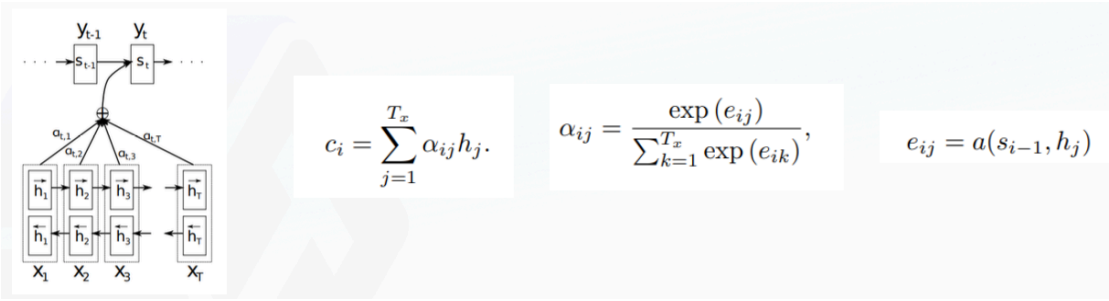
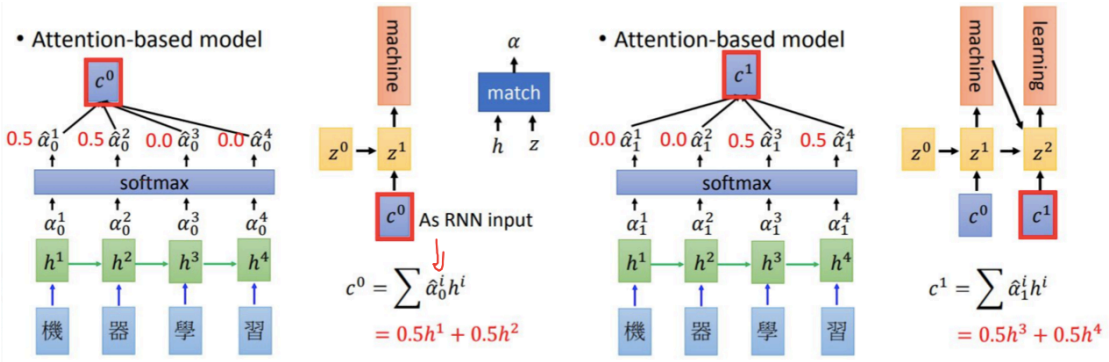
Week2

本笔记聚焦在 Attention 机制上。

1.Attention 概览

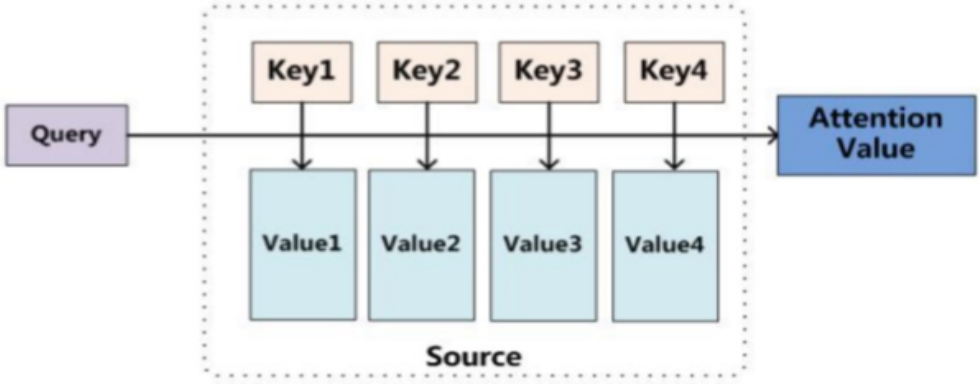
Attention 来源于人类的视觉注意力机制，即人一般不会关注整体，而是根据需要观察某个重要的部分。Attention 赋予模型对于重要性的区分和辨别能力。它的本质是一系列的权重分配，如在多特征或多模型中，若某些特征或者模型比较重要，那么赋予它们更多的权重。在 Attention-based 模型中，$a_{i,j}$ 为注意力权重向量，它来自上一层输出和当前 h 输出的函数。然后把 a_{ij}(注意力权重向量)和当前 h 输出相乘，结果 (此结果强调了某些值，弱化了某些值) 作为 RNN 输入。Attention 权重的最终值是通过机器学习方式获得。





2.KQV（key, query, value）直观解释

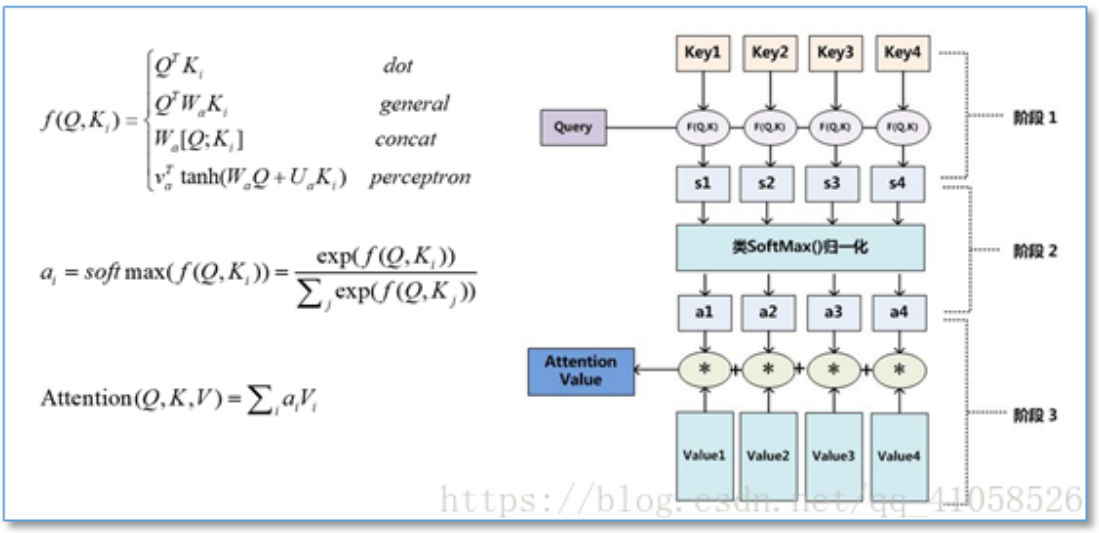
Attention 函数的本质可以被描述为一个查询（query）到一系列（键 key-值 value）对的映射，如下图:



在计算 attention 时主要分为三步:

a.将 query 和每个 key 进行相似度计算得到权重，常用的相似度函数有点积，拼接，感知机

等；
b.一般是使用一个 softmax 函数对这些权重进行归一化；
c.将权重和相应的键值 value 进行加权求和得到最后的 attention。
目前在 NLP 研究中，key 和 value 常常都是同一个，即 key=value。



3.Self-Attention

Self-Attention 可以捕获同一个句子中单词之间的一些句法特征。引入 Self-Attention 后会更容易捕获句子中长距离的相互依赖的特征
如果是 RNN 或者 LSTM，需要依次序序列计算，对于远距离的相互依赖的特征，要经过若干时间步步骤的信息累积才能将两者联系起来，而距离越远，有效捕获的可能性越小。
但是 Self-Attention 在计算过程中会直接将句子中任意两个单词的联系通过一个计算步骤直接联系起来，所以远距离依赖特征之间的距离被极大缩短，有利于有效地利用这些特征。
除此外，Self-Attention 对于增加计算的并行性也有直接帮助作用。这是为何 Self-Attention 逐渐被广泛使用的主要原因。

3.1Self-Attention 细节：

每个 word 创建三个向量 q1, k1, v1（Query, Key, Value），计算注意力得分 q1*k1，q1*k2，然后除以 64 的开方（8），得到的值进行 softmax 标准化。再让 Value 乘以对应的 softmax 值，sum 所有 value，产生一个单词的 self-attention.