Acoustic Scene Classification Using Reduced MobileNet Architecture

Jun-Xiang Xu[1], *Tzu-Ching Lin*[1], *Tsai-Ching Yu*[1], *Tzu-Chiang Tai*[2], *and Pao-Chi Chang*[1]
[1]Department of Communication Engineering, National Central University, Jhongli, Taiwan
[2]Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan

Abstract—Sounds are ubiquitous in our daily lives, for instance, sounds of vehicles or sounds of conversations between people. Therefore, it is easy to collect all these soundtracks and categorize them into different groups. By doing so, we can use these assets to recognize the scene. Acoustic scene classification allows us to do so by training our machine which can further be installed on devices such as smartphones. This provides people with convenience which improves our lives. Our goal is to maximize our validation rate of our machine learning results and also optimize our usage of hardware. We utilize the dataset from IEEE Detection and Classification of Acoustic Scenes and Events (DCASE) to train our machine. The data of DCASE 2017 contains 15 different kinds of outdoor audio recordings, including beach, bus, restaurant etc. In this work, we use two different types of signal processing techniques which are Log Mel and HPSS (Harmonic-Percussive Sound Separation). Next we modify and reduce the MobileNet structure to train our dataset. We also make use of fine-tuning and late fusion to make our results more accurate and to improve our performances. With the structure aforementioned, we succeed in reaching the validation rate of 75.99% which is approximately the seventh highest performing group of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge 2017, with less computational complexity comparing with others having higher accuracy. We deem it a worthy trade-off.

Keywords—DCASE 2017, MobileNet, seventh highest.

I. Introduction

Acoustic scene classification has caught people's attention these years. Aside from the traditional image recognition, soundtracks provide more detail information than pictures do. Our motivation is that even though there are already many kinds of models, we can still attempt to integrate different combinations to optimize our performances. By trying various feature extraction techniques and structures, such as VGG or ResNet, we finally determine the MobileNet model to achieve our work.

There are many possible or existing applications about Acoustic Scene Classification such as security systems and monitoring applications. Different from speech generated from human beings, the wider range of environmental sounds are more challenging than the former ones. The reason is that there are more unexpected and unrecognizable trifling sounds in our daily living environments. In this work we use the label existed in the previous trained data to improve our performance on classifying environmental sounds. Furthermore, we cited the concept of arithmetic ensemble, most using in the competition as a key of success, which could help us obtain the best conclusion being calculated from all the experimental results through calculating.

II. Related Technologies

According to the paper which ranked the third place of DCASE 2017 [1], the overall structure is linked two different architectures. Firstly, one of the models is inputted single mel-spectrogram after the background subtraction processes. On the other hand, the other is used for paired input such as left-right (LR), mid-side (MS), and harmonic-percussive sound separation (HPSS). The two individual channels are processed with the same convolutional block which contains traditional CNN structure. Finally, the authors combine the two similar models before the last fully-connected layer. Afterwards, the paper adopts the ensemble method of iteration which aims to find the optimal weight. Consequently, we are motivated by the strategies and architecture mentioned above. So we come up with our own model which adopt the MobileNet with Log Mel and HPSS feature extractions.

A. Data Augmentation

First of all, data augmentation techniques for deep learning, specific to images, are widely used to increase dataset size via transformations. However, in the paper [1], compared to conventional training process, the proposed method achieves more significant results. By applying each augmentation, the accuracy for distinct class have improved individually.

The data augmentation of audio soundtracks differs from that of images which containing time stretching, pitch shifting and adding random noise. With the use of these techniques, we can effectively increase our training validation rate and make our model perform more steadily.

B. MobileNet

In our paper, we utilize the advantage of MobileNet, which is based on a streamlined architecture that uses depthwise separable convolutions [3]. Depthwise separable convolution have become popular in DNN models recently, for two reasons. First of all, they have fewer parameters than regular convolutional layer, and thus are less prone to overfitting. Secondly, with fewer parameters, they also require less operations to compute, making it cheaper and faster. According to the paper [4], MobileNets are able to operate effectively

978-1-5386-6857-3/18/$31.00 ©2018 IEEE
DOI 10.1109/ISM.2018.00038

across a wide range of applications by trading off a reasonable amount of accuracy to reduce size and latency, Therefore, in our work, we are motivated to try the MobileNet model on acoustic scene classification.

C. Late Fusion

Late fusion solves the problem of different prediction results generated from classifiers which each trained with a specific feature [5]. The basic approach to late fusion is to estimate a fixed weight for each classifier and then use a weighted summation of the prediction scores as the fusion result. It is inappropriate to hypothesize that classifiers have same prediction capabilities on different samples. Thus, in order to mitigate prediction errors, it is necessary to estimate the fusion weights for each sample rather than using fixed weights.

Ensemble methods play an important role in late fusion, which contains building a set of classifiers and then categorize new data by taking a vote of their own predictions [6]. There are many ensemble approaches, including the original Bayesian averaging, Bagging and boosting etc.

III. The Proposed Architecture

A. Feature Extraction

In our work, we utilize two different kinds of feature extraction techniques in order to improve performances.

1. Log Mel

   Mel scale [7] is a frequency-binning method based on the human ear' s frequency resolution. The Mel scale tends to imitate the human ear in terms of the manner with which frequencies are sensed and resolved, which human beings are more sensitive to the difference between frequencies in low pitches. By changing our raw audio data into spectrogram with a Y- axis of mel scale, we can use the advantage of detail information provided by spectrograms which varies with time. Next, we are able to send our spectrograms into our MobileNet structure.

2. HPSS (Harmonic-Percussive Sound Seperation)

   By decomposing sounds, we can get two different components: harmonic and percussive [2]. HPSS algorithms aims to separate drum sounds from mixture music. From the paper [8], we acknowledged that it is possible to use a simple and fast algorithm specifically for the harmonic/percussive separation based on the anisotropy of them on spectrograms. In our work, we use the Python code presented in Librosa to decompose our dataset and also turn them into spectrograms.

B. Data augmentation

In the paper [2], we realize that there are varieties of methods to accomplish data augmentation of sounds. The followings are the techniques we conduct in our work. These are completed before we transfer our dataset into spectrograms.

- **Random Noise**: We add random Gaussian distribution nosies to original soundtracks in order to enlarge our dataset.
- **Time Stretching**: There are two ways: slowing down or speeding up. We speed up our audio files 1.2 times faster than originals while keeping the same pitch.
- **Pitch Shifting**: We lower our pitches and keep the duration unchanged. Therefore, the audio samples are pitch shifted by {-1, -2} (in semitones).
- **Time Shifting**: We delay our soundtrack then cut and put the segment which is beyond the time interval in the beginning of the sound.

C. MobileNet

We have come up with the architecture based on MobileNet structure that contains a parallel structure including two different feature extraction results which is Log Mel and HPSS respectively.

D. Fine-Tuning

By replacing and retraining the classifier on top of the ConvNet, and also fine-tuning the weights of the pre-trained network via back propagation, we are able to take advantage of the pre-trained weights done by others. We initially froze the upper layer but later found out that it did not result in better performances. Therefore, we then cut out 9 layers of the original MobileNet model and trained them with a stochastic gradient descent (SGD) optimizer [9], which we lowered the learning rate to improve our performance.

E. Late Fusion

Because there are two different results from Log Mel and HPSS, we transformed each classification result into a 15-dimension array, then adopt the mean ensemble strategy which averages two model prediction probability and obtains a more reliable result. The reason why we did not use the dynamic fraction of late fusion is that we figured out the difference between it and the arithmetic mean method is slight and even no comparison. Thereby, we decided to set equal probabilities respectively. By doing so, the outcome of our structure will be evenly distributed without over-relying on a particular result.

F. Overall Structure

The overall architecture is illustrated in Fig.1 which is composed of two different feature extraction techniques. Three MobileNet models are trained individually with different preprocessing methods. Next, these prediction scores are ensembled before calculating to the probabilities of each detected scenes. Figure 2 shows the MobileNet blocks, where the middle one indicates the structure inside each MobileNet layer. Besides, the left and right demonstrate two models with distinctive preprocessing strategies. We use the standard model of MobileNet in [3], which includes the details demonstrated in Fig.2. The middle block in Fig.2 clearly indicates that we employed batch normalization (BN) [10] which normalize the output of previous convolution layer with additional shifting or scaling and followed by rectified linear unit (ReLU) after both depthwise convolution and pointwise convolution. We did not adopt an inverse arrangement of convolution layer and activation function because we are not sure about the performance of the pre-activation concept proposed in residual network [11] will operate well with MobileNet models.

After introducing inside of our MobileNet models, we move on to the left and right hand side of the blocks in Fig. 2. In the right block, we proposed two 9-layer MobileNets with HPSS. After training, the two results concatenated and been reshaped

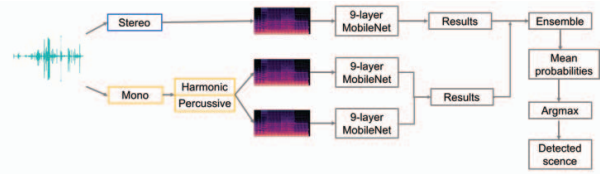

Fig. 1. Overall structure of the proposed architecture.

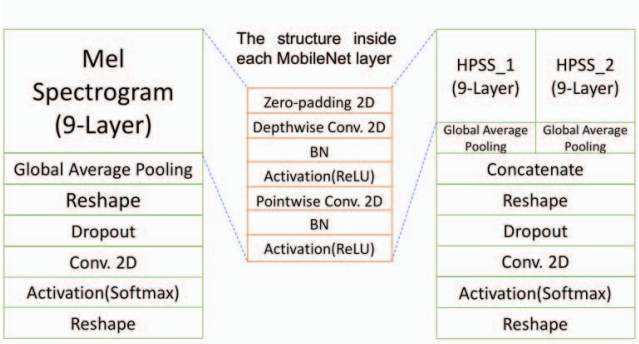

Fig. 2. MobileNet blocks.

into a 1024-dimension vector. Afterward, we added a dropout function with a parameter 0.5, which is able to avoid overfitting effectively. Then we applied a convolutional layer to be our classifier and the results were reshaped again into a 15-dimension array which represents fifteen different probabilities of scenes in our dataset. The left block in Fig.2 also indicates a parallel model but with feature extraction Log Mel instead. Next, according to Fig.1, we ensemble the two results. An ensemble is a combination of models whose predictions are integrated by different mechanism [12]. Therefore, by combining two outcomes from each structure (aforementioned in 3.5), we are able to acquire prediction results more accurately and credibly.

IV. Experiments

In our experiments, we determined the parameter of batch size as 8 because we found it is more suitable than the number 16 or 32 in this task. This is because larger batch size could not lead to a better accuracy through stochastic gradient descent (SGD) optimizer. We gradually declined our learning rate from 0.01 and figured out that our proposed structure resulted in a satisfactory accuracy rate when it is 0.001. The reason is that high learning rates would ruin the pre-trained weights easily. Although the number of epoch while training will increase, the performance of low learning rate is better significantly.

Fig.3 and Fig.4 showed each scene validation accuracy rate of HPSS and Log Mel features respectively. It is obvious that the model with Log Mel has a higher performance than that with HPSS in most scenes, except for residential and tram sounds. The total rate of each preprocessing method is 68.6% and 72.59% respectively. It is also worth noticing that in some places, the accuracy rate is relatively low even though we had utilized different feature extraction techniques. For instance, the validation accuracy rates of the scene library are the lowest in each figures. This is explicable because there are no obvious features in the library which are distinguishable to machines, and some different noises make our model problematic. Thus, it is difficult to precisely make correct predictions. Besides, the combination of sounds generated by both human beings and living creatures in the park has also influenced the extracted features which therefore resulted in inaccurate predictions. Judging from the factors aforementioned, we look forward to taking advantage of Generative Adversarial Network which can make our model more robust and is able to maintain great performance even the features of the training data cannot be resolved efficiently.

In addition, we attempted to examine the effectiveness of late fusion. We found the accuracy of late fusion in Fig.5 had significantly increased as well as the rates in all kinds of scenes and it also achieved an entire result of 75.99%.

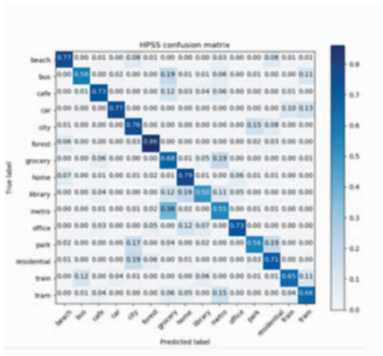

Fig. 3. Confusion matrix of HPSS.

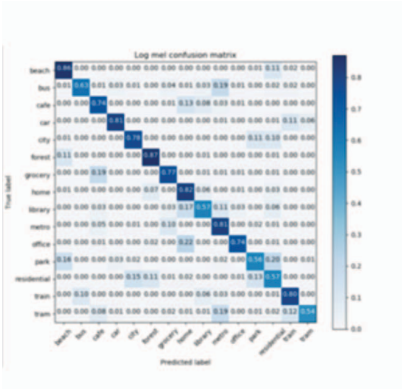

Fig. 4. Confusion matrix of log mel.

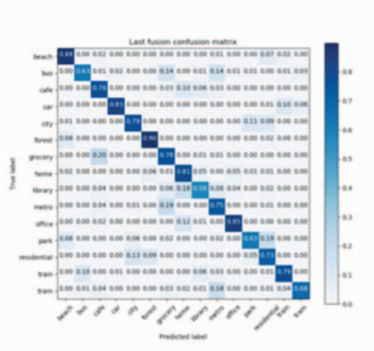

Fig. 5. Confusion matrix of late fusion.

Even we got a higher rate after finishing data augmentation, the accuracy was still beneath our satisfaction. This phenomenon results from the fact that more layers in a model does not always lead to better performance. The reason is that when there are lots of layers in a deep learning network, the gradients adjacent to the input layer cannot be updated effectively, which is known as vanishing gradient. Although it is able to be solved by adopting ReLU activation function, the results are not always contented. Therefore, we then attempted to reduce layers from the original MobileNet model and have come up with a best accuracy rate when there are nine layers remaining in our proposed structure.

TABLE I. LOG MEL FEATURE EXTRACTION VALIDATION ACCURACY.

Model / Data	Original model	Five layers	Seven layers	Nine layers
Original data(average)	≈ 66%	≈ 64%	≈ 68%	≈ 71%
After data augmentation	≈ 66%, steady	≈ 65%, steady	≈ 70%, steady	≈ 72%, steady

TABLE II. HPSS FEATURE EXTRACTION VALIDATION ACCURACY.

Model / Data	Original model	Five layers	Seven layers	Nine layers
Original data(average)	≈ 66%	≈ 60%	≈ 65%	≈ 67%
After data augmentation	≈ 67%, steady	≈ 61%, steady	≈ 66%, steady	≈ 68%, steady

V. CONCLUSION

According to our experiment results, we have finally come up with a parallel MobileNet model which performs the best when they are both reduced to 9 layers in each of the feature extraction structure (HPSS and Log Mel). Also, we made effort to gradually change our parameters and found the optimization of them. Thus, according to the details elaborated above, we finally succeed in elevating our validation accuracy rates. In the outlook, we would like to figure GAN to generate more training data for the purpose of stabilizing our models and gaining better achievement.

REFERENCES

[1] Yoon chang Han, Jeong soo Park, Kyogu Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification", Detection and Classification of Acoustic Scenes and Events 2017.

[2] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification", IEEE Signal Processing Letters, Nov,2016.

[3] L. Sifre. "Rigid-motion scattering for image classification". PhD thesis, Ph. D. thesis, 2014.

[4] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto and Hartwig Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications", Google Inc., Apr, 2017.

[5] Dong Liu, Kuan-Ting Lai, Guangnan Ye, Ming-Syan Chen, Shih-Fu Chang, "Sample-specific late fusion for visual category recognition", 2013.

[6] T.G. Dietterich, "Ensemble methods in machine learning", 1st Int. Workshop on Multiple Classifier Systems, vol. 1857, pp. 1-15, 2000.

[7] Chin Kim On, Paulraj M. Pandiyan, Sazali Yaacob, Azali Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition", 2006 International Conference on Computing & Informatics, Pages: 1 - 5, 2006.

[8] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in Signal Processing Conference, 2008 16th European. IEEE, Pages: 1–4, 2008.

[9] Guojing Cong, Onkar Bhardwaj, "A Hierarchical, bulk-synchronous stochastic gradient descent algorithm for deep-learning applications on GPU clusters", 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pages: 818- 821,2017

[10] Vignesh Thakkar, Suman Tewary, Chandan Chakraborty, "Batch normalization in convolutional neural networks — a comparative study with cifar-10 data", 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), pages: 1-5, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks", 2016.

[12] Dai Wei, Juncheng Li, Phuong Pham, Samarjit Das, Shuhui Qu, "Acoustic scene recognition with deep neural networks (DCASE challenge 2016)", Detection and Classification of Acoustic Scenes and Events 2016, Sep, 2016.