Large-scale Video Analysis and Understanding

Zhongwen Xu

August 2017

Centre for Artificial Intelligence
Faculty of Engineering and Information Technology
University of Technology Sydney

Advisor: Prof Yi Yang

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

©Zhongwen Xu, 2017

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by an Australian Government Research Training Program Scholarship.

Production Note:

Signature of Student: Signature removed prior to publication.

Date: Aug 16, 2017

Abstract

Video understanding is a complex task in computer vision, which requires not only recognizing objects, persons, and scenes, but also capturing and remembering the changes of visual content along time. Rapid development in building blocks like image classification task in recent years provides great opportunities for accurate and efficient video understanding. Based on deep convolutional neural networks and recurrent neural networks, various kinds of deep learning applications on video understanding have been studied. In this thesis, I present my research on large-scale video analysis and understanding in three major aspects: video representation learning, recognition with limited examples, and vision & language. Representation and features are the most important part for vision tasks, since it is very general and can be used for classification task, detection task and also tasks for structural prediction like vision and language. We begin with video classification from multimodal features, which are hand-crafted features from different streams, *i.e*. vision and audio. For representation learning, we investigate aggregation methods to generate video representation from frame features. Significant improvements over classical pooling methods have been demonstrated. In addition, we propose a hierarchical recurrent neural network to learn the hierarchical structure for video. Going beyond supervised learning, we develop a sequence model to learn from reconstruction of future and past features based on the current sequences, showing that unlabeled videos can help learning good and generalizable video representation. We explore the problem of recognition with limited examples, which tries to tackle the situation that we cannot obtain enough data to train the model. The encouraging results show that it is feasible to obtain good performance with only a few examples for the target class. Except for the video classification task which only outputs labels for the video, we also seek for richer interaction between machine and human on vision content via natural language. We consider two major forms of vision and language tasks, the first is video captioning, *i.e.*, to automatically generate caption to describe the given video sequence, and video question answering, *i.e.*, to answer questions related to the presented video sequence. Finally, I conclude the thesis with some future directions on video understanding.

Dedicated to my parents, Guorong and Shuying

Contents

Acknowledgments

I am very grateful to have worked with many wonderful people in my PhD study, who have provided so many insightful discussions and collaborations on my research and various kinds of great help in my study and my life.

First and foremost I would like to thank my advisor Prof. Yi Yang. None of this would have been possible without him. Thanks a lot for giving me the chance to visit Carnegie Mellon University when I was a junior student in Zhejiang University, which has completely changed my life and made the initial direction leading to this thesis. It is hard to believe that it has been more than 5 years since then. Thanks a lot to making me determined to pursue a PhD degree when I knew nothing. It has been a long journey following Yi from CMU to UQ, then to UTS but it has always been pleasant experience to be under his supervision. Yi gave me enough freedom and strong encouragement to pursue my dream about research and about my life, which is the great fortune in my PhD study. Yi also generously allowed me to divide my time in different research groups to experience different styles of research and have the great chances to learn from different great people, which are the most important experiences in my PhD. Yi's insistence on practical and useful research has shaped important part of my view on research. Moreover, Yi stands as a model to teach me about how to treat family, friends and students, and how to manage a group to work together. Yi is always there when I need some help. I owe Yi more than I can express here.

I would like to thank Alexander Hauptmann, my host professor at Carnegie Mellon University. Alex's focus on video analysis for decades is always inspiring to my own research. Alex's encouragement to think bigger and to do innovative research has been influencing me. It has always been pleasant to talk with Alex.

I would like to thank Ivor Tsang, who provided ideas and great help to my very first two papers. Ivor was always patient to help me derive equations and waited until late night to polish my papers. Without Ivor's help, I could not be able to learn to conduct independent research in my PhD.

I would like to thank Shuicheng Yan, my host professor at National University of Singapore, from whose group I learnt the basics of deep learning and ran codes on GPUs first time in my life. The knowledge I learnt from Shuicheng and his group is invaluable to the rest of my PhD study. Shuicheng's diligence to work encourages me to move forward.

I would also like to thank Fei Wu and Yueting Zhuang, my honor class advisors in Zhejiang University, who gave me the first research papers to read when I was a sophomore. I could not imagine my current life without their help many years ago.

I am deeply indebted to Pierre Sermanet and George Toderici, my intern hosts at Google Brain. Pierre's confidence on research and his great encouragement when I encountered difficulties in my project is memorable. Thanks to Pierre for introducing me to the area of deep reinforcement learning and offering me various kinds of help along my internship.

George always gave me the most useful pointers towards Google's infrastructure and video processing related stuff. I would also like to thank Vincent Vanhoucke for giving me the privileged opportunity to intern at Brain and thank George's great help for applying for an additional intern slot for me.

I am also deeply indebted to Hado van Hasselt and David Silver, my intern hosts at DeepMind. I am so grateful that Hado moved to the next desk to me for offering immediate help at any time. The daily conversations about the projects and about reinforcement learning helped me learn very quickly. Hado was always very patient and helpful, no matter what naive questions I asked. I would never forget Dave's great vision about research and his encouragement to take grand challenges. I still remember in the first week of my internship and in the discussions about my potential internship projects, Dave said, "AlphaGo was from an internship project, never be afraid to work on hard problems.". It was a great experience from watching Dave's lectures on RL to working in Dave's RL team as an intern.

This research is supported by an Australian Government Research Training Program Scholarship. Thanks to Australian Government for this scholarship support. I am also thankful to all the funding agencies which supported my research.

Thanks to all my friends and collaborators in Yi's group, who have made my PhD life so pleasant: Yan Yan, Xiaojun Chang, Xingzhong Du, Linchao Zhu, Pingbo Pan, Ke Ning, Hehe Fan, Zhedong Zheng, Yanbin Liu, Wenhe Liu, Yutian Lin, Guoliang Kang, Xuanyi Dong, Zheng Liang, and Zongting Lv. Special thanks to Yan Yan for all the accompanies in the four years, and to Linchao Zhu for his amazing implementation ability in all of our collaborated projects and the wonderful discussions about frontier research. Many thanks to other friends at UTS, especially Bo Han, Donna Xu, Yali Du, and Yuangang Pan. Thanks to the friends and collaborators on the ALADDIN project at CMU: Shoou-I Yu, Zhigang Ma, Yang Cai, Huan Li, Zhenzhong Lan, and Xuanchong Li. Special thanks to Shoou-I Yu and Zhigang for helping me a lot in the early years of my research. Thanks to all the members in Shuicheng's group at NUS, especially Kang Zhang, Jiashi Feng, Si Liu, Yunchao Wei, Qiang Chen, Min Lin, Junshi Huang, Yang Xu, and Xiaodan Liang. Special thanks to Qiang Chen for pointing me very helpful knowledge about deep learning in the first year of my PhD which benefits a lot to my research, and to Min Lin and Jiashi Feng for the long-term discussions about research.

Thanks to the intern fellows at Brain, especially Takeru Miyato, Laurent Dinh, and Maithra Raghu. Thanks to all the great help from TensorFlow team at Brain, which solved countless engineering problems for me. Thanks to all the other team members at Reinforcement Learning team in DeepMind, including Tom Schaul, Arthur Guez, Matteo Hessel, Joseph Modayil, John Quan, Dan Horgan, Andre Barreto and Helene Hjelmvik, whose wide knowledge about reinforcement learning and bright ideas made me completely impressive. Thanks to all the researchers and engineers who offered help in both internships.

Finally, I would like to thank my parents Guorong Xu and Shuying Mai for their endless support and love. I am grateful that my parents always support my decisions in my life and give me care, power and encouragement. I would also like to thank my lovely younger sister Yuting Xu for her belief in me. I am very fortunate to have such a great family. The weekly video chat with my family is always the most pleasant moment in my PhD life.