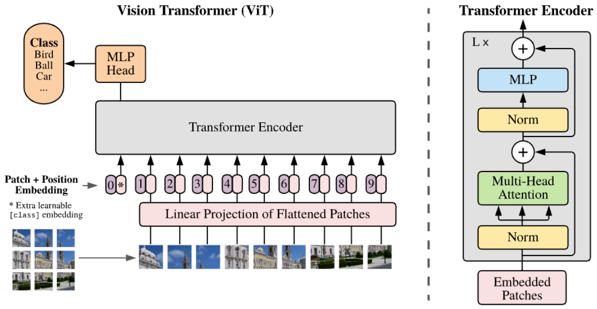
Tranformers For Image Recogniniton At Scale

The image splitted into pathces and a series of linear embeddings are generated from them, which in turn is given to transformer as its input. Patches are treated as tokens.

Transformers lack the inductive biases inherent to CNNs such as translation equivariance and locality as a result they do not generalize well without sufficient data. Yet, large scale data compansates this lack of inductive bias.

It is computationally infeasible for attention at the pixel granularity, as the cost is quadratic in ht enumber of pixels.



1. Architecture

The standart Tranformer recieves as input a 1D sequence of token embeddings. To handle 2D images, we reshape the image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $\boldsymbol{x}_p \in \mathbb{R}^{\mathbb{N}\times(P^2\cdot C)}$ where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = \frac{HW}{P^2}$ is the resulting number of patches, which also swerves as the effective input sequence length for the transformer. A constant latent vector size of D is used across the layers of the transformer. So the flattened patches are linearly mapped to this dimensionality before inputting.

A learnable embedding is preprended to the sequence of embeddings obtained from the patches. This embedding's last representation at the output of the encoder is used as the representation of the entire image. This embedding is then given to a classification head (a MLP). During pretraining this MLP has a single hidden layeri during fine-tuning its simply a linear layer.

A standart learnable 1D positional encoding is used, turns out using 2D-aware encoding doesn't change the preformance.

For CNNs locality, two-dimensional neighborhood structure as well as translation equivariance are foundational architectural biases, inherently built into the design of convolutional layers. In ViT, two-dimensonal neighborhood structure is completely lost except for the fact that patches are formed in squares and in fine-tuning the positions embeddings are adjusted accordingly. MLP provides some locality and translation equivariance.

It is also possible to use feature maps of CNNs as tokens after patching their outputs.

Typically ViT is pre-trained on large datasets and fine-tuned on downstream tasks. For fine-tuning the prediction head is removed and a linear layed added which maps from the transformer diension of D into the task dimension K. When the resolution changes pretrained position embeddings may no longer be useful. To adjust for the resolution 2D interpolation of the pretrained embeddings is done according to their locations in the original image.