PRUNING AND QUANTIZATION FOR LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Arshdeep Singh[1], *Dhanunjaya Varma Devalraju*[2], *Padmanabhan Rajan*[3]

Indian Institute of Technology, Mandi, India
Email: $\{d16006^1, S18023^2\}$@students.iitmandi.ac.in, [3]padman@iitmandi.ac.in

ABSTRACT

This technical report describes the IITMandi_AudioTeam's submission for DCASE 2021 ASC Task 1, Subtask Low-Complexity Acoustic Scene Classification with Multiple Devices. This report aims to design low-complexity systems for acoustic scene classification by eliminating filters in a pre-trained convolution neural network. A filter pruning strategy is opted, which consists of three steps. Step 1 aims to identify the redundant filters which have low-norm. Step 2 explicitly removes the redundant filters and their connecting feature maps from the unpruned network to give a pruned network. Step 3 involves fine-tuning of the pruned network to regain performance. Further, the trained parameters are quantized to 16-bit. On DCASE-2021 task 1A development dataset, the proposed framework reduces 68% parameters with competitive performance.

***Index Terms*—** Acoustic scene classification, Low-complexity, Convolution neural network.

1. INTRODUCTION

Convolutional neural networks (CNNs) perform state-of-the-art as compared to traditional hand-crafted methods in many domains [1]. However, it may be difficult to deploy large-size CNNs on constrained devices such as mobile phones, internet of things devices (IoT) etc. This is owing to their high computational cost during inference and requirement of more memory [2, 3]. Thus, the issue of reducing the size and the computational cost of large-scale networks has drawn a significant amount of attention in the research community.

In this report, we aim to prune or compress a well-trained CNN which gives a competitive performance as that of the unpruned CNN. For this, the filter pruning methods have opted since they produce a structured pruned network that is easy to deploy and gives effective speed-up in contrast to that of weight pruning methods. A filter in a given intermediate layer is defined as "redundant" if it has a low norm. We hypothesize that such redundant filters produce low activation or response, and are relatively less effective in providing useful information than other filters for classification. The rest of the report is organized as follows. In section 2, the proposed methodology is explained. Experimental setup and submitted entries are included in Sections 3 and 4 respectively.

The rest of the report is organized as follows. In section 2, the proposed methodology is explained. Experimental setup and submitted entries are included in Section 3 and 4 respectively.

2. PROPOSED METHODOLOGY

In this section, we describe various steps to achieve a pruned network.

Step 1: Given a pre-trained network, identify importance scores for each filters in a given intermediate layer using l_1-norm based method [4] or GM-based method [5].

- *l_1-norm based method [4]*: In this method, the norm of each filter is computed, and is used as a significant score to quantify their filter importance. A filter with low-norm indicates relatively less importance than other filters.
- *Geometric-median (GM)* based method [5]: In this method, geometric median of all filters is computed, which represents the common information of all filters. After this, the difference between the geometric median of all filters and a given filter is computed, which represents the significance score corresponding to that filter. A low significance score indicates that the filter represents common information, and hence can be neglected.

Step 2: Given a pruning ratio[1], select top few important filters and remove other filters alongwith their connecting feature maps explicitly from the network. The elimination process can be in one-shot, which means all the connection across various intermediate layers are removed in one-shot and then the pruned network is fine-tuned. In iterative pruning process, the connections are eliminated from a given layer, then the network is fine-tuned. This process is again performed for other layers.

Step 3: Fine-tune the pruned network.

Step 4: Quantization of the trained parameters to 16-bit.

3. EXPERIMENTAL SETUP

Dataset used: The dataset used for the task is TAU Urban Acoustic Scenes 2020 Mobile, development dataset [6]. It consists of 13962 training and 2970 testing examples of 10-seconds length. Each example is transformed into time-frequency representations to produce log-mel band energy based representations of size (40 × 500) with 40 mel-bands and 500 time frames.

Unpruned network: The unpruned network consists of DCASE-2021 baseline network [6, 7] which is trained on DCASE-2021 task 1A dataset. The network comprises of three convolutional layers (L1 to L3) which are followed by a fully connected layer (Dense). The input to the network is log mel-band energies of size (40 × 500) with 40 mel-bands and 500 time frames as computed for an audio of 10-second length. The size of the network is 90.3 KB[2] with 46246 number of parameters. The performance of the network is measured

[1] Pruning ratio is defined as number of filters to be eliminated from the network or in a given intermediate layer.

[2] CNN model size, number of total and non-zero parameters are computed using the script *model_size_calculation.py*

Table 1: Various CNN models submitted for DCASE-2021 task1a challenge.

Entry No.	Network	Architecture (L1-L2-L3-Dense)	Parameters	Size (KB)	Accuracy (%)	log-loss
NA	Unpruned	16-16-32-100	46246	90.1	48.59	1.425
1	Singh_29KB	16-8-16-32	14754	28.8	47.74	1.383
2	Singh_53KB	16-8-32-100	27166	53.06	48.48	1.394
3	Singh_74KB	16-16-24-100	38110	74.43	49.05	1.395
4	Singh_71KB	16-12-32-100	36818	71.44	48.65	1.413

in terms of accuracy and the log-loss metrics. The trained network gives 48.59% accuracy and 1.425 log-loss.

Fine-tuning of the pruned network The pruned network is fine-tuned for 200 epoch with Adam optimizer. An early stopping criterion is used on validation loss during fine-tuning process. It is observed that the pruned network takes 30-50 epochs to converge the unpruned network performance. Apart from architecture, dropout after various layers are kept similar to that of the unpruned network.

4. SUBMITTED ENTRIES

In this section, a detail of various submitted models is described.

- **Singh_29KB:** The pruned model is generated in an iterative pruning manner using three pruning levels after applying l_1-norm based pruning. **(Level 1 pruning)** First, L2 layer is pruned at 50% pruning rate, and the pruned network is fine-tuned. **(Level 2 pruning)** Next, the pruned network thus obtained is again pruned by eliminating L3 layer 50% unimportant connection. **(Level 3 pruning)** Next, the pruned network thus obtained is fine-tuned by reducing the number of units in the dense layer to 32. The weights of subnetwork (L1 to L3 layer) are initialized as obtained after fine-tuning process in Level 2 pruning.
- **Singh_53KB:** The pruned model is generated in an one-shot pruning manner after applying GM-based pruning. L2 layer is pruned at 50% pruning ratio.
- **Singh_74KB:** The pruned model is generated in an one-shot pruning manner after applying GM-based pruning. L3 layer is pruned at 25% pruning ratio.
- **Singh_71KB:** The pruned model is generated in an one-shot pruning manner after applying GM-based pruning. L2 layer is pruned at 25% pruning ratio.

5. CHALLENGE SUBMISSION

We submit four results obtained using the four pruned networks (1)-(4) as given in Table 1 as a final submission for evaluation dataset. The following filenames are used in the submission.

1. Singh_IITMandi_task1a_1 : Predictions generated by network Singh_29KB.
2. Singh_IITMandi_task1a_2 : Predictions generated by network Singh_53KB.
3. Singh_IITMandi_task1a_3 : Predictions generated by network Singh_74KB.
4. Singh_IITMandi_task1a_4 : Predictions generated by network Singh_71KB.

The above trained pruned network, codes can be found at this link [3] .

6. CONCLUSION

This report focuses on low-complexity system for acoustic scene classification. A filter pruning and quantization procedure is applied to obtain compressed, accelerated, and low-size CNN. The proposed framework shows promising results in terms of reduction in parameters and performance.

7. REFERENCES

[1] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[4] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[5] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang, "Filter pruning via geometric median for deep convolutional neural networks acceleration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4340–4349.

[6] I. Martín-Morató, T. Heittola, A. Mesaros, and T. Virtanen, "Low-complexity acoustic scene classification for multi-device audio: analysis of dcase 2021 challenge systems," *arXiv preprint arXiv:2105.13734*, 2021.

[7] 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

[3] Link: Pruned models, script for pruning, fine-tuning.