

Strategies for Practical Deep Learning

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Lukas Tuggener

under the supervision of
Jürgen Schmidhuber and Thilo Stadelmann

11 2024

Dissertation Committee

Luca Maria Gambardella Università della Svizzera Italiana, Switzerland
Rolf Krause Università della Svizzera Italiana, Switzerland
Benjamin F. Grewe Eidgenössische Technische Hochschule Zürich, Switzerland
Martin Jaggi École Polytechnique Fédérale de Lausanne, Switzerland

Dissertation accepted on 29 11 2024

Signiert von:

E68CA477620F40F...

Research Advisor

Jürgen Schmidhuber

Signiert von:

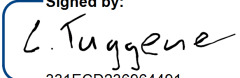
414F744C42464AF...

Co-Advisor

Thilo Stadelmann

PhD Program Director
The PhD program Director Walter Binder

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Signed by:

331EGD236964491...

Lukas Tuggener
Lugano, 29 11 2024

Abstract

In recent years, deep neural networks have achieved breakthrough results in diverse domains, from computer vision and natural language processing to game playing and life sciences. However, harnessing the full power of this technology in practical applications remains challenging. In this thesis, we explore strategies to address the challenges of applying deep learning to real-world pattern recognition problems. We tackle multiple practical problems, such as Optical Music Recognition (OMR), automated machine learning (AutoML), or the design of robust neural network architectures. In the context of OMR, we introduce two datasets, DeepScores and DeepScoresV2, the largest and most complete OMR datasets to date. Based on this data, we develop the first object detection method capable of handling the challenges of written music and methods to harden neural networks against the effects of degraded real-world data more than doubling detection performance on messy, degraded data. We then investigate the current state of AutoML, introduce a novel method for AutoML and extract design patterns for resource-constrained AutoML settings. In the latter parts of this thesis, we focus on the underlying issues that often cause neural networks to generalize poorly to real-world data. We first investigate the dataset dependency of modern CNN architectures. We show through an extensive empirical study that ImageNet alone is not sufficient to judge the power of CNN architectures and propose strategies for developing more universal evaluation methods. Finally, we tackle the lack of rotation invariance in modern vision systems and introduce a novel bio-inspired paradigm that significantly enhances the rotational robustness and outperforms the current state of the art by 19%.

Acknowledgements

I want to express my deepest gratitude to my supervisors Professors Jürgen Schmidhuber and Thilo Stadelmann, for guiding me throughout my academic journey and giving me the freedom to find my path while challenging me to become the best researcher I can be. Under their leadership, I was able to grow tremendously not only as a researcher but also as a person.

To my colleagues at CAI and InIT, I am very thankful for all the collaborations, discussions and the positive and stimulating environment they have created. Special thanks to Mohammadreza Amirian and Ismail Elezi for treading the treacherous waters of early PhD life with me.

Special thanks to Larissa for her love, patience, and support. She has been a tremendous source of strength for me. I thank her for standing by my side in all my hardest moments and celebrating the good ones with me. I am eternally grateful to have her by my side.

To my parents and sisters, I am thankful for their unconditional support. They provided me with the stable foundation upon which I was able to build and to this day give me a deep sense of calm and safety.

My dear friends Dominic, Pascal and Thomas I thank for their companionship, for all the laughs and heartfelt discussions we had throughout the years. They helped me get my mind off things even in the most stressful times.

Lastly, I want to thank all the healthcare workers who took care of me during the time when I could not.

Contents

Contents	vii
1 Introduction	1
1.1 Motivation	2
1.2 Research goal	3
1.3 Organization	4
1.4 List of contributions	5
1.4.1 First authorship	5
1.4.2 Co-authorship	6
2 Background	7
2.1 Strategies for practical NN design	7
2.2 Optical music recognition	8
3 Synthetic data as a strategy to overcome data scarcity: DeepScores	13
3.1 Introduction	13
3.2 <i>DeepScores</i> in the context of other datasets	14
3.2.1 Comparisons with computer vision datasets	17
3.2.2 Comparisons with OMR datasets	18
3.3 The <i>DeepScores</i> dataset	20
3.3.1 Quantitative properties	20
3.3.2 Flavors of ground truth	21
3.3.3 Dataset construction	23
3.4 Anticipated use and impact	24
3.4.1 Unique challenges	24
3.4.2 Towards next-generation computer vision	25
3.5 Discussion	26

4 Task-informed neural architecture design: The deep watershed detector	29
4.1 Introduction and problem statement	29
4.2 Related work	31
4.3 Deep watershed detection	32
4.3.1 Retrieving object centers	34
4.3.2 Object class and bounding box	34
4.3.3 Network architecture and losses	35
4.4 Experiments and results	36
4.4.1 Used datasets	36
4.4.2 Network training and experimental setup	37
4.4.3 Results	40
4.5 Discussion	42
5 Extending synthetic data with information and augmentation: DeepScoresV2	45
5.1 Introduction	46
5.2 The DeepScoresV2 dataset	47
5.2.1 Oriented bounding boxes	48
5.2.2 Extended symbol set	50
5.2.3 Additional features	50
5.3 Baseline results on DeepScoresV2	53
5.3.1 Reference experimental setup	53
5.3.2 Deep watershed detector	54
5.3.3 Faster R-CNN	54
5.3.4 Evaluation and discussion	54
5.4 The <i>RealScores</i> data	55
5.5 ScoreAug	57
5.6 Conclusion and future work	60
6 Automated machine learning under constrained resources	63
6.1 Introduction	63
6.2 Impact of practical machine learning	64
6.3 Survey of the current state of automated machine learning	65
6.4 Portfolio Hyperband	69
6.4.1 Intermediate conclusions	72
6.5 AutoDL for Vision and the sponge effect	72
6.5.1 Base architecture selection	73
6.5.2 Classifier design	74

6.5.3	The sponge effect in multiple modalities	76
6.5.4	Intermediate conclusions	77
7	Robust neural network design: Is it enough to optimize CNN architectures on ImageNet?	79
7.1	Introduction	80
7.2	Related work	82
7.3	Datasets	84
7.4	Experiments and results	85
7.4.1	Experimental setup	85
7.4.2	Experimental results	88
7.4.3	Impact of the number of classes	89
7.4.4	Identifying drivers of difference between datasets	91
7.5	Discussion and intermediate conclusions	94
8	Efficient rotation invariance in computer vision tasks: Artificial mental rotation	99
8.1	Introduction	100
8.2	Related work	101
8.3	Artificial mental rotation module	102
8.3.1	AMR module for CNNs	104
8.3.2	AMR module for ViTs	104
8.3.3	Motivation for add-on design	105
8.4	Experiments	105
8.5	Validity of self-supervised training built on artificial rotation	110
8.6	Application to a novel downstream task: semantic segmentation	113
8.7	Limitations and future work	113
8.8	Intermediate conclusions	114
9	Conclusions	115
9.1	Summary	115
9.2	Future directions	116
A	Extended Result	119
A.1	Is it enough to optimize CNN architectures on Imagenet?	119
A.1.1	Verifying the numerical robustness of our study	119
A.1.2	Additional ablation studies	125
A.2	Mental Rotation	129
A.2.1	Investigating on Stanford Cars and Oxford Pet	129
A.2.2	Are the base network features useful for rotation estimation?	129

A.2.3 Re-digitized Images	132
B Artifacts	135
B.1 DeepScoresV2 and RealScores	135
B.2 Is it enough to optimize CNN architectures on Imagenet?	135
B.3 Mental Rotation	135
Bibliography	139
List of Figures	161
List of Tables	169