DCASE 2017 ACOUSTIC SCENE CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK IN TIME SERIES

Biho Kim

Hyundai Motor Group R&D Division
Gyeonggi-do, Republic of Korea
biho@hyundai.com

ABSTRACT

This technical paper presents our approach for the acoustic scene classification (ASC) task in DACSE2017 challenge. We propose combination of recently deep learning algorithm for classification sequence of audio. We stack dilated causal convolution which is efficient for time series signal without recurrent structure and use SELU activation unit instead batch-normalization. Based on this, various experiments were evaluated on the ASC development dataset. The results were analyzed from different perspectives and the best accuracy score obtained by our system on 75.9% .

.

Index Terms— Dilated causal convolution, SELU, WaveNet

1. INTRODUCTION

In ASC task, it is important to explore frequency dependency or mel-band dependency, but it is more important to explore relationship between frames in terms of classifying scenes. From this perspective, there is many approach. Although recurrent neural network is most success in many areas such as word2vec, translation, but it is not widely used in ASC task. There are variety of reasons, one of them is that noisy training data makes RNN difficult to learn. Instead, there are many architectures that is similar role without recurrent. We pay attention to this, we have examined many structures that can deal with dependency, and finally we have defined the following structure.

2. PROPOSED METHOD

For this task, we employ CNN with residual and skip connections in WaveNet also self-normalizing activation unit called SELU [1], [2]. WaveNet is well-known model based residual block and dilated convolutions neural net instead of LSTM-RNN. It shows that stack of dilated causal convolution layers can effectively capture long range contexts without RNN structure and residual block makes stacked this layers trained. In original paper, it only deals with raw audio signal, the other papers show possibility using other features [3]. Another point is self-normalizing neural networks. Batch normalization which is well-known technique addressing internal covariate shift is widely used for image recognition based on CNN. SELU introduced by G. Klambauer is similar with RELU containing batch normalization [2].

In our model, we stacked residual block introduced in WaveNet which consists of two types of CNN contained residual and skip connections except first layer for frequency dependency training each time stamp. Gated activation unit which is combination of tanh and sigmoid connect 1x1 convolution, this is used for not only skip connections but also deeper training with residual. This throughout skip connection is used for softmax output. Another difference with original block, we used SELU activation unit all layer except above mentioned tanh and sigmoid layer.

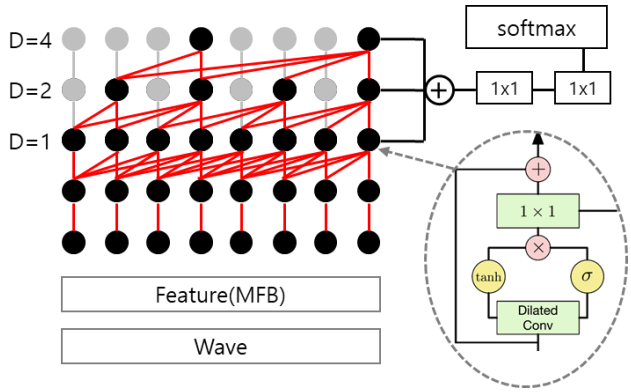

Figure 1: The architecture of the proposed method.

Overall structure is represented in Figure 1. First layer is used for training frequency dependency. After first layer, the above mentioned block is used with alpha dropout [2]. Original paper used non-overlap dilated causal convolution, but we use overlap dilated causal convolution. Overlap structure is similar to semi fully-connected structure, but causal structure makes current input more directly trained. In our test set, causal overlap structure mostly better than fully-connected and non-overlap case.

3. EVALUATION

For our evaluation, we use the DCASE 2017 development data set. The data set consists of 15 different classes which are bus, café/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office. Residential area, train, tram, urban park.

The model is evaluated according to a four-fold cross-validation scheme. For the evaluation of each fold, per-class accuracies are calculated on the test set on a segment-wise level. Final output is calculated by the total number of segments belonging to the class. Finally, the overall accuracy is calculated by averaging the four-fold accuracy.

For training, all cases use 10% validation set for early stopping which check no improvement more than 2 epochs. Almost cases are stopped after 10 epochs except a few case. In each system, first layer and softmax output layer are all the same, only middle layers are different neural networks which use activation unit with tanh and sigmoid in above mentioned residual block and it is written in table. All layers use 64 filters except output layer.

Structure	Input len.	Max dilated len.	Kernel len.	Accuracy
Baseline	5	-	-	74.8%
Fully-connected	4	-	4	73.9%
Fully-connected	8	-	8	73.8%
Fully-connected	12	-	12	73.5%
Dilated causal	4	4	2	73.1%
Dilated causal	4	4	4	73.4%
Dilated causal	8	4	2	74.5%
Dilated causal	8	4	4	74.6%
Dilated causal	**8**	**4**	**8**	**75.9%**
Dilated causal	8	8	2	74.1%
Dilated causal	8	8	8	73.8%

Table 2: The output of the experiments using MFB.

Overall outputs are represented in Table 1. Notable points are as follows. First, Fully-connected structure is similar output regardless of kernel length. Rather, as the kernel length increases, it gets a little worse. Second, unlike original paper, non-overlap kernel structure which has small length compared to dilation is not so good. Between fully overlap (~=fully connected) and almost non-overlap (~=small length kernel compared to dilation), the best results were obtained.

We also study some case using MFCC input. But there is no particular performance difference. Outputs are represented in Table2.

Structure	Input len.	Max dilated len.	Kernel len.	Accuracy
Baseline	5	-	-	74.8%
Fully-connected	4	-	4	75.0%
Fully-connected	12	-	12	75.0%
Dilated causal	8	4	4	74.5%
Dilated causal	8	4	8	73.7%

Table 2: The output of the experiments using MFCC.

4. CONCLUSIONS

Our work shows many possible structure based on CNN for ASC task. In many approaches, we reach 75.9% accuracy on the DACSE 2017 development dataset. In many approaches and many trials, we observed the robustness of our structure. We believe that our proposed architecture is robust and is good performance in evaluation dataset.

5. REFERENCES

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.

[2] Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S., 2017. Selfnormalizing neural networks. arXiv preprint arXiv:1706.02515.

[3] Kim and Park. Speech-to-Text-WaveNet. 2016. GitHub repository. https://github.com/buriburisuri/.