HIERARCHY-AWARE LOSS FUNCTION ON A TREE STRUCTURED LABEL SPACE FOR AUDIO EVENT DETECTION

Arindam Jati[1], *Naveen Kumar*[†], *Ruxin Chen*[2], *Panayiotis Georgiou*[1]

[1]University of Southern California, Los Angeles, CA, USA
[2] SONY Interactive Entertainment LLC, San Mateo, CA, USA

ABSTRACT

The paper introduces a hierarchy-aware loss function in a Deep Neural Network for an audio event detection task that has a bi-level tree structured label space. The goal is not only to improve audio event detection performance at all levels in the label hierarchy, but also to produce better audio embeddings. We exploit the label tree structure to preserve that information in the hierarchy-aware loss function. Two different loss functions are separately employed. First, a triplet loss with probabilistic multi-level batch mining is introduced. Second, a quadruplet learning method is applied, which is a special case of generalized triplet learning for bi-level label taxonomy. The training is performed in a multi-task learning framework by jointly optimizing cross entropy based loss and hierarchy-aware loss function. The proposed method is found to outperform the baseline cross entropy based models at both levels of the hierarchy. The multi-task model is also able to learn better audio representations as observed in our clustering experiments. Moreover, the model is shown to transfer well when an out-of-domain dataset is used for evaluation.

Index Terms— Hierarchical audio event detection, metric learning, triplet loss, quadruplet loss, convolutional neural network.

1. INTRODUCTION

Audio Event Detection (AED) deals with understanding and recognizing the semantic class (generally human annotated) associated with an audio recording. Recently it has gained popularity [1, 2, 3] due to its numerous applications in surveillance [4, 5], audio content understanding and retrieval [6], context detection [7], and even health monitoring systems [8].

Generally, an audio events ontology is hierarchical in nature [2, 9, 10] because inherently it is easier for humans to identify (and hence annotate) first the coarse class of the audio (*e.g.,* vehicle), and then the fine class (*e.g.,* bus). This paper focuses on detecting audio events that have a hierarchical relationship in their human annotated label space. We have two complementary objectives:

1. Train a model that can identify the audio events satisfactorily at all levels in the label hierarchy, possibly by exploiting the hierarchical label taxonomy.
2. The model should be able to produce a distinctive audio *embedding* [11] or representation [12] that tries to follow the label hierarchy in a lower dimensional manifold with respect to some distance measure.

One classical approach [13] for AED is GMM fitting on the "bag of frames" modeling of the audio recording and applying KL divergence measure between different GMM models. MFCC features are popular for this kind of approaches. Chu *et al.* [14] proposed a matching pursuit technique for finding useful time-frequency features to complement MFCC features, and obtained improved performance. Later, time series models such as HMMs were employed to better use context over time [15]. A detailed survey of non-deep learning approaches can be found in [1]. Recently Deep Neural Network (DNN) [16] has shown promise for AED. A DNN based approach was shown to outperform GMM in [17] for classifying 61 audio classes. In [18], a Convolutional Neural Network (CNN) was shown to extract robust features for noisy AED task. Some other works [19, 20] also showed potential of CNN in extracting robust features directly from the spectrogram for AED. In 2017, Google released an AED corpus, AudioSet [2] containing $\sim 1.8M$ 10s excerpts from YouTube videos, which is much larger than the previous datasets. Again, CNN based models like ResNet-50 [21] gave quite satisfactory performance (0.959 AUC) [3] for classifying 485 audio classes on AudioSet.

Surprisingly, little work can be found in the field of AED that directly address the problem of hierarchical classification in audio event taxonomy [10, 9]. Xu *et al.* [22] proposed a DNN based multi-task learning method to solve this problem for a dataset having 3 coarse classes and 15 fine classes. Pre-training the DNNs [22] separately for coarse and fine classes was found to be helpful before applying the weighted multi-task (coarse- and fine-level classifications) cross entropy objective function. But, the number of audio classes were very limited in this work.

In this paper, our application requires dealing with an unprecedented number ($\sim$ 5K) of specific AED classes defined on a hierarchical ontology similar to AudioSet. The goal of learning a distinctive audio manifold guides us to employ a loss function that can harness the hierarchy information of the label space, and force the embeddings to follow that through some imposed similarity constraints during DNN training. The employed hierarchy-aware loss functions are found to be complementary with standard cross entropy loss, and together, in a multi-task learning setting, they improve the AED performance at both higher and lower levels in the taxonomy.

The rest of the paper is divided into the following sections. Section 2 introduces the hierarchical audio event dataset and the motivation behind applying hierarchy-aware loss function on it. Section 3 describes the methodology. Experimental settings are reported in Section 4. Results and discussions are provided in Section 5. Finally conclusions are drawn and future directions are given in Section 6.

2. HIERARCHICAL AUDIO EVENTS

2.1. Dataset

Our manually labeled hierarchical Audio Event (AE) dataset has 183 AE classes, and each class has one or more AE subclass(es). An

This work was supported by SONY Interactive Entertainment LLC.

[†]The work was done while the co-author was associated with SONY.

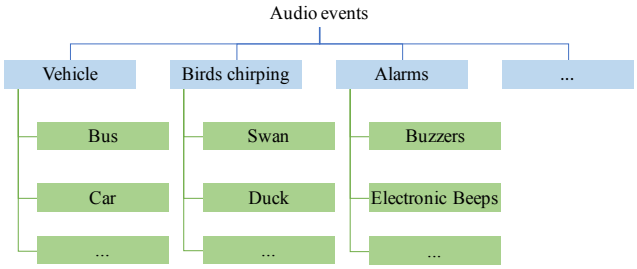

Fig. 1: An example of the hierarchical audio events class structure in our dataset.

example of the class hierarchy is shown in Fig. 1. We will represent the dataset labels in two levels: "coarse" and "fine". So, there are total 183 coarse labels (*e.g.,* vehicle, birds chirping *etc.* in Figure 1). A fine class label carries information about both AE class (*e.g.,* vehicle) and subclass (*e.g.,* bus). For example, fine label, c_{kl} denotes k^{th} class and l^{th} subclass of that class. There are total 4721 unique fine labels. The dataset has around 230K audio samples with variable durations. Each audio stream has only one audio event. The dataset was manually recorded and annotated by professional sound engineers.

2.2. Motivation to use hierarchy-aware loss

The complementary objectives, as introduced in Section 1, direct us to learn fine-grained feature representation [23] in the data, that can help comparing the audio events at different levels in their taxonomy. A common approach is to learn distance metrics in the embedding space by imposing similarity constraints during training. Pairwise contrastive loss [24] and triplet loss [11] are some of the popular metrics in the field of face recognition. But, these loss functions do not consider hierarchical label structures as in the case of our current AED problem. To solve this issue, we propose a variant of the standard triplet learning by deploying a knowledge driven probabilistic triplet mining that utilizes the hierarchy information while sampling the triplets. In [23], the authors introduced a *generalized triplet* learning, which has the inherent ability to work on label tree structures. We also employ this method in our AED task. We build a multi-task learning [25] framework where a multi-objective loss function, combining the hierarchy-aware loss and cross entropy loss, has been employed to train the DNN, inspired by some recent successes in the computer vision field [23, 26].

3. HIERARCHY-AWARE LOSS ON TREE STRUCTURED LABEL SPACE

3.1. Problem formulation

Let, $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a dataset of N variable length audio samples having following labels in a bi-level hierarchy.

$$\text{Coarse: } \mathcal{L}_C = \{y_1, y_2, \ldots, y_N\}, y_i \in \{1, 2, \ldots, C\} \quad (1)$$

$$\text{Fine: } \mathcal{L}_F = \{z_1, z_2, \ldots, z_N\}, z_i \in \{1, 2, \ldots, F\} \quad (2)$$

Here, C and F are number of coarse and fine classes respectively ($F >> C$). Please note that, as mentioned in Section 2.1, $z_i = c_{kl}$ carries information for both AE class and subclass. Our complementary objectives (as introduced in Section 1) can now be defined:

1. Train a non-linear mapping $\mathcal{M}$ such that it maximizes the classification accuracy at both label spaces $\mathcal{L}_C$ and $\mathcal{L}_F$.
2. Model $\mathcal{M}$ should also provide an intermediate mapping, $f(\mathbf{x}) \in \mathbb{R}^d, \forall \mathbf{x} \in \mathcal{D}$ and $||f(\mathbf{x})||_2^2 = 1$, that tries to project audio on a manifold such that their mutual Euclidean distances obey the order as found in the label space hierarchy. To explain, let $\mathbf{x}_1, \mathbf{x}_2 \in S_1 \subset G_1$, $\mathbf{x}_3 \in S_2 \subset G_1$, $\mathbf{x}_4 \in S_3 \subset G_2$. Here, G_i and S_i denote class and subclass respectively. Then, ideally in the embedding space, $||f(\mathbf{x}_1) - f(\mathbf{x}_2)||_2^2 < ||f(\mathbf{x}_1) - f(\mathbf{x}_3)||_2^2 < ||f(\mathbf{x}_1) - f(\mathbf{x}_4)||_2^2$.

Note that the l_2 normalization of the embeddings is required to constrain them to lie on the d-dimensional hypersphere [11]. Here, $\mathcal{M}$ is the DNN with *softmax(.)* outputs. The embedding layer output, $f(\mathbf{x})$ is connected to multiple *softmax(.)* outputs (with C or F output units depending on coarse- or fine-level training) through a single linear layer.

3.2. Baseline DNN

We want to analyze the effect of introducing the hierarchy-aware loss functions on standard cross entropy learning for AED task. We separately train the baseline DNN (Section 4.1 describes the architecture) with two Cross Entropy (CE) loss functions. First (will be called "CE coarse"), we train it with coarse labels and the following cross entropy objective function:

$$\begin{aligned} &\underset{g}{\text{argmin}}\, L_{\text{CE coarse}}\left(g(\mathbf{x}), y\right) \\ &= \underset{g}{\text{argmin}}\, \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp\left(g\left(\mathbf{x}_i, y_i\right)\right)}{\sum_{j=1}^{C}\exp\left(g\left(\mathbf{x}_i, j\right)\right)} \end{aligned} \quad (3)$$

Here, $g(\mathbf{x}_i, y_i) = g_i(f(\mathbf{x}_i))$, denotes the final output of the DNN for the i^{th} class before applying *softmax(.)*. So, this model is unable to give fine representations, but it will help comparing the benefits of hierarchy-aware loss at coarse level. The second model ("CE fine") is trained with the following fine grained cross entropy loss:

$$\begin{aligned} &\underset{g}{\text{argmin}}\, L_{\text{CE fine}}\left(g(\mathbf{x}), z\right) \\ &= \underset{g}{\text{argmin}}\, \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp\left(g\left(\mathbf{x}_i, z_i\right)\right)}{\sum_{j=1}^{F}\exp\left(g\left(\mathbf{x}_i, j\right)\right)} \end{aligned} \quad (4)$$

This model will be able to predict at both levels.

3.3. Hierarchy-aware Loss

3.3.1. Balanced triplet loss

Triplet learning [11], inspired from contrastive pairwise learning [24], works on triplets $(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) \in \mathcal{T}$ that are sampled from the training data $\mathcal{D}$. Here, the anchor, $\mathbf{x}_i^a$ and the positive sample, $\mathbf{x}_i^p$ are from the same class; but the negative sample, $\mathbf{x}_i^n$ is from a different class. The global optimum in training data should find $f(.)$ such that, in the embedding space, they satisfy the constraint:

$$\begin{aligned} ||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)||_2^2 + \alpha < ||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)||_2^2, \\ \forall(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) \in \mathcal{T} \end{aligned} \quad (5)$$

Here $\alpha > 0$ is a margin parameter. This is done by minimizing the following loss function:

$$\underset{f}{\text{argmin}}\, \frac{1}{N_T}\sum_{i=1}^{N_T} L_{\text{Tri}}\left(f(\mathbf{x}_i^a), f(\mathbf{x}_i^p), f(\mathbf{x}_i^n), \alpha\right) \quad (6)$$

where $N_T = |\mathcal{T}|$, and

$$\begin{aligned} &L_{\text{Tri}}\left(f(\mathbf{x}_i^a), f(\mathbf{x}_i^p), f(\mathbf{x}_i^n), \alpha\right) = \\ &max(||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)||_2^2 + \alpha - ||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)||_2^2, 0) \end{aligned} \quad (7)$$

While sampling a triplet for training the DNN for hierarchical AED task, one has the option to pick the negative sample either from the same coarse class as the anchor but a different subclass, or from a completely different coarse class from the anchor. Mining triplets according to the uniform distribution does not consider the hierarchical label structure, and most of the negative samples are mined (with a probability of $\frac{C-1}{C} >> \frac{1}{C}$) from a coarse class that is different from the anchor's coarse class. So, the model mostly learns the coarse representation, and not the fine one. To alleviate the problem, we perform triplet *negative* mining in a probabilistic way so that the model encounters around 50% triplets where the negative sample comes from the same class but different subclass. If $\mathbf{x}_i^a \in S_i \subset G_j$, then we choose $\mathbf{x}_i^p$ s.t. $\mathbf{x}_i^p \in S_i$. Here, $S_i, \forall i = 1, \ldots, F$ is the fine set containing $\mathbf{x}_i^a$, and $G_j, \forall j = 1, \ldots, C$ is the coarse set that contains $\mathbf{x}_i^a$. The negative exemplar, $\mathbf{x}_i^n$ is randomly sampled conditioned on a Bernoulli distribution as follows:

$$\begin{aligned}\mathbf{x}_i^n = \mathbb{I}(r=0) \times \{\mathbf{x}_i^n : \mathbf{x}_i^n \in G_j \text{ and } \mathbf{x}_i^n \notin S_i\} \\ + \mathbb{I}(r=1) \times \{\mathbf{x}_i^n : \mathbf{x}_i^n \notin G_j\}\end{aligned} \tag{8}$$

Here, $r \sim Ber(0.5)$, and $\mathbb{I}(.)$ is the indicator function.

3.3.2. Quadruplet loss

The basic idea here is to generalize the triplet learning across multiple levels in the hierarchy [23]. For a bi-level tree like ours, a quadruplet, $(\mathbf{x}_i^a, \mathbf{x}_i^{p+}, \mathbf{x}_i^{p-}, \mathbf{x}_i^n) \in \mathcal{Q}$ is sampled from $\mathcal{D}$, s.t., if $\mathbf{x}_i^a \in S_i \subset G_j$, then $\mathbf{x}_i^{p+} \in S_i$ and $\mathbf{x}_i^{p-} \notin S_i$ but $\mathbf{x}_i^{p-} \in G_j$. On the other hand, the negative is chosen s.t. $\mathbf{x}_i^n \notin G_j$. The ultimate goal is to satisfy the following constraint in the embedding space:

$$\begin{aligned}||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^{p+})||_2^2 + \alpha < ||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^{p-})||_2^2 + \beta, \\ < ||f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)||_2^2, \quad \forall (\mathbf{x}_i^a, \mathbf{x}_i^{p+}, \mathbf{x}_i^{p-}, \mathbf{x}_i^n) \in \mathcal{Q}\end{aligned} \tag{9}$$

which is achieved by minimizing the objective function:

$$\begin{aligned}&\underset{f}{\text{argmin}} \frac{1}{N_Q} \sum_{i=1}^{N_Q} L_{\text{Quad}} \left(f(\mathbf{x}_i^a), f(\mathbf{x}_i^{p+}), f(\mathbf{x}_i^{p-}), f(\mathbf{x}_i^n), \alpha, \beta\right) \\ &\triangleq \underset{f}{\text{argmin}} \frac{1}{N_Q} \sum_{i=1}^{N_Q} \big[L_{\text{Tri}} \left(f(\mathbf{x}_i^a), f(\mathbf{x}_i^{p+}), f(\mathbf{x}_i^{p-}), \alpha - \beta\right) \\ &\qquad + L_{\text{Tri}} \left(f(\mathbf{x}_i^a), f(\mathbf{x}_i^{p-}), f(\mathbf{x}_i^n), \beta\right) \big]\end{aligned} \tag{10}$$

Here, $L_{\text{Tri}}(.)$ is as defined in Equation (7), $N_Q = |\mathcal{Q}|$, and α and β are two margin parameters satisfying $\alpha > \beta > 0$.

3.4. Multi-task learning

As mentioned in Section 2.2, we train the DNN in a multi-task learning environment (inspired by [23]) by imposing a joint objective function to minimize:

$$L_{\text{Multi}} = \lambda L_{\text{HAL}} + (1-\lambda) L_{\text{CE fine}} \tag{11}$$

where, $\lambda \in [0, 1]$ and, L_{HAL} is the hierarchy-aware loss function and it can be either L_{Tri} or L_{Quad}.

4. EXPERIMENTAL SETTING

4.1. DNN architecture

Inspired from the performance of deep CNN models in flat AED tasks [3], we employ a slightly modified version of recently proposed ResNet-18 model [21]. We change the input layer of the basic ResNet-18 model to confront to the single channel spectrogram inputs, and the output layer depending on number of audio classes. We replace the final average pooling layer by a fully connected layer that goes to a 512 dimensional embedding layer. We perform l_2 normalization of the embeddings during training and evaluation. The baseline model is trained end-to-end with cross entropy losses as described in Section 3.2. The multi-task model is trained with joint objectives as shown in Equation (11).

4.2. Data split

A dataset splitting has been performed to produce disjoint (in terms of audio files) train ($\sim 185K$), validation ($\sim 22.5K$) and test ($\sim 22.5K$) sets. Early stopping [16] has been used based on the validation set performance for model selection. We should mention that the chance accuracies of a majority guess classifier for coarse and fine classification tasks are 4.65% and 1.1% respectively.

4.3. Features and parameters

64 dimensional mel spectrogram features have been extracted from single channel audio streams having sampling rate of 48KHz using moving window of $42.67ms$ (2048 samples) length and $10.67ms$ (512 samples) shift. Online batch mining is employed during training for fetching triplets or quadruplets. Random windows of 100 feature frames have been generated, and 100×64 dimensional samples are fed into the input CNN layer of the model for training. We do not implement hard negative sampling [11] for triplet or quadruplet mining to increase the training speed. Instead we use a large batch size of 1024 samples to increase the probability of finding some hard negative exemplars during random sampling. We use 8 GPUs for training with data parallelism. We employ Adam optimizer with a learning rate of 10^{-3} and l_2 regularization penalty of 10^{-6}. For triplet loss, we have chosen $\alpha = 0.1$. For quadruplet loss, we have picked $\alpha = 0.2$ and $\beta = 0.1$. The weighting factor, λ in Equation (11), is chosen to be 0.5 based on the validation set performance.

5. RESULTS AND DISCUSSIONS

5.1. Classification

Table 1 shows the performance of different methods for AED in terms of classification accuracies. The predictions on a test audio stream are generated by taking mean over all the posterior probabilities (non-overlapping sliding windows of 100 frames). 'Top k' accuracy calculates the classification accuracy by observing whether the true class is in the top k predictions. From the first two rows of Table 1, we can see that the Cross Entropy (CE) loss on fine labels gives a better Top 1 accuracy (even at the coarse level) than a CE loss on the coarse labels only. This indicates that the model has the ability to learn the fine labels and the knowledge of fine labels results in a better representation for classification (even at the coarse level).

The training using the multi-task loss function (Equation (11)) provides much better results at both levels than only CE supervision. We hypothesize that the hierarchy-aware loss helps to learn a better embedding space by reducing intra-cluster distance and increasing inter-cluster distance (in other words, following the constraints in Equation (5) and (9) in the embedding space), and in turn helps the CE loss. It can also be thought as a regularization term along with standard CE. We achieve **3.14%** and **3.88%** absolute improvements for Top 1 coarse (183-way), and fine (4721-way) classifications respectively.

Comparison between balanced triplet and quadruplet learning shows almost similar performance with quadruplet leading by a small margin in case of classification with fine labels. This might be

Table 1: Audio events classification accuracies of different models (CE=Cross Entropy)

		Coarse-level			Fine-level		
Model	Objective function (Section 3)	Top 1	Top 3	Top 10	Top 1	Top 3	Top 10
CE Coarse	Equation (3)	71.46%	85.82%	**94.30%**	N/A	N/A	N/A
CE Fine	Equation (4)	75.58%	85.79%	92.35%	65.82%	77.30%	84.44%
Multi-task (CE+Triplet)	Equation (6),(11)	**78.72%**	87.42%	93.39%	69.57%	79.76%	86.23%
Multi-task (CE+Quadruplet)	Equation (10),(11)	78.53%	**87.44%**	93.32%	**69.70%**	**79.92%**	**86.42%**

Table 2: Clustering evaluation of the test embeddings generated by different models (see Section 5.2 for metric acronyms)

	Coarse-level							Fine-level						
Model	H	C	V	ARI	AMI	Intra CD	Inter CCD	H	C	V	ARI	AMI	Intra CD	Inter CCD
CE fine	0.471	0.435	0.452	0.106	0.366	7.137	4.957	0.791	0.765	0.777	0.078	0.242	0.318	**1.190**
Multi-task (CE+Quadruplet)	**0.517**	**0.479**	**0.497**	**0.131**	**0.417**	**6.942**	**5.023**	**0.792**	**0.771**	**0.781**	**0.088**	**0.265**	**0.308**	1.153

happening because of the bi-level constraints (Equation (9)) imposed on the quadruplet loss (Equation (10)). The rest of the experiments are performed with the quadruplet loss model.

5.2. Clustering audio embeddings

To evaluate the effect of introducing hierarchy-aware loss on learning better manifold and in turn producing more compact audio embeddings, we cluster the embeddings of the test audio using K-means clustering (K=183 for coarse-, K=4721 for fine-level clustering). Table 2 shows different metrics evaluating the clustering performed on the embeddings generated by baseline DNN with fine level cross entropy and the multi-task algorithm with quadruplet loss. The same parameter settings are used for K-means for both the algorithms. The metrics are explained below (all of them implemented in scikit-learn [27]):

- Homogeneity (H), Completeness (C) and V-measure (V) [28]: If the true classes are known, then perfect homogeneity (1.0) occurs when each cluster contains samples from a single class. Perfect completeness ensures all members of a single class to stay inside a single cluster. V-measure is the harmonic mean of these two metrics.
- Adjusted Random Index (ARI): Given true labels and predicted cluster labels, ARI estimates a similarity between them ignoring possible permutations [29]. It varies from 0 to 1, and higher value is better.
- Adjusted Mutual Information (AMI): Mutual information between true and predicted cluster labels with a normalization to account for chance. A higher value is preferable.
- Intra Cluster Distance (Intra CD): This does *not* require a clustering algorithm, but only needs true class (or, cluster) labels. It simply calculates the average distance between all points inside a cluster, and takes average among all clusters. Note that the embeddings are l_2 normalized before computing this (and Inter CCD) metric(s). A *lower* value is better.
- Inter Cluster Centroids Distance (Inter CCD): Average distance between all cluster centroids. It is also independent of the clustering algorithm. A *higher* value is preferable.

We can see from Table 2 that the multi-task model outperforms the baseline DNN in all metrics at coarse level, and all except one metric (inter CCD) at fine label. The lower value of inter CCD at fine level might be coming from the quadruplet constraint as mentioned in Equation (9). The model learns to separate the fine level clusters, but not too much because it also tries to keep them under the parent coarse level cluster.

Table 3: Classification accuracies for AED on Greatest Hits dataset

	Coarse-level			Fine-level		
Model	Top 1	Top 3	Top 10	Top 1	Top 3	Top 10
Rand-init	64.89%	89.10%	98.64%	59.47%	85.19%	97.45%
Pre-trained	**74.86%**	**92.93%**	**99.20%**	**69.04%**	**90.29%**	**98.49%**

5.3. Transfer learning

To measure the transfer ability [30] of the model, we evaluate its performance on the Greatest Hits dataset [31]. It contains audio visual data of different *actions* (hit, scratch and other) performed on different *objects* (*e.g.,* dirt, glass, leaf *etc.*), and also their *reactions* (*e.g.,* deform, scatter *etc.*). We utilize the audio part of the dataset, and all 17 objects and 2 actions (hit and scratch) serve as class and subclass respectively. We do not use reaction because a bi-level hierarchy is more well aligned with the scope of this paper. Following the notations introduced in Section 2.1, an object creates a coarse class, and object and action together create a fine class. So, we generate 17 coarse and 34 fine classes. A random (80%,10%,10%) data split has been performed to produce disjoint train, validation, and test sets. Table 3 compares the test performances of a ResNet-18 trained from random initialization (referred as *Rand-init* in the table), and from our quadruplet based pre-trained multi-task model. For Top 1, we get around **10%** absolute improvement at both coarse and fine levels.

6. CONCLUSION AND FUTURE DIRECTIONS

The paper dealt with a bi-level hierarchical audio event detection task. We introduced hierarchy-aware loss functions that learn from the tree structured label ontology for achieving better classification performance at all levels and to produce more distinctive audio embeddings. A multi-task learning framework was built with cross entropy loss and the hierarchy-aware loss . Two different hierarchy-aware loss functions were employed. First, a modified triplet loss with a probabilistic multi-level batch balancing strategy. Second, quadruplet learning suitable for labels having bi-level tree structure. The classification and clustering experiments showed the efficacy of the employed method. The evaluation on the Greatest Hits dataset showed the model's ability to transfer to a different domain.

An obvious extension of the work would be to apply the employed methods in AED tasks having deeper label structures. Unsupervised learning of hierarchical audio events and their mixtures might also be an interesting problem to attack in the future due to the availability of large amounts unlabeled audio events data.

7. REFERENCES

[1] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] Jort F Gemmeke, Daniel PW 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[3] Shawn Hershey, Sourish Chaudhuri, Daniel P.W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "CNN architectures for large-scale audio classification," *ICASSP*, pp. 131–135, 2017.

[4] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, Fabio Antonacci, and Augusto 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.

[5] Pradeep K Atrey, Namunu C Maddage, and Mohan S Kankanhalli, "Audio based event detection for multimedia surveillance," in *Proceedings of ICASSP*, 2006, vol. 5.

[6] Yao Wang, Zhu Liu, and Jin-Cheng Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, no. 6, pp. 12–36, 2000.

[7] Wen-Huang Cheng, Wei-Ta Chu, and Ja-Ling Wu, "Semantic context detection based on hierarchical audio models," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*. ACM, 2003, pp. 109–115.

[8] Tian Hao, Guoliang Xing, and Gang Zhou, "isleep: unobtrusive sleep quality monitoring using smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2013, p. 4.

[9] Maria Niessen, Caroline Cance, and Danièle Dubois, "Categories for soundscape: toward a hybrid classification," in *Inter-Noise and Noise-Con Congress and Conference Proceedings*. Institute of Noise Control Engineering, 2010, vol. 2010, pp. 5816–5829.

[10] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[12] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] Jean-Julien Aucouturier, Boris Defreville, and Francois Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[14] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[15] Xiaodan Zhuang, Xi Zhou, Mark A Hasegawa-Johnson, and Thomas S Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[16] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[17] Oguzhan Gencoglu, Tuomas Virtanen, and Heikki Huttunen, "Recognition Of Acoustic Events Using Deep Neural Networks," *Proc. 22nd EUSIPCO*, pp. 506–510, 2014.

[18] Haomin Zhang, Ian McLoughlin, and Yan Song, "Robust sound event recognition using convolutional neural networks," in *ICASSP*. IEEE, 2015, pp. 559–563.

[19] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *ICASSP*, 2018, pp. 121–125.

[20] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, "Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition," *Interspeech 2016*.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Yong Xu, Qiang Huang, Wenwu Wang, and Mark D. Plumbley, "Hierarchical learning for DNN-based acoustic scene classification," *Detection and Classification of Acoustic Scenes and Events*, September 2016.

[23] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang, "Embedding label structures for fine-grained feature representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1114–1123.

[24] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*. IEEE, 2005, vol. 1, pp. 539–546.

[25] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al., "Deep face recognition.," in *BMVC*, 2015, vol. 1, p. 6.

[27] Pedregosa *et. al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] Julia Bell Hirschberg and Andrew Rosenberg, "V-measure: a conditional entropy-based external cluster evaluation," Proceedings of EMNLP, 2007.

[29] Lawrence Hubert and Phipps Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[30] Sinno Jialin Pan, Qiang Yang, et al., "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[31] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman, "Visually indicated sounds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2405–2413.