

GLGait: A Global-Local Temporal Receptive Field Network for Gait Recognition in the Wild

Anonymous Author(s)

Submission Id: 461

A PGTA DETAILS

In this section, we detail the memory and computation complexity of Pseudo Global Temporal Self-Attention (PGTA) and Spatio-Temporal Multi-Head Self-Attention [1, 2] (MHSA).

First, we introduce the formula of MHSA. Given $\mathbf{x}_{in} \in \mathbb{R}^{L \times T \times C}$ as input, where L is $H \times W$, the formula of MHSA is as follows:

$$\text{Reshaping } \mathbf{x}_{in}, \quad \mathbb{R}^{L \times T \times C} \rightarrow \mathbb{R}^{\frac{LT}{P_t P_l} \times P_l P_t C} \quad (23)$$

$$[\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i] = \mathbf{x}_{in} \mathbf{U}_{qkv}^i, \quad \mathbf{U}_{qkv}^i \in \mathbb{R}^{P_l P_t C \times 3D} \quad (24)$$

$$\mathbf{A}_i = \text{softmax}(\mathbf{q}_i \mathbf{k}_i^T / \sqrt{D}), \quad \mathbf{A}_i \in \mathbb{R}^{\frac{LT}{P_l P_t} \times \frac{LT}{P_l P_t}} \quad (25)$$

$$\mathbf{x}_i = \mathbf{A}_i \mathbf{v}_i, \quad \mathbf{x}_i \in \mathbb{R}^{\frac{LT}{P_l P_t} \times D} \quad (26)$$

$$\mathbf{x}_g = [\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_k] \mathbf{U}_{msa}, \quad \mathbf{U}_{msa} \in \mathbb{R}^{k \cdot D \times P_l P_t C} \quad (27)$$

$$\text{Reshaping } \mathbf{x}_g, \quad \mathbb{R}^{\frac{LT}{P_l P_t} \times P_l P_t C} \rightarrow \mathbb{R}^{L \times T \times C} \quad (28)$$

where token size in MHSA is $P_l P_t \times C$. Specifically, since the tokens are separated in high-level feature maps rather than video frames, we set token size as $P \times C$ and patch size as P for simplicity.

Secondly, we explain how to calculate the memory and computation complexity in MHSA and PGTA.

For memory complexity, in Equation (6), Equation (9), Equation (24), and Equation (27), it is based on the input channels and output channels (patch size P , C , and D). In MHSA, the input channels are $P_l \times P_t \times C$ and the output channels are D , thus MHSA memory complexity is $O(P_l P_t C D)$ in Equation (15). In PGTA, we reduce the memory computation in two aspects. 1) we separate the spatial dimension from the patch, and the memory complexity is reduced from $O(P_l P_t C D)$ to $O(P_t C D)$. 2) we separate the patch size from the token, reducing the memory complexity from $O(P_t C D)$ to $O(C D)$ in Equation (17).

For computation complexity, in Equation (6), Equation (9), Equation (24), and Equation (27), it is based on feature map size and output channels, and the token size does not affect the computation complexity, which is $O(L T C D)$ of both MHSA and PGTA in Equation (14), thus we only consider the computation complexity in Equation (7), Equation (8), Equation (25), and Equation (26). In these equations, the complexity is mainly based on the token number and feature channels, even having quadratic complexity to the number of tokens. In MHSA, the token number is $\frac{LT}{P_l P_t}$, thus the computation is $O(\frac{L^2 T^2}{P_l^2 P_t^2} D)$ in Equation (16). In PGTA, we separate the spatial dimension from the patch, reducing the computation complexity from $O(\frac{L^2 T^2}{P_l^2 P_t^2} D)$ to $O(L \frac{T^2}{P_l^2} D)$ (P_l^2 is less than L in our experiments). To reduce the memory computation, we separate the patch size from the token, resulting in the computation complexity increasing P_t times, from $O(L \frac{T^2}{P_l^2} D)$ to $O(L P_t \frac{T^2}{P_l^2} D)$ in Equation (18). Even though, the

computation complexity in PGTA is also less than MHSA as shown in Table 7, performing a good memory-computation cost trade-off.

Finally, given the specific token size $P \times C$ and D in Equation (15), Equation (16), Equation (17), and Equation (18), the memory and computation complexity can be calculated.

REFERENCES

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision*. 6836–6846.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*.