NON-NEGATIVE MATRIX FACTORIZATION-CONVOLUTION NEURAL NETWORK (NMF-CNN) FOR SOUND EVENT DETECTION

Technical Report

Teck Kai Chan,[1,2], Cheng Siong Chin[1]*

[1]Newcastle University Singapore
Faculty of Science, Agriculture, and Engineering
Singapore 599493
{t.k.chan2, cheng.chin}@newcastle.ac.uk

Ye Li[2]

[2]Visenti Pte Ltd (A Brand of Xylem)
3A International Business Park
Singapore 609935
ye.li@xyleminc.com

ABSTRACT

The main scientific question of this year DCASE challenge, Task 4 - Sound Event Detection in Domestic Environments, is to investigate the types of data (strongly labeled synthetic data, weakly labeled data, unlabeled in domain data) required to achieve the best performing system. In this paper, we proposed a deep learning model that integrates Convolution Neural Network (CNN) with Non-Negative Matrix Factorization (NMF). The best performing model can achieve a higher event based F1-score of 30.39% as compared to the baseline system that achieved an F1-score of 23.7% on the validation dataset. Based on the results, even though synthetic data is strongly labeled, it cannot be used as a sole source of training data and resulted in the worst performance. Although, using a combination of weakly and strongly labeled data can achieve the highest F1-score, but the increment was not significant and may not be worthwhile to include synthetic data into the training set. Results have also suggested that the quality of labeling unlabeled in domain data is essential and can have an adverse effect on the accuracy rather than improving the model performance if labeling was not done accurately.

Index Terms— Non-negative matrix, convolutional neural network, DCASE 2019

1. INTRODUCTION

The primary objective of a Sound Event Detection (SED) system is to identify the type of sound source present in an audio clip or recording and returns the onset and offset of the identified source. Such a system has great potential in several domains such as activity monitoring, environmental context understanding, and multimedia event detection [1], [2]. However, there are several challenges associated with SED in real life scenarios.

Firstly, in real-life scenarios, different sound event can occur simultaneously [2]. Secondly, the presence of background noise could complicate the identification of sound event within a particular time frame [3]. This problem is further aggravated when the noise is the prominent sound source resulting in a low Signal to Noise Ratio (SNR). Thirdly, each event class is made up of different sound sources, e.g. a dog bark sound event can be produced from several breeds of dogs with different acoustic characteristics [1]. Finally, to achieve the best results, SED detection algorithm may require strongly labeled data where the occurrence of each event with its onset and offset are known with certainty during the model development phase. While such data are useful, collecting them is often time-consuming and sizes of such dataset are often limited to minutes or a few hours [3], [4]. In certain scenarios such as an approaching vehicle, the onset and offset time is ambiguous due to the fade in and fade out effect [5] and is subjective to the person labeling the event.

On the other hand, there exist a substantial amount of data known as the weakly labeled data where only the occurrence of an event are known without any offset or onset annotations. While it seems like the core information is missing, previous implementations proposed in the annual Detection and Classification of Acoustic Scenes and Events (DCASE) challenge that utilized only weakly labeled data had achieved a certain level of success [6]-[8]. Although a large number of different SED system were proposed in the past, a majority of them were mainly based on Gaussian Mixture Model (GMM) [9], Hidden Markov Model (HMM) [10] or the use of dictionaries constructed using NMF [11-13]. However, due to the rising success of deep learning in other domains [14-17], deep learning for SED development is now a norm and has been shown to perform slightly better than established methods [1]. Riding on the success of deep learning, this paper proposed a deep learning model that integrates NMF and CNN which can provide an approximate

* This work was supported by the Economic Development Board-Industrial Postgraduate Programme (EDB-IPP) of Singapore under Grant BH180750 with Visenti Pte. Ltd.

strong label to the weakly labeled data. Results have shown that proposed system achieved a higher event based F1-score of 30.39% as compared to the baseline system that achieved an F1-score of 23.7% on the validation dataset.

2. RELATED WORKS

In the recent years, SED development has been overwhelmed with the use of deep learning algorithms particularly the use of CNN or Convolutional Recurrent Neural Network (CRNN). This phenomenon was also reflected in the 2018 DCASE challenge, where almost all participants for Task 4 (Large-scale weakly labeled semi-supervised sound event detection in domestic environments) proposed the use of CRNN. As discussed in [1], CNN has the benefit of learning filters that are shifted in both time and frequency while Recurrent Neural Network (RNN) has a benefit of integrating information from the earlier time windows. Thus, a combined architecture has the potential to benefit from two different approaches that suggest its popularity.

The CRNN architecture proposed by Cakir et al. [1] first extracted features through multiple convolutional layers (with small filters spanning both time and frequency) and pooling in the frequency domain. The features were then fed to recurrent layers, whose features were used to obtain event activity probabilities through a feedforward fully connected layer. Evaluation over four different datasets had also shown that such a method has a better performance as compared to CNN, RNN and other established SED system. However, such a system would require a large amount of annotated data for training.

Lu [8] proposed the use of Mean Teacher Convolution System that won the DCASE Task 4 challenge with an F1 score of 32.4%. In their system, context gating was used to emphasize the important parts of audio features in frames axis. Mean-Teacher semi-supervised method was then applied to exploit the availability of unlabeled data to average the model weights over training steps. Although, this system won the 2018 challenge, there is still a large room for improvement.

3. SYSTEM OVERVIEW

3.1. Dataset Description

In this year DCASE challenge, the types of dataset available for training can be summarized in Table 1.

3.2. Audio Processing

In this system, training inputs are mel-frequency scaled. This is because they can provide a reasonably good representation of signal's spectral properties. At the same time, they also provide reasonably high inter-class variability to allow class discrimination by many different machine learning approaches [18].

In this paper, audio clips were first resampled to 32 kHz that were suggested to contain the most energies [19]. Moreover, segments containing higher frequency may not be useful for event detection in daily life [8]. A short-time fast Fourier transform with a Hanning window size of 1024 samples and a hop size of 500 samples was used to tabulate the spectrogram. After that, a mel filter bank of 64 and bandpass filter of 50 Hz to 14 kHz was applied to obtain the mel spectrogram to be used as input to the training model. Finally, a logarithm operation was applied to obtain the log mel spectrogram.

Class	Synthetic	Weakly Labeled	Unlabeled
Speech	2132	550	
Dog	516	214	
Cat	547	173	
Alarm/Bell Ringing	755	205	
Dishes	814	184	
Frying	134	171	
Blender	540	134	
Running Water	157	343	
Vacuum Cleaner	204	167	
Electric Shaver / Toothbrush	230	103	
	Number of Events		**Number of Clips**
	6032	2244	14412

Table 1. Given dataset for Task 4

3.3. Non-negative Matrix Factorization

The NMF popularized by Lee and Seung [20] is an effective method to decompose a non-negative matrix, $M \in \mathbb{R}^{\geq 0, L \times N}$, into two non-negative matrices, $W \in \mathbb{R}^{\geq 0, L \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$. Where R is the number of components. Therefore, it can be represented as

$$M \approx WH \quad (1)$$

Where W can be interpreted as the dictionary matrix and H can be interpreted as the activation matrix. These two matrices can be randomly initialized and updated through the multiplicative rule given as [20]

$$W \leftarrow W \otimes \frac{W^T \frac{M}{WH}}{W^T 1} \quad (2)$$

$$H \leftarrow H \otimes \frac{\frac{M}{WH} H^T}{1H^T} \quad (3)$$

W is commonly extracted on isolated events to form a dictionary and SED is performed by applying a threshold on the activation matrix obtained from the decomposition of the test data [12]. Since NMF only works on non-negative matrix, it was applied on the mel spectrogram prior to the logarithm operation. Thus, M represent the mel spectrogram with L as the number of mel bins and N as the number of frames. In this paper, instead of consolidating W to form the dictionary. We find the H to indicate which frames of each audio clip are activated (above a pre-defined threshold) to label the weakly labelled data so that the weakly labelled data becomes an approximated strongly labelled data.

3.4. Convolutional Neural Network

The CNN used in this system is modified based on the one proposed in [19]. Kong et al. [19] proposed four different CNN with a different number of layers and- pooling operators and found that the nine layers CNN with max pooling operator achieved the best performance. In this paper, we are interested in finding out whether with the inclusion of NMF, will a shallower CNN produce a comparable or even a better result.

Proposed	Kong [19]
Input : log-mel spectrogram	
$\begin{pmatrix} 5\times5@64 \\ BN, ReLU \end{pmatrix}$	$\begin{pmatrix} 3\times3@64 \\ BN, ReLU \end{pmatrix}\times2$
2 x 2 Max Pooling	
$\begin{pmatrix} 5\times5@128 \\ BN, ReLU \end{pmatrix}$	$\begin{pmatrix} 3\times3@128 \\ BN, ReLU \end{pmatrix}\times2$
2 x 2 Max Pooling	
$\begin{pmatrix} 5\times5@256 \\ BN, ReLU \end{pmatrix}$	$\begin{pmatrix} 3\times3@256 \\ BN, ReLU \end{pmatrix}\times2$
2 x 2 Max Pooling	
$\begin{pmatrix} 5\times5@512 \\ BN, ReLU \end{pmatrix}$	$\begin{pmatrix} 3\times3@512 \\ BN, ReLU \end{pmatrix}\times2$

Table 2. CNN architectures

In this paper, a 5 layers CNN with max pooling operator is proposed. In this architecture, it consists of 4 convolutional layers of kernel size 5 x 5 with a padding size of 2 x 2 and strides 1 x 1. This architecture is almost similar to Kong et al. [19] except for the kernel size and the number of layers. For both architectures, Binary Cross Entropy is adopted as the loss function which is similar to the loss function adopted in [19] given as

$$l_{BCE(p,y)} = \sum_{k=1}^{K} [y_k \ln(p_k) + (1-y_k)\ln(1-p_k)] \quad (4)$$

3.5. System Flow

In this year DCASE challenge, Task 4 - Sound Event Detection In Domestic Environments, is specifically organized to investigate the types of data (strongly labeled synthetic data, weakly labeled data, unlabeled in domain data) required to achieve the best performing system. Therefore, the flow of the proposed system depends on the types of data used. While strongly labeled and weakly labeled data can be used readily, unlabeled data require a model to be trained in advance so that its content can be tagged and be used as training data. The flow of our proposed system can be summarized in Fig. 1.

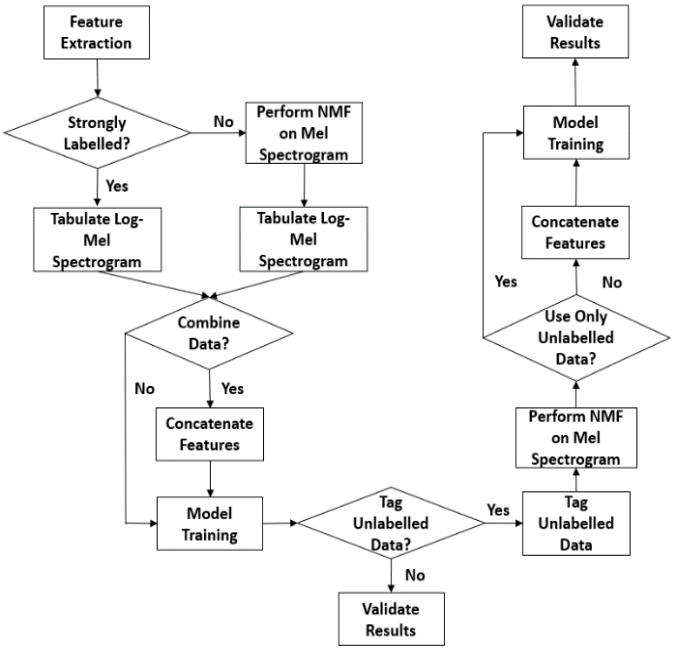

Fig. 1. Flowchart of proposed architecture

4. RESULTS AND DISCUSSION

Based on the proposed system flow, we tested the accuracy of our proposed architecture using the different combination of data on the given validation data that is a mixture of DCASE 2018 task 4 test set, and evaluation set consisting of 1168 audio clips with 4093 events. Based on the results shown in Table 3, the model trained using both weakly labeled data and synthetic data achieved the highest accuracy as compared to using a single type of data. It is surprising to find that strongly labeled synthetic data was not able to achieve higher accuracy than weakly labeled data. Whereas, a combination of data can increase the accuracy of the model. However, the increment in the event based F1

score was only 0.66%. Therefore, it may not be worthwhile to include synthetic data if synthetic data cannot be generated easily.

With the available of a large set of unlabeled in domain data, the next phase of the test was to tag (classify) the unlabeled data using the model trained using only weakly labeled data and the model trained using weakly labeled and strongly labeled synthetic data. As the performance of a model trained using synthetic data was not up to expectation, it was not utilized in the next phase to label the unlabeled data.

Based on Table 4 and 5, results have shown that using only unlabeled in domain data or training a model with the inclusion of unlabeled in domain data labeled using different models, accuracy decreases. This could be due to the quality of unlabeled data being labeled. As seen in the results, if unlabeled in domain data were labeled using a model with higher accuracy, then the validation results would be higher. Thus, it can be deduced that if unlabeled data was not properly labeled, it can have an adverse effect on the accuracy rather than improve the model performance.

Type of Data	Weakly Labeled	Strongly Labeled Synthetic Data	Weakly Labeled and Strongly Labeled Synthetic Data
Event Based F1	29.73%	15.27%	30.39%
Segment Based F1	55.79%	43.59%	57.66%

Table 3. Accuracy using different types of data

Type of Data	Unlabeled Data	Weakly Labeled Data and Unlabeled Data
Event Based F1	25.47%	27.2%
Segment Based F1	45.88%	48.52%

Table 4. Accuracy of model with unlabeled data labeled using model with F1 29.73%

In Table 6, the comparison of results was made between the best performing model trained using weakly labeled data and synthetic strongly labeled data, Kong et al. [19] model and baseline model. Although the proposed model can achieve a better event based F1 score as compared to both other models, it has a lower segment based F1-score as compared to Kong et al. [19]. It may be due to the way how NMF was utilized. In this system, NMF was used to find H that indicates when the event was activated when the calculated H of certain frames were above a predefined threshold. However, if the clip contains multiple events, then NMF will indicate that those frames above a predefined threshold belong to all the events present in the audio. Therefore, it may be worthwhile to investigate the use of source separation before the application of NMF.

Type of Data	Unlabeled Data	Weakly Labelled, Strongly Labeled Synthetic Data and Unlabeled Data
Event Based F1	26.64%	27.84%
Segment Based F1	47.13%	50.92%

Table 5. Accuracy of model with unlabeled data labeled using model with F1 30.39%

Finally, the best four models are chosen to be submitted for the DCASE task 4 challenge, namely, 1) model trained using weakly labeled and strongly labeled synthetic data, 2) model trained using weakly labeled data, 3) model trained using only all data, 4) model trained using weakly labeled and unlabeled data.

Type of Data	Proposed	Kong et al. [19]	Baseline
Event Based F1	**30.39%**	24.1%	23.7%
Segment Based F1	57.66%	**63.0%**	55.2%

Table 6. Comparison of results

5. CONCLUSION

In this paper, a five layers CNN with the use of NMF was proposed for DCASE 2019 task 4. The proposed system was able to achieve an event based F1-score of 30.39% and segment based accuracy of 57.66% as compared to the baseline model that has an event-based F1-score of 23.7% and segment based accuracy of 55.2%. However, there is still room for improvement, particularly in the aspect of source separation that may very well helps in the accuracy of sound event detection.

6. ACKNOWLEDGMENT

We would like to thank the organizers for their technical support especially Nicolas Turpault, Romain Serizel and also Kong Qiuqiang from Surrey University for his prompt replies in regards to our questions on his system.

7. REFERENCES

[1] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Trans Audio, Speech, an Language Process.*, vol. 25, no. 6, pp. 1291-1303, Jun. 2017.

[2] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "Duration-Controlled LSTM for Polyphonic Sound Event Detection," *IEEE/ACM Trans Audio, Speech, an Language Process.*, vol. 25, no. 11, pp. 2059-2070, Nov. 2017.

[3] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley "Sound Event Detection and Time–Frequency Segmentation from Weakly Labelled Data," *IEEE/ACM Trans Audio, Speech, an Language Process.*, vol. 27, no. 4, pp. 777-787, Apr. 2019.

[4] B. McFee, J. Salamon, and J. P. Bello, "Adaptive Pooling Operators for Weakly Labeled Sound Event Detection," *IEEE/ACM Trans Audio, Speech, an Language Process.*, vol. 26, no. 11, pp. 2180-2193, Apr. 2018.

[5] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A Joint Separation-Classification Model For Sound Event Detection of Weakly Labelled Data," *2018 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 321-325.

[6] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound Event Detection In Multichannel Audio Using Spatial and Harmonic Features,"

[7] S. Adavanne, P. Pertila, and T. Virtanen, "Sound Event Detection Using Spatial Features and Convolutional Recurrent Neural Network," *Detection and Classification of Acoustics Scenes and Events 2017*, Munich, Germany, Nov. 2017, pp. 1-5.

[8] J. Lu, "Mean Teacher Convolution System For DCASE 2018 Task 4," *Detection and Classification of Acoustics Scenes and Events 2018*, Shanghai, China, Jul. 2018, pp. 1-5.

[9] D. Su, X. Wu, L. Xu, "GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection," *2010 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 4890-4893.

[10] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen "Acoustic Event Detection In Real Life," *18th European Signal Process. Conf.*, Aalborg, Denmark, Aug. 2010, pp. 1267-1271.

[11] O. Dikmen, and A. Mesaros, "Sound Event Detection Using Non-Negative Dictionaries Learned From Annotated Overlapping Events," *2013 IEEE Workshop Applications Signal Process. Audio Acoustics*, New Paltz, New York, Oct. 2013, pp. 1-4.

[12] V. Bisot, S. Essid, and G. Richard, "Overlapping Sound Event Detection With Supervised Nonnegative Matrix Factorization," *2017 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 31-35.

[13] T. Komatsu, Y. Senda, and R. Kondo, "Acoustics Event Detection Based on Non-Negative Matrix Factorization With Mixtures of Local Dictionaries and Activation Aggregation," *2016 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 2259-2263.

[14] Z. Md. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems," *IEEE Commun. Surveys Tutorials*, vol. 19, no. 4, pp. 2432-2455, 2017.

[15] Z. Liu, Z. Jia, C. Vong, S. Bu, J. Han, and X. Tang, "Capturing High-Discriminative Fault Features for Electronics-Rich Analog System via Deep Learning," *IEEE Trans. Indust. Inform.*, vol. 13, no. 3, pp. 1213-1226, Jun. 2017.

[16] M. He and D. He, "Deep Learning Based Approach for Bearing Fault Diagnosis," *IEEE Trans. Indust. Applications*, vol. 53, no. 3, pp. 3057-3065, Jun. 2017.

[17] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017-5032, Dec. 2015.

[18] A. Mesaros, T. Heittola , E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Trans Audio, Speech, an Language Process.*, vol. 26, no. 2, pp. 379-393, Feb. 2018.

[19] Q. Kong, Y. Cao, T. Iqbal, Yong Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio-tagging, sound event detection spatial localization: DCASE 2019 baseline systems," arXiv: 1904.03476, pp. 1-5.

[20] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, Oct. 1999.