

IMPROVEMENTS IN LEARNING CAPACITIES OF CONVOLUTIONAL NEURAL NETWORKS

BISHSHOY DAS



**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

DECEMBER 2023

© Indian Institute of Technology Delhi (IITD), New Delhi, 2023

IMPROVEMENTS IN LEARNING CAPACITIES OF CONVOLUTIONAL NEURAL NETWORKS

by

BISHSHOY DAS

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

DECEMBER 2023

*To my beloved mother
and my beloved father*

THESIS CERTIFICATE

This is to certify that the thesis entitled “*Improvements in Learning Capacities of Convolutional Neural Networks*” being submitted by Bishshoy Das to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of the degree of Doctor of Philosophy is the record of the bonafide research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted either in part or in full to any other university or institute for the award of any degree or diploma.

Prof. Sumantra Dutta Roy

Professor, Dept. of Electrical Engineering

Indian Institute of Technology Delhi

Hauz Khas, New Delhi, Delhi - 110016

DECLARATION

I hereby certify that :-

1. The work contained in the thesis is original and has been done by me under the guidance of my supervisor.
2. The work has not been submitted to any other institute for any other degree or diploma.
3. I have followed the guidelines provided by the institute in preparing the thesis.
4. I have conformed to ethical norms and guidelines while writing the thesis.
5. Whenever I have used materials (data, models, figures, text) from other sources, I have given due credit by citing them in the text of the thesis, and giving their details in the references, and taken permission from the copyright owners of the sources whenever necessary.



Bishshoy Das, 05.12.2023

Dept. of Electrical Engineering

Indian Institute of Technology Delhi

Hauz Khas, New Delhi, Delhi - 110016

ACKNOWLEDGEMENT

I would like to thank Professor Sumantra Dutta Roy for his assistance and guidance during my PhD. I learned a great deal from him in the classroom, but I learned much more from our interactions and talks outside of the classroom.

I would also want to thank Professors Shiv Dutt Joshi and Brejesh Lall for their insightful recommendations and methodical analyses, which helped me advance my thoughts when I was stuck.

I would also want to thank Professor Manas Kamal Bhuyan of Indian Institute of Technology Guwahati, who inspired and encouraged me to pursue an academic career.

Finally, I would like to thank my colleagues and friends Milton Mondal, Martin Cheerangal, Gopala Rao D, Siranjeevi Gurumani, Karthikeya G S, Himanshu Padole, and Santosh Kumar Gedela for their intellectual and emotional support, without which I would not have been able to endure my long PhD trip. During the course of my PhD, I also served as a mentor to numerous M. Tech students, who offered me the opportunity to instruct them on fundamental ideas and help them through their projects. Regarding this, I would like to thank Srijeet Chatterjee, Hareesh Kumawat, Sakshi Agrawal, Asmita Nandkumar Patil, Abhinav Gaur, and Panchal Sanket Hitendrakumar for their stimulating discussions.

ABSTRACT

In this thesis, we explore different methods of improving the utilization of a convolutional neural network’s learning capacity. We define what we mean by the learning capacity of a neural network, since no standard definition of learning capacity exists in the literature hitherto. Based on our definition of learning capacity, we formulate two scenarios in which the learning capacity of a CNN can get significantly affected, (i) Intra-network learning and (ii) Inter-network learning. In both cases, we explore in-depth as to what is the cause behind the lack of utilization of learning capacities. Then, we formulate targeted solutions to solve for each of the discovered lacunae. For example in our explorations with intra-network learning, we discover that in non-BatchNormalized networks, some filters or even some layers completely goes to zero during training, a phenomena which we call zeroing out. This leads to a drop in the utilization of the network’s trainable parameters and hence an under-utilization of learning capacities. We address this phenomena by proposing **PowerGrad Transform** that modifies the gradient before the backward pass and with that we mitigate the zeroing out phenomena and improve final accuracies of non-BatchNormalized convolutional neural networks. We also apply PowerGrad Transform to BatchNormalized networks. We observe that PowerGrad transform can help the network learn parameter that are outside of the scope of standard gradient descent. Even with high loss values, networks can achieve higher final convergence accuracies in both training and test datasets. This leads us to believe that transforming the

gradient through PowerGrad transform leads the network to better utilize its learning capacity.

In our explorations of inter-network learning, we see that when two or more models are trained from random initialization, often the final converged weights have significant overlaps leading to a lower utilization of pool learning capacity. Ensembles of such networks do not improve the performance (or marginally improves) the performance over any one single network in the same pool. This points to the observation that the total learning capacity of the pool as a whole is severely under-utilized. In order to remedy this, we propose adversarial training through **Feature Difference Losses**, a method in which all feature vectors are shared amongst worker networks in a pool, and each worker adversarially pushes each other to learn different feature maps, leading to a forced outcome where at convergence each network arrives at vastly different optima on the loss landscape. This in turn improves ensemble performance as it leads to a diversity in the networks' features, which directly translates to an increased utilization of the pool's learning capacity.

सार

इस थीसिस में, हम एक कन्वोलुशनल न्यूरल नेटवर्क की सीखने की क्षमता के उपयोग में सुधार के विभिन्न तरीकों का पता लगाते हैं। हम परिभाषित करते हैं कि न्यूरल नेटवर्क की सीखने की क्षमता से हमारा क्या मतलब है, क्योंकि अब तक साहित्य में सीखने की क्षमता की कोई परिभाषा मौजूद नहीं है। सीखने की क्षमता की हमारी परिभाषा के आधार पर, हम दो परिदृश्य तैयार करते हैं जिनमें सीएनएन की सीखने की क्षमता महत्वपूर्ण रूप से प्रभावित हो सकती है, (i) इंट्रा-नेटवर्क लर्निंग और (ii) इंटर-नेटवर्क लर्निंग। दोनों ही मामलों में, हम गहराई से पता लगाते हैं कि सीखने की क्षमताओं के उपयोग की कमी के पीछे क्या कारण है। फिर, हम प्रत्येक खोजी गई कमी को हल करने के लिए लक्षित समाधान तैयार करते हैं। इंट्रा-नेटवर्क लर्निंग के साथ हमारे अन्वेषणों में उदाहरण, हम पाते हैं कि गैर-बैच-नॉर्मलाइज्ड नेटवर्क में, कुछ फिल्टर या यहां तक कि कुछ लेयर प्रशिक्षण के दौरान पूरी तरह से शून्य हो जाती हैं, इस घटना को हम जीरो आउट हो जाना कहते हैं। इससे नेटवर्क के प्रशिक्षण योग्य मापदंडों के उपयोग में गिरावट आती है और सीखने की क्षमताओं का कम उपयोग होता है। हम इस घटना को हल करने के लिये पावरग्रेड ट्रांसफॉर्म का प्रस्ताव करते हैं जो बैकवर्ड पास से पहले ग्रेडिएंट को संशोधित करता है और इसके साथ हम जीरो आउट घटना को कम करते हैं और गैर-बैच नॉर्मलाइज्ड कन्वोलुशनल न्यूरल नेटवर्क की अंतिम सटीकता में सुधार करते हैं। हम पावरग्रेड ट्रांसफॉर्म को बैच नॉर्मलाइज्ड नेटवर्क पर भी लागू करते हैं। हम देखते हैं कि पावरग्रेड ट्रांसफॉर्म नेटवर्क को बेहतर पैरामीटर सीखने में मदद कर सकता है जो मानक ग्रेडिएंट डिसेंट के दायरे से बाहर हैं। उच्च हानि मूल्यों के साथ भी, नेटवर्क प्रशिक्षण और परीक्षण डेटासेट दोनों में उच्च अंतिम अभिसरण सटीकता प्राप्त कर सकते हैं। यह हमें यह विश्वास दिलाता है कि पावरग्रेड ट्रांसफॉर्म के माध्यम से ग्रेडिएंट को

बदलने से नेटवर्क अपनी सीखने की क्षमता का बेहतर उपयोग कर सकता है।

इंटर-नेटवर्क लर्निंग के हमारे अन्वेषणों में, हम देखते हैं कि जब दो या दो से अधिक मॉडलों को रैंडम इनिशियलाइज़ेशन से प्रशिक्षित किया जाता है, तो अक्सर अंतिम कनभारजड वेइट्स में महत्वपूर्ण ओवरलैप होते हैं जिससे पूल लर्निंग कैपेसिटी का कम उपयोग होता है। इस तरह के नेटवर्क के समूह एक ही पूल में किसी एक नेटवर्क पर प्रदर्शन में सुधार (या मोटे तौर पर सुधार) नहीं करते हैं। यह इस अवलोकन की ओर इशारा करता है कि समग्र रूप से पूल की कुल सीखने की क्षमता का बहुत कम उपयोग किया जाता है। इसका समाधान करने के लिए हम फीचर डिफरेंस लॉस के माध्यम से प्रतिकूल प्रशिक्षण का प्रस्ताव करते हैं, एक विधि जिसमें सभी फीचर वेक्टर को एक पूल में श्रमिक नेटवर्क के बीच साझा किया जाता है, और प्रत्येक कार्यरत नेटवर्क प्रतिकूल रूप से एक-दूसरे को विभिन्न फीचर मैप्स सीखने के लिए दबाव डालता है, जिससे एक मजबूर परिणाम होता है जहां अभिसरण पर प्रत्येक नेटवर्क नुकसान के परिदृश्य पर काफी अलग ऑप्टिमा पर पहुंचता है। यह बदले में समूह प्रदर्शन में सुधार करता है क्योंकि यह नेटवर्क की फीचर्स में विविधता लाता है, जो सीधे पूल की सीखने की क्षमता के उपयोग में वृद्धि करता है।

TABLE OF CONTENTS

	Thesis Certificate	i
	Declaration	ii
	Acknowledgement	iii
	Abstract	iv
	संर	vi
	List of Figures	xi
	List of Tables	xii
1	Introduction	1
1.1	Learning Capacity	2
1.2	Motivation	6
1.3	Contributions	7
1.4	Objective Formulation	8
1.4.1	Inter-network learning capacity	9
1.4.2	Intra-network learning capacity	10
1.5	Organization of the Thesis	12
2	Mathematical Preliminaries	14
2.1	Convolutional layers	14
2.2	Pooling layers	17
2.3	Residual connections	18
2.4	Activation layers	19
2.5	Normalization layers	20

2.6	Fully-connected layers	21
2.7	Softmax layer	22
2.8	Loss functions	22
2.9	Hyperparameters & Regularization	23
2.10	Backpropagation & Optimization	24
3	Related Works	29
4	Knowledge Diversification	35
4.1	Motivation	35
4.2	Introduction	37
4.3	Method	38
4.3.1	Feature Difference Loss functions	40
4.3.2	Training Phases	42
4.3.3	Phase 0	42
4.3.4	Phase 1 & 2	43
4.3.5	Phase 3	44
4.3.6	Phase 4	46
4.4	Experiments	46
4.5	Ablation studies	52
4.6	Discussions	54
5	PowerGrad Transform	57
5.1	Motivation	57
5.2	Introduction	58
5.3	Method	63
5.4	Analysis	65

5.4.1	Properties of PGT	65
5.4.2	Effect on Logits	68
5.4.3	Effect on Weight Gradients	70
5.4.4	Effect on Class Separation	72
5.5	Experiments	73
5.5.1	BatchNormalized networks	73
5.5.2	Non-BatchNormalized networks	75
5.5.3	Grid search and Effects on Loss	82
5.6	Ablation studies	84
5.6	Implementation	85
5.7	Discussions	86
6	Conclusion	89
6.1	Knowledge Diversification	90
6.2	PowerGrad Transform	92
7	List of Publications	95
8	References	97
	Author's Curriculum Vitae	118

LIST OF FIGURES

1	Problem 1: Improving inter-network learning capacity	10
2	Problem 2: Improving intra-network learning capacity	11
3	Architecture of the ensemble head network	40
4	Architecture of four networks trained with FDL	41
5	Training pipeline of two networks trained with FDL	44
6	Networks used for FDL experimentation.	49
7	Filters vs Accuracy of simple ConvNet with and w/o FDL	49
8	Multi-network FDL ensemble accuracies	51
9	Several plots for FDL ablation studies	54
10	Feature map differences with and without FDL	55
11	Gradient alteration types	61
12	Effect of PGT on probability distributions	65
13	Analysis of PGT at the fully connected layer	73
14	PGT Training/Test accuracy vs. baseline for ResNet-50	76
15	ResNet-18 architecture with layer indices	80
16	The Zeroing Out phenomena	81
17	Norm vs. Iteration plot of layer 5 of non-BN ResNet-18	82
18	Experiment demonstrating the efficacy of PGT	83
19	Effects of PGT: various statistical measures	86
20	PGT Grid search for various values of α	87
21	Python code of PowerGrad Transform based on PyTorch	90

LIST OF TABLES

1	FDL: Experiments on MNIST	50
2	Comparisons of ensemble methods vs. FDL on CIFAR datasets	51
3	Comparisons of ensemble methods vs. FDL on ImageNet-1K datasets	51
4	Multi-network FDL ensembles with CIFAR-10	52
5	Multi-network FDL ensembles with CIFAR-100	52
6	FDL experiment hyperparameters	53
7	Ablation studies using FDL (1)	54
8	Ablation studies using FDL (2)	55
9	PGT: Comparison table for various CNNs	78
10	PGT: Results for non-BN ResNets	85
11	PGT: Ablation studies	89