MILES: Visual BERT Pre-training with Injected Language Semantics for Video-text Retrieval Appendix

Yuying Ge[1] Yixiao Ge[2] Xihui Liu[4] Jinpeng Wang[5]
Jianping Wu[6] Ying Shan[2] Xiaohu Qie[3] Ping Luo[1]

[1]The University of Hong Kong [2]ARC Lab, [3]Tencent PCG [4]UC Berkeley
[5]National University of Singapore [6]Tsinghua University
yuyingge@hku.hk {yixiaoge, yingsshan, tigerqie}@tencent.com
xihui.liu@berkeley.edu jinpengwang@u.nus.edu
jianping@cernet.edu.cn pluo@cs.hku.hk

1 Visualization

1.1 Local Visual Semantics Capture

We visualize the self-attention map from the video encoder through computing the self-attention of the [CLS] token in the last block. As shown in Fig. 1, compared to the model without MVM, our pre-trained model pays high attention to those significant local regions in the video. For example, in the right column of the second row, our model is highly focused on the boat area as well as the duck in the lake while the model without MVM only takes notice of the duck, where the boat region is essential to describe the video content for retrieving this video with the given query text. Pre-training the model with MVM can capture local visual semantics and enhance the fine-grained video context understanding.

1.2 Fine-grained Video-text Alignment

We also visualize the cross-modality alignment between text and video tokens by calculating the similarity map between features embedded from the text encoder and video encoder. Fig. 2 shows that compared with the model without MVM, our pre-trained model aligns words with corresponding visual regions accurately. For example, in the right column of the first row, visual features of the area where the woman's hand holds the phone shows large similarity with the text features of the word "phone" in our model, while the visual features of irrelevant background area are highly similar to the word "phone" in the model without MVM. Performing MVM using video-text aligned features as the reconstruction targets effectively trains the model to capture video-text alignment with the "dual-encoder" architecture.

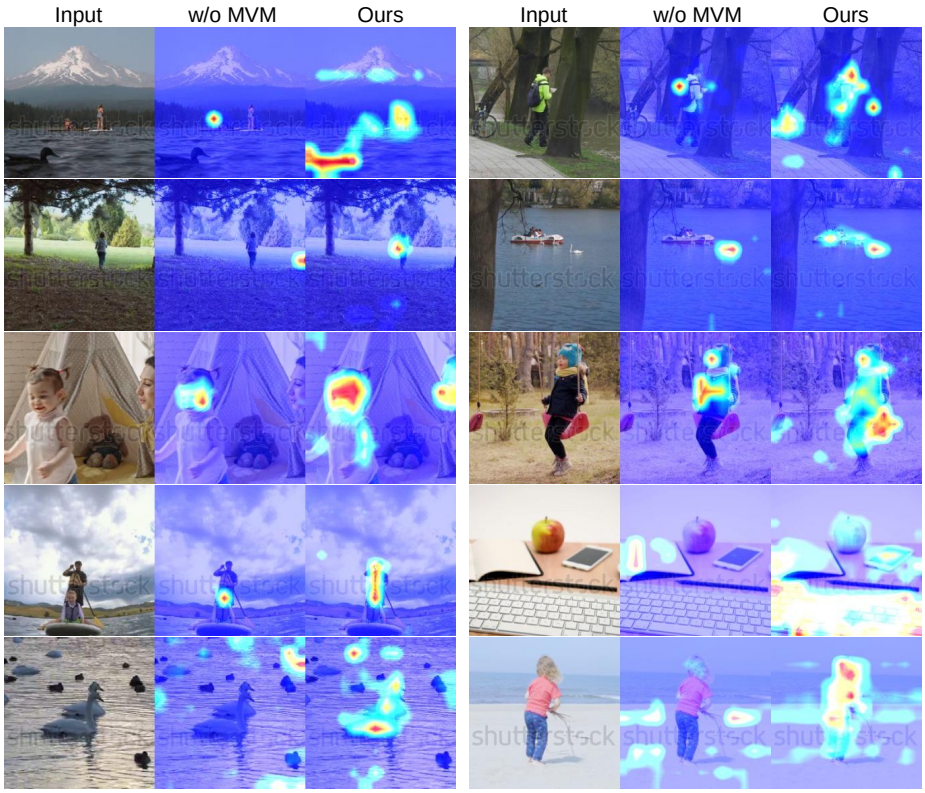

Fig. 1. The visualizations of the self-attention from the video encoder. Compared with the model that is not trained with MVM, our pre-trained model pays high attention to those significant local regions in the video (*e.g.* a duck in the left column and a bicycle in the right column of the first row), showing that MVM can promote the model to capture fine-grained visual semantics.

2 Clip-based Pre-training

Due to the dominant success of CLIP [6] in image-language representation learning, which pre-trains a model with 400 million image-text pairs, recent work [4] uses the pre-trained CLIP model as the backbone in video-language pre-training for retrieval. We also pre-train a model initialized from CLIP weights following the setting of [4] on CC3M and WebVid-2M. Specifically, we use the pre-trained CLIP (ViT-B/32) to initialize the video encoder, text encoder and the snapshot video encoder. As shown in Table. 1, our CLIP-initialized pre-trained model achieves better results for text-to-video retrieval on three datasets with both the zero-shot and fine-tune evaluation protocols. Performing MVM with the video-text aligned features as the reconstruction targets also benefits CLIP-based video-text pre-training for downstream retrieval.

Fig. 2. The visualizations of the local cross-modality alignment between the text token (the words on both sides) and video tokens. Compared with the model that is not trained with MVM, our pre-trained model aligns words with corresponding visual regions accurately (*e.g.* the region above the man's hand shows large similarity between the word "cup" in the left column of the first row), which indicates that the pretext task of MVM can effectively train the model to enhance local video-text alignment.

3 Detailed Model Architecture

Our model consists of a video encoder, a text encoder and an snapshot video encoder. We follow [2] to build the video encoder upon the structure of TimeSformer [3] with a minor modification. Specifically, TimeSformer is made up of a stack of Divided Space-Time Attention blocks, where each block first computes temporal attention by comparing each patch with all the patches at the same spatial location in all frames, and then computes spatial attention by comparing each patch with all the patches in the same frame. Different from TimeSformer, which contains a residual connection between the input of each block and the output of the temporal attention, we establish a residual connection between the input of each block and the output of the spatial attention following [2] for more stable training. The text encoder is built upon the structure of DistilBERT [7],

Table 1. Text-to-video retrieval results of models initialized from CLIP [6] weights on different datasets under zero-shot (top) and fine-tune (bottom) evaluation, where **higher** R@k and **lower** MdR (Median Rank) and MnR (Mean Rank) are better.

Method	MSR-VTT					MSVD					LSMDC				
	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CLIP-straight [5]	31.2	53.7	64.2	4.0	-	37.0	64.1	73.8	3.0	-	11.3	22.7	29.2	56.5	-
CLIP4Clip [4]	32.0	57.0	66.9	4.0	34.0	38.5	66.9	76.8	2.0	17.8	15.1	28.5	36.4	28.0	117.0
Ours	**33.1**	**59.0**	**69.9**	**3.0**	**25.4**	**46.8**	**77.0**	**85.8**	**2.0**	**8.2**	**15.5**	**30.3**	**38.8**	**23.0**	**94.8**
CLIP4Clip [4]	43.1	70.4	**80.8**	2.0	16.2	46.2	76.1	84.6	2.0	10.0	20.7	38.9	47.2	13.0	65.3
Ours	**44.3**	**71.1**	80.2	**2.0**	**14.7**	**53.6**	**81.3**	**89.9**	**1.0**	**5.8**	**22.5**	**42.9**	**50.7**	**9.5**	**57.2**

which is a multi-layer bidirectional transformer. The snapshot video encoder has exactly the same architecture of the video encoder.

4 Comparing Model Size and Complexity

We analyze the size and the complexity of the model through calculating the number of parameters and FLOPs (higher FLOPs indicate that the model requires more computation costs). As shown in Table. 2, although the snapshot encoder in our method increases the number of parameters and computational costs in pre-training to provide reconstruction targets for MVM, it is not retained for downstream retrieval, rendering an efficient "dual-encoder architecture with comparable model size and complexity while achieves higher performance.

Table 2. Comparison of model size and complexity in pre-training and downstream retrieval. "R@10" denotes the evaluation results of zero-shot text-to-video retrieval on MSR-VTT.

Method	#params (M)		FLOPs (G)		R@10
	train	inference	train	inference	
VATT [1]	414.9	327.0	1004.8	792.0	29.7
Frozen [2]	180.9	180.9	771.0	771.0	51.6
Ours	295.1	180.9	1533.4	771.0	56.9

5 Fine-tuning with MVM.

When evaluating the model for downstream retrieval with the fine-tune protocol, we simply use the contrastive objective to tune our pre-trained model and outperform the existing methods by a large margin. We further add the pretext task of masked visual modeling to fine-tune our pre-trained model with the training set of MSR-VTT and achieve better performance as shown in Table. 3. We can conclude that MVM is effective to optimize the pre-trained model towards stronger representations with the domain specific training data.

Table 3. Ablation studies on fine-tuning downstream retrieval with MVM. Results of text-to-video retrieval on MSR-VTT are reported.

MVM	R@1↑	R@5↑	R@10↑	R@50↑	MnR↓
×	37.7	63.6	73.8	89.8	24.2
√	**37.8**	**63.6**	**74.6**	**90.5**	**23.9**

References

1. Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021. 4
2. Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2021. 3, 4
3. Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv preprint arXiv:2102.05095*, 2021. 3
4. Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 4
5. Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021. 4
6. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 4
7. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3