Attention

Friday, June 4, 2021 12:57 AM

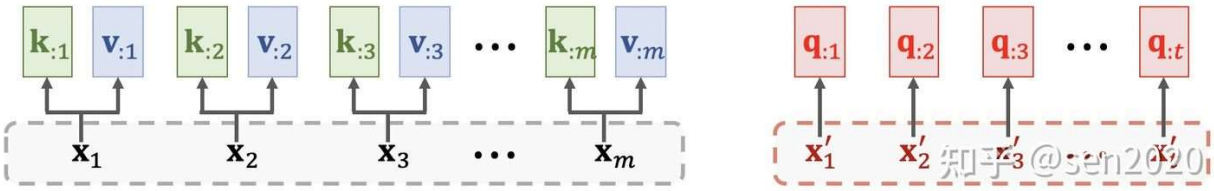
Attention用在Sequence-to-sequence，有两个序列。对左边的序列做线性变换，得到key和value。对右边的序列做线性变换，得到query。

主要可以参考blog：
https://blog.csdn.net/tg229dvt5i93mxaq5a6u/article/details/78422216

这个博客包含了seq2seq，seq2seq+attention，transformer
https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

- Keys and values are based on encoder's inputs $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$.
- Key: $\mathbf{k}_{:i} = \mathbf{W}_K \mathbf{x}_i$.
- Value: $\mathbf{v}_{:i} = \mathbf{W}_V \mathbf{x}_i$.

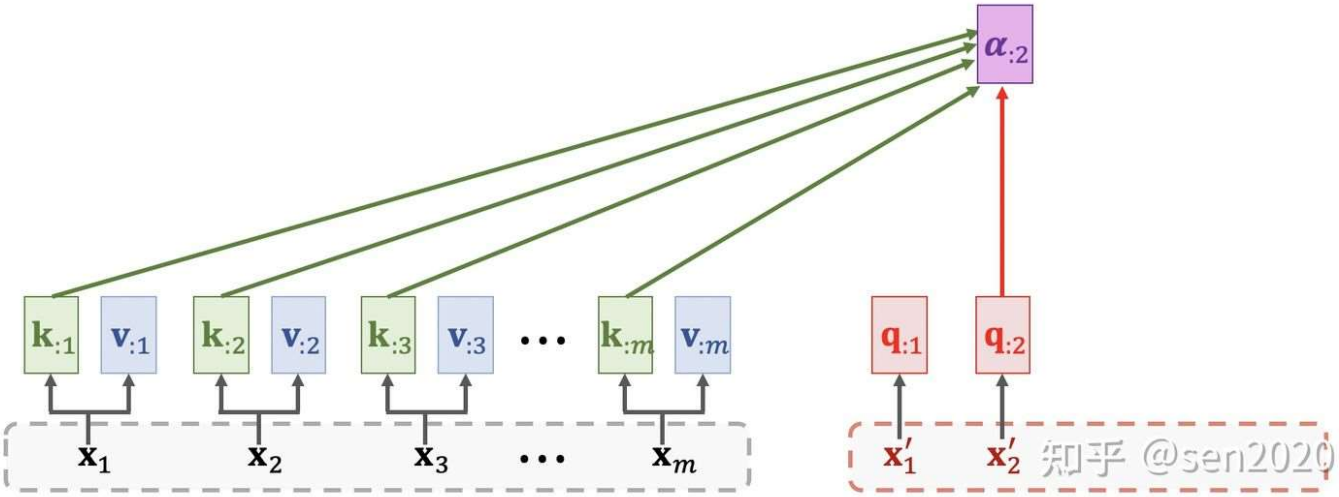
- Queries are based on decoder's inputs $\mathbf{x}'_1, \mathbf{x}'_2, \cdots, \mathbf{x}'_t$.
- Query: $\mathbf{q}_{:j} = \mathbf{W}_Q \mathbf{x}'_j$.



query的意思是"去匹配key"。

key的意思是"被query匹配"。

用query和key共同计算权重alpha。



用权重alpha对value做加权平均，得到context vector c。