DATA-EFFICIENT LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING MOBILENET VARIANT

Technical Report

Wei Gao and Ivan Lee

UniSA STEM,
University of South Australia, Adelaide, Australia

ABSTRACT

This report describes our submission on the task of low-complexity acoustic scene classification of the DCASE 2024 challenge. To meet the system complexity limitations of the task, we trained a single MobileNet variant fitting for all five pre-defined data folds. The training was optimised towards a focal loss function that helped on hard misclassified samples. The models were deployed with FP16 precision for the sake of efficient inference.

***Index Terms*—** MobileNet, Focal Loss

1. INTRODUCTION

The DCASE Challenge for the research problem of acoustic scene classification focuses on categorising short audio clips into ten specific scene types including both indoor and outdoor scenarios. The 2024 version [1] introduces hard constraints, limiting the maximum memory allowed for model parameters to 128 kB and capping the number of multiply-accumulate operations (MACs) at 30 million for inferring a one-second audio clip as provided. Also, the development data set is updated to contain five explicitly defined folds that aims to encourage solutions handling limited amount of labeled data.

2. METHODS

2.1. Focal Loss Function

Based on our previous attendances of the challenge [2, 3], the uses of focal loss [4] have demonstrated calibration effects for poorly classified samples so as to improve the overall classification results. The focal loss function is defined as

$$\mathrm{FL}(p_t) = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$$

where p_t is the predicted probability for the class, α_t is a balancing factor designed to address the issue of class imbalance during training. γ is a focusing parameter leading to the modulating factor $(1 - p_t)^{\gamma}$ that can be tuned against the predicted probability p_t. It turns small and decreases the loss if the classier is too confident on some classes, and vice visa. Such modifications adding to the standard cross-entropy loss help focus more on predicting challenging samples.

Given that the DCASE dataset is well-curated to ensure class balance, the balancing factor α_t was not configured for the experiments. The focusing parameter γ was set to 2 for optimal performance as determined by the results of the conducted experiments.

2.2. Training Set-up

Our submitted system was training a MobileNet variant [5] and adapted from the official baseline implementation [6]. The system required 119.5 kB of memory after quantization to 16-bit precision and 29.43 MMACs for inference.

Similar strategy of extracting acoustic features was also deployed except that a slightly shorter hopping window of 472 was applied while generating spectrograms. This leaded to a resulting input features with 256 frequency bins and 68 time samples.

For training the models on smaller data folds such as 5% subset, 10% subset and 25% subset, a batch of 78 and a learning rate of 0.005 were chose. For the 50% and 100% subsets, the batch size was set to 130 and a learning rate of 0.01 was applied.

3. RESULTS

Table 1 compares our results with the official baseline.

4. REFERENCES

[1] F. Schmid, P. Primus, T. Heittola, A. Mesaros, I. Martín-Morató, K. Koutini, and G. Widmer, "Data-efficient low-complexity acoustic scene classification in the dcase 2024 challenge," 2024.

[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] W. Gao and M. D. McDonnell, "Acoustic scene classification using deep residual networks with focal loss and mild domain adaptation," DCASE2020 Challenge, Tech. Rep., June 2020.

[4] 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.

[5] A. Howard, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, nov 2019, pp. 1314–1324.

[6] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, "Distilling the knowledge of transformers and CNNs with CP-mobile," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, 2023, pp. 161–165.

Table 1: Class-wise results using each of five subsets of the development dataset. **B** denotes the official baseline results on the particular training data split whereas **S** denotes the results of our system. **Avg Accuracy** specifies the class-wise macro-averaged accuracy of individual development subset.

Categories	5%B	5%S	10%B	10%S	25%B	25%S	50%B	50%S	100%B	100%S
Airport	34.77%	34.76%	38.50%	35.13%	41.81%	38.54%	41.51%	46.65%	46.45%	47.56%
Bus	45.21%	47.77%	47.99%	56.26%	61.19%	61.81%	63.23%	71.98%	72.95%	78.68%
Metro	30.79%	35.48%	36.93%	41.48%	38.88%	42.77%	43.37%	49.15%	52.86%	52.25%
Metro Station	40.03%	45.79%	43.71%	39.79%	40.84%	43.67%	48.71%	52.28%	41.56%	51.81%
Park	62.06%	70.77%	65.43%	65.91%	69.74%	72.25%	72.55%	73.29%	76.11%	80.26%
Public Square	22.28%	26.36%	27.05%	29.76%	33.54%	39.69%	34.25%	42.45%	37.07%	40.84%
Shopping Mall	52.07%	49.22%	52.46%	47.74%	58.84%	62.42%	60.09%	55.85%	66.91%	57.30%
Street Pedestrian	31.32%	28.45%	31.82%	42.65%	30.31%	42.55%	37.26%	36.06%	38.73%	45.92%
Street Traffic	70.23%	63.36%	72.64%	72.55%	75.93%	75.42%	79.71%	82.22%	80.66%	77.97%
Tram	35.20%	40.84%	36.41%	47.56%	51.77%	52.12%	51.16%	52.63%	56.58%	63.20%
Avg Accuracy	42.40% $\pm$0.42	44.28%	45.23% $\pm$1.01	47.89%	50.29% $\pm$0.87	53.12%	53.19% $\pm$0.68	56.26%	56.99% $\pm$1.11	59.58%