C02029: Doctor of Philosophy
CRICOS Code: 009469A
PhD Thesis: Information Technology
May 2021

Automated Deep Learning:
A Study on Neural Architecture Search

Miao Zhang

School of Biomedical Engineering
Faculty of Eng. & IT
University of Technology Sydney
NSW - 2007, Australia

Automated Deep Learning:
A Study on Neural Architecture Search

A thesis submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Information Technology

by

Miao Zhang

to

School of Biomedical Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

May 2021

Abstract

Deep Learning (DL) has shown its superiority in various research areas in recent years, including computer vision, natural language processing, and autonomous driving. Through designing different deep neural networks (DNNs), deep learning techniques have achieved the state-of-the-art performance in numerous real-world applications. Deep neural network has become the first choice for most researchers when solving different machine learning problems. However, the performance of deep neural networks is very sensitive to the structures, and engineers need to choose or design appropriate network structures through tedious and repeated experiments so that deep neural networks can reach their potentials for different problems. Automated Deep Learning (AutoDL) aims to build a better deep learning model in a data-driven and automated manner, so that most practitioners in deep learning can also build a high-performance machine learning model, with being relieved from a labor-intensive and time-consuming neural network design process. AutoDL can bring new research ideas to deep neural networks, and lower the threshold of deep learning in various research areas through automated neural network design.

The process of automated neural network design is termed as Neural Architecture Search (NAS). As the name suggested, the goal of NAS is to automatically design deep neural networks without human intervention. Most recent works on NAS adopt a *weight-sharing* paradigm to find competitive architectures with greatly reducing the computational complexity. Instead of separating training architectures, weight sharing strategy encodes the whole search space as a supernet, and all neural networks directly inherit weights from the supernet for evaluation without needing to be trained from scratch. Pioneer studies on weight-sharing NAS follow two sequential steps. They first adopt an architecture sampling controller to sample architectures for the supernet training. Then, a heuristic search method is adopted to search promising architectures over a discrete search space based on the trained supernet. Since only the supernet is trained for once, this paradigm is also called as *one-shot NAS*. To further improve the efficiency, later researches further employ the continuous relaxation to make the neural architecture differentiable, so that gradient descent can be used to optimize the architecture with respect to validation accuracy, and this paradigm is also referred to as *differentiable NAS*. This thesis focuses on the two specific research directions: *one-shot NAS* and *differentiable NAS*.

Most state-of-the-art one-shot NAS methods use the validation accuracy based on inheriting weights from the supernet as the stepping stone to search for the best performing architecture, adopting a bilevel optimization pattern with assuming this validation accuracy approximates to the test accuracy after re-training. However, recent works have found that there is no positive

correlation between the above validation accuracy and test accuracy for these weight-sharing methods, and this reward based sampling for supernet training also entails the rich-get-richer problem. To handle this deceptive problem, **Chapter 2** presents a new approach, **E**fficient **N**ovelty-driven **N**eural **A**rchitecture **S**earch (EN^{2}AS), to sample the most abnormal architecture to train the supernet. Specifically, a single-path supernet is adopted, and only the weights of a single architecture sampled by our novelty search are optimized in each step to reduce the memory demand greatly. Experiments demonstrate the effectiveness and efficiency of our novelty search based architecture sampling method.

Although one-shot NAS significantly improves the computational efficiency, it also introduces multi-model forgetting during the supernet training, where the performance of previous architectures degrades when sequentially training new architectures with partially-shared weights. To overcome such catastrophic forgetting, **Chapter 3** formulates the supernet training in the one-shot NAS as a constrained optimization problem of continual learning that the learning of current architecture should not degrade the performance of previous architectures during the supernet training. We propose a Novelty Search based Architecture Selection (**NSAS**) loss function and demonstrate that the posterior probability could be calculated without the strict assumption when maximizing the diversity of the selected constraints. Extensive experiments demonstrate that our method enhances the predictive ability of the supernet in one-shot NAS and achieves remarkable performance on CIFAR-10, CIFAR-100, and PTB with efficiency.

Existing works on differentiable NAS adopt a bilevel optimization to alternatively optimize the supernet weights and architecture parameters after relaxing the discrete search space into differentiable, to further improve the efficiency. However, there is non-negligible incongruence in this simple transformation, and it is hard to guarantee that the differentiable optimization in the continuous latent space is equivalent to the optimization in the discrete space. In **Chapter 4**, we utilize a variational graph autoencoder to injectively transform discrete architecture space into an equivalently continuous latent space, to resolve the incongruence. We further devise a probabilistic exploration enhancement method to encourage intelligent exploration during the architecture search in latent space. The catastrophic forgetting is an inevitable problem in weight-sharing NAS, which deteriorates supernet predictive ability and makes the bilevel optimization inefficient in differentiable NAS. This paper proposes an architecture complementation method to relieve this deficiency in differentiable NAS. We analyze the effectiveness of the proposed method in differentiable NAS, and a series of experiments have been conducted to compare the proposed method with state-of-the-art differentiable NAS methods.

Despite notable benefits on computational efficiency from differentiable NAS, more recent works find that existing differentiable NAS techniques struggle to outperform naive baselines, yielding deteriorative architectures as the search proceeds. Rather than directly optimizing the architecture parameters, **Chapter 5** formulates the neural architecture search as a distribution learning problem through relaxing the architecture weights into Gaussian distributions. By leveraging the recently-proposed natural-gradient variational inference (NGVI), the architecture distribution can be easily optimized based on existing codebases without incurring more memory and computational consumption. We demonstrate how the differentiable NAS benefits from Bayesian principles, enhancing exploration and improving stability. The experimental results on benchmark datasets confirm the significant improvements the proposed framework can make. Furthermore, to enhance the searched architectures' transferability in the complicated search space, we propose a simple yet effective depth-aware differentiable neural architecture search. Specifically, we achieve state-of-the-art results on the NAS-Bench-201 and NAS-Bench-1Shot1 benchmark datasets. Our best architecture in the DARTS search space also obtains competitive test errors with 2.37%, 15.72%, and 24.2% on CIFAR-10, CIFAR-100, and ImageNet datasets, respectively.

While much has been discussed about several potentially fatal factors in DARTS, the architecture gradient, a.k.a. hypergradient, has received less attention. In **Chapter 6**, we tackle the hypergradient computation in DARTS based on the implicit function theorem, making it only depends on the obtained solution to the inner-loop optimization and agnostic to the optimization path. To further reduce the computational requirements, we formulate a stochastic hypergradient approximation for differentiable NAS, and theoretically show that the architecture optimization with the proposed method, named iDARTS, is expected to converge to a stationary point. Comprehensive experiments on two NAS benchmark search spaces and the common NAS search space verify the effectiveness of our proposed method. It leads to architectures outperforming, with large margins, those learned by the baseline methods.

Author's declaration

I, *Miao Zhang* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Biomedical Engineering*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia, 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.

Production Note:
SIGNATURE: Signature removed prior to publication.
[Miao Zhang]

DATE: 13th May, 2021
PLACE: Sydney, Australia

Dedication

To my dear parents, brother, and lovely Shufen …

ACKNOWLEDGMENTS

I would like to express my earnest thanks to my supervisors, A/Professor Steven Su, Dr. Shirui Pan, Professor Huiqi Li, and Dr. Steve Ling. Without the tremendous support and guidance from them, this thesis cannot be finished. First of all, I would like to sincerely thank my main supervisor, Professor Su, who led me into the realm of machine learning and deep learning. In the last three years of studying with Professor Su, I have learned a lot of research skills and theoretic foundation knowledge. Furthermore, Professor Su's broad research horizon and rigorous research attitude have benefited me a lot, and these will also have a positive impact in my future work and study. His optimistic and open-minded character are what I have been learning. Dr. Shirui Pan is my associate supervisor. From Dr. Shirui Pan, I learned the rigor and patience in research. His hard-working and self-motivation have set a good example for me.

I also sincerely thank the other mentors in Australia and China, including Dr. Steve Ling, Dr. Xiaojun Chang, and Professor Huiqi Li. I would also like to thank all the staffs in the School of Biomedical Engineering, University of Technology Sydney. Their help makes my research and life in UTS much easier. I will remember them all in my heart.

I would also like to thank all mates at UTS who all had a positive impact on my life and research, Taoping Liu, Li Wang, Yanhao Zhang, Huan Yu, Yongbo Chen, Jiaheng Zhao, Fang Bai, Kairui Guo, Wentian Zhang, Yao Huang, Juan Lyu, Hairong Yu, Ye Shi, Wei Huang. I will also express my gratitude to my friends, Xinyu Lin, Zuyi Chen, Fangzhou Liu, Hantang Liu, Lin Tian, Ming Wei, Qian Li, Huasu Jin, Jiawei Chen, Haocheng Liu, Kang Liu, Mingkun Pei. Their accompany and encouragement also give me a lot of support during my PhD research life.

Lastly, and above all, I would like to thank my parents, brother and people loved me for their selfless love and support. I wish them health and happiness.

List of Publications

Related to the Thesis :

- Chapter 2:

  1. Miao Zhang, Huiqi Li, Shirui Pan, Taoping Liu, Steven Su, One-Shot Neural Architecture Search via Novelty Driven Sampling, In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2020 [162]. ***CORE Rank A****

- Chapter 3:

  2. Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Steven Su, Overcoming Multi-Model Forgetting in One-Shot NAS with Diversity Maximization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020 [159]. ***CORE Rank A****
  3. Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, and Steven Su, One-Shot Neural Architecture Search: Maximising Diversity to Overcome Catastrophic Forgetting. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020 [161]. ***CORE Rank A****

- Chapter 4:

  4. Miao Zhang, Huiqi Li, Shirui Pan, Xiaojun Chang, Zongyuan Ge, Steven Su, Differentiable Neural Architecture Search in Equivalent Space with Exploration Enhancement. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020 [158]. ***CORE Rank A****
  5. Miao Zhang, Steven Su, Shirui Pan, Xiaojun Chang, Huiqi Li, Gholamreza Haffari, Differentiable Neural Architecture Search in Equivalent Space with Enhancing Exploration and Relieving Forgetting. Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. ***CORE Rank A****

- Chapter 5:

  6. Miao Zhang, Steven Su, Shirui Pan, Xiaojun Chang, Li Wang, Gholamreza Haffari, Differentiable Neural Architecture Search via Bayesian Learning Rule. Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. ***CORE Rank A****

- Chapter 6:

  7. Miao Zhang, Steven Su, Shirui Pan, Xiaojun Chang, Huiqi Li, Ehsan Abbasnejad, Gholamreza Haffari, iDARTS: Differentiable Architecture Search with Stochastic Implicit Gradients. Accepted by *International Conference on Machine Learning (ICML)*, 2021 [165]. ***CORE Rank A****

OTHERS :

- **8. Miao Zhang**, Huiqi Li, Steven Su, High Dimensional Bayesian Optimization via Supervised Dimension Reduction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, ***CORE Rank A****
- **9. Miao Zhang**, Huiqi Li, Juan Lyu, Steve Ling, Steven Su, Hyperparameter Optimization with Non-stationary Kernel for CNN based Lung Nodule Classification. In *IEEE Transaction on Evolutionary Computing (TEvC)*, 2021, ***CORE Rank A****
- **10. Miao Zhang**, Steven Su, Shirui Pan, Xiaojun Chang, Gholamreza Haffari, Differentiable Architecture Search Without Training Nor Labels: A Pruning Perspective. In Submitting to *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021, ***CORE Rank A****

TABLE OF CONTENTS

List of Figures

LIST OF TABLES