperform other adaptation approaches in few-shot scenarios when the number of examples per classes is less than 16. As the number of shots increases, the proposed Treff adapter can still have a competitive result to the finetuning approach. We find that our empirical study is consistent with some previous findings on different datasets [6, 7]. Our experimental results demonstrate that the proposed Treff adapter outperforms zero-shot classification by a large margin with only a small set of annotated data. In addition, our Treff adapter finetuned with half the number of shots can have a comparable performance with state-of-the-art fully-supervised learners. We speculate this work would shed light on domain adaptation in data-limited scenarios. The contributions in this work are summarised below:

- We introduce the Treff adapter to further improve CLAP performance with a small set of examples while preserving its ability to zero-shot classification. To our knowledge, this is the first work to study audio-language models in the few-shot audio domain.
- We propose CALM to retrieve labels of the test examples by measuring the affinity between test and support embeddings.

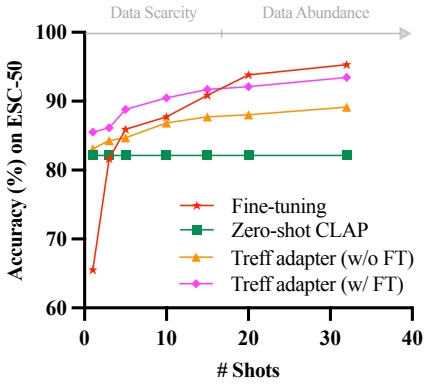


Figure 1: ***Adaptation accuracy on the ESC-50 dataset v.s. numbers of shots across various models***. *Our Treff adapters outperform other adaptation approaches in data-scarce scenarios and are comparable to finetuning with abundant data. We find that current approaches to domain adaptation yield a poor performance in low-shot settings. Finetuning is even worse than zero-shot knowledge transfer when fewer than 3 examples are available, largely due to overfitting. This motivates our work on few-shot domain adaptation.*

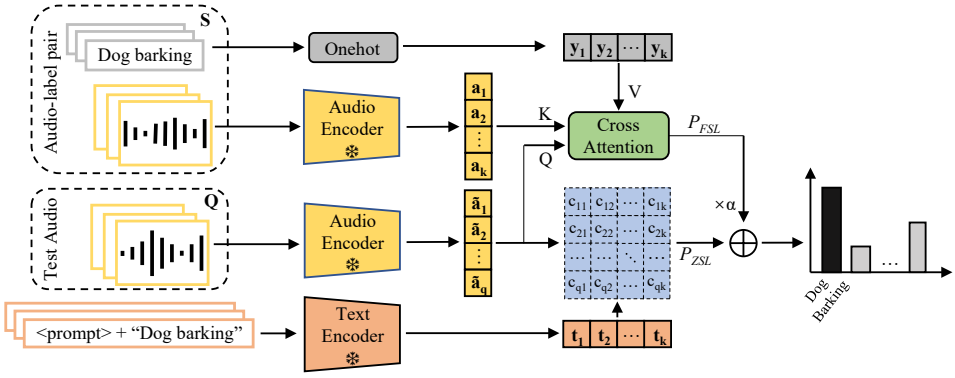


Figure 2: ***Overall framework of Treff adapter*** *to classify q audio clips in an n-way k-shot setting. For the few-shot classification results P_{FSL}, support embeddings $\mathbf{a}$ and test embeddings $\tilde{\mathbf{a}}$ are extracted with a frozen audio encoder separately. Test embeddings $\tilde{\mathbf{a}}$ then attend to support embeddings $\mathbf{a}$ and their corresponding one-hot labels $\mathbf{y}$ in the CALM. To get the zero-shot transferred knowledge P_{ZSL}, candidate labels concatenated with a fixed prompt (e.g., "This is a sound of ") are encoded by the frozen language encoder to get text embeddings $\mathbf{t}$. Coefficient $\{\mathbf{c}_{qk}\}$ is then measured by calculating cosine similarity between test audio embedding $\tilde{\mathbf{a}}_q$ and text embedding $\mathbf{t}_k$. The overall result is yielded by combining P_{FSL} and P_{ZSL} with a learnable parameter α.*

We show that the CALM can efficiently learn from few-shot settings while maintaining the semantic information of the pretrained feature representation.

- We devise a cosine initialisation strategy for CALM. We find that our Treff adapter benefits from this strategy in the few-shot settings even without training. The proposed Treff adapter in training-free version outperforms the zero-shot classifier by over 5% accuracy on the ESC-50 dataset.

2. Related work

Contrastive cross-modality pretraining. CLAP [4, 5] follows the idea of Contrastive Image-Language Pretraining (CLIP) [8] where contrastive learning is applied to train a cross-modal retrieval model with a large number of image-text pairs. The cross-modal retrieval model is trained so that multi-modal embeddings from the same pair are moved closer, otherwise, pushed away. Therefore, it can classify audio clips by measuring the similarity between multi-modal embeddings without additional examples (referred as zero-shot learning, ZSL). Such cross-modality models, however, cannot benefit from extra labelled data during inference. To further adapt cross-modality retrieval models with available examples, learnable prompts are devised to reduce CLIP's reliance on prompt engineering [9, 10]. Still, these adaptation methods require a large amount of labelled data for finetuning. The training-free adapter for CLIP (TIP-adapter) was proposed to learn a cache model to map test audio to their labels and initialise the cache model with support audio embeddings to avoid training [11]. It initialised the weights of the cache model with the values of support embeddings for rapid learning. The TIP-adapter, however, inevitably jeopardises the semantic information of support examples in the gradient-based optimisation. Cross-modal few-shot learning is designed where the examples from other modalities are treated as additional examples [12]. Although this cross-modal method showed promising results in few-shot classification, it requires updating linear classifiers and multiple encoders for the best performance, which takes additional computational resources.

Metric-based few-shot learning. Few-shot learning (FSL) aims to train a generic encoder via n-way k-shot problems where a classifier to predict the probability distribution of test data (queries) over n different classes and each class has k labelled objects (support shots or shots) for reference. Let $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ be a data point in the dataset $\mathcal{D}$. Suppose $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{nk}$ be a (small) support set of nk examples. Given a test audio clip $\tilde{\mathbf{x}}_q \in \mathcal{Q}$ where $\mathcal{Q}$ is the test set, a few-shot paradigm is expected to predict categorical probabilities $P(\tilde{\mathbf{y}}_q)$:

$$P(\tilde{\mathbf{y}}_q) = \mathcal{M}_\theta(\tilde{\mathbf{y}}_q | \tilde{\mathbf{x}}_q, \mathcal{S}) \tag{1}$$

where $\mathcal{M}_\theta$ is any mapping function and θ represents the model parameters. Metric-based FSL aims to cluster data points as per their categories in the emebedding space. Matching Networks map a set of support data and queries to their labels [1]. Prototypical networks learn prototypes to represent different classes and measure the l_2 distance between queries and prototypes in the embedding space [2]. Subsequent works attempted to apply prototypical networks to multi-label cases [13]. Siamese networks take two different data as input and learn the distance between them explicitly [14].

3. Treff adapter for CLAP

3.1. Overall framework

We hereby introduce the Treff adapter, a simple adaptation method to boost audio classification performance of the CLAP model by leveraging few-shot learning.

As shown in Fig. 2, the Treff adapter makes use of audio-label pairs and zero-shot knowledge by aggregating logits of FSL and ZSL. For FSL, the Treff adapter extracts audio embeddings $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{nk}] \in \mathbb{R}^{nk \times d}$ and text embeddings $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n] \in \mathbb{R}^{n \times d}$ where $\mathbf{a}$ and $\mathbf{t}$ are column vectors of $\mathbf{A}$ and $\mathbf{T}$, respectively, and d is the dimension of the embedding in the shared audio-language space. The test audio embedding $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \ldots, \tilde{\mathbf{a}}_q] \in \mathbb{R}^{q \times d}$ is likewise extracted by the audio encoder and then mapped to its label with the support audio-label pairs using CALM. The resulted affinity matrix $\mathbf{S} = \{s_{ij}\}$ is then combined with one-hot labels of support audio for few-shot classification. In the other branch, the cosine distances between the query audio embedding $\tilde{\mathbf{A}}$ and the text embedding $\mathbf{T}$ are measured and utilised for zero-shot classification explicitly. The overall inference for test audio $\tilde{\mathbf{y}}_q$ aggregates logits from zero-shot learning (P_{ZSL}) and few-shot learning (P_{FSL}):

$$P_{overall}(\tilde{\mathbf{y}}_q) = P_{ZSL}(\tilde{\mathbf{y}}_q | \tilde{\mathbf{x}}_q, \mathbf{T}) + \alpha \cdot P_{FSL}(\tilde{\mathbf{y}}_q | \tilde{\mathbf{x}}_q, \mathcal{S}) \tag{2}$$

where α is a trainable parameter and we initialise it with 1.0 in our experiments.

Our Treff adapter has two core designs: CALM and cosine initialisation. We illustrate how the CALM structure retrieves the label of a query audio clip using a small amount of support audio examples in Section 3.2; and in Section 3.3, we show that our Treff adapter can function as a similarity measurement without finetuning by adopting a simple initialisation.

3.2. CALM

Inspired by recent progress in attention mechanisms [15], we devise CALM by using the KQV-attention [15]. Different from common cross-attention modules where K and V are from the same modality [16, 17], CALM uses bi-modality information from labelled examples. Figure 3 compares structures of the cache model [11] and our CALM where Q and K denote the query and support audio embeddings, respectively, and V denotes the one-hot labels of the support examples. The query embedding $\tilde{\mathbf{a}}_i$ and the support embedding $\mathbf{a}_j$ are normalised and then fed into a trainable linear layer, separately

$$\tilde{\mathbf{e}}_i = f_{\mathbf{W}}(\tilde{\mathbf{a}}_i) = \frac{\tilde{\mathbf{a}}_i}{|\tilde{\mathbf{a}}_i|}\mathbf{W}^{\mathsf{T}}, \tag{3}$$

$$\mathbf{e}_j = f_{\mathbf{W}}(\mathbf{a}_j) = \frac{\tilde{\mathbf{a}}_j}{|\mathbf{a}_j|}\mathbf{W}^{\mathsf{T}}, \tag{4}$$

where $\mathbf{W}$ is the weight in the trainable layer. We note that the linear layer shares the weights across these two embeddings. The motivation is that sharing the weights reduces the trainable parameters in the finetuning. The linear layer maps the audio embedding to a new space where the embeddings from the same class get closer to each other. The affinity coefficient is then calculated by

$$s_{ij} = \tilde{\mathbf{e}}_i^{\mathsf{T}}\mathbf{e}_j, \tag{5}$$

The output of the CALM is obtained by

$$\mathbf{o}_i = \sum_j \varphi(s_{ij}) \cdot y_j, \tag{6}$$

where $\mathbf{o}_i \in \mathbb{R}^n$ is the probability distribution over the i-th test audio example; $\varphi(\cdot)$ is a scaling function controlling the sharpness of the similarity coefficient

$$\varphi(x) = \exp(b(1-x)) \tag{7}$$

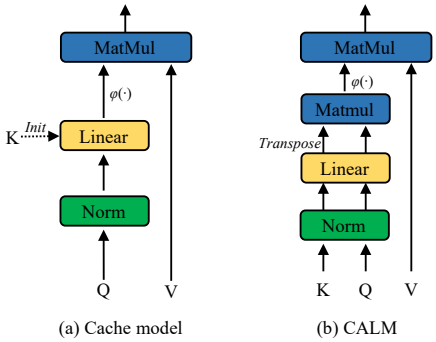
where b is a temperature factor, set empirically as 5.5 in our experiments.

3.3. Training-efficient adaptation

Training-free Treff adapter. The Treff adapter endows CLAP models with the ability to learn new, domain-specific knowledge from a small set of labelled examples. We prove that the Treff adapter can boost the model performance in a few-shot setting even *without* gradient-based optimisation. Inspired by the TIP-adapter [11], a generic encoder pretrained on a large-scale database should be good enough to cluster data points as per their semantic information. We replace the weights of the trainable linear layer with an identity matrix when no fine-tuning is available. Therefore, coefficients s_{ij} in the affinity matrix would be

$$s_{ij} = \left(\frac{\tilde{\mathbf{a}}_i}{|\tilde{\mathbf{a}}_i|}\right)^{\mathsf{T}}\frac{\mathbf{a}_j}{|\mathbf{a}_j|}, \tag{8}$$

In this case, the non-trainable version of CALM is actually a cosine-similarity function between the query and support embedding (i.e., $\cos(\tilde{\mathbf{a}}_i, \mathbf{a}_j)$). We found that even the cosine-similarity scores between the audio embeddings are capable of indicating how close those audio examples are.



(a) Cache model (b) CALM

Figure 3: ***Cache model (left) and our proposed CALM (right).*** *The Q and K values are the query and support audio embeddings respectively, while the V values are the one-hot labels associated to support audio examples.*

Finetuning the Treff adapter. While the training-free Treff adapter outperforms the performance of zero-shot knowledge transfer, it is still not comparable with finetuning and other gradient-optimization methods (e.g., the TIP-adapter [11]). We thus optimise the weights of the Treff adapter with the support examples only. Specifically, we initialise the weights of trainable linear layer with the identity matrix and calculate the affinity matrix among the support audio embeddings for the training. The weights of the linear layer are updated using the cross-entropy between the predicted logits and the ground truth of these support examples. Finally, we apply the updated weights to infer the labels of the test audio clips.

3.4. Relations with existing works

Relation with TIP-adapter. The TIP-adapter was proposed for the image domain whereas our Treff adapter is proposed for audio. In addition, our Treff adapter differs from the TIP-adapter on how the trainable parameters are finetuned. While the TIP-adapter applied a cache model to learn key values in the few-shot knowledge retrieval, our Treff adapter trains a linear layer to transform both query and support audio embeddings to a new space where audio embeddings belonging to the same class get closer than those in the original space. Therefore, the TIP-adapter will lose the information from the support embeddings due to the gradient-optimization while our Treff adapter preserves the semantic information of the support examples during the training. The experimental results in Table 2 showed that the information retained is beneficial, especially when the computational resource is limited. Meanwhile, when initialising the learnable parameters with appropriate methods and fixing them in the training, both our Treff adapter and TIP-adapter reduce to a cosine similarity measure.

Relation with Matching Network. While both Matching Networks and Our Treff adapter attempt to retrieve query labels with the affinity matrix between query and support embedding, the matching network [1] aims to train a generic audio encoder from scratch. Therefore, matching network keeps the affinity calculation fixed throughout the training process. In contrast, benefiting from pretrained audio-language models, the Treff adapter optimises the affinity measurement in the finetuning. This makes the Treff adapter yield a better performance than the matching network with less training resources.

Relation with Siamese network. Siamese networks [14] attempted to find a gradient-optimised distance measure between

two input examples. Our Treff adapter, however, transforms the embedding to a semantic space where the clusters of data points become more compact. In addition, the Siamese network learns implicitly the feature representation for which our proposed methods leverage a pretrained model.

4. Experiments and results

4.1. Experiment setup

We compare our methods with several sound datasets, including ESC-50 [18], FSDKaggle2018 [19], and FSD-FS [20]. Among them, FSD-FS is a dedicated subset of FSD50K [21] for multi-label few-shot audio classification.

To implement experiments, we resampled the audio recordings at 44.1kHz. The window length was about 20ms with 50% overlap, and the number of Mel bank filters was fixed to 64. Log-Mel spectrograms were used as input for few-shot learning methods. We followed the few-shot settings as per [13, 20]. More details can be found in our released code [1].

4.2. Experiment results

Table 1: ***Accuracy (%) of different methods for the ESC-50 and FSDKaggle18K datasets.*** *ZS denotes zero-shot performance, and FT is short for few-shot performance. * indicates our reproduced results. The best result on each dataset is marked with underline, and the best adaptation methods are highlighted in* ***bold****.*

Model	ESC-50	FSDKaggle18K
Topline systems		
AudioCLIP (FT) [3]	97.15	-
CLAP (FT) [4]	95.30*/96.70	-
HTS-AT [22]	97.00	-
Baseline systems		
VGGish (ZS) [23]	33.00	-
AudioCLIP (ZS) [3]	69.40	-
CLAP (ZS) [4]	82.60	-
Treff adapter (w/o FT)	87.75	71.03
TIP-adapter (w/ FT) [11]	91.45*	83.23*
Treff adapter (w/ FT)	**92.21**	**86.58**

Comparison with fine-tuning. Table 1 shows the comparison between the proposed Treff adapter against the state-of-the-art following both fine-tuning (topline systems) and the zero-shot transferring (baseline systems) approach. We show the results of CLAP adapters trained with 16 examples here, as the improvement is getting marginal when the amount of support data increases (see more details in Figure 1). Our Treff adapter outperforms the zero-shot learning methods by a large margin and even achieves a comparable performance with the fine-tuning methods. In addition, the proposed Treff adapter outperforms the TIP-adapter, which was also adopted as a CLAP adapter, by 0.8% absolute in terms of accuracy. This is likely because our Treff adapter benefits from the preserved semantic information in the support embedding whereas the TIP-adapter discards it via gradient-based optimisation.

Comparison with few-shot learning. Table 2 shows the accuracy of the proposed Treff adapter against the metric-based few-shot learning and the TIP-adapter. Both our Treff adapter and the TIP-adapter achieve better results than other metric-based learning algorithms by a large margin. The reason may be that these two adapters utilise a large-scale pretrained model as their encoder. Our proposed Treff adapter outperforms the TIP-adapter by 0.71 percentage points in terms of accuracy, indicating that CALM learns task-specific knowledge while preserving the knowledge from CLAP. Table 3 compares the proposed Treff-adapter with other cross-modality few-shot methods on the ImageNet-ESC [12]. It can be observed that the Treff adapter and the TIP-adapter outperform the cross-modality few-shot learning by a large margin as they are able to make use of zero-shot knowledge transferring explicitly while the cross-modality FSL discards it gradually in the parameter optimisation.

Table 2: ***Accuracy (%) of various methods in the few-shot settings.*** *The best results are highlighted in* ***bold****.*

	ESC-50 (%)		FSD-FS (%)
	5-way	12-way	15-way
ProtoNet [2]	88.18	77.70	33.02
MatchNet [1]	86.83	71.81	-
HPN [13]	88.65	78.65	-
Treff adapter (w/o FT)	97.49	94.68	68.34
TIP-adapter [11]	97.52	95.58	69.45
Treff adapter	**98.53**	**96.29**	**70.59**

Table 3: ***Accuracy (%) of cross-modality few-shot learning.*** *The best results are highlighted in* ***bold****. Xmodal FSL denotes cross-modality FSL method.*

Dataset	Method	1-shot	2-shot	4-shot
ImageNet-ESC-19	Xmodal FSL	35.70	45.90	51.60
	TIP-adapter	**86.45**	87.89	87.11
	Treff adapter	85.26	**88.03**	**87.63**
ImageNet-ESC–27	Xmodal FSL	35.00	43.50	48.50
	TIP-adapter	**86.39**	**87.31**	87.03
	Treff adapter	86.02	87.04	**89.35**

5. Conclusion

We presented the Treff adapter, a training-efficient adapter for CLAP, to boost zero-shot classification performance by making use of a small set of labelled data. Specifically, we designed CALM to retrieve the probability distribution of text-audio clips over classes using a set of audio-label pairs and combined it with CLAP's zero-shot classification results. Furthermore, we designed a training-free version of the Treff adapter by using CALM as a cosine similarity measure. Experiments showed that the proposed Treff adapter is comparable and even better than fully-supervised methods and adaptation methods in low-shot and data-abundant scenarios. While the Treff adapter shows that combining large-scale pretraining and rapid learning of domain-specific knowledge is non-trivial for obtaining generic representations for few-shot learning, it is still limited to audio classification tasks. In the future, we will explore how to use audio-language models in diverse audio domains.

6. Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/T518086/1]. The research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT, http://doi.org/10.5281/zenodo.438045.

[1] https://github.com/JinhuaLiang/lam4fsl

7. References

[1] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

[2] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

[3] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending Clip to Image, Text and Audio," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 976–980, iSSN: 2379-190X.

[4] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," Jun. 2022, arXiv:2206.04769 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2206.04769

[5] Y. Wu*, K. Chen*, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[6] 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.

[7] J.-C. Gagnon-Audet, R. P. Monti, and D. J. Schwab, "AWE: Adaptive weight-space ensembling for few-shot fine-tuning," in *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. [Online]. Available: https://openreview.net/forum?id=rrMIPIboZL

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[9] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022. [Online]. Available: https://doi.org/10.1007/s11263-022-01653-1

[10] ——, "Conditional Prompt Learning for Vision-Language Models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification," in *Computer Vision – ECCV 2022*, ser. Lecture Notes in Computer Science, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 493–510.

[12] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models," Jan. 2023, arXiv:2301.06267 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2301.06267

[13] J. Liang, H. Phan, and E. Benetos, "Leveraging Label Hierachies for Few-Shot Everyday Sound Recognition," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, Nov. 2022.

[14] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," in *ICML Deep Learning Workshop*, vol. 2. Lille, 2015.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A General Architecture for Structured Inputs & Outputs," Mar. 2022. [Online]. Available: https://openreview.net/forum?id=fILj7WpI-g

[17] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a Visual Language Model for Few-Shot Learning," Nov. 2022, arXiv:2204.14198 [cs]. [Online]. Available: http://arxiv.org/abs/2204.14198

[18] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*. Brisbane Australia: ACM, Oct. 2015, pp. 1015–1018. [Online]. Available: https://dl.acm.org/doi/10.1145/2733373.2806390

[19] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose Tagging of Freesound Audio with Audioset Labels: Task Description, Dataset, and Baseline," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Nov. 2018, pp. 69–73.

[20] J. Liang, H. Phan, and E. Benetos, "Learning from Taxonomy: Multi-label Few-Shot Classification for Everyday Sound Recognition," Dec. 2022, arXiv:2212.08952 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2212.08952

[21] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[22] 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, iSSN: 2379-190X.

[23] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, and others, "CNN Architectures for Large-scale Audio Classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.