A SYSTEM FOR DCASE CHALLENGE USING 2018 CRNN WITH MEL FEATURES

Technical Report

Hanyu Zhang

Beijing University of
Posts and Telecommunications
zhanghanyu94@bupt.edu.cn

Shengchen Li

Beijing University of
Posts and Telecommunications.
Shengchen.li@bupt.edu.cn

ABSTRACT

For the Acoustic Scene Classification task (General-purpose audio tagging of Freesound content with AudioSet labels) of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2018), we propose a method to classify 41 different acoustic events using a Convolutional Recurrent Neural Network (CRNN) with log Mel spectrogram. First, the waveform of the audio recordings is transformed to log Mel spectrogram and MFCC. The convolutional layers are then applied on the log Mel spectrogram and Mel-frequency cepstral coefficients to extract high level features. The features are fed into the Recurrent Neural Network (RNN) for classification. On the official development set of the challenge, the best MAP is 0.8613, which increases 6.83% compared with the baseline.

1. THE PROPOSED SYSTEM

1.1. Setup

The task employs a subset of Freesound platform [1] annotated by a vocabulary of 41 labels from Google's AudioSet Ontology [2]. The subset covering a wide range of human and animal sounds, musical instruments and common everyday environmental sounds. State-of-the-art audio classification methods usually transform the waveform to the time-frequency representation. Mel-frequency cepstral coefficients and Log-Mel filter banks are both he most commonly used time-frequency representation in audio classification, so they are used as features in our system. Each chunk has a fixed length by padding or truncating. In the proposed system, we use three gated convolutional recurrent neural network blocks (dropout in each layer is 0.3), which with learnable gated linear unit's non-linearity applied on the features, and the feed-forward neural network has 41 output nodes corresponding to each audio event class.

1.2. Results

The experimental results on development dataset are shown in Table 1. We observe that the MAP in both using log Mel spectrogram and MFCC as the input are significant improvement compared with the baseline which use log Mel spectrogram convolutional recurrent neural network (three layers) 0.793 [3]. The best MAP is observed when using log Mel spectrogram. Table 2 shows the MAP of the different event for the log Mel features with the best MAP 0.8613 and the MFCC features with the MAP 0.8342. It can be seen from the table that the MAP of some events is very low. This is because the audio duration of these events is too short or too long, and a lot of information is lost when features are padding or truncating. Using MFCC as input can alleviate this problem to a certain extent. Later we will find a method to combine the Log Mel and MFCC's advantage to improve the system. The MAP of our approach on evaluation dataset is 0.802, outperforming the baseline 0.704.

Table 1: Performance of the different input features.

Input feature	MAP
log Mel	0.8613
MFCC	0.8342
log Mel +MFCC	0.791

Table 2: The MAP of different event of the development set for the proposed system with log Mel and MFCC.

Results with log Mel

Acoustic	0.8916	Gunshot_or_gunfire	0.5639
Applause	0.9657	Harmonica	0.9287
Bark	0.8707	Hi-hat	0.9051
Bass_drum	0.8213	Keys_jangling	0.8852
Burping_or_eructation	0.8403	Knock	0.9141
Bus	0.9444	Laughter	0.9327
Cello	0.9564	Meow	0.695
Chime	0.7814	Microwave_oven	0.5751
Clarinet	0.9612	Oboe	0.9463
Computer_keyboard	0.9416	Saxophone	0.9215
Cough	0.8143	Scissors	0.4178
Cowbell	0.8027	Shatter	0.9103
Double_bass	0.8991	Snare_drum	0.732
Drawer_open_or_close	0.6641	Squeak	0.642
Electric_piano	0.8772	Tambourine	0.9277
Fart	0.8726	Tearing	0.9543
Finger_snapping	0.6006	Telephone	0.6395
Fireworks	0.8656	Trumpet	0.9117
Flute	0.8448	Violin_or_fiddle	0.9281
Glockenspiel	0.878	Writing	0.8831
Gong	0.8839		

Results with MFCC

Acoustic	0.9073	Gunshot_or_gunfire	0.5247
Applause	0.9273	Harmonica	0.868
Bark	0.8396	Hi-hat	0.7653
Bass_drum	0.6732	Keys_jangling	0.8881
Burping_or_eructation	0.7535	Knock	0.8708
Bus	0.9714	Laughter	0.9481
Cello	0.9518	Meow	0.9436

Chime	0.9259	Microwave_oven	0.4363
Clarinet	0.9714	Oboe	0.9629
Computer_keyboard	0.9357	Saxophone	0.9215
Cough	0.7853	Scissors	0.519
Cowbell	0.9402	Shatter	0.7353
Double_bass	0.7613	Snare_drum	0.8047
Drawer_open_or_close	0.3171	Squeak	0.8202
Electric_piano	0.8647	Tambourine	0.8568
Fart	0.7631	Tearing	0.6381
Finger_snapping	0.9214	Telephone	0.8603
Fireworks	0.8656	Trumpet	0.8297
Flute	0.8448	Violin_or_fiddle	0.9026
Glockenspiel	0.9724	Writing	0.8342
Gong	0.9601		

2. REFERENCES

[1] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), pp 486-493. Suzhou, China, 2017.

[2] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in ICASSP, 2017.C. D. Jones, A. B. Smith, and E. F. Roberts, “A sample paper in conference proceedings,” in Proc. IEEE ICASSP, 2003, vol. II, pp. 803-806.

[3] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. General-purpose tagging of freesound audio with audioset labels: task description, dataset, and baseline. Submitted to DCASE2018 Workshop, 2018.