Supplementary material for Glance-and-Gaze Vision Transformer

Anonymous Author(s)
Affiliation
Address
email

This document contains the supplementary material for "Glance-and-Gaze Vision Transformer". The primary goal of the supplementary material is to present more details about network architectures and ablation studies. More detailed implementations can be found in the code supplementary files.

Table 1: Configuration details of GG-Transformer. P_i, C_i, M_i, K_i, N_i, α_i refer to embedding patch size, channel number, Glance size, Gaze kernel size, block number, and MLP expansion ratio, respectively.

	Output Size	Layer Name	GG-Tiny	GG-Small
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding	$P_1 = 4; C_1 = 96$	
		GG-MSA block	$\begin{bmatrix} M_1 = 7 \\ K_1 = 9 \\ N_1 = 3 \\ \alpha_1 = 4 \end{bmatrix} \times 2$	$\begin{bmatrix} M_1 = 7 \\ K_1 = 9 \\ N_1 = 3 \\ \alpha_1 = 4 \end{bmatrix} \times 2$
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Patch Embedding	$P_2 = 2; C_2 = 192$	
		GG-MSA block	$\begin{bmatrix} M_2 = 7 \\ K_2 = 5 \\ N_2 = 6 \\ \alpha_2 = 4 \end{bmatrix} \times 2$	$\begin{bmatrix} M_2 = 7 \\ K_2 = 5 \\ N_2 = 6 \\ \alpha_2 = 4 \end{bmatrix} \times 2$
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding	$P_3 = 2; C_3 = 384$	
		GG-MSA block	$\begin{bmatrix} M_3 = 7 \\ K_3 = 3 \\ N_3 = 12 \\ \alpha_3 = 4 \end{bmatrix} \times 6$	$\begin{bmatrix} M_3 = 7 \\ K_3 = 3 \\ N_3 = 12 \\ \alpha_3 = 4 \end{bmatrix} \times 18$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Patch Embedding	$P_4 = 2; C_4 = 768$	
		GG-MSA block	$\begin{bmatrix} M_4 = 7 \\ K_4 = 3 \\ N_4 = 24 \\ \alpha_4 = 4 \end{bmatrix} \times 2$	$\begin{bmatrix} M_4 = 7 \\ K_4 = 3 \\ N_4 = 24 \\ \alpha_4 = 4 \end{bmatrix} \times 2$

A Network Architectures

The detailed network architectures for GG-Tiny and GG-Small are summarized in Table 1. Each network consists of 4 stages as most hierarchical CNNs and ViTs adopted. We follow Swin-Transformer [1] in terms of network depth and width to ensure fair comparisons.

B Attention for the Gaze Branch

A natural choice is to also adopt self-attention for implementing the Gaze branch. Therefore, we conduct experiments by using local window attention [1] as the Gaze branch. Note that, unlike

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Table 2: Comparison among different self-attentions.

	Top-1
W& SW-MSA [1]	78.50%
MSA	79.79%
Glance+Gaze (DWConv)	80.28%
Glance+Gaze (Attn)	79.07%

depthwise convolution, a self-attention variant of the Gaze branch cannot be integrated with the Glance branch into the same Transformer block while keeping the overall model size and computation cost at the same level. To ensure a fair comparison, we use two consecutive Transformer blocks where one is Glance attention and another is Gaze attention.

Results are summarized in Table 2. All results are obtained based on Swin-T trained for 100 epochs on ImageNet.

We adopt W& SW-MSA (*i.e.*, Swin-T [1]) as the baseline for all variants, which achieves 78.50% top-1 accuracy on ImageNet validation set. Although W& SW-MSA enjoys a linear complexity to the input size, it sacrifices the accuracy as a trade-off. Specifically, we replace the self-attention in all Transformer blocks of stage 3 and 4 with MSA (we also tried to replace stage 1 or 2, yet it is not trainable with out-of-memory problem), which leads to a 1.29% improvement on accuracy. This may indicate that W& SW-MSA is more efficient with linear complexity compared with MSA which has quadratic complexity, yet the performance is degraded. Notably, when adopting the proposed Glance and Gaze mechanism instead, which shares a same complexity of W& SW-MSA, can achieves much better performance, where the Glance+Gaze (Attn) improves the performance by 0.57%, and Glance+Gaze (DWConv) (*i.e.*, GG-T) by 1.78%, which is even higher than MSA by 0.49%.

Using either convolution or self-attention to implement the Gaze branch can both improve the performance compared to [1], illustrating the effectiveness of the Glance and Gaze designs. However, using self-attention is inferior to depth-wise convolution with a degrade of 1.21%, which may indicate that convolution is still a better choice when it comes to learning local relationships. Besides, using depth-wise convolution as Gaze branch can also naturally be integrated into the Transformer block with Glance attention, thus makes it more flexible in terms of network designs.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.