Audio-Visual Fusion Model for Video Captioning by Utilizing Deep Neural Networks

Chenxiao Guan, *Department of Computer Science, University of Rochester*

Introduction

Video captioning has been a challenging problem in the computer vision field as the difficulties to extract the features represented by the video since most models have treated videos as a plain sequence of visual contents, while ignoring the multimodality (e.g frames and audios). This project aims to utilize deep convolutional and recurrent neural network structures, along with newly designed tricks and mechanisms, to elegantly combine both visual and auditory information inside the video, and help an Artificial Intelligent odel to understand the video content better, in order to generate a more convincing sentence to describe what is happening in the video.

Method

The generic approach of captioning a video is presented in Figure 1. A raw video will be separated into visual frames and raw wav file. These two sequences of features will then be fed into two LSTM encoders to learn the temporal structures of the video, and the outputs are elegantly combined by the child-sum unit and multi-level attention mechanism. Finally, features are fed into a LSTM decoder to generate the caption of the video. All models are implemented by the pytorch deep learning framework.

Audio MFCCs are extracted for each one-second frame size, each frame is separated into 32 segments and 20 MFCCs are extracted for each segment, and feed into the audio LSTM encoder. Each video will extract 80 video frames into jpg images, and feed into a ImageNet pretrained deep CNN to extract a 2048 dimensional feature map and then feed into the visual LSTM encoder.

The mini-batch sizes were set to 128 for the vanilla mean pool model, and 32 for the Multi-level Attention model. The dimension of all LSTMs are set to 1024. Models were trained for 500 epochs, by using the ADAM optimizer with a learning rate of 10e-4, along with a LR decay rate of 0.8 for each 100 epochs.

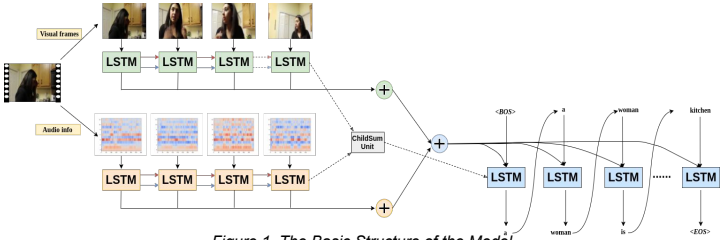

Figure 1. The Basic Structure of the Model

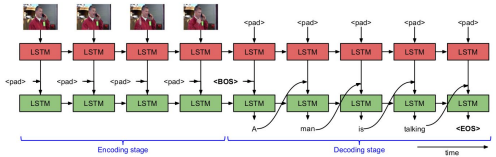

Figure 2. Vanilla S2VT[4]

Result Analysis

Below is the comparison between some current state-of-the-art results with my own implementations. Figure 2 above demonstrates the structure of the vanilla S2VT.

Methods	BLEU@4	METEOR	CIDEr-D	ROUGE-L
MA-LSTM (G+C+A)(childsum)[1]	36.5	26.5	41.0	59.8
hLSTMat (R)[2]	38.3	26.3	-	-
R+MCNN+MCF-matrix multiply [3]	38.1	27.2	42.1	-
My own methods				
S2VT (without audio)[4]	29.3	25.2	-	54.5
Vanilla mean-pool	33.7	26.6	41.4	57.9
Multi-level Attention	35.8	26.3	40.2	57.9

Table 1. Implementations comparison (scores higher the better)

All results are reported by using the MSR-VTT dataset, which is a popular dataset for general video understanding model training purpose. For all 10,000 videos with caption, the dataset was separated into 8000/1000/1000 for train, validate and test.

By looking at the table presented on the left, one can notice that there is a significant improve on the performance of the model, when I added the audio feature as a helper for the model. Compare with all state-of-the-art results, there is still some distance in performance between my model with theirs, given the fact that my model is much simpler, easy to run and takes less space.

For the future improvement, there is definitely some other implementations to do, including come up with my own fancy structure of processing the data. But, as long as it is concerned for this project, my main goal was to see if there will be any improvement for the video captioning model to outperform the vanilla model without audio features, when I take audio features into account.

Acknowledgements

This project was directed under professor Chenliang Xu at the University of Rochester, Department of Comp- uter Science. Special Thanks to Yapeng Tian, who is my Ph.D. mentor under the same group and helped me a lot in this project.

References

[1]Xu, Jun, et al. "Learning Multimodal Attention LSTM Networks for Video Captioning." Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017.

[2]Song, Jingkuan, et al. "Hierarchical LSTM with adjusted temporal attention for video captioning." arXiv preprint arXiv:1706.01231 (2017).

[3]Wu, Aming, and Yahong Han. "Multi-modal Circulant Fusion for Video-to-Language and Backward." IJCAI. 2018.

[4]Venugopalan, Subhashini, et al. "Sequence to sequence-video to text." Proceedings of the IEEE international conference on computer vision. 2015.

This research was funded by a generous gift from the Xerox Corporation and the Office of the President.