FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization

An Image is worth 16x16 words (VIT) is the revolutionary transformer paper in the era of computer vision. That was achieved state of the art performance on image classification, segmentation and detection. Previously attention was just used last layer after convolution. After that paper skip the used of convolution and used transformer to extract feature from image. But computationally it was very expensive and recent many works were proposed lower the compute and memory requirement. FastViT is one of them which proposed hybrid vision transformer architecture that obtain state of the art accuracy-latency trade-off. To do that they introduce several methods like RepMixer, train time overparameterization and large kernel convolution to boost accuracy.

In majority of the hybrid architecture token mixers were self-attention based then Metaformer introduce Pooling a simple and efficient candidate for mixing token. Pooling means average pooling from CNN which have no learnable parameter and doesn't require quadric memory like self-attention. In this paper they used RepMixer which is inherited from ConvMixer. In convMixer they used depthwise convolution layer with non-linear activation function as well as batch normalization but in RepMixer they simply ignore non-linear activation function. The main benefit of this design is it can be reparametrized at inference time to single depthwise convolution. To check the validation of RepMixer they used pooling from Metaformer-S12 with embedding dimensions [64, 128, 256,512] and see repmixer significantly lower the latency 25.1% on 384 x384 resolution and 43.9 on 1024x1024 large resolution.

To improve more efficiency, parameter count, FLOPs they replace k x k convolution with its factorized version such as you have model like (input-> KxK convolution -> output) to change this used (input-> 1xK depthwise conv -> Kx1 depthwise conv -> 1x1 pointwise conv ->output). The depthwise convolution capture spatial features within channel and pointwise convolution adjust the number of channel and introduce non-linearity. The computational cost in this layer is lower than the rest of the network.

Already discussed that self-attention token mixer is computationally expensive and an efficient approach is improving the receptive field on early stage that do not used self-attention. They also introduce depth-wise large kernel convolution[7x7] in Feed Forward Network and patch embedding layer. While a variant of depthwise large kernel size is very competitive with the variant self-attention layer in terms of modest latency, Top-1 accuracy for both variant is same but in the case of parameter large kernel have 8.8 M with the 1.4 ms latency and self-attention have 12.4 M with latency 1.5 ms on mobile. The architecture of their FFN block same as ConvNext block with a few key differences. They used batch normalization instead of layer normalization coz it can be fused preceding layer at inference also not need additional reshape operation to obtain appropriate tensor layout.

They experimented on several task like image classification, semantic segmentation & detection and 3D hand mesh estimation. In image classification they used ImageNet dataset and train their model with 300 epochs using AdamW optimizer with weight decay 0.05 peak learning rate 10^-3 and batch size 1024. The result is 3.4x faster than CMT a recent hybrid transformer model, 4.9x faster than EfficientNet and 1.9 faster than ConvneXt on the mobile device on the same result and similar latency the model achieved 4.2 better Top-1 accuracy than MobileOne. For object detection and instant segmentation on MS coco dataset using Mask-RCNN head the model attain comparable performance with CMT-S with 4.3x lower backbone latency. On semantic segmentation on ADE20k the model performs 5.2% better over poolformer-M36 with 1.5x lower backbone latency. On 3D hand mesh estimation task on FreiHand dataset the model is 1.9x faster than MobileHand and 2.8x faster than recent state of the art MobRecon when benchmark on GPU.

In addition to accuracy metrics, they study with robustness to corruption and out of distribution samples which doesn't correlate well with accuracy, and they did that. Through structure reparameterization the model has lowest memory cost with reducing run time specially on high resolution image. Having better performance than diffidence SOTA model in difference task.