Attention is all you need

## Scaled Dot-Product Attention

一种特殊的Attention结构，输入三个矩阵：Q(query)、K(key)、V(vale)。



**计算方式：**

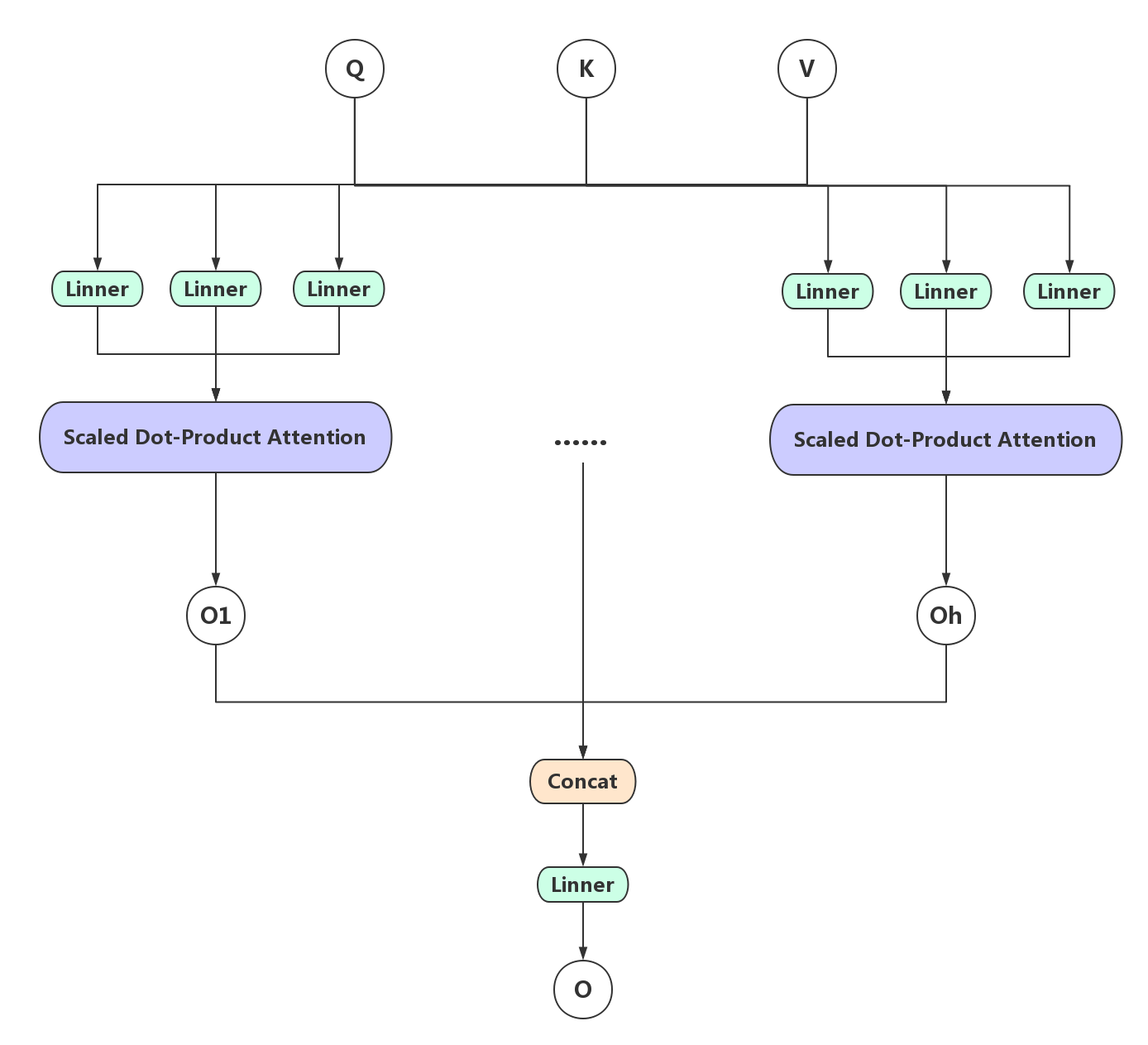
计算公式：

最终得到一个的矩阵。

其中：Scale层是正规化，Mask层实现一步的除法操作。

## Multi-HeadAttention

并行多个Scaled Dot-Product Attention模块，然后将结果融合



**各层说明：**

Linner: 线性映射层，即Dense层

Concat: 矩阵融合操作，将若干个矩阵和在一起，Concat(n\*m, n\*m)=n\*2m，最后再使用一个Dense层将维度还原为n\*m。

## 使用上述模型进行搭建翻译网络



**说明：**

Add : 矩阵加法

Norm : 正规化

Positional Encoding: 对位置进行编码，将位置坐标变成一个向量

训练时：输入到Encoding中的Outputs就是标签数据

测试时：输入到Encoding中的Outputs是零初始化的一组tensor