

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 Non-rigid Structure-from-Motion.” (**Under Review**)

Contents

Certificate	ii
Abstract	iii
Acknowledgments	iv
List of Publications	v
List of Figures	xi
1 Introduction	1
1.1 Egocentric Action Recognition	2
1.2 Text-video Retrieval	4
1.3 Efficient video recognition	4
2 Literature Survey	6
2.1 Deep Video Recognition	6
2.2 Egocentric Action Recognition	7
2.3 Human-Object Interaction	8
2.4 Visual Attention	9
2.5 Efficient Computing	9
2.6 Text-Video Retrieval	10
2.7 VLAD Encoding	11
3 Symbiotic Attention for Egocentric Action Recognition	12

3.1	Introduction	12
3.2	Method	17
3.2.1	Overview	17
3.2.2	Preliminaries	19
3.2.3	Object-centric Feature Alignment	20
3.2.4	Symbiotic Attention	21
3.2.5	Training and Objectives	25
3.3	Experiments	25
3.3.1	Datasets	25
3.3.2	Experiment Settings	26
3.3.3	The Effectiveness of SAOA	28
3.3.4	Comparison with State-of-the-art Results	34
3.3.5	EPIC-Kitchens Action Recognition Challenge 2020	38
3.3.6	Visualization	39
3.4	Summary	40
4	Recognizing Active Objects in Egocentric Videos via In-	
	teractive Prototype Learning	42
4.1	Introduction	42
4.2	Interactive Prototype Learning	44
4.2.1	Overview	44
4.2.2	Verb Classification	46
4.2.3	Noun Classification	47
4.2.4	Training and Inference	50
4.3	Experiment	51

4.3.1	Dataset	51
4.3.2	Implementation Details	52
4.3.3	Comparison with State of the Arts	53
4.3.4	Ablation Studies	57
4.3.5	Qualitative Results	58
4.4	Summary	60
5	Parallel Sampling Network: A Differentiable Framework with Context Relation Mining for Efficient Video Recognition	61
5.1	Introduction	61
5.2	The Proposed Approach	64
5.2.1	Problem Setting	64
5.2.2	Parallel Video Sampling Network	65
5.2.3	Training Objectives	69
5.2.4	Inference Strategy	71
5.3	Experiments	71
5.3.1	Experimental Setup	71
5.3.2	Ablation Studies	74
5.3.3	Comparison with the State-of-the-Art	78
5.3.4	Analysis on the sampling scores.	80
5.3.5	Qualitative Results	82
5.4	Summary	83
6	T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval	84

6.1	Introduction	84
6.2	Method	87
6.2.1	Overview	87
6.2.2	Video Representations	88
6.2.3	Text Representation	90
6.2.4	Local Alignment	90
6.2.5	Global Alignment	92
6.3	Experiments	93
6.3.1	Experimental Details	93
6.3.2	Comparison to State-of-the-art	94
6.3.3	Ablation Study	96
6.3.4	Qualitative Results	99
6.4	Summary	101
7	Conclusion and Future Directions	103
	Bibliography	105

List of Figures

3.1	The illustration of the object-centric feature alignment. For verb classification, the spatial location provided by the detector can possible reduce the object-irrelevant motions. Local motion features aligned with object features serve as possibly action candidates. For noun classification, global alignment inject the local object features into the context-aware global feature. These location-aware feature candidates from the two branches are beneficial to the subsequent meticulous reasoning.	13
3.2	The proposed SAOA framework. Our framework consists of three feature extractors and one interaction module. The detection model generates a set of local object features and location proposals. This location-aware information is injected to the two branches by an object-centric alignment method For the Verb branch, the feature map is locally aligned with the objects by combining the local motion features with corresponding object detection features. For the Noun branch, the object features are aligned with the global noun representation. Subsequently, the fused features from each branch interact with the global feature from the other branch by a symbiotic attention mechanism. The two object-centric feature matrices are first normalized by a cross-stream gating operation. After that, the matrices are attended by the other branch to select the most action-relevant information. The outputs of SAOA are used to classify the verb and noun, respectively.	18

3.3	The illustration of symbiotic attention on the noun branch. The object-centric noun feature matrix is first normalized by the global verb feature. After that, the feature matrix interacts with the global verb feature to generate attention weights. The final noun representation is the weighted sum of the normalized object-centric features.	23
3.4	Qualitative results of our SAOA I3D (Flow) model. The colored boxes show the top-5 detected regions and the numbers are the corresponding attention weights generated by our action-attended relation module. Red indicates the failure case.	39
4.1	The motivation of Interactive Prototype Learning (IPL) framework is to collaboratively learn judicious location-aware spatio-temporal features for more accurate <i>noun</i> (active object) classification.	43
4.2	Our Interactive Prototype Learning (IPL) framework. The feature map of size $T \times H \times W \times C$ is extracted from the last convolutional layer of the 3D CNN backbone. To facilitate the interaction between the verb branch and the noun branch, we introduce a set of <i>verb prototypes</i> shared across the two branches. A background prototype is introduced to filter the action-irrelevant information from the spatio-temporal feature map. Each prototype is a C -dimensional vector and is random initialized during training. Verb prediction is obtained by computing the cosine similarity between the average pooled verb feature and the verb prototypes. For noun prediction, the feature map is decomposed and grouped by soft-assigning each feature to the prototypes. We select the most relevant K groups based on verb predictions to generate the final noun representation. The 3D CNN backbone and IPL are jointly trained in an end-to-end manner.	45

4.3	Qualitative results of our IPL model. We illustrate the sum of assignments on the top-K verb prototypes for each feature vector on the spatio-temporal feature map. For each input clip, we uniformly sample four frames and plot the corresponding assignment map. Higher assignment values shows in red. We also print the predicted label and the ground-truth label above the images (Green for correct predictions and Red for failure cases).	59
5.1	An overview of our proposed parallel sampling network. Given an input video, we pre-sample N candidate frames. The sampler CNN processes the N frames in a parallel manner. The features are fed into a Context Relation Mining (CRM) module to produce N importance scores. We illustrate three instantiations of the CRM module: Non-local Block, Encoder-Decoder TCN, and Vanilla TCN. After that, we utilize the Top-K sampling strategy to select the highest M scores and their corresponding frame indices. The classification model only takes the sampled M frames as input and produces M prediction vectors. Finally, the prediction vectors are multiplied by the selected M weights and averaged as the final prediction.	64
5.2	Context Relation Mining Module Instantiations.	68
5.3	Mean Average Precision vs. Computational Cost on AcitivityNet. Comparison with state-of-the-art methods.	79
5.4	The left is the histogram of relative sampling location. The right is the histogram of scores produced by the sampler.	81

5.5	Visualization of the Uniformly Sampled 10 frames and the 10 frames sampled by our PSN. The ground truth actions for the videos are “Swimming”, “Playing beach volleyball” and “Futsal”, respectively. The red box indicates that the frame is irrelevant to the action class empirically.	82
6.1	Global and local alignment between texts and videos.	85
6.2	Our T2VLAD framework for text-video retrieval. “TA” is a temporal aggregation method. For simplicity, we use a max-pooling operation to aggregated each expert.	88
6.3	Visualization of the assignment weights. We take Video 7060 in the MSRVT 1K-A test set as an example. We plot the text assignments to the three centers as black lines. The thickness of the line indicates the relative value on the same center. The numbers next to the line are the assignment values. We only illustrate the top text assignments for better visualization. The Top-10 frames (the padding features have been removed.) correspond to the appearance features assigned to the centers are shown at the bottom.	100
6.4	The text-video retrieval results on the MSRVT 1K-A test set. The left are the videos ranked by our T2VLAD, and the right are the results from the model only with global alignment.	101