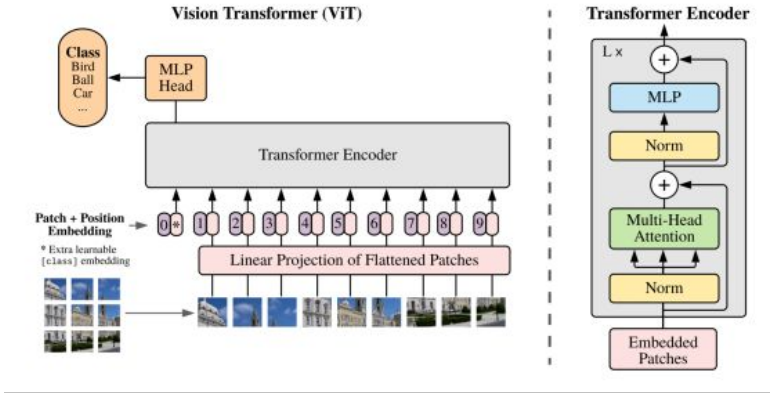
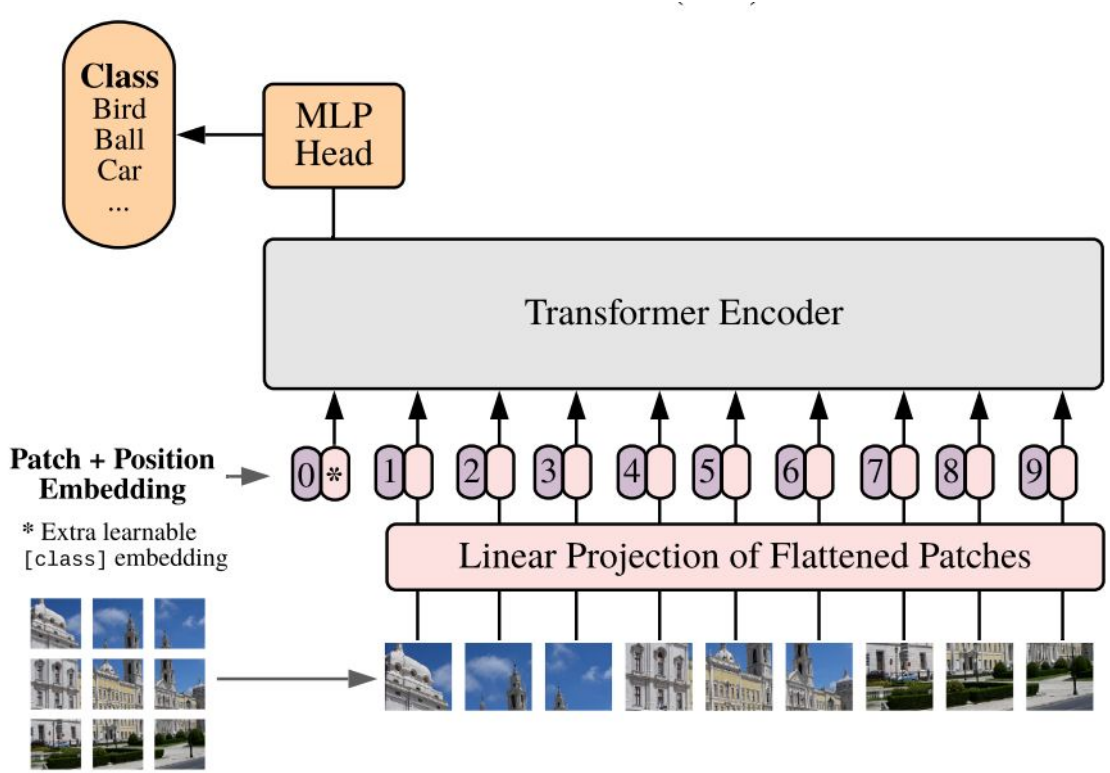
# VISION TRANSFORMERS




IDIOT DEVELOPER

# What is Vision Transformer?



# Input Image Processing

**Input Image**

**Size:** 512 x 512 x 3

Transform

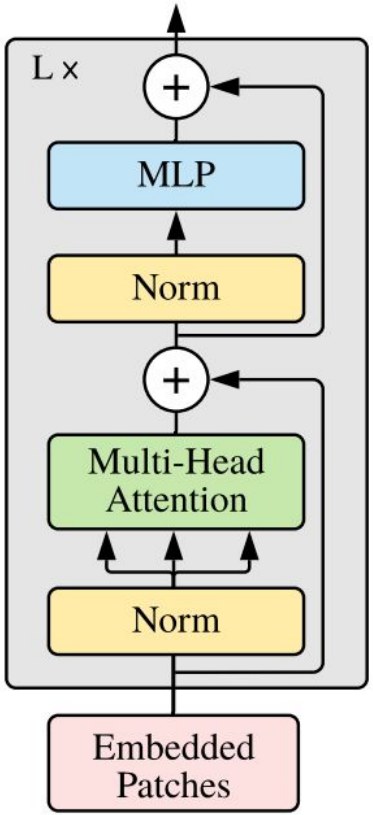
**Patches**

**Size:** 32 x 32 x 3
**Num Patches:** 256

# Input Image Processing

| THEORY | EXAMPLE |
| --- | --- |
| **Input Image:** H x W x C.<br>**Patch Size:** Ph x Pw<br>**Number of patches (N):** (H x W) / (Ph x Pw)<br>**Transformed Input:** (N, Ph x Pw x C)<br><br>H x W = Image height x width<br>C = Image channels<br>Ph x Pw = Patch height x width<br>N = Number of patches | **Input Image:** 512 x 512 x 3<br>**Patch Size:** 32 x 32<br>**Number of patches (N):**<br>= (512 x 512) / (32 x 32)<br>= 256<br>**Transformed Input:**<br>= (256, 32 x 32 x 3)<br>= (256, 3072) |

# 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 |