1Q-08

Research on Video Captioning with a Late Fusion Based Multimodal Transformer Network

Fei BAO† Takaaki ISHIKAWA‡ Hiroshi WATANABE†‡
†Department of Communications and Computer Engineering, Waseda University
‡Global Information and Telecommunication Institute, Waseda University

1. Introduction

Video captioning has drawn great attention in recent years as the intersectional task of computer vision and natural language processing. It can automatically generate descriptions for a given video with a natural human language. As the video is a combination of different modalities, not only appearance features from visual contents but audio features and textual features are also taken into the consideration to understand the video. However, these multiple modality features are simply concatenated before input into the video caption model. This leads to a higher computational cost and even quadratic regarding the input length of features. In addition, visual features are usually obtained from the deep layer of a pre-trained backbone but languages are not. It means that the model thereby needs to concentrate most of the computation on the linguistic representations.

To mitigate the above issues, we propose a Transformer based model that handles different modalities with separated encoders and fuses them at the decoder side.

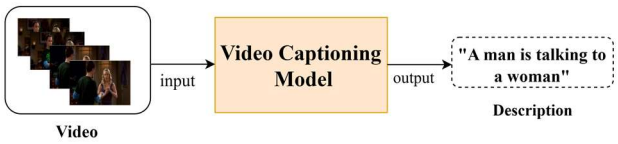

Figure 1. Illustration of the video captioning task

2. Related Works

2.1 Attention mechanism

Vaswani et al. [1] proposed the Transformer, a convolution-free and recurrence-free neural network relying entirely on an attention mechanism to compute representations of a sequence. The self-attention mechanism in the Transformer model is implemented using scaled dot-production, which projects the input sequence into matrix queries(Q), keys(K), values(V) and subsequently calculates the attention weights by taking the dot product of each query and paired keys. Then, attention weights are normalized by applying a softmax function to them and proceeding to obtain the final attention score by multiplying them with values. In practice, the attention function is applied to the entire sets of queries simultaneously and the matrix of outputs can be performed as:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

Intuitively, the computational complexity is $O(n_A^2 \cdot d)$, which denotes that it highly depends on the length of the input sequence A. Therefore, it turns to $O((n_A + n_B)^2 \cdot d)$ if visual sequence A and linguistic sequence B are directly concatenated before the encoder.

2.2 Dataset and baseline

Lei et al. [2] introduced the TV show Captions (TVC) dataset, a large-scale multimodal video captioning dataset. TVC contains 108K video clips paired with subtitles from 6 TV shows across diverse genres. Besides, each video clip is on average 13.4 seconds in length and has 2 or 4 manually annotated descriptions.

In addition, they also proposed an early fusion based Multimodal Transformer (MMT) [2] as a baseline model on the TVC dataset. To efficiently understand the interaction between two modalities and produce an accurate caption, visual features and subtitles are concatenated simply and then input into the encoder for learning context from different modalities. Although it is a convenient way to utilize multiple modalities, computational complexity is also increased for the longer sequence after

Research on Video Captioning with a Late Fusion Based Multimodal Transformer Network
†Fei BAO ‡Takaaki ISHIKAWA ††Hiroshi WATANABE
†Department of Communications and Computer Engineering, Waseda University
‡Global Information and Telecommunication Institute, Waseda University

concatenation, and self-context in the linguistic section cannot be explored completely due to the early fusion.

3. Proposed Method

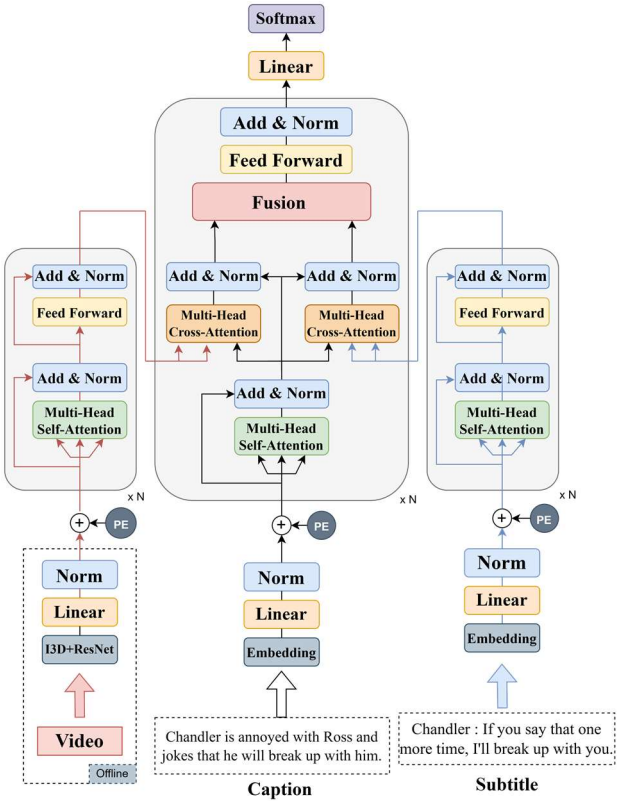

Figure 2. The framework of our proposed method

As shown in Fig. 2, visual and subtitle sequences are processed in separate encoders and video features are extracted beforehand to save the GPU memory. The visual encoder and the decoder have 2 layers while the subtitle encoder has different numbers of layers ranging from 2 to 8 for the experiment. Two types of sequences are summed at fusion layer in the decoder after implementing cross-attention computation with the ground truth. Therefore, the computational complexity can be reduced to $O((n_A^2 + n_B^2) \cdot d)$ as each self-attention score of two modalities is computed individually.

4. Experiment results

The inputs of the video encoder including extracted appearance and motion features from videos are summed and normalized, just like in the MMT. We conducted several experiments to test using different numbers of layers in the subtitle encoder. The results using 2 layers and 4 layers are shown in Table 1 and Fig. 3, it demonstrates that proposed method outperforms the MMT in terms of all four evaluation metrics. In particular, we found that the model has the highest value for CIDEr-D when the linguistic encoder has 4 layers.

Table 1. Comparison with the baseline on the TVC *val* set

	BLEU-4 ↑	METEOR ↑	Rouge-L↑	CIDEr-D ↑
MMT [2]	10.53	16.61	32.35	44.39
Ours (2-layers)	10.79	**16.72**	32.57	45.61
Ours (4-layers)	**11.19**	16.54	**32.67**	**45.65**

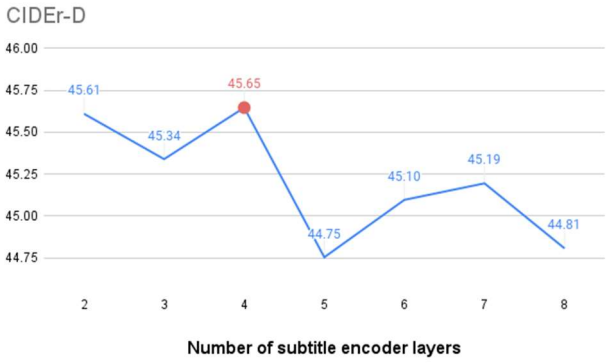

Figure 3. The value of CIDEr-D regarding to increasing layers of the subtitle encoder.

5. Conclusion

In this research, we investigated the effect of different multimodal fusion strategies in a Transformer network on video captioning tasks. We discovered that using distinct encoders for multiple modalities and fusing later tends to perform better and have a lower computational complexity compared to using a single encoder. Additionally, we found that focusing more on the linguistic modality leads to better results.

References

[1] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin: “Attention Is All Your Need”, In NIPS, pp.6000-6010, Dec. 2017

[2] J. Lei, L. Yu, T. Berg, M. Bansal: “TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval”, In ECCV, pp.447–463, Aug. 2020