CNN BASED SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

Sangwon Lee, Seungtae Kang, Gil-Jin Jang

Kyoungpook National University
School of Electronics Engineering
80 Daehakro, Bukgu, Daegu, Korea
lsw0767@knu.ac.kr, cdef3456@naver.com, gjang@knu.ac.kr

ABSTRACT

Convolution neural networks (CNNs) have achieved great successes in many machine learning tasks such as classifying visual objects or various audio sounds. In this report, we describe our system implementation for acoustic scene classification task of DCASE 2018 based on CNN. The classification accuracies of the proposed system are 72.4% and 75.5% on development and leaderboard datasets, respectively.

Index Terms— Acoustic Scene Classification, Convolution Neural Networks, Dropout, Ensemble learning.

1. INTRODUCTION

These days, recurrent neural networks (RNNs) are generally employed in audio classification tasks such as audio tagging, acoustic scene classification, and sound event detection [4, 5]. Because RNN models are actively models temporal relationships of the given observations, they are by nature effective for learning sequential data. However, the uncertainties or observation errors of the input data are accumulated over time in the RNN forward process, RNNs often become unstable and hard to learn reliable models. On the contrary, convolutional neural networks (CNNs) use kernel windows of fixed sizes, so that it removes long-term, accumulated effects but local properties of the input data. Therefore, we built CNN based system for acoustic scene classification (ASC) in task 1A of detection and classification of acoustic scenes and events. We used shallow and wide CNN models with our novel methods.

This paper is organized as follows: we briefly describe the general idea and implementation details of our novel methods in Section 2. In Section 3, the classification experimental results on DCASE 2018 [1] are shown, followed by conclusion of the report in Section 4.

2. OUR SUBMISSION

2.1. Dropout

Dropout was proposed only for fully connected layers to substitute node outputs with zeros, and the selection is done randomly and independently with nodes [6]. In this report, we applied dropout to convolution kernel weights to be suited to convolution layers. The convolution kernel window coefficients are dropped (zero substitution) with given keep probability.

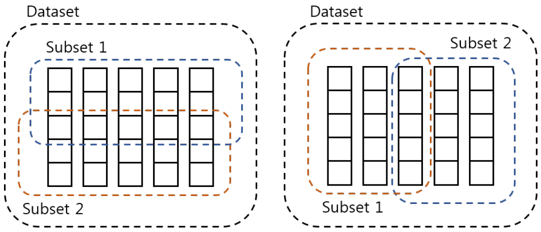
2.2. Partial Label Ensemble

Ensemble method [3] is a well-known technique to combine different model outputs to achieve improve performance from the individual model outputs. Majority rule is one of popular ones, which selects alternatives which have the largest number of classification labels. With this rule, system can avoid misprediction from the minority or outliers. To apply the majority rule, each model should be trained with different datasets to ensure the independence among the trained models. Cross-validation is one way to do that, by splitting dataset into data-wise evenly and selecting some parts of them. Cross-validation requires all of original classes, so when the amount of data is too small for a certain label due to the data splitting, it may lead model underfitting. To avoid problem, we split the dataset class-wisely with the same number of classes per subset. The proposed cross-validation policy is described in Figure 1.

3. EXPERIMENTAL SETUP

3.1. Feature Extraction

We first converted the audio sound to mono channel, and extracted log-scale, mel-band energies, which is identical



(a) data-wise (b) class-wise

Figure 1: Data split policy for the proposed ensemble learning. Each column represents a class, and each row represents sample data.

to the baseline system. The temporal width and stride of the convolution kernel window 0.08 ms and 0.4 ms, respectively. The number of extracted band energies is 250, followed by a zero padding along time axis to make a squared feature shape. As a result, we use 250x250 spectral image as an input to the network.

3.2. Model Description

Layer Name	Kernel Size	Output Shape
Input	-	(250, 250, 1)
Conv	11x11	(250, 250, 64)
Maxpool	2x2	(125, 125, 64)
Conv	9x9	(125, 125, 128)
Maxpool	5x5	(25, 25, 128)
Conv x 2	5x5	(25, 25, 256)
Maxpool	5x5	(5, 5, 256)
Conv x 2	3x3	(5, 5, 512)
Avgpool	5x5	(1, 1, 512)
Dense	-	(512)
Dropout	-	(512)
Dense	-	(10)

Table 1: CNN structure: on every Conv layer, batch normalization, relu activation, and drop are applied.

Our system is shallow and wide CNN which has 6 more layers than baseline system. Batch normalization and l_2 normalization are employed as a preprocessing. ReLU activation function [9] is applied right after every batch normalization. The combination of cross entropy and scaled mean squared error is used for loss function, and we use momentum algorithm to minimize loss function. The model is trained for 200 epochs, with 0.0125 initial learning rate with batch size 16. During training, learning rate is divided by 10 at every 80 epochs. Table 1 shows our CNN structure.

3.3. Evaluation Results

Dataset	Model Name	Accuracy
Dev fold1	Baseline	59.7%
	Dropout	72.4%
leaderboard	Dropout	74.3%
	Dropout_Ens	74.0%
	Dropout_PL_Ens	75.5%

Table 2: Classification accuracies of the development and leaderboard sets with dropout, partial label ensemble method. 'Dropout_Ens' is applying dropout and Ensemble learning, and 'Dropout_PL_Ens' is applying the proposed partial label ensemble learning.

Table 2 shows the classification accuracies on the provided development dataset fold1 and leaderboard dataset. By applying dropout, the accuracy was greatly improved by 6.6%. According to the results in Table 2, first two rows, we choose dropout for leaderboard dataset experiments. The bottom two compare the conventional Ensemble method with the proposed partial label ensemble method. The proposed ensemble method also improves the performance by 1.5% as well. The final performance of the proposed method was 75.5%.

4. CONCLUSION

In this paper, we describe our CNN based system for task 1A of DCASE 2018. We applied our novel methods, and achieved accuracies 72.4% and 75.5% respectively on development and leaderboard dataset. This is significant improvement compared to baseline system which achieved accuracy of 62.5% on leaderboard dataset. Moreover, we did not use any data augmentation method. So, with data augmentation, the performance is expected to be further improved.

Acknowledgments

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding.

5. REFERENCES

[1] `http://dcase.community/challenge2018/`.

[2] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” J. Mach. Learn. Res., 15(1):1929–1958, January 2014.

[3] Deng, Li and Platt, John C., “Ensemble deep learning for speech recognition,” In INTERSPEECH 2014, 1915–1919.

[4] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E, “Imagenet classification with deep convolutional neural networks,” In Advances in neural information processing systems, pp. 1097–1105, 2012.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CoRR, abs/1409.1556, 2014.

[6] S. Zagoruyko and N. Komodakis, “Wide residual networks,” arXiv preprint arXiv:1605.07146, 2016.

[7] Sergey Ioffe and Christian Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” In ICML 2015.

[8] Martin Abadi, Ashish Agarwal, Paul Barham and many, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” arXiv preprint arXiv:1603.04467, 2016.

[9] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, 2010.