UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Collaborative Dual-Stream Modeling for Video Understanding

by Xiaohan Wang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Yi Yang

University of Technology Sydney
Faculty of Engineering and Information Technology

February 2021

Certificate of Authorship/Originality

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

I also certify that this thesis has been written by me. Any help that I have received in my research and in the preparation of the thesis itself has been fully acknowledged. In addition, I certify that all information sources and literature used are quoted in the thesis. This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

February 2021

ABSTRACT

Collaborative Dual-stream Modeling for Video Understanding

by

Xiaohan Wang

Most existing video recognition systems classify the input video to coarse-grained labels with single-stream architectures or combine multi-modal predictions by simple late fusion. However, real-world video applications usually require understanding complex human-object interactions and fine-grained content. It expects a video analysis system to be able to conduct meticulous reasoning. Besides, the urgent need for multi-modal alignment and communication among different models requires multi-stream video modeling, which is beyond single-stream architectures' capacity.

In this thesis, I argue that we should tackle video understanding with collaborative dual-stream modeling in several challenging scenarios. The interaction between the different information in videos can encourage the video understanding system to exploit the spatio-temporal relation. The idea has been applied to three tasks. First, for egocentric action recognition, symbiotic attention mechanism and interactive prototype learning scheme are developed to explore the relationship between the motion stream and appearance stream. Second, we design a T2VLAD framework for text-video retrieval to align the text stream and video stream. Third, for efficient video recognition, the communication between the lightweight model and heavyweight model is enabled by a parallel sampling network to sample more salient frames. Extensive experiments on popular video datasets demonstrate the effectiveness of the proposed approaches.

Dissertation directed by Professor Yi Yang

School of Computer Science

Acknowledgements

I would like to thank my principal supervisor Prof. Yi Yang. He motivates me to set a high standard for my research and gives considerate advice on both research and career planning. Without his continuous and invaluable support, the Ph.D. degree can not be attainable.

I would like to thank my co-supervisor Dr. Linchao Zhu. He introduced me to the video understanding area and taught me a lot of research skills. He gave inspirational advice on my research and helped me a lot on paper writing and presentations.

I want to thank my main collaborators, Mr. Yu Wu at UTS, Dr. Heng Wang at FaceBook AI, Dr. Ping Liu at Singapore A*STAR and Mr. Haitian Zeng at Baidu Research. Thank you all for discussing with me on interesting ideas and helping me with paper writing.

Thanks to all my colleagues and friends in UTS and Baidu Research.

Finally, I would like to thank my parents Mr. Xun Wang and Ms. Xuechun Wang for their unconditional love over so many years. I want to thank my girlfriend Ms. Yang Su for her love, support, and encouragement during the past eight years.

Xiaohan Wang

Sydney, Australia, 2021.

List of Publications

Journal Papers

J-1. **X. Wang**, L. Zhu, Y. Wu and Y. Yang, "Symbiotic Attention for Egocentric Action Recognition with Object-centric Alignment," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2020.3015894. (**Published**)

J-2. Y. Wu, L. Zhu, **X. Wang**, Y. Yang, F. Wu, "Learning to Anticipate Egocentric Actions by Imagination," in *IEEE Transactions on Image Processing*. (**Accepted**)

Conference Papers

C-1. **X. Wang**, Y. Wu, L. Zhu, Y.Yang, "Symbiotic Attention with Privileged Information for Egocentric Action Recognition," in *AAAI*, pp. 12249-12256, 2020. (**Published**)

C-2. **X. Wang**, Y. Wu, L. Zhu, Y.Yang, "Baidu-UTS Submission to the EPIC-Kitchens Action Recognition Challenge 2019," in *CVPR Workshop*, 2019. (**Published**)

C-3. **X. Wang**, L. Zhu, H. Wang, Y.Yang, "What Are You Cutting? Recognizing Active Objects in Egocentric Videos via Interactive Prototype Learning." (**Under Review**)

C-4. **X. Wang**, L. Zhu, Y.Yang, "T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval," (**Under Review**)

C-5. **X. Wang**, L. Zhu, P. Liu, Y.Yang, "Parallel Sampling Network: A Differentiable Framework with Context Relation Mining for Efficient Video Recognition." (**Under Review**)

C-6. H. Zeng, Y. Dai, X. Yu, **X. Wang**, Y.Yang, "Learning Deep Clustered Representation in Shape Space for for Non-rigid Structure-from-Motion." (**Under Review**)

Contents

List of Figures