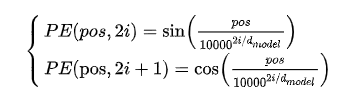
transformer答疑

1. 如何理解transformer的位置向量？



pos代表一个句子中的一个词。

如：我 想 谈恋爱

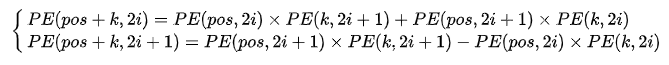
PE(0,2i) 代表位置为0的词我这个词的词向量的第2i个位置

PE(0,2i+1) 代表位置为0的词我这个词的词向量的第2i+1个位置

位置向量的目的是：希望模型学习到词与词之间的相对位置。

为啥这个公式可以表示词与词之间的位置信息呢？



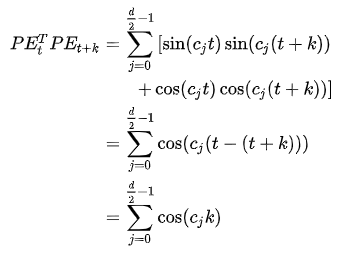


这个公式的含义是：

句子中的第pos+k个位置，某个词之后的第k个位置的词，2i是这个词对应的词向量的第2i个值，可以表示为第pos个位置的词向量的线性表示。

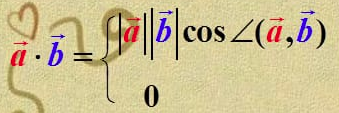
但是这个线性表示引入的东西比较多，包括第k个位置词向量的2i+1个值和2i个值，还有第pos个位置词的2i和2i+1的位置。

这样做一个线性表示就能表示词与词之间的相对位置吗？



怎样计算两个向量之间的距离：

因为每个位置向量的模都是一样的大小。跟词向量维度有关，如果维度为512，那么模的大小为256^0.5



衡量距离可以用向量和向量之间的角度来衡量。因此点积就可以衡量向量之间的距离。

可以看出第k个位置词和第t+k个位置之间的距离是k的函数。

因此那种表示是可以衡量相对位置的。

但是，思考另外一个问题：



这个公式意味着词与词之间的距离一致时便无法分清先后顺序。

例如：我 想 谈恋爱。针对想这个词，我 和 谈恋爱这两个词是无法定位其先后顺序的。

同时，位置向量需要通过长途跋涉才会到达decoder。

在第一个encoder中就出现了



这个意味着中间加了个矩阵作为乘法，虽然说最终的结果是一个标量，但是w是一个可以学习得到的参数，是具有可变性的，可变的东西意味着不再是一个仅仅是关于k的一个自变量的函数，而是更加复杂，更加没有规律性。

这可能就是transformer中位置信息做的不够细致的一个原因吧。后面会有很多人进行改进。