Video Paragraph Captioning using Hierarchical Recurrent Neural Networks

Haonan Yu[1*] **Jiang Wang**[3] **Zhiheng Huang**[2*] **Yi Yang**[3] **Wei Xu**[3]
[1]Purdue University
haonan@haonanyu.com
[2]Facebook
zhiheng@fb.com
[3]Baidu Research - Institute of Deep Learning
{wangjiang03,yangyi05,wei.xu}@baidu.com

We consider the problem of video captioning, *i.e.*, generating one or multiple sentences to describe the content of a given realistic video. This is an extension to the image captioning problem in that videos have an additional temporal dimension compared to static images. Thus in addition to appearance features, it is important for a video-captioning method to employ motion features to recognize events in videos, and analyze temporal relations between consecutive sentences if multiple sentences are to be generated. Most existing approaches (*e.g.*, [1, 8]) address the problem by using the sequence-to-sequence framework [6] , which first extract deep convolutional features from video frames, encode them to a compact representation, and decode it to generate a sentence. The encoding and decoding are usually performed with the Recurrent Neural Network (RNN) or the Long Short-Term Memory (LSTM) Network. While good results were obtained, experiments were performed on datasets where only a simple sentence is required to be generated for each short video clip. It still remains largely unexplored that how to generate multiple sentences or a paragraph for a long video.

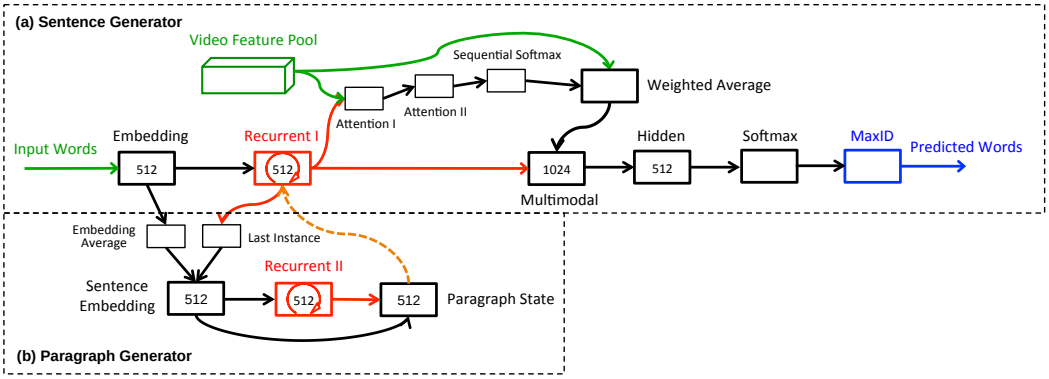
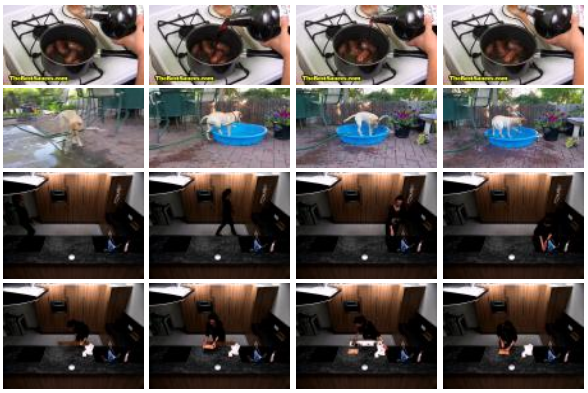


Figure 1: The proposed hierarchical RNN framework for video paragraph captioning. We extract both appearance features (*e.g.*, VggNet [5]) and motion features (*e.g.*, C3D [7] and/or Dense Trajectories [9]). These features, together with the output of the sentence recurrent layer, are input into a multimodal layer to generate the next word given the current word. The initial state of the sentence recurrent layer is set by the output of the paragraph generator. The output word is decided by the maxid layer which takes the maximal entry in the softmax layer's output.

Inspired by the recent progress of document modeling in computational linguistics [3], we propose a hierarchical-RNN framework for describing a long video with a paragraph consisting of multiple sentences. The intuitive idea behind our hierarchical framework is that we want to exploit the temporal dependency between sentences in a paragraph, so that when producing a paragraph, the sentences are not generated independently. Instead, the generation of one sentence might be af-

*This work was done while the authors were at Baidu.



A man is pouring oil into a pot.

A dog is playing in a bowl.

The person walked into a kitchen.
The person went to the fridge.
The person walked over to the sink.
The person washed the carrot.

The person opened a drawer.
The person took out a cutting board and a knife.
The person cut the cucumber on the cutting board.

Figure 2: Examples of generated sentences. The first two rows are on the YouTube2Text dataset and the last two rows are on the TACoS-Multilevel dataset.

fected by the semantics context provided by the previous sentences. For example, a sentence like *the person peeled the potatoes* is more likely to occur, than the sentence *the person turned on the stove*, after the sentence *the person took out some potatoes from the fridge*.

Towards this end, our hierarchical framework consists of a sentence generator and a paragraph generator. At the low level, the sentence generator produces a single short sentence that describes a specific time interval and video region. The semantic meaning of the generated sentence is captured by the output state sequence of the sentence generator. At the high level, the paragraph generator takes this state sequence as input, and uses the recurrent layer to output a paragraph state, which is then used to reset the initial state of the sentence generator. Figure 1 illustrates the overall framework.

We evaluate our approach on two public benchmark datasets: YouTube2Text [2] and TACoS-Multilevel [4]. YouTube2Text contains 1,970 open-domain videos and approximately 80,000 annotated sentences with a vocabulary of 12,766 words, while TACoS-Multilevel contains 185 cooking videos and approximately 40,000 annotated sentences with a vocabulary of 2,864 words. The experiment on YouTube2Text demonstrates a BLEU@4 score of 0.499 (with the previous highest 0.453). The experiment on TACoS-Multilevel demonstrates a BLEU@4 score of 0.305 (with the previous highest 0.288). These results suggest that our method is the new state of the art for video captioning. Figure 2 shows some examples of the generated sentences.

References

[1] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. 2014.

[2] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, T. D. R. Mooney, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV'13 Int. Conf. on Computer Vision 2013*, December 2013.

[3] J. Li, M. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *CoRR*, abs/1506.01057, 2015.

[4] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition (GCPR)*, September 2014.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[6] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[7] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. C3D: generic features for video analysis. *CoRR*, abs/1412.0767, 2014.

[8] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko. Sequence to sequence - video to text. *CoRR*, abs/1505.00487, 2015.

[9] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, June 2011.