ACOUSTIC SCENE CLASSIFICATION USING DEEP RESIDUAL NETWORKS WITH FOCAL LOSS AND MILD DOMAIN ADAPTATION

Technical Report

Wei Gao and Mark D. McDonnell

Computational Learning Systems Laboratory,
UniSA STEM,
University of South Australia, Mawson Lakes SA 5095, Australia

ABSTRACT

This technical report describes our approach to Tasks 1a in the 2020 DCASE acoustic scene classification challenge. We have incorporated few more training techniques based on our previous contest entries. One was replacing cross-entropy loss with focal loss which aims to focus on poor-classified samples while reducing the loss on well-classified samples with high probability; another methods used was to add an auxiliary binary classifier to serve the purpose of domain adaptation.

Index Terms— deep residual network; focal loss; domain adaptation

1. INTRODUCTION

This paper describes our four submission entries to the Task 1a in the DCASE2020 IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events. The aim of Task 1a is to build a multi-category classifier that can identify the background scene out of ten choices from pre-recorded audio clips in which they were collected [1]. The increased variety of recording devices used in this task has made the task more complicated compared to last year's, which poses challenge to the generalisation capabilities of the proposed models.

Models were trained and validated using a development set, and tested and evaluated on an separate evaluation set. We used our previous contest model as baseline [2] which demonstrated excellent generalisation to unknown devices in the evaluation set. In additional to our baseline setup, we also investigated the use of focal loss and mild domain adaptation in order to handle device mismatch between the training set and the others.

Code for training our models using Keras [3] is available at `https://github.com/emilywg/DCASE2020-Task1` .

2. EXPERIMENT SETUP

2.1. Acoustic Feature Extraction

Similar to [2], We used the LibROSA library[1] to generate the acoustic features. We calculated 128 log-mel energies under the original sampling rate of 44.1KHz for each time slice by taking 2048 FFT points with 50% overlap. The mel scale was defined using HTK formula [4]. The resulting spectrograms were of size 128 frequency bins, 423 time samples and 3 channels with each representing log-mel spectrograms, its delta features and its delta-delta features respectively.

2.2. The Baseline Model

We trained a 18-layer pre-activation ResNet served as the baseline. The details of the model was reported in [2], which explained the architecture design (see Section 2.1, 3.4 & 3.5), the design of splitting high and low frequencies (see Section 3.3), and the methods used for regularisation and data augmentation (see Section 2.2 & 3.6). Please refer to the paper as cited above.

For training the baseline using the similar approach in [2], we used backpropagation and stochastic gradient descent, with a batch size of 32, momentum of 0.9, and the cross-entropy loss function. Each network was trained for 310 epochs using a learning rate schedule with warm restart that resets the learning rate to its maximum value of 0.1 after $10, 30, 70$ and 150 epochs, and then decays according to a cosine pattern to 1×10^{-5}. It was shown by [5] and verified by [6] that this approach can provide improvements in accuracy on image classification relative to using stepped schedules.

2.3. Focal Loss

The focal loss [7] was initially designed to handle severe data imbalance in object detection problem. It adds a modulating factor to the cross-entropy loss so as to increases the sensitivity of the model to classify hard samples. The use of focal loss has demonstrated calibration effect even in balanced dataset as shown in [8], therefore we assume it may help with the problem of device mismatching in this task.

2.4. Domain Adaptation

We added an auxiliary binary classifier which tried to identify whether the audio inputs were processed by the major device (as source domain) or by other devices (as target domain). We think it serves as mild domain adaptation as the gradients in the models were gaining domain information through back-propagating the binary loss. Also, we applied weighting to these two losses during training where the Total Loss = CE Loss – $0.1 \times$ Adaptation Loss.

W.G. was supported by an Australian Government Research Training Program Scholarship.

[1] `https://librosa.github.io/librosa/`

Table 1: Results for Task 1a on official development data fold with 2970 samples in validation set.

Categories	Submission ID			
	1	2	3	4
airport	56.5%	57.2%	56.9%	56.5%
bus	82.8%	79.7%	81.8%	82.4%
metro	77.7%	73.4%	70.7%	75.0%
metro station	76.0%	71.3%	72.7%	73.7%
park	90.9%	90.5%	90.5%	91.5%
public square	49.8%	50.1%	52.1%	51.8%
shopping mall	63.9%	69.6%	72.3%	70.3%
street pedestrian	55.5%	58.5%	56.2%	56.5%
street traffic	88.8%	89.8%	89.5%	90.2%
tram	75.0%	77.4%	71.0%	77.4%
Avg accuracy	71.7%	71.8%	71.4%	72.5%
# of parameters	4.31M	4.31M	4.31M	12.93M

3. RESULTS

Table 1 shows the raw accuracy of each categories and the balanced accuracy. Figure 1 shows the confusion matrix of the ensemble model.

Our four submission entries are listed as following:

- Submission 1: The baseline model
- Submission 2: The baseline model except using focal loss
- Submission 3: The baseline model with an auxiliary binary classifier
- Submission 4: The ensemble of above three

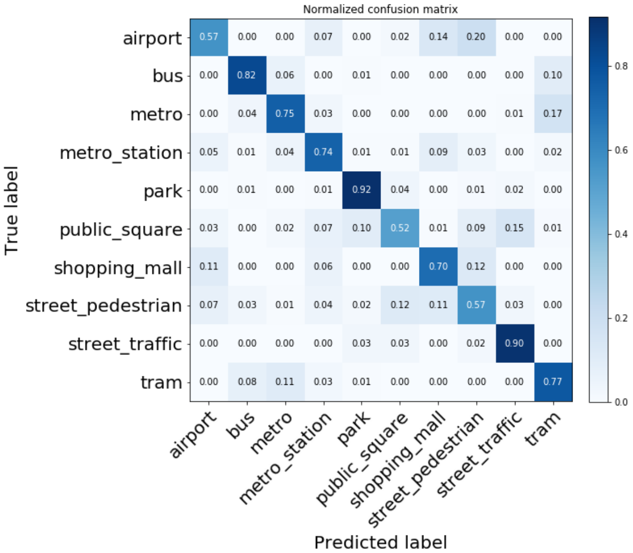

Figure 1: Normalized confusion matrix showing the classification accuracy for the ensemble model.

4. REFERENCES

[1] 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.

[2] M. D. McDonnell and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 141–145.

[3] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[4] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[5] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with restarts," *CoRR*, vol. abs/1608.03983, 2016. [Online]. Available: http://arxiv.org/abs/1608.03983

[6] M. D. McDonnell, "Training wide residual networks for deployment using a single bit for each weight," 2018, in Proc. ICLR 2018; arxiv: 1802.08530.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[8] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania, "Calibrating deep neural networks using focal loss," 2020, in arxiv.2002.09437.