Video Captioning System with Excellent Multi-Modal Feature

Sihan Chen[1,2], Dongze Hao[1,2], Jiawei Liu[1,2], Zijia Zhao[1,2], Longteng Guo[1,2], Xinxin Zhu[1], 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, haodongze2021@ia.ac.cn, liujiawei2020@ia.ac.cn, zhaozijia2021@ia.ac.cn
longteng.guo@nlpr.ia.ac.cn, xinxin.zhu@nlpr.ia.ac.cn, jliu@nlpr.ia.ac.cn

1 APPROACH

Our video captioning system takes a general encoder-decoder paradigm, in which the encoder extracts rich semantic features from the input videos and the decoder translate the abstract representation into natural language sentences. Based on the basis framework, we address the video captioning challenge from three key perspectives: excellent feature, multi-modal fusion and augmented training data.

1.1 Excellent Feature

For each video, we sample 30 frames equal-intervally and exploit four kinds of features including appearance, motion, region and audio features.

Appearance Feature. We use a FixResNeXt-101 32x48d network pretrained on ImageNet-1k dataset(86.4% top-1 accuracy) to extract the 2048-dim appearance feature for each sampled frame, resulting in a 30×2048 representation.

Motion Feature. For each sampled frame we take 64 frames around it as the input to a irCSN-152 network pretrained on Kinetics-400 dataset(82.8% top-1 accuracy) to extract the 2048-dim motion feature, resulting in a 30×2048 representation.

Region Feature. We sample 5 frames equal-intervally and we use the pretrained VinVL model to detect the object regions. For each sampled frame, we pick the top-10 confidence proposal feature, resulting in a 50×2048 representation.

Audio Feature. We use a CNN14 model pretrained on AudioSet dataset to extract the audio feature for each video and resize it to a 30×2048 representation via linear interpolation.

1.2 Multi-Modal Feature Fusion

The overall architecture of our video captioning network is depicted in Figure 1. We use a multi-path XLAN network to encode and decode the multi-modal features. The multi-modal features are first embedded into the same 1024-dim semantic space and then passed into independent encoders. At the every decoding time step, the multi-modal 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 LSTM as query and multi-modal features as keys to achieve four context features via X-Linear attention mechanism, and then 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 due to its simplicity and effectiveness.

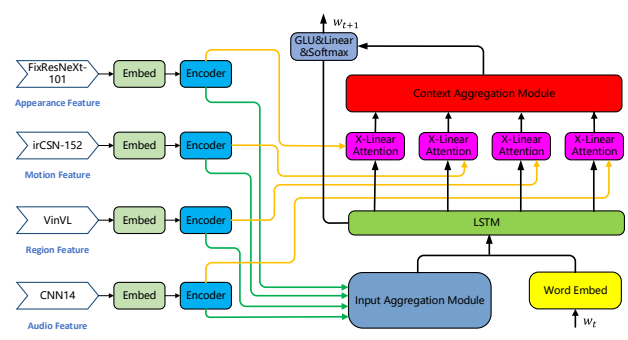

Figure 1: The details of model architecture.

1.3 Training Details

We first train the network with cross-entropy loss and then finetune it with the self-critical sequence training method. The reward is designed as the summation of BLEU and CIDEr score.

Another challenge is that the data distribution of the online testset is not consistent with the MSR-VTT dataset. We relieve this problem by introducing more training data. Specifically, we haven't use the ACTION dataset due to its large difference compared to the MSR-VTT dataset in the aspect of video length and annotation style. Instead, we additionally utilize the VATEX dataset and train models based on the mixed dataset of MSR-VTT and VATEX. We found it experimentally helpful for the reason that some videos of the online testset is closer to the VATEX's distribution.

Model	BLEU4	METEOR	CIDEr	SPICE
trained with MSR-VTT	26.69	19.38	30.23	6.79
trained with mixed dataset	22.88	20.51	31.54	7.82
model ensemble	26.13	**20.86**	**35.10**	**7.85**

Table 1: Experiment results on the online testset. All models are finetuned with self-critical sequence training method. All models are tested with beam search method and the beam size is 3.

2 RESULT

As shown in the first row of Table 1, the well-designed multi-modal features endow our model powerful capacity to understand video contents. Otherwise, taking the VATEX dataset as additional training data indeed improves three metrics, but we observe a large performance drop regarding to the BLEU score. It is reasonable because the predicted sentences become longer and the vocabularies get richer due to introducing the VATEX dataset, but the BLEU metric is based on the precision. We address this problem by ensembling 4 models trained with the mixed dataset using different random seeds and 1 model trained with the MSR-VTT dataset.