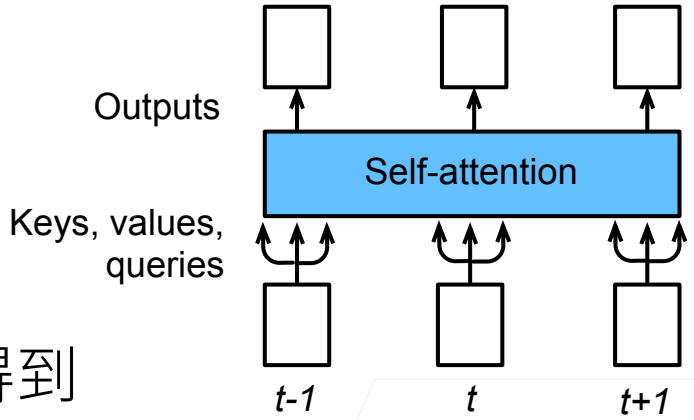
自注意力和位置编码

动手学深度学习 v2
李沐・AWS

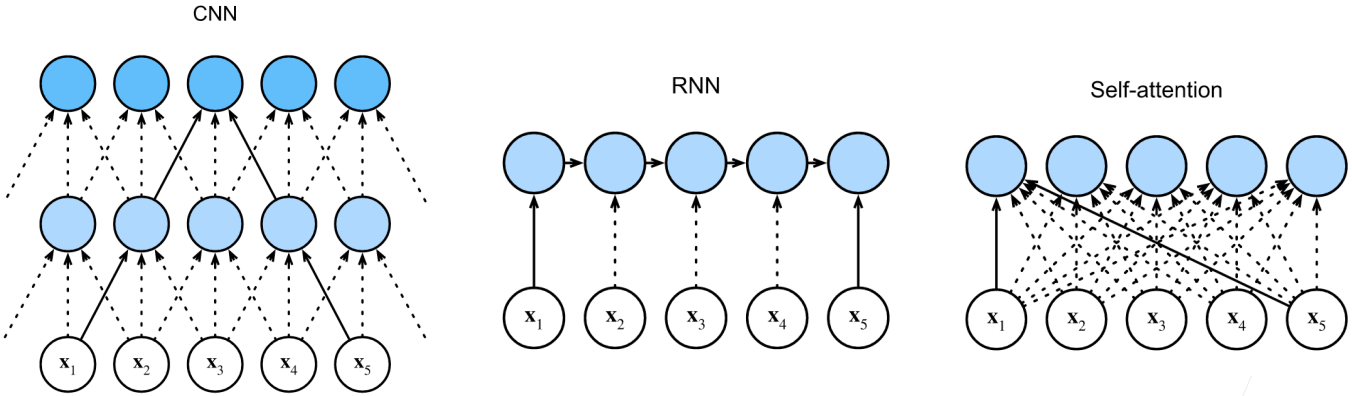
自注意力

- 给定序列 $\mathbf{x}_1, \ldots, \mathbf{x}_n, \forall \mathbf{x}_i \in \mathbb{R}^d$
- 自注意力池化层将 $\mathbf{x}_i$ 当做key，value，query来对序列抽取特征得到 $\mathbf{y}_1, \ldots, \mathbf{y}_n$，这里

$$\mathbf{y}_i = f(\mathbf{x}_i, (\mathbf{x}_1, \mathbf{x}_1), \ldots, (\mathbf{x}_n, \mathbf{x}_n)) \in \mathbb{R}^d$$



跟CNN，RNN对比



	CNN	RNN	自注意力
计算复杂度	O(knd^2)	O(nd^2)	O(n^2d)
并行度	O(n)	O(1)	O(n)
最长路径	O(n/k)	O(n)	O(1)

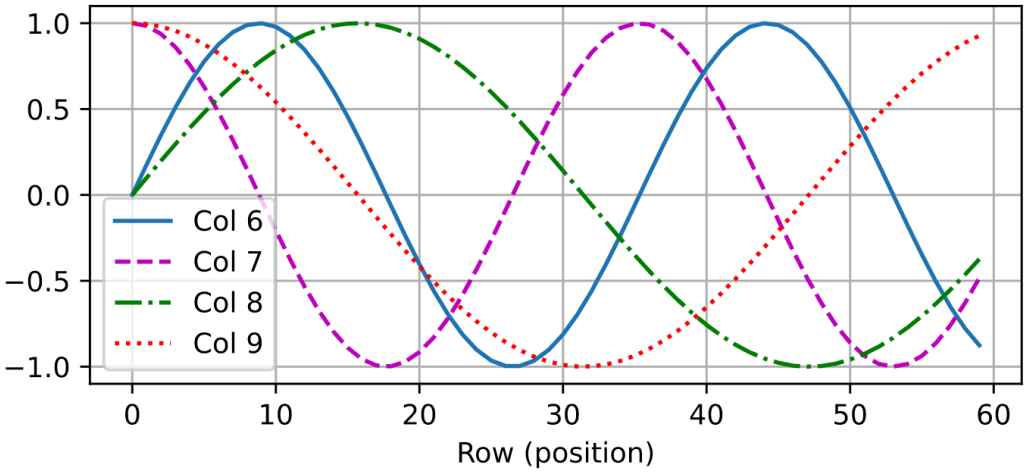
位置编码

- 跟CNN/RNN不同，自注意力并没有记录位置信息
- 位置编码将位置信息注入到输入里
  - 假设长度为 n 的序列是 $\mathbf{X} \in \mathbb{R}^{n \times d}$，那么使用位置编码矩阵 $\mathbf{P} \in \mathbb{R}^{n \times d}$ 来输出 $\mathbf{X} + \mathbf{P}$ 作为自编码输入
- $\mathbf{P}$ 的元素如下计算：

$$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right), \quad p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right)$$

位置编码矩阵

- $\mathbf{P} \in \mathbb{R}^{n \times d}$：$p_{i,2j} = \sin\left(\frac{i}{10000^{2j/d}}\right), \quad p_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/d}}\right)$

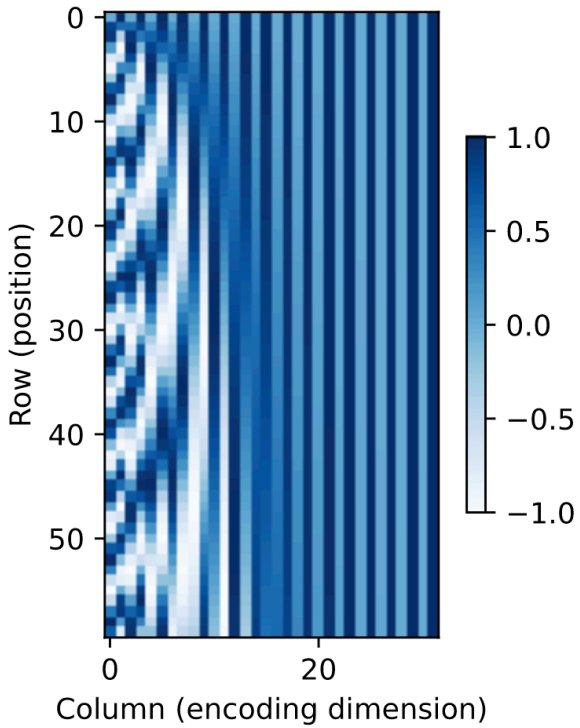


绝对位置信息

计算机使用的二进制编码

```
0 in binary is 000
1 in binary is 001
2 in binary is 010
3 in binary is 011
4 in binary is 100
5 in binary is 101
6 in binary is 110
7 in binary is 111
```

位置编码矩阵



相对位置信息

- 位置于 $i+\delta$ 处的位置编码可以线性投影位置 i 处的位置编码来表示
- 记 $\omega_j = 1/10000^{2j/d}$，那么

$$\begin{bmatrix} \cos(\delta\omega_j) & \sin(\delta\omega_j) \\ -\sin(\delta\omega_j) & \cos(\delta\omega_j) \end{bmatrix} \begin{bmatrix} p_{i,2j} \\ p_{i,2j+1} \end{bmatrix} = \begin{bmatrix} p_{i+\delta,2j} \\ p_{i+\delta,2j+1} \end{bmatrix}$$

投影矩阵
跟 i 无关

总结

- 自注意力池化层将 $\mathbf{x}_i$ 当做key，value，query来对序列抽取特征
- 完全并行、最长序列为1、但对长序列计算复杂度高
- 位置编码在输入中加入位置信息，使得自注意力能够记忆位置信息