# XlanV Model with Adaptively Multi-Modality Feature Fusing for Video Captioning

Yiqing Huang
Tsinghua University
huangyq95@163.com

Qiuyu Cai
Beijing University of Posts and Telecommunications
caiqiuyu@bupt.edu.cn

Siyu Xu
Shanghai Ocean University
1759236@st.shou.edu.cn

## 1 INTRODUCTION

Some recent work explores the structure of X-Linear Attenion networks [1], which exploit static object feature to facilitate image captioning, but the situation of video captioning is more challenging. Unlike still images, video contains both static feature and dynamic feature as the length of video varies. Therefore, we propose to adaptively fuse these two kinds of features to make better utilization of video features in video captioning.

The main contributions of our method are as follows:

- We propose XlanV model to introduce X-Liner Attention networks into video captioning.
- Adaptively fusing multi-modality features to enhance video captioning.

## 2 METHOD

### 2.1 Feature Extraction

Our model utilizes two kinds of features. 1) We transform each video into 40 frames of images and leverage a pre-trained ResNet-152 network to extract 40x2048 static features. 2) We extract 160 frames of images from the video and exploit the action classification network I3D to map each 8 frames of images to a 1024d feature.

### 2.2 XlanV Model for Video Captioning

The overall paradigm of our model, which leverages the X-Linear Attention network [1] as the backbone framework, is shown in Fig. 1. Two encoders are implemented to encode the static feature and the dynamic features respectively. In the LSTM decoder, our model adaptively fuse these two kinds of features.

### 2.3 Adaptive Multi-modality Fusion

Denoting the attended static feature and dynamic feature as $\hat{\boldsymbol{v}}_t^r$ and $\hat{\boldsymbol{v}}_t^i$ respectively, the formulation of adaptive multi-modality fusing can be formulated as follows, where $\mathbf{W}$ are trainable parameters and ';' denotes concatnation.

$$\hat{\boldsymbol{v}}_t^{fuse} = \alpha * \hat{\boldsymbol{v}}_t^r + (1-\alpha) * \hat{\boldsymbol{v}}_t^i \quad (1)$$

$$\alpha = sigmoid(\mathbf{W}[\hat{\boldsymbol{v}}_t^r; \hat{\boldsymbol{v}}_t^i]) \quad (2)$$

Thus, our model is capable of adaptively weighting these two features and makes better utilization of one kind of feature when the other one is not so useful at the current time step.

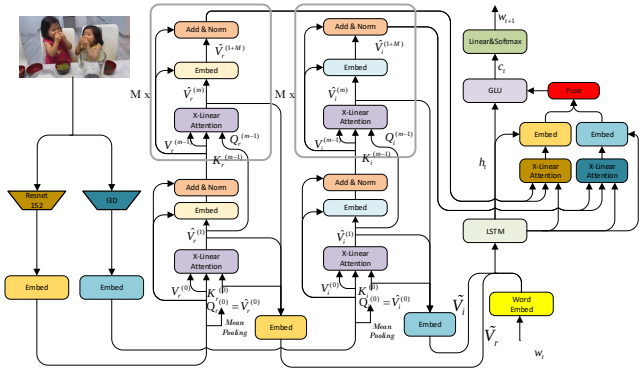


**Figure 1: The overall framework of our model. Due to space limitation, only one time step of the LSMT decoder is shown.**

**Table 1: Video Captioning Result**

| Model | Bleu_4 | METEOR | CIDEr-D |
|---|---|---|---|
| Bi-LSTM | 0.1504 | 0.1672 | 0.1898 |
| XlanV | 0.1764 | 0.1614 | 0.1928 |
| XlanV+SCST | 0.1950 | 0.1710 | 0.2391 |
| XlanV+SCST+ens | 0.2115 | 0.1739 | 0.2443 |

## 3 RESULT

Due to the large difference between the test data and the GIF dataset, the GIF dataset was not used in the training process. We only use the MSR-VTT dataset, which is more similar to the test video, as training and validation dataset.

Table 1 shows the test performance of the XlanV model. The comparison baseline uses the results of a Bi-LSTM with attention mechanism network. We train our model with both cross-entropy loss and the reinforcement learning based SCST [2]. In addition, ensembling multi-model can obtain better results.

## 4 CONCLUSION

In this paper, we introduce a structure of X-linear Attention network for video captioning, which fully integrates video features by adaptively fusing multi-modality video features.

## REFERENCES

[1] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-Linear Attention Networks for Image Captioning. In *CVPR*. 10971–10980.
[2] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*. 7008–7024.