DEEP NEURAL NETWORK BASELINE FOR DCASE CHALLENGE 2016

Qiuqiang Kong, Iwona Sobieraj, Wenwu Wang, Mark D. 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 Scene 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. In Task 1 we obtained accuracy of 81.0% using Mel + DNN against 77.2% by using Mel Frequency Cepstral Coefficients (MFCCs) + Gaussian Mixture Model (GMM). In Task 2 we obtained F value of 12.6% using Mel + DNN against 37.0% by using Constant Q Transform (CQT) + Nonnegative Matrix Factorization (NMF). In Task 3 we obtained F value of 36.3% using Mel + DNN against 23.7% by using MFCCs + GMM. In Task 4 we obtained Equal Error Rate (ERR) of 18.9% using Mel + DNN against 20.9% by using MFCCs + GMM. Therefore the DNN improves the baseline in Task 1, 3, and 4, although it is worse than the baseline in Task 2. This indicates that DNNs can be successful in many of these tasks, but may not always perform bette than the baselines.

Index Terms— 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 around them (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 of 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] 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 at detecting sound events in synthetic mixture (e.g. “door slam”, “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 environment.

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 Coefficients (MFCCs) 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 proposed by Lee *et al.* [4] uses 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 in [7], where 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, MFCCs + GMM is a standard method to detect whether or not tags occur in the audio [12]. Recently Convolution Neural Networks (CNNs) have been used for audio tagging in [13].

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

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

ants include CNNs and Recurrent Neural Networks (RNNs). In this paper, we propose to use the same features and the 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. They are developed with the assumption that sounds are produced by glottal pulse passing through vocal tract filter. However, with MFCCs some useful information about the sound may be lost, which restricts its ability for recognition and classification. In recent years, Mel Filter Bank Features have been widely used in speaker recognition [14]. Other features such as Constant Q Transform (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].

2.2. DNN structure

The DNN we used in our experiment is a fully connected neural network with 3 hidden layers. As the bag of frames feature cannot 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 Tasks 2, 3, and 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 is used since it is generally faster than Stochastic Gradient Descend (SGD). The DNN structure is shown in Figure 1.

3. EXPERIMENTS

In this section we evaluate the performance of Mel-filter bank features plus DNN on DCASE Challenge 2016 Tasks 1 - 4 on ASC, AED and audio tagging. We use 40 Mel-filter bank features. Then we apply the DNN shown in Figure 1 to all of the four tasks. These systems are implemented in python. The source code can be found in Task 1[3], Task 2[4], Task 3[5], Task 4[6]. Our DNN implementation is based on *Hat*[7], which

[2]https://github.com/librosa
[3]https://github.com/qiuqiangkong/DCASE2016_Task1
[4]https://github.com/qiuqiangkong/DCASE2016_Task2
[5]https://github.com/qiuqiangkong/DCASE2016_Task3
[6]https://github.com/qiuqiangkong/DCASE2016_Task4
[7]https://github.com/qiuqiangkong/Hat

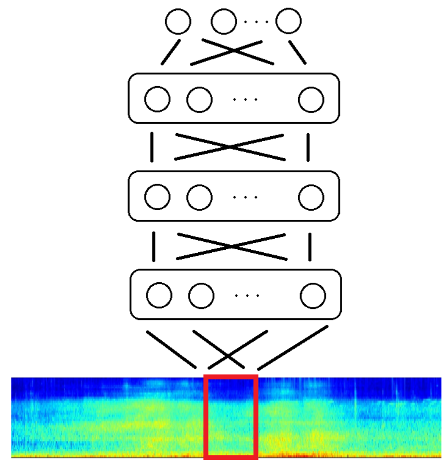

Figure 1: DNN used for Task 1 - 4

is an open source deep learning framework built on top of *Theano*[8].

3.1. Task 1: Acoustic Scene Classification

The TUT Acoustic scenes 2016 datasets [19] 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 10^{-3} at beginning then is tuned to 10^{-4} after 30 epochs. The maximum number of epochs is set to 100. Time consumption is 3 s/epoch on Tesla 2090. The results are shown in Table 1. NG is the abbreviation of Not Given. Dev., Test means development dataset and private dataset, respectively.

Table 1: Accuracy of Task 1

	Chunk based acc. (Dev.)	Segment based acc. (Dev.)	Segment based acc. (Test)
MFCCs + GMM (Baseline)	NG	72.5%	77.2%
Mel + DNN	**63.3%**	**76.4%**	**81.0%**

From this table, it can be observed that using the Mel + DNN obtains an accuracy of 81.0%, outperforms MFCCs + GMM baseline (77.2%) in test dataset. Detailed results of development set on each fold are shown in Table 2.

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

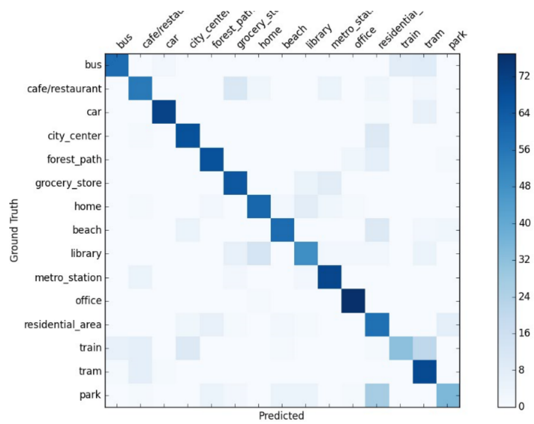

Figure 2: Confusion matrix of segment based accuracy in Task 1.

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

	Frame based acc.	Segment 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 different folds (development dataset) are different, with frame based accuracy ranging from 61.5% to 65.2% and segment based accuracy ranging from 70.7% to 80.1%. This indicates the dataset is not homogeneous. The overall confusion matrix is shown in Figure 2. We can see that "park" is easily mis-recognized as "residential area". This may result from the fact these scenes share similar features, which are difficult to classify using the bag of words model.

3.2. Task 2: AED in Synthetic Audio

A dataset provided by IRCCYN Ecole Centrale de Nantes is used in Task 2 [19]. The 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)[9] 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 (one processor). Results are shown in Table 3.

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

Table 3: F value of Task 2

EBR	F value (Dev.)	F value (Test)
CQT + NMF (Baseline)	**41.6%**	**37.0%**
Mel + DNN	17.4%	12.6%

Table 3 shows that using Mel + DNN yields an F value of 12.6% which is worse than CQT + NMF baseline (37.0%). One possible explanation for this underperformance is that without data augmentation, DNN is not good at classifying the samples with additive noise, while the NMF based technique has better ability in modeling sounds with additive noise. Detailed results on development dataset on different EBR levels of -6, 0, +6 dB are shown in Table 4.

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 [19] is used in this task. Audio in the dataset is a subset of TUT Acoustic scenes 2016 dataset (used for task 1). 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 (Dev.)	Residential area (Dev.)	Average (Dev.)	Average (Test)
MFCCs + GMM (baseline)	15.9%	31.5%	23.7%	34.3%
Mel + DNN	**29.2%**	**47.0%**	**38.1%**	**36.3%**

Table 5 shows that for real life event detection using Mel + DNN yields an F value of 36.3%, which outperforms MFCCs + GMM baseline (23.7%). Detailed results on development dataset 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 (Dev.)	(Test)
MFCCs + GMM (baseline)	**21.0%**	20.9%
Mel + DNN	20.9%	**18.9%**

Table 7 shows that we obtain Equal Error Rate (ERR) of 18.9% using Mel + DNN, which is similar to MFCCs + GMM baseline (20.9%). Detailed results on development dataset 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. In summary, in Task 1, Mel + DNN is better than MFCCs + GMM (accuracy of 81.0% against 77.2%). In task 2, Mel + DNN is worse than the CQT + NMF baseline (F value 12.6% against 37.0%). In task 3, Mel + DNN is better than the MFCCs + GMM baseline (F value 36.3% against 23.7%). In task 4, Mel + DNN is better than MFCCs + DNN baseline (18.9% against 20.9%). 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 ”Making Sense of Sounds”, CVSSP, University of Surrey and China Scholarship Council (CSC), European Union’s H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n 642685 MacSeNet. MP is partly supported by EPSRC grante EP/N014111/1.

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 *IEEE International Conference on Emerging Signal Processing Applications (ESPA), 2012*, 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 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 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.

[19] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *24th European Signal Processing Conference*, 2016.