Computer Vision_DiscussionQuiz_08

1. It's challenging to adapt a language-based transformer model to a vision problem since visual features vary in magnitude because visual features are more difficult to adjust to size variations than word tokens, which are the essential blocks in a language model. The computational difficulty of semantic segmentation tasks conducted via image processing can also be increased by the high resolution of pictures.

2. The link between the picture size and the computational complexity of self-attention in the vision-transformer model is that ViT creates a single low-resolution feature map, as the computational cost of ViT is proportional to the image size. And in contrast to ViT's global self-attention, the swin transformer solves this problem by performing self-attention within every window locally.

3. For self-attention, Swin Transformer achieves linear computational complexity because the photos are divided into parts by these windows. There are a fixed number of patches in each division. As opposed to a quadratic rise in the case of a standard vision transformer, the running time rises as the picture size grows.

4. For an image of h x w patch size, the difference in computing complexity of a global multi-head self-attention and a window-based MSA is as follows.

   The image's complexity is 4hwC2 + 2(hw)2 C, as mentioned in paper 2. We can get the complexity of 4hwC2 + 2M2hwC by switching to a window-based technique with a patch size of h x w. Because the M2 coefficient is a constant for the patch size, it has no effect on the complexity we may deduce that the hw part's complexity is linear. The computational complexity of this window-based technique is lower. And for high quality pictures, the global multi-head self-attention has a high quadratic-based complexity, making it computationally costly.

5. Swin transformer architecture's patch merging layer are blended with nearby 2 x 2 patches. To establish a hierarchy, the same patch merging is done. This results in proportional concatenation of the network's layers.

6. Differences between DeiT and original ViT model are as follows:
   a. One main difference is that DeiT model may be generalized by merely training on ImageNet without additional information.

b. DeiT employs CNN models from the ResNets family that have previously been trained. CNNs form superior teacher networks than ViT because of their inductive bias.

7. In terms of D, h, and d, the architecture of DeiT-B, DeiT-S, and DeiT-Ti differs as for any architecture, the little d remains the same. For all of the architectures in comparison, the d is D/h, which is 64. DeiT-input B's embedding is 768 bytes long, whereas DeiT-is S's 384 bytes long and DeiT-is Ti's 192 bytes long. DeiT-B has 12 heads, DeiT-Ti has three heads, and DeiT-S has six.

8. The embedded vector utilized in the DeiT transformer model is the distillation token. The result of the instructor network is used instead of the label's usual output. The distillation token outperforms the teacher network and typical class token-based networks when used appropriately.

9. The reasons author provided to argue that ConvNets are better teacher models for distillation-based training of DeiT model are: the networks are appropriate for use in teacher models as Translational invariance and localization are examples of robust inductive biases in CNNs. In comparison to non-distillation models, this improves the network's performance. The CNN networks' benefits are used into the DeiT model's distillation-based training.

10. To compete with state-of-the-art ConvNet models, ViT required a large quantity of data, but DeiT does not. To solve this problem, the writers employed random erasing, Autoaugmen, and randaugment approaches. When compared to fresh samples in a bigger dataset, the data augmentation utilized does not yield many new features. Also proposed a knowledge distillation approach in which like other transformer networks the pre-trained ConvNet models, may also act as instructors.

11.

a. Relative position bias: It is a bias matrix created in self-attention that depicts the relative location of each token within the window in relation to the others.

b. Soft distillation: The objective function is provided by the logits of the student and instructor models, which determine the distillation output. The Kullback-Leibler divergence between the instructor and student softmax outputs is minimized using a soft distillation technique.

c. Hard-label distillation: The hard label distillation is similar to the soft label distillation, only it uses hard labels instead of soft labels to represent the result of the teacher system.

Note: https://towardsdatascience.com/swin-vision-transformers-hacking-the-human-eye-4223ba9764c3

Papers – Training data-efficient image transformers & distillation through attention

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows