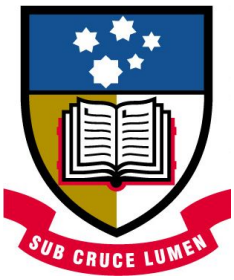
Deep Learning for Fine-Grained Visual Recognition



Teng Li

School of Computer Science

The University of Adelaide

A thesis submitted for the degree of

Master of Philosophy

28/03/2016

Contents

Abstract

Fine-grained object recognition is an important task in computer vision. The cross-convolutional-layer pooling method is one of the significant milestones in the development of this field in recent years. Based on the method, we conducted a number of experiments on a new fine-grained car dataset - CompCars. The corresponding experiments illustrate its applicability and effectiveness on this newly-designed dataset. Meanwhile, based on the experiments, we found out that pooling the most distinguishable regions like car logos and headlights areas in the indicator maps, which usually have higher activations, with the local features in the same regions can achieve better results than those by pooling the whole indicator maps with the corresponding local features. Therefore, we conjecture that better performance may be achieved if we have more powerful indicator maps or pooling channels that can better highlight these distinguishable regions.

Based on the above hypothesis and inspired by the cross-convolutional-layer pooling, next we propose the Spatially Weighted Pooling (SWP) method, which is a simple yet effective pooling strategy to improve fine-grained classification performance. SWP learns a dozen of pooling channels or spatial encoding masks that aggregate local convolutional feature maps with learned spatial importance information and produce more discriminative features. It can be seamlessly integrated into existing convolutional neural network (CNN) architectures such as the deep residual network. It also allows end-to-end training. SWP has few parameters to learn, usually in several hundreds, therefore does not introduce much computational overhead.

SWP has shown significant capability to improve fine-grained visual recognition performance by simply adding it before fully-connected layers in off-the-shelf deep convolutional networks. We have conducted comprehensive experiments on a number of widely-used fine-grained datasets with a variety of deep CNN architectures such as Alex networks (AlexNet), VGG networks (VGGNet) and the deep residual networks (ResNet). By integrating SWP into ResNet (ResNet-SWP), we achieve state-of-the-art results on three fine-grained datasets and the MIT67 indoor scene recognition dataset. With ResNet152-SWP models, we obtain 85.2% on the bird dataset CUB-200-2011 without bounding-box annotations and 87.4% with bounding-box, 91.2% on FGVC-aircraft, 94.1% on Stanford-cars with bounding-box information and 82.5% on the MIT67 dataset.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature: **Date:** 28/03/2017

Acknowledgements

First I would like to thank Prof. Chunhua Shen, my principal supervisor during my study of the master degree. His dedication to research inspired me a lot, and will definitely influence me throughout my research career in the coming years. Under his supervision, I have learnt not only specific knowledge in the field, but also how to think as a mature researcher, including the ability to find unsolved challenges and address them in novel ways.

I would also like to thank staff and visitors in the Australian Center for Visual Technologies (ACVT). As a non-native English speaker, I made grammar errors in academic manuscripts now and then. I would like to thank my co-supervisor, Dr. Guosheng Lin, for his time and effort in helping my research. In addition, I would like to thank Dr. Lingqiao Liu, a research fellow in our research center. He gave me endless support and advice to help me develop my proposed pooling method which is also a cornerstone of this thesis. Other staffs in the ACVT have given me thoughtful suggestions on research. In a word, I am very lucky to be a member of the ACVT.

As a master by research student, I spent most time with PhD students who all have helped me a lot on both research and life. Having friends with different culture backgrounds really enriches my life experience.

Finally, thanks go to the excellent support from staffs in the School of Computer Science at the University of Adelaide. Julie Mayo and Sharyn Liersch are administrative staffs of our school. Thanks also go to our Head of School, Prof. Katrina Falkner, and our Postgraduate Coordinator Prof. Ian Reid I am really grateful for what they have done.

List of Figures