UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Understanding deep learning through ultra-wide neural networks

by

Wei Huang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2021

Certificate of Authorship/Originality

I, Wei Huang, declare that this thesis, is submitted in fulfilment of the requirements for the award of PhD, in the Faculty of Engineering and IT at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference 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 Austraian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication. Date: 09/09/2021

Acknowledgements

In the first place, I want express my sincere gratitude to my supervisor, professor Richard Xu, for his unwavering support and invaluable instructions. Prof. Xu has brought me into the topics of non-parametric Bayesian and deep learning. He taught me a lot of machine learning knowledge by insisting on teaching every week. By revising the paper along with the discussion, he taught me how to write scientific papers. His profound knowledge and rigorous scientific research spirit deeply influenced the research path I would not have been able to carry out research projects without his valuable suggestions.

I would like to extend my sincere thanks to my co-supervisor Prof. Massimo Piccardi for helping with my candidature assessment, and I am greatly inspired by his warm support.

I am grateful to my collaborator Mr. Weitao Du who has been providing valuable support on mathematical techniques and proofs. We have been best friends since undergraduate students. We have spent plenty of time on discussion and collaborating. I have improved a lot in mathematical proofs. I sincerely appreciate my collaborator Mr. Yayong Li for his considerable discussion on the graph neural network. I am thankful for what I have learned from him during our collaborations.

I would like to offer my special thanks to all other members of our machine learning lab, especially, Mr. Chunrui Liu, Mr. Yunce Zhao, Mr. Ziyue Zhang, Mr. Xuan Liang, Mr. Chen Deng, Dr. Caoyuan Li, Miss Ying Li, Miss Sophie Yuan, Mr. Chris Markos, Dr. Jason Traish, Miss Ningkai Xiao, Miss Erica Huang, Mr. Haodong Chang, and Dr. Shawn Jiang.

My sincere thanks go to my friends for a happy time with them during my Ph.D. study. Those thanks especially go to Dr. Haihan Sun, Dr. Ye Shi, Miss Xuan Wang, Miss Ye Su, Mr. Feng Shan, Mr. Haiming Qian, Mr. Leijie Zhang, Dr. Hongwen Yu, Dr. Tianyu Yang, Dr. Wentian Zhang, Miss Maral Ansari, Mr. Li Wang and Miss Chen Wang.

Finally, I would like to express my forever thanks to my parents for their selfless love and eternal support. My parents will always be my strong backing.

ABSTRACT

Understanding deep learning through ultra-wide neural networks

by

Wei Huang

Deep learning has been responsible for a step-change in performance across machine learning, setting new benchmarks in a large number of applications. However, the existing accounts fail to resolve why deep learning can achieve such great success. There is an urgent need to address the deep learning theory caused by the demand of understanding the principles of deep learning. One promising theoretical tool is the infinitely-wide neural network. This thesis focuses on the expressive power and optimization property of deep neural networks through investigating ultra-wide networks with four main contributions.

We first use the mean-field theory to study the expressivity of deep dropout networks. The traditional mean-field analysis adopts the gradient independence assumption that weights used during feed-forward are drawn independently from the ones used in backpropagation, which is not how neural networks are trained in a real setting. By breaking the independence assumption in the mean-field theory, we perform theoretical computation on linear dropout networks and a series of experiments on dropout networks. Furthermore, we investigate the maximum trainable length for deep dropout networks through a series of experiments and provide a more precise empirical formula that describes the trainable length than the original work.

Secondly, we study the dynamics of fully-connected, wide, and nonlinear networks with orthogonal initialization via neural tangent kernel (NTK). Through a series of propositions and lemmas, we prove that two NTKs, one corresponding to Gaussian weights and one to orthogonal weights, are equal when the network width

is infinite. This suggests that the orthogonal initialization cannot speed up training in the NTK regime. Last, with a thorough empirical investigation, we find that orthogonal initialization increases learning speeds in scenarios with a large learning rate or large depth.

The third contribution is characterizing the implicit bias effect of deep linear networks for binary classification using the logistic loss with a large learning rate. We claim that depending on the separation conditions of data, the loss will find a flatter minimum with a large learning rate. We rigorously prove this claim under the assumption of degenerate data by overcoming the difficulty of the non-constant Hessian of logistic loss and further characterize the behavior of loss and Hessian for non-separable data.

Finally, we demonstrate the trainability of deep Graph Convolutional Networks (GCNs) by studying the Gaussian Process Kernel (GPK) and Graph Neural Tangent Kernel (GNTK) of an infinitely-wide GCN, corresponding to the analysis on expressivity and trainability, respectively. We formulate the asymptotic behaviors of GNTK in the large depth, which enables us to reveal the dropping trainability of wide and deep GCNs at an exponential rate.

List of Publications

Journal Papers

J-1. **Wei Huang** and Ricahrd Yi Da Xu, “Gaussian Process Latent Variable Model Factorization for Context-aware Recommender Systems," 2019. Submitted to *Pattern Recognition Letters*.

Conference Papers

C-1. **Wei Huang**, Richard Yi Da Xu, Weitao Du, Yutian Zeng, and Yunce Zhao, “Mean field theory for deep dropout networks: digging up gradient backpropagation deeply,” *ECAI 2020, 24th European Conference on Artificial Intelligence.*

C-2. **Wei Huang**, Weitao Du, and Richard Yi Da Xu, “On the neural tangent kernel of deep networks with orthogonal initialization,” 2020. Submitted to IJCAI 2021.

C-3. **Wei Huang**, Weitao Du, Richard Yi Da Xu, and Chunrui Liu, “Implicit bias of deep linear networks in the large learning rate phase,” 2020. Submitted to IJCAI 2021.

C-4. **Wei Huang**, Yayong Li, Weitao Du, Richard Yi Da Xu, Jie Yin, and Ling Chen, “Wide Graph Neural Networks: Aggregation ProvablyLeads to Exponentially Trainability Loss,” 2021. Submitted to ICML 2021.

Contents

List of Figures

Abbreviation

ML - Machine Learning

DL - Deep Learning

FCN - Fully-Connected Network

MLP - Multi-Layer Perceptro

GP - Gaussian Process

GPK - Gaussian Process Kernel

GNTK - Graph Neural Tangent Kernel

GD - Gradient Descent

Resnet - Residual Network

ReLU - Rectified Linear Unit

Erf - Error Function

FIM - Fisher Information Matrix

CLT - Central Limit Theorem

MSE - Mean Squared Error

SVM - Support Vector Machine

GCN - Graph Convolutional Network