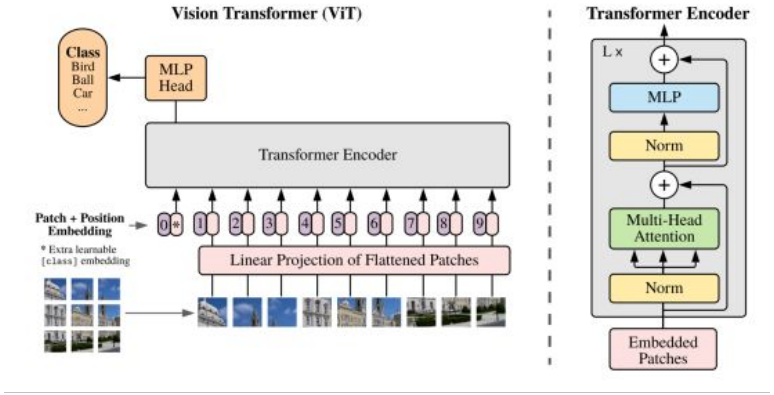
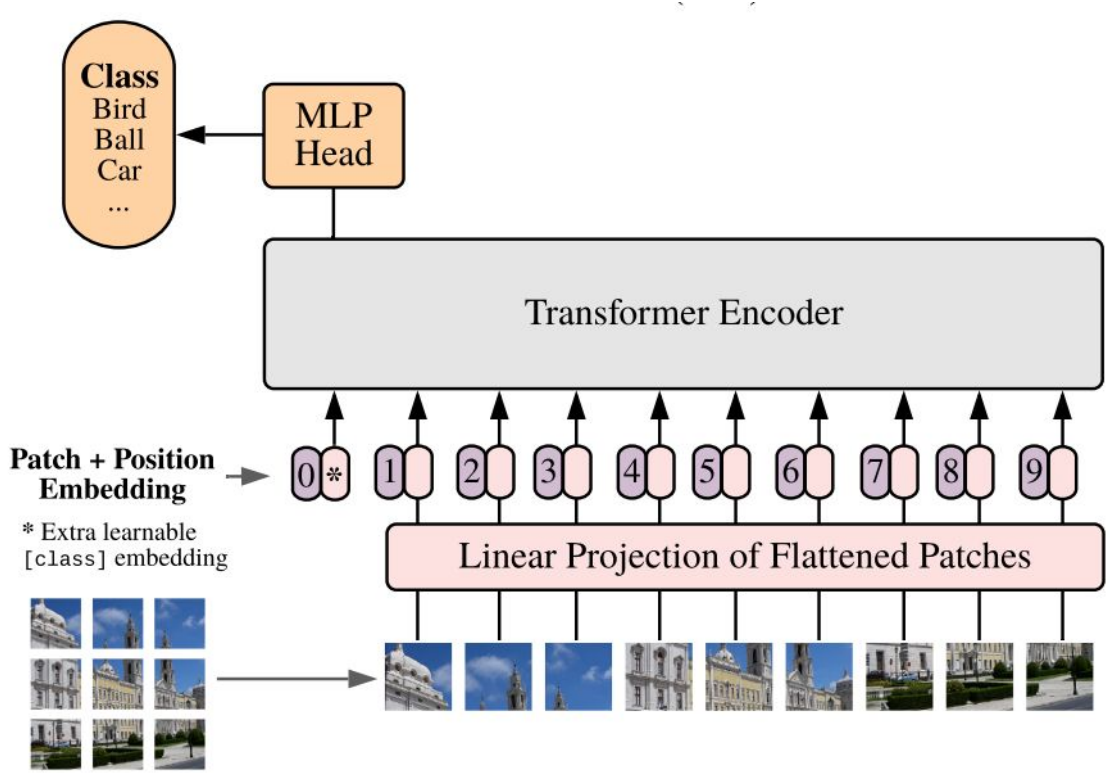
VISION TRANSFORMERS



IDIOT DEVELOPER

What is Vision Transformer?



Input Image Processing

Input Image

Size: 512 x 512 x 3

Transform

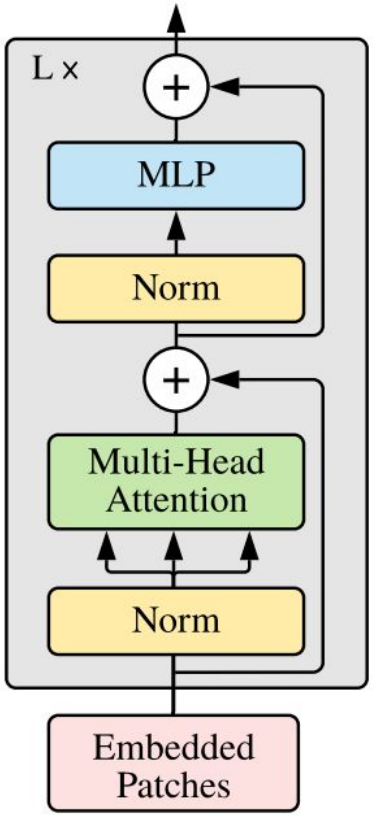
Patches

Size: 32 x 32 x 3
Num Patches: 256

Input Image Processing

THEORY	EXAMPLE
Input Image: H x W x C. **Patch Size:** Ph x Pw **Number of patches (N):** (H x W) / (Ph x Pw) **Transformed Input:** (N, Ph x Pw x C) H x W = Image height x width C = Image channels Ph x Pw = Patch height x width N = Number of patches	**Input Image:** 200 x 200 x 3 **Patch Size:** 25 x 25 **Number of patches (N):** = (200 x 200) / (25 x 25) = 64 **Transformed Input:** = (64, 25 x 25 x 3) = (256, 1875)

Transformer Encoder



Norm: Layer Normalization.

MLP: Uses GELU activation function.

ViT Variants

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M