- 文章概要
- 架构与细节
- 对比分析

Vision Transformer 架构

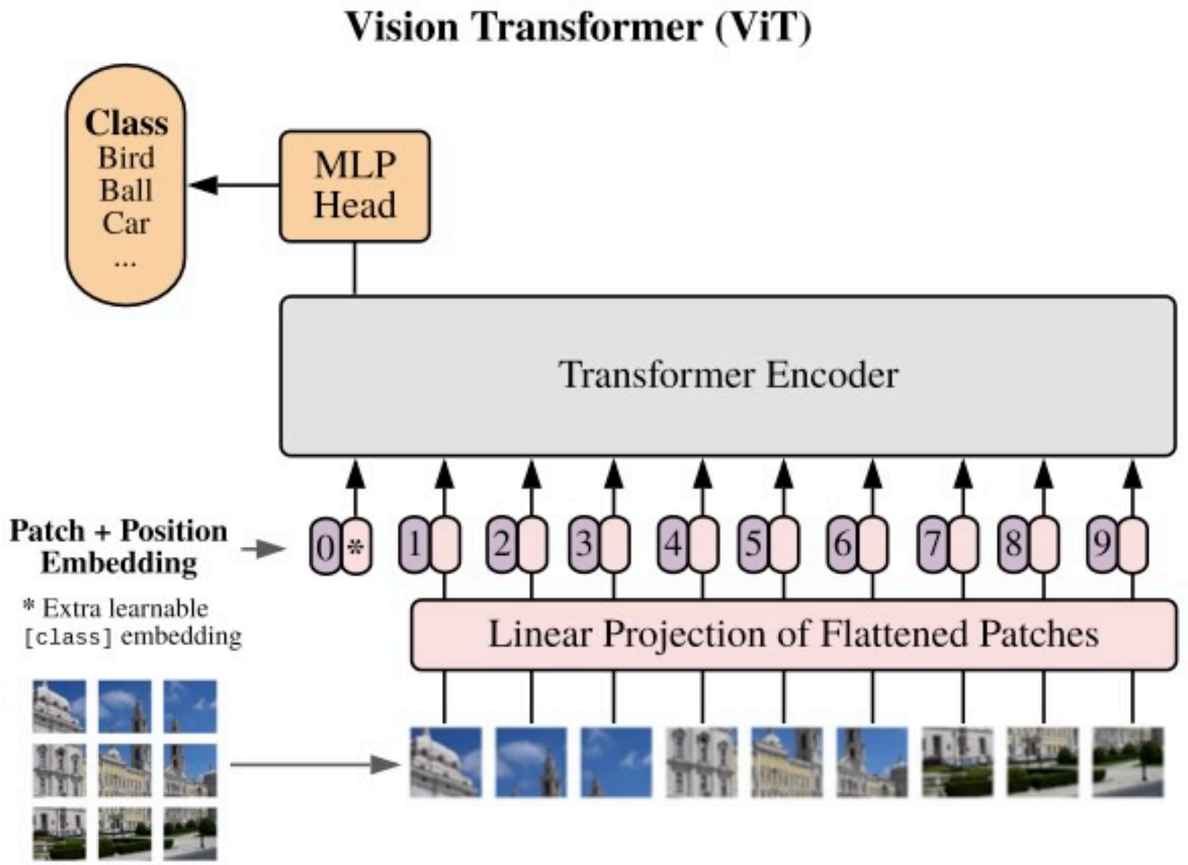

图 1　模型概览

- Linear Projection of Flattened Patches

  将 Patches (图像块) 序列化、通过全连接神经网络投射为 1 维向量并嵌入位置编码、加上 token 为 [class] 的同维度向量作为 Transformer Encoder 的输入

- Transformer Encoder

  利用自注意力机制与全连接神经网络提取输入向量中包含的信息

- MLP Head

  以 Transformer Encoder 的 [class] 输出作为输入，通过 MLP (多层感知机 - 全连接神经网络) 计算并输出图像类别