Face Pyramid Vision Transformer - Supplementary

Khawar Islam[1]
https://khawar-islam.github.io
Muhammad Zaigham Zaheer[2]
https://zaighamz.com
Arif Mahmood[3]
arif.mahmood@itu.edu.pk

[1] FloppyDisk.AI, Pakistan
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3] Information Technology University, Pakistan

1 Model Details

The proposed FPVT parameters are described as follows: For the i_{th} stage, p_i is the patch-size, c_i is the number of output channel, l_i is the number of layers in encoder, r_i is the reduction-ratio in F-SRA, h_n is the number of heads, e_i is the expanding-ratio of convolutional FFN.

Following the design principles of SwinT [4] and PyramidT [5], we utilize the small number of output channels in shallow stages and focus the major computational resource in the middle stages. To provide instances of FPVT, we present only one model of our method which is presented in Table. 1. The number of parameters of FPVT is smaller than ResNet-18 [2], IR-18 [1], IR-SE-18 [3].

Stages	Output Size	Layer Name	OPVT
1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding	$P_1 = 7;\ C_1 = 64$
		Transformer Encoder	$\begin{bmatrix} R_1 = 8 \\ N_1 = 1 \\ E_1 = 4 \end{bmatrix} \times 2$
2	$\frac{H}{8} \times \frac{W}{8}$	Patch Embedding	$P_2 = 3;\ C_2 = 128$
		Transformer Encoder	$\begin{bmatrix} R_2 = 4 \\ N_2 = 2 \\ E_2 = 4 \end{bmatrix} \times 2$
3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding	$P_3 = 3;\ C_3 = 256$
		Transformer Encoder	$\begin{bmatrix} R_3 = 2 \\ N_3 = 4 \\ E_3 = 4 \end{bmatrix} \times 2$
4	$\frac{H}{32} \times \frac{W}{32}$	Patch Embedding	$P_4 = 3;\ C_4 = 512$
		Transformer Encoder	$\begin{bmatrix} R_4 = 1 \\ N_4 = 8 \\ E_4 = 4 \end{bmatrix} \times 2$

Table 1: Calculated settings and the design principles follow the same rules of PVT [5]. e denotes MLP ratio, whereas, r represents resolution, and n denotes the number of heads.

2 Inference Speed

We evaluate the inference speed of our proposed FPVT architecture, in order to present its feasibility under limited computational resources on real-time applications. We compare the FPVT speed with general ViT models on LFW dataset.The proposed FPVT provides a better recognition accuracy with the inference speed of general ViTs is 0.37s per image whereas our FPVT achieves 0.32s.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2019.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2016.

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2018.

[4] 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. *International Conference on Computer Vision*, 2021.

[5] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *International Conference on Computer Vision*, 2021.