CLIP4Caption: CLIP for video caption

Zhanyu Wang, Mingkang Tang, Zhenhua Liu
Kandian Content Center, Tencent
{zhanyuwang,mktang,edinliu}@tencent.com

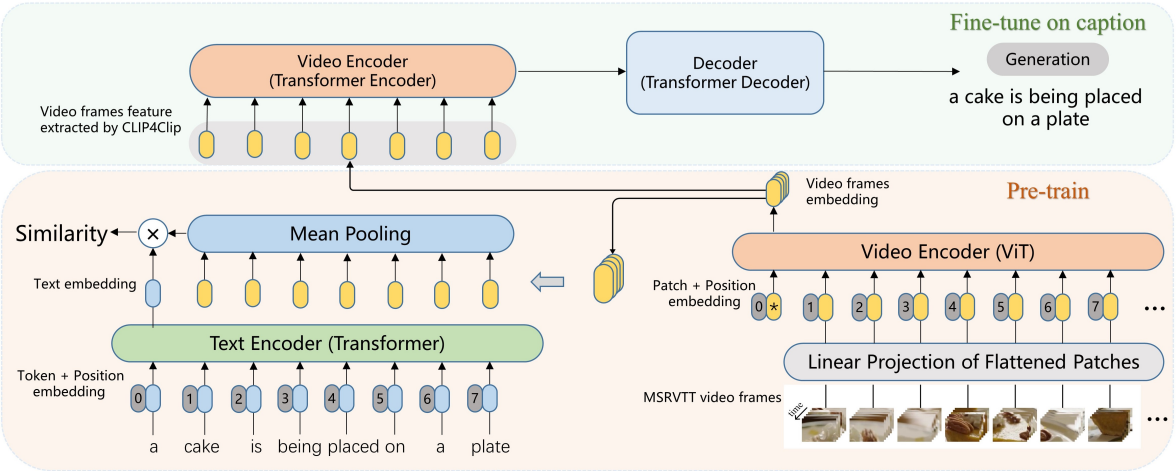

Figure 1: Overview of the proposed framework which comprises of a CLIP-based video encoder and a text transformer decoder.

1 INTRODUCTION

CLIP4Clip [2] model transferred from CLIP [3] has demonstrated its power in video-text retrieval task. In this paper, we propose CLIP4Caption, which is built upon CLIP4Clip and fine-tuned in a well-pretrained transformer decoder [1].

Our contributions are summarized as follows.

• We use CLIP4Clip as our video feature extractor to enforce our model to learn strongly text-correlated video features for text generation.

• We leverage the pre-trained weights of UniVL [1] while greatly simplified its structure to better fitting video captioning task.

• We design a novel ensemble mechanism for caption.

2 METHODOLOGY

2.1 Pre-training using CLIP4Clip

We first pre-train our CLIP4Clip model on MSR-VTT dataset, and then extract frame embedding at 1 fps using the backbone inside CLIP4Clip, resulting $n \times 512$ dynamic features $F_v = ViT(v)$ for each video, where n is the number of frames.

2.2 Finetune on video captioning

F_v is then input to a one-layer transformer Video Encoder(TE) to obtain the enhanced feature $F_{ve} = TE(F_v)$, and then fed into a three-layer Transformer Decoder (TD) to produce caption $t = TD(F_{ve})$ for each video. We initialize TE and TD with the weights pretrained in UniVL [1] and train the model with only cross-entropy loss.

2.3 Ensemble strategy

Considering the predicted captions t of one video from n different models as T, the importance score for ith caption S_i is calculated as: $S_i = metric(ref = [T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_n], hpy = T_i)$, where $i \in [1, n]$ and $metric(\cdot)$ is the captioning metric, such as BLEU, CIDEr, etc. The predicted caption with biggest score S is selected as the final output.

Table 1: Performances of CLIP4Caption

Model	B@4	M	C	S
Uni-VL on MSR-VTT	42.2	28.8	49.9	6.49
CLIP4Caption (single) on MSR-VTT	46.1	30.7	57.7	7.56
CLIP4Caption (single) on test	22.6	18.1	27.7	5.93
CILP4Caption (ensemble) on test	**23.7**	**19.6**	**31.2**	**7.53**

3 RESULT

Only the MSR-VTT dataset is used as the training and validation dataset since it is more similar to the test video. We take UniVL as the baseline model and report its captioning performance on MSR-VTT in Table 1, comparing with our CLIP4Caption model. It is very clear that our CLIP4Caption model performes better, and achieves SOTA on MSR-VTT. We also report our results on the test data with single and ensemble of CLIP4Caption models.

REFERENCES

[1] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353* (2020).

[2] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860* (2021).

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. (2021). arXiv:cs.CV/2103.00020