Self-Attention for Seq2Seq model

2021年2月4日 14:24

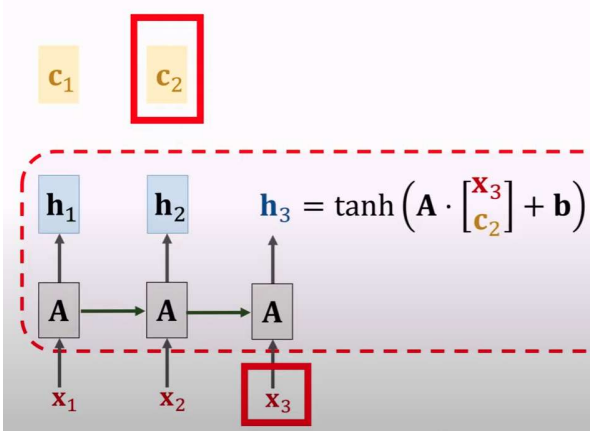
1. **Self-Attention for Sequence-to-Sequence**

- 链接：https://www.youtube.com/watch?v=Vr4UNt7X6Gw

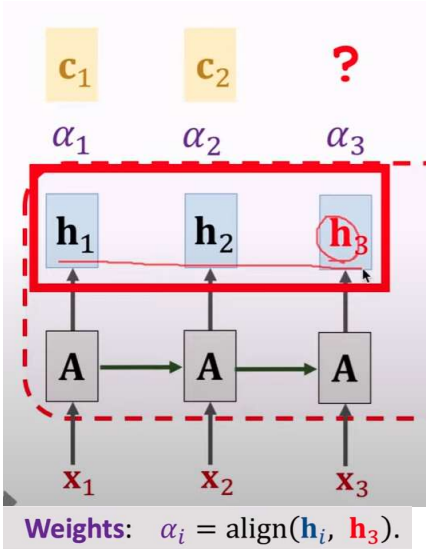
前面两个模型的最主要的问题：如果训练的数据集不够，而多层lstm叠加让embedding层的参数过多，就会出现过拟合的现象，导致实验结果不理想

- 思路

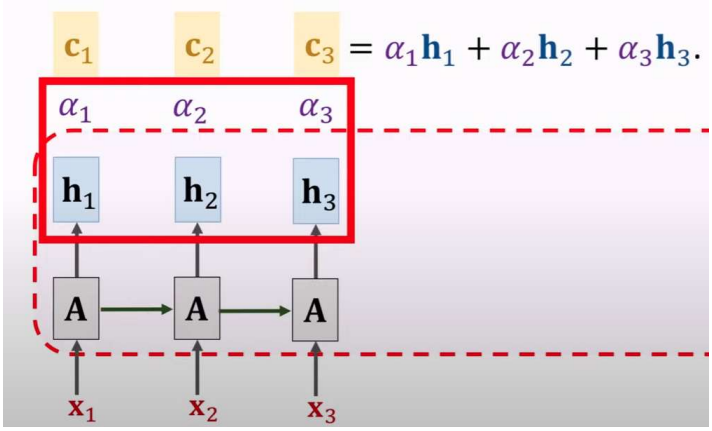
主要用在encoder部分，自己给自己加了attention



1. 初始状态为 c_0和h_0都是0向量，所以在这里没有写出来
2. 左侧的图表示的是隐藏状态h的更新

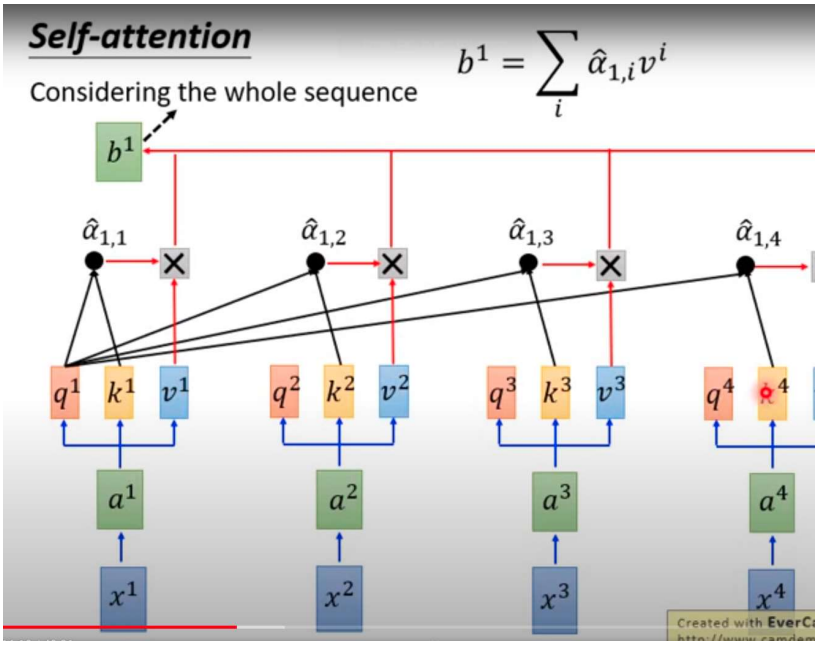


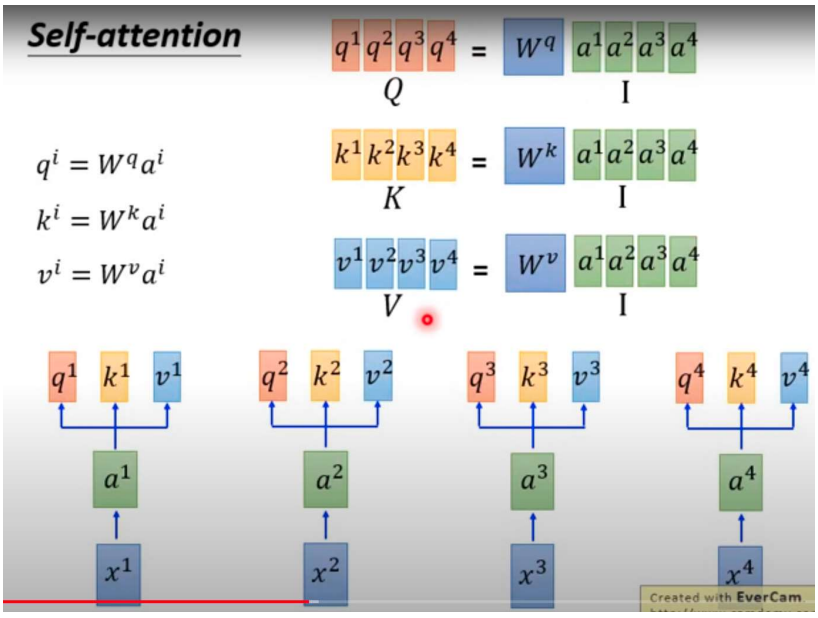
1. 参数α_i分别是根据每个隐藏状态和最新状态之间的相关性计算出来的
2. 左侧的图表示的是隐藏状态h的更新
3. Align 表示的是相关性！

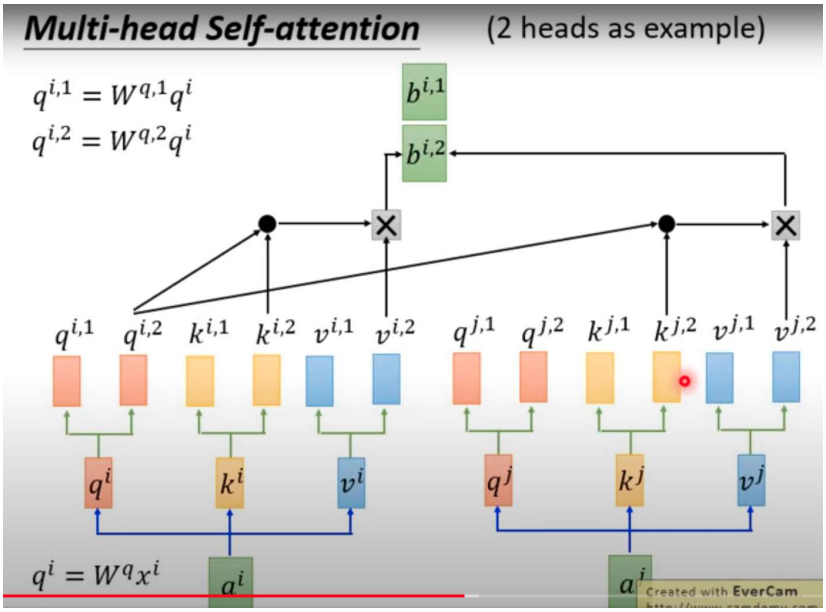


1. 参数 c 表示隐藏状态和权重参数之间加权平均

Self-attention for Transformer






Multi-head Self-attention
(2 heads as example)
$q^{i,1} = W^{q,1}q^i$
$q^{i,2} = W^{q,2}q^i$
$b^{i,1}$
$b^{i,2}$
$q^{i,1}$ $q^{i,2}$ $k^{i,1}$ $k^{i,2}$ $v^{i,1}$ $v^{i,2}$ $q^{j,1}$ $q^{j,2}$ $k^{j,1}$ $k^{j,2}$ $v^{j,1}$ $v^{j,2}$
q^i k^i v^i q^j k^j v^j
$q^i = W^q x^i$
a^i
a^j
Created with EverCam