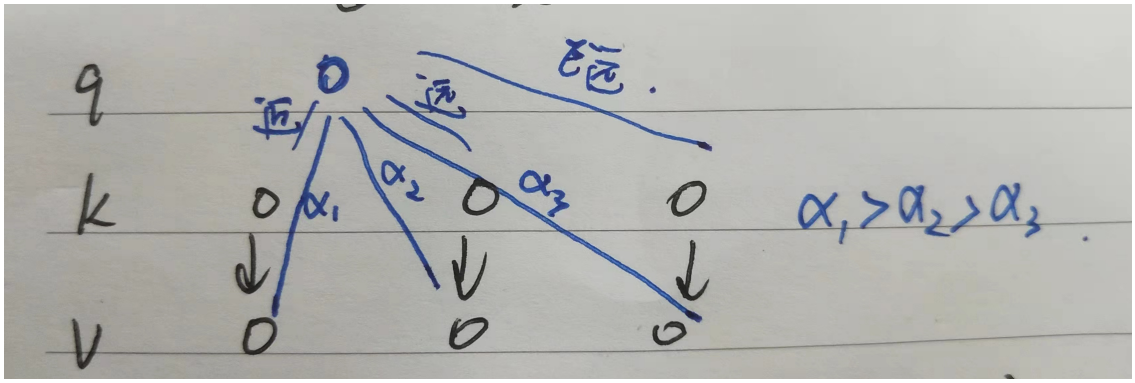
一、Transformer 阅读随记

1. 论文的首次提出是针对机器翻译任务
2. 主要 是将RNN中的循环层替换 成了 多头自注意力 （Multi-headed Self-attention）
    - 自注意力 即 q k v 都是一样的
    - 利用attention 抓取序列中的信息
    - 采用多头的原因：可以类比CNN中同时使用**多个滤波器**的作用，直观上讲，多头的注意力**有助于网络捕捉到更丰富的特征/信息，也就是综合利用各方面的信息/特征。**
3. 采用了LayerNorm
    - LayerNorm： 针对每个样本做Norm
    - BatchNorm：针对每一个batch中的每个feature做Norm
    - 采用原因：在时序的序列模型中，每个样本的长度可能会发生变化，若采用BatchNorm会产生不稳定的均值和方差，而使用LayerNorm后，方差和均值均比较稳定。
4. 解码器多了一个带掩码的多头自注意力
    - 避免看到t时刻及之后的时序信息
5. 前馈神经网络（Feed-Forward Networks）
    - 可以看成一个MLP
    - 多头注意力都是线性变换，而线性变换的学习能力不如非线性变换强。因为采用Feed-Forward通过激活函数的方式，强化表征能力。
6. 注意力



给一个q 去和 不同的 k 去比较，相似度越高，该k对应的v的权重越大。

7. Transformer 和 RNN 的异同
    - 异：**如何传递序列的信息**
        - RNN 是把上一个时刻的信息输出传入下一个时候做输入。
        - Transformer 通过一个 attention 层，去全局的拿到整个序列里面信息，再用 MLP 做语义的转换。
    - 同：**语义空间的转换 + 关注点**
        - 用一个线性层 or 一个 MLP 来做语义空间的转换。
        - 关注点：怎么有效的去使用序列的信息。