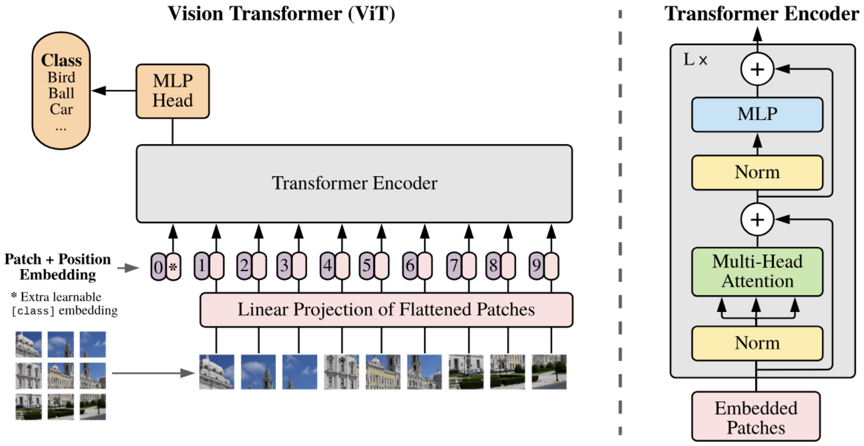
ViT

A Preprint



Authors suggest next:

- Image is split into fixed size patches
- Lineary embed each of them
- Add position embeddings
- feed the resulting sequence of vectors to a standard Transformer encoder

It is important to note that for classification task authors use the standard approach of adding an extra learnable "classification token" to the sequence.

As the transformers handle input as 1D sequence of token embeddings, images are reshaped into a sequence of 2D flattened patches $\quad x^{H\times W\times C} \to x^{N\times(P^2\cdot C)}, \quad N = \dfrac{HW}{P^2}$

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. N: number of patches

The Transformer uses constant latent vector size D through all of its layers, so authors flatten the patches and map to D dimensions with a trainable linear projection:

$$\mathbf{x}_E = \mathbf{x}_{\text{patch\_flat}} \cdot W_{\text{proj}} + \mathbf{b}_{\text{proj}}$$

Later it is fed as:

$$z_0 = [\mathbf{x}_{\text{class}},\ \mathbf{x}_1^{pE},\ \mathbf{x}_2^{pE},\ \dots,\ \mathbf{x}_N^{pE}] + E_{\text{pos}}, \quad E \in \mathbb{R}^{(P^2\cdot C)\times D}, \quad E_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$$

where x_i^E are the patch embeddings obtained from the linear projection

Positional embeddings are also added to retain positional information: standard learnable 1D position embeddings

The classification head is implemented by a MLP with one hidden layer at pre-training time and by a single linear layer at fine-tuning time.

The MLP contains two layers with a GELU non-linearity.

$$z_0 = [\mathbf{x}_{\text{class}},\ \mathbf{x}_1^{pE},\ \mathbf{x}_2^{pE},\ \ldots,\ \mathbf{x}_N^{pE}] + E_{\text{pos}}, \tag{1}$$

$$E \in \mathbb{R}^{(P^2 \cdot C) \times D}, \quad E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$z'^{(\ell)} = \text{MSA}(\text{LN}(z^{(\ell-1)})) + z^{(\ell-1)}, \quad \ell = 1, \ldots, L \tag{2}$$

$$z^{(\ell)} = \text{MLP}(\text{LN}(z'^{(\ell)})) + z'^{(\ell)}, \quad \ell = 1, \ldots, L \tag{3}$$

$$\mathbf{y} = \text{LN}(z_0^{(L)}) \tag{2}$$

This difference in inductive bias explains why ViTs require large datasets for training. Without strong built-in assumptions about images, ViTs need to learn these properties from a vast amount of data to perform well.

Due to this fact the amount of data is very importnant for ViT model