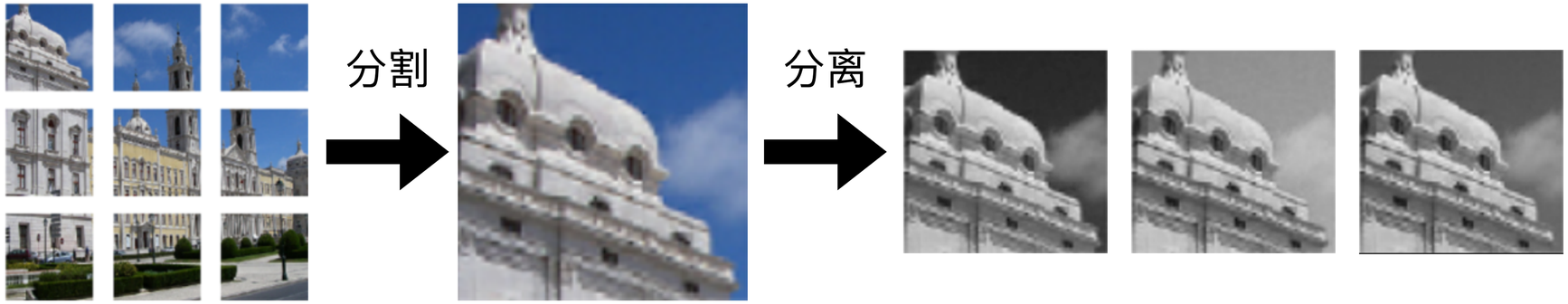
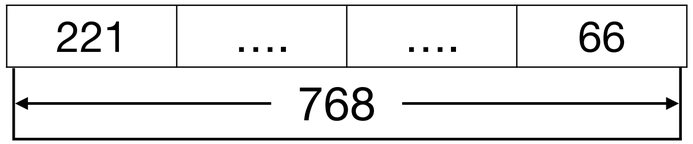
Linear Projection of Flattened Patches



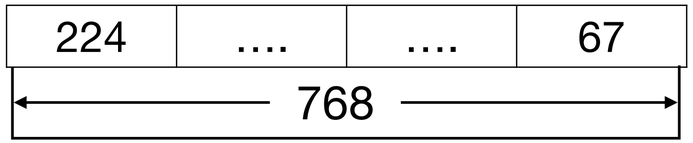
$x \in \mathbb{R}^{H \times W \times C}$: image　　$x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$: patch　　C : Channels

序列化并线性投射



x_p

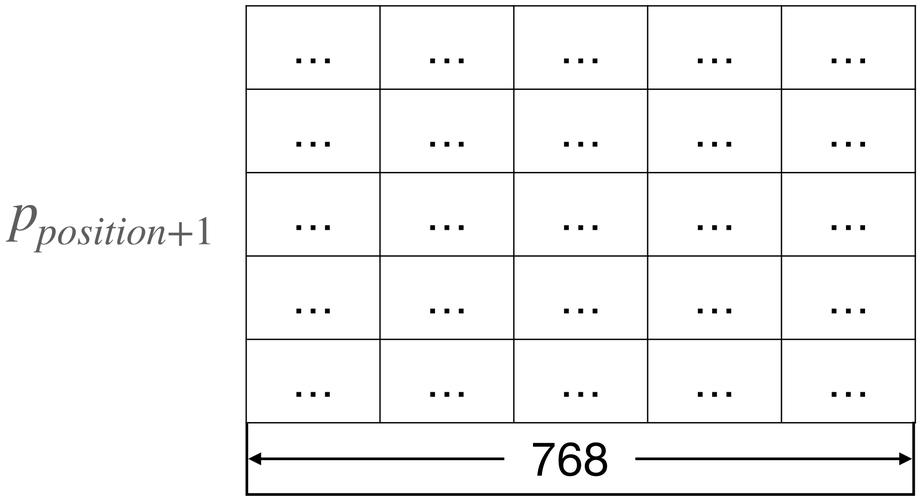
嵌入位置信息



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

Position Embedding



Vision Transformer Position Embedding 矩阵

$$PE(pos,2i) = sin(pos/10000^{2i/d_{model}}) \text{ .....式 1}$$

$$PE(pos,2i+1) = cos(pos/10000^{2i/d_{model}}) \text{ .....式 2}$$

原始 Transformer Position Embedding 生成公式

- Position Embedding 在 NLP 中代表词汇间的顺序信息，在 Vision Transformer 中代表图像块的位置信息
- Vision Transformer 的 Position Embedding 通过训练得来，与原始 Transformer 中的方法不同