Attention-based Resnet-18 model for Acoustic Scene Classification

Technical Report

Nisan Aryal

Gachon University
PRML Lab, 1342 Seongnamdaero, Sujeng-gu,
Gyeonggido 13120, Republic of Korea.
nisanaryal123@gmail.com

Sang Woong Lee †

Gachon University
PRML Lab, 1342 Seongnamdaero, Sujeng-gu,
Gyeonggido 13120, Republic of Korea.
slee@gachon.ac.kr

ABSTRACT

This technical report describes our approach to solve Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 challenge task1a. Resnet-18 with attention model and Openl3 embedding are used to solve the acoustic scene classification problem. The model shows 59.6% accuracy in the training and validation split of the development set, which is 5.5% higher than that of the baseline network.

Index Terms— Acoustic scene classification, Resnet, CBAM, Openl3 embedding

1. INTRODUCTION

DCASE challenge has introduced acoustic scene classification with mismatched devices since 2018 [1]. Task 1a of DCASE 2020 challenge expands previous challenges by introducing the dataset recorded using 4 different devices and 11 mobile devices (synthesis data based on the original recordings). This submission consists of a system for the classification of Acoustic Scene based on Openl3 embedding [2] and Resnet-18 [3] with the convolutional block attention module (CBAM) [4].

2. METHOD

2.1. Model

We use an attention-based Resnet-18 network as a classifier. CBAM is an attention module that sequentially infers attention maps along the channel and spatial dimensions, and the attention maps are multiplied to the input feature to get a refined feature.

Channel attention focuses on ‘what is meaningful’, given an input to the convolution layer. To find the channel attention, at first max pooling and average pooling are performed to the input feature, then it is passed through a shared multilayer perceptron with one hidden layer. The two features obtained after multilayer perceptron are added and sigmoid function is applied. The final result is multiplied with the input feature to obtain channel attention.

After applying channel attention, spatial attention is applied. Spatial attention focuses on ‘where is an informative part’. In order to compute spatial attention, max pooling and average pooling operation are applied along the channel axis, and two outputs are concatenated. The concatenated feature is passed through a convolution operation and a sigmoid function. The feature is multiplied with the input feature to calculate spatial attention.

In the original Resnet-18 architecture, we change the kernel of the first convolutional layer to 3 * 3 from 7 * 7 and the final layer to 10 to match the classes of acoustic scene classification dataset.

2.2. Preprocessing

We use Openl3 embedding as an input to the system. The input representation and embedding size for the calculation of Openl3 embedding was ‘mel256’ and 512. The content type used in openl3 for the submission 1 and 2 of task1a is ‘env’ and for the submission 3 and 4 is ‘music’. For the experiment on the training and validation set we use ‘env’ content type.

2.3. Data Augmentation and Loss Function

Mixup [5] and SpecAugment [6] are used as data augmentation techniques, and Focal Loss [7] is used as a loss function. Mixup is a data augmentation technique in which we combine two inputs and their label together. This helps the network to better understand the features and thus increase the performance. The value of alpha was 0.2 for this experiment. SpecAugment uses time masking, frequency masking, and time stretching techniques to augment the data. We use both of time masking and frequency masking. Time-stretching is not used as it is computationally heavy and does not contribute much in the learning. Focal loss adds a modulating factor to the cross-entropy loss. We use gamma as 2 and alpha as 0.2 in the focal Loss. Time masking and Frequency making are used in about 30% of the training data randomly.

† Corresponding Author

This work was supported by the Technology Development Program (S2735244) funded by the Ministry of SMEs and Startups (MSS, Korea).

3. EXPERIMENTAL SETUP AND DATASET

3.1. Dataset

The dataset for the DCASE 2020 challenge task1a is TUT urban acoustic scenes 2020 mobile, development dataset. It has 10 classes of 10 seconds audio recorded over different cities in Europe. The classes consist of airport, shopping mall, metro station, street pedestrian, public square, street traffic, tram, bus, metro, and park. In this dataset, the audio is from 3 real devices and 6 simulated devices.

All the submitted model for the challenge is trained on the development set (train split and the extra data) and is tested on the validation split. No additional data is used to train the network.

3.2. Experimental Setup

The CBAM model is trained in Titan X (Pascal) using Pytorch with Adam optimizer. The learning rate and batch size are used as 0.001 and 20, respectively. The model is run for 300 epochs. Torch Audio is used for SpecAugment augmentation with the value for frequency masking as 40 and for time masking as 100.

4. RESULT

Table 1: Class-wise accuracy of our model and the baseline model.

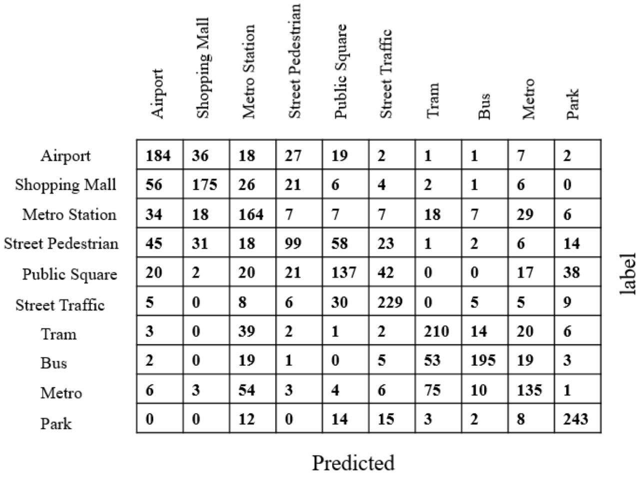
Class	**Baseline (%)**	**Ours (%)**
Airport	45.0	**61.2**
Shopping Mall	48.3	**59.9**
Metro Station	53.0	**55.2**
Street Pedestrian	29.8	**33.3**
Public Square	44.9	**46.1**
Street Traffic	79.9	**77.1**
Tram	52.2	**70.7**
Bus	62.9	**65.6**
Metro	**53.5**	45.5
Park	71.3	**81.8**
Average	**54.1**	**59.6**

An average accuracy of our model is 59.6%, as shown in Table 1, it is 5.5% higher than that of the baseline model. Our model shows better accuracy in every class except Metro in which the baseline had higher accuracy. The confusion matrix for the test set is given in Figure 1.

In Table 1, the proposed model shows the highest accuracy in Park which is 81.8%, this is about 10% higher than that of the baseline and got the lowest in Metro, which is 45.5%, and this result is 8% less than that of baseline. From figure 1, we can see that the model misclassified Metro as Tram and Metro Station. The lowest accuracy among all the classes is of Street Pedestrian at 33.3%.

5. CONCLUSION

Attention-based Resnet-18 model was able to obtain 59.6% accuracy, which is 5.5% more than that of the baseline model. Both baseline and our model used Openl3 embedding. This result showed that Resnet with attention model is very effective in acoustic scene classification.



	Airport	Shopping Mall	Metro Station	Street Pedestrian	Public Square	Street Traffic	Tram	Bus	Metro	Park
Airport	**184**	**36**	**18**	**27**	**19**	**2**	**1**	**1**	**7**	**2**
Shopping Mall	**56**	**175**	**26**	**21**	**6**	**4**	**2**	**1**	**6**	**0**
Metro Station	**34**	**18**	**164**	**7**	**7**	**7**	**18**	**7**	**29**	**6**
Street Pedestrian	**45**	**31**	**18**	**99**	**58**	**23**	**1**	**2**	**6**	**14**
Public Square	**20**	**2**	**20**	**21**	**137**	**42**	**0**	**0**	**17**	**38**
Street Traffic	**5**	**0**	**8**	**6**	**30**	**229**	**0**	**5**	**5**	**9**
Tram	**3**	**0**	**39**	**2**	**1**	**2**	**210**	**14**	**20**	**6**
Bus	**2**	**0**	**19**	**1**	**0**	**5**	**53**	**195**	**19**	**3**
Metro	**6**	**3**	**54**	**3**	**4**	**6**	**75**	**10**	**135**	**1**
Park	**0**	**0**	**12**	**0**	**14**	**15**	**3**	**2**	**8**	**243**

Figure 1: Confusion Matrix for the validation split

6. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” In *Proceedings of 2018 Workshop on the Detection and Classification of Acoustic Scenes and Events (DCASE2018)*, 2018, pp. 9-13.

[2] J. Cramer, H-.H. Wu, J. Salamon, and J. P. Bello, “Look, listen and learn more: design choices for deep audio embeddings,” In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[3] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778

[4] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, “CBAM: Convolutional Block Attention Model,” In *Proceedings of the European Conference on Computer Vision (ECCV),* 2018, pp. 3-19.

[5] H. Zhang, M. Cisse, Y. N. Dauphin and D. L. Paz, “mixup: Beyond empirical risk minimization,” In *proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[6] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.

[7] T. Y. Lin, P, Goyal, R. Girshick, K. He, and P. Dollár. “Focal loss for dense object detection.” In *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2017, pp. 2980-2988.