CONSISTENCY LEARNING BASED ACOUSTIC SCENE CLASSIFICATION WITH RES-ATTENTION

Technical Report

Mengfan Cui, Fan Kui, Liyong Guo

Northwestern Polytechnical University
Computer Science Dept
1640291556@qq.com, 1145921404@qq.com, 839019390@qq.com

ABSTRACT

In this report, we propose a consistency learning based method with different data augmentation methods to tackle Acoustic Scene Classification task1a in the DCASE2021 Challenge. Classification of data from multiple devices (real and simulated) targeting generalization properties of systems across a number of different devices and focusing on low-complexity solutions. Consistency learning is used to reduce the embedding distance of the augmented sample and the original sample. With the consistency learning, the algorithm is robust with device variances. For low-complexity and high-accuracy, a Res-Attention structure which combines residual structure with separable convolution layer and attention layer is proposed. On Task1a development dataset, the presented method gets 69.71% accuracy (0.87 log CrossEntropy loss) with the model size 93.3KB by using int8 quantization.

Index Terms— Acoustic Scene Classication, low-complexity, multi-devices, consistency learning, data augmentation

1. INTRODUCTION

With the development of computer acoustic analysis, the task of acoustic scene classification has received more and more attention. Since the DCASE competition was held in 2013, there will be a related event set up every time to promote the development of the field [1] [2] [3].

Acoustic Scene Classification(ASC) is a topic to tell the category of recorded audio. The main idea is to recognize real-world sounds into a given set of environments Class, such as subway station, street traffic, or public square. Acoustic scene has a large amount of sound information and rich content. This makes accurate scene prediction difficult and thus becomes an interesting research question. Therefore, ASC has always been an attractive study For decades, the detection and classification of acoustic scenes and events (DCASE) challenges [4] have provided Benchmarking data and a competitive platform for the promotion of sound scenes Research and analysis. The main challenge of ASC is that the recordings are not enough to cover different devices and different situations. So we can only use the limited dataset to make the algorithm generalize on un-seen data.

In DCASE 2021, Task 1 has two different subtasks [5][**?**][4]. The report mainly focuses on the task 1a. The key goal is to design a device-invariant system that can classify ten scene audio recorded by different equipment, no need to use any equipment information for the evaluation phase. At the mean time, it focuses on low-complexity solutions, the model size should be less than 128KB.

We describe the system we submitted for DCASE 2021 task1a. For task 1a, we constructed a acoustic scene Classification system by combined different data augmentation methods with consistency learning to improve the robustness and reduce the device dependence. The proposed model with Res-Attention structure significantly improves the performance of ASC. Then the int8 quantization method is used to compress the trained model. In this way, the model can be compressed to 1/4 of the original size.

2. METHODS

We proposed consistency learning based acoustic scene classification with Res-Attention. The training and test phase are shown in Figure 1 and Figure 2. Every training sample is augmented by a random set of operations, such as reverb, filtering etc. The processed data and the original data should get similar embeddings with consistency learning restriction.

2.1. Model

The model structure is shown in Figure 3. The Res-Attention block is shown in Figure 4. In the Res-Attention block, the conv2d layer is replaced by depthwise convolution layer and pointwise convolution layer [6]. At the mean time, a self-attention block [7] [8] is added to the residual structure to capture the whole information.

2.2. Data Augmentation

To improve the robustness of our model, we applied different data augmentation methods:

1. Reverb: To enlarge the dataset and improve the variance, we simulate different room impulse response signal [9] to convolve with the original data.
2. Filtering : To mimic the recording device difference, a random high-pass, low-pass filter or band-pass filter is used to enhance or suppress random frequency subband.
3. Random Gain : To randomly adjust the gain of recordings.
4. SpecAugment[10]: consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps.

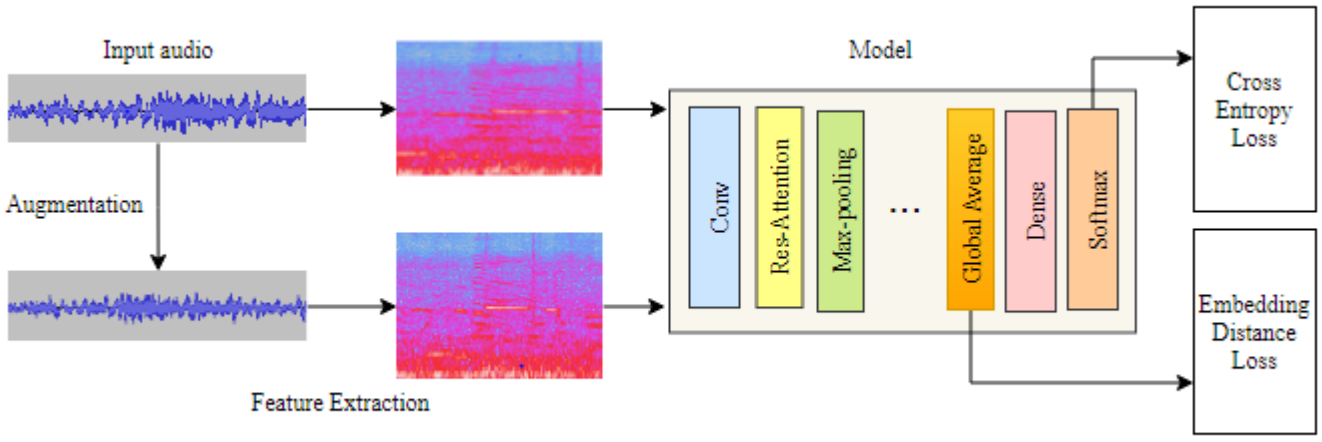

Figure 1: The training phase of proposed method.

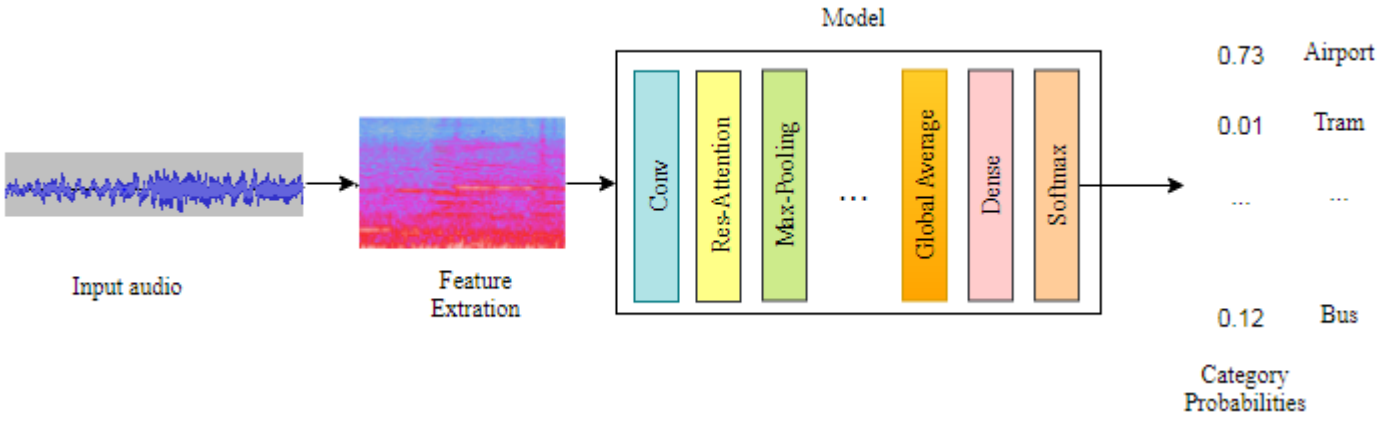

Figure 2: The test phase of proposed method.

2.3. Training Loss

The training loss consists two parts. The first part is CrossEntropy classification loss $L_{CrossEntropy}$. And the second part is embedding distance loss $L_{EmbeddingDistance}$.

$$L_{CrossEntropy} = \sum(-y * log(f(x))) \quad (1)$$

where x represents the input feature, $f(x)$ means the model inference which is used to get the predicted probabilities for each class. y is the onehot encoding of the target.

$$L_{EmbeddingDistance} = ||embedding - embedding_{aug}||^2 \quad (2)$$

where $embedding$ is the output of the global average pooling layer, $embedding_{aug}$ corresponds to the augmented data e augmented data. The EmbeddingDistance loss is mainly used to improve the robustness of the algorithm.

3. TRAINING

We train and evaluate our proposed architecture on DCASE2021 development dataset which consists of acoustic scene data recorded in different European cities. All data has the scene label (one of 10 scenes: e.g. 'airport' or 'shopping mall') and city label (one of 12 cities: e.g. 'lyon'). In the experiments, we used 9,185 segments and 4,185 segments for the training and evaluation. The development dataset is provided by DCASE2021. Our proposed algorithm was implemented using PyTorch with a GeForce GTX TITAN X GPU with 12Gb RAM. Every 10 second-long audio input sampled as 44.1kHz, we extracted 128 log-mel features(window size 2048, hop length 882, fft number 2048) for the input feature of the network. The network was trained by the SGD optimizer with 300 epochs. The learning rate was set to 0.1, exponentially decaying values from 0.1 to 0.001.

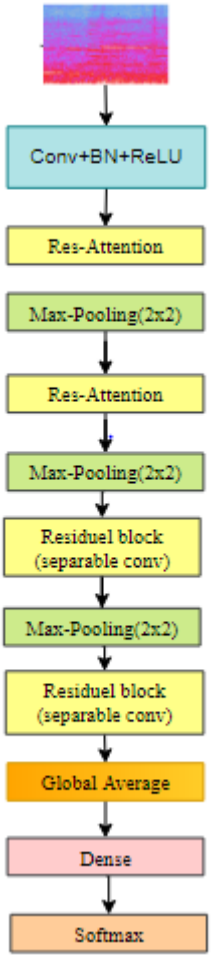

Figure 3: The proposed method architecture.

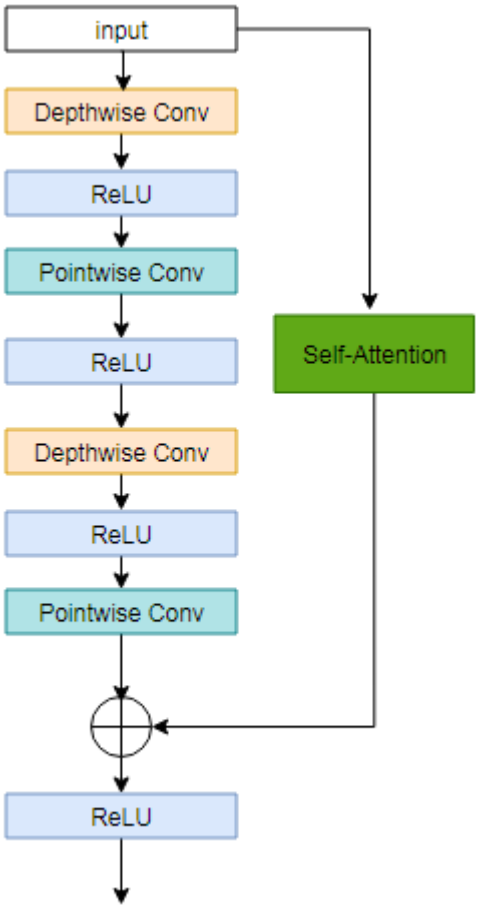

Figure 4: The proposed res-attention structure.

4. RESULTS

This report has presented a robust acoustic scene classification method for DCASE2021 task1a. As a result in Table 3, we achieve 69.71% accuracy on development dataset and the confusion matrix is shown in Figure 5. The class-wise and device-wise accuracy in shown in Table 1 and Table 2.

Table 3: The result of proposed method.

	CrossEntropy loss	accuracy
DCASE2021 Task 1a Baseline	1.473	47.7%
ours	0.87	69.71%

5. REFERENCES

[1] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "Cp-jku submissions for dcase-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, vol. 6, pp. 5024–5028, 2016.

[2] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "Dcase 2016 acoustic scene classification using convolutional neural networks," in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99.

[3] M. Wan, R. Wang, B. Wang, J. Bai, C. Chen, Z. Fu, J. Chen, X. Zhang, and S. Rahardja, "Ciaic-asc system for dcase 2019 challenge task1," *Tech. Rep., DCASE2019 Challenge*, 2019.

[4] http://dcase.community/challenge2021/.

[5] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

[6] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

Table 1: The class-wise accuracy.

Accuracy	airport	bus	metro	metro station	park	public_square	shopping mall	street pedestrian	street_traffic	tram
baseline	40.5%	47.1%	51.9%	28.3%	69.0%	25.3%	61.3%	38.7%	62.0%	53.0%
ours	61.14%	79.12%	63.97%	70.37%	85.18%	54.21%	58.92%	54.54%	88.55%	81.08%

Table 2: The device-wise accuracy.

Accuracy	a	b	c	s1	s2	s3	s4	s5	s6
baseline	63.9%	52.2%	56.3%	44.2%	43.9%	44.5%	38.5%	40.6%	38.2%
ours	77.3%	67.8%	73.5%	70.3%	64.2%	70.0%	67.6%	68.2%	68.5%

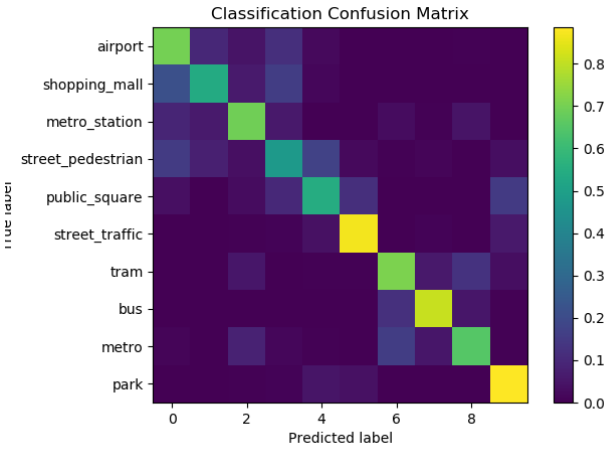

Figure 5: The confusion matrix.

[8] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[10] 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.