PIDRo: Parallel Isomeric Attention with Dynamic Routing for Text-Video Retrieval Supplementary Material

Peiyan Guan[1]*, Renjing Pei[2]†, Bin Shao[2], Jianzhuang Liu[2],
Weimian Li[2], Jiaxi Gu[2], Hang Xu[2], Songcen Xu[2], Youliang Yan[2], Edmund Y. Lam[1]†
The University of Hong Kong[1], Huawei Noah's Ark Lab[2]
{pyguan, elam}@eee.hku.hk, {peirenjing, shaobin3, liu.jianzhuang, liweimian,
xusongcen,yanyouliang}@huawei.com, {imjiaxi, chromexbjxh}@gmail.com

1. More Details about the Parallel Isomeric Attention

We conduct an interaction between the spatial transformer of the S-T frame branch (S-T) and the temporal transformer of the T-S patch branch (T-S). Except for the last layer, the encoding of each layer in the spatial transformer is added to the encoding corresponding to the same layer in the temporal transformer. Given a video clip $V \in \mathbb{R}^{N_f \times H \times W \times 3}$ with N_f sampled frames and each frame being divided into N_p non-overlapping patches, its encoding at layer l of the spatial transformer of S-T is denoted as:

$$Z_l^{S-T} = [z_{l,\,1}^{S-T}, z_{l,\,2}^{S-T}, \cdots, z_{l,\,N_f}^{S-T}] \in \mathbb{R}^{(N_p+1)\times N_f \times d}, \tag{1}$$

where $z_{l,\,i}^{S-T} \in \mathbb{R}^{(N_p+1)\times d}$ is the encoding of i-th frame, which includes the encodings of a [cls] token and all patches of this frame. Similarly, its encoding at layer l of the temporal transformer of T-S is represented as:

$$Z_l^{T-S} = [z_{l,\,1}^{T-S}, z_{l,\,2}^{T-S}, \cdots, z_{l,\,N_p}^{T-S}] \in \mathbb{R}^{(N_f+1)\times N_p \times d}, \tag{2}$$

where $z_{l,\,j}^{T-S} \in \mathbb{R}^{(N_f+1)\times d}$ is the encoding of j-th patch cube, which contains the encodings of a [cls] token and all patches of this cube. We conduct the interaction between the two branches on the encodings of the patches only (excluding the [cls] tokens). For simplicity, in the following, we still use $Z_l^{S-T} \in \mathbb{R}^{N_p \times N_f \times d}$ and $Z_l^{T-S} \in \mathbb{R}^{N_f \times N_p \times d}$ to represent the encodings of the patches without the [cls] tokens at layer l of the spatial and temporal transformers, respectively. We permute the first and second dimensions of Z_l^{S-T} to make it have the same shape as Z_l^{T-S}, and then add the permuted Z_l^{S-T} and Z_l^{T-S}, which is represented as:

$$\overline{Z}_l^{T-S} = Perm(Z_l^{S-T}) + Z_l^{T-S}, \tag{3}$$

where $Perm(\cdot)$ denotes the permutation operation and $\overline{Z}_l^{T-S}$ is used as the input to layer $l+1$ of the temporal transformer. Such interaction enables us to leverage CLIP's spatial attention knowledge to enhance the temporal transformer of T-S.

*This work was done during an internship at Huawei.
†Corresponding authors: Renjing Pei, Edmund Y. Lam

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
Support Set [5]	34.7	59.9	70.0	3.0	-
Straight-CLIP [6]	59.9	85.2	90.9	**1.0**	-
TeachText-CE+ [1]	27.1	55.3	67.1	4.0	-
CLIP4Clip-meanP [4]	56.6	79.7	84.3	**1.0**	7.6
CLIP4Clip-seqTransf [4]	62.0	87.3	92.6	**1.0**	4.3
PIDRo (ours)	**62.8**	**89.3**	**94.4**	**1.0**	**4.2**

Table 1. v2t results on the MSVD dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
MMT [2]	12.3	28.6	38.9	22.5	77.1
TeacherText-CE+ [1]	17.5	36.0	45.0	14.3	-
Straight-CLIP [6]	6.8	16.4	22.1	73.0	-
CLIP4Clip-meanP [4]	20.6	39.4	47.5	13.0	56.7
CLIP4Clip-seqTransf [4]	20.8	39.0	48.6	12.0	54.2
PIDRo (ours)	**22.8**	**42.3**	**51.3**	**9.0**	**47.5**

Table 2. v2t results on the LSMDC dataset.

2. More Video-to-Text Retrieval Results

We present the video-to-text (v2t) retrieval results of different methods on the MSVD, LSMDC, ActivityNet and DiDeMo datasets in Tables 1, 2, 3 and 4, respectively. As can be seen, the proposed PIDRo exhibits obvious improvements across different datasets compared to the previous methods, which validates the robustness of our method.

3. More analysis of the dynamic routing module

We design the dynamic routing (DR) module for fine-grained information redistribution within a sentence. It allows for many network options and an MLP is employed

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [3]	17.7	46.6	-	6.0	24.4
MMT [2]	28.9	61.1	-	4.0	17.1
TeacherText-CE+ [1]	23.0	56.1	-	4.0	-
Support Set [5]	28.7	60.8	-	**2.0**	-
CLIP4Clip-seqTransf [4]	41.4	73.7	85.3	**2.0**	6.7
PIDRo (ours)	**42.2**	**74.2**	**86.0**	**2.0**	**6.3**

Table 3. v2t results on the ActivityNet dataset.

Methods	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
CE [3]	15.6	40.9	-	8.2	42.4
TeachText-CE+ [1]	21.1	47.3	61.1	6.0	-
CLIP4Clip-seqLSTM [4]	42.4	69.2	79.2	**2.0**	11.8
CLIP4Clip-meanP [4]	42.5	70.6	80.2	**2.0**	11.6
PIDRo (ours)	**47.5**	**74.7**	**83.6**	**2.0**	**8.3**

Table 4. v2t results on the DiDeMo dataset.

in our work. Here we use a 2-layer transformer as an alternative to the DR module for comparison. Each layer of this transformer has the same structure as that of the text transformer of CLIP. We remove the T-S patch branch and use "Base_model+DR " as the evaluation model. The v2t retrieval results on MSR-VTT of these two options are presented in Table 5. As can be seen, our MLP gives better retrieval result. This transformer achieves 47.0% R@1 that is lower than 47.5% R@1 of the MLP, which verifies the superiority of our method.

DR	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
Transformer	47.0	74.1	**82.9**	**2.0**	13.6
MLP(ours)	**47.5**	**74.4**	**82.9**	**2.0**	**13.3**

Table 5. Analysis of the DR structure.

References

[1] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *ICCV*, 2021. 1, 2

[2] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1, 2

[3] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 2

[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. 1, 2

[5] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 1, 2

[6] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *MCPR*, 2021. 1