

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, **Efficient Novelty-driven Neural Architecture Search (EN²AS)**, 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 Publications	xi
List of Figures	xvii
List of Tables	xix
1 Introduction	1
1.1 Background: Deep Learning and Automated Deep Learning	1
1.1.1 Deep Learning	1
1.1.2 Automated Deep Learning	3
1.2 Literature Review	6
1.2.1 Performance Prediction	6
1.2.2 Weights Generation	7
1.2.3 Weights Sharing	7
1.2.4 NAS Search Spaces	9
1.3 Motivations and Challenges	13
1.3.1 Motivations	13
1.3.2 Challenges	13
1.4 Thesis Contributions	15
1.5 Thesis Structure	17
I One-Shot Neural Architecture Search	19
2 One-Shot NAS via Novelty Driven Sampling	21
2.1 Introduction	21
2.2 Problem Definition and Preliminaries	23
2.2.1 Neural Architecture Search	23
2.2.2 One-Shot Neural Architecture Search	24

TABLE OF CONTENTS

2.2.3	Novelty Search	24
2.3	Efficient Novelty-driven Neural Architecture Search	24
2.3.1	Single Path Supernet Training based on Novelty Search	25
2.3.2	Model Selection	26
2.4	Experimental Result	27
2.4.1	Architecture Search for Convolutional Cells	28
2.4.2	Architecture Search for Recurrent Cells	29
2.4.3	Empirical Comparison with Baselines	30
2.4.4	Experiments on Benchmark Dataset	33
2.5	Chapter Summary and Discussion	33
3	Overcoming Multi-Model Forgetting in One-Shot NAS	35
3.1	Introduction	35
3.2	Preliminaries	38
3.2.1	Catastrophic Forgetting	38
3.2.2	Multi-model Forgetting	38
3.3	Novelty Search based Architecture Selection Loss Function	40
3.3.1	Problem Formulation	40
3.3.2	Constraints Selection based on Novelty Search	40
3.3.3	The NSAS Loss Function	42
3.3.4	From Weight Plasticity Loss (WPL) to NSAS	42
3.3.5	One-Shot NAS with Novelty Search based Architecture Selection	43
3.4	Experimental Result	44
3.4.1	Experimental Results on Common Search Space	44
3.4.2	Experimental Results on NAS-Bench-201	50
3.5	Chapter Summary and Discussion	55
II	Differentiable Neural Architecture Search	57
4	Differentiable Neural Architecture Search with Exploration Enhancement	59
4.1	Introduction	59
4.2	Problem Definition and Preliminaries	62
4.2.1	Weight-Sharing NAS	62
4.2.2	Differentiable NAS	64

4.3	Exploration Enhancing Neural Architecture Search with Architecture Completion	67
4.3.1	Exploration Enhancement in the Latent Space	67
4.3.2	Overcoming Multi-Model Forgetting through Architecture Completion	70
4.3.3	Regularization based Differentiable NAS	73
4.4	Experimental Result	74
4.4.1	Experiments on the Benchmark Dataset	74
4.4.2	Experiments on DARTS Search Space	80
4.5	Chapter Summary and Discussion	84
5	Differentiable Neural Architecture Search via Bayesian Learning Rule	87
5.1	Introduction	87
5.2	Preliminaries	89
5.2.1	Differentiable Neural Architecture Search	89
5.2.2	Distribution learning based NAS	91
5.2.3	Deep Learning with Bayesian Principles	92
5.3	Bayesian Learning Rule for Neural Architecture Search (BaLeNAS)	93
5.3.1	Formulating NAS as Distribution Learning	93
5.3.2	Natural-Gradient Variational Inference for NAS	94
5.3.3	Implicit Regularization with MCMC Sampling	95
5.3.4	Depth-Aware Regularization for BaLeNAS	96
5.4	Experimental Result	98
5.4.1	Experiments on Benchmark Datasets	98
5.4.2	Experiments on DARTS Search Space	100
5.4.3	Ablation Study of MCMC on NAS-Bench-201	102
5.4.4	Ablation Study on the Effect of Exploration	102
5.4.5	Tracking of the Hessian norm	104
5.5	Chapter Summary and Discussion	104
6	Differentiable Architecture Search with Stochastic Implicit Gradients	107
6.1	Introduction	107
6.2	Preliminaries: Hypergradient Approximation in DARTS	109
6.2.1	One-step Unrolled Differentiation	110
6.2.2	Reverse-mode Back-propagation.	111
6.3	Differentiable Architecture Search with Stochastic Implicit Gradients	112

TABLE OF CONTENTS

6.3.1	iDARTS: Implicit gradients differentiation.	112
6.3.2	Stochastic Approximations in iDARTS	113
6.3.3	Stochastic Approximation of Hypergradient	115
6.3.4	Differentiable Architecture Search with Stochastic Implicit Gradients	116
6.4	Experimental Result	117
6.4.1	Reproducible Comparison on NAS-Bench-1Shot1	118
6.4.2	Reproducible Comparison on NAS-Bench-201	119
6.4.3	Experiments on DARTS Search Space	121
6.4.4	Ablation study on the number of approximation terms	122
6.5	Chapter Summary and Discussion	124
7	Conclusions and Future Work	127
7.1	Summary of Thesis	127
7.2	Limitations of Thesis and Future Work	131
A	Appendix	133
A.1	Proof of Lemma 1	133
A.2	Proof of Lemma 2	134
A.3	Proof of Lemma 3	135
A.4	Proof of Lemma 5	136
A.5	Proof of Corollary 1	137
A.6	Proof of Theorem 1	137
A.7	Proof of Corollary 2	138
A.8	Proof of Lemma 7	139
A.9	Proof of Theorem 2	140
	Bibliography	143

LIST OF FIGURES

FIGURE	Page
1.1 Illustrations of neural network structures in the early stage [57, 78].	2
1.2 Typical structures of modern deep neural networks [77, 128, 132].	3
1.3 Description of DARTS convolutional (middle) and recurrent (right) search space.	10
1.4 Example architectures in NAS-Bench-101 search space	11
1.5 Search Space in NAS-Bench-201.	12
1.6 Framework of the thesis.	17
2.1 Best cell structures found by EN ² AS.	27
2.2 Validation accuracy of sampled architecture and fixed architectures during the supernet training for GDAS (dash lines) and EN ² AS (solid lines).	31
2.3 The τ metric and mean test accuracy for architectures obtained through different architecture sampling methods.	32
3.1 Left: The general process of one-shot NAS. First, the search space is defined as a supernet containing all candidate architectures. Then a single path of the supernet (an architecture) is trained in each step of the supernet training process. Promising architectures are selected based on the validation accuracy of weights inherited from the trained supernet without the need for training from scratch. Right: The validation accuracy for four different architectures during the supernet training. The solid lines ("Arch") are the accuracies returned using weights inherited from the supernet; the dashed lines ("Arch-R") are the accuracies after retraining. . . .	36
3.2 NSAS loss function ensures that the learning of current architecture will not deteriorate the performance of previous architectures in the constraint subset. . .	41
3.3 The best found cells with NSAS and NSAS-C on CIFAR-10.	48

LIST OF FIGURES

3.4	The validation accuracy during supernet training for four different architectures with RandomNAS-NSAS and GDAS-NSAS. The solid lines (“Arch”) indicate the validation accuracy with weights inherited from the supernet, and the dashed lines (“Arch-R”) represent the validation accuracy after retraining.	49
3.5	The Kendall Tau metric (τ) of architecture ranking based on weight sharing and retraining.	49
3.6	(a) The architecture ranking differences between retraining and inheriting weights from a trained supernet with RandomNAS, RandomNAS-NSAS, GDAS, and GDAS-NSAS (from left to right, respectively). (b) The mean retraining validation accuracy for the architectures found through different methods.	49
4.1	The framework of the proposed E ² NAS.	65
4.2	Example of obtaining α_i^c through our architecture complementation.	71
4.3	Sigmoid-type function for the hyperparameter γ with the training epochs based on Eq.(4.16).	76
4.4	Analysis of architecture complementation on NAS-BENCH-201 dataset and DARTS search space [40, 95].	78
4.5	Best found cells with E ² NAS and E ² NAS-R on CIFAR-10.	81
5.1	Comparison of node connection in original DARTS and <i>depth-aware</i> DARTS. . .	97
5.2	Validation and test error of BaLeNAS and DARTS on the search space 3 of NAS-Bench-1Shot1.	100
5.3	The best normal cells discovered by BaLeNAS with and without depth regularization.	101
5.4	(a) The depth of the searched cells during the architecture search with and without depth-aware regularization. (b) The validation performance of searched architectures by BaLeNAS and BaLeNAS w/o on ImageNet.	102
5.5	The ratio of skip-connection the searched normal cells during the architecture search in the DARTS space.	103
5.6	Trajectory of the Hessian norm in DARTS space.	104
6.1	Validation and test errors of iDARTS with different T and DARTS on the search space 3 of NAS-Bench-1Shot1.	118
6.2	Hyperparameter analysis of iDARTS on the NAS-Bench-201 benchmark dataset.	120
6.3	The best cells discovered by iDARTS on the DARTS search space.	122
6.4	Ablation study on K for iDARTS with $T = 1$ and $T = 5$ on NAS-Bench-1Shot1.	123

LIST OF TABLES

TABLE	Page
1.1 Summarize of common search spaces in NAS	10
2.1 Comparison results with state-of-the-art weight sharing NAS methods on CIFAR-10, CIFAR-100 and ImageNet.	28
2.2 Comparison results with state-of-the-art NAS approaches on PTB and WT2.	28
2.3 Comparison with two baselines on NAS-Bench-201 dataset.	32
3.1 Results with the existing NAS approaches on CIFAR-10 and CIFAR-100.	45
3.2 Results with manual-designed architectures and NAS approaches on the ImageNet dataset.	46
3.3 Results of one-shot NAS baselines on NAS-Bench-201.	51
3.4 Analysis of one-shot NAS with various settings for β and M on the NAS-Bench-201 dataset.	51
3.5 Analysis of the one-shot NAS with constraint selection strategies on CIFAR-10.	53
3.6 Analysis of the one-shot NAS with constraint selection strategies on CIFAR-100.	53
3.7 Analysis of the one-shot NAS with constraint selection strategies on ImageNet-16-120.	53
4.1 Comparison results with state-of-the-art NAS approaches on NAS-Bench-201.	74
4.2 Analysis of E^2 NAS with different γ on NAS-Bench-201.	76
4.3 Analysis of one-shot NAS with different ϵ settings on CIFAR-10. We set a fixed $\gamma = \text{Sig}_\gamma(10)$ for our E^2 NAS in this experiment.	79
4.4 The CIFAR-10 test accuracy for our E^2 NAS with different ϵ and γ settings.	80
4.5 The searching time of differentiable NAS baselines.	80
4.6 Comparison results with state-of-the-art weight-sharing NAS approaches.	82
5.1 Comparison results with state-of-the-art weight-sharing NAS approaches.	99

LIST OF TABLES

5.2	Comparison results with different MCMC number for BaLeNAS on NAS-Bench-201.	103
6.1	Comparison results with NAS baselines on NAS-Bench-201.	117
6.2	Comparison results with state-of-the-art weight-sharing NAS approaches.	121
6.3	Ablation study on K for iDARTS with on NAS-Bench-201.	124