Vision Transformer 架构

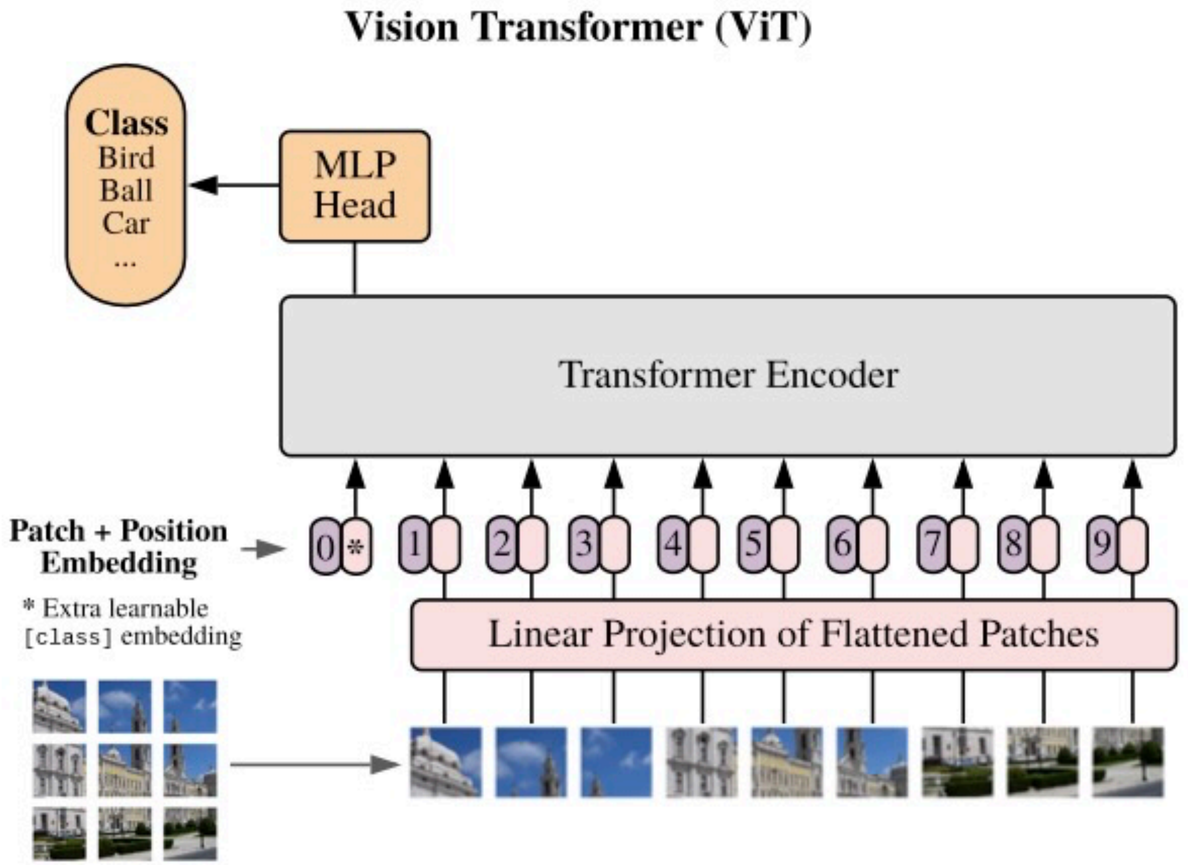


图 1　模型概览

- 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

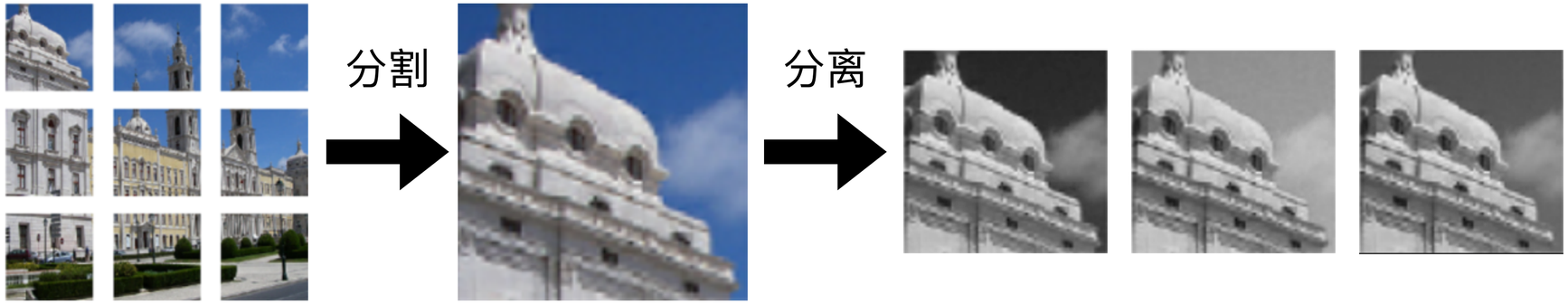
- Transformer Encoder

  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

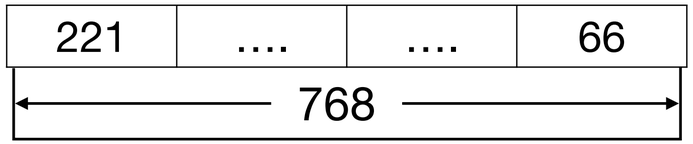
  以 Transformer Encoder 的 [class] 输出作为输入，通过 MLP (多层感知机 - 全连接神经网络) 计算并输出图像类别

Linear Projection of Flattened Patches



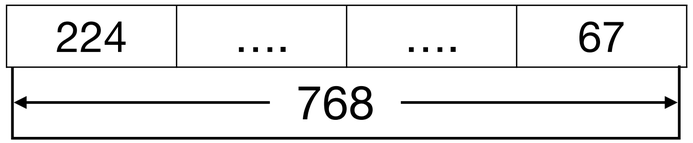
$x \in \mathbb{R}^{H \times W \times C}$: image　　$x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$: patch　　C : Channels

序列化并线性投射



x_p

嵌入位置信息



x_p with position embedding

- 结合神经网络，将图像信息转化为适用于 Transfomer Encoder 的向量形式，类似于自然语言处理中的词向量化，能够起到特征提取的作用