MODEL OF B-CNN AND WAVENET IN TASK 2

Technical Report

Zhesong Yu

Peking University
Institute of computer science & technology
Beijing, China
yzs@pku.edu.cn

ABSTRACT

In DCASE 2018 task 2, I attempted to use the WaveNet model in raw audio and Bilinear-CNN in MFCC to solve the classification. I trained the two model independently, finally ensemble the result of two model and get the mAP 91.9%.

Index Terms— B-CNN, WaveNet

1. INTRODUCTION

The system has two parts, one is BCNN in MFCC, another is WaveNet in raw audio. B-CNN is a model proposed to solve Image fine classification task, and WaveNet is a model used to music generation. And the propose of this paper is just to see the performance of the two model in music classification task.

2. DATA PROCESSING

In WaveNet model, we need the raw audio, and in B-CNN model, we need the processed MFCC.

Raw audio: sample the raw audio in 16kHz, and trim the silence part in top and head of the audio. Then I randomly cut about 39k samples (the number of samples fit the WaveNet model) about 2.5 seconds. If there is not enough samples, I just pad zeros ahead. Because raw audio is typically stored as a sequence of 16-bit integer values, so a model need to output 65536 probabilities per time step to model all possible values. To make this more tractable, we apply a μ-law companding transformation to the data, and then quantize it to 256 possible values:

$$f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

where -1 < x_t < 1 and $\mu = 255$.

MFCC : sample the raw audio in 44.1kHz, and trim the silence part in top and head of the audio. Finally get MFCC from 'librosa', which frame length is 512. To fit the B-CNN, we firstly get 60x180 MFCC, 60 is the dim of MFCC and 180 is the number of frame. 180 frames is around 2 second. Then we cut 60x180 into 3 parts, and stack them, finally we get MFCC of 3x60x60 as the input features. We did it to make it more like RGB pictures. What's more, we also get 3x80x80 and 3x100x100 MFCC as inputs.

Because there are lots of train data unverified, so I use a simple CNN model to verified the train data. Firstly, I take the verified data as train dataset to train a model, then I use the model to judge the other train data. I label the unverified data which is predicted right in top 3 as verified, and expand the train dataset. Do it over and over again, until there are less than 10 audios predicted right in top3. Finally, I take 8337 audio as train set.

3. WAVENET IN RAW AUDIO

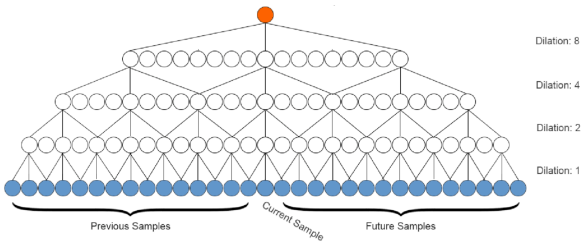
WaveNet is a powerful generative approach to probabilistic modeling of raw audio. When applied to text-to-speech, it yields state-of-the-art performance. Although WaveNet is designed to solve the music generation problem, it can also solve problems such as speech denoising, source separation and so on. So I attempt to use it in our task.

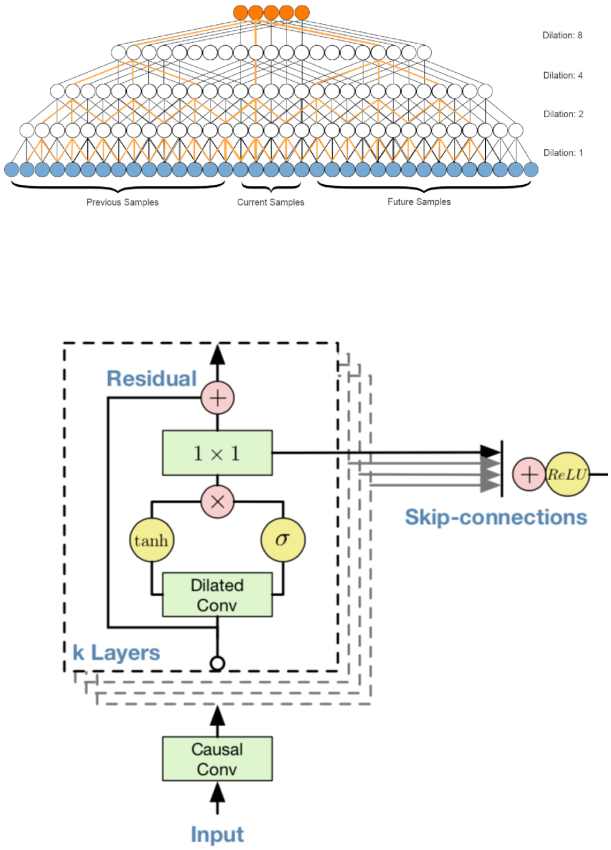
The model we used is a little different with the original WaveNet. We use non-causality structure, and increase the filter length to 3 and perform symmetric padding at each dilated layer. And we have 4 stacks, each stack has 12 layers, so the receptive field of every output is 32761 samples, around 2s.

Then we design the audio input length is 38910, so it generate (38910 – 32761 + 1) 6150 vectors every time. Then we add an average pool layer and some convolution layers to classify it to 41 classes(task2 classes number).

The model training is a hard work. We took 10% from train set as the valid set, and used 4 GPU, and it cost almost 30 hours to fit the dataset well. And it's accuracy is 67% in valid set. Because of the time it cost, we only trained one WaveNet model.

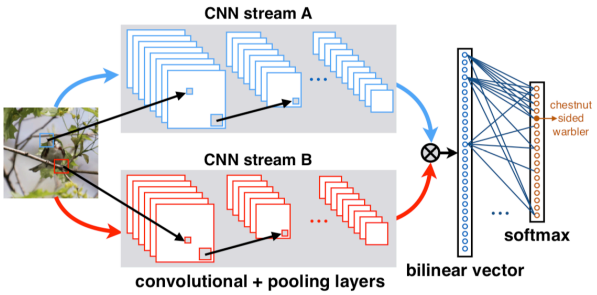
Some details of WaveNet are shown in the figures below.





4. BILINEAR-CNN IN MFCC

Bilinear CNN is designed for image fine classification task. As the figure shown below, an image is passed through CNNs A and B, and their outputs at each location are combined using the matrix outer product and average pooled to obtain the bilinear feature representation. This is passed through a linear and softmax layer to obtain class predictions.



In our model, we fully shared the parameters of CNN A and CNN B, it also means that we just use one CNN and use the matrix outer product to CNN's output with its transposition. We use ResNet34 as convolutional. The shape of input data is 3x60x60, after ResNet34 which cut out the pooling layer and fully connection layers, we get the shape of [512 x 2 x 2], then we reshape it to [512x4] and matrix outer product to get [512x512]. Finally, we add a fully connection layer to get 41 classes.

And then, I change the shape of the input data to 3x80x80 and 3x100x100 to train other models. Finally we got 3 MFCC model.

5. ENSEMBLE

In WaveNet model, we only trained one model, and I just simply duplicated its prediction 10 times. In B-CNN, we used 10-flod cross valid, and we designed two scales model (60x60 and 80x80), so we got 30 B-CNN predictions(each model has 10 flod), each got 76% accuracy on valid set. Finally, we ensemble the 40 models to predict the test set, and got mAP 91.9% in the end.

6. REFERENCES

[1] Van Den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio[C]//SSW. 2016: 125.

[2] Lin T Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1449-1457.

[3] Rethage D, Pons J, Serra X. A Wavenet for speech denoising[J]. arXiv preprint arXiv:1706.07162, 2017.