Machine Learning for Robotics: **Vision Transformer**

Prof. Navid Dadkhah Tehrani

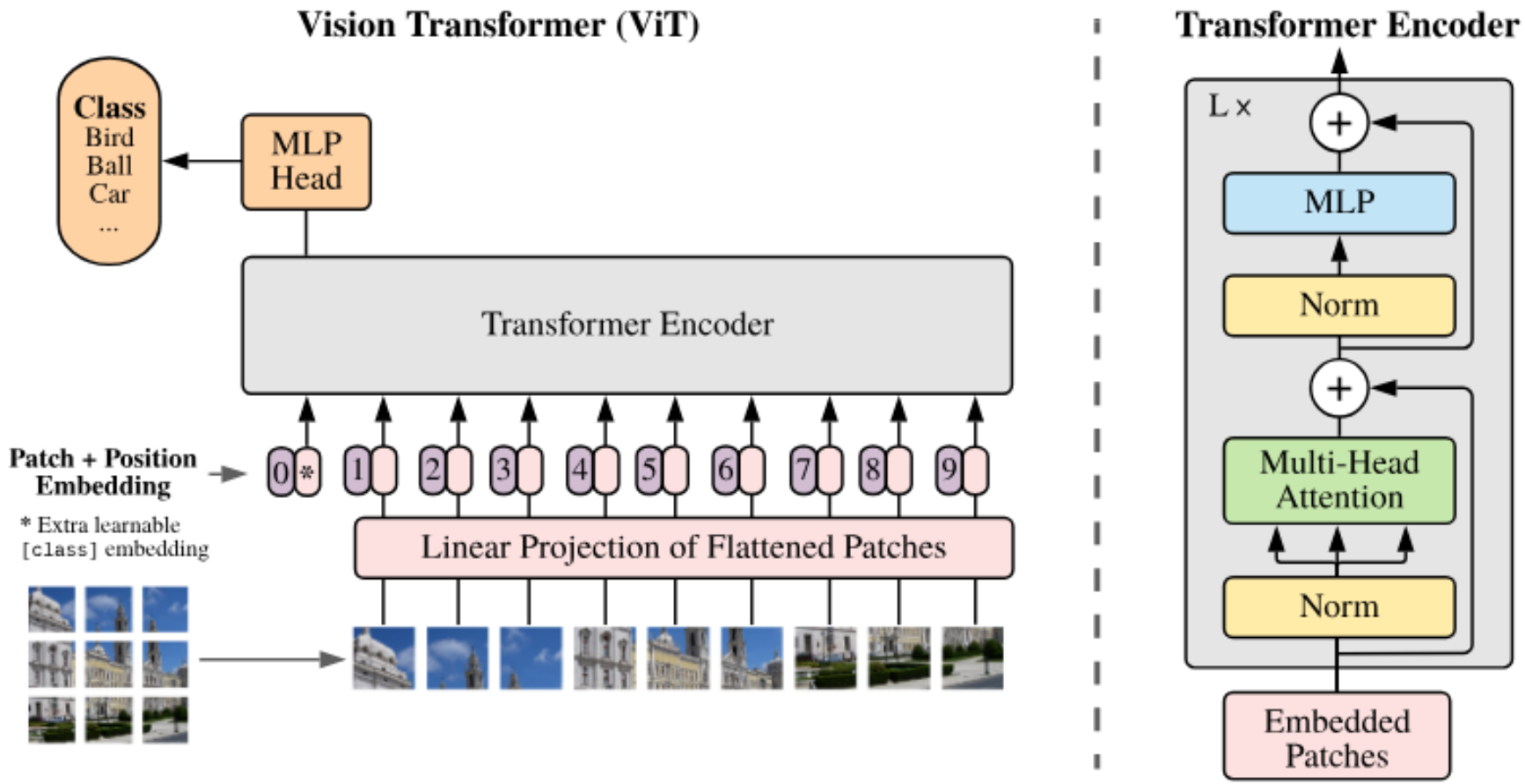
Vision Transformer (ViT) from Google brain team was introduced in [1]. It was the first successful attempt to use transformers, which was widely use for language processing, in computer vision.

ViT is a CNN-free architecture, and they showed that it has comparable performance with the state-of-art CNN-based models.

There are many transformer-based architecture that are inspired by this paper and achieved state of the art performance.

As we mentioned earlier, many real-world robotics architectures has CNN, MLP, RNN, Transformers all combined.

[1] Alexey Dosovitskiy, et. al. , An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D},\ \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0)$$

$x \in R^{H \times W \times C} \rightarrow x_p \in R^{N \times (P^2 C)}$

N: number of patches $= \frac{HW}{P^2}$

(P, P): resolution of each image patch

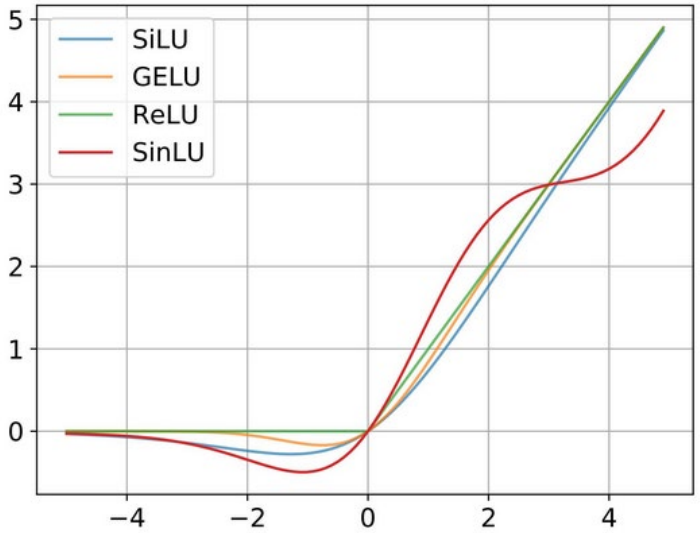
C: number of color channels

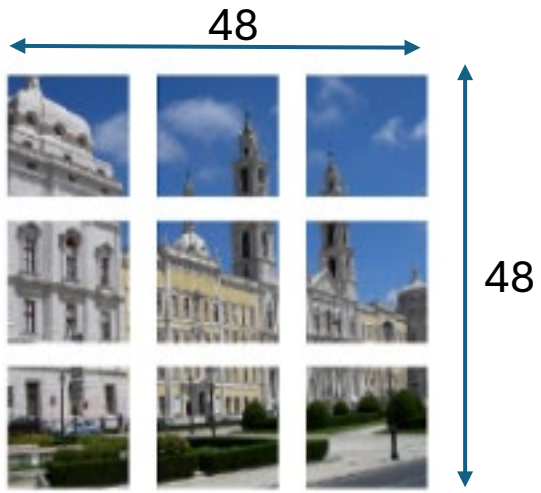
D: path embedding dimension

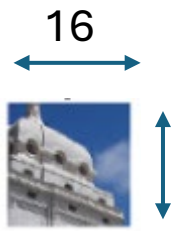
MSA: multi-head self attention

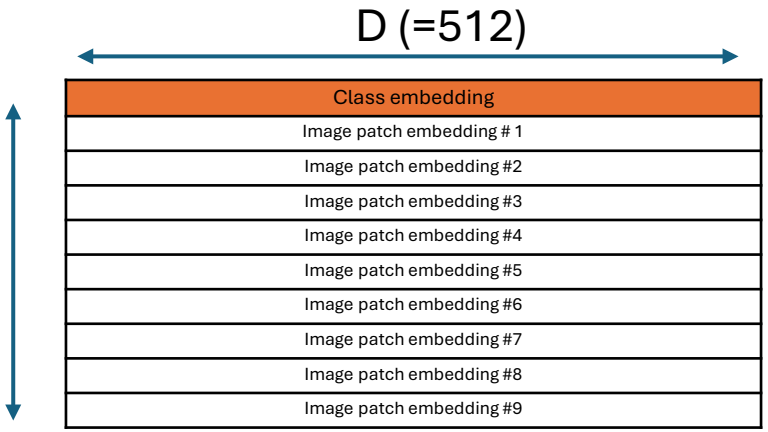
LN: layer norm

The MLP has GELU non-linearity









When trained on mid-sized datasets such as ImageNet without strong regularization, these models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome may be expected: Transformers 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.

However, the picture changes if the models are trained on larger datasets (14M-300M images). We find that large scale training trumps inductive bias. Our Vision Transformer (ViT) attains excellent results when pre-trained at sufficient scale and transferred to tasks with fewer datapoints. When pre-trained on the public ImageNet-21k dataset or the in-house JFT-300M dataset, ViT approaches or beats state of the art on multiple image recognition benchmarks. In particular, the best model reaches the accuracy of 88.55% on ImageNet, 90.72% on ImageNet-ReaL, 94.55% on CIFAR-100, and 77.63% on the VTAB suite of 19 tasks.

Lack of Inductive biases: CNNs have two specific inductive biases:

- Translation equivariance: This means that CNNs are inherently good at recognizing objects regardless of their position in the image. For example, if a dog is in the top left corner or the center of the image, CNNs are designed to detect it similarly.
- Locality: CNNs process local regions of an image (e.g., small patches) and gradually build up to larger contexts. This helps them capture local spatial relationships effectively.

Transformers, on the other hand, were originally designed for translation and do not have these inherent biases.
They process global relationships in the data without assuming any local structure,
which makes them flexible but less effective at generalizing from smaller datasets without specific adaptations.