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

Peiyan Guan¹; Renjing Pei²; Bin Shao²; Jianzhuang Liu²; Weimian Li²; Jiaxi Gu²; Hang Xu²; Songcen Xu²; Youliang Yan²; Edmund Y. Lam¹†
The University of Hong Kong¹, Huawei Noah’s Ark Lab²
{pyuan, 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:

\[
Z^{S-T}_l = z^{S-T}_{l,1} , z^{S-T}_{l,2} , \ldots , z^{S-T}_{l,N_f} \in \mathbb{R}^{(N_f+1) \times N_f \times d},
\]  
(1)

where \( z^{S-T}_{l,i} \in \mathbb{R}^{(N_f+1) \times d} \) 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:

\[
Z^{T-S}_l = z^{T-S}_{l,1} , z^{T-S}_{l,2} , \ldots , z^{T-S}_{l,N_p} \in \mathbb{R}^{(N_p+1) \times N_p \times d},
\]  
(2)

where \( z^{T-S}_{l,j} \in \mathbb{R}^{(N_f+1) \times d} \) 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 \( Z^{S-T}_l \in \mathbb{R}^{N_p \times N_f \times d} \) and \( Z^{T-S}_l \in \mathbb{R}^{N_f \times N_p \times d} \) 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 \( Z^{S-T}_l \) to make it have the same shape as \( Z^{T-S}_l \), and then add the permuted \( Z^{T-S}_l \) and \( Z^{T-S}_l \), which is represented as:

\[
Z^{T-S}_l = \text{Perm}(Z^{S-T}_l) + Z^{T-S}_l,
\]  
(3)

where \( \text{Perm}(\cdot) \) denotes the permutation operation and \( Z^{T-S}_l \) 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.

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


