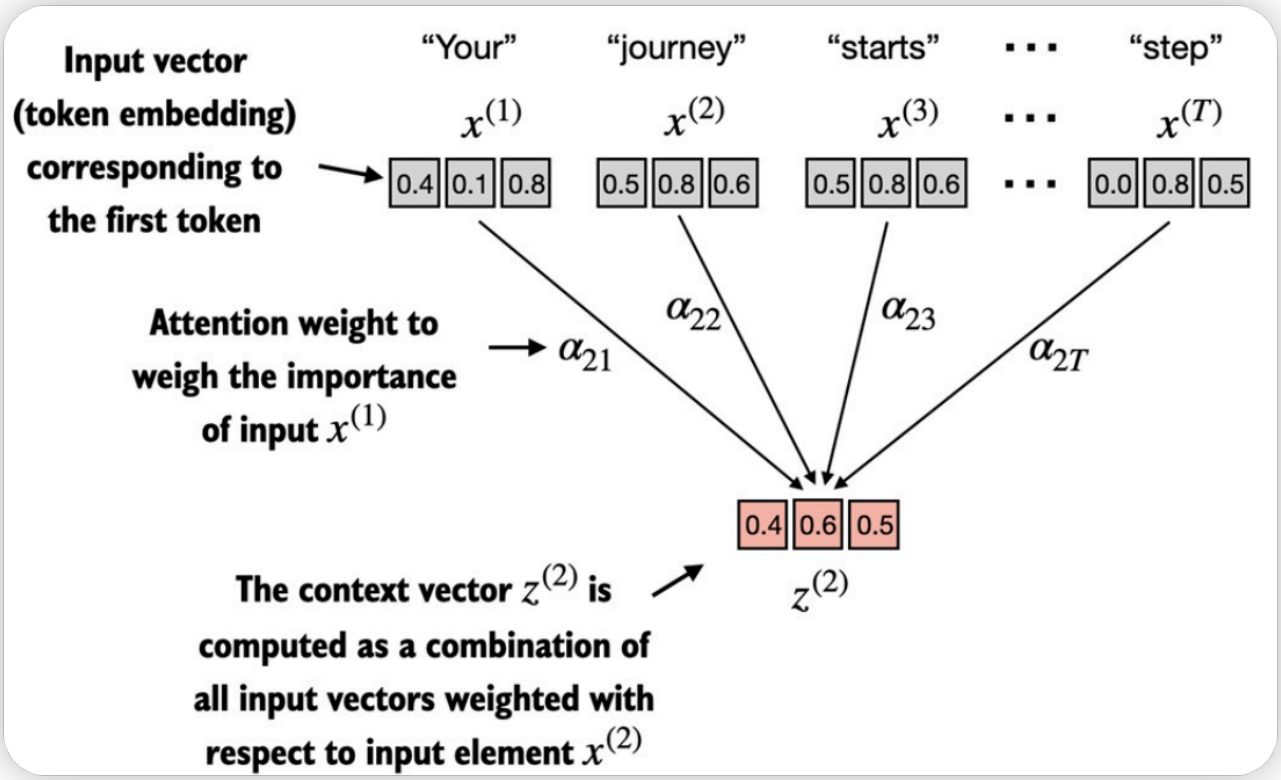
步骤

1. 计算注意力权重（attention weights）：通过输入向量
    1. 通过计算与每个输入的点积来计算：`attn_scores = inputs @ inputs.T`
    2. 然后再进行softmax归一化得到：`attn_weights = torch.softmax(attn_scores, dim=1)`
2. 计算上下文向量（context vector）：通过注意力权重和输入向量
    1. `attn_weights @ inputs`

exp

exp - 为输入元素 $x^{(2)}$ 计算上下文向量 $z^{(2)}$：



通过矩阵乘法计算代码如下：

```
# 注意力得分：直接计算输入向量之间的点积
attn_scores = inputs @ inputs.T
# 注意力权重：归一化注意力得分
attn_weights = torch.softmax(attn_scores, dim=1)
all_context_vecs = attn_weights @ inputs
```