TJU-MM Submission to Pre-training for Video Understanding Challenge 2021

Anonymous submission
Tianjin University

Anonymous submission
Tianjin University

Anonymous submission
Tianjin University

Anonymous submission
Tianjin University

Anonymous submission
Tianjin University

1 METHOD

As shown in Fig 1, our method includes feature extraction, transformer encoder, cross-encoder, and decoder.

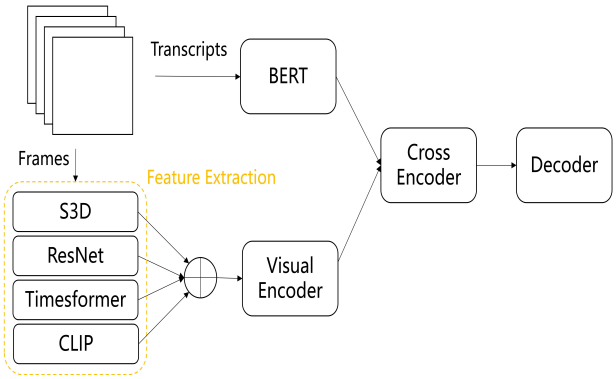

Figure 1: Illustration of our proposed framework.

Pre-processing. To obtain more informative features, we employ three types of feature extractors. Specifically, we adopt the Bert-base [2] to extract text information for all input transcripts during training. For the visual information, we use the CLIP [4] model to extract frame-level features, the S3D [7] and TimeSformer [1] are employed to extract clip-level features. The S3D and TimeSformer are pre-trained on the large-scale HowTo100M dataset [5] and the S3D is trained with the MIL-NCE loss. We sample videos at 1 *fps* and resize the spatial size of each frame to 224×224. The maximum number of sampled frames is 64, and the maximum input words is 48.

Model Details. We employ five pre-training objectives proposed in UniVL [3]: (1) video text joint, (2) conditioned masked language model, (3) conditioned masked frame model, (4) video-text alignment, and (5) language reconstruction. The transformer encoder consists of 6 attention blocks. The cross-encoder can combine visual and text features and fed them into the decoder.

2 EXPERIMENTS

2.1 Details.

We pre-train our model on the ACTION dataset [6] and fine-tune it on the MSR-VTT dataset [8]. During pre-training, we train our model with the multi-stage learning and a warm-up strategy. The initial learning rate is 1e-5.

Table 1: The performance of video caption task on MSR-VTT

	Models	B@4	M	C	S
validation dataset	UniVL[3]	41.79	28.94	50.04	
	UniVL (s3d+CLIP)	45.28	29.92	54.72	
	UniVL (s3d+CLIP+Res+ViT)	**45.98**	**29.93**	**55.93**	
test online	UniVL s3d+CLIP	21.32	17.26	23.55	5.55
	UniVL (s3d+CLIP+Res+ViT)	21.37	17.32	23.84	5.45
	model result fusion	**22.80**	**18.87**	**27.95**	**6.40**

2.2 Results.

We show the performance of our method in Table 1.The UniVL model with s3d and CLIP feature fuision are evaluated on validation dataset and test server, and it's BLUE-4 performance is 45.28 and 21.32 respectively. the BLUE-4 performance of UniVL(s3d+CLIP+Res +ViT) is 45.98 on validation dataset and 21.37 on test server.By ensembling more models like UniVL(s3d+CLIP) and UniVL(s3d+CLIP +Res+ViT), the BLUE-4 can be further boosted.

REFERENCES

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? *arXiv preprint arXiv:2102.05095* (2021).
[2] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
[3] 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).
[4] 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).
[5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*. 2630–2640.
[6] Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. 2020. Auto-captions on GIF: A Large-scale Video-sentence Dataset for Vision-language Pre-training. *arXiv preprint arXiv:2007.02375* (2020).
[7] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision*. 305–321.
[8] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5288–5296.