Video Caption Generation

Ngoc Ta
University of New Mexico
Albuquerque, NM 87131, USA
nta@unm.edu

Aidean Sharghi
University of Central Florida
Orlando, FL 32816, USA
aidean.sharghi@gmail.com

Dr. Mubarak Shah
University of Central Florida
Orlando, FL 32816, USA
shah@crcv.ucf.edu

Center for Research in Computer Vision
University of Central Florida

Abstract

Dense video captioning is a relatively new and challenging problem in computer vision which aims at both localizing and describing all events in a video. This project focus on reproducing the results by using a state-of-art model Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning (Bidirectional SST) (Fig. 1) [3]. *Furthermore, we construct a model to generate sentences for small video clips using traditional LSTM-based encoder and decoder* (Fig. 3). *The datasets that we use are the ActivityNet Captions* [1] *and UT Ego datasets.*

Figure 1: Example of dense video captioning using Bidirectional SST model (upper row: input video; bottom row: temporally localized sentences generated by dense video captioning

1. Introduction

For the human, describing what happened in a video is not a challenge problem. For the machine, extracting the meaning from video pixels and generate them into meaningful human-produced translations is a very difficult task. Dense video captioning can be decomposed into two parts: event detection and event description. Existing methods that can be addressed to these problems is to use event proposal, captioning module, and exploit the way to combine them. Therefore, the main contribution of our work on this dense video captioning task is to create a new architecture that unify the temporal localization of event proposals and sentence generation.

2. Related Works

2.1 Temporal Action Proposals

Similar to the approach of SST [3], we take the long sequence training problem and generate proposals in a single pass without dividing the input into temporal windows or short overlapping clips. However, this method cannot produce long proposals nor exploit future context of the video. In contrast, the Bidirectional SST model we are using is able to tackle these issues.

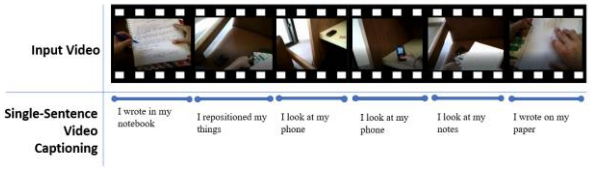

Figure 3: Example of single-sentence video captioning (upper row: input video; bottom row: sentences generated from each clip of the videos

2.2 Video Captioning

In this work, we also discover the variety of different temporal attention mechanisms that have been adopted to use in the video captioning module. In two of the methods, Bidirectional SST [3] and M3 [7], the dynamic attention mechanism that have been used to fuse visual features and the context vectors. Zhou et al. [5,6] introduces another approach that employ the cross-module attention and self-attention in their captioning module. A new problem to Dense Video Captioning, the Weakly Supervised Dense Event Captioning in Videos is recently introduced. In this work, Duan et al. presents the new idea of employing weakly supervised in videos with the use of attention mechanism to get the better generated captions.

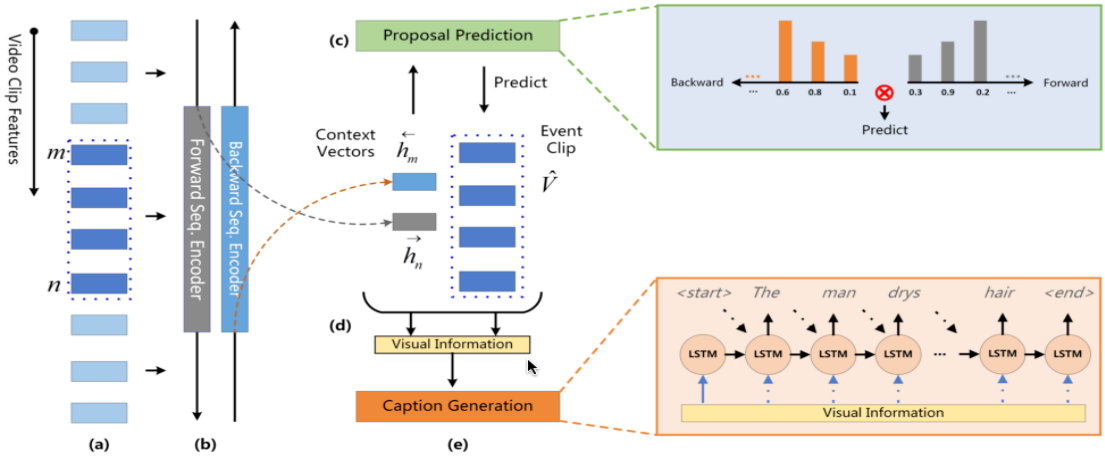

Figure 2. The main framework of our proposed method. (a) A video input is first encoded as a sequence of visual features (e.g., C3D). (b) The visual features are then fed into our bidirectional sequence encoder (e.g., LSTM). (c) Each hidden state from the forward/backward seq. encoder will be fed into the proposal module. The forward/backward seq. encoders are jointly learned to make proposal predictions. (d) Hidden states at the boundary of a detected event ($\overrightarrow{\mathbf{h}}_n$, $\overleftarrow{\mathbf{h}}_m$) will be served as context vectors for the event. The context vectors and detected event clip features are then fused together and served as visual information input. We detail the fusion methods in Section 3.2.2. (e) The decoder LSTM translates visual input into a sentence.

3. Dataset

We use both the ActivityNet Captions and UT Egocentric datasets. (1) ActivityNet Captions dataset originally comes with 20k YouTube untrimmed videos from real life where each video is 120 seconds long in average. Most of the videos comes with over 3 annotative events which are described in natural human-written sentences with corresponding start/end time. The challenges of this dataset are the large variety of video scenes and occurrences. The dataset is divided into a training set of 10024 videos, a validation set of 4926 videos, and a test set of 5044 videos, respectively. From the test split, we first withhold ground-truth annotations for competition, then train our model on the validation set and report the test set's final result. (2) The University of Texas at Austin Egocentric (UT Ego) dataset contains 4 videos captured in a natural, uncontrolled setting from head-mounted cameras. The videos are varied from 3-5 hours long and captured at 15 frames per second with the resolution of 320x480. The topics of these videos include activities such as eating, cooking, driving, shopping, and attending lecture.

4. Approaches

In this section we introduce two different frameworks to first reproduce the state-of-the-art result from the novel Bidirectional SST model (Fig. 2) and second to construct a new model to generate sentences for small-interval video clips

This custom single-sentence video captioning model is trained and tested on the UT Ego dataset while ActivityNet Captions dataset is used with the Bidirectional SST model.

4.1. Dense Video Captioning using Bidirectional SST model

4.1.1 Proposal Module

The proposal module is used to determine the temporal regions that likely contain actions or events. We first extract visual features from video frames of the ActivityNet Captions dataset through the use of a 3D CNN. Principal Component Analysis (PCA) is then applied to reduce the feature dimensionality (from 4096 to 500). These visual features are encoded by LSTM to get confidence scores, and backward pass is employed to improve event proposals. We feed the input sequence to the backward sequence encoder in the reverse order to obtain event proposals, the corresponding confidences scores, and a hidden state representation. Consequently, we fuse the two sets of confidence scores to select the proposal predictions with the highest confidence.

4.1.2 Captioning Module

Context vectors, the proposal hidden states that encode the past and future context event information of the detected proposals, are then fused with the encoded proposal detected event visual clip features. This serves as the visual information input. Finally, the LSTM-based decoder translates visual input into natural event descriptions. The total loss is calculated by summing proposal loss, captioning loss, and λ which is simply set to 0.5.

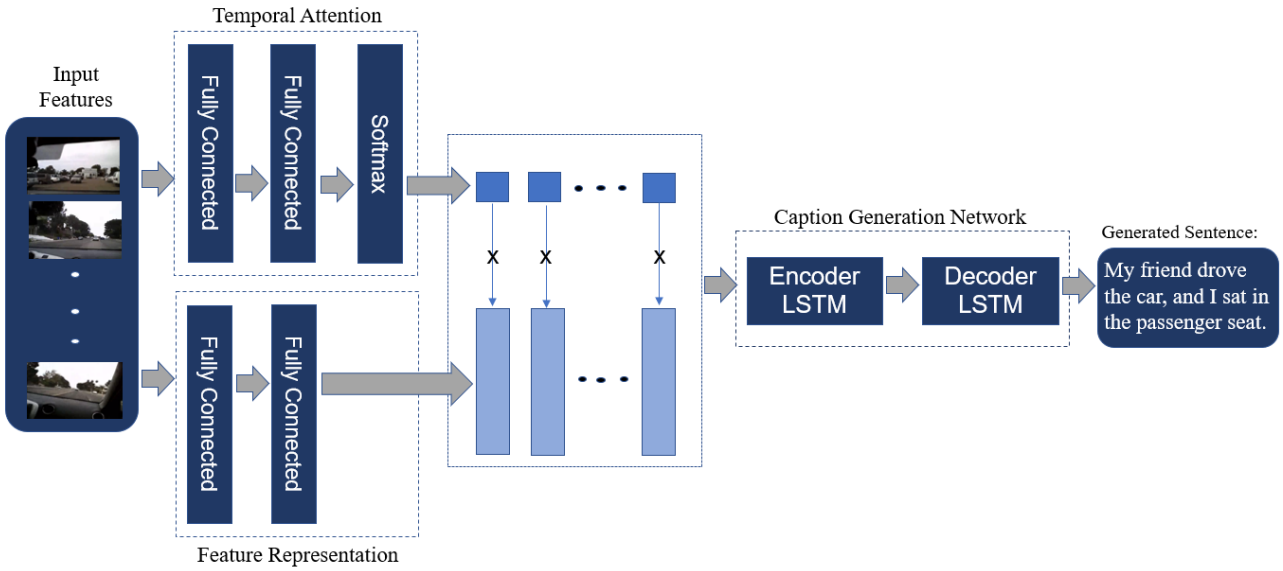

Figure 4. Single Sentence Video Caption Model

4.1.3 C3D Feature Extractor

In order to generate descriptions from any random video, we use C3D feature extractor to extract video features and use them as the inputs in caption generation model. First, we extract frames from the videos. Then visual features can be extracted from each 16 frames. The visual information features are outputted with the dimensional of 4096.

4.2. Single Sentence Video Captioning

The single sentence video captioning model (Fig. 4) generates one sentence for each video clip in the UT Ego dataset. All videos in the dataset are first uniformly divided into 5-second-long clips. From each clip, we extract visual features and use them as an input sequence. The model splits the sequence input into two halves to complete two tasks: performing temporal attention and obtaining better feature representations. We then perform multiplication on the outputs of each stage. The final product with the size of 512x5 is fed into the LSTM-based encoder to get proposal predictions. Finally, decoder LSTM is manipulated to translate these predictions into natural sentences.

5. Details

5.1 Dense Video Captioning using Bidirectional SST model

To make sure that all ground truth proposals are included, we run the network continuously in a single stream over very long input video sequences, without dividing the input into short overlapping clips or temporal windows for batch processing. At first, in order to ensure the network run properly, we train the proposal module for over 5 epochs. Then the whole model is trained in an end-to-end manner setting batch size to 1, number of anchors to be 128, and the number of proposals to be 100. Adam optimization algorithm is used with the learning rate of 0.001.

For dense event captioning, we use Meteor metric to evaluate how well the model is performing. These metric computes score for implicit word-to-word matches between machine-produce translations and human-produce reference translations. We average the meteor scores for the whole captioning system at tIou thresholds of 0.3, 0.5, 0.7, and 0.9.

5.2 Single Sentence Video Captioning

The original input with the size of 5x1024 is first fed to the network then the input is splitted into two 6x512 sub-inputs. We take one of the inputs to perform temporal attention by feeding it into two fully connected layers and perform softmax to get the output with the size of 6x1. Another sub-input is passed through two other fully connected layers to get better feature representations. We multiply the outputs from these two branches above together and feed them to the LSTM-based sequence encoder and decoder to get the final generated sentences.

6. Results

6.1. Quantitative Results from using Bidirectional SST model

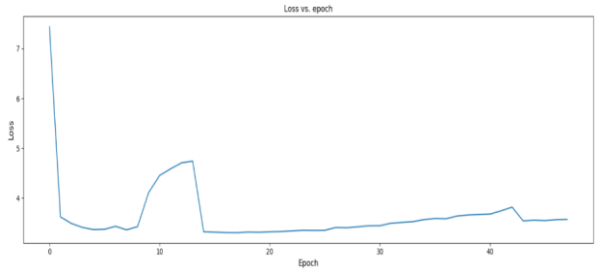

Figure 5: Loss vs epoch

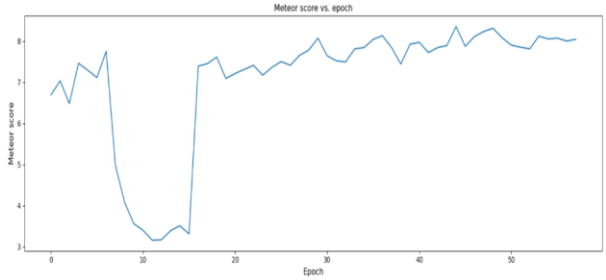

Figure 6: Meteor score vs epoch

Figure 5 and 6 indicate the loss values and meteor scores vs the number of epochs from training the validation set. Even though the values of meteor score have abruptly dropped around the 10th epoch that associated with the escalation in the graph of loss vs epoch. Overall, the values of meteor score are increasing. The results from figure 7 demonstrate the superiority of Bidirectional SST model in both localizing and describing events in dense video caption generation.

Video	Ground Truth	Proposal Captions
	A weight lifting tutorial is given	A man is seen speaking to the camera while holding a large exercise stick and leads into him moving on a mat
	The coach helps the guy in red with the proper body placement and lifting technique	A man is seen standing on a mat and moving himself around while looking to the camera.
		A man is seen standing on a mat and leads into him moving on a piece of exercise equipment.
	Men on horseback sit with polo clubs in hand	They are playing a game of polo
	The crowd disburses across the field	They are playing a game of curling
	The men begin playing polo	They are playing a game of hockey

Figure 7: Qualitative results from Bidirectional SST model

6.2. Quantitative Results from using Single Sentence Video Captioning Model

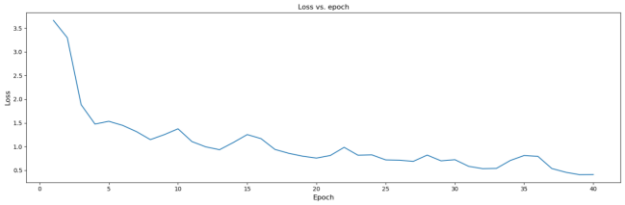

Figure 8: Loss vs epoch

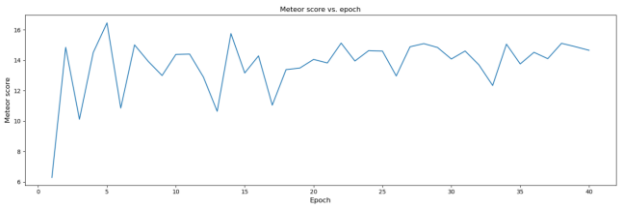

Figure 9: Meteor score vs epoch

Figure 9 demonstrates that although the meteor scores change inconsistently at the first few epochs, but we can see the line rises smoother after epoch 18.

Video	Ground Truth	Proposal Captions
	I looked at my notes	I looked a the books
	I wrote in my notebook	I looked at the instruction
	I used my laptop	I looked at the table and watch TV
	I looked at my laptop	I looked at the TV

Figure 10: Qualitative results from single sentence video captioning model

The result from figure 10 shows that our model can produce both good and bad sentences. One possible solution to improve the generated captions in our model is to train longer on the validation set.

8. Conclusion

In this project, we tackled the problem of video caption generation. We used a challenging datasets of dense event videos and long videos. We have successfully implementing the code and reproduce the result that had been reported in the research paper. Finally, we introduced our new model which is based on the LSTM and showed that our model can produce well-written sentences when trained on UT Ego dataset.

References

[1] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In ICCV, 2017

[2] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Singlestream temporal action proposals. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 6373–6382. IEEE, 2017.

[3] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. arXiv preprint arXiv:1804.00100, 2018.

[4] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. arXiv preprint arXiv:1804.08274, 2018.

[5] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. arXiv preprint arXiv:1804.00819, 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. NIPS, 2017.

[7] Wang, J.; Wang, W.; Huang, Y.; Wang, L.; and Tan, T. 2016. Multimodal memory modelling for video captioning. arXiv preprint arXiv:1611.05592.

[8] Jiang Zhu, Wei Zou, Zheng Zhu. End-to-end video-level representation learning for action recognition. arXiv preprint arXiv:1812.03849

[9] Bairui Wang, Lin Ma, Wei Zhang, Wei Liu. Reconstruction network for video captioning. arXiv preprint arXiv: 1803.11438

[10] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, Junzhou Huang. Weakly supervised dense event captioning in videos. arXiv preprint arXiv:1812.03849

[11] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, William Yang Wang. Video captioning via hierarchical reinforcement learning. arXiv preprint arXiv:1711.11135

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri. C3D: Generic features for video analysis. arXiv preprint arXiv:1412.0767