Self Attention and ShiftViT

Paper 1: Exploring self-attention for image recognition

Paper 2: Do you even need Attention? A stack of feed-forward layers does surprisingly well on ImageNet.

Paper 3: ShiftViT: When Shift Operation Meets Vision Transformer: An Extremely Simple Alternative to Attention Mechanism

1. Describe pairwise selfattention as discussed in paper 1 (talk about each term in eq uation 2). Also, explain briefly all the forms of the relation function δ, explored in this paper.

2. Why position encoding is used in paper 1. Describe the steps for the position encoding. Also, discuss how the relative position information is calculated in this context?

3. Describe patchwise self-attention (talk about equation 4). Also, explain all the forms of the relation function δ briefly.

4. Describe the architectural changes introduced in paper 2 as compared to Vision T ransformer (ViT)model. What is the argument of the author for NOT attributing th e attention layers as the most important factor for superior performance of the Transformer-based models?

5. According to paper 3, many researchers believe that the attention mechanism in transformers has two significant strengths over convolution operation in CNN. What are those strengths?

6. How does the shift block works as described in paper 3? What is the outcome of the shifting the input features as described in paper 3?

7. Explain how shift operation approximate K x K convolution?

8. The shift block in paper 3 uses partial shift operation, describe each step of this operation for a given input tensor. How does vacant pixels are handles? How much shift step is used in paper 3?

9. How is the performance of ShiftViT affected by: (a) changing the expand ratio of MLP (b) percentage of shifted channels, and (c) shifted pixels

10. Describe briefly: (a) Rotation invariance (paper 1)
    (b) Permutation invariant (paper 1)
    (c) Cardinality invariant (paper 1)