Submission for Vatex Competition: A Simple Baseline

Tao Jin*, Xinglu Wang*
The College of Information Science and Electronic Engineering, Zhejiang University
Hangzhou 310027, China
{jint_zju, xingluwang}@zju.edu.cn

Abstract

The submission reports our experimental results in Vatex Competition at ICCV 2019. We propose a simple multi-task baseline model with a shared encoder and two independent decoders for both English video captioning task and Chinese video captioning task without reinforcement learning. The multimodal features of video and text are fully utilized.

1. Introduction

The goal of traditional video captioning task is to generate descriptions for video data in only one language, such as English, Chinese. We therefore call it monolingual video captioning. Recently, [6] builds a multilingual video captioning dataset, in which each video has both English and Chinese descriptions. In addition, [6] proposes three baseline models with motion features. However, such baseline models ignore rich multimodal information in video data. Existing methods of monolingual video captioning have verified the advantage of the information. In this paper, we develop a multimodal baseline model for multilingual video captioning.

2. Method

Our proposed model consists of one shared encoder and two separate decoders for both English and Chinese, which is simialr to [6]. In the encoder, the extracted multimodal features are processed by Bi-LSTM. In the decoder, we employ the multimodal attention mechanism at each time step.

2.1. Encoder

We extract global temporal features from video data, including image, motion, and audio features. These features are sent to Bi-LSTM which has a capability to process sequence data, capturing the features of forward and backward directions with two opposite LSTMs.

In detail, we use Inception-ResNet-v2[4] to extract image features, we use I3D[1] to extract motion features, and we use VGGish[2] to extract audio features.

2.2. Decoder

We utilize a unidirectional LSTM in the decoder. At each time step, we combine the output of LSTM and the features provied by the encoder to predict words. In addition, we consider the importance of text modality and utilize the information from previous $t-1$ words at time step t. Therefore, we utilize 4 modalities in the multimodal attention mechanism.

The structure of the decoder is similar to [3] and [5], we implement a hierarchical attention mechanism for the multimodal features.

3. Experiments

3.1. Dataset

The vatex captioning dataset contains 34991 video clips. The time length of a video clip is about 10-25 seconds and each video clip is annotated with about 10 English sentences and 10 Chinese sentences. We follow the previous work and take 25991 videos for training, 3000 for validation, and 6000 for testing. In the experiments, we four common metrics, BLEU4, ROUGE, METEOR, and CIDEr.

3.2. Preprocessing

Since the original videos are not provided, we download them with youtube-dl[1] module which is efficient. However, some videos are unavailable now.

we sample video data for extracting image features. Each video is sampled into 80 frames. If the number of video frames is more than 80, we sample at the regular intervals. Otherwise, we add paddings at the end of the video. For extracting motion features, we divide the raw video data into video chunks centered on 80 sampled frames at the first step. Each video chunk includes 64 frames.

* equal contributions

[1] https://github.com/activitynet/ActivityNet/tree/master/Crawler/Kinetics

Table 1. The results on test set

method	English				Chinese			
	BLEU4	**ROUGE**	**METEOR**	**CIDEr**	**BLEU4**	**ROUGE**	**METEOR**	**CIDEr**
Motion only	28.4	47.0	21.7	45.1	24.9	51.7	29.8	35.0
Multimodal+text	31.2	48.6	22.7	49.2	26.1	52.4	30.4	37.7

3.3. Experimental Details

The hidden state size is 512 for all LSTMs. The attention layer size is also 512 for both temporal and object attentions. We add layer normalization to the LSTM decoder. The dropout rate of both input and output of the LSTM decoder is 0.5. In the training stage, we use Adam algorithm to optimize the loss function; the learning rate is set to 0.001 and halved when the CIDEr score of validation set does not increase for 4 epochs. In the testing stage, we use beam search method with beam size 5. The English and Chinese word vector matrices are initialized by pre-trained word2vec. Each word is represented as a 300-dimension vector.

3.4. Results

The results of our proposed model are a little better than the unimodal baseline as shown in Table 1.

4. Conclusion

There are still many points to dig on multilingual video captioning, such as multi-task learning, dual reinforcement learning. In the future, we will try these directions.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[2] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[3] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4203–4212. IEEE, 2017.

[4] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[5] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *arXiv preprint arXiv:1804.05448*, 2018.

[6] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *arXiv preprint arXiv:1904.03493*, 2019.