James Thurlow
Biplav Srivastava
CSCE 580/581
Dec 05 2023

Research Paper Summary: An Image is Worth 16x16 Words

Summary

The paper “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” introduces a new method for image processing: the Vision Transformer (ViT). Typically, transformers are used for natural language processing, which they are quite successful at. The authors demonstrate transformer’s abilities to process images similarly to how they process natural language. They show that under certain circumstances, transformers can process images more efficiently and with less bias than convolutional neural networks (CNN’s), which are typically used for image recognition.

Key Contributions

The biggest finding of this paper is that conventional image processing done by CNNs can be replaced by pure transformers when applied directly to sequences of image patches.

Similarly to how words are processed in natural language models by being vectorized and fed into a transformer encoder, images are split into fixed size “patches” in a process known as image to sequence conversion. The performance of ViT directly correlates to the size of the training dataset. On smaller datasets such as ImageNet, ViT has reasonably accurate performance, but when a large dataset is used such as JFT-300M the performance increases dramatically.

The study found that the inductive bias found in CNN’s can be mitigated by a ViT with a sufficiently large dataset. The paper states that ViT’s have less inductive bias because Cnn’s have inherent biases like locality and translation variance, which ViT learns from data, and can be mitigated through a large training dataset. Because of these lack of biases, ViT’s show great potential for self-supervised learning.

In terms of performance, ViT’s have the same or better efficiency as CNNs if they are pre-trained on large data sets. ViT also has higher accuracy than traditional CNNs when utilized for smaller tasks. The findings of the paper found this to be true across a variety of configurations, scenarios, and training data.

Critique

I agree with Sankalp to a large degree that the positives of this paper are it's novel approach to image recognition. The potential efficiency savings of ViT's over CNNs is a step forward for the field of image recognition. Unlike the "Exploring The Ultimate Brain" paper which I also reviewed, this paper is not a demonstration of a established model's capability, but rather a new proposal. The paper is very data driven and very objective with it's findings, using a wide range of training data and applications to test the feasibility of transformers for image processing. The potential for self supervised learning is another great innovation.

The downsides to ViT is it's lack of biases, which means a dependency on large datasets to pick up on patterns. The paper also focuses solely on image classification and ignores other image processing fields such as object detection.

Overall, I thought it was a fascinating paper.