LARGE-SCALE WEAKLY SUPERVISED SOUND EVENT DETECTION (DCASE CHALLENGE 2017)

Shao-Yen Tseng[†*], *Juncheng Li*[‡]

[‡] Robert Bosch LLC, Research and Technology Center, USA
[†] University of Southern California, Department of Electrical Engineering, USA

ABSTRACT

State-of-the-art audio event detection (AED) systems fully rely on supervised-learning based on strongly labeled data. The dependence on strong labels severely limits the scalability of AED work. Large-scale manually annotated datasets are difficult and expensive to collect [1], whereas weakly labeled data could be much easier to acquire. In weakly labeled data, we only need to determine whether an event in the recording is present or absent. This not only makes manual labelling significantly easier but also makes automatically inferring labels from online multimedia or audio meta-information (titles, tags, etc) possible [2]. This work employs a subset of Google's AudioSet [3], which is a large number of weakly labeled YouTube video excerpts. The subset focuses on transportation and warning sounds and consists of 17 sound events divided into two categories: *Warning* and *Vehicle*.

We perform experiments on 3 sets of features, including standard Mel-frequency cepstral coefficients (MFCC) and log-Mel spectrograms and pre-trained embeddings extracted from a deep convolutional network. Our system employs multiple instance learning (MIL) [4] approaches to deal with weak labels by bagging them to positive or negative bags. We apply 4 models, Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Convolutional Deep Neural Network. Using the late-fusion approach, we improve the performance of the baseline audio tagging (Subtask A) F1 score 13.1% by 18.1%.

The embeddings extracted by the convolutional neural networks significantly boosts the performance of all the models.

Index Terms— weak labels, audio event detection (AED), multiple instance learning (MIL), late fusion

1. INTRODUCTION

Increasingly, machines in various settings are equipped with hearing capabilities. There will be a huge benefit brought by human-like audio event detection in domains such as self-driving cars, smart cities and related areas. Unlike human speech, environmental sounds are much more diverse and span a wide range of frequencies. Sound events that happen in these settings are usually sporadic and smeared with different noises. Existing works on audio event detection (AED) [5][6] rely heavily on the supervised learning paradigm which requires strongly labeled datasets. Given the huge difficulty and heavy resource requirement to collect such datasets, there are only few that are available to the public and they are often of very limited size [1][7].

This motivates the community to explore weakly labeled dataset. Weak labels only need to determine whether an event in the recording is present or absent. This greatly reduces the resource needed for collecting such dataset. In this work, we use a subset of Google's AudioSet [3]. We extract several feature representations based on signal processing methods and neural networks: Mel-frequency cepstral coefficients(MFCC), log Mel-Spectrum and CNN based embeddings. Here, all the training instances are loaded in the multiple instance learning (MIL) framework. We apply state-of-the-art deep learning models to perform classification.

2. DATASET

2.1. Google Audio Set

The DCASE 2017 Challenge Task 4 dataset is a subset of Google's "AudioSet: An Ontology And Human-Labeled Dataset For Audio Events".

To collect AudioSet Google worked with human annotators who listen, analyze and verify the sounds they hear within YouTube clips. To facilitate faster accumulation of examples for all classes, Google relies on available YouTube meta-information and content-based search to nominate candidate video segments that are likely to contain the target sound. A detailed description of the data annotation procedure is available in [3].

2.2. Unbalanced Labels

The challenge subset contains 17 sound event classes divided into two categories : *Warning* and *Vehicle* sounds. Each sample is a 10 second audio excerpt from a YouTube video and

[*]This work has been done during an internship at Robert Bosch LLC

is assigned one or more class label(s). The number of samples per class is highly imbalanced with an imbalance ratio of 1:94 for minority to majority class samples. In total the training dataset contains 51,172 samples which is approximately 142 hours of audio. The class label names and number of samples in each class are shown in Table 1.

Class Label Name	Number of Samples
Warning Sounds	
Train horn	441
Truck horn	407
Car alarm	273
Reversing beeps	337
Ambulance siren	624
Police siren	2,399
Fire truck siren	2,399
Civil defense siren	1,506
Vehicle Sounds	
Bicycle	2,020
Skateboard	1,617
Car	25,744
Car passing	3,724
Bus	3,745
Truck	7,090
Motorcycle	3,291
Train	2,301

Table 1. Class labels and number of samples per class.

Table 1 shows that the number of car and truck sample are way more than other rare classes such as reverse beeps and car alarm. Meanwhile, the number of every classes in the provided validation set is balanced. This invariably drive the training of the model biased towards those majority classes. In order to address this issue, we take 2 approaches: (1) We assigned a weight penalty (number of total training samples / number of of each class) on the predicted probability of each class. (2) We force to balance the number of training samples of each class to 3010. As for the classes which already have more than 3010 sample, we randomly select a subset of 3010 samples; As for the classes which are short of 3010 samples, we augment the number of samples by adjusting the tempo and speed of the recordings.

3. FEATURE REPRESENTATIONS

We perform experiments on 3 types of features—standard Mel-frequency cepstral coefficients (MFCC), log-Mel spectrograms and pre-trained embeddings extracted from a deep convolutional network.

3.1. MFCC Features

We take 23 Mel-frequency (excluding the 0th) cepstral coefficients over window length 20 ms. We augment the feature with first and second order differences using 60 ms window, resulting in a 61-dimension vector.

3.2. Log-Mel Spectrogram Images

Spectrogram images were generated over different window lengths based on individual models. We use a frame size of 25ms with 10ms shift in the short-time Fourier transform and integrate the power spectrogram into 64 mel-spaced frequency bins. A log-transform is then applied to the power spectrogram to generate the spectrogram images.

3.3. Audio Embeddings

Similar to [8] we generate audio embeddings by training a CNN to give frame-wise predictions of the clip label. The outputs from the penultimate layer of the CNN are then extracted and used as audio embeddings. We use the first three convolutional groups from VGG-16 [9] and two fully-connected hidden layers of size 1024 and 512 for the embedding CNN. Batch normalization is added after each convolutional layer. The final embedding size is 512.

Since frame-wise training of the instances results in the data being badly labeled the final model selection of the embedding CNN is crucial in generating meaningful embeddings. We use the maximum of frame-wise predictions as the predicted clip label and evaluate the embedding CNN at the clip-level using held-out validation data.

4. MULTIPLE INSTANCE LEARNING

4.1. MIL Framework

The task of detecting sound events using weakly labeled training data can be described as a multiple instance learning problem. In MIL, labels are assigned to *bags* of *instances* without explicitly specifying the relevance of the label to individual *instances*. All that is known is one or more *instances* within the *bag* contribute to the *bag* label. Applying this framework to our task, we view each YouTube audio clip as a *bag* of *instances* $B_i = \{x_{ij}\}$ where each *instance* x_{ij} is an audio segment j of shorter duration. We then assign all the labels of the clip to the bag so that each bag has the label $Y_i = \{y_{in}\}$ where y_{in} indicates the presence of sound event n. The goal of the MIL problem is then to classify labels of unseen *bags* given only the *bag* and label pairs (B_i, Y_i) as training data.

In our implementation we generate the *instances* by segmenting the original YouTube clip into non-overlapping 1-second segments. We calculate the log-mel spectrogram of each segment as well as its first delta and use these as input

features to our proposed MIL framework using neural networks.

4.2. MIL using Neural Networks

Since each instance is a 64 by 100 log-mel spectrogram image we employ convolutional neural networks to handle such large dimensional input. We use the first two conv groups from VGG-16 [9] as the feature extraction layers and add two fully-connected layers of size 3072 and 1024. As this task is a multi-label problem we use a sigmoid layer as the output. To obtain a prediction for the entire bag we take the maximum over all instances for each class using a max pooling layer. That is

$$Y_i' = \{\max_j f(x_{ij})_n\}$$

where $f(x_{ij})_n$ is the CNN output for class n on instance j.

During training we use cross entropy as the objective function and share all weights of the CNN for each instance. Figure 1 shows the architecture of the proposed MIL framework using CNN.

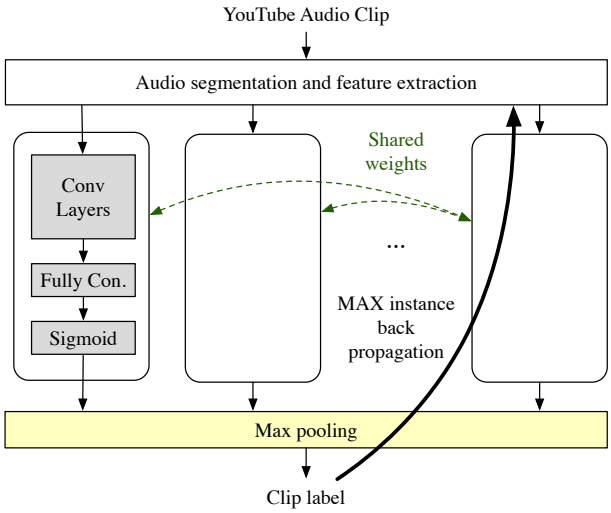

Fig. 1. Architecture of MIL using CNN.

Our model infers that for a certain class, the highest scoring instances are most important and contribute directly to the corresponding bag label. The training of the neural network to identify these important instances is an expectation maximization (EM) approach. However there are two possible issues which may result from this model. The first is that as with most EM methods, system performance highly depends on the initialization point. With a bad initialization point the model chooses the wrong instance as being indicative of the class label and optimizes on irrelevant input. These types of errors would be hard to recover from if there is high variation for each individual sound event. A second issue is that by using a max pooling layer over all instances back-propagation will only propagate through the maximum scoring instance. This may result in some instances being ignored for most of the training. While this focus on relevant instances only is the central idea of MIL it greatly reduces robustness to noise which occurs intermittently in the audio. Therefore we propose the use of audio embeddings to handle the above issues. By using audio embeddings as features we postulate that sound events as well as noise conditions can be better represented which can improve the performance of the MIL framework.

The final MIL system is similar in architecture to the MIL-CNN but uses audio embeddings as features for each instance. In addition, the convolutional layers are replaced with fully-connected layers as we no longer deal with images. The best system has four hidden layers using relu activation function and layer sizes 512, 512, 256 and 128.

5. EXPERIMENTAL RESULTS

The best F1-score achieved by the MIL system using CNN (MIL-CNN) on a two-fold cross-validation setup was 22.4%. However by using audio embeddings as features and only a DNN as classifier (MIL-DNN) the performance improved to 28.7% which is 15.6 percentage points above the official baseline. Using an ensemble of models with different hyper-parameters the F1-score further improved to 31.1%. For the ensemble we used per-class weighted voting based on validation accuracy. Table 2 shows the performance of the different models.

Model	Precision	Recall	F1-Score
Official Baseline	12.2%	14.1%	13.1%
MIL-CNN	19.6%	26.1%	22.4%
MIL-DNN w/ Embed.	22.0%	41.3%	28.7%
Ensemble	28.6%	46.0%	**35.3%**

Table 2. F1-Score on Testing Set for different models.

6. CONCLUSION

In this work we explored methods of training models for large-scale sound event detection using weakly-supervised data and proposed a multiple instance learning framework using deep neural networks. We showed that by using audio embeddings pre-trained on all the data we can achieve higher performance with a simple DNN model in the MIL framework. The audio embeddings were extracted from a CNN trained to give frame-wise predictions for the weakly labeled data. While the performance of this CNN is poor, the embeddings generated by this model can be used as features to drastically improve performance in a MIL framework. We

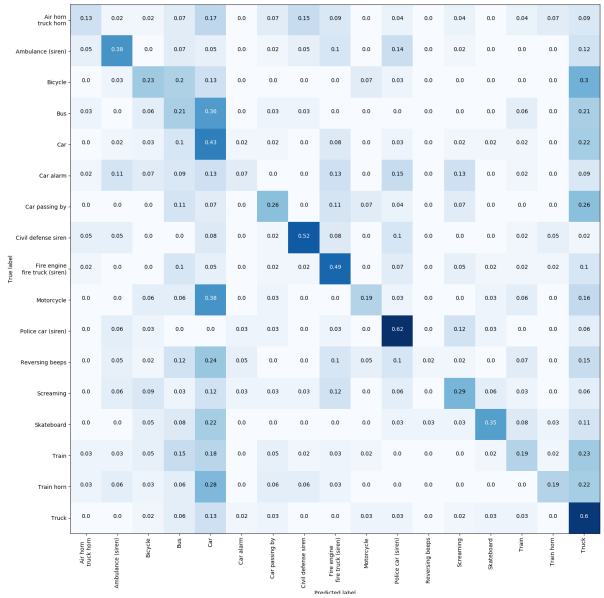

Fig. 2. Confusion matrix for the proposed MIL system.

postulate that audio embeddings map data into an acoustically meaningful high-dimensional space which is more indicative of the sound events. However, we also observed that selection of the embedding model is pivotal in the final system performance. We expect that with better model selection and embedding generation the MIL framework can obtain much higher performance.

7. REFERENCES

[1] 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.

[2] Bhiksha Raj and Anurag Kumar, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 3475–3482.

[3] 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[4] Oded Maron and Tomás Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[5] Jose Portelo, Miguel Bugalho, Isabel Trancoso, Joao Neto, Alberto Abad, and Antonio Serralheiro, "Non-speech audio event detection," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1973–1976.

[6] Onur Dikmen and Annamaria Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

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

[8] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson, "CNN Architectures for Large-Scale Audio Classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," 09 2014.