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

wangjun19930314@bupt.edu.cn

shengchen.li@bupt.edu.cn

1 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.

2 Proposed Architecture

The MLPs used in this paper has an input layer, certain number of hidden layers and an output layer. The total number N in a MLP is

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

- L: the number of hidden layers
- D: the number of input neurons
- T: the number of output neurons
- H the width of MLPs

The depth of MLPs is set to a dedicated value first and the width of MLPs is calculated by equation (1). the number of hidden layers are varies to test the importance of depth and width of MLPs.

3 Experiment and results

The number of parameters in MLPs is set to three pre-selected numbers: 12K, 80K, 120K, which resulting models are referred as 12K models, 80K models and 120K models respectively:

- 12K models, DCASE2017 baseline system [1];
- 80K models, our prior work [2] in DCASE 2017;
- 120K models, for further verification.

The base performed models are:

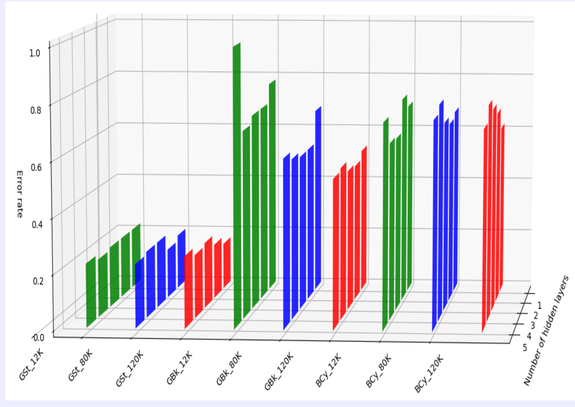
- For $N = 12K$, $L = 4 \implies ER = 0.51$
- For $N = 80K$, $L = 2 \implies ER = 0.47$
- For $N = 120K$, $L = 1 \implies ER = 0.44$

The performance of the MLPs are evaluated by the event based on Error Rate (ER), i.e.

$$ER = \frac{S + D + I}{E} \quad (2)$$

- I: insertions
- D: deletions
- S: substitutions
- E: the number of reference events

Event-based metrics [3] compare system output and corresponding reference event by event. The general result is shown in the figure below.



Results of with 12K,80K and 120K models, where 'BCy', 'GBk', 'GSt' represent 'baby cry', 'glass break', 'gunshot' respectively.

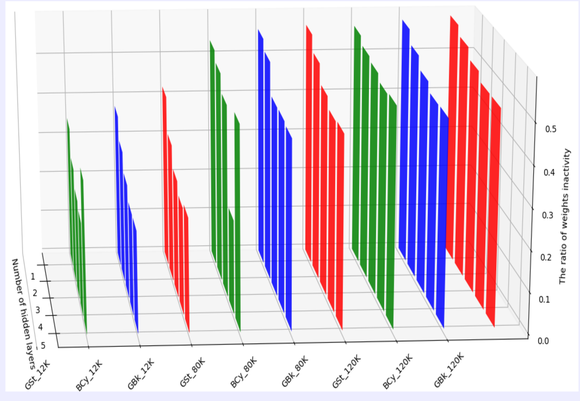
From the figure above, we find two potential effecting fators:

- Shallower MLPs outperform the deeper MLPs when there are enough parameters are engaged in MLPs.
- 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.

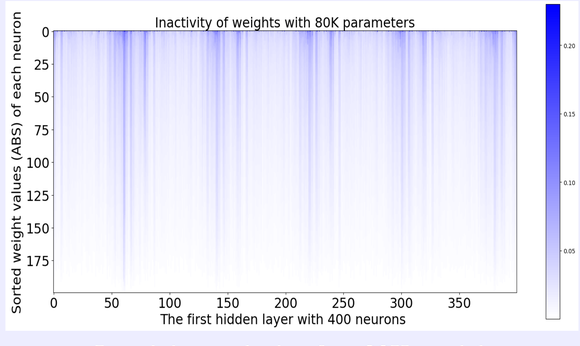
4 Discussion

The inactivity ratio ($\mathcal{I}$) of weights (w) in MLPs is examined to show how acoustic events are detected.

$$\mathcal{I} = \frac{\text{count}(|w_{ij}| < 0.01)}{N} \quad (3)$$



Inactivity with 12K, 80K and 120K parameters.



Inactivity analysis of an 80K model

From the figure above, we find two potential effecting factors:

- The MLPs with more parameters are less active. A deeper and narrower MLP usually has lower activity ratio of weights thus is more active in general and vice versa.
- 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 Conclusions

- Shallow neural network can outperform the neural network with deep architectures.
- With the same number of parameters, a deep MLP cannot guarantee the success of rare audio event detection in synthesised audio.
- The best performance of the hardware may be a shallower DNN rather than a deeper DNN when the budget of memory is adequate.

6 References

[1] Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, et al. DCASE 2017 challenge setup: Tasks, datasets and baseline system, DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events, Nov. 2017.

[2] Jun Wang and Shengchen Li, Multi-frame concatenation for detection of rare sound events based on deep neural network , November, 2017.

[3] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, Metrics for polyphonic sound event detection, Applied Sciences, Vol. 6, No. 6, pp. 162−178, May 2016.