The Description of the Algorithm Evaluated in the Pre-training for Video Captioning Challenge

Lanxiao Wang
lanxiao.wang@std.uestc.edu.cn

Chao Shang
shangc@std.uestc.edu.cn

Heqian Qiu
hqqiu@std.uestc.edu.cn

Taijin Zhao
zhtjww@std.uestc.edu.cn

Benliu Qiu
qbenliu@std.uestc.edu.cn

Hongliang Li
hlli@uestc.edu.cn

University of Electronic Science and Technology of China
Chengdu, China

1 Full name and abbreviated name of the algorithm

Multi-stage Tag Guidance Network (MTGNet).

2 Description of the algorithm

Our method MTGNet is designed as illustrated in Figure 1. Specifically, we adopt a variety of feature extraction models to process the video (eg., I3D, Inception-V2, ResNeXt101 and Faster-RCNN). To make features more robust for complex scenes, we follow the idea of Delving Deeper into the Decoder Model[1] and apply the Tag network into the backbone and optimize it further.

Taking into account the prevention of overfitting and time and efficiency issues, the entire training process is divided into two stages of training. The first stage trains all data, and the second stage introduces a random dropout. Note that only the first stage training is performed during the GIF pre-training process. Furthermore, we used CNN-based network to pick out the best candidate results.

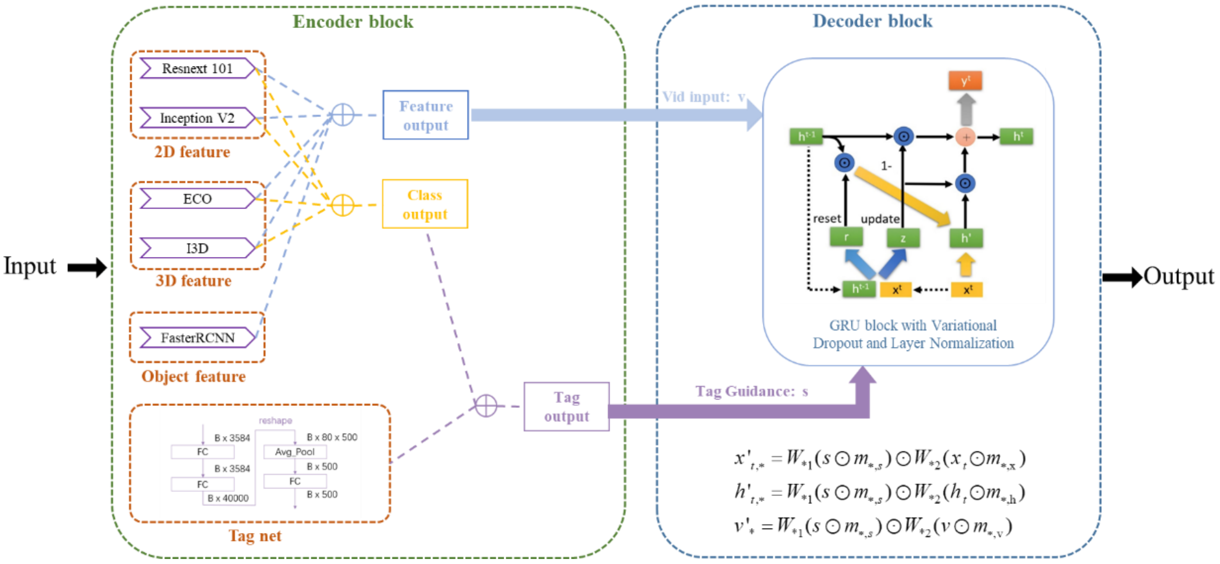

Figure 1 : The overview of Multi-stage Tag Guidance Network

3 Experimental environments

This algorithm was evaluated according to the metrics as specified in the Pre-training For Video Captioning Challenge 2020.

- Information about the training set (e.g., Auto-captions on GIF, MSRVTT, MSVD dataset. No additional training datasets were used).
- Information about pre-trained models: we use the I3D model pre-trained on Kinetics, the Inception-V2 model and the ResNeXt101 model pre-trained on ImageNet, the Faster-RCNN model pre-trained on Visual Genome and the ECO model pre-trained on Kinetics. (But not every result uses all the above features to generate)
- In our experiments, we train MTGNet in Figure 1 using multi-stage training and use GIF, MSVD and MSRVTT to train Tag-Net as Guidance. Overall, we adopt GIF to pretrain MTGNet and then fine-tune it with MSR-VTT.

Referring

[1] Chen, Haoran , J. Li , and X. Hu . "Delving Deeper into the Decoder for Video Captioning." (2020).