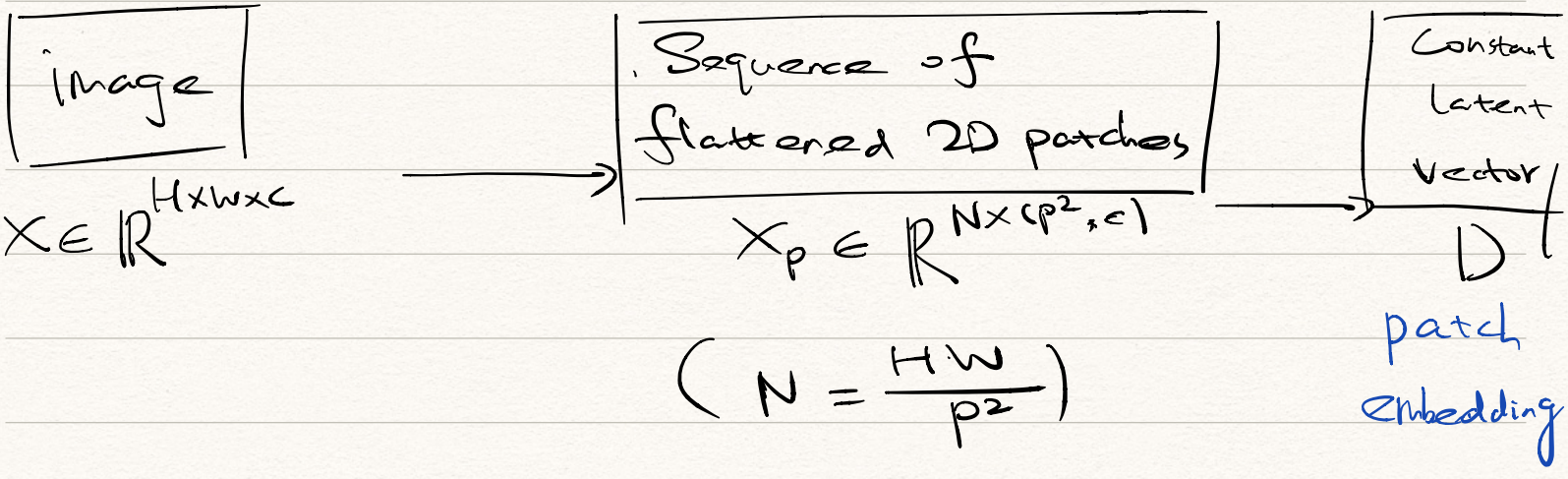
Transformer

Pros:
- computational efficiency
- scalability
- can handle arbitrary sequence lengths

Cons: lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data.

Vision Transformer:



Constant
Latent vector
D

1D
Position Embeddings

} input → Transformers

MSA: Multiheaded Self-attention

MLP: GELU non-linearity

$$z_0 = [\, x_{class} ; x_p^1 E ; x_p^2 E ; x_p^2 E ; \cdots ; x_p^N E \,] + E_{pos}$$

$$z'_l = MSA(\, LN(\, z_{l-1} \,)) + z_{l-1}$$

$$z_l = MLP(\, LN(\, z'_l \,)) + z'_l$$

$$y = LN(\, z_L^0 \,)$$

$$E \in \mathbb{R}^{(P^2 \times C) \times D} \; ; \; E_{pos} \in \mathbb{R}^{(N+1) \times D} \; ; \; l = 1, \ldots, L$$

Attach a zero-initialized $D \times k$ feed forward layer

Fine-tuning Accuracies: capture the performance of each model after fine-tuning it on the respective dataset.

Few-shot accuracies: obtained by solving a regularized linear regression problem that maps the (frozen) representation of a subset of training images to $\{-1, 1\}^k$ target vectors.

(for fast on-the-fly evaluation where fine-tuning would be too costly)