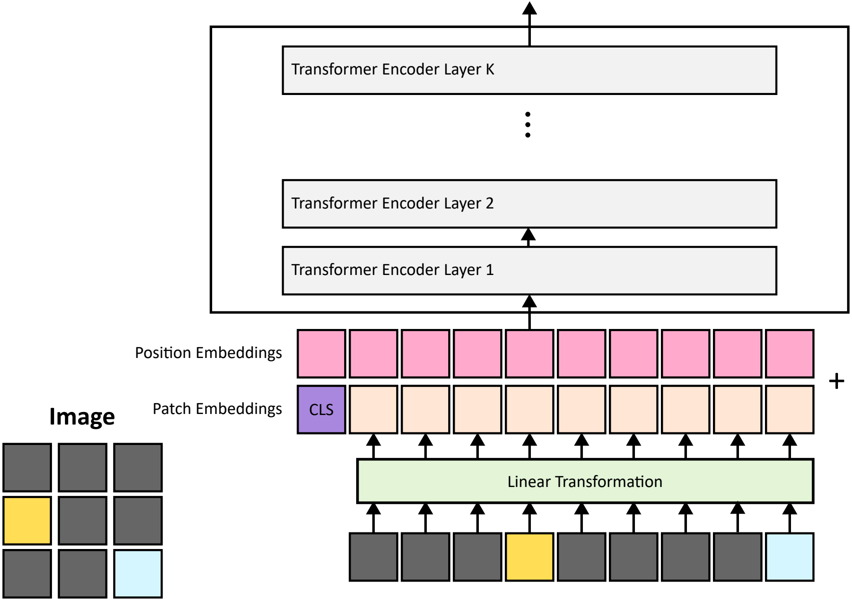
Vision Transformer (ViT)

Dr. Alireza Aghamohammadi

Vision Transformer Architecture


Transformer Encoder Layer K
Transformer Encoder Layer 2
Transformer Encoder Layer 1
Position Embeddings
Patch Embeddings
CLS
+
Image
Linear Transformation

Vision Transformer Concept

- CNNs are not essential; a pure transformer can be directly applied to sequences of image patches for effective image classification.
- Divide an image into patches and apply a linear transformation to these patches. Add positional embeddings and use this as the input to the transformer.
- Treat image patches similarly to tokens (words) in NLP applications.
- The transformer's input consists of a sequence of vectors.
- For classification, add an extra learnable CLS vector to the sequence.
- Given an image $X \in \mathbb{R}^{H \times W \times C}$, where H is height, W is width, and C is the number of channels, each patch $X_p^i \in \mathbb{R}^{P^2 \cdot C}$, where P is the patch size.
- The linear transformation is $X_p^i W = Z_i \in \mathbb{R}^D$, where $W \in \mathbb{R}^{P^2 \cdot C \times D}$ and D is the dimension of the transformed patch.