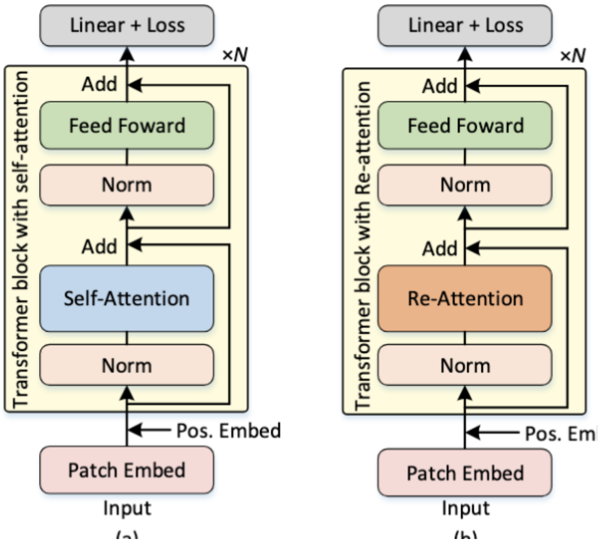
DeepViT

A Preprint

During research authors found out that main problem of training ViT models is their depth, specifically, attention maps that becomes similar on deep layers (such as 32 etc). Practice shows, that ViT with 32 layers has worse perfomance that ViT with 24 layers, so by making Re-attention authors fix this problem and make ViT improve even when depth is beyond 10 layers. Authors replace MHSA with suggested Re-attention which core idea is to regenerate attention maps by exchanging the information from different attention heads in a learnable manner.



$A \in \mathbb{R}^{H \times T \times T}$, $\quad H$ – attention and number of SA heads respectively.

$$Attention(Q; K; V) = \text{Softmax}(\frac{QK^T}{\sqrt{D}})V, \quad \text{where}\sqrt{D} \text{ is a scaling factor based on depth}$$

Specifically, improvment of ViT stops after 24 blocks!

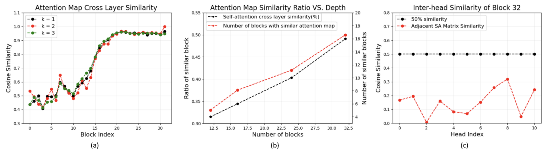


Figure 3: (a) The similarity ratio of the generated self-attention maps across different layers. The visualization is based on ViT models with 32 blocks pre-trained on ImageNet. For visualization purpose, we plot the ratio of token-wise attention vectors with similarity in Eqn. (2) larger than the average similarity within nearest k transformer blocks. As can be seen, the similarity ratio is larger than 90% for blocks after the 17th one. (b) The ratio of similar blocks to the total number of blocks increases when the depth of the ViT model increases. (c) Similarity of attention maps from different heads within the same block. The similarity between different heads within the blocks is all lower than 30% and they present sufficient diversity.

To measure similarity between attention masks authors use cosine similarity. By measuring distance on different layers of model results show that attention masks become similar as model goes deeper.

Coming to Re-attention, the core idea is that different heads of attention focus on different things. Authors combine it by adding a learnable transformation matrix which is later multiplied by attention maps. This helps to re-weight and create new connection between all the heads and get new maps that do have a diversity.

$$\text{Re-Attention}(Q,K,V) = \text{Norm}\left(\Theta^\top \left(\text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)\right) V\right)$$

During experiments hidden dimension is set to 384, image size is 224 × 224 and 3 epochs for learning rate warmup.

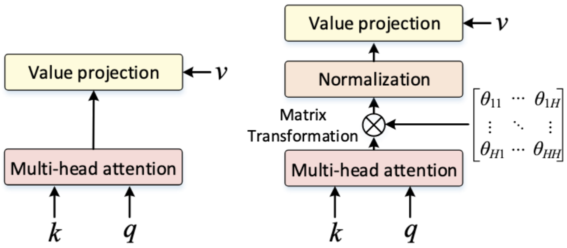

Value projection
v
Multi-head attention
k
q
Value projection
v
Normalization
Matrix Transformation
$\begin{bmatrix} \theta_{11} & \cdots & \theta_{1H} \\ \vdots & \ddots & \vdots \\ \theta_{H1} & \cdots & \theta_{HH} \end{bmatrix}$
Multi-head attention
k
q

Figure 7: (**Left**): The original self-attention mechanism; (**Right**): Our proposed re-attention mechanism. As shown, the original attention map is mixed via a learnable matrix Θ before multiplied with values.