Video Captioning with Pretraining Techniques

Sihan Chen[1,2], Xinxin Zhu[1], Dongze Hao[1,2], Wei Liu[1,2]
Jiawei Liu[1,2], Zijia Zhao[1,2], Longteng Guo[1,2], Jing Liu[1,2]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
chensihan2019@ia.ac.cn, xinxin.zhu@nlpr.ia.ac.cn, haodongze2021@ia.ac.cn, liuwei2019@ia.ac.cn
liujiawei2020@ia.ac.cn, zhaozijia2021@ia.ac.cn, longteng.guo@nlpr.ia.ac.cn, jliu@nlpr.ia.ac.cn

1 APPROACH

Video captioning aims to generate the textual description of video contents, in which the key tasks are to understand video semantics from the multi-modal features and express its semantics with natural language. In this work, we design two kinds of pretraining schemes to enhance the multi-modal feature representations of videos. First, we explore some single-modality pretraining models to extract visual, audio, and motion features, respectively, and then fuse these features to generate suitable textual description, in which the multi-modal fusion module and the textual decoder are learned from the training dataset of the challenge. Second, we directly import a multi-modality pretrained model for video feature encoding and decoding, in order to inherit the cross-modality correlations in the pretrainer, and further fine-tune the model in the task of video captioning. Moreover, we augment the challenge trainset with the VATEX dataset to further improve the model generalization capacity. In the implementation, we first train the network with cross-entropy loss and then finetune it with the self-critical sequence training method, where the reward is designed as the summation of BLEU and CIDEr score.

Single-Modality Pretrained Feature Fusion. The framework of the model trained with the single-modality pretrained feature fusion techbique is depicted in Figure 1. We use a multi-path XLAN network to encode and decode the multi-modality features including appearance, motion, region and audio features, which are extracted by a FixResNeXt-101 network pretrained on the ImageNet-1k dataset, a irCSN-152 network pretrained on the Kinetics-400 dataset, a pretrained VinVL model and a CNN14 model pretrained on the AudioSet dataset, respectively. The multi-modality features are first embedded into the same latent space and then passed into independent XLAN encoders. At the every decoding time step, the multi-modality global features are aggregated into a single vector and then passed into the LSTM together with the word embedding. We take the hidden state of the decoder LSTM as query and the encoded multi-modality features as keys to achieve four context features via X-Linear attention mechanism, after that we use the hidden state and the aggregated context feature to predict the next token via a linear classifier. The operation of the two aggregation modules can be selected from concatenation, average pooling and additional attention. We have tried different optional combinations of the two aggregation modules and finally we take concatenation as the input aggregation module and average pooling as the context aggregation module for its best performance on the validation set.

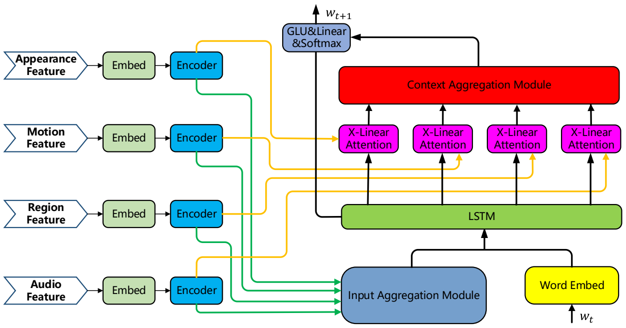

Figure 1: The framework of model trained with single-modality pretrained feature fusion technique.

Multi-Modality Pretrained Model Finetuning. We utilize a pretrained Omni-Perception Pre-Trainer model[1](OPT) based on an encoder-decoder transformer architecture and finetune it on the MSR-VTT dataset. Specifically, we pass the appearance and region features into the OPT's vision encoder and the audio feature into OPT's audio encoder, after that multi-modality features interact with each other through the cross-modal encoder and are translated into predicted sentences via OPT's decoder.

2 RESULT

As shown in Table 1, based on the proposed single-modality pretrained feature fusion technique, introducing more training data, i.e., the VATEX dataset indeed help our model generalize better. Otherwise, ensemble serveral models trained with two pretraining techniques can further enhance the generalization capacity and achieve the best results.

Model	Training Data	BLEU4	METEOR	CIDEr	SPICE
SMPFF	M*	25.73	18.87	28.24	6.49
MMPMF	M*	24.28	18.58	27.78	6.19
SMPFF	M	26.69	19.38	30.23	6.79
SMPFF	M+V	22.88	20.51	31.54	7.82
ENSEMBLE	-	26.13	**20.86**	**35.10**	**7.85**

Table 1: Experiment results on the test server. All models are finetuned with self-critical sequence training method. All models are tested with beam search method and the beam size is 3. SMPFF: Single-Modality Pretrained Feature Fusion, MMPMF: Multi-Modality Pretrained Model Finetuning, M*: standard trainset of MSR-VTT(6513 videos),M: MSR-VTT(10000 videos), V: standard trainset and testset of VATEX(31991 videos).

References

[1] Jing Liu, Xinxin Zhu, Fei Liu, Longteng Guo, Zijia Zhao, Mingzhen Sun, Weining Wang, Jinqiao Wang, and Hanqing Lu. 2021. OPT: Omni-Perception Pre-Trainer for Cross-Modal Understanding and Generation. *arXiv preprint arXiv:2107.00249* (2021).