FROM CNNS TO TRANSFORMERS: RESNET AND VISION TRANSFORMER

Prepared by: George Emil

Supervised by: Dr. Amjad Bakry

Technical Report

From CNNs to Transformers: ResNet and Vision Transformer

Paper 1: The CNN Peak

Title: Deep Residual Learning for Image Recognition (ResNet)

Authors: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (Microsoft Research)

Conference: CVPR 2016 — Best Paper Award

1. Motivation: The Degradation Problem

Before 2015, most people believed that increasing the depth of Convolutional Neural Networks (CNNs) would automatically lead to better feature representations and higher classification accuracies. However, experiments showed that deeper networks often hit a performance plateau and eventually got worse. Moreover, it was found that both training and validation errors went up, so the problem was not due to overfitting.

This issue referred to as the Degradation Problem underscores the basic difficulties the networks have with the optimization of deeper architectures. Specifically, ordinary deep networks have trouble learning identity mappings, which is why deeper models are much more difficult to optimize than shallower ones.

2. Architectural Innovation: Residual Learning

To solve the problem of degradation, the authors introduced residual learning by means of residual blocks. Rather than directly learning a desired mapping H(x), the network is changed to learn a residual function F(x) such that:

$H(x) = F(x) + x$

Here, x stands for the input feature map, F(x) is the residual mapping learned by the stacked layers, and the identity shortcut connection makes it possible for x to be added straight to the output.

By facilitating the network's learning of functions that are near identity mappings, this formulation streamlines optimization. Additionally, the vanishing gradient issue is successfully mitigated by the shortcut connections, which allow gradient flow to continue uninterrupted during backpropagation.

3. Network Variants: Bottleneck Blocks vs. Basic

Depending on the network depth, the ResNet architecture introduces two different kinds of residual blocks:

- Basic Block (ResNet-18/ResNet-34): Comprised of two consecutive 3x3 convolutional layers.
- Bottleneck block (ResNet-50, ResNet-101, and ResNet-152) consists of a final 1x1 convolution for dimensionality restoration, a 3x3 convolution for spatial processing, and a 1x1 convolution for dimensionality reduction.

Very deep networks are made possible by the bottleneck design, which also lowers computational costs.

ResNet-152 won the ILSVRC 2015 competition with a top-5 classification error of 3.57% on the ImageNet benchmark. On the CIFAR-10 dataset, the authors also successfully trained networks with depths greater than 1,000 layers, proving that optimization was no longer the main bottleneck in deep CNN training.

Paper 2: The Transformer Shift

Title: An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale (ViT)

Authors: Alexey Dosovitskiy et al. (Google Brain)

Conference: ICLR 2021

1. Motivation: Moving Beyond Inductive Bias

Convolutional Neural Networks are largely based on inductive biases such as locality and translation equivariance that have been a major reason for their success in vision tasks. Nonetheless, these assumptions can be considered as model constraints.

The Vision Transformer (ViT) goes against this approach by completely getting rid of the convolutional inductive biases and depending on global self, attention alone. The authors believe that if there are large enough datasets, Transformers can learn spatial relationships just from data.

2. Methodology: Image Patching

Transformers have no ability to process raw image grids directly because of the quadratic complexity of self-attention. ViT splits the input image into fixed-size patches, usually 16×16 pixels, to get around this restriction. To create a sequence appropriate for Transformer processing, each patch is flattened and handled as a token.

This method preserves global contextual information while lowering computational complexity.

3. Architecture Details

The Vision Transformer structure is almost the same as a BERT, style Transformer encoder. Linearly projecting the flattened patch embeddings into the latent space and adding learnable positional embeddings to them for retaining the spatial information are the two steps involved. A classification token ([CLS]) is added at the beginning of the sequence and the final representation of the token is used for image, level classification..

4. Key Results and the Role of Data Scale

Experimental results show that ViT underperforms CNN, based models on mid, sized datasets such as ImageNet-1k due to insufficient inductive bias. However, when pre, trained with large, scale datasets like ImageNet-21k or JFT, 300M, ViT can outperform the best CNNs.

Besides that, ViT is able to take advantage of the highly optimized matrix multiplications on modern GPUs and TPUs which often lead to better computational efficiency when compared with the convolution, based architectures.

References

1. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

2. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in International Conference on Learning Representations (ICLR), 2021.