LEARNING FROM TAXONOMY: MULTI-LABEL FEW-SHOT CLASSIFICATION FOR EVERYDAY SOUND RECOGNITION

Jinhua Liang[⋆] *Huy Phan*[‡◇] *Emmanouil Benetos*[⋆†]

[⋆] Centre for Digital Music, Queen Mary University of London, UK
[†] The Alan Turing Institute, UK [‡] Amazon Alexa, Cambridge, MA, USA

ABSTRACT

Humans categorise and structure perceived acoustic signals into hierarchies of auditory objects. The semantics of these objects are thus informative in sound classification, especially in few-shot scenarios. However, existing works have only represented audio semantics as binary labels (e.g., whether a recording contains *dog barking* or not), and thus failed to learn a more generic semantic relationship among labels. In this work, we introduce an ontology-aware framework to train multi-label few-shot audio networks with both relative and absolute relationships in an audio taxonomy. Specifically, we propose label-dependent prototypical networks (LaD-ProtoNet) to learn coarse-to-fine acoustic patterns by exploiting direct connections between parent and children classes of sound events. We also present a label smoothing method to take into account the taxonomic knowledge by taking into account absolute distance between two labels w.r.t the taxonomy. For evaluation in a real-world setting, we curate a new dataset, namely FSD-FS, based on the FSD50K dataset and compare the proposed methods and other few-shot classifiers using this dataset. Experiments demonstrate that the proposed method outperforms non-ontology-based methods on the FSD-FS dataset.

***Index Terms*—** Few-shot learning, multi-label classification, audio taxonomy, everyday sound recognition

1. INTRODUCTION

Everyday sound recognition is to classify types of sound events in a recording. It is a core task of machine listening and involves many practical applications, such as smart cities [1, 2] and bioacoustics [3]. While many works in the past years have succeeded in recognising sound events using large amounts of labelled data [4, 5], these methods are not always suitable to real-world scenarios where it takes great effort to gather sufficient amounts of annotated data for each category or there exist sound events of unknown classes in the inference stage.

Recently, some works proposed the use of few-shot learning in everyday sound classification [6, 7, 8]. These classifiers can rapidly learn new acoustic patterns with a small set of labelled examples, largely due to their different training objective. However, they are still restricted to using the ground truth as a binary attribute (e.g., whether the recording contains *dog barking* or not), instead of capturing the audio semantics in labels. Fig.1 showcases an example of four predictions: *Hoot*; *Bird*; *Chirp*; and *Water* for examples with the ground-truth *Hoot*. It can be observed that the three false positives are not equally "wrong" as their semantic meaning varies. In other words, the prediction with *Water* is more "wrong" than *Bird* and *Chirp*. This suggests that labels cannot be assumed to be independent with each other. Thus, one question arises: *can we leverage the relationship between labels to improve audio encoders in data-scarce scenarios?*

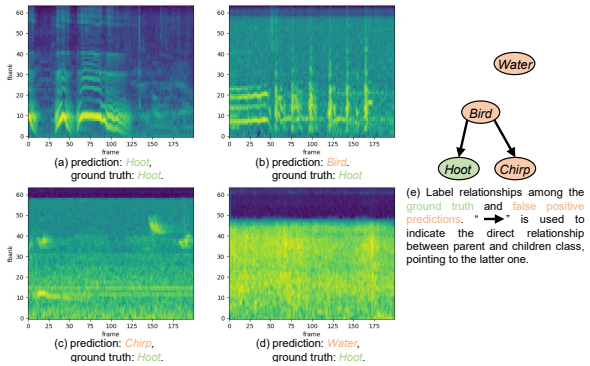

Fig. 1. **Audio labels convey more information than a binary attribute**. Given the ground-truth *Hoot*, the false positives (b-d) reflect different extents where the model captures the audio semantics of audio input. As shown in (e), *Bird* in prediction (b) is the parent class of the ground-truth, *Chirp* in the prediction (c) is a sibling class, and *Water* in the prediction (d) is an irrelevant class.

In this work, we introduce an ontology-aware framework, namely label-dependent prototypical network (LaD-ProtoNet), to learn audio semantics from abstract to fine-grained levels using both relative and absolute label relationships in an audio taxonomy. To that end, we first convert a multi-label classification task to multiple single-label classification tasks. Particularly, when both a parent and a child class in the taxonomy are present in the ground-truth, LaD-ProtoNet takes the classification task for the parent class with a higher priority. The network will thus learn audio semantics from abstract to fine-grained levels. In addition, we propose taxonomy-informed (Ti) label embedding, a label smoothing method that encodes pairwise label distance w.r.t the ontology [9] into the ground truth. Experiments show that the LaD-ProtoNet alone outperforms the non-ontology-based methods by a large margin. When combined with Ti-embedded label, LaD-ProtoNet can yield an even better performance. The contributions of this paper are three-fold:

i) We introduce LaD-ProtoNet to exploit the label relationship in sound recognition. The network handles the classification task associated with more abstract label with a higher priority than that associated with a more fine-grained label.
ii) We propose Ti label embedding to encode label distance in the ontology, improving the model's performance with negligible computational cost.
iii) We curate a new, large-scale database, FSD-FS, for multi-label few-shot audio classification. Different from existing datasets, FSD-FS is publicly available, making it a useful dataset for benchmarking few-shot audio classification.

[◇]The work was done when H. Phan was at School of Electronic Engineering and Computer Science, Queen Mary University of London, UK and The Alan Turing Institute, UK and prior to joining Amazon.

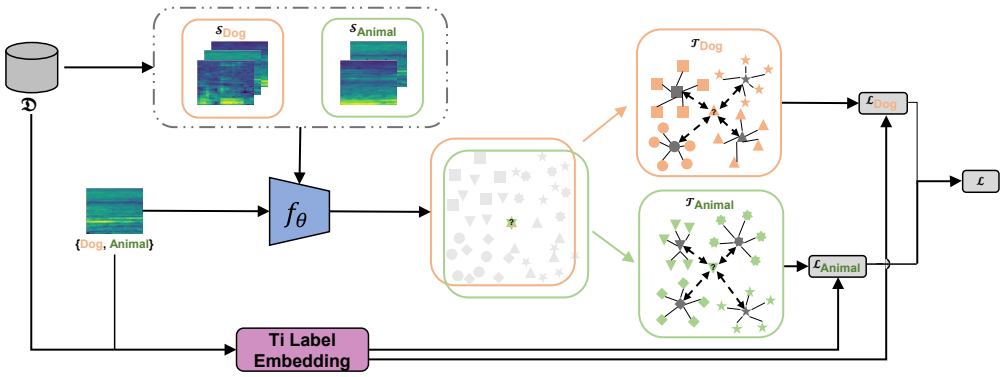

Fig. 2. **Ontology-aware learning**. Suppose a query example is associated with two labels: *Dog* and *Animal*, where *Animal* is the parent class of *Dog*. $\mathcal{T}_{Dog}$ and $\mathcal{T}_{Animal}$ are formed and two support sets $\mathcal{S}_{Dog}$ and $\mathcal{S}_{Animal}$ are sampled from the training set $\mathcal{D}$, respectively, to train the network. On each task, Ti-embedded labels are derived to take into account the taxonomic knowledge into the ground-truth labels that is incorporated into the loss function for network training.

2. RELATED WORK

2.1. Few-shot everyday sound recognition

There exist some works applying few-shot learning to everyday sound recognition [8, 10, 6, 11, 12]. Heggan *et al.* implemented various few-shot algorithms in some everyday sound datasets for single-label few-shot classification [8]. Targeting the multi-label few-shot problem, Wang *et al.* curated a synthesized dataset, FSD-MIX-CLIPS and FSD-MIX-SED, and compared model performance by controlling some generative factors in FSD-MIX-SED [6]. Cheng *et al.* adapted existing single-label few-shot algorithms to multi-label classification by proposing a One-vs.-Rest strategy [11]. They then evaluated their methods on the AudioSet [9] dataset. Shi *et al.* used meta-learning algorithms as well as linear regression on AudioSet and found that meta learning performed better than other few-shot methods [12]. We note that while Cheng and Shi both conducted experiments using AudioSet, their results are not comparable since AudioSet is not released to the public directly and neither of them detailed how the database was adapted for few-shot learning.

2.2. Prototypical networks

Given a training set $\mathcal{D} = \{(\mathbf{x}_i, c_i)\}_{i=1}^{|\mathcal{D}|}$ where $\mathbf{x}_i$ denotes the feature vector, $c_i \in \mathcal{C}$ denotes the discrete label of the i-th example, and $\mathcal{C}$ denotes the label set of C classes, $\mathcal{C} = \{1, \ldots, C\}$. Prototypical networks are trained with a series of "N-way K-shot" classification problems formed from the training set $\mathcal{D}$.

For such a "N-way K-shot" problem, the classification task is formed by three ingredients: (i) a label subset $\mathcal{C}_s$ of N classes sampled from $\mathcal{C}$, (ii) K examples (known as support examples) sampled from $\mathcal{D}$ for each class in $\mathcal{C}_s$, and (iii) Q examples (known as query examples) further sampled from $\mathcal{D}$ for each class in $\mathcal{C}_s$. For a class $n \in \mathcal{C}_s$, let $\mathcal{S}_n$ be the subset of support examples belonging to this class, $|\mathcal{S}_n| = K$. The prototype $\mathbf{a}_n$ of class n is then derived as the mean of embedding vectors of the support examples in $\mathcal{S}_n$. Formally,

$$\mathbf{a}_n = \frac{1}{K} \sum_{(\mathbf{x},c) \in \mathcal{S}_n} f_\phi(\mathbf{x}), \tag{1}$$

where f denotes the embedding mapping realized by the model whose parameters are denoted collectively as ϕ.

Given a query example $\mathbf{x}_q$, the model performs classification by producing a probability distribution over N classes in $\mathcal{C}_s$ based on a softmax over distances between $\mathbf{x}_q$ and the N prototypes in the embedding space. More specifically, the probability that $\mathbf{x}_q$ is classified as class $n \in \mathcal{C}_s$ is calculated as

$$p_\phi(\hat{y}_q = n|\mathbf{x}_q) = \frac{\exp(-d(f_\phi(\mathbf{x}_q), \mathbf{a}_n))}{\sum_{j \in \mathcal{C}_s} \exp(-d(f_\phi(\mathbf{x}_q), \mathbf{a}_j))}, \tag{2}$$

where $\hat{y}_q$ is the predicted label for $\mathbf{x}_q$, d is a distance function, such as ℓ_2 or cosine distance. The network is trained to minimize the negative log-probability of the true class over the $N \times Q$ query examples:

$$\mathcal{L}(\phi) = \sum\nolimits_{(\mathbf{x},c) \in \mathcal{Q}} - \log p_\phi(\hat{y} = c|\mathbf{x}), \tag{3}$$

where $\mathcal{Q}$ is the set of query examples, $|\mathcal{Q}| = N \times Q$.

Although prototypical networks perform well in many applications [13, 14], they are not suitable to multi-label few-shot classification directly where "N-way K-shot" problems are hard to formulate since labels often co-occur with each other (i.e., multi-label setting).

2.3. Label smoothing

Label smoothing originates from the idea of knowledge distillation where soft labels are derived from one-hot ground-truth by an ensemble system [15]. It helps models avoid over-confidence in the training process. Szegedy *et al.* simplified this technique by replacing pre-trained models with a uniform distribution [16]. Bertinetto *et al.* incorporated semantic information into the ground-truth by considering distances between different classes in a taxonomy [17]. The distance between two labels is measured by counting the intermediate nodes between them. However, their method is not suitable for a hierarchical label set as it cannot embed labels from different levels of the hierarchy. This work improves this taxonomy-aware label smoothing technique for multi-label few-shot settings by adopting a different distance measure.

3. PROPOSED METHOD

We propose an ontology-aware framework to train prototypical networks for multi-label few-shot settings and to take into account label relationships in a given taxonomy. The proposed framework is illustrated in Fig. 2 and contains two core techniques: LaD-ProtoNet and Ti label embedding. LaD-ProtoNet takes a multi-label example as input and converts the multi-label classification task into multiple single-label tasks w.r.t. the input labels, enabling training prototypical networks for few-shot settings. Rather than treating the single-label

tasks equally, LaD-ProtoNet puts more importance to those associated with more abstract labels (i.e., in higher levels of the hierarchy) so that it can learn patterns from coarse to fine-grained levels. In addition, the Ti-embedded label takes into consideration label relationships in the taxonomy by incorporating the label distances into the ground-truth.

3.1. LaD-ProtoNet

Let $(\mathbf{x}_q, \mathbf{y}_q)$ be a query example in the training set $\mathcal{D}$. Note that we are dealing with multi-label classification here, thus, $\mathbf{y}_q \in \{0,1\}^C$ is a multi-hot encoding vector. Assume that there are M positive classes present in $\mathbf{y}_q$, we denote the set of these M positive classes as $\mathcal{M}$. In order to deal with multiple labels for few-shot learning settings with prototypical networks, in LaD-ProtoNet, we first convert the multi-label classification task into M single-label classification tasks, $\{\mathcal{T}_m\}_{m\in\mathcal{M}}$, as follows.

For the task $\mathcal{T}_m$, in addition to the positive class $m \in \mathcal{M}$, we randomly sample $N-1$ classes from $\mathcal{C} \setminus \mathcal{M}$, which will serve as the negative classes, resulting in the label set $\mathcal{C}_m$, $|\mathcal{C}_m| = N$, for the classification task $\mathcal{T}_m$. Subsequently, for each class in $\mathcal{C}_m$, we sample (without replacement) K examples from $\mathcal{D}$, making $N \times K$ examples for the support set $\mathcal{S}_m$. Let $\mathcal{S}_m^n$ denote the subset of $\mathcal{S}_m$ corresponding to a class $n \in \mathcal{C}_m$. A prototype is then derived for each class n as

$$\mathbf{a}_m^n = \frac{1}{|\mathcal{S}_m^n|} \sum\nolimits_{(\mathbf{x},\mathbf{y})\in\mathcal{S}_m^n} f_\theta(\mathbf{x}). \tag{4}$$

Note that due to multi-label, $|\mathcal{S}_m^n|$ is not necessarily equal to K as each support example can be associated with more than one class in the label set $\mathcal{C}_m$. We then calculate l_2 distance between the query example $\mathbf{x}_q$ and the N prototypes in the embedding space and produce the probability distribution over the N classes in $\mathcal{C}_m$ via a softmax. Specifically, the probability for $\mathbf{x}_q$ to be classified as class $n \in \mathcal{C}_m$ is calculated as

$$p_\phi(\hat{y}_q = n|\mathbf{x}_q) = \frac{\exp(-d(f_\theta(\mathbf{x}_q), \mathbf{a}_m^n))}{\sum_{j\in\mathcal{C}_m} \exp(-d(f_\theta(\mathbf{x}_q), \mathbf{a}_m^j))}, \tag{5}$$

where d denotes the l_2 distance function. The loss induced by the task $\mathcal{T}_m$ on the query example $\mathbf{x}_q$ is calculated as

$$\mathcal{L}_m(\phi) = -\log p_\phi(\hat{y}_q = m|\mathbf{x}_q). \tag{6}$$

The network is optimized to minimize the total loss induced by all M tasks on the query example $\mathbf{x}_q$:

$$\mathcal{L}(\phi) = \sum\nolimits_{m\in\mathcal{M}} \mathcal{L}_m(\phi), \tag{7}$$

where $\mathcal{L}_m(\phi)$ is given in (6).

While the above method converts a multi-label classification problem to multiple single-label classification tasks and enables training prototypical networks for multi-label few-shot settings, it assumes the independence between the labels, and thus, ignores the label relationship during training. To further take into account the label relationship w.r.t a taxonomy, we incorporate parent-children relationship in the training objective in (7) and re-write it as

$$\mathcal{L}_{(\phi)} = \sum\nolimits_{m\in\mathcal{M}} \max\Big(\mathcal{L}_m(\phi), \mathbf{1}\left(\mathcal{P}\left(m\right), \mathcal{M}\right) \mathcal{L}_{\mathcal{P}(m)}(\phi)\Big), \tag{8}$$

where $\mathcal{P}(m)$ is the function mapping a child label m to its parent label in the taxonomy and

$$\mathbf{1}(\mathcal{P}(m), \mathcal{M}) = \begin{cases} 1 & \text{if} \quad \mathcal{P}\left(m\right) \in \mathcal{M} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

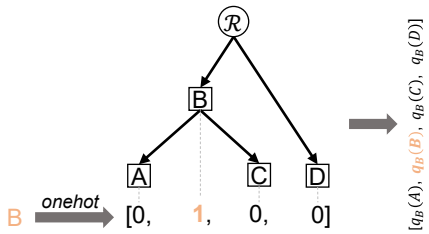

Fig. 3. **Ti label embedding**. $\mathcal{R}$ represents the root node. Ti-embedded label encodes the taxonomic knowledge into the label via label smoothing. It then measures the taxonomy distance between a label in $\mathcal{C}_m = \{A, B, C, D\}$ to the posible label B and produces the probability distribution via a softmax over the distances.

With the integration of parent-children relationship in (8), the network will be optimised using the loss $\mathcal{L}_m(\phi)$ associated with the child label m only if it can perform better on $\mathcal{L}_{\mathcal{P}(m)}(\phi)$ associated with its parent $\mathcal{P}(m)$.

It is interesting to find that eq. (8) behaves as a simple aggregation of the task losses when the set $\mathcal{M}$ does not contain any parent-children label pairs and that One-vs.-Rest selection strategy in [11] is a special case of the proposed LaD-ProtoNet. It should be noted that LaD-ProtoNet can be easily extended to multi-level ontology by decomposing the tree structure into parent-children pairs.

3.2. Taxonomy-informed label smoothing

In addition to parent-children relationships, we take into account the audio taxonomic knowledge via label smoothing. Fig. 3 gives an illustrative example on how the Ti label embedding works. Assume that for the task $\mathcal{T}_m$ in the LaD-ProtoNet framework in Section 3.1, the positive label is B, i.e., $m = B$. Furtheremore, assume that we have sampled three negative classes A, C, and D. That is, the label set $\mathcal{C}_m = \{A, B, C, D\}$. We derive the Ti-embedded label encoding vector $\mathbf{q}_B = (q_B(A), q_B(B), q_B(C), q_B(D)) \in [0,1]^N$ ($N = 4$ in this case) for the query example, where

$$q_B(\ell) = \frac{\exp(-\beta d_t(\ell, B))}{\sum_{j\in\mathcal{C}_m} \exp(-\beta d_t(j, B))}, \tag{10}$$

for $\ell \in \mathcal{C}_m$. Here, β is a temperature factor controlling how class probabilities are distributed among the labels. $d_t(i, j)$ denotes the taxonomy distance which is measured by the number of edges between two nodes i and j. In that way, the negative classes A, C, and D are assigned with nuance probabilities larger than 0 while the probability of the positive class B is slightly lower than 1.

For generalization, with a derived Ti-embedded label $\mathbf{q}_m$ w.r.t the positive label m, the loss in (6) is re-written as:

$$\mathcal{L}_m(\phi) = \sum_{n\in\mathcal{C}_m} -q_m(n) \log p_\phi(\hat{y}_q = n|\mathbf{x}_q). \tag{11}$$

Please note that the Ti-embedded label encoding vector approaches a one-hot encoding vector when β increases and approaches a uniform distribution when β decreases.

4. EXPERIMENTS

4.1. FSD-FS dataset

We curated a multi-label few-shot database, namely FSD-FS, by adapting the FSD50K dataset [19]. We inherited the taxonomy from FSD50K and excluded part of them to avoid the issue when there are multiple paths to travel from a node to the root node of the taxonomy.

Table 1. Comparison of different methods in terms of mAP, AUC, F1-score with 0.95 confidence. The best results are highlighted in **bold**.

	validation set (%)			evaluation set (%)		
	mAP ↑	AUC ↑	F1-score ↑	mAP ↑	AUC ↑	F1-score ↑
Baseline [18]	33.02±1.04	83.73±0.80	37.32±0.75	34.75±1.39	84.81±0.97	39.29±1.37
one-vs.rest [11]	38.71±1.06	86.07±0.29	41.65±0.64	38.71±2.00	86.71±1.41	42.82±1.93
LaD-ProtoNet (β=15)	39.36±0.90	**86.10±0.33**	42.04±0.58	**40.33±1.57**	**87.10±0.73**	43.82±1.21
LaD-ProtoNet (β=30)	39.71±0.56	85.77±0.67	42.16±0.98	40.05±0.58	87.04±0.58	**43.97±0.29**
LaD-ProtoNet (β=45)	**39.98±0.51**	86.01±0.30	**42.47±0.57**	39.68±0.75	86.97±0.31	43.40±0.74

The rendered FSD-FS spans across 143 classes and contains 43,805 raw audio recordings. Following [20], we split the label set with the ratio 7:2:1, resulting in 98 classes in the base set, 30 classes in the validation set, and 15 classes in the evaluation set. More details can be found in the supplemental material[1] and data repository[2].

4.2. Experimental setup

We used prototypical networks as the baseline. The models (i.e., the baseline and the proposed LaD-ProtoNet models) were trained with 15-way classification tasks to match the evaluation condition where only 15 classes are available. Note that we excluded irrelevant labels whose classes are not sampled in a "N-way K-shot" problem.

In all the experiments, the audio recordings were sampled at 44.1kHz. Log-Mel spectrogram was used as input. A spectrogram was extracted from an audio recording using a window length of 20ms with 50% overlap, and 64 Mel filters. In addition, the spectrograms were z-normalized along each Mel bin. We applied a 8-layer convolutional neural network (CNN) as audio encoder to all the few-shot learner for a fair comparison. Details of the network architecture can be found in our available implementation[3].

4.3. Experiment results

Table 2. Ablation study on FSD-FS evaluation split. LaD denotes the label-dependent structure, and Ti represents the Ti-embedded label.

LaD	Ti	mAP ↑	AUC ↑	F1-score ↑
✓		39.34±1.24	86.92±0.62	43.20±0.88
	✓	39.79±0.67	**87.23±0.28**	43.41±0.57
✓	✓	**40.33±1.57**	87.10±0.73	**43.82±1.21**

Table 1 compares the performance of different methods in terms of mAP, AUC, and F1-score. It can be seen that our proposed LaD-ProtoNets with Ti-embedded labels obtain better performance than both the baseline and the one-vs.rest method over all the evaluation metrics. This indicates that capturing the relationships between labels does help models learn useful features for classification.

Table 2 shows the ablation study of the proposed framework on the evaluation split of FSD-FS. It achieves the best performance in terms of mAP and F1-score with the combination of LaD-ProtoNet and the Ti-embedded label. We should note that the standalone LaD-ProtoNet has the relative parent-children relationship integrated while coupling with the Ti-embedded labels, the network is able to leverage the absolute relationship between any two labels in the taxonomy. This implies that the two techniques are compliment to each other.

The effect of β to the mAP metric of the models coupled with Ti-embedded labels is shown in Fig. 4. The best mAP is highest on the evaluation and validation set with β=30, 45, respectively. We note

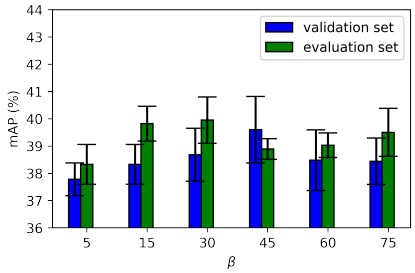

Fig. 4. Performance comparison of the models coupled with the Ti-embedded label with various β values.

that the model with β=45 performs worse on the evaluation set than the one with β=30 while it achieves the best mAP in the validation set. This is a sign of outfitting, suggesting that models' generalisation deteriorates when the Ti-embeded labels approximate the one-hot encoding vector.

5. CONCLUSION

We proposed the LaD-ProtoNet framework for multi-label few-shot audio classification. In the framework, a multi-label classification task was converted into multiple single-label tasks that makes few-shot learning feasible. LaD-ProtoNet then took into account the parent-children relationships in a given sound taxonomy and purposed the model training so that the task associated with a parent label (in more abstract level) was handled with higher importance than the one associated with a child node (in more fine-grained level). Beyond the parent-children relationship, we further proposed Ti label embedding to encode knowledge of the audio taxonomy into the ground-truth labels by considering the taxonomic distance between a label pair. Evaluations conducted on a newly curated dataset, FSD-FS, showed that the proposed framework outperformed the baseline by 1.34% absolute in terms of mAP.

Although this work proves that taxonomy knowledge can benefit few-shot sound classification, it still needs a predefined audio ontology which restricts the method to a close-world knowledge. In future work, we will explore approaches to leverage label correlations without predefined taxonomies in multi-label few-shot audio classification.

6. ACKNOWLEDGEMENTS

J. Liang is supported by the Engineering and Physical Sciences Research Council [grant number EP/T518086/1]. E. Benetos is supported by a RAEng/Leverhulme Trust Research Fellowship [grant number LTRF2223-19-106]. 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/LaD-ProtoNet/blob/main/assets/appendix.pdf
[2] https://zenodo.org/record/7557107
[3] https://github.com/JinhuaLiang/LaD-ProtoNet

7. REFERENCES

[1] Graham Dove, Charlie Mydlarz, Juan Pablo Bello, and Oded Nov, "Sounds of New York city," *Interactions*, vol. 29, no. 3, pp. 32–35, 2022.

[2] Andrew Mitchell, Emmeline Brown, Ratneel Deo, Yuanbo Hou, Jasper Kirton-Wingate, Jinhua Liang, Alisa Sheinkman, Christopher Soelistyo, Hari Sood, Arin Wongprommoon, Kaiyue Xing, Wingyan Yip, and Francesco Aletta, "Deep learning techniques for noise annoyance detection: Results from an intensive workshop at the Alan Turing Institute," *The Journal of the Acoustical Society of America*, vol. 153, no. 3_supplement, pp. A262–A262, Mar. 2023.

[3] Veronica Morfi, Inês Nolasco, Vincent Lostanlen, Shubhr Singh, Ariana Strandburg-Peshkin, Lisa F. Gill, Hanna Pamula, David Benvent, and Dan Stowell, "Few-Shot Bioacoustic Event Detection: A New Task at the DCASE 2021 Challenge.," in *DCASE*, 2021, pp. 145–149.

[4] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020, Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[5] Yuan Gong, Yu-An Chung, and James Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[6] Yu Wang, Nicholas J Bryan, Justin Salamon, Mark Cartwright, and Juan Pablo 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)*. 2021, pp. 36–40, IEEE.

[7] Jinhua Liang, Huy Phan, and Emmanouil 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.

[8] Calum Heggan, Sam Budgett, Timothy Hospedales, and Mehrdad Yaghoobi, "MetaAudio: A Few-Shot Audio Classification Benchmark," in *Artificial Neural Networks and Machine Learning – ICANN 2022*, Elias Pimenidis, Plamen Angelov, Chrisina Jayne, Antonios Papaleonidas, and Mehmet Aydin, Eds., Cham, 2022, pp. 219–230, Springer International Publishing.

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 776–780, IEEE.

[10] Jinhua Liang, Xubo Liu, Haohe Liu, Huy Phan, Emmanouil Benetos, Mark D. Plumbley, and Wenwu Wang, "Adapting Language-Audio Models as Few-Shot Audio Learners," in *Proc. INTERSPEECH 2023*, 2023, pp. 276–280.

[11] Kai-Hsiang Cheng, Szu-Yu Chou, and Yi-Hsuan Yang, "Multi-label Few-shot Learning for Sound Event Recognition," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, Sept. 2019, pp. 1–5, ISSN: 2473-3628.

[12] Bowen Shi, Ming Sun, Krishna C. Puvvada, Chieh-Chi Kao, Spyros Matsoukas, and Chao Wang, "Few-Shot Acoustic Event Detection Via Meta Learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 76–80, ISSN: 2379-190X.

[13] Junying Huang, Fan Chen, Keze Wang, Liang Lin, and Dongyu Zhang, "Enhancing Prototypical Few-Shot Learning By Leveraging The Local-Level Strategy," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1660–1664, ISSN: 2379-190X.

[14] Ren Li, Jinhua Liang, and Huy Phan, "Few-Shot Bioacoustic Event Detection: Enhanced Classifiers for Prototypical Networks," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, Nov. 2022.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the Knowledge in a Neural Network," in *NIPS 2014 Deep Learning Workshop*, Mar. 2015.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord, "Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 12503–12512, ISSN: 2575-7075.

[18] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical Networks for Few-shot Learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

[19] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier 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.

[20] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M. Bronstein, "LaSO: Label-Set Operations Networks for Multi-Label Few-Shot Learning," 2019, pp. 6548–6557.