SE-RESNET WITH GAN-BASED DATA AUGMENTATION APPLIED TO ACOUSTIC SCENE CLASSIFICATION

Technical Report

Jeong Hyeon Yang, Nam Kyun Kim, and Hong Kook Kim[*]

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Gwangju 61005, Korea
{didrn43, skarbs001, hongkook}@gist.ac.kr

ABSTRACT

This report describes our contribution to the development of audio scene classification methods for the DCASE 2018 Challenge Task 1A. The proposed systems for this task are based on data augmentation through generative adversarial network (GAN)-based data augmentation and various convolutional networks such as residual networks (ResNets) and squeeze-and-excitation residual networks (SE-ResNets). In addition to data augmentation, SE-ResNets are revised so that they operate on the log-mel spectrogram domain, and the numbers of layers and kernels are adjusted to provide better performance on the task. Finally, the ensemble method is applied using a four-fold cross-validated training dataset. Consequently, the proposed audio scene classification system improves classwise accuracy by 10% compared to the baseline system through the Kaggle competition in acoustic scene classification.

Index Terms—Acoustic scene classification, generative adversarial net, data augmentation, squeeze-and-excitation residual network

1. INTRODUCTION

Acoustic scene classification (ASC) is a task of classifying an environment using recorded audio signals. This task can both support human cognitive abilities and assist in determining situations where visual information is unavailable. In addition, ASC can help to detect an acoustic event in noisy environments by removing noise classified to a specific environment because it is relatively easier to remove a known background noise rather than unknown noise. Image classification has been studied for a long time, and the use of neural networks has been improving classification accuracy each year through competition. Thus, we first try to apply networks that have performed well at ImageNet classification to audio classification. There have been several attempts to use neural networks for ASC. A convolutional neural network (CNN)-based approach was proposed for music classification [1]. In addition, the CNN with batch normalization [2] was applied to ASC [3]. It is known that a neural network with many layers shows superior performance as the amount of data increases, but if a network is deeper, the gradient exploding of the back propagation becomes a problem [4]. However, the residual neural network (ResNet) [5] solved this problem through an identity skip connection. Furthermore, squeeze-and-excitation blocks are proposed to improve the overall performance of ResNets [6].

We are interested in the deeper architecture of the neural network and inspired by deep residual network, we propose several neural networks based on ResNet and SE-ResNet for DCASE 2018 Challenge Task 1A. In order to mitigate the overfitting problem of neural networks as in this task, where the neural network should be trained for a given development dataset, data augmentation using the generative adversarial network (GAN) [7, 8] is introduced to diversify the training data. Specifically, the data augmentation is performed by modifying the structure of WaveGAN [9] to generate a direct wave rather than to generate a modified feature for training data. During this work, a wave generated by WaveGAN is evaluated by either an ResNet- or an SE-ResNet-based classifier, and then it is chosen as an augment data if it is correctly classified.

Following this Introduction, Section 2 describes the overall proposed ResNet-based system for ASC, and Section 3 describes the modified WaveGAN to diversifying the training data. After that, Section 4 evaluates the performance of the proposed system. Finally, Section 5 concludes this report.

2. PROPOSED RESNET-BASED ASC SYSTEM

This section describes the proposed ResNet-based ASC system applied to DACSE 2018 Challenge Task 1A. The proposed systems are constructed by modifying ResNet and SE-ResNet, as shown in Figs. 1(a) and 1(b), respectively. The following subsections give more description on feature extraction, residual blocks, SE blocks and model architectures.

2.1 Feature extraction

Each audio sample is divided into frames of 40 ms in length with a 20 ms overlap, where the sampling rate is set to 48 kHz. Instead of using stereo signal, the mono signal is obtained by averaging

[*] This work was supported in part by the Ministry of Trade, Industry & Energy (MOTIE, Korea), under Industrial Technology Innovation Program (No. 10063424, Development of distant speech recognition and multi-task dialog processing technologies for in-door conversational robots).

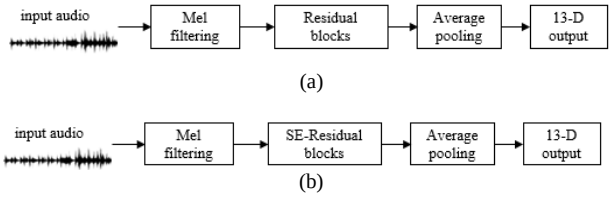

Figure 1: Overall architectures of (a) ResNet-rev and (b) SE-ResNet-rev.

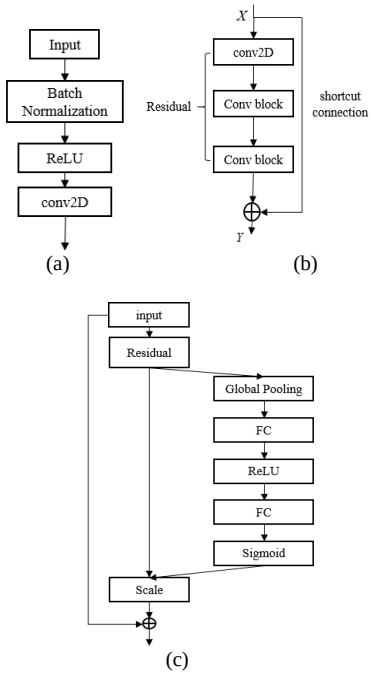

Figure 2: Detailed diagrams of blocks used in ResNet-rev and SE-ResNet-rev; (a) convolutional block, (b) residual block, and (c) residual with a squeeze-and-excitation block.

the left- and right-channel signals so that it contains the binaural information. After that, a 2,048-point short-time Fourier transform (STFT) is applied to each audio frame. Finally, 40-dimensional log-mel spectrum is extracted as input features for ASC. In addition, in order to generate waves by WaveGAN, audio signals are down-sampled into 16 kHz due to a trade-off between the efficiency and complexity of WaveGAN and then the generated wave of WaveGAN is up-sampled back into 48 kHz. The subsequent process is the same as above.

2.2 Convolutional and residual block

The convolutional and residual blocks in this report are shown in Figs. 2(a) and 2(b), respectively. While CNN-based networks have been popular because their performance is better than other architectures, they have problems in that they are hard to converge and have larger memory to train. Thus, ResNets have been proposed to train very deep CNNs [10]. ResNet is a block-wise stacked architecture of the same shape, and each block in ResNet

Table 1: Network architectures of the proposed ASC systems.

Network	ResNet(50)-rev	ResNet(152)-rev	SE-ResNet(152)-rev
Conv block	5×5, 32, stride (2, 2)		
Pooling layer	3×3, max pool, stride (2,2)		
Residual block 1	$\begin{bmatrix}3\times3 & 32\\ 3\times3 & 32\end{bmatrix}\times3$	$\begin{bmatrix}3\times3 & 32\\ 3\times3 & 32\end{bmatrix}\times3$	$\begin{bmatrix}3\times3 & 32\\ 3\times3 & 32\end{bmatrix}\times3$
Residual block 2	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times4$	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times8$	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times8$
Residual block 3	$\begin{bmatrix}3\times3 & 128\\ 3\times3 & 128\end{bmatrix}\times6$	$\begin{bmatrix}3\times3 & 128\\ 3\times3 & 128\end{bmatrix}\times36$	$\begin{bmatrix}3\times3 & 128\\ 3\times3 & 128\end{bmatrix}\times36$
Residual block 4	$\begin{bmatrix}3\times3 & 256\\ 3\times3 & 256\end{bmatrix}\times3$	$\begin{bmatrix}3\times3 & 256\\ 3\times3 & 256\end{bmatrix}\times3$	$\begin{bmatrix}3\times3 & 256\\ 3\times3 & 256\end{bmatrix}\times3$
Pooling layer	Average pool		

contains direct connections between the output of a lower layer and the inputs of a higher layer.

2.3 Squeeze-and-excitation block

The squeeze-and-excitation (SE) block is briefly depicted by the flow from global pooling to sigmoid and scaling in Fig. 2(c). The SE block reduces the dimensions with global pooling and expands through scaling.

2.4 Model architecture

Table 1 shows network architectures of the proposed different systems. In the table, $n\times n$ means the kernel size of a convolutional block, and the numbers such as 32, 64, 128, and 256 correspond to the number of kernels. Also, 'xx' in ResNet(xx)-rev and SE-ResNet(xx)-rev is the total number of convolution layers in the residual block repetitions and convolutional blocks. In addition, SE-ResNet(152)-rev also has the same structure as ResNet(152)-rev, but the SE-residual block shown in Fig 2(c) is substituted for the residual block shown in Fig. 2(b). Unless otherwise noted, the stride is 1. SE-ResNet(152)-rev can be trained using training data with augmented data through WaveGAN, which is referred to as SE-ResNet(152)-rev-aug.

2.5 Ensemble method

In order to improve the performance of ASC, we trained four individual model for each proposed system with 4-fold cross-validation from training set. Four models are linearly combined for ensemble classification.

3. DATA AUGMENTATION USING WAVEGAN

Tables 2 and 3 show the GAN-based wave generator and discriminator architectures, respectively. The latent dimension is set to 200 and both the input and generated output sizes are all set to 32,768, which results in wave segments of around two seconds long with a sampling rate of 16 kHz. Then, 5 wave segments are

Table 2: WaveGAN generator architecture.

Operation	Kernel Size	Output Shape
Input z~Uniform(-1,1)	(200,512d)	(n,200)
Dense 1		(n,512d)
Reshape		(n,32,16d)
ReLU		(n,32,16d)
Trans Conv1D (Stride=4)	(25,16d,8d)	(n,128,8d)
ReLU		(n,128,8d)
Trans Conv1D (Stride=4)	(25,8d,4d)	(n,512,4d)
ReLU		(n,512,4d)
Trans Conv1D (Stride=4)	(25,4d,2d)	(n,2048,2d)
ReLU		(n,2048,2d)
Trans Conv1D (Stride=4)	(25,2d,d)	(n,8192,d)
ReLU		(n,8192,d)
Trans Conv1D (Stride=4)	(25,d,c)	(n,32768,c)
Tanh		(n,32768,c)

Table 3: WaveGAN discriminator architecture.

Operation	Kernel Size	Output Shape
Input x or G(z)	(25,c,d)	(n,32768,c)
Conv1D (Stride=4)		(n,8192,d)
LReLU (α=0.2)		(n,8192,d)
Phase Shuffle(n=2)		(n,8192,d)
Conv1D (Stride=4)	(25,d,2d)	(n,2048,2d)
LReLU (α=0.2)		(n,2048,2d)
Phase Shuffle(n=2)		(n,2048,2d)
Conv1D (Stride=4)	(25,2d,4d)	(n,512,4d)
LReLU (α=0.2)		(n,512,4d)
Phase Shuffle(n=2)		(n,512,4d)
Conv1D (Stride=4)	(25,4d,8d)	(n,128,8d)
LReLU (α=0.2)		(n,128,8d)
Phase Shuffle(n=2)		(n,128,8d)
Conv1D (Stride=4)	(25,8d,16d)	(n,32,16d)
LReLU (α=0.2)		(n,32,16d)
Reshape		(n,512d)
Dense	(512d,1)	(n,1)

merged into a wave of 10 seconds long. Throughout this approach, a total of waves generated by WaveGAN is 1,728 and they are used for SE-ResNet(152)-rev-aug.

4. PERFORMANCE EVALUATION

4.1 Dataset

As a training set, 8,640 audio clips in total were provided for DCASE 2018 Challenge Task 1A, where there were 10 audio scenes such as airport, shopping mall, metro station, street, pedestrian public square, street traffic, tram, bus, metro, park [11]. Each scene had 864 audio clips and each audio clip of 10 seconds long was sampled at 48 kHz with 24-bit resolution in stereo. Also, there were 1,200 audio clips provided as a test set.

4.2 Experiment setting and ensemble method

Each of the proposed systems with different configurations, as shown in Table 1, was trained with the mini-batch ADAM optimization algorithm to minimize the categorical cross-entropy criterion. The training data was divided into 4 folds. Each fold was then used once as a validation, while the nine remaining folds were used for training. Finally, an ensemble classifier was obtained by linearly combining four models.

Table 4: Classwise accuracy with 50% of the test dataset for several ensemble models. Note that the baseline system is trained with the training subset (6122 segments) without considering 4-fold ensemble. Exceptionally, A feature extracted from 80-dimensional log-mel spectrum instead of 40 is used for training SE-ResNet(152)-rev*.

Model	Classwise Accuracy (%)
Baseline	62.50
ResNet(50)-rev	69.83
ResNet(152)-rev	70.33
SE-ResNet(152)-rev	71.12
SE-ResNet(152)-rev*	**72.50**
SE-ResNet(152)-rev-aug	70.50

4.3 Results and discussion

Table 4 compares the classwise accuracy of different configurations of the proposed ASC system. Note here that all the accuracies were measured on a four-ensemble model. As shown in the table, ResNet-rev improved classwise accuracy as the number of layers increased. In particular, ResNet(152)-rev gave relative classwise improvement of 12.5%, compared to the baseline [12]. Next, the classwise accuracy of SE-ResNet-rev was evaluated and it was higher than that of ResNet-rev when the number of layers was 152. Finally, the training data were augmented by using WaveGAN. Consequently, the classwise accuracy of SE-ResNet(152)-rev-aug was 70.5%. This was because the up-sampled wave could be less accurate and both the GAN-based generator and discriminator might be poorly trained.

5. CONCLUSION AND FUTURE WORK

This report proposed audio scene classification methods by using ResNet or SE-ResNet with GAN-based data augmentation, and they were applied to DCASE 2018 Challenge Task 1A. It was shown from the evaluation of the proposed methods on 50% of the test data that SE-ResNet(152)-rev* had classwise accuracy of 72.50% which was higher of 10% than that of the baseline. Throughout this work, we observed that WaveGAN could make further improvement but the processing time of WaveGAN was higher than we expected. Thus, GAN-based data augmentation is still working on and we are going to present the detailed results on this for other workshop or conference.

6. REFERENCES

[1] S. Dieleman and B. Schrauwen, “End-to-end learning for music audio,” in *Proc. IEEE ICASSP*, 2014, pp. 6964–6968.

[2] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proc. 32nd International Conference on International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[3] M. Valenti, S. Squartini, A. Diment, G. Parascandolo, and T. Virtanen, “A convolutional neural network approach for acoustic scene classification,” in *Proc. International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1547–1554.

[4] G. Philipp, D. Song, and J. G Carbonell, “Gradients explode – deep networks are shallow – ResNet explained,” arXiv:1712.05577, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” arXiv:1709.01507, 2017.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[8] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, “Deep neural network based learning and transferring mid-level audio features for acoustic scene classification,” in *Proc. IEEE ICASSP*, 2017, pp.796–800.

[9] C. Donahue, J. McAuley, and M. Puckette, “Synthesizing audio with generative adversarial networks,” arXiv:1802.04208, 2018.

[10] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “CNN architectures for large-scale audio classification,” arXiv:1609.09430, 2016.

[11] T. Heittola, A. Mesaros, and T. Virtanen, “TUT urban acoustic scenes 2018, development dataset,” http://doi.org/10.5281/zenodo.1228142.

[12] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, “General-purpose tagging of Freesound audio with audioset labels: task description, dataset, and baseline,” submitted to *DCASE 2018 Workshop*, arXiv:1807.09902, 2018.