NITS-VC System for VATEX Video Captioning Challenge 2020

Presented by
Alok Singh
alok.rawat478@gmail.com

Other Authors
Dr. Thoudam Doren Singh
thoudam.doren@gmail.com
NIT, Silchar, India

Prof. Sivaji Bandyopadhyay
sivaji.cse.ju@gmail.com
NIT Silchar, India

Department of Computer Science & Engineering
National Institute of Technology
Silchar, Assam, India

June 16, 2020
Contents

1. Introduction
2. Objective
3. System Description
4. Experimental Setup
5. Results
6. Conclusion
Introduction

Video Captioning
- Short and informative textual description of the content, event and action of the video.

Applications of Video Captioning
- Effective video indexing and retrieval.
- Video guided Machine Translation (MT) [1].
- Video sentiment analysis.
- Aid for visually impaired people.
What is video?

- A sequence of frames with a specific frame rate accompanied by an audio track.

Figure: Sequence of frames in a video.

- Multiple scenes containing multiple events and actions.
Objective of the challenge

- For a given video, we have to generate a suitable caption based on the content, events and action in the video.

Statistics of Dataset

| Dataset Split      | #Videos | #English Captions | #Chinese Captions |
|--------------------|---------|-------------------|-------------------|
| Training           | 25,991  | 259,910           | 259,910           |
| Validation         | 3,000   | 30,000            | 259,910           |
| Public test set    | 6,000   | 30,000            | 30,000            |
| Private test set   | 6,287   | 62,780            | 62,780            |
VATEX-2020: System Description

- For this task a traditional encoder-decoder based approach is used.
- The encoder-decoder framework based on the concept of encoding the video into a context vector \( c_t \) and decoded them using a suitable decoder.
- Objective function of encoder-decoder based framework.

\[
y_{\theta^*} = \arg\max_{\theta} \sum_{(V,y)} \log p(y|V; \theta) \tag{1}
\]

where \( \theta \) are the parameters of the model, \( V \) is a video and \( y = \{y_1, y_2 \ldots y_t\} \).
Encoder

- C3D (3D Convolutional Neural Network) pre-trained on Sports-1M dataset [2, 3]
  1. Firstly, the video is evenly segmented into \( n \) segments in the interval of 16.
  2. A visual feature vector \( f = s_1, s_2...s_n \) for video is extracted.
  3. Feature reduction using average pooling with filter size 5.
For the decoding, two Long Short Term Memory (LSTM) recurrent network are used.

1. An Embedding layer is used to get a dense representation for each word in the input caption.
2. The first LSTM takes the output of embedding layer as an input and an encoded visual feature vector as an initial stage.
3. For the second LSTM, the visual feature vector concatenated with the output of embedding layer.
4. Finally, element wise product is preformed between the output from both LSTM.
5. The unrolling procedure of system is given below:

\[ \tilde{y} = W_e X + b_e \]  

\[ \tilde{z}_1 = LSTM_1(\tilde{y}, h_i) \]  

\[ \tilde{z}_2 = LSTM_2([\tilde{y}; f_a]) \]  

\[ y_t = softmax(\tilde{z}_1 \odot \tilde{z}_2) \]
Experimental Setup

- Each caption is concatenated by two special marker \(<\textit{BOS}>\) and \(<\textit{EOS}>\).
- The maximum number of words in a caption is upto 30, and masking with zero.
- 15K words with most occurrence are retained, for out-of-vocabulary words, a special tag UKN is used.
- Cross-entropy loss function is used with ADAM optimizer and learning rate is set to $2 \times 10^{-4}$.
- Dropout of 0.5 is used and the hidden units of both LSTMs are set to 512 units, batch size is 64.
# Results

**Table**: Performance of the system on public dataset

| Evaluation Metrics | Proposed System on public test set | Proposed System on private test set |
|--------------------|------------------------------------|-------------------------------------|
| CIDEr              | 0.24                               | 0.27                                |
| BLEU-1             | 0.63                               | 0.65                                |
| BLEU-2             | 0.43                               | 0.45                                |
| BLEU-3             | 0.30                               | 0.32                                |
| BLEU-4             | 0.20                               | 0.22                                |
| METEOR             | 0.18                               | 0.18                                |
| ROUGE-L            | 0.42                               | 0.43                                |
Conclusion

We have used encode-decoder based video captioning framework for the generation of English captions. Our system scored 0.20 and 0.22 BLEU-4 score on public and private video captioning test set respectively.
X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, “Vatex: A large-scale, high-quality multilingual dataset for video-and-language research,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 4581–4591, 2019.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in Proceedings of the IEEE international conference on computer vision, pp. 4489–4497, 2015.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732, 2014.
Thank you