Attention and Transformers

Paper 1: Attention is all you need

Paper 2: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

1. What is the architectural similarities and difference between encoder and decoder modules of the transformer model as described in paper 1?

2. Define Attention as per paper 1. Describe briefly the two types of attention discussed in paper 1.

3. Discuss the reasons presented by the author for applying scaling in dot-product attention.

4. State the three different ways the transformer uses multi-head attention?

5. Why does the author use a sinusoid for positional encoding. Give at least two reasons.

6. Training a transformer model for classification task on ImageNet dataset underperforms compared to the state-of-the-art ResNet models, why? Explain the reasons.

7. What are positional embeddings as discussed in paper 2? Author prefers 1D PE over 2D positional embedding for images. What is the argument of the author in paper 2 for not using hand-crafted 2D-aware embedding for images?

8. To check how the transformer model uses self-attention capabilities, compute average distance image space based on attention weights – attention distance. Explain how ’attention distance’ is analogous to receptive field size in CNNs.

9. Describe briefly. (a) Autoregressive model (paper 1), (b) Positional encoding (paper 1), (c) patch embedding (paper 2), (d) Inductive bias (paper 2).