DeiT and Swin Transformer

Paper 1: Training data-efficient image transformers (DeiT) & distillation through attention

Paper 2: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

1. As discussed in paper 2, visual elements vary substantially in scale. Explain why the variation in the scale of visual elements makes it difficult to adapt a language-based transformer model to a vision task?

2. What is the relationship between the computational complexity of self-attention in vision-transformer
(ViT) model with the image size? What is the solution proposed by swin-transformer to tackle this problem?

3. How does Swin Transformer achieves linear computational complexity for self-attention? Explain in a few sentences

4. Discuss the difference in computation complexity of a global multi-head self-attention (MSA) and a window-based MSA on an image of h x w patch size.

5. What is the patch merging layer in Swin transformer architecture? Describe how it operates?

6. How DeiT model is different from the original ViT model. List at least two differences.

7. In paper 1, many versions of DeiT is trained, explain the difference in architecture between DeiT-B, DeiT-S, and DeiT-Ti in terms of D, h, and d as discussed in paper 1.

8. Define distillation token as used in paper 1. How does the proposed distillation to ken is used for training DeiT? Explain in detail.

9. What reasons does the author provide to argue that ConvNets are better teacher models for distillation-based training of DeiT model?

10. ViT needed huge amount of data to reach at par with state-of-the-art ConvNet models, but DeiT does not need that much data. How did the authors tackle this problem? What specific methods were used to achieve this? name them.

11. Describe briefly. (a) Relative position bias (paper 2) (b) Soft distillation (paper 1) (c) Hard-label distillation (paper 1)