UNIVERSITY
OF TECHNOLOGY
SYDNEY

Accelerating Deep Convolutional Neural Networks via Filter Pruning

by Yang He

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Yi Yang

University of Technology Sydney
Faculty of Engineering and Information Technology
July 2021

Certificate of Authorship/Originality

I, Yang He, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Yang He

Production Note:
Signature removed prior to publication.

Date: 1/July/2021

Acknowledgements

First, I would like to express my tremendous gratitude to my supervisor, Professor Yi Yang, I really appreciate his mentoring and cultivating in my research and my life. Although I changed my major for my PhD study and had little knowledge in the area computer science when I entered the group, Yi has always been patient with me and given me quite a lot guidance. Every time I encountered difficulties, Yi is always here to help me. It is really my fortune to have Yi as my supervisor.

I am grateful for all my colleagues and teammates in Yi's group. I convey my gratitude to my co-supervisor, Prof. Liang Zheng, for his help and advises during this research. I feel so lucky to met Prof. Yanwei Fu, Guoliang Kang and Xuanyi Dong that they teached me quite a lot about doing research and writing code when I am a novice of the research field. It is so important to express my gratitude to Ping Liu, who is reliable and heart-warming, and gives me many kinds of help and advice. I am happy to collaborate with many creative teammates in our team and I really appreciate the kind and useful suggestions given by them. Moreover, I am thankful for Prof. Hanwang Zhang for his guidance and advises during my visiting to NTU. I have met quite a lot friends at NTU and thanks for your support. I also thank Qi Yao for helping me with my PhD study.

Finally, I want to express my special gratitude to my parents, who give me endless love, help, and encouragement in my life. How lucky to be your child.

Thanks for all the people that ever helped me and encouraged me.

Yang He
Sydney, Australia, 2021.

List of Publications

Journal Papers

J-1. **Yang He**, Xuanyi Dong, Guoliang Kang, Yanwei Fu, Chenggang Yan, and Yi Yang, “Asymptotic Soft Filter Pruning for Deep Convolutional Neural Networks,” IEEE Transactions on Cybernetics, vol. 50, no. 8, pp. 3594–3604, 2019.

Conference Papers

C-1. **Yang He**, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang, “Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks,” in International Joint Conference on Artificial Intelligence (IJCAI), pp. 2234–2240, 2018.

C-2. **Yang He**, Ping Liu, Ziwei Wang, Zhilan Hu, Yi Yang,“Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4340–4349, 2019.

C-3. **Yang He**, Yuhang Ding, Ping Liu, Linchao Zhu, Hanwang Zhang, Yi Yang, “Learning Filter Pruning Criteria for Deep Convolutional Neural Networks Acceleration,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2009–2018, 2020.

Contents

Certificate ii

Acknowledgments iii

List of Publications iv

List of Figures x

List of Tables xvi

Abstract xix

1 Introduction 1

1.1 Background 1

1.2 Thesis Organization 3

2 Related Works 5

2.1 Matrix Decomposition 5

2.2 Low Precision 6

2.3 Weight Pruning 6

2.4 Filter Pruning 7

2.5 Neural Architecture Search 8

2.6 Vision Transformer Compression 9

2.7 Applications 9

3 Soft Filter Pruning 11

3.1 Introduction 11

3.2 Related Works 13
3.3 Methodology 15
3.3.1 Preliminaries 15
3.3.2 Soft Filter Pruning (SFP) 16
3.3.3 Computation Complexity Analysis 20
3.4 Evaluation and Results 22
3.4.1 Benchmark Datasets and Experimental Setting 22
3.4.2 ResNet on CIFAR-10 22
3.4.3 ResNet on ILSVRC-2012 23
3.4.4 Ablation Study 25
3.5 Conclusion 27

4 Asymptotic Soft Filter Pruning 28

4.1 Introduction 28
4.2 Related Work 31
4.2.1 Matrix Decomposition 31
4.2.2 Low Precision 31
4.2.3 Weight Pruning 32
4.2.4 Filter Pruning 32
4.3 Methodology 33
4.3.1 Preliminary 33
4.3.2 Pruning with Hard Manner 35
4.3.3 Pruning with Soft Manner 36
4.3.4 Asymptotic Soft Filter Pruning (ASFP) 37
4.3.4.1 Asymptotic Filter Selection 38

4.3.4.2 Filter Pruning 39

4.3.4.3 Reconstruction 39

4.3.4.4 Obtaining Compact Model 40

4.3.5 Pruning Strategy for Convolutional Network 40

4.3.6 Computation Complexity Analysis 41

4.3.6.1 Theoretical Speedup Analysis 41

4.3.6.2 Realistic Speedup Analysis 42

4.4 Experiment 42

4.4.1 Benchmark Datasets and Experimental Setting 42

4.4.2 VGGNet on CIFAR-10 45

4.4.3 ResNet on CIFAR-10 47

4.4.4 ResNet on ILSVRC-2012 48

4.4.5 Comparing SFP and ASFP 49

4.4.6 Ablation Study 51

4.4.6.1 Filter Selection Criteria 51

4.4.6.2 Varying Pruned FLOPs 52

4.4.6.3 Selection of the Pruned Layers 52

4.4.6.4 Sensitivity of the ASFP Interval 53

4.4.6.5 Sensitivity of Parameter D of ASFP 53

4.5 Conclusion 53

5 Filter Pruning via Geometric Median 54

5.1 Introduction 54

5.2 Related Works 57

5.3 Methodology 58

5.3.1 Preliminaries 58

5.3.2 Analysis of Norm-based Criterion 59

5.3.3 Norm Statistics in Real Scenarios 60

5.3.4 Geometric Median 62

5.3.5 Filter Pruning via Geometric Median 62

5.3.6 Theoretical and Realistic Acceleration 65

5.3.6.1 Theoretical Acceleration 65

5.3.6.2 Realistic Acceleration 65

5.4 Experiments 68

5.4.1 Experimental Settings 68

5.4.2 Single-Branch Network Pruning 69

5.4.3 Multiple-Branch Network Pruning 69

5.4.4 Ablation Study 72

5.4.5 Feature Map Visualization 73

5.5 Conclusion 74

6 Learning Filter Pruning Criteria 75

6.1 Introduction 75

6.2 Related Work 78

6.3 Methodology 80

6.3.1 Preliminaries 80

6.3.2 Learning Filter Pruning Criteria 81

6.3.2.1 Pruning Criteria 81

6.3.2.2 Criteria Space Complexity 83

6.3.2.3 Differentiable Criteria Sampler 83

6.4 Experiments 86

6.4.1 Experimental Setting 86

6.4.2 ResNet on CIFAR-10 89

6.4.3 ResNet on CIFAR-100 91

6.4.4 ResNet on ILSVRC-2012 92

6.4.5 More Explorations 92

6.5 Conclusion 95

7 Conclusion and Future Work 96

7.1 Conclusion 96

7.2 Future Work 97

Bibliography 99

List of Figures

1.1 Neural network acceleration is about breaking the existing “wall” between algorithm and hardware to fit machine learning algorithms into the resource-constrained hardware platforms. 1

3.1 **Hard Filter Pruning** *v.s.* **Soft Filter Pruning**. We mark the pruned filter as the green dashed box. For the hard filter pruning, the pruned filters are always ***fixed*** during the whole training procedure. Therefore, the model capacity is reduced and thus harms the performance because the dashed blue box is useless during training. On the contrary, our SFP ***allows*** the pruned filters to be updated during the training procedure. In this way, the model capacity is recovered from the pruned model, and thus leads a better accuracy. 12

3.2 Overview of SFP. At the end of each training epoch, we prune the filters based on their importance evaluations. The filters are ranked by their ℓ_p-norms (purple rectangles) and the small ones (blue circles) are selected to be pruned. After filter pruning, the model undergoes a reconstruction process where pruned filters are capable of being reconstructed (i.e., updated from zeros) by the forward-backward process. **(a)**: filter instantiations before pruning. **(b)**: filter instantiations after pruning. **(c)**: filter instantiations after reconstruction. 16

3.3 Accuracy of ResNet-110 on CIFAR-10 regarding different hyper-parameters. (Solid line and shadow denotes the mean and standard deviation of three experiment, respectively.) 26

4.1 Hard filter pruning *v.s.* soft filter pruning. We mark the pruned filter as the orange dashed box. For the hard filter pruning, the pruned filters are always ***fixed*** during the whole training procedure. Therefore, the model capacity is reduced and thus harms the performance because the dashed blue box is useless during training. On the contrary, our soft pruning method allows the pruned filters to be ***updated*** during the training procedure. In this way, the model capacity is recovered from the pruned model and thus leads a better accuracy. 29

4.2 Pruning and training schedule of HFP and SFP. Before training, we first select some filters with pre-defined importance evaluations. HFP directly deletes these filters before training, while for SFP, those are set to zero and kept. During training (epoch 1 to N), the model size is smaller than the original one for HFP. While for SFP, the zero value filters (filter 2 and 5) become non-zero after training epoch 1. Then we evaluate the importance of filters again and prune filter 3 and 4. The model size would not be reduced but be the same as the original one. When training is finished, the final pruned model is the model at epoch N for HFP. While for SFP, we delete the zero value filters (filter 3 and 6) at epoch N to get the final pruned model. 34

4.3 Overview of HFP (first row), SFP (second row) and ASFP (third row). **(a)**: filter instantiations before pruning. **(b)**: filter instantiations after pruning. **(c)**: filter instantiations after reconstruction. The filters are ranked by their ℓ_p-norms and the small ones (purple rectangles) are selected to be pruned. 35

4.4 Pruning residual block with pruning rate 50%. Red and green number means the remaining output channel number and input channel number after pruning, respectively. “BN” and “ReLU” represents the batch norm layer and non-linear layer, respectively. . 41

4.5 Asymptotically changed pruning rate when the goal pruning rate is 30%. Three blue points are the three pairs to generate the exponential function of pruning rate (the solid curve). 46

4.6 Model performance regarding different ratio of pruned FLOPs. The green line indicates the model without pruning. The blue and the orange lines represent the model under ASFP and SFP, respectively. . 48

4.7 The training process of ResNet-18 and ResNet-50 and on ImageNet regarding SFP and ASFP. The solid blue line and red dashed line indicate the accuracy of the model before and after pruning, respectively. The black line is the performance gap due to pruning, which is calculated by the accuracy after pruning subtracting that before pruning. 50

4.8 Ablation study of ASFP. (Solid line and shadow denote the mean and standard deviation of three experiments, respectively.) 50

5.1 An illustration of (a) the pruning criterion for norm-based approach and the proposed method; (b) requirements for norm-based filter pruning criterion. In (a), the green boxes denote the filters of the network, where deeper color denotes larger norm of the filter. For the norm-based criterion, only the filters with the largest norm are kept based on the assumption that smaller-norm filters are less important. In contrast, the proposed method prunes the filters with redundant information in the network. In this way, filters with different norms indicated by different intensities of green may be retained. In (b), the blue curve represents the ideal norm distribution of the network, and the v_1 and v_2 is the minimal and maximum value of norm distribution, respectively. To choose the appropriate threshold $\mathcal{T}$ (the red shadow), two requirements should be achieved, that is, the norm deviation should be large, and the minimum of the norm should be arbitrarily small. 55

5.2 Ideal and Reality of the norm-based criterion: (a) Small Norm Deviation and (b) Large Minimum Norm. The blue dashed curve indicates the ideal norm distribution, and the green solid curve denotes the norm distribution might occur in real cases. 59

5.3 Norm distribution of filters from different layers of ResNet-110 on CIFAR-10 and ResNet-18 on ILSVRC-2012. The small green vertical lines and blue curves denote each norm and Kernel Distribution Estimate (KDE) of the norm distribution, respectively. 61

5.4 Accuracy of ResNet-110 on CIFAR-10 regarding different hyper-parameters. Solid line and shadow denotes the mean values and standard deviation of three experiments, respectively. 72

5.5 Input image (left) and visualization of feature maps (right) of ResNet-50-conv1. Feature maps with red bounding boxes are the channels to be pruned. 74

6.1 (a) Previous filter pruning methods manually select a criterion and apply it to all layers; (b) our pruning method learns appropriate criteria for different layers based on the filter distribution. In the blue dashed box, the solid boxes of different colors denote different pruning criteria. The yellow boxes without shadow correspond to unpruned layers of the network, while the ones with shadow are the layers pruned by a selected pruning criterion. 75

6.2 Criteria forward and backward in the network. Grey boxes are the normal filters. The probability distribution of criteria for three layers are initialized, as shown in the big orange shadow. After pruning with four criteria, we obtain four “pruned versions” for every layer, which are denoted as boxes in purple, green, orange, and blue color. These filters are utilized to conduct *criteria forward*. Then we get the *criteria loss* on the validation set to update the “criteria distribution”. 82

6.3 Criteria forward within a layer. Boxes of different colors indicate the different pruning criteria. First, we evaluate the importance of the filter based on different criteria. Second, we prune the filter with small importance scores and get four versions of pruned layers with various probabilities. After that, the output feature map is the aligned weighted sum of four feature maps of the pruned layers. . . . 83

6.4 Visualization of the learned criteria and kept filters for ResNet-56 on CIFAR-10. The grey strip indicates the layers before pruning. The blue, orange and green color denote ℓ_1-norm, ℓ_2-norm and geometric median criteria, respectively. For example, the bottom green strip means that for all the 64 filters in 55_{th} layer, GM criterion is automatically selected to prune half of those filters, base on the filter distribution on that layer. 91

6.5 Visualization of the conventional and adversarial criteria for ResNet-56 on CIFAR-10. The grey strip indicates the layers before pruning. Different blue and green colors represent different pruning criteria. 93

6.6 The learned criteria during training the criteria sampler. The L1, L2, and GM denote conventional ℓ_1-norm, ℓ_2-norm, and geometric median criteria, respectively. 94

List of Tables

3.1 Comparison of pruning ResNet on CIFAR-10. In “Fine-tune?” column, “Y” and “N” indicate whether to use the pre-trained model as initialization or not, respectively. The “Accu. Drop” is the accuracy of the pruned model minus that of the baseline model, so negative number means the accelerated model has a higher accuracy than the baseline model. A smaller number of ”Accu. Drop” is better. 21

3.2 Comparison of pruning ResNet on ImageNet. “Fine-tune?” and ”Accu. Drop” have the same meaning with Tab. 3.1. 23

3.3 Comparison on the theoretical and realistic speedup. We only count the time consumption of the forward procedure. 24

4.1 Overall performance of pruning ResNet on CIFAR-10. 43

4.2 Pruning from scratch and pre-trained VGGNet on CIFAR-10. “FT” means “fine-tuning” the pruned model. 46

4.3 Overall performance of pruning ResNet on ImageNet. 47

4.4 Comparison of the theoretical and realistic speedup. We only count the time consumption of the forward procedure. 49

4.5 Accuracy of CIFAR-10 on ResNet-110 under different pruning rate with different filter selection criteria. 51

5.1 Comparison of pruned ResNet on CIFAR-10. In “Fine-tune?” column, “✓” and “✗” indicates whether to use the pre-trained model as initialization or not, respectively. The “Acc. ↓” is the accuracy drop between pruned model and the baseline model, the smaller, the better. 66

5.2 Comparison of pruned ResNet on ILSVRC-2012. “Fine-tune?” and ”acc. ↓” have the same meaning with Table 3.1. 67

5.3 Pruning pre-trained VGGNet on CIFAR-10. “w.o.” means “without” and “FT” means “fine-tuning” the pruned model. 70

5.4 Pruning scratch VGGNet on CIFAR-10. “SA” means “sensitivity analysis”. Without sensitivity analysis, FPGM can still achieve comparable performances comparing to [1]; after introducing sensitivity analysis, FPGM can surpass [1]. 70

5.5 Comparison on the theoretical and realistic acceleration. Only the time consumption of the forward procedure is considered. 71

6.1 Different categories of filter pruning algorithms. “W” and “A” denote the weight-based and activation-based criteria. “O” and “G” indicate the one-shot and greedy pruning. 78

6.2 Comparison of the pruned ResNet on CIFAR-10. In “Init pretrain” column, “✓” and “✗” indicate whether to use the pre-trained model as initialization or not, respectively. The “Acc. ↓” is the accuracy drop between pruned model and the baseline model, the smaller, the better. A negative value in “Acc. ↓” indicates an improved model accuracy. 87

6.3 Comparison of the pruned ResNet on ImageNet. “Init Pretrain” and ”acc. ↓” have the same meaning with Table 3.1. 89

6.4 Comparison of the pruned ResNet-56 on CIFAR-100. 92

6.5 Analysis of adversarial criteria. “w Adv” and “w/o Adv” denote containing the adversarial criteria or not, respectively. 94

ABSTRACT

Accelerating Deep Convolutional Neural Networks via Filter Pruning

by

Yang He

The superior performance of deep Convolutional Neural Networks (CNNs) usually comes from the deeper and wider architectures, which cause the prohibitively expensive computation cost. To reduce the computational cost, works on model compression and acceleration have recently emerged. Among all the directions for this goal, filter pruning has attracted attention in recent studies due to its efficacy. For a better understanding of filter pruning, this thesis explores different aspects of filter pruning, including pruning mechanism, pruning ratio, pruning criteria, and automatic pruning. First, we improve the pruning mechanism with soft filter pruning so that the mistaken pruned filters can have a chance to be recovered. Second, we consider the asymptotic pruning rate to reduce the sudden information loss in the pruning process. Then we explore the pruning criteria to better measure the importance of filters. Finally, we propose the automatic pruning method to save human labor. Our methods lead to superior convolutional neural network acceleration results.

Dissertation directed by Professor Yi Yang

ReLER, Australian Artificial Intelligence Institute, School of Software