ViT

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021)
16k

Transformer: $O(n^2)$ computational complexity too expensive for all pixels

Previously: CNN + attention

Our work: a pure transformer applied directly to sequences of image patches

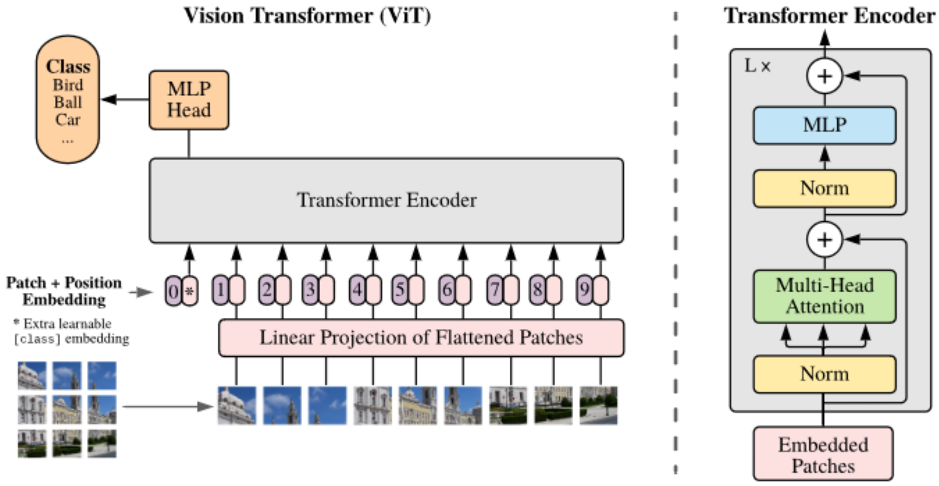

Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

image → patch embedding (linear projection of flattened patches) + position embedding + extra learnable class embedding (cls token for classification output) → transformer encoder → MLP head → class

Image: 224 * 224

N = 224^2 / 16^2 = 196

Patch: D = 16 * 16 * 3=768

Sequence (k, q, v): 197 * 768 (196 images + 1 cls token)

Transformers lack inductive biases inherent to CNNs, such as translation equivariance and locality.

Perform worse on small datasets. SOTA on larger datasets.