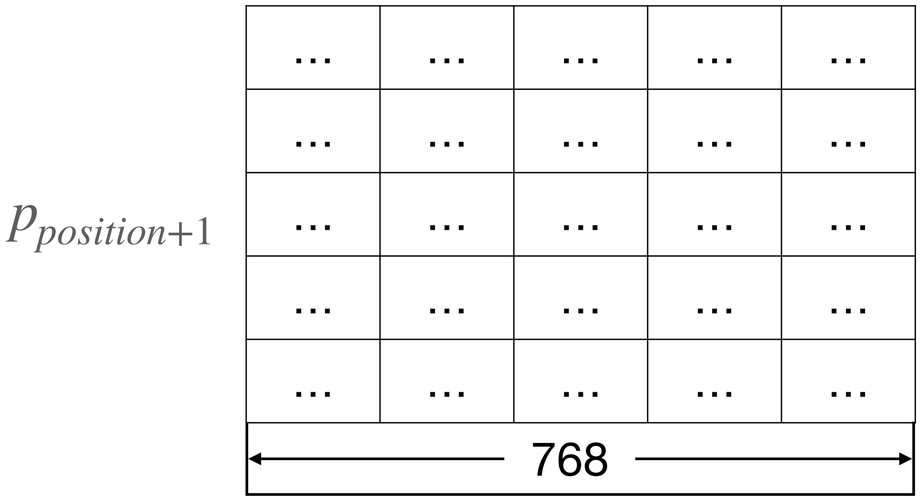
Position Embedding



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

Position Embedding

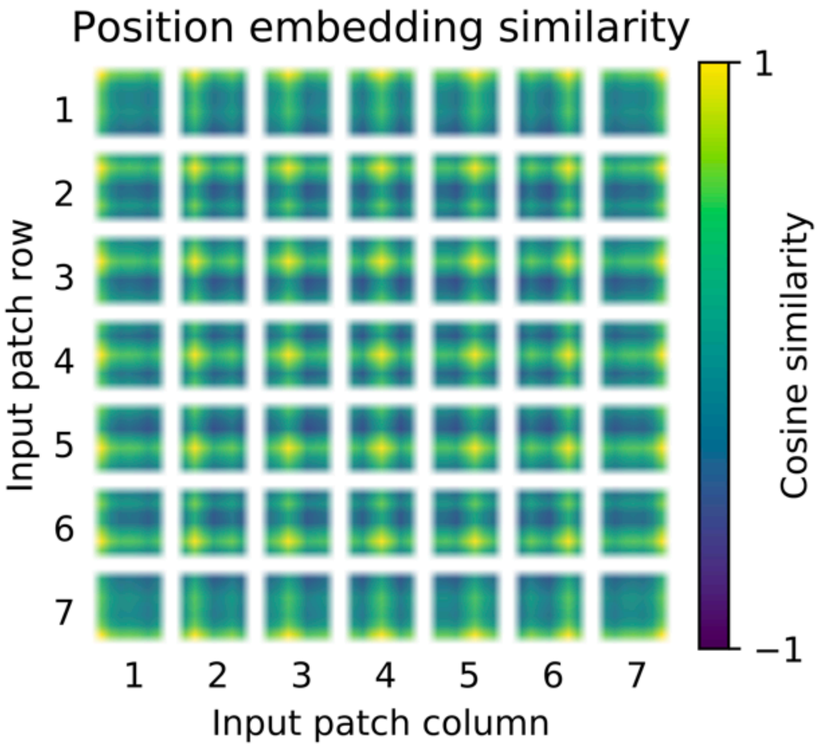


图 2　相似程度可视化

- 对训练后的 Position Embedding 两两间的相似程度进行可视化
- 可视化结果显示所有 Position Embedding 与其自身和近邻向量的相似性较高，可以证明模型能够学习到较准确的二维位置信息