ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746118

HYBRID ATTENTION-BASED PROTOTYPICAL NETWORKS FOR FEW-SHOT SOUND CLASSIFICATION

You Wang, David V. Anderson

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

ABSTRACT

In recent years, prototypical networks have been widely used in many few-shot learning scenarios. However, as a metric-based learning method, their performance often degrades in the presence of bad or noisy embedded features, and outliers in support instances. In this paper, we introduce a hybrid attention module and combine it with prototypical networks for few-shot sound classification. This hybrid attention module consists of two blocks: a feature-level attention block, and an instance-level attention block. These two attention mechanism can highlight key embedded features and emphasize crucial support instances respectively. The performance of our model was evaluated using the ESC-50 dataset and the $noise$ESC-50 dataset. The model was trained in a 10-way 5-shot scenario and tested in four few-shot cases, namely 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot. The results demonstrate that by adding the hybrid attention module, our model outperforms the baseline prototypical networks in all four scenarios.

***Index Terms*—** Few-shot Learning, sound classification, hybrid attention, deep learning

1. INTRODUCTION

Acoustic environments can have profound impact on our daily life. Therefore, many applications for the perception and recognition of surrounding sound events have emerged, including audio surveillance [1], hearing aids [2], and smart cars [3], etc. Recently, deep learning methods have achieved great success and their performance is superior to traditional approaches. However, deep learning techniques are usually highly data dependent, and when the dataset is small or there is a lack of labelled training samples, perfomance can degrade significantly. To tackle this problem, few-shot learning was proposed and has been continuously gaining interest.

Among all the few-shot learning methods, embedding learning has become one of the most widely used in the field of acoustics. Embedding learning algorithms [4, 5] embed data samples to a feature space equipped with a distance metric such that within-class distances tend to be small and between-class distances tend to be large. Some typical metric-based embedding learning networks include convolutional Siamese neural networks [6], matching networks [7], and prototypical networks [8]. Matching networks [7] firstly introduced episodic training, and within each episode, training is performed in the same way as testing. Prototypical networks [8] also uses episodic training and tend to perform better than other related models in many different tasks. However, metric-based embedding learning often suffers from bad feature vectors and for prototypical networks, outliers in the support instances will also negatively influence the performance. Attention then comes into play and can be used to deal with these issues.

In audio classification tasks, attention is often used to emphasize certain temporal, channel, or spectral features. In [9], Yu *et al.* proposed a multi-level temporal attention model for weakly labelled audio classification. In [10], Zhang *et al.* introduce both channel and temporal attention into a CRNN model. In our previous work [11], we proposed a multi-channel temporal attention CNN model that fully exploits the relevant temporal information in different feature channels. Recent years have seen an increasing focus on the combination of audio attention and few-shot learning. Zhang *et al.* [12] proposed an attentional graph neural network for few-shot audio classification. Wang *et al.* [13] introduced a few-shot continual learning framework with an attention-based weight generator. And Chou *et al.* [14] proposed an attentional similarity module and plugged it into metric-based learning methods.

In this paper, we used prototypical networks as our baseline few-shot learning approach, and our contributions are threefold. Firstly, we present a slightly modified version of our previous work [11] and use it as a feature-level attention block to focus on essential feature regions. Secondly, we introduce an instance-level attention block to diminish the effect of those support instances that are far away from others, and combine it with the previous block as a hybrid-attention module. Finally, we incorporate the hybrid-attention module into prototypical networks. Experimental results on ESC-50 and *noise*ESC-50 datasets show improved performance.

2. METHODOLOGY

Figure 1 shows the overall architecture of our proposed model. It consists of three main parts: a backbone CNN

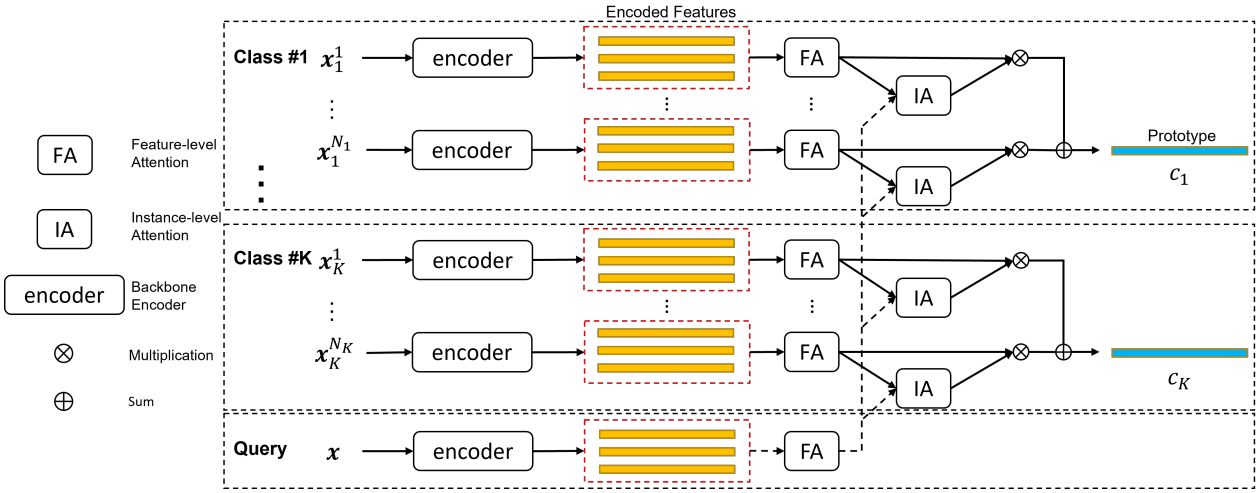

Fig. 1. The architecture of our proposed model.

network as a feature encoder, a feature-level attention block and an instance-level attention block. These three parts are combined together and integrated into the prototypical networks few-shot learning framework.

2.1. Prototypical Networks

Prototypical networks are one of the most popular metric-based few-shot learning methods. In the few-shot learning setting, assume we are given a training set $D = \{(\mathbf{x_i}, y_i)|i = 1, ..., N\}$ with N samples and C classes, where each $y_i \in \{1, ..., C\}$. In each training episode, a support set $S = \{S_k|k = 1, ..., K\}$ with K randomly selected classes is formed, where S_k contains N_k samples that belong to class k. This is called a K-way N_k-shot learning task. After that, a disjoint query set is created and the goal is to classify each query sample into one of the K support classes. The key idea of prototypical networks is the computation of prototypes to represent each class by averaging the encoded feature vectors of support samples in each class:

$$\mathbf{c}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} f_\theta(\mathbf{x}_i^k), \quad k \in \{1, ..., K\} \tag{1}$$

where $f_\theta(\cdot)$ stands for the encoder with learnable parameters θ, and $\mathbf{x}_i^k$ stands for a support sample in class k. For this paper, $f_\theta(\cdot)$ represents a backbone CNN network. Given a distance metric $d(\cdot)$, the posterior probability that a query sample $\mathbf{x}$ belongs to class k is then denoted as:

$$P_\theta(y = k|\mathbf{x}) = \frac{\exp(-d(f_\theta(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_\theta(\mathbf{x}), \mathbf{c}_{k'}))} \tag{2}$$

As in the implementation of [14], we use squared Euclidean distance as our distance metric.

2.2. Hybrid Attention

Our proposed hybrid attention module contains two blocks, a feature-level attention block to emphasize key features and an instance-level attention block to focus on crucial support instances and diminish the effect of outliers.

2.2.1. Feature-level Attention

Since each class in the support set only has a few samples and the prototypes are solely determined by the features extracted from these samples, the performance of prototypical networks is highly dependent on the discriminative power of the encoded feature vectors. There are certain parts of the features that contain more relevant information than others, and the feature-level attention block is introduced to focus on those parts. For audio classification, temporal attention has been widely used because of the unique temporal structure of audio signals [11]. In our previous work [11], we proposed a multi-channel temporal attention model, and the same idea is used here with a slight modification as our feature-level attention block. Figure 2(a) shows the block diagram of the feature-level block and Figure 2(b) shows the architecture of the multi-channel temporal attention. To create the attention vectors, the encoded features firstly go through a 3×3 convolutional layer with 1×1 padding, and then the *softmax* function is used on the time dimension to make the sum of each vector 1—this is changed from standard normalization as in our previous work [11]. This multi-channel structure creates a unique attention vector for each channel and is able to fully exploit the temporal information across channels.

2.2.2. Instance-level Attention

In the original prototypical networks, the prototypes are computed by averaging all support instance embeddings within

each class, which means it is assumed that the contribution of each support instance is equal. However, as shown in Figure 3(a), if one of the support instances is an outlier or is much farther away from the query sample than the other instances, it can cause a deviation on the resultant prototype and thereby causing incorrect classification. Therefore, inspired by [15], we propose to introduce an instance-level attention block to alleviate this problem. The core idea is to generate an attention score for each support instance based on its relationship with the query sample and replace Eq. 1 with a weighted average:

$$\mathbf{c}_k = \sum_{i=1}^{N_k} \beta_i^k f_\theta(\mathbf{x}_i^k) \tag{3}$$

where β_i^k is the score assigned to a support instance in class k, and it is modelled as follows:

$$\beta_i^k = \frac{e_i^k}{\sum_{n=1}^{N_k} e_n^k} \tag{4}$$

$$e_i^k = sum\{\sigma(f_\phi(f_\theta(\mathbf{x}_i^k)) \circ f_\phi(f_\theta(\mathbf{x})))\} \tag{5}$$

where $f_\phi(\cdot)$ is a fully connected layer, $\sigma(\cdot)$ is $sigmoid$ function, and $sum\{\cdot\}$ is the summation over all elements of a vector. The purpose of Eq. 5 is to compute the similarity between a support sample and the query sample, therefore it can be seen from Eq. 4 and Eq. 5 that the weight of a support sample is query-dependent, meaning that the contribution of each support sample depends on the incoming query sample instead of having an independent "quality." Moreover, the reason why we choose standard normalization over $softmax$ in this case is that $softmax$ will dramatically enlarge the difference between the weights of the support samples and may result in overly biased prototypes. As an illustration shown in Figure 3(b), by assigning lower scores to bad instances, the model can become more capable of assigning the correct label to the query sample.

3. EXPERIMENTAL SETUP

3.1. Datasets

Our experiments were conducted using two datasets: ESC-50 and *noise*ESC-50. The ESC-50 dataset [16] has 2000 5-second-long audio recordings with a sample rate of 44.1kHz, organized into 50 balanced classes. The *noise*ESC-50 dataset was created in [14] by mixing the clean ESC-50 samples with random acoustic scenes from DCASE2016 dataset [17] as additive background noise. The relatively small size of ESC-50 makes it a good candidate for few-shot sound classification.

3.2. Data Preparation

By using the same splitting strategy as in [14], we randomly selected 35 classes for training, 5 classes for 5-way 5-shot validation, and the remaining 10 classes for testing. Also according to [14], for both ESC-50 and *noise*ESC-50, the audio clips were firstly downsampled from 44.1kHz to 16kHz, and then log mel-spectrograms with 128 mel bins were extracted. The window size is 2048 and the hopsize is 497, resulting in a 128×160 feature map for each data sample. The input features were then z-score normalized using the mean and standard deviation of the training set before being fed into the model.

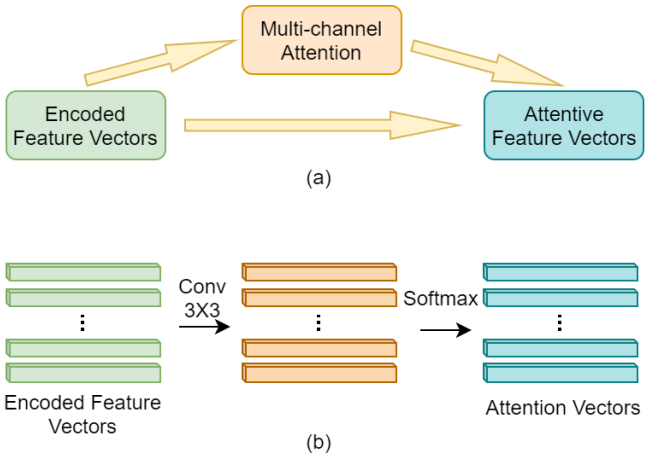

Fig. 2. (a) Block diagram the feature-level attention block. (b) Structure of the multi-channel attention.

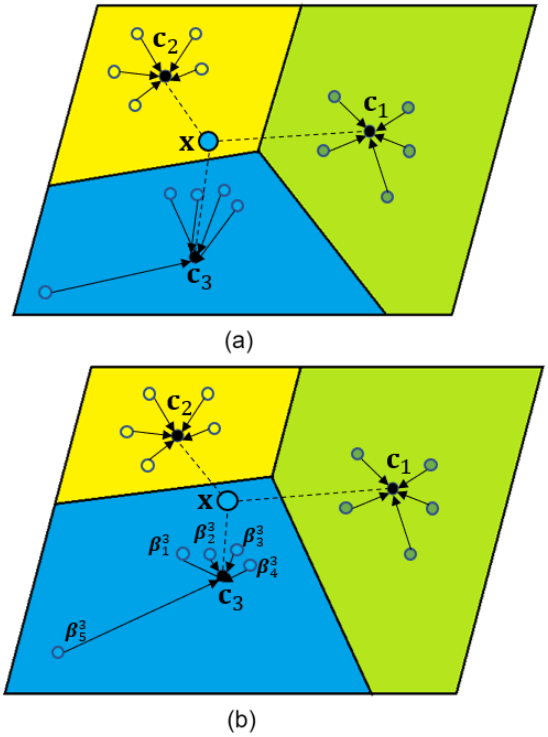

Fig. 3. (a) Prototypical networks without instance-level attention. (b) Prototypical networks with instance-level attention.

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Prototypical Networks	64.40±1.38%	83.83±0.92%	47.83±1.10%	71.00±1.04%
Proto-FA (Ours)	**71.18±1.23%**	89.60±1.08%	**57.08±1.43%**	78.48±1.51%
Proto-HA (Ours)	–	**90.35±0.83%**	–	**80.08±1.31%**

Table 1. Few-shot sound classification accuracies for prototypical networks with or without attention module on ESC-50.

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Prototypical Networks	61.53±0.40%	81.03±0.57%	45.90±0.28%	64.98±0.52%
Proto-FA (Ours)	**71.35±0.90%**	88.00±0.63%	**56.55±1.22%**	78.55±0.75%
Proto-HA (Ours)	–	**88.78±0.45%**	–	**79.08±1.12%**

Table 2. Few-shot sound classification accuracies for prototypical networks with or without attention module on *noise*ESC-50.

3.3. Backbone Network and Training Settings

The structure of our backbone network was inspired by the work of Kumar *et al.* [18]. It contains 3 basic blocks and each block consists of a 3×3 convolutional layer, batch normalization, ReLU activation, and a max pooling layer consecutively. The kernel sizes for these 3 max pooling layers are 8×2, 8×2, and 2×1, respectively. The numbers of channels for these 3 convolutional layers are 128, 256, and 384. We trained the network using Adam with cross-entropy loss and a starting learning rate as 0.001. The total number of epoch is 60, and the learning rate is decayed by a factor of 10 after every 20 epochs. We also used a 1e-4 weight decay as in [14]. In addition, according to [8], it is beneficial to use more ways during training than during testing, therefore we trained our model in a 10-way 5-shot scenario.

4. RESULTS AND DISCUSSION

The few-shot audio classification accuracies for the test sets of ESC-50 and *noise*ESC-50 are shown in Table 1 and Table 2 respectively. The best model was selected based on the 5-way 5-shot accuracies of the 5 validation classes. For each dataset, our model was run 10 times and the average accuracy is reported along with the 95% confidence intervals. Proto-FA stands for prototypical networks with the feature-level attention module, and Proto-HA stands for prototypical networks with the hybrid attention module including both the feature-level and the instance-level blocks. Since the instance-level attention assigns weights to multiple support samples, we only focus on the 5-shot cases when comparing the performances involving hybrid attention.

Table 1 compares the performance of prototypical networks with or without attention on ESC-50. It can be seen that by adding the feature-level attention block, there is a significant improvement for all four scenarios. Especially for 10-way 1-shot and 10-way 5-shot cases, where the performance gain is 9.25% and 7.48% respectively, which indicates that our multi-channel temporal attention block is capable of highlighting key features to make data samples more distinguishable. Moreover, the hybrid attentional prototypical networks perform best in 5-shot scenarios, showing that the instance-level attention block is able to focus on crucial support samples when computing the prototypes. Note, that in the 1-shot case, the instance-level attention block performs no function since there is only a single support sample and its weight will always be 1. Therefore 1-shot results for Proto-HA are not reported.

The results for *noise*ESC-50 are shown in Table 2. For *noise*ESC-50, the Proto-FA model enhances the performance and the improvement is even greater than the performance gain of Proto-FA with ESC-50. This suggests that emphasizing relevant features can largely diminish the effect of noise. Similarly, when using the Proto-HA model, the performance is further enhanced for the 5-shot cases, showing that in the noisy setting, the instance-level attention module is still able to lessen the contribution of those bad support instances. It should be noted that the improvement provided by the Proto-HA model for the 10-way 5-shot case on *noise*ESC-50 is less than the the improvement seen on ESC-50. The reason for this might be that when there are good and bad support instances, the instance-level attention can assign higher weights to those good ones, but when all the support and query samples are degraded, the advantage might not be as big.

5. CONCLUSIONS

In this paper, we proposed a hybrid attention-based prototypical networks for few-shot sound classification. The hybrid attention module contains a feature-level attention block to highlight important features and an instance-level attention block to focus on crucial support instances. The experimental results on ESC-50 and *noise*ESC-50 show performance improvement over the baseline prototypical networks and the instance-level attention block helps to elevate the performance for 5-shot cases. Our future work includes further fine-tuning our model and testing our model on different datasets.

6. REFERENCES

[1] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2015.

[2] Enrique Alexandre, Lucas Cuadra, Manuel Rosa, and Francisco Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2249–2256, 2007.

[3] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[4] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675–678.

[5] Michael D Spivak, *A comprehensive introduction to differential geometry*, Publish or perish, 1970.

[6] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*. Lille, 2015, vol. 2.

[7] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[8] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[9] Changsong Yu, Karim Said Barsim, Qiuqiang Kong, and Bin Yang, "Multi-level attention model for weakly supervised audio classification," *arXiv preprint arXiv:1803.02353*, 2018.

[10] Zhichao Zhang, Shugong Xu, Shunqing Zhang, Tianhao Qiao, and Shan Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130327–130339, 2019.

[11] You Wang, Chuyao Feng, and David V Anderson, "A multi-channel temporal attention convolutional neural network model for environmental sound classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 930–934.

[12] Shilei Zhang, Yong Qin, Kewei Sun, and Yonghua Lin, "Few-shot audio classification with attentional graph neural networks.," in *Interspeech*, 2019, pp. 3649–3653.

[13] Yu Wang, Nicholas J Bryan, Mark Cartwright, Juan Pablo Bello, and Justin Salamon, "Few-shot continual learning for audio classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 321–325.

[14] Szu-Yu Chou, Kai-Hsiang Cheng, Jyh-Shing Roger Jang, and Yi-Hsuan Yang, "Learning to match transient sound events using attentional similarity for few-shot sound recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 26–30.

[15] Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun, "Hybrid attention-based prototypical networks for noisy few-shot relation classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6407–6414.

[16] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[17] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[18] Anurag Kumar, Maksim Khadkevich, and Christian Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 326–330.