Review of “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”

Riyad Bin Rafiq

1. Paper summary

Transformer has been used successfully in the field of natural language processing, but the computer vision field has its limited applications. Attention with convolution is used in some research. But this paper [1] presented a pure transformer without any reliance on CNN and this transformer was called Vision Transformer (ViT). This newly proposed transformer obtained best results based on different datasets compared to state-of-the-art convolutional networks.

2. Contribution

2.1 Not relied on convolution

This is not the first paper where self-attention has been used in the image recognition task. But this paper first introduced a transformer for image recognition tasks without relying on the convolutional layers. In my opinion, this is one of the significant contributions of the paper.

2.2 Different evaluation

The paper evaluated the learning capabilities of ResNet, Vision Transformer and the hybrid. Moreover, different datasets such as ImageNet, JFT were used for pretraining and CIFAR-10/100, Oxford-IIIT Pets were used for transferring the model to learn various tasks. The results from these evaluations provide the robustness of the VIsion Transformer.

2.3 ViT Inspection

The authors provided internal representations of ViT by visualizations. These visualizations help to understand how the layer, encoding and attention weights have an impact on image classification. Moreover, the visualizations help to realize the components of ViT.

3. Critique

3.1 Large dataset is needed

ViT needs a large dataset to show satisfactory performance. Otherwise, it shows a few percentage points below ResNet of comparable size. But it’s not always possible to manage such resources for training the network. In my opinion, this is one of the biggest disadvantages of ViT.

3.2 Few-shot learning?

The authors reported the results based on fine-tuning or few-shots. They explained the methods of fine-tuning whereas they didn’t mention the procedure of few-shots in detail. As few-shots were a part of the research, this topic should be clear.

Reference

1. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv [cs.CV]. 2020. Available: http://arxiv.org/abs/2010.11929