COMPARING THE INFLUENCE OF DEPTH AND WIDTH OF DEEP NEURAL NETWORK BASED ON FIXED NUMBER OF PARAMETERS FOR AUDIO EVENT DETECTION

Jun Wang, Shengchen Li

Beijing University of Posts and Telecommunications
No. 10, Xi Tu Cheng Road, Beijing, 100876

ABSTRACT

Deep Neural Network (DNN) is a basic method used for the rare Acoustic Event Detection (AED) in synthesised audio. The structure of DNNs including Multi-Layer Perceptron (MLP) and Recurrent Neural Network (RNN) for AED tasks has rather fewer hidden layers compared with computer vision systems. This paper tries to demonstrate that a DNN with more hidden layers does not necessarily guarantee a better performance in AED tasks. Taking the rare AED in synthesised audio with MLPs as an example and simulating a fixed budget of memory in an embedded system, various structures of MLPs are tested with fixed number of parameters engaged. Comparing the importance of neuron numbers in a hidden layer (i.e. the width of DNNs) and the importance of layer numbers in DNNs (i.e. the depth of DNNs) for AED tasks, the performance of the candidate DNN systems are evaluated by the event-based error rate. The results illustrate that a shallower network may outperform a deeper network when enough parameters are engaged and a larger number of parameters introduces a better performance in general.

Index Terms— Deep neural network, shallow neural network, audio event detection

1. INTRODUCTION

The purpose of Audio Event Detection (AED) is to identify the sound events in audio recordings, including estimating the onset and offset of sound events and giving the label for each event. There are many applications of audio event detection such as multimedia indexing [1], intelligent monitoring system in living environment [2], scene classification and recognition [3], etc.

Towards the automatic audio event detection and tagging in polyphonic audio, rare audio event detection in synthesised audio forms an initial stage of AED research. To detect a synthesised rare audio event in a piece of polyphonic audio, various types of Deep Neural Networks are used. Regardless of specific structure of DNNs used, the DNNs used for AED tasks generally have less layers than the best performed DNNs for computer vision tasks. In the field of computer vision, neural networks are generally have a number of hidden layers. For example, Szegedy et al [4] proposed a deep convolutional neural network (CNN) with 22 layers to achieve the classification and detection of images. Simonyan et al [5] found that a CNN with 16-19 weights layers introduces a significant improvement to image recognition.

For AED tasks, the best performed DNNs usually have fewer hidden layers compared with the DNNs used for computer vision tasks: Choi et al [6] propose a noise reduction approach to enhance mel-band energy feature in DNN-based system with only 3 hidden layers. Hayashi et al [7] propose a new method of recognising daily human activities based on a DNN classifier with only 3 hidden layers. Salamon and Bello [8] proposed for acoustic modelling using DNNs with 4 hidden layers. However, the speech processing, there are research [9] asserting a DNN with more layers could help the performance of system. As there are no literatures reporting the DNNs used for AED tasks prefer more hidden layers, an experiment is presented in this paper to test whether a DNN with more hidden layers is likely to introduce the success of AED tasks.

For easier presentation, the number of neurons in a layer is defined as the width of a layer. A wider layer has more neurons. The depth of a DNN is defined as the number of hidden layers in a DNN. A deeper neural network has more hidden layers. In this paper, the importance of width and depth of a DNN in AED tasks are demonstrated with fixed number of parameters engaged in a MLP for rare audio event detection in synthesised audio.

The constrains applied in the paper, the fixed number of parameters for a MLP, has a scenario in real life. If a system for AED is implemented with a Field Programmable Gate Array (FPGA) chip whose budget of memory and logic resources are engaged. The most suitable MLP for the system should be the MLP that makes full use of all resources and introduces the best performance. Thus the full use of memory makes the number of parameters in MLPs unchanged while different structures of MLPs are implemented by the system.

When the number of parameters in a DNN system are fixed, there are two typical ways to design a MLP for AED tasks: a deeper but narrower network or a shallower but wider network. more hidden layers with fewer neurons in each layer and fewer To our knowledge, there are few research focusing

on the performance of two types of MLPs with fixed number of parameters engaged. In this paper, a pilot experiment is presented to demonstrate the importance of depth and width of a MLP by evaluating the performance of MLPs in rare audio event detection tasks with synthesised audio.

The reminder of the paper is organized as follows. Section 2 introduces the proposed architecture of this paper. Section 3 describes the dataset and experiment setup and analyzes experimental results. Section 4 draws the conclusion of our work.

2. PROPOSED ARCHITECTURE

2.1. Multi-Layer Perceptron

MLP is a commonly used method for AED tasks [10] despite the simple structure and limited accuracy. The MLPs used in this paper has an input layer, certain number of hidden layers and an output lyaer. The input of the MLP is the mel-band energies with 40 bands for a concatenation of five successive frames of a piece of audio. Each frame of audio lasts 40 ms with 44.1KHz sampling rate and there are 20 ms overlap between the neighbour frames. The output layer consists of softmax activation function with two units. The activation and inactivation of the target rare audio event are seen as two states that are mutually exclusive (i.e. only one state is valid at any time).

For the hidden layers, the Rectified Linear Unit (ReLU) [11] is used as the activation function for better performance and faster convergence [12]. By setting the number of hidden layers L, the number of parameters in MLP N, the number of output neurons T and the width of MLPs H follows

$$N = (L-1) * H^2 + (D + T + L) * H + T \qquad (1)$$

As N and T is fixed in this experiment, the width H can be calculated by a given L. In this experiment, the number of hidden layers are varies to test the importance of depth and width of MLPs for rare audio event detection in synthesised audio.

To compare the importance of depth and width of the MLPs, whose number of parameters are engaged, in rare audio event detection. The depth of MLPs is set to a dedicated value first and the width of MLPs is calculated by equation (1). Hence by setting the total number of parameters N in MLPs and the number hidden layers L, the structure of a MLP is decided. For simplicity, we use $\{N, L\}$ to represent a MLP who has L hidden layers and N parameters in total.

The number of parameters in MLP is set to 12K, 80K as these are the number of parameters in the DCASE 2017 baseline system [13] and our prior works [14], where both system is based on MLPs. And the total number of parameters is set to 120K to further study the content of this paper.

2.2. Post-processing

With a predefined threshold, the activation of rare audio event engaged is determined by the outputs of the activation neuron in the output layer of the MLPs. Based on the knowledge that the audio event is continues in time domain, a median filter is applied to the results of thresholding process to remove possible outliers of the output. The window of median filter is successive 5 samples of the output of the thresholding process, which is long enough to fix some outliers without affecting the detection of rare audio events.

3. EXPERIMENT AND RESULTS

3.1. Dataset

In this paper, the detection of rare sound events provided by the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2017) [13] is presented in order to analyse the influence of the width and depth of the neural network on performance. The term rare indicates that every target sound event to be detected could occur at most once within a half-minute period. In this AED task, there are three target sound events to be detected: baby cry, gun shot and glass break. The audio data consists a mixture of a piece of natural background sound and a rare audio event synthesised artificially with different Signal-Noise Ratios (SNR) at a random timing point.

In this experiment, the possible rare audio event to detected is known but there might be no rare events in the given piece of audio. As a result, three separate MLPs with the identical structure are to be developed for the detection of three target sound events.

The number of rare event audio for synthesis used in this experiment varies for each target rare audio events as shown in Table 1. The instances of rare audio used for training audio synthesis will not be used for the synthesis process of testing audio. On average, the duration of rare audio event is short compared with the background audio. The mean value for the duration of the sound event in development datasets is shown in Table 2. In all 1121 background audios, 844 of them are used for training datasets and 277 of them are used for test datasets

Events	Train	Test	Total
Baby Cry	106	42	148
Glass Break	96	43	133
Gun Shot	134	53	187

Table 1. The number of instances of rare audio events for synthesis.

The mixtures of rare audio events and background audio are generated with an engaged algorithm [13]. In this paper, there are 500 mixtures synthesised for every target class of

Events	Train	Test
Baby Cry	2.41s	1.85s
Glass Break	1.36s	0.72s
Gun Shot	1.43s	1.04s

Table 2. The duration of sound events in development datasets.

rare audio event in the training dataset and another 500 mixtures are synthesised to form a test dataset. The probability of the existence of target event is 0.5, that is, 250 synthetic pieces of audio have target events and the remaining pieces audio have no target events present. The combination of the rare sound instance and the background sound are randomly selected. The event-to-background ratios (EBR) which was defined as a ratio of average Root Mean Squared Error (RMSE) value calculated over the duration of the event are also randomly selected from predefined probability value.

3.2. Setup

A typical cross-validation method is used in this experiment. The training dataset is divided into two subsets, where 90% of the data is used for training MLPs and the remaining 10% of the data is used for validating the resulting models to prevent overfitting problem. The performance of the final MLPs are evaluated by the test dataset only.

In order to make full use of the computing resources, tensorflow [15] framework is used to build deep neural networks in this paper. The parameters of the network were initialized by random values sampled from zero-mean normal distribution. Three DNN-based systems are trained by using back-propagation with cross-entropy loss function, correct labels and estimated labels. A stochastic gradient descent algorithm [16] is performed using Adam algorithm optimization [17] in mini-batches to improve learning convergence. Dropout technique [18] is used to prevent overfitting problem.

3.3. Results

For evaluation of system performance, evaluation of results includes event-based error rate (ER) as metrics described in [19] in the experiment of this paper. Event-based metrics compare system output and corresponding reference event by event. The ER is the total number of insertions I, deletions D and substitutions S relative to the number of reference events E.

$$ER = \frac{S + D + I}{E} \tag{2}$$

The calculation of ER is calculated by equation (2).

With the MLPs used in this experiment, the event based error rate (ER) of all three target class of audio events are listed in Table 3.

(a) $N = 12K$

$\{N, L\}$	BCy	GBk	GSt	Avg
$\{12K, 1\}$	0.70	0.22	0.78	0.56
$\{12K, 2\}$	0.75	0.22	0.71	0.56
$\{12K, 3\}$	**0.63**	0.22	0.71	0.52
$\{12K, 4\}$	0.64	**0.21**	**0.68**	0.51
$\{12K, 5\}$	0.74	0.23	1.00	0.65

(b) $N = 80K$

$\{N, L\}$	BCy	GBk	GSt	Avg
$\{80K, 1\}$	0.68	0.20	0.68	0.52
$\{80K, 2\}$	**0.66**	**0.18**	0.56	0.47
$\{80K, 3\}$	0.69	0.24	**0.56**	0.49
$\{80K, 4\}$	0.78	0.24	0.58	0.53
$\{80K, 5\}$	0.75	0.23	0.61	0.53

(c) $N = 120K$

$\{N, L\}$	BCy	GBk	GSt	Avg
$\{120K, 1\}$	**0.62**	**0.17**	0.53	0.44
$\{120K, 2\}$	0.70	0.20	**0.50**	0.48
$\{120K, 3\}$	0.74	0.24	0.51	0.49
$\{120K, 4\}$	0.78	0.23	0.55	0.52
$\{120K, 5\}$	0.72	0.26	0.54	0.51

Table 3. Event-based error rate of the rare audio event detection in synthesised audio with different structures of MLPs engaged. In the representation $\{N, L\}$, N represents the total number of parameters in MLPs and L represents the number of hidden layer in MLPs.'BCy' represents "baby cry", 'GBk' represents "glass break", 'GSt' represents "gunshot" and 'Avg' represents the average performance. A smaller value indicates a better performance of MLP.

As shown in Table 3, the error rate of classification does not decrease with the depth of neural network increasing when the total numbers of network are fixed as 12K, 80K and 120K respectively. When the total number of parameters is fixed as 12K, the best performed MLP detecting "glass break" and "gunshot" has 4 hidden layers whereas the best performed MLP detecting "baby cry" has 3 hidden layers instead. When the total number of parameters is set to 80K, the best performed MLP for "baby cry" and "glass break" has 2 hidden layers and the best MLP for detecting "gunshot" has 3 hidden layers. With 120K parameters in the MLPs, there are 2 hidden layers in the best MLP detecting "gunshot" and there is only 1 hidden layer in the best MLP detecting "baby cry" and "glass break". In general, a shallower MLP structure is preferred by a MLP with more parameters.

Comparing the best results of MLPs with different number of parameters engaged, the MLPs with more parameters outperforms the MLPs with fewer parameters. Thus we can say the shallower MLPs can outperform the deeper MLPs when there are enough parameters are engaged in MLPs.

Considering the results presented in Table 3, when imple-

menting a hardware solution for rare audio event detection system, a deeper MLP may not improve the accuracy of audio event detection tasks with fixed budget of memory and computation source especially when the budget of memory is adequate.

4. DISCUSSION

Shown by the results of experiment, a MLP with more parameters engaged prefer a shallower structure. Especially for the MLP with 120K parameters, the MLP with only 1 hidden layer performs best in general whereas for small MLPs who has 12K parameters, the best performed MLP has 4 layers in general, which is deeper than the MLP with 120K parameters.

This fact suggests that given the total number of parameters engaged, the best performed MLPs can detect rare audio events in two different ways. If the shallower MLPs outperforms the deeper MLPs, the success of audio event detection depends on a group of direct mapping between inputs and outputs thus requires more parameters in MLPs. In the case that the deeper MLPs is better than shallower MLPs, the mapping between inputs and outputs is hierarchically established.

To demonstrate the proposed hypothesis, the inactivity ratio of weight in MLPs are calculated. If the absolute value of a weight is less than 0.01, the weight between two neurons is said to be *inactive*. The ratio of weights inactivity is defined as ratio between the number of inactivate weights and the total number of parameters.

From the results presented in Table 4, the inactivity ratio of weights in MLPs only depends on the depth and the width of the MLP. The MLPs with more parameters are less active. Moreover, a deeper and narrower MLP usually has lower activity ratio of weights thus is more active in general and vice versa.

The results can demonstrate that the way that the MLP detects rare audio events with different number of parameters engaged is different. As a result, a deeper MLP does not always outperform a shallower MLP for detecting rare audio events in synthesised audio due to different number of parameters in the MLPs especially there are more parameters engaged in the MLPs.

5. CONCLUSION

In this paper, we study the influence when the size of layers and the number of layers are varied so as to freeze the total number of parameters. In the case of different numbers of parameters engaged, we found that shallow neural network can outperform the neural network with deep architectures. Thus given a constraint on total number of parameters, a deep MLP cannot guarantee the success of rare audio event detection in synthesised audio.

From the results presented, given a hardware platform whose source budget is engaged, the best performance of the hardware may be introduced by a shallower DNN rather than a deeper DNN when the budget of memory is adequate.

(a) $N = 12K$

$\{N, L\}$	Baby Cry	Glass Break	Gun Shot
$\{12K, 1\}$	0.377	0.424	0.348
$\{12K, 2\}$	0.324	0.339	0.283
$\{12K, 3\}$	0.280	0.288	0.243
$\{12K, 4\}$	**0.247**	**0.261**	**0.229**
$\{12K, 5\}$	0.248	0.277	0.369

(b) $N = 80K$

$\{N, L\}$	Baby Cry	Glass Break	Gun Shot
$\{80K, 1\}$	0.564	0.573	0.538
$\{80K, 2\}$	0.539	0.534	0.513
$\{80K, 3\}$	0.462	0.488	0.470
$\{80K, 4\}$	**0.464**	**0.462**	**0.229**
$\{80K, 5\}$	0.459	0.468	0.494

(c) $N = 120K$

$\{N, L\}$	Baby Cry	Glass Break	Gun Shot
$\{120K, 1\}$	0.583	0.592	0.569
$\{120K, 2\}$	0.550	0.569	0.550
$\{120K, 3\}$	0.520	0.543	0.541
$\{120K, 4\}$	**0.498**	**0.522**	**0.525**
$\{120K, 5\}$	0.500	**0.522**	0.531

Table 4. The ratio of weights inactivity with different number of parameters N engaged in MLPs. For presentation $\{N, L\}$, N represents the total number of parameters in MLPs and L represents the number of hidden layers. A smaller number indicates that the weights in MLPs are more active.

6. REFERENCES

[1] Dongqing Zhang and Dan Ellis, "Detecting sound events in basketball video archive," *Department of Electronic Engineering, Columbia University*, 2001.

[2] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, Jan. 2013.

[3] 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*, 2009, pp. 1096–1104.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *arXiv:1409.4842*, pp. 1–9, June 2014.

[5] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICRL)*, pp. 1–14, 2015.

[6] Inkyu Choi, Kisoo Kwon, Soo Hyun-Bae, and Nam Soo-Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction," 2016.

[7] Pierre Laffitte, David Sodoyer, Charles Tatkeu, and Laurent Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2016, IEEE.

[8] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[9] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 437–440, 2011.

[10] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 International Joint Conference on Neural Networks (IJCNN)*. July 2015, IEEE.

[11] Vinod Nair and Geoffrey E. Hinton , "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th International Conference on Machine Learning*, , no. 3, pp. 807–814, 2010.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks," *AISTATS '11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, 2011.

[13] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, Nov. 2017.

[14] Jun Wang and Shengchen Li, "Multi-frame concatenation for detection of rare sound events based on deep neural network," , no. November, 2017.

[15] Martín Abadi, Paul Barham, Jianmin Chen, Zhi feng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiao qiang Zheng, and Google Brain, "Tensorflow: A system for large-scale machine learning," *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pp. 265–284, Nov. 2016.

[16] Pieter Abbeel, "Policy gradient," *Control*, pp. 1–6, 2008.

[17] Asmelash Teka Hadgu, Aastha Nigam, and Ernesto Diaz-Aviles, "Large-scale learning with adagrad on spark," in *2015 IEEE International Conference on Big Data (Big Data)*. Oct. 2015, IEEE.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout : A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[19] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162–178, May 2016.