

FEW-SHOT ACOUSTIC EVENT DETECTION VIA META LEARNING

Bowen Shi^{1*}, Ming Sun², Krishna C. Puvvada², Chieh-Chi Kao², Spyros Matsoukas², Chao Wang²

¹Toyota Technological Institute at Chicago

²Amazon

bshi@ttic.edu, {mingsun,puvvav,chiehchi,matsouka,wngcha}@amazon.com

ABSTRACT

We study few-shot acoustic event detection (AED) in this paper. Few-shot learning enables detection of new events with very limited labeled data. Compared to other research areas like computer vision, few-shot learning for audio recognition has been under-studied. We formulate few-shot AED problem and explore different ways of utilizing traditional supervised methods for this setting as well as a variety of meta-learning approaches, which are conventionally used to solve few-shot classification problem. Compared to supervised baselines, meta-learning models achieve superior performance, thus showing its effectiveness on generalization to new audio events. Our analysis including impact of initialization and domain discrepancy further validate the advantage of meta-learning approaches in few-shot AED.

Index Terms— Acoustic event detection, few-shot learning, meta learning

1. INTRODUCTION

Acoustic event detection (AED) is the task of detecting whether certain events occur in an audio clip. It can be applied in many areas such as surveillance [1, 2], and recommendation systems [3]. Current state-of-the-art AED models are data-hungry and large number of labeled data is needed to achieve high performance on detecting target event [4], which causes problem for detecting events where labels are limited. One potential solution is semi-supervised learning which incorporates unlabeled data for model training. Multiple methods belonging to this category [5, 6, 7] have been explored for AED. However those approaches fail to generalize to few-shot scenarios where only very few samples (e.g., < 10) are available for target event. Few-shot learning for AED is useful in practice not only because there exists large variety of audio events and labels of rare events are extremely limited. Besides, some audio events like doorbell sounds can be different for different households and there is no universal definition for such event. Efficient few-shot learning facilitates the personalization of AED in real life.

Few-shot learning tasks have been increasingly studied in literature and often rely on meta-learning approaches including MAML (Model-Agnostic Meta-Learning)[8] and Prototypical networks [9]. Most such works are done in computer vision [10, 11] or natural language recognition[12] while very little work has been done in audio-related tasks. [13] is the only existing work on sound recognition to our knowledge. However it studies multi-way classification while we focus on multi-label classification (detection) here. To our knowledge few-shot detection problem has not been thoroughly studied before, though it bears similarity with multi-way few-shot

classification. Formulation of this problem will be discussed in detail in section 2. Besides we will also study several typical meta-learning approaches under few-shot AED setting and compare them with traditional supervised models in different scenarios (e.g., domain shift).

2. METHODS

2.1. Basic AED model

For AED, given feature of audio signal \mathbf{x} (e.g. Log-mel Filterbank Energy) the task is to train a model f to predict a multi-hot vector $\mathbf{y} \in \{0, 1\}^C$ with C being the size of event set and y_c being a binary indicator whether event c is present in \mathbf{x} . It is a multi-label classification problem and prediction $f(\mathbf{x})$ is not a distribution over event set since multiple events can occur in \mathbf{x} . In supervised setting, we train model f using cross-entropy loss (see equation 1), where w_c is the penalty on mis-classifying true positive samples of class c . w_c serves the purpose of balancing losses between positive and negative instances and is tuned as hyper-parameter in practice.

$$L(f, D) = - \sum_{(\mathbf{x}, \mathbf{y}) \in D} \sum_{c=1}^C \{w_c y_c \log f^c(\mathbf{x}) + (1 - y_c) \log(1 - f^c(\mathbf{x}))\} \quad (1)$$

2.2. Few-shot AED setup

We aim to obtain a model that can be trained to perform good detection given few labeled examples for new event. Suppose we are given two non-overlapping sets of classes (events) \mathcal{C}_{train} and \mathcal{C}_{test} . The task is to train a detection model f using labeled data of classes from \mathcal{C}_{train} and to evaluate on data of classes in \mathcal{C}_{test} .

Few-shot learning problem is commonly tackled in a meta-learning setting, which is also called episodic setting in literature [14]. It operates on meta datasets where each element is a dataset and thus instantiates a task, which is multi-label classification here. For two tasks, events (classes) to detect can differ. Models are trained with *meta-training set* and will be tested on *meta-test set*. There is *no class overlap* in the two sets. Similar to supervised learning, there often exists a *meta validation set* which is used for hyper-parameter tuning and model selection. This special setup allows model to generalize across classes since it is trained with a set of tasks while each consists of different classes.

Now we formally define the setup. A dataset \mathcal{D} is a set of $\langle \text{audio}, \text{events} \rangle$ pairs. A meta-set \mathcal{M} is a set of datasets. Each dataset \mathcal{D} instantiates a K -way binary classification problem where K is number of events. \mathcal{D} is divided into two parts, “training” subset \mathcal{S} and “test” subset \mathcal{Q} . To avoid confusion with meta-training and

*Work done at Amazon

Algorithm 1 Construction of training, validation and test sets for K -way N -shot detection. N_S^-, N_Q^- are number of negative samples for support and query set respectively. All sampling is without replacement. $\text{SAMPLE}(X, N, C)$ denotes a set of N elements of event set C chosen uniformly at random from set X

Input: Dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. Whole set of events \mathcal{C} (label set).

Output: Training/validation/test set $\mathcal{M}_{train}/\mathcal{M}_{val}/\mathcal{M}_{test}$
 Split \mathcal{C} into three disjoint sets $\mathcal{C}_{train}, \mathcal{C}_{val}, \mathcal{C}_{test}$
 Split \mathcal{D} into $\mathcal{D}_{train}, \mathcal{D}_{val}, \mathcal{D}_{test}$ according to $\mathcal{C}_{train}, \mathcal{C}_{val}, \mathcal{C}_{test}$
for x in $\{train, val, test\}$ **do**
 for t in $\{1, \dots, |\mathcal{M}_x|\}$ **do**
 Randomly sample K events $\mathcal{C}_t = \{c_1, c_2, \dots, c_K\}$ from \mathcal{C}_x
 for k in $\{1, 2, \dots, K\}$ **do**
 $\mathcal{S}_t \leftarrow \mathcal{S}_t \cup \text{SAMPLE}(\mathcal{D}_x, N, \{c_k\})$
 $\mathcal{Q}_t \leftarrow \mathcal{Q}_t \cup \text{SAMPLE}(\mathcal{D}_x, Q, \{c_k\})$
 end for
 $\mathcal{S}_t \leftarrow \mathcal{S}_t \cup \text{SAMPLE}(\mathcal{D}_x, N_S^-, \mathcal{C}_x \setminus \mathcal{C}_t)$
 $\mathcal{Q}_t \leftarrow \mathcal{Q}_t \cup \text{SAMPLE}(\mathcal{D}_x, N_Q^-, \mathcal{C}_x \setminus \mathcal{C}_t)$
 $\mathcal{M}_x \leftarrow \mathcal{T} = (\mathcal{S}_t, \mathcal{Q}_t)$
end for
end for

meta-test set mentioned above, \mathcal{S} and \mathcal{Q} are herein called support and query set respectively. Under this setting, model f is trained with meta-training set \mathcal{M}_{train} and evaluated with meta-test set \mathcal{M}_{test} . More concretely, $\forall \mathcal{D}_i \in \mathcal{M}_{test}$, model f is required to make prediction on query set \mathcal{Q}_i (no label) given its support set \mathcal{S}_i (with label). We also define *shots* N as the number of positive samples in support set \mathcal{S} . A typical value of N would be $1 \sim 5$ in few-shot setting. Note we only constrain number of positive samples per event, which is different from multi-way few-shot classification problems where N refers to number of labeled data for each class. This is for practical considerations since negative samples for one particular event can be much easier to acquire. From a typical labeled dataset (e.g., AudioSet), meta datasets can be constructed through sampling. Details on the construction process is shown in algorithm 1.

2.3. Approaches to few-shot AED

Few-shot learning does not prescribe a specific training procedure. We will discuss in the following how to tackle this problem under current setup from two aspects: (1). with traditional supervised methods, (2). with meta-learning approaches.

2.3.1. Supervised Baselines

Before diving into meta-learning, it is important to explore solutions based purely on supervised learning. A natural approach to exploit \mathcal{M}_{train} would be to train a model f over the whole training set \mathcal{M}_{train} by aggregating all training classes in \mathcal{C}_{train} . At test time, f will be fine-tuned with the support set of new task \mathcal{T} . Commonly, f is prone to over-fitting at fine-tuning stage since \mathcal{S} only has few samples per event. A natural extension will be to fine tune only part of model. Formally if pre-trained model $f = [f_\theta, f_\phi]$, where f_θ is the feature extraction part and f_ϕ is classifier (commonly last linear layer), we would freeze f_θ and fine-tune only f_ϕ at test time. We refer the above two baselines as **FT-All** and **FT-Linear** respectively.

In addition to above two fine-tuning baselines, we also consider following supervised approach, which classifies query sample based on its distance to support samples in feature space. Formally, given

a new task $\mathcal{T} = (\mathcal{S}, \mathcal{Q}) \in \mathcal{M}_{test}$, $\forall \mathbf{x}^{(q)} \in \mathcal{Q}$, we define following distance as in equation 2, which characterize how far $\mathbf{x}^{(q)}$ is from being positive and negative.

$$d_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S}) = \frac{\sum_{(\mathbf{x}^{(s)}, \mathbf{y}) \in \mathcal{S}} \text{dist}(f_\theta(\mathbf{x}^{(s)}), f_\theta(\mathbf{x}^{(q)}))}{\sum_{(\mathbf{x}^{(s)}, \mathbf{y}) \in \mathcal{S}} 1_{y_k=t}} \quad (2)$$

$t \in \{0, 1\}, k \in \{1, \dots, K\}$

$d_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S})$ is average distance of query sample $\mathbf{x}^{(q)}$ to positive or negative samples of event k in support set in the feature space. $\text{dist}(\cdot, \cdot)$ in equation 2 is a distance metric (e.g., L2, cosine), which can be tuned as hyper-parameter. For ease of evaluation, we further convert $d_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S})$ to probability $p_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S})$ using softmax function:

$$p(y_k = t | \mathbf{x}^{(q)}; \theta, \mathcal{S}) = \frac{\exp(-d_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S}))}{\exp(-d_k^0(\mathbf{x}^{(q)}; \theta, \mathcal{S})) + \exp(-d_k^1(\mathbf{x}^{(q)}; \theta, \mathcal{S}))} \quad (3)$$

Fine-tuning, which might potentially leads to over-fitting due to the small size of support set, is avoided in this method. Another motivation for this method is that embedding space learned from \mathcal{M}_{train} also provides meaningful representations for samples of unseen classes in \mathcal{M}_{test} . We call this baseline **NN**, which is short for nearest neighbor, as it classifies new data based on its neighboring data points.

2.3.2. Meta-learning Approaches

Meta-learning based models are trained end-to-end for the purpose of learning to build classifiers from few examples. At training time, we compute probability $p(\mathbf{y}^q | \mathbf{x}^{(q)}, f, \mathcal{S})$ for $\mathbf{x}^{(q)} \in \mathcal{Q}$. Since labels of query data ($\mathbf{y}^{(q)}$) are known at training time, we can define loss function based on the ground-truth and prediction, thus being able to train model f end-to-end via gradient descent. Models vary by the manner in which the conditioning on support set is realized. Main approaches we experimented are described below. We will use same notations as above.

Prototypical Networks This approach is a simple integration of NN-baseline into the end-to-end meta-learning framework. Given task $\mathcal{T} = (\mathcal{S}, \mathcal{Q})$ we compute probability of query samples $\mathbf{x} \in \mathcal{Q}$ being positive or negative on event k , $p(y_k = t | \mathbf{x}; \theta, \mathcal{S})$, as in equation 2 and 3. At training time we use it to compute cross-entropy loss:

$$L(\theta; \mathcal{T}) = - \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{Q}} \sum_{k=1}^K w_k \log p(y_k | \mathbf{x}; \theta, \mathcal{S}) \quad (4)$$

Model f_θ is updated via gradient descent. At test time, we follow equation 2 and 3 to compute $p(y_k = t | \mathbf{x}; \theta, \mathcal{S})$ for any new data \mathbf{x} .

Above approach is an adaptation of prototypical network [9] from multiway classification to multi-label classification setting. Main difference between ours and [9] lies in the way of computing distance $d_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S})$. In [9], the embeddings of data of class k from support set are averaged as the prototype for that class and distance $d_k^t(\mathbf{x}^{(q)}; \theta, \mathcal{S})$ is between query sample and prototype. While here we compute distance of query sample $\mathbf{x}^{(q)}$ to every support sample and then take the mean distance. In short, we average distance instead of averaging embedding. Such a modification is because negative samples of any event can come from

much larger number of other classes. Mean embedding of negative samples might not be a good representation for negative class.

MetaOptNet MetaOptNet is short for meta-learning via differentiable convex optimization, which is originally proposed in [8]. Similar to multi-prototypical network, MetaOptNet learns a classifier in feature space. However, a linear SVM classifier is learned instead of nearest neighbor:

$$\begin{aligned} \min_{\mathbf{w}} \min_{\boldsymbol{\xi}} \frac{1}{2} \sum_{k=1}^K \sum_{t \in \{0,1\}} \|\mathbf{w}_k^t\|^2 + \lambda \sum_{k=1}^K \xi^k \\ \text{s.t. } \mathbf{w}_k^{y_k} f_{\theta}(\mathbf{x}) - \mathbf{w}_k^t f_{\theta}(\mathbf{x}) \geq 1 - \delta_{y_k,t} - \xi^k \\ \forall k \in \{1, \dots, K\}, t \in \{0, 1\}, (\mathbf{x}, \mathbf{y}) \in \mathcal{S} \end{aligned} \quad (5)$$

We solve equation 5 with a differentiable QP solver [15] so that f_{θ} can be learned in an end-to-end way. At test time, we solve equation 5 to score new data $\mathbf{x}^{(q)}$ and transform it to probability through equation 6.

$$p(y_k = t | \mathbf{x}^{(q)}; \theta, \mathcal{S}) = \frac{\exp(\mathbf{w}_k^t f_{\theta}(\mathbf{x}^{(q)}))}{\exp(\mathbf{w}_k^1 f_{\theta}(\mathbf{x}^{(q)})) + \exp(\mathbf{w}_k^0 f_{\theta}(\mathbf{x}^{(q)}))} \quad (6)$$

MetaOptNet is explored here because SVM naturally fits binary classification setting. Additionally, it has achieved state-of-the-art results in few-shot image classification benchmarks [10, 16, 17, 18].

MAML Model Agnostic Meta-Learning (MAML) [19] is another popular meta-learning framework. Different from Prototypical network and MetaOptNet where classifier is trained on an embedding space, MAML learns initialization parameters θ_0 and ϕ_0 from meta training set \mathcal{M}_{train} such that the model can perform well on query set after a few steps of gradient descent. Support set \mathcal{S} are used to calculate loss used for gradient computation. Suppose model f is initialized as f_{θ_0, ϕ_0} , let $\theta_N, \phi_N = \text{GD}(\theta_0, \phi_0; L, \mathcal{S}, N)$ be the model parameters updated through N steps of gradient descent where the loss function is L (same as equation 1) computed on support set \mathcal{S} . We solve optimization problem defined as equation 7, which minimizes the cross-entropy loss of f_{θ_N, ϕ_N} on query set \mathcal{Q} .

$$\min_{\theta_0, \phi_0} L(\theta_N, \phi_N; \mathcal{Q}) = \min_{\theta_0, \phi_0} L(\text{GD}(\theta_0, \phi_0; L, \mathcal{S}, N); \mathcal{Q}) \quad (7)$$

Note only initial model parameters θ_0, ϕ_0 are updated throughout training process while $\theta_{1:N}, \phi_{1:N}$ are just intermediate variables. Training in MAML is made possible by unrolling gradient descent steps. Given data $\mathbf{x}^{(q)} \in \mathcal{Q}$ at test time, we can compute the overall probability vector as $p(\mathbf{y} | \mathbf{x}^{(q)}; \mathcal{S}) = f_{\theta_N, \phi_N}(\mathbf{x}^{(q)})$ where $\theta_N, \phi_N = \text{GD}(\theta_0, \phi_0; L, \mathcal{S}, N)$.

3. EXPERIMENTAL SETUP

We use Audioset [20] for experiments. In total, there are 5.8k hours of audios from 527 sound classes in Audioset. The large number of sound events provides good test-bed for few-shot learning. The whole set of event classes of Audioset forms tree-like structure. We only select ‘‘leaf’’ nodes to ensure fine granularity of audio events. Besides, audios with annotation accuracy¹ less than 80% are not selected. In total we obtained 19,841 10-second audio clips from 142 events. Following our setup, they are randomly split into meta training set with 99 events, meta validation set with 21 events and meta

test set with 21 events. Table 1 shows event examples in each partition.

	Train	Val	Test
Events	'Sneeze', 'Spray', 'Electric guitar', 'Choir', 'Cheering'	'Glockenspiel', 'Toilet flush', 'Sink', 'Vacumm cleaner'	'Ukulele', 'Clapping', 'Toot', 'Purr', 'Racing'

Table 1. Examples of events in train, validation and test set

For audio pre-processing, we compute log Mel-filterbank energy feature for each audio clip. It is calculated with window size of 25 ms and hop size of 10 ms. The number of mel coefficients is 64, which gives us log-mel spectrogram feature of size 1000×64 for each audio clip. Global CMVN (Cepstral Mean and Variance Normalization) is applied on feature map before it is fed as input to the model.

We use convolutional neural network (CNN) with 4 blocks as our backbone model architecture. Each block is composed of 3×3 convolution, ReLU, Batch Norm and 3×3 max-pooling layers. We tuned CNN architecture on meta validation set and did not find improvements from using deeper networks like ResNet [21].

For few-shot setup, we experimented on 1-shot and 5-shot setting, as is commonly done in few-shot benchmarks [10, 16, 17]. The number of ways is set to 5. The query set is composed of 15 positive samples per way. On the other hand, we sampled 10 and 50 negative data for support set respectively in 1-shot and 5-shot setting. The number of negative data is 150 for query set in both settings. Here the negative data are audios where none of 5 target events occur. We randomly sample 5000 training tasks, 200 validation tasks and 200 test tasks. For evaluation, we measure Area Under Curve (AUC) for Receiver Operating Characteristic (ROC) curve. AUC score averaged over tasks in test set is reported. Higher AUC is better.

For supervised baselines, we pre-train CNN model (same architecture as above) with all training data by doing 99-way detection. A small subset (10%) is held out for purpose of hyper-parameter tuning. After pre-training, the last linear layer is replaced by a randomly initialized linear layer with output dimension being 5, which is equal to number of ways. Either the whole model (**FT-whole**) or the linear layer (**FT-Linear**) is fine-tuned for 20 epochs on the test task. For NN baseline, the distance metric is chosen among {L2, cosine, dot} and we find cosine distance works best. For **MAML**, the model runs through 5 steps of gradient descent both at training and testing time. For prototypical network and MetaOptNet, number of detection ways is tuned to be 5 during training.

4. RESULTS

4.1. Main Results

Figure 1 shows the performance of 6 methods on test set. Three meta-learning approaches outperform all supervised baselines on both 1-shot and 5-shot setting. Meta-learning approaches are encouraged to generalize to unseen events as new task is sampled for training in each step. Compared to supervised baselines, it is more effective on capturing the relationship between different events.

On the other hand, performance order of three supervised approaches are: FT-All < FT-Linear < NN, which implies less parameter tuning leads to better performance. In few-shot setting where labeled data are extremely scarce for one task, parameter fine-tuning often induces overfitting. As data amount increases we notice the gap between heavily fine-tuned approach and moderately tuned approach is reduced (FT-All vs. FT-Linear on 5-shot). Similar trend is shown on the side of meta-learning (MAML vs. Prototypical network/MetaOptNet).

¹According to <https://research.google.com/audioset/dataset/index.html>

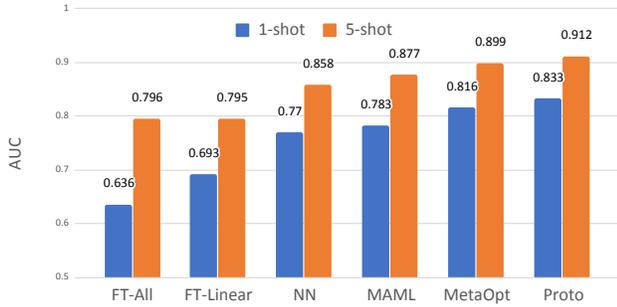


Fig. 1. Mean AUC on test set of 3 supervised baselines and 3 meta-learning approaches. Higher is better

We also notice NN achieves comparable performance to meta-learning approaches despite its simplicity. This implies embedding function f_θ learned from detecting known events also equips audios with meaningful features on distinguishing unseen events because event classes are correlated with each other. For instance, “orchestra” and “electric guitar” both belong to music sound. A model trained on detecting “orchestra” sound captures characteristics of music, which can be used in detecting “electric guitar”. However, such model tends to overfit to over fine-grained features of known events. Model capacity on detecting distant events (e.g., those from non-music branch) will diminish. This issue can be mitigated by prototypical network, which also does nearest neighbor classification but learns embedding function f_θ through sampling from large pool of tasks, which helps generalization to new events.

4.2. Analysis

Impact of domain mismatch In default setup, we randomly split the events, under which the distribution of audio events in training and test set are the same. Here we created a setting where testing events come from a different distribution. For instance, all testing events belong to animal sounds while training events are household sounds. Such a setup is for testing robustness of meta-learning methods. More concretely, based on original setup we select one target domain (e.g., music), which is child of root node in Audioset ontology. All leaf nodes under target domain will be treated as new test set. Those events and their associated audios are removed from original training set. We re-run experiments for all 6 approaches under this setup. For purpose of comparison, we also evaluate same models with test set from which target events are removed. We experimented “music” and “animal” domains. Results are shown in table 2.

AUC	1-shot			5-shot		
	In-domain	Music	Animal	In-domain	Music	Animal
FT-All	0.645	0.653	0.568	0.739	0.743	0.639
FT-Linear	0.72	0.716	0.612	0.798	0.785	0.670
NN	0.727	0.709	0.613	0.824	0.798	0.695
MAML	0.715	0.651	0.599	0.826	0.780	0.666
MetaOpt	0.773	0.699	0.640	0.856	0.81	0.740
Proto	0.796	0.712	0.644	0.875	0.824	0.749

Table 2. Impact of domain discrepancy on different models. **Note:** “In-domain” numbers are different from those of figure 1 as both training and evaluation sets differ.

Through the comparison of in-domain and target domain evaluation from table 2, all methods deteriorate due to domain mismatch.

Results on “music” are slightly better than those on “animal”, which is mainly because in Audioset music sounds are common and some patterns exist in other sounds as well such as bell ringing. On the other hand animal sounds such as “dog growling” rarely resemble other sounds. The gain of meta-learning approaches over supervised baselines are diminished due to domain mismatch. This implies the potential overfitting issue in meta-learning. Meta-learning models learn to utilize the correlation between classes. However, if all training classes come from one domain, the model tends to overfit to that particular domain and performance on new domain would drop. We notice that FT-linear achieves better results than MAML, which suggests learned features can still benefit detection in different domain. This can also be seen from overall high performance of feature-based approaches including NN, MetaOptNet and prototypical network.

Does pre-training help? Compared to random guess, supervised baselines lead to much better results. Such a fact shows benefit of pre-training since all three baselines are built on the pre-trained detector. Meta-learning models are trained on individual tasks where labeled data in each task are of small amount. We want to see whether pre-training with whole training set would also provide good starting point for meta-learning. We follow same pre-training steps as supervised baselines, which is pre-training a 99-way detector with whole training set. Model parameters except the last layer are used as initial parameters for each method. Results of pre-training vs. non-pretraining for three meta-learning approaches are shown in table 3.

AUC	MAML (+P)	MetaOpt (+P)	Proto (+P)
1-shot	0.783 (0.788)	0.816 (0.823)	0.833 (0.827)
5-shot	0.877 (0.841)	0.899 (0.899)	0.912 (0.912)

Table 3. Supervised pre-trained initialization vs. randomly initialization for meta-learning models (+P: with pre-training)

We do not find consistent gain from pre-training according to table 3. By learning with sub-tasks in training set, meta-learning models implicitly learns detector conditioned on support set despite of limited number of labeled data in each task. Initialization has little effect on the final model. Another possible reason is that our backbone CNN is of small size thus there is little overfitting in learning new task.

5. CONCLUSION

We formulated and studied few-shot acoustic event detection by comparing typical meta-learning and supervised approaches. Through experimentation, we find meta-learning approaches outperforms their supervised counterparts, which shows the effectiveness of its training setup on generalization to new audio events. Besides, we find methods based on feature learning (e.g., NN baseline, prototypical network) outperform others. Thus one line of our future work is to incorporate unlabeled audio data for representation learning in few-shot AED. Domain discrepancy has negative impact on performance of all methods we experimented. However, we can still achieve gains from some meta-learning methods. How to train an AED model more robust to domain mismatch under few-shot setting remains to be investigated.

6. REFERENCES

- [1] M. Cristani, M. Mecego, and V. Murino, “Audio-visual event recognition in surveillance video sequences,” *IEEE Transactions on Multimedia*, vol. 9, pp. 257–267, 2 2007.

- [2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS*, 2007, pp. 21–26.
- [3] P. Cano, M. Koppenberger, and N. Wack, "Content-based music audio recommendation," in *ACM International Conference on Multimedia*, 2005, pp. 211–213.
- [4] C. Kao, W. Wang, M. Sun, and C. Wang, "R-crn: Region-based convolutional recurrent neural network for audio event detection," in *Interspeech*, 2018.
- [5] B. Shi, M. Sun, C. Kao, V. Rozgic, S. Matsoukas, and C. Wang, "Semi-supervised acoustic event detection based on tri-training," in *ICASSP*, 2019.
- [6] J. Lu, "Mean teacher convolution system for dcase 2018 task 4," Tech. Rep., DCASE2018 Challenge, 2018.
- [7] L. Lin and X. Wang, "Guided learning convolution system for dcase 2019 task 4," Tech. Rep., DCASE2019 Challenge, 2019.
- [8] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.
- [9] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017.
- [10] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017.
- [11] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015.
- [12] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauero, H. Wang, and B. Zhou, "Diverse few-shot text classification with multiple metrics," in *NAACL*, 2018.
- [13] S. Chou, K. Cheng, J. Jang, and Y. Yang, "Learning to match transient sound events using attentional similarity for few-shot sound recognition," in *ICASSP*, 2019.
- [14] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *NeurIPS*, 2016.
- [15] B. Amos and J. Kolter, "Optnet: Differentiable optimization as a layer in neural networks," in *ICML*, 2017.
- [16] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. Tenenbaum, H. Larochelle, and R. Zemel, "Meta-learning for semi-supervised few-shot classification," in *ICLR*, 2018.
- [17] S. Ravi, H. Bertinetto, J. Henriques, P. Torr, and A. Vedaldi, "Meta-learning with differentiable closedform solvers," in *ICLR*, 2019.
- [18] B. Oreshkin, P. Rodrigues, and A. Lacoste, "Tadam: Task dependent adaptive metric for improved fewshot learning," in *NeurIPS*, 2018.
- [19] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *CVPR*, 2019.
- [20] 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 *ICASSP*, 2017.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.