Title: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Context and objective:

The paper covers the possibility to apply Transformers to image recognition tasks where convolutional architectures are still dominating. Efforts were made to combine (and replace in some cases) CNN-like architectures with self-attention, resulting in some theoretical efficiency without being scaled effectively on modern hardware accelerators due to the use of specialized attention patterns. Inspired by the success of scaling Transformers in NLP, the experiment involves applying a standard Transformer directly to images with minimal modifications.

Keywords:

Transformers, CNN, Patches, Self-Attention, Position Embeddings.

Challenges:

- Direct application of self-attention mechanisms to images results in quadratic costs in terms of pixel interactions, making it impractical for realistic input sizes.
- Some approaches require intricate engineering for efficient hardware acceleration.

Vocable:

K is the number of downstream classes.
D is the dimension of the latent representation of the flattened 2D patches

Contribution:

The contribution of this paper is to apply a standard Transformer with minimal modifications to present an alternative to the convolutional neural network, turning an image processing problem into a token prediction problem.
The experimented model can be divided into three separate parts:
The image embeddings: it starts by reshaping the image into a sequence of flattened 2-d patches, these patches are then mapped into a latent space of D dimensions by applying a trainable linear projection.
Standard learnable 1-d position embedding methods are used to add position embeddings to the patch embeddings in order to retain positional information.
The transformer encoder: The Transformer encoder consists of alternating layers of multi headed self-attention and Multi-Layer Perceptron (MLP) blocks. Layer normalization is applied before every block, and residual connections are used after every block.
The classification head: a classification head is implemented by a MLP with one hidden layer during pre-training and a single linear layer during fine-tuning.

Another alternative architecture to raw image patches has been proposed, it is a hybrid approach where patches are extracted from the feature maps of a CNN to form the input sequence. These patches are then fed to the patch embedding method.

For fine-tuning, The pre-trained prediction head is removed and a zero-initialized $D \times K$ feedforward layer is attached. K is the number of classes.

For high resolution images, it is preferable to fine-tune than pre-train. An advantage for this approach is that the patch size remains the same, resulting in a larger effective sequence length.
On the other hand, pre-trained position embeddings may lose their significance with a change in resolution. This is where 2D interpolation of pre-trained position embeddings comes into play during fine-tuning to to adapt to the new resolution. This adjustment ensures that the position embeddings correspond accurately to their locations in the original image.

Experiments and Results:

When trained on mid-sized datasets like ImageNet without strong regularization, the models achieve modest accuracies, a few percentage points below comparable ResNets.
Performance improves significantly when models are trained on larger datasets (14M-300M images).
ViT attains excellent results when pre-trained at sufficient scale and transferred to tasks with fewer data points.
Pre-training on ImageNet-21k or JFT-300M datasets leads to ViT approaching or beating state-of-the-art on multiple image recognition benchmarks.
Best model achieves 88.55% accuracy on ImageNet, 90.72% on ImageNet-ReaL, 94.55% on CIFAR-100, and 77.63% on the VTAB suite of 19 tasks.

Conclusion:

The novelty of this paper lies in presenting a distinctly different approach to large-scale image recognition compared to contemporary state-of-the-art practices. It achieves this by introducing a straightforward yet effective vision transformer that adapts the original Transformer with minimal modifications: patch embedding, a class token, and a class head.
Despite the effectiveness and the interesting empirical results, it is important to note that this paper does not contribute technical innovations, as it essentially refines existing state-of-the-art Transformers.
While the presented vision transformer is promising, there are complexities still to be worked out. It may face challenges in tasks requiring pixel-level predictions, such as segmentation tasks or 3D vision tasks, as this paper focuses on image recognition tasks.
ViTs have shown promise in few-shot learning scenarios, where models are trained to perform well on tasks with very limited labeled data.

What's interesting about this paper:

The main interesting point in this work is its potential to apply NLP techniques to computer vision tasks. Given the recent advancements in NLP over the past two years, there exists a parallel opportunity for progress in computer vision. This supplementary advancement can complement domain-specific techniques, as exemplified by Meta's success with its Segment Anything Models.
The simplicity of the adaptability of Transformers inspires researchers, and most interestingly practical engineers to explore different fields when facing challenges looking for adequate methods to their specific projects.