FILTER PRUNING FOR COMPACT AND EFFICIENT CONVOLUTIONAL NEURAL NETWORKS

MILTON MONDAL

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2023

FILTER PRUNING FOR COMPACT AND EFFICIENT CONVOLUTIONAL NEURAL NETWORKS

by

MILTON MONDAL

DEPARTMENT OF ELECTRICAL ENGINEERING

submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2023

Dedicated to

My Teachers

Family

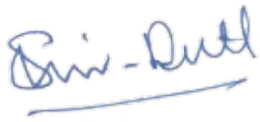
&

Friends

CERTIFICATE

This is to certify that the thesis entitled, "**Filter Pruning for Compact and Efficient Convolutional Neural Networks** ", being submitted by **Mr. Milton Mondal**, to the Indian Institute of Technology Delhi, for the award of degree of **Doctor of Philosophy** in the Department of Electrical Engineering is a record of bonafide research work carried out by him under our supervision and guidance. He has fulfilled the requirements for the submission of the thesis, which to the best of our knowledge has reached the required standard.

The material contained in the thesis has not been submitted in part or full to any other University or Institute for the award of any degree or diploma.



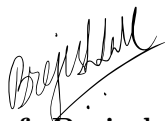
Prof. Shiv Dutt Joshi

Department of Electrical Engineering

Indian Institute of Technology Delhi

New Delhi -110016

Date: April 2023



Prof. Brejesh Lall

Department of Electrical Engineering

Indian Institute of Technology Delhi

New Delhi -110016

Dr. Pushpendra Singh

Department of Electronics & Communication Engineering

National Institute of Technology Hamirpur

Himachal Pradesh - 177005

ACKNOWLEDGEMENTS

I would like to start by thanking my supervisors, Prof. Shiv Dutt Joshi, Prof. Brejesh Lall, and Dr. Pushpendra Singh, for their support, guidance, and encouragement throughout my Ph.D. journey. I am grateful to Prof. Shiv Dutt Joshi, from whom I have learned a lot of subjects with a deeper understanding beyond my research. His trust in my abilities and his encouragement to continue my research with ease, have been a source of inspiration to me. I would also like to extend my sincere gratitude to Prof. Brejesh Lall, who helped me in every possible way to conduct my experiments for advanced deep models. Also, his insightful feedback and constructive criticism have helped me to improve my research. I am obliged to Dr. Pushpendra Singh for his time and effort in guiding me, especially regarding manuscript preparation and revision. My thanks also go to all my SRC members who provided valuable feedback during my semester presentations.

This Ph.D. would not have been possible without the support of my parents, Mr. Rajendra Nath Mondal and Mrs. Anima Mondal, and all my family members. Their unwavering love, sacrifice, and emotional support have been the driving force behind my academic journey. It is a true blessing for me to have such a wonderful family. I am also grateful to my friends, including Himanshu, Bishshoy, and Ayan, for the beautiful memories we have shared. I count myself fortunate to have a friend like Himanshu, who has helped me throughout my academic journey, and with whom I have shared numerous enjoyable moments. My special thanks to Bishshoy, with whom I have discussed the most during my research, and he always supported me whenever I required any assistance.

I would like to express my deep gratitude to my yoga friends Akash, Hardik, Hemant, Sunaina, Shilpa, Mansi, and my yoga teacher, Ms. Haimanti Mukhopadhyay, for their invaluable influence in helping me to lead a healthy and happy life. I am indebted to all my spiritual masters for their teachings and for guiding me to become a better person every day. Lastly, I am eternally grateful to the entire universe for the love, support, and opportunities it has bestowed upon me to make a positive impact on the world.

New Delhi — Milton Mondal

27 April 2023 — (2016EEZ8498)

ABSTRACT

Convolutional neural networks (CNNs) have become the primary tool for solving most computer vision tasks. The networks have become deeper over time to extract more abstract features for improving generalization performance. Currently, workstations and GPU servers only can use these deep CNNs due to their capacity to store a large number of parameters and execute a high number of floating point operations (FLOPs). The main objective of this thesis is to design compact and efficient CNNs. The first question that comes to our mind is whether all filters present in a network are essential for a task or not, or if they are important, are their importance the same or different? To address these questions, we observe the performance of the networks after pruning filters from them. We find that with retraining the pruned model can recover the performance degradation. A lot of filter-pruning algorithms have been proposed in recent years. However, the literature lacks a clear understanding of whether we should prune a filter at a uniform rate or non-uniformly from each layer. To find the answer, we conduct several experiments while pruning the same percent filters from different layers. We observe that the performance drop in each case is different, demonstrating that different layers are sensitive to filter pruning differently. We also discover that existing methods manually set the pruning fraction per layer by observing the layer-wise sensitivity for each layer. However, this manual task of determining the pruning fraction for each layer becomes extremely tedious for very deep models. We solve this problem by designing the filter importance in such a manner that they become comparable across the layers.

We also eliminate some of the major problems of existing filter pruning methods. A filter pruning method can be either feature-dependent or feature-independent. Existing feature-dependent methods do not consider the class label of a training example when calculating filter importance using feature map activations. However, our observations indicate that a feature can have high intensity for one class but not for others, indicating that the feature is useful for that class. Our proposed method, Global Filter Importance based Adaptive Pruning (GFI-AP) resolves this issue by defining filter importance based on the strength of class-specific feature maps. GFI-AP outperforms feature-dependent methods that do not consider class-specific feature map strength. Similarly, the problem in existing feature-independent filter pruning methods is that they determine the filter index that needs to be pruned based only on that filter weight. However, when we prune

a filter from a network, the corresponding channel of all filters in the next layer is also eliminated to maintain structural consistency. Thus, unlike existing methods, our feature-independent pruning method, Filter Pruning by Successive Layers (FPSL) prunes a filter through successive layers analysis. FPSL removes a filter from a layer so that the feature maps of the subsequent layer remain close to their unpruned state. Additionally, FPSL eliminates the need for layer-by-layer retraining, manual per-layer pruning percentage selection, and intensive hyperparameter search for finetuning. In a CNN, there is always a possibility of redundancy in the number of filters used as we do not ensure that different filters should pick up different information of the input. Thus, pruning helps in eliminating those unimportant filters from the base model. Here we propose an alternate approach Multi-band CNN (M-CNN) which modifies the architecture design so that different filters capture complementary bands of the input signal. It is an attempt to prevent the generation of redundant filters in the base model. Incorporating multi-band filtering into CNN has provided a novel means of building a compact and efficient network. The proposed M-CNN maintains the model performance but reduces the number of filters that needs to be trained by a factor of four for the first or last convolutional layer. The advancements and improvements achieved by our proposed methods over existing approaches would enable the deployment of compact and efficient deep learning models in edge devices, allowing them to be more beneficial for vision and other related tasks.

सार

कनवॉल्यूशनल न्यूरल नेटवर्क् (CNN) अधिकांश कंप्यूटर दृष्टि कार्यों को हल करने के लिए प्राथमिक उपकरण बन गए हैं। बेहतर परिणाम पाने के लिए समय के साथ नेटवर्क गहरा हो गया है ताकि और अच्छा फीचर्स पाया जाये। वर्तमान में, वर्कस्टेशन और जीपीयू सर्वर केवल बड़ी संख्या में पैरामीटर्स को संग्रहीत कर पाते है और बड़ी संख्या में फ्लोटिंग पॉइंट ऑपरेशंस (FLOPs) को निष्पादित करने की क्षमता रखते हे इस कारण इन गहरे सीएनएन का उपयोग कर सकते हैं। इस थीसिस का मुख्य उद्देश्य कॉम्पैक्ट और कुशल सीएनएन डिजाइन करना है। सबसे पहला प्रश्न जो हमारे दिमाग में आता है वह यह है कि क्या किसी नेटवर्क में मौजूद सभी फ़िल्टर किसी कार्य के लिए आवश्यक हैं या नहीं, या यदि वे महत्वपूर्ण हैं, तो क्या उनका महत्व समान है या अलग है? इन सवालों के समाधान के लिए, हम नेटवर्क से फिल्टर (filter) की छंटाई के बाद उनके प्रदर्शन का निरीक्षण करते हैं। हम पाते हैं कि छंटनी किए गए मॉडल को फिर से प्रशिक्षित करने से प्रदर्शन में गिरावट को ठीक किया जा सकता है। हाल के वर्षों में बहुत सारे फिल्टर–प्रूनिंग एल्गोरिदम प्रस्तावित किए गए हैं। हालाँकि, मौजूदा तरीके में इस बात की स्पष्ट समझ का अभाव है कि क्या हमें प्रत्येक परत से एक समान दर पर या गैर–समान रूप से फ़िल्टर को छंटाई करनी चाहिए। उत्तर खोजने के लिए, हम अलग–अलग परतों से समान प्रतिशत फ़िल्टर की छंटाई करते हुए कई प्रयोग करते हैं। हम देखते हैं कि प्रत्येक परतों में प्रदर्शन में गिरावट अलग है, यह प्रदर्शित करता है कि अलग–अलग परतें अलग–अलग तरीके से फ़िल्टर छंटाई के प्रति संवेदनशील हैं। हम यह भी पता लगाते हैं कि मौजूदा तरीके प्रत्येक परत के लिए परत–वार संवेदनशीलता को देखते हुए मैन्युअल रूप से प्रति परत छंटाई अंश निर्धारित करते हैं। हालाँकि, प्रत्येक परत के लिए छंटाई अंश निर्धारित करने का यह मानव आधारित कार्य बहुत गहरे मॉडल के लिए अत्यंत कठिन हो जाता है। हम फिल्टर के महत्व को इस तरह से

डिजाइन करके इस समस्या को हल करते हैं कि वे परतों के बीच तुलनीय हो जाते हैं। हम मौजूदा फिल्टर प्रूनिंग विधियों की कुछ प्रमुख समस्याओं को भी समाप्त कर देते हैं। फिल्टर छंटाई विधि फीचर (feature) –निर्भर या फीचर –स्वतंत्र हो सकती है। फ़ीचर सक्रियण का उपयोग करके फ़िल्टर महत्व की गणना करते समय मौजूदा फ़ीचर–निर्भर विधियाँ प्रशिक्षण उदाहरण के वर्ग पर विचार नहीं करती हैं। हालाँकि, हमारी टिप्पणियों से संकेत मिलता है कि एक फ़ीचर, एक वर्ग के लिए उच्च तीव्रता का हो सकती है, लेकिन दूसरों के लिए नहीं, यह दर्शाता है कि यह फ़ीचर उस वर्ग के लिए उपयोगी है। हमारी प्रस्तावित विधि, ग्लोबल फ़िल्टर इंपोर्टेंस–बेस्ड एडेप्टिव प्रूनिंग (GFI–AP) वर्ग–विशिष्ट फीचर सक्रियण की ताकत के आधार पर फ़िल्टर महत्व को परिभाषित करके इस समस्या को हल करती है। जीएफआई–एपी फीचर–निर्भर तरीकों से बेहतर प्रदर्शन करता है जो वर्ग–विशिष्ट फीचर की ताकत पर विचार नहीं करते हैं। इसी तरह, मौजूदा फ़ीचर–स्वतंत्र फ़िल्टर प्रूनिंग विधियों में समस्या यह है कि वे फ़िल्टर महत्व का निर्धारण करते हैं केवल उस फ़िल्टर के आधार पर। हालाँकि, जब हम किसी नेटवर्क से छँटाई करते हैं, तो नेटवर्क का संरचना को बनाए रखने के लिए अगली परत में सभी फ़िल्टर के संबंधित चैनल को भी हटा दिया जाता है। इस प्रकार, मौजूदा तरीकों से अलग, हमारी फीचर–स्वतंत्र प्रूनिंग विधि, सक्सेसिव लेयर्स द्वारा फ़िल्टर प्रूनिंग (FPSL) क्रमिक परतों के विश्लेषण के माध्यम से एक फिल्टर की छंटाई करती है। FPSL एक परतों से एक फिल्टर हटाता है ताकि बाद की परतों के फीचर उनकी पहले का स्थिति के करीब रहें। इसके अतिरिक्त, FPSL परत दर परत पुनर्प्रशिक्षण, मानव आधारित प्रति–परत छंटाई प्रतिशत चयन, और फ़ाइनट्यूनिंग (fine–tuning) के लिए गहन हाइपरपैरामीटर (hyperparameter) खोज की आवश्यकता को समाप्त करता है। CNN में, उपयोग किए जाने वाले फ़िल्टरों की संख्या में हमेशा अतिरेक की संभावना होती है क्योंकि हम यह सुनिश्चित नहीं करते हैं कि विभिन्न फ़िल्टर अलग–अलग जानकारी प्राप्त करें। इस प्रकार, छंटाई मॉडल से उन महत्वहीन फिल्टरों को हटाने में मदद करती है। यहां हम एक वैकल्पिक दृष्टिकोण मल्टी–बैंड सीएनएन (M–CNN) का प्रस्ताव करते हैं जो मॉडल निर्माण को संशोधित करता है ताकि विभिन्न फिल्टर इनपुट सिग्नल के पूरक बैंड को कैप्चर कर सकें। यह बेस मॉडल में निरर्थक फिल्टर के निर्माण को रोकने का एक प्रयास है। CNN में मल्टी–बैंड फ़िल्टरिंग को शामिल करने से एक कॉम्पैक्ट और कुशल नेटवर्क बनाने का एक नया साधन उपलब्ध हुआ है। प्रस्तावित M–CNN मॉडल के प्रदर्शन को बनाए रखता है, लेकिन पहले या अंतिम संकेंद्रित परत

के लिए चार के एक कारक द्वारा प्रशिक्षित किए जाने वाले फिल्टर की संख्या को कम करता है। मौजूदा दृष्टिकोणों पर हमारे प्रस्तावित तरीकों द्वारा हासिल की गई उन्नति और सुधार बढ़त उपकरणों में कॉम्पैक्ट और कुशल गहन मॉडल की तैनाती को सक्षम करेगा, जिससे वे दृष्टि और अन्य संबंधित कार्यों के लिए अधिक फायदेमंद हो सकेंगे।

Contents

List of Abbreviations

AP Adaptive Pruning

APN Adaptively Pruned Network

BLAS Basic Linear Algebra Subprograms

CNN Convolutional Neural Network

DCT Discrete Cosine Transform

DNN Deep Neural Network

FCFN Fully Connected Feedforward Neural network

FLOPs Floating Point Operations

FPSL Filter Pruning by Successive Layers

GFI-AP Global Filter Importance based Adaptive Pruning

GPF Global Pruning Fraction

GPU Graphics Processing Unit

M-CNN Multi-band Convolutional Neural Network

PCA Principal Component Analysis

RPF Restricted Pruning Fraction

UP Uniform Pruning

UPN Uniformly Pruned Network

List of Figures

List of Tables