注意力如何计算 $$A(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V$$

好的，我来用一个具体的简单例子，逐步解释**注意力机制（Attention）**是如何计算的。

问题背景：

我们有两个输入单词（Token）：

- x_1：代表单词"Thinking"
- x_2：代表单词"Machines"

每个输入会生成三个向量：**查询向量（query, q）**、**键向量（key, k）**和**值向量（value, v）**。这些向量通过线性映射生成，并用于计算注意力。

步骤 1：计算查询与键的点积

首先，我们计算每个查询向量 q 与所有键向量 k 之间的**点积（Dot Product）**，以衡量查询和键的相关性。结果如下：

- $q_1 \cdot k_1 = 112$（表示第一个单词的查询向量与自己的键向量的相关性）
- $q_1 \cdot k_2 = 96$（表示第一个单词的查询向量与第二个单词的键向量的相关性）

同样地，计算第二个单词的查询与所有键的点积：

- $q_2 \cdot k_1 = 96$
- $q_2 \cdot k_2 = 120$

步骤 2：缩放点积并通过 Softmax 归一化

接下来，缩放点积的结果（通常是除以键向量的维度的平方根），并通过 **Softmax 函数** 将这些点积值转化为概率分布：

- 对于第一个单词 q_1，点积结果是 [112, 96]，除以缩放因子 $\sqrt{d_k}$ 后，得到 [14, 12]。
- 然后通过 Softmax 计算，得到权重 [0.88, 0.12]。

同样地，计算 q_2 的权重：

- 点积结果是 [96, 120]，除以缩放因子后得到 [12, 15]。
- 通过 Softmax 计算得到权重 [0.12, 0.88]。

步骤 3：计算加权和值向量的和

最后，使用从步骤 2 得到的权重对值向量进行加权平均。对于 q_1 和 q_2 的加权结果：

- **对于** q_1：我们将权重 [0.88, 0.12] 乘以相应的值向量 v_1 和 v_2，并相加，得到输出 z_1。
- **对于** q_2：权重 [0.12, 0.88] 乘以 v_1 和 v_2，并相加，得到输出 z_2。

总结：

1. **查询与键的点积**衡量了输入单词之间的相关性。
2. **通过 Softmax 归一化**，将相关性转化为权重。
3. **使用权重对值向量加权**，计算出最终的注意力结果。

这个机制使得每个输入单词根据其与其他单词的相关性，获得相应的注意力权重，从而影响它们在输出中的表示。