DEEP NEURAL NETWORK BASELINE FOR DCASE CHALLENGE 2016

Qiuqiang Kong, Iwnoa Sobieraj, Wenwu Wang, Mark Plumbley

Centre for Vision, Speech and Signal Processing, University of Surrey, UK
{q.kong, iwona.sobieraj, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

The DCASE Challenge 2016 contains tasks for Acoustic Acene Classification (ASC), Acoustic Event Detection (AED), and audio tagging. Since 2006, Deep Neural Networks (DNNs) have been widely applied to computer visions, speech recognition and natural language processing tasks. In this paper, we provide DNN baselines for the DCASE Challenge 2016. For feature extraction, 40 Mel-filter bank features are used. Two kinds of Mel banks, *same area bank* and *same height bank* are discussed. Experimental results show that the *same height bank* is better than the *same area bank*. DNNs with the same structure are applied to all four tasks in the DCASE Challenge 2016. In Task 1 we obtained accuracy of 76.4% using Mel + DNN against 72.5% by using Mel Frequency Ceptral Coefficient (MFCC) + Gaussian Mixture Model (GMM). In Task 2 we obtained F value of 17.4% using Mel + DNN against 41.6% by using Constant Q Transform (CQT) + Nonnegative Matrix Factorization (NMF). In Task 3 we obtained F value of 38.1% using Mel + DNN against 26.6% by using MFCC + GMM. In task 4 we obtained Equal Error Rate (ERR) of 20.9% using Mel + DNN against 21.0% by using MFCC + GMM. Therefore the DNN improves the baseline in Task 1 and Task 3, and is similar to the baseline in Task 4, although is worse than the baseline in Task 2. This indicates that DNNs can be successful in many of these tasks, but may not always work.

Index Terms— Mel-filter bank, Deep Neural Network (DNN), Acoustic Scene Classification (ASC), Acoustic Event Detection (AED), Audio Tagging

1. INTRODUCTION

Sounds carry a large amount of information about our everyday environment. Humans can perceive the sound scene where they stay (busy street and office, etc.), and recognize individual sound events (car passing by and footsteps). Although image classification and detection have been popular in recent years, audio classification and detection have not attracted a similar level attention. In the past years, CLEAR 2007 was a challenge on detecting events and activities [1]. The DCASE Challenge 2013 [2] contained challenge for scene classification and synthetic acoustic classification . The DCASE Challenge 2016[1] held by Tampere University has four tasks in acoustic related problems. Task 1 is Acoustic Scene Classification (ASC), the goal of which is to classify a test recording into one of the predefined classes that characterize the environment in which it was recorded - for example “park”, “home”, “office”. Task 2 is Acoustic Event Detection (AED) in Synthetic audio, which aims to detect synthetic polyphonic sound events (eg. “doorslam”, “human speaking”) that are present within an audio. Task 3 is Sound Event Detection in Real Life Audio. In contrast to Task 2, it aims to detect acoustic events in real life, such as “bird singing”, “car passing by”. Task 4 is Domestic Audio Tagging, the goal of which is to perform multi-label classification on short recordings collected in a domestic environments.

ASC and AED are intimately related to industry applications. They have applications in audio indexing [3], audio classification [4], audio tagging [5], audio segmentation [6], surveillance, military and public abnormal event detection [7], etc. In previous work, Mel Frequency Ceptral Coefficient (MFCC) and Gaussian Mixture Model (GMM) were used for ASC [8]. McLoughlin *et al.* improved on this result by using auditory features and Deep Neural Network (DNN) classifier [9]. Unsupervised learning used by Lee *et al.* [4] proposed to use convolutional deep belief networks to learn audio features. In AED, the Constant Q Transform (CQT) and Nonnegative Matrix Factorization (NMF) are widely used to detect sound events in a recording [10]. Hidden Markov Models (HMM) with Viterbi decoding have been proposed [7], a universal background model (UBM) is used to model background sound. In [11], a Bidirectional Long Short Term Memory (BLSTM) is proposed, which yields better result than the HMM. In audio tagging, MFCC and GMM is a standard method to detect whether or not tag occurrs in the audio [12]. Recently Convolution Neural Networks (CNNs) has been used for audio tagging in [13].

This work is aimed at providing DNN baseline for all four tasks of the DCASE Challenge 2016. The reminder of the paper is organized as follows. Section 2 discusses related works. Section 3 describes the deep DNN structure. Section 4 are experimental results we obtained on Task 1 - 4 of DCASE Challenge 2016. Section 5 draws conclusion of our

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/

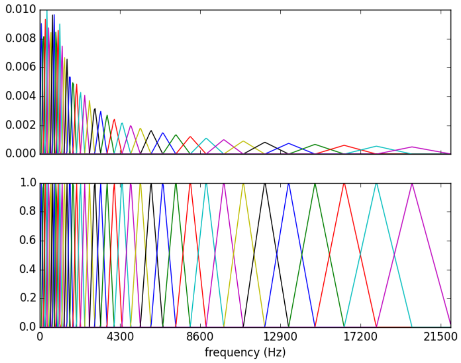

Figure 1: Upper: Mel-filter bank with same bank area (*librosa*). Lower: Mel-filter bank with same bank height (*voicebox*)

work and future research.

2. DEEP NEURAL NETWORKS

DNNs have been widely used in Computer Vision (CV), Natural Language Processing (NLP), *etc.* since 2006. Their variants include CNNs, Recurrent Neural Networks (RNNs). In this paper, we propose to try same features and same structures of DNN for all of the four tasks in the DCASE Challenge 2016. This is aimed at evaluating how DNN performs in these tasks compared with original baseline methods, as well as providing a baseline for other researchers to compare.

2.1. Features

In audio processing, MFCCs are widely used in speech recognition. However, MFCCs are developed inspired by the human speech production process, which assumes sounds are produced by glottal pulse passing through vocal tract filter. However, MFCCs discard useful information about the sound, which restricts its ability for recognition and classification. In recent years, Mel Bank Features have been widely used in speaker recognition [14]. Other features such as CQT [15] are used in music related tasks, which has good resolution in low frequency. In this paper, we apply Mel-filter bank features with 40 channels to all of the four tasks.

Features extraction code is based on *librosa*[2]. The original Mel-bank extracted by *librosa* is shown in upper part of Figure 1. However, this kind of Mel-filter bank is designed for speech analysis. Experimental results in Section 4 show that using reweighted mel-bank with same height (shown in lower part of Figure 1, which is same with *voicebox*[3] in matlab) performs better than the original Mel-filter bank.

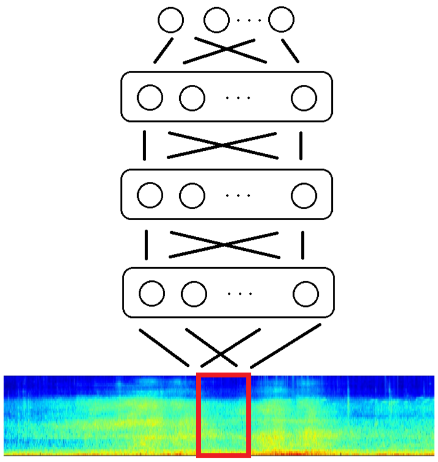

Figure 2: DNN used for Task 1 - 4

2.2. DNN structure

The DNN we use in our experiment is a fully connected neural network with 3 hidden layers. As the bag of frames feature can not capture time dependency, the input to the DNN is taken as a concatenation of 10 frames mel-bank features so there are 400 input nodes (10 frames * 40 Mel-filter banks). We use 500 hidden units per layer. Relu [16] activation function is used. For Task 1, softmax output and categorical cross-entropy loss function are used. For Task 2, Task 3, and Task 4, binary output and binary cross-entropy function are used. Dropout [17] with value of 0.1 is used to avoid overfitting. Rmsprop [18] optimizer are used since it is generally faster than Stochastic Gradient Descend (SGD). The DNN structure is shown in Figure 2.

3. EXPERIMENTS

In this section we evaluate the performance of Mel-filter bank features + DNN on DCASE Challenge 2016 Task 1 - 4 on ASC, AED and audio tagging. We use 40 Mel-filter bank features shown in lower part of Figure 1. Then we apply DNN shown in Figure 2 to all of the four tasks. These systems are implemented in python. The souce code can be found in Task 1[4], Task 2[5], Task 3[6], Task 4[7]. DNN implementation is based on *HAT*[8], which is an open source deep learning framework built on top of *Theano*[9].

[2] https://github.com/librosa

[3] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[4] https://github.com/qiuqiangkong/DCASE2016_Task1

[5] https://github.com/qiuqiangkong/DCASE2016_Task2

[6] https://github.com/qiuqiangkong/DCASE2016_Task3

[7] https://github.com/qiuqiangkong/DCASE2016_Task4

[8] https://github.com/qiuqiangkong/Hat

[9] http://deeplearning.net/software/theano/

3.1. Task 1: Acoustic Scene Classification

TUT Acoustic scenes 2016 dataset is used in this task. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. Each recording contains 30-second segments. There are altogether 15 classes with 4 fold cross validation. For training DNN, the batch size is set to 100. Rmsprop (Section 3.2) learning rate is set to 1e-3 at beginning then is tuned to 10^{-3} after 30 epochs. The maximum epochs is set to 100. Time consumption is 3 s/epoch on Tesla 2090.

We compare the results of Mel-filter bank with the same bank area (upper part of Figure 1) and the Mel-filter bank with the same bank height (lower part of Figure 1). The results are shown in Table 1.

Table 1: Accuracy of Task 1

	Frame based acc.	Event based acc.
MFCC + GMM (Baseline)	-	72.5%
Mel (same bank area) + DNN	39.6%	46.5%
Mel (same bank height) + DNN	**63.3%**	**76.4%**

From this table, it can be observed that using the Mel + DNN with the same bank height obtains accuracy of 76.4%, outperforms MFCC + GMM baseline (72.5%). However, the Mel-filter bank with the same bank area is much worse, with an accuracy of 46.5%. This may result from environmental sound need to be emphasised in high frequency. Normalization of input may help but is not implemented in our experiment and need to be further researched. In the reminder of the paper, we use 40 Mel-filter bank with same bank area as feature. Detailed results on each fold are shown in Table 2.

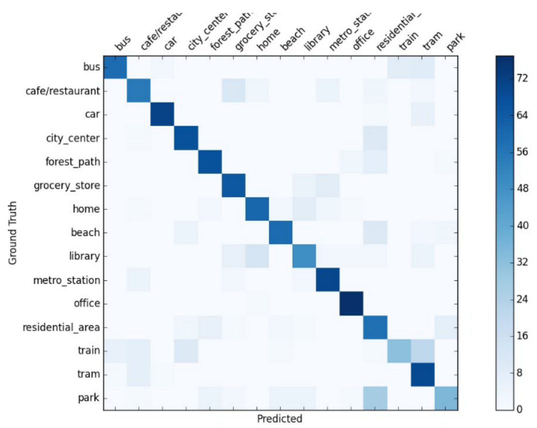

Figure 3: Confusion matrix of event based accuracy in Task 1.

Table 2: Fold wise accuracy of Task 1 using Mel + DNN

	Frame based acc.	Event based acc.
fold 1	65.2%	80.0%
fold 2	61.5%	70.7%
fold 3	62.0%	74.8%
fold 4	64.6%	80.1%
average	**63.3%**	**76.4%**

Table 2 shows that the accuracy of scene classification in different folds is not homogenious, with frame based accuracy ranging from 61.5% to 65.2% and event based accuracy ranging from 70.7% to 80.1%. The overall Confusion matrix is shown in Figure 3. We can see that "park" is easily recognized as "residential area". This may result from these scenes share similar features, which is difficult to classify using bag of words model.

3.2. Task 2: AED in Synthetic Audio

Audio provided by IRCCYN Ecole Centrale de Nantes is used in Task 2. Training set includes 11 classes of sound events. There are 20 samples provided for each sound event class in the training set, plus a development set consisting of 18 minutes of synthetic mixture material in 2 minute length audio files. The event-to-background ratio (EBR)[10] is set to -6, 0, +6 dB. In this task, we set the Rmsprop learning rate to 10^{-3}, the batch size to 20, the number of epochs to 20, respectively. Binary output and sigmoid cost function are used. Time consumption in Tesla 2090 GPU is 0.1 s/epoch. Results are shown in Table 3.

Table 3: F value of Task 2

	F value
CQT + NMF (Baseline)	**41.6%**
Mel + DNN	17.4%

Table 3 conveys that using Mel + DNN yields an F value of 17.4% which is worse than CQT + NMF baseline (41.6%). One possible explanation for this underperformance is that DNN is not good at classifying samples that it has not seen, with NMF has better generalization ability in classifying unseen samples. Detailed results on different EBR levels of -6, 0, +6 dB are shown in Table 4.

[10] http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio

Table 4: Fold wise F value of Task 2 using Mel + DNN

	F value
-6 dB	16.0%
0 dB	17.6%
+6 dB	18.8%
Average	**17.4%**

3.3. Task 3: AED in Real Life Audio

The TUT Sound events 2016 dataset is used in this task. Audio in the dataset is a subset of TUT Acoustic scenes 2016 dataset (used for task 1). Sound events in the TUT Sound events 2016 dataset consists of recordings from two acoustic scenes: Home (indoor) and Residential area (outdoor). In this task, we set the Rmsprop learning rate to 10^{-3}, the batch size to 20, the number of epochs to 50. Results are shown in Table 5.

Table 5: F value of Task 3

	Home	Residential area	Average
MFCC + GMM (baseline)	18.1%	35.2%	26.6%
Mel + DNN	**29.2%**	**47.0%**	**38.1%**

Table 5 shows that for real life event detection using Mel + DNN yields an F value of 38.1%, which outperforms MFCC + GMM baseline (26.6%). Detailed results on each fold are shown in Table 6.

Table 6: Fold wise F value of Task 3 using Mel + DNN

	Home	Residential area
fold 1	28.0%	62.4%
fold 2	28.8%	34.5%
fold 3	22.3%	43.7%
fold 4	37.5%	47.5%
average	**29.2%**	**47.0%**

3.4. Task 4: Domestic audio tagging

The CHiMe-Home dataset is used in Task 4 . The objective of this task is to perform multi-label classification on 4-second audio chunks. There are 7 labels occurring in audio segments including child speech and adult male, *etc*. Binary output and binary cross-entropy loss function are used because the labels can occur simultaneously. We set the Rmsprop learning rate to 10^{-3}, the batch size to 500, the number of epoch to 100. Cross validation with 4 folds is used. Results are shown in Table 7.

Table 7: F value of Task 4

	EER
MFCC + GMM (baseline)	**21.0%**
Mel + DNN	20.9%

Table 7 shows that we obtain Equal Error Rate (ERR) of 20.9% using Mel + DNN, which is similar to MFCC + GMM baseline (21.0%). Detailed results on four folds are shown in Table 8.

Table 8: Fold wise EER of Task 4 using Mel + DNN

	EER
fold 1	19.3%
fold 2	15.6%
fold 3	26.3%
fold 4	22.4%
average	**20.9%**

4. CONCLUSION

In this paper, we have applied the same DNN structure to Task 1 - 4 in the DCASE Challenge 2016 as a DNN baseline for future research. We compared the Mel-filter bank features with the same bank area and Mel-filter bank features with the same height. Experimental results show that the mel-filter bank feature with the same height performs much better in Task 1. In summary, in Task 1, Mel + DNN is better than MFCC + GMM (accuracy 76.5% against 72.5%). In task 2, Mel + DNN is worse than the CQT + NMF baseline (F value 17.4% against 41.6%). In task 3, Mel + DNN is better than the MFCC + GMM baseline (F value 38.1% against 26.6%). In task 4, Mel + DNN is similar to MFCC + DNN baseline (20.9% against 21.0%). We publish our codes of Task 1 - 4 and hope this will attract interests from other institutions to do further research.

5. ACKNOWLEDGMENT

This work is sponsored by CVSSP, University of Surrey and China Scholarship Council (CSC). MP is partly supported by EPSRC granted EP /NO. 14111/ 1

6. REFERENCES

[1] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, R Travis Rose, Martial Michel, and John Garofolo. The clear 2007 evaluation, multimodal technologies for perception of humans: International evaluation workshops clear 2007 and rt 2007, baltimore, md, usa, may 8-11, 2007, revised selected papers, 2008.

[2] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[3] Rui Cai, Lie Lu, Alan Hanjalic, Hong-Jiang Zhang, and Lian-Hong Cai. A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):1026–1039, 2006.

[4] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104, 2009.

[5] Mohit Shah, Brian Mears, Chaitali Chakrabarti, and Andreas Spanias. Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices. In *Emerging Signal Processing Applications (ESPA), 2012 IEEE International Conference on*, pages 99–102. IEEE, 2012.

[6] Gordon Wichern, Jiachen Xue, Harvey Thornburg, Brandon Mechtley, and Andreas Spanias. Segmentation, indexing, and retrieval for environmental and natural sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):688–707, 2010.

[7] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–13, 2013.

[8] Burak Uzkent, Buket D Barkana, and Hakan Cevikalp. Non-speech environmental sound classification using svms with a new set of features. *International Journal of Innovative Computing, Information and Control*, 8(5):3511–3524, 2012.

[9] Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, and Wei Xiao. Robust sound event classification using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):540–552, 2015.

[10] Arnaud Dessein, Arshia Cont, and Guillaume Lemaitre. Real-time detection of overlapping sound events with non-negative matrix factorization. In *Matrix Information Geometry*, pages 341–371. Springer, 2013.

[11] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Recurrent neural networks for polyphonic sound event detection in real life recordings. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444. IEEE, 2016.

[12] Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley. Chime-home: A dataset for sound source recognition in a domestic environment. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pages 1–5. IEEE, 2015.

[13] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

[14] Li Deng, Ossama Abdel-Hamid, and Dong Yu. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6669–6673. IEEE, 2013.

[15] Judith C Brown and Miller S Puckette. An efficient algorithm for the calculation of a constant q transform. *The Journal of the Acoustical Society of America*, 92(5):2698–2701, 1992.

[16] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[17] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[18] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.