

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 excited about. He also provided tremendous help on building up my research career, and offered great kindness to my personal life. He also taught 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 intellectual conversations with them.

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

Lastly I would like to thank 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 (IJ-CAI)* (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

Certificate	iii
Abstract	iv
Acknowledgments	v
List of Publications	vi
List of Figures	xii
1 Introduction	1
1.1 Video Feature Learning	1
1.2 Video and Language	2
1.3 Contributions	3
2 Literature Review	5
2.1 Video Classification	5
2.1.1 Convolutional Networks for Video Classification	6
2.1.2 Recurrent Networks for Video Classification	7
2.2 Bridging Vision and Language	8
2.2.1 Video Captioning	8
2.2.2 Video Question Answering	8
2.3 Few-shot Video Classification	10
2.3.1 Memory-Augmented Neural Networks	11
3 Bidirectional Multirate Reconstruction for Temporal Mod-	

eling in Videos	13
3.1 Introduction	13
3.2 Multirate Visual Recurrent Models	15
3.2.1 Multirate Gated Recurrent Unit	15
3.2.2 Unsupervised Video Sequence Reconstruction	18
3.2.3 Complex Event Detection	21
3.2.4 Video Captioning	23
3.3 Results	23
3.3.1 Complex Event Detection	23
3.3.2 Video Captioning	28
3.4 Conclusion	32
4 Uncovering the Temporal Context for Video Question Answering	34
4.1 Introduction	34
4.2 Dataset Collection and Task Definitions	37
4.2.1 Dataset and QA Pair Generation	38
4.2.2 Task Definitions and Analysis	40
4.3 The Proposed Approach	42
4.3.1 Learning to Represent Video Sequences	44
4.3.2 Dual-Channel Learning to Rank	48
4.4 Results	51
4.4.1 Evaluation of Describing the Present	51
4.4.2 Evaluation of Inferring the Past and Predicting the Future	56
4.4.3 Limitations and Future Work	57

4.5	Conclusion	58
5	Few-Shot Object Recognition from Machine-Labeled Web Images	60
5.1	Introduction	60
5.2	Proposed Approach	64
5.2.1	Preliminaries	64
5.2.2	Model Overview	65
5.2.3	Model Components	68
5.2.4	Training	72
5.2.5	Inference	72
5.3	Experiments	72
5.3.1	Preprocessing	73
5.3.2	Model Specifications	73
5.3.3	Datasets	74
5.3.4	Few-shot Learning with Human-labeled annotations	74
5.3.5	Few-shot Learning with Machine-labeled Annotations	77
5.3.6	Hyperparamter Study	79
5.4	Conclusion	80
6	Compound Memory Networks for Few-shot Video Classification	81
6.1	Introduction	81
6.2	Few-shot Video Classification Setup	84
6.3	Compound Memory Network	85
6.3.1	Multi-saliency Embedding Function	86

6.3.2	Compound Memory Structure	87
6.3.3	Training	91
6.4	Experiments	92
6.4.1	Datasets	92
6.4.2	Implementation Details	92
6.4.3	Evaluation	94
6.4.4	Ablation Study	98
6.5	Conclusion	99
7	Future Works	100
	Bibliography	101

List of Figures

3.1	Frame sampling rate should vary in accordance with different motion speed. In this example, only the last three frames have fast motion. The dashed arrow corresponds to a fixed sampling rate, while the solid arrow corresponds to multiple rates.	14
3.2	We illustrate the two modes in the mGRU. In the slow to fast mode, the state matrices \mathbf{V}_* are block upper-triangular matrices and in the fast to slow mode, they are block lower-triangular matrices.	17
3.3	Unrolled mGRU. In the example, the state is divided into three groups and the slow to fast mode is shown. At each step t , groups satisfying $(t \text{ MOD } T_i) = 0$ are activated (cells with black border). For example, at step 2, group 1 and group 2 are activated. The activated groups take the frame input and previous states to calculate the next states. For those that are inactivated, we simply pass the previous states to the next step. Group 1 is the fastest and group 3 is the slowest with larger T_i . The slow to fast mode is the mode by which the slower groups pass the states to the faster groups.	20
3.4	The model architecture of unsupervised video representation learning. In this model, two decoders are used to predict surrounding contexts by reconstructing previous frames and next frame sequences. The “<G0>” input, which is a zero vector, is used at step 0 in the decoder. During training, one of the two decoders is used with a probability of 0.5 for reconstruction.	22

4.1	Questions and answers about the past, the present and the future. Our system includes three subtasks, which infer the <i>past</i> , describe the <i>present</i> , and predict the <i>future</i> , while <i>only the current frames are observable</i> . Best viewed in color	35
4.2	t-SNE visualization of word embeddings for each category learned from word2vec model.	37
4.3	Examples of QA pairs for different categories and levels of difficulty. The words colored in green are correct answers, and the difficult candidates are marked in red.	41
4.4	Distribution of question types for each dataset	44
4.5	Distribution of question lengths for each dataset	45
4.6	Distribution of answer lengths for each dataset	46
4.7	The encoder-decoder model (top): encoder state of last time step is passed to three decoders for reconstruction. Learn to answer questions (bottom): encoder state of last time step is passed to the ranking module which selects an answer based on the visual information	47
4.8	Illustration of dual-channel learning to rank	49
4.9	The effectiveness of dual-channel learning to rank. We conduct experiments on the <i>Present-Easy</i> task to showcase. $\lambda = 0$ corresponds to using the sentence channel only and $\lambda = 1$ corresponds to using the word channel only	55
4.10	Example results obtained from our model. Each candidate has a score corresponding to a clip. Correct answers are marked in green while failed cases are in red	56

5.1	Given a large vocabulary of labels and their corresponding images, we conduct few-shot learning on a novel category which is not in the vocabulary and only has a handful of positive examples. The image examples in the vocabulary are stored in the external memory of our model, and the image example from the novel category queries the external memory. Our model returns helpful information according to visual similarity and LSTM controllers. The retrieved information, i.e., visual features and their corresponding labels, are combined to classify this query image example.	62
5.2	An illustration of our proposed model. Best viewed in color.	65
5.3	Sample images from the OpenImages dataset. Annotations on the images are shown in the bottom. The annotations listed are “label id”, “label name”, “confidence” tuples.	75
5.4	We show the query results returns from the external memory. The scores are the softmax probabilities. Only top-3 results are shown. . .	78
6.1	The setting of the few-shot video classification. There are two non-overlapping datasets in this figure, i.e., meta-training and meta-testing. The meta-training set is for meta-learning and the meta-testing set is for evaluating the generalization performance on novel categories. The network is trained in an episodic way and each episode has a support set and a query example.	82
6.2	Illustration of the input embedding model. The embedding function generates the multi-saliency descriptor \mathbf{Q} , which is flattened and normalized to a query vector.	87

6.3	Our CMN structure. A video is first mapped to a matrix representation via the multi-saliency embedding function. This hidden representation is then vectorized and normalized as a query vector, which performs a nearest neighbour search over the abstract memory. The most similar memory slot is retrieved and the label stored in the value memory will be used as the prediction. The constituent key memory contains the matrix representations of the inputs, while the abstract memory is constructed on top of the stacked constituent keys.	88
6.4	Illustration of the update rule for CMN.	90
6.5	Per class accuracy on the 5-way 1-shot setting. We show the accuracies of 24 classes on the meta-testing set.	95
6.6	We illustrate the inference procedure. There are 5 classes and the memory has 16 slots. Two different update rules will be used depending on the query results.	97