# Multi-head Multi-dimension Feature Enhancement for Video Captioning

Chengbo Dong, Xinru Chen, Aozhu Chen, Fan Hu, Fangming Zhou, Zihan Wang
dongchengbo@ruc.edu.cn, chen_xinru1999@163.com
AIMC, Renmin University of China
Beijing, China

## 1. Method

### 1.1. Dataset

Large differences exist between provided pretraining dataset action-GIF [7] and test data as well as MSR-VTT in terms of video length, caption style and so on. In such consideration, we select more than 9k videos according to referenced factors from action-GIF as well as MSR-VTT, tgif [5] and vatex [9] as training part.

### 1.2. Model

For each video, we apply visual extractors including 2D models (irCSN [4], ResNext [6], CLIP [8]) and 3D models ( TimesFormer [2], X3D [3]). In order to better capture the relationships among the various video features, we learn from the idea of multi-head self-attention in the Transformer and propose a multi-head multi-dimension self-attention structure, as proposed in Fig.1, to ensemble the features and feed them into the decoder. The decoder consists of a top-down attention LSTM and a language LSTM which are almost the same with [1].

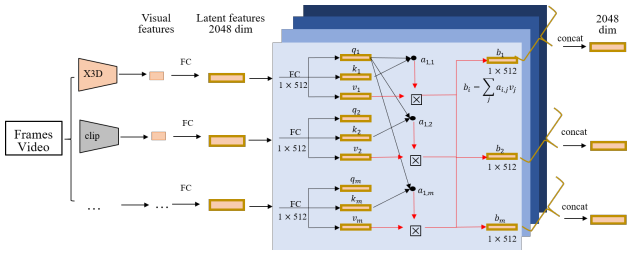


Figure 1. **Multi-head multi-dimensional self-attention module.** Visual features are enhanced through it before sent into the captioning decoder.

### 1.3. Training Details

We apply a self-critical learning method rewarding CIDer, Metoer and Bleu_4. A pretrained SentenceBert is used to measure the similarity between generated caption and the ground truth, which is utilized as another reward item. All reward items weight the same. Adam optimizer together with a decayed learning rate is utilized. Beam search size is set as 5 during decoding. We train eight models with different rewards, features and training settings in total. For each video, we choose the caption with the highest retrieval score among all its predictions using a pretrained video-text retrieval model.

## 2. Results

According to the text server, our best result on testset is proposed in Table 1.

Table 1. The performances of our model on testset of 2021 PRE-TRAINING FOR VIDEO CAPTIONING.

| Bleu_4 | METEOR | ROUGE_L | CIDEr | SPICE |
|---|---|---|---|---|
| 20.66 | 20.13 | 43.68 | 30.18 | 7.40 |

## References

[1] P. Anderson et al. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1

[2] G. Bertasius et al. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1

[3] C. Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 1

[4] D. Ghadiyaram et al. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019. 1

[5] L. Li et al. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, 2015. 1

[6] D. Mahajan et al. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 1

[7] Y. Pan et al. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training. *arXiv preprint arXiv:2007.02375*, 2020. 1

[8] A. Radford et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020, 2021.*, 2021. 1

[9] X. Wang et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 1