UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

# Video Representation Learning with Deep Neural Networks

by

**Linchao Zhu**

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

**Doctor of Philosophy**

Sydney, Australia

2019

# 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.

Linchao Zhu

Production Note:
Signature removed prior to publication.

ABSTRACT

**Video Representation Learning with**

**Deep Neural Networks**

by

Linchao Zhu

Despite the recent success of neural networks in image feature learning, a major problem in the video domain is the lack of sufficient labeled data for learning to model temporal information. One method to learn a video representation from untrimmed videos is to perform unsupervised temporal modeling. Given a clip sampled from a video, its past and future neighboring clips are used as temporal context, and reconstruct the two temporal transitions, i.e., present→past transition and present→future transition, which reflect the temporal information in different views. In this thesis, the two transitions are exploited simultaneously by incorporating a bi-direction reconstruction which consists of a backward reconstruction and a forward reconstruction. To adapt an existing model to recognize a new category which was unseen during training, it may be necessary to manually collect hundreds of new training samples. Such a procedure is rather tedious and labor intensive, especially when there are many new categories. In this thesis, a classification model is proposed to learn from a few examples in a life-long manner. To evaluate the effectiveness of the learned representation, extensive experiments are conducted on multimedia event detection, image classification, video captioning, and video question answering.

Dissertation directed by Professor Yi Yang

Centre for Artificial Intelligence, School of Software

# Acknowledgements

First and foremost, I would like to thank my supervisor Professor Yi Yang. I am extremely grateful for his patience and support. He guided me on any research directions I was exited about. He also provided tremendous help on building up my research career, and offered great kindness to my personal life. He also teached me how to work with colleagues, which is valuable to my future career. Thanks to Prof Alexander G. Hauptmann, my advisor when I visited Carnegie Mellon University, from whom I learned how to do research for real-world applications. Thanks to Heng, Du, and Laura, my supervisors when I interned at Facebook Research, from whom I learned critical thinking and how to perform research in a systematic way.

I would also like to thank my colleagues at University of Technology Sydney. I would like to thank Xiaojun Chang, Xuanyi Dong, Hehe Fan, Qianyu Feng, Qingji Guan, Yang He, Wenhe Liu, Yanbin Liu, Ping Liu, Yutian Lin, Peike Li, Fan Ma, Jiaxu Miao, Pingbo Pan, Yu Wu, Xiaohan Wang, Zhongwen Xu, Yan Yan, Zongxin Yang, Fengda Zhu, Hu Zhang, Zhong Zhun, Zhedong Zheng, Liang Zheng, Xiaolin Zhang, and many others. I was really fortunate to work with them and participate in intellectuall conversations with them.

I would also like to thank Data to Decision CRC for supporting my research.

Lastly I would like to thanks my mother Supin Chen and Qiming Zhu for their support and love throughout the years.

Linchao Zhu

Sydney, Australia, 2019.

# List of Publications

## Journal Papers

J-1. **Zhu, L.**, Xu, Z., Yang, Y. and Hauptmann, A.G., 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3), pp.409-421.

J-2. Gan, C., Yang, Y., **Zhu, L.**, Zhao, D. and Zhuang, Y., 2016. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1), pp.61-77.

## Conference Papers

C-1. **Zhu, L.**, Xu, Z. and Yang, Y., 2017, July. Bidirectional Multirate Reconstruction for Temporal Modeling in Videos. In *Computer Vision and Pattern Recognition* (CVPR), 2017 IEEE Conference on (pp. 1339-1348). IEEE. **Spotlight**.

C-2. **Zhu, L.** and Yang, Y., 2018, September. Compound Memory Networks for Few-Shot Video Classification. In *European Conference on Computer Vision* (pp. 782-797). Springer, Cham.

C-3. **Zhu, L.***, Xu, Z.*, and Yang, Y., 2017, July. Few-Shot Object Recognition from Machine-Labeled Web Images. In *Computer Vision and Pattern Recognition* (CVPR), 2017 IEEE Conference on (pp. 5358-5366). IEEE. **Spotlight**. (* indicates equal contribution)

C-4. Fan, H., **Zhu, L.** and Yang, Y., 2019. Cubic LSTMs for Video Prediction. In *The Thirty-Third AAAI Conference on Artificial Intelligence* (AAAI).

C-5. Fan, H., Xu, Z., **Zhu, L.**, Yan, C., Ge, J. and Yang, Y., 2018. Watching a Small Portion could be as Good as Watching All: Towards Efficient Video

Classification. In *International Joint Conference on Artificial Intelligence* (IJCAI) (Vol. 2, No. 5, p. 6).

C-6. Wu, Y., **Zhu, L.**, Jiang, L. and Yang, Y., 2018. Decoupled Novel Object Captioner. In 2018 *ACM Multimedia* Conference on Multimedia Conference (pp. 1029-1037). ACM.

C-7. Dong, X., **Zhu, L.**, Zhang, D., Yang, Y. and Wu, F., 2018, October. Fast Parameter Adaptation for Few-shot Image Captioning and Visual Question Answering. In 2018 *ACM Multimedia* Conference on Multimedia Conference (pp. 54-62). ACM.

# Contents

# List of Figures