Semantic Tag Boosted XlanV Model for Video Captioning

Yiqing Huang*
Tsinghua University
huang-yq17@mails.tsinghua.edu.cn

Hongwei Xue*
University of Science and Technology of China
gh051120@mail.ustc.edu.cn

1 INTRODUCTION

Recent works verified that the semantic tags, like the classes of the images, are beneficial in the visual language pre-training for image related tasks [1]. Therefore, we boost the champion model XlanV [2] of the last year with the semantic tags in the 2^{nd} competition this year. We also propose various pre-training tasks and modules to effectively exploit the semantic tags.

The main contributions of our method are as follows:

- We propose semantic tag boosted XlanV model to introduce semantic tags into video captioning.
- We design several pre-training tasks and novel structures to handle the task of video captioning

2 METHOD

2.1 Feature Extraction

Our Semantic Tag boosted XlanV (ST-XlanV) model utilizes four kinds of features.

2 kinds of visual features: We transform each video into 40/20 frames of images and leverage pre-trained ResNet-152/S3D networks to extract 40x2048/12x1024 static/dynamic features.

2 kinds of semantic features: The semantic tags are obtained as a by-product when we extract the visual features. We encode the top-1 class of the video in both static and dynamic perspective to formulate the semantic tags. In the pre-training process, we also incorporate the captions as an input language feature.

2.2 Semantic Tag Boosted XlanV Model

The overall paradigm of our model, which leverages the X-Linear Attention network [4] as the backbone framework, is shown in Fig. 1. Four encoders are implemented to encode the four kinds of features respectively. To better leverage the sequential information of the visual features, we also adopt trainable positional encodings in the encoders for static and dynamic features. The encoded features are then concatenated and sent to a cross-attention module to further explore the cross-modal interactions between these features. The LSTM decoder finally utilizes the crossed output to formulate plausible captions.

2.3 Pre-training Tasks

We explore three pre-training tasks to exploit the ACTION dataset.

Mask Language Modeling: Similar to BERT, we randomly mask 15% words in the input caption and predict these words in pre-training.

Tag Alignment Detection: We randomly replace the semantic tags of the current video with other tags with a chance of 50% and predict whether the tags have been replaced.

*Equal Contribution.

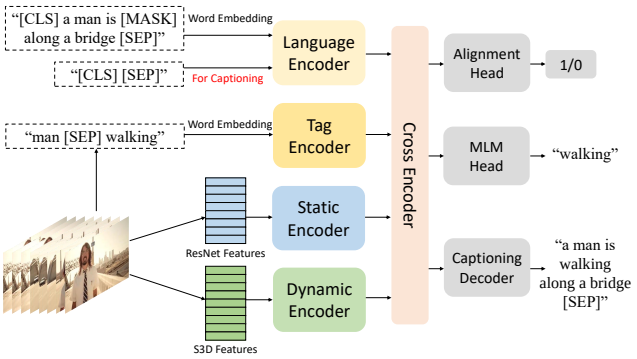

Figure 1: The overall framework of our model. Language, semantic tags, static and dynamic visual features are separately encoded and fused in a cross encoder. In captioning task, all words are invisible for language encoder.

Table 1: Performance comparisons on online testing server.

Model	B@4	M	C	S
TransED$_{RL}$ [3]	19.5	16.8	23.9	5.4
ST-XlanV$_{RL+ens}$	20.93	17.35	24.42	5.60

Video Captioning: We include the auto-regressive decoder with cross-entropy objective to perform typical captioning.

3 RESULT

Table 1 shows the test performance of the semantic tag boosted XlanV model and official baseline. We train our model with both cross-entropy loss and the reinforcement learning based SCST. In addition, ensembling multi-models can obtain better results.

4 CONCLUSION

In this paper, we introduce a structure of Semantic Tag boosted XlanV model for video captioning. We adopt effective pre-training tasks and novel network structures to fully exploit the semantic tags and enhance video captioning.

REFERENCES

[1] Xiujun Li et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.
[2] Yiqing Huang et al. 2020. XlanV Model with Adaptively Multi-Modality Feature Fusing for Video Captioning. In *ACMMM*.
[3] Yingwei Pan et al. 2020. Auto-captions on GIF: A Large-scale Video-sentence Dataset for Vision-language Pre-training. *arXiv preprint arXiv:2007.02375* (2020).
[4] Yingwei Pan et al. 2020. X-Linear Attention Networks for Image Captioning. In *CVPR*.