Interspeech 2018
2-6 September 2018, Hyderabad

Multiple Instance Deep Learning for Weakly Supervised Small-Footprint Audio Event Detection

Shao-Yen Tseng[†], *Juncheng Li*[‡γ], *Yun Wang*[γ], *Florian Metze*[γ], *Joseph Szurley*[‡], *Samarjit Das*[‡]

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

shaoyent@usc.edu, {junchenl,yunwang,fmetze}@cs.cmu.edu,
{joseph.szurley,samarjit.das}@us.bosch.com

Abstract

State-of-the-art audio event detection (AED) systems rely on supervised learning using strongly labeled data. However, this dependence severely limits scalability to large-scale datasets where fine resolution annotations are too expensive to obtain. In this paper, we propose a small-footprint multiple instance learning (MIL) framework for multi-class AED using weakly annotated labels. The proposed MIL framework uses audio embeddings extracted from a pre-trained convolutional neural network as input features. We show that by using audio embeddings the MIL framework can be implemented using a simple DNN with performance comparable to recurrent neural networks.

We evaluate our approach by training an audio tagging system using a subset of AudioSet, which is a large collection of weakly labeled YouTube video excerpts. Combined with a late-fusion approach, we improve the F1 score of a baseline audio tagging system by 17%. We show that audio embeddings extracted by the convolutional neural networks significantly boost the performance of all MIL models. This framework reduces the model complexity of the AED system and is suitable for applications where computational resources are limited.

Index Terms: audio event detection, weakly-supervised learning, multiple instance learning

1. Introduction

Increasingly, devices in various settings are equipped with auditory perception capabilities. The inclusion of acoustic signals as an extra modality brings robustness to a system and offers improved performance in many tasks. This benefit can be attributed to the omnidirectional nature of acoustic signals which provides a valuable cue for detecting events in various applications. For example, [1] analyzed audio signals to monitor the conditions of industrial tools, and in [2] a water leakage detection system using sound recordings of water pipes was proposed. Such systems are able to run in real-time and at a lower cost as capturing audio is much less expensive than distributing specialized physical sensors throughout the environment. In addition, acoustic signals can provide informational cues that are hard to or cannot be captured by other modalities. A common example is the detection of alarms or sirens in a driving scenario with smart cars. Very often, sources of these warning sounds may be visually occluded and these events are only detectable using auditory perception [3][4]. Many of these applications also have a requirement of real-time operation using low computational resources. This is a major challenge since, unlike human speech, environmental sounds are much more diverse and span a wider range of frequencies. Audio events that occur in these settings are also usually sporadic and corrupted by noise.

Previous works on AED have relied on training models using a supervised learning paradigm which requires strongly labeled data [5][6]. However, given the difficulty and high resource requirement of annotating large datasets there are only a few datasets that are publicly available and are often of limited size [7][8]. Motivated by this, many recent works have explored the use of weakly labeled data for training AED systems. One successful approach is to transform the audio into time-frequency representations and apply a convolutional recurrent neural network to tag or classify the entire clip [9][10]. These methods, however, are unsuitable for real-time applications as the recurrent and subsequent pooling layers require the full clip to be parsed before a decision can be made. In addition, the complexity and computation time of these models are quite high. Another approach for learning with weak labels is to treat segments in an audio clip as a *bag of instances* and apply multiple instance learning [11]. The MIL model assumes independent labels for each instance and accounts for the uncertainty of the weak labels by assigning a positive bag label only if there is *at least one* positive instance. Evidently, this paradigm is more suitable for portable applications as the classifier can be applied to individual instances which is ideal for real-time operation.

In this work, we propose to enhance the framework for multi-class MIL using convolutional audio embeddings. Different from prior works, our proposed architecture addresses the issue of building low complexity models with a small footprint for real-time applications. We propose the use of audio embeddings as input features and show that by using pre-trained embeddings the MIL model can be implemented with a simple DNN architecture. The use of audio embeddings also significantly improves AED accuracy compared to random initialization. Our proposed architecture removes the need for complex CNN structures or recurrent layers which drastically reduces model complexity and is suitable for portable applications with low computational resource and real-time requirements.

2. Multiple Instance Learning

2.1. MIL Framework

The task of detecting audio events using weakly labeled training data can be formulated as a multiple instance learning problem [12]. 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 audio clip i as a *bag* of *instances* $B_i = \{x_{ij}\}$ where each *instance* x_{ij} is an audio segment j of shorter dura-

10.21437/Interspeech.2018-1120

tion. We then assign all the labels of the clip to the bag so that each bag has the label $Y_i = \{y_{in}\}$ where $y_{in} = 1$ indicates the presence of audio 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 this we work we implement the MIL framework using neural networks.

2.2. MIL using Neural Networks

In our implementation we generate instances by segmenting the audio clip into non-overlapping 1-second segments and taking the time-frequency representations. The segment size was chosen as a balance between number of total instances and coverage of audio events. 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 spectrogram. We also use the first delta as an additional input channel.

Since the spectrogram can be viewed as an image we employ convolutional layers for feature extraction. We reference CNN architectures proven to have good performance in the field of computer vision. Specifically, we use the first three conv groups from VGG-16 [13] and add two fully-connected layers of size 3072 and 1024. Batch normalization is added after each convolutional layer. The ReLU activation function is used in all layers. As our goal is a multi-label system we apply a sigmoid activation function and view the outputs as independent posterior probability estimates for each class. We use a reduced version of the full VGG model because (1) we are exploring compact models for portable applications and (2) the subset dataset does not contain enough samples to train large models without overfitting.

To obtain a prediction for the entire bag we adopt a naïve approach and assign the label of the maximum scoring instance to the bag. The motivation behind this is in part due to the fact that since instances in a continuous audio clip are not i.i.d. many MIL algorithms are not applicable [11]. However this approach is still beneficial as it allows us to train an instance classifier which can be applied in a real-time scenario.

Using this approach, the final bag label is obtained using a max pooling layer. That is

$$\widehat{Y}_i = \{\hat{y}_{in}\} = \{\max_j f_n(x_{ij})\}$$

where $f_n(x_{ij})$ is the predicted probability of class n on instance x_{ij}.

The multi-class MIL loss can then be defined as simply the cross entropy loss summed over all the classes, which is

$$J_i = -\sum_n \left(y_{in} \log \hat{y}_{in} + (1 - y_{in}) \log \left(1 - \hat{y}_{in}\right)\right)$$

In order to address class imbalance we apply a weight to the MIL loss proportional to the inverse frequency of each class. During back-propagation only the gradient from the maximally scoring instance is calculated and used for updating weights. An interesting fact is that as each class has its own max pooling layer, errors originate from different instances between classes. Figure 1 shows the architecture of the proposed MIL framework using CNN.

2.3. MIL using Audio Embeddings

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 similar to 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 audio 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. We propose that the use of pre-trained audio embeddings can alleviate the above issues. By using audio embeddings as features we postulate that audio events as well as noise conditions can be better represented which can improve the performance of the MIL framework.

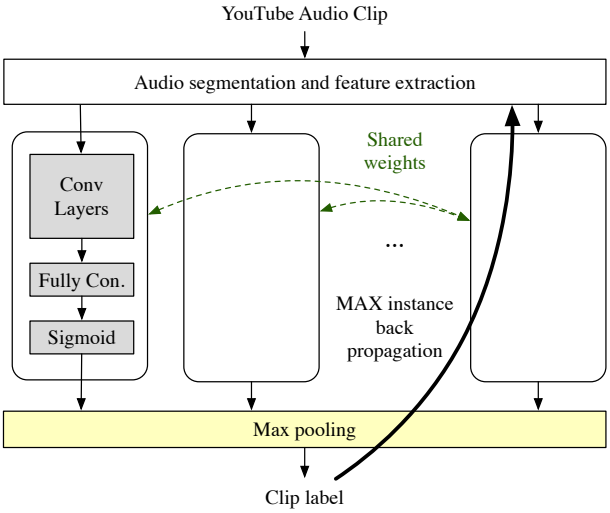

Figure 1: *Architecture of MIL using CNN. Back-propagation is performed along the MAX instance for each class.*

Similar to [14] we generate audio embeddings by training a CNN to give frame-wise predictions of the clip label. The input features are 128-bin log-mel spectrograms computed over 1-second segments of audio by short-time Fourier transform. We use the clip label as targets for all 1-second segments in the audio clip. The outputs from the penultimate layer of the CNN are then extracted and used as input to the MIL framework. We use the same CNN structure described in the previous section but add an additional fully-connected layer of size 512 to generate the final audio embedding. Since frame-wise training of the instances results in badly labeled data, 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 select the CNN model with the best performance at the clip-level using held-out validation data.

The final MIL system is similar in architecture to the MIL-CNN but uses audio embeddings as features for each instance. The convolutional layers are replaced with fully-connected layers as we no longer deal with images. The best performing system has four hidden layers using a ReLU activation function with layer sizes of 512, 512, 256 and 128. The final architecture of the MIL framework is shown in Figure 2.

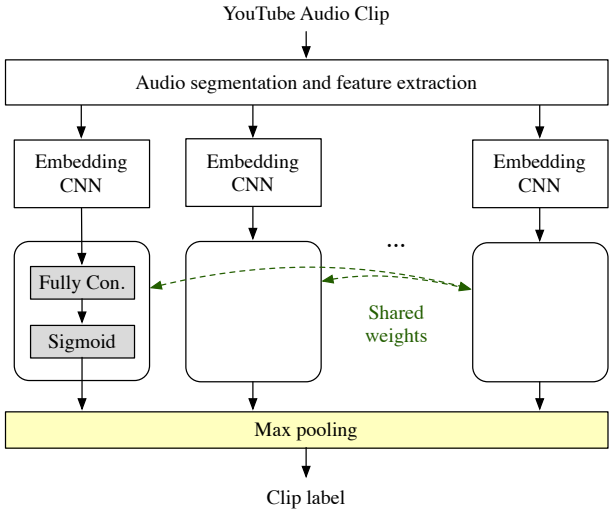

Figure 2: *Architecture of MIL using audio embeddings.*

3. Dataset & Challenges

3.1. Dataset

We evaluated our models using a subset of Google's AudioSet [15]. AudioSet is an extensive collection of 10-second YouTube clips annotated over a large number of audio events. This dataset contains 632 audio event classes and over 2 million sound clips, however as a proof of concept we refer to a subset released by the DCASE 2017 challenge [16].

The challenge subset contains 17 audio event classes divided into two categories : *Warning* and *Vehicle* sounds. These audio events are highly focused on transportation scenarios and is primed towards evaluating AED systems for self-driving cars, smart cities and related areas. The subset contains 51,172 samples which is around 142 hours of audio. The class names and number of samples per class are shown in Table 1.

Class Name	Samp #	Class Name	Samp #
Warning Sounds		*Vehicle Sounds*	
Car alarm	273	Skateboard	1,617
Reversing beeps	337	Bicycle	2,020
Air/Truck horn	407	Train	2,301
Train horn	441	Motorcycle	3,291
Ambulance siren	624	Car passing by	3,724
Screaming	744	Bus	3,745
Civil defense siren	1,506	Truck	7,090
Police siren	2,399	Car	25,744
Fire engine siren	2,399		

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

3.2. Challenges of the Dataset

The main challenge of the dataset is the noisiness of YouTube data. As clips are user submitted and mostly recorded using consumer devices in real life environments, audio events are often far-field and corrupted with a variety of noise, including human speech, music, wind noise, etc. Another challenge is the variability of audio events. Even within class, the characteristic of an audio event can vary drastically. An example of this is the use of different types of sirens by different regions which would make it hard to differentiate between *ambulance* and *fire truck sirens*. In short, it is possible that each label type encompasses all possible global variations of that category.

Finally, the number of samples per class is also highly imbalanced in the subset dataset. The imbalance ratio of the least occurring to most occurring class is 1:94. While this issue can be alleviated through machine learning techniques, the inherent shortage of information in minority classes may result in bad generalization of those classes.

4. Experimental Setup and Results

In all experiments we used cross entropy as the loss function and the Adam optimizer [17] to perform weight updates. To handle class imbalance the loss function was weighted inversely proportional to the number of samples for each class. For model selection of the embedding CNN we adopted a clip-level validation scheme. The posterior class probabilities were averaged over all instances in a clip and the model with the best clip tagging accuracy was selected to generate audio embeddings.

We compared our MIL framework to an MLP baseline from the DCASE challenge [16]. The best F1-score achieved by our MIL system using a CNN architecture on a two-fold cross-validation setup was 22.4%. Using audio embeddings as features and only a DNN as classifier the performance improved to 31.4% which is 20.5% absolute improvement from the DCASE baseline. We compared to an MIL framework where the DNN classifier is replaced with a 3-layer Bi-LSTM RNN and found that results were comparable to DNNs. We also applied late-fusion to models with different hyper-parameters using a weighted majority voting scheme which improved the F1-score further to 35.3%. The weights of the voting scheme were based on model validation accuracy. Finally, we show that the performance of our MIL framework improves to 46.5% using embeddings from AudioSet. These embeddings are part of AudioSet and trained with a CNN architecture from [14] using the YouTube-8M dataset [18]. Table 2 shows the performance and parameter number of the different models.

The confusion matrix for the proposed MIL system is shown in Figure 3. Although there is high confusability in the *Car* class, which may be due to the imbalance of labels, the system is still able to distinguish between classes with relative accuracy.

Model	Prec.	Rec.	F1	Param #
Development set				
Baseline [16]	7.9	17.6	10.9	13K
MIL-CNN	19.6	26.1	22.4	29M
MIL-RNN-Embed	23.7	38.1	29.2	6.5M
MIL-DNN-Embed	25.4	41.3	31.4	**700K**
Ensemble	28.6	46.0	35.3	-
MIL-DNN-AudioSet	41.9	52.2	46.5	**700K**
Evaluation set				
Baseline [16]	15.0	23.1	18.2	13K
Ensemble	31.6	39.7	35.2	-

Table 2: *Comparisons of precision, recall, F1-score (%), and number of parameters for the various models.*

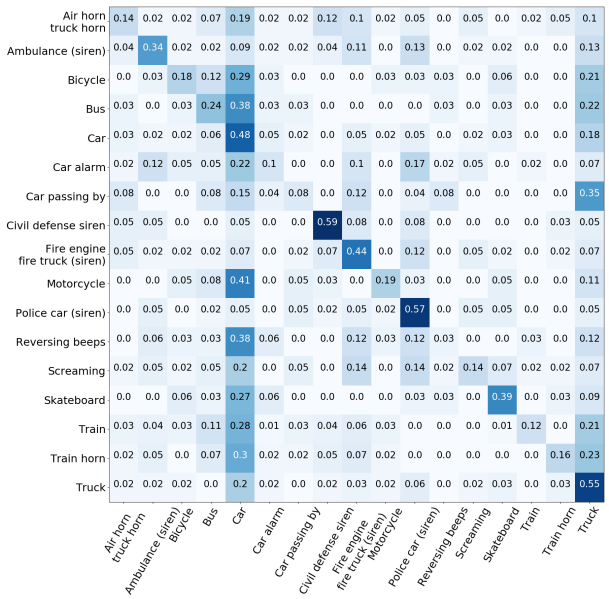

Figure 3: *Confusion matrix for the proposed MIL system.*

5. Discussion

While our framework is not state-of-the-art [19], which achieves a single-model F1-score of 54.2%, a more fair comparison would be with models without recurrent layers, such as [20], which has an F1-score of 49.0%. Even so, a direct comparison is of limited value as our proposed method mainly aims to address two major issues in deploying AED to real-life scenarios: model complexity and real-time operation. Our proposed method reduces model complexity by removing the need of recurrent layers and is suitable for applications where computational resources are limited. Under similar performance conditions the MIL system using DNN reduces the number of parameters by a factor of almost 10 compared to a 3-layer Bi-LSTM RNN. In terms of evaluation runtime, the DNN model is also up to 5 times faster than RNNs. The DNN model is able to handle 2,500 samples per second compared to 500 samples with RNN using an NVIDIA GTX-1080 GPU.

In addition, by using independent instance classifiers our system is able to run in real-time and give running predictions of audio events. This property is crucial when applying AED in smart cars as events such as sirens and horns have to be detected as soon as they occur. With recurrent networks or even CNNs requiring full length inputs this mode of operation would not be possible.

Finally, as shown by the gain in performance through the use of AudioSet embeddings, the MIL system can easily be improved through transfer learning of other sound events. An interesting observation from our experiments is that joint optimization of the pre-trained embedding CNN with the MIL loss did not improve performance much above random initialization. This shows that audio embeddings already contain rich acoustic information and can be trained in a task-independent manner. The separation of embedding and classifier training means that we can take advantage of additional labels in large-scale weakly-supervised data and learn embeddings independently. However, we also observed that selection of the embedding model is pivotal in the final system performance and not all embeddings are as useful.

6. Conclusions

In this work we proposed a small-footprint multiple instance learning framework using deep neural networks for audio event detection which can be trained using large-scale weakly-supervised data. We showed that by using pre-trained audio embeddings we can achieve good performance with a simple DNN model in an MIL framework. 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 the performance of an MIL framework. Further improvements were achieved by using embeddings from AudioSet which were trained with more data and additional labels. We postulate that audio embeddings map data into an acoustically meaningful high-dimensional space which is more indicative of audio events. Using these embeddings we can achieve a good trade-off between model size and performance.

In future work, we hope to apply our model to the entire AudioSet for a truly large-scale weakly-supervised MIL framework. With the introduction of additional data as well as class labels we expect the audio embeddings to contain richer representations which can further improve performance of AED in smart cars.

7. References

[1] I. Ubhayaratne, M. Pereira, Y. Xiang, and B. Rolfe, "Audio signal analysis for tool wear monitoring in sheet metal stamping," *Mechanical Syst. and Signal Process.*, vol. 85, 2017.

[2] S. Seyoum, L. Alfonso, S. J. van Andel, W. Koole, A. Groenewegen, and N. van de Giesen, "A Shazam-like household water leakage detection method," in *Proc. of the XVIII Int. Conf. on Water Distribution Syst. (WDSA)*, vol. 185, no. Supplement C, 2016.

[3] F. Meucci, L. Pierucci, E. D. Re, L. Lastrucci, and P. Desii, "A real-time siren detector to improve safety of guide in traffic environment," in *Proc. of the European Signal Process. Conf. (EUSIPCO)*, 2008.

[4] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2013.

[5] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2009.

[6] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Proc. of the IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA)*, 2013.

[7] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. of the ACM Int. Conf. on Multimedia and Expo (ICME)*, 2014.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. of the European Signal Process. Conf. (EUSIPCO)*, 2016.

[9] E. Çakr, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 25, no. 6, 2017.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2017.

[11] B. Babenko, "Multiple instance learning: algorithms and applications," *PubMed/NCBI Article*, 2008.

[12] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proc. of the ACM Conf. on Multimedia (MM)*, 2016. [Online]. Available: http://doi.acm.org/10.1145/2964284.2964310

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[15] 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 *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2017.

[16] R. Badlani, A. Shah, and B. Elizalde, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," DCASE2017 Challenge, Tech. Rep., 2017.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[18] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016. [Online]. Available: http://arxiv.org/abs/1609.08675

[19] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *CoRR*, vol. abs/1710.00343, 2017. [Online]. Available: http://arxiv.org/abs/1710.00343

[20] S.-Y. Chou, J.-S. Jang, and Y.-H. Yang, "FrameCNN: A weakly-supervised learning framework for frame-wise acoustic event detection and classification," DCASE2017 Challenge, Tech. Rep., September 2017.