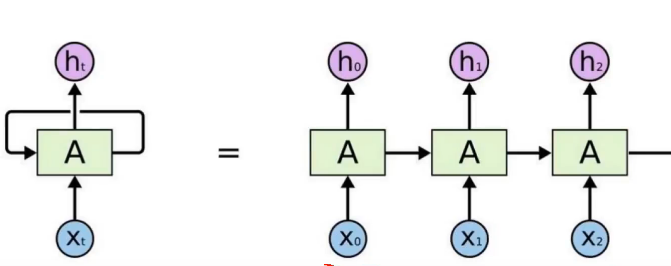
# 传统问题

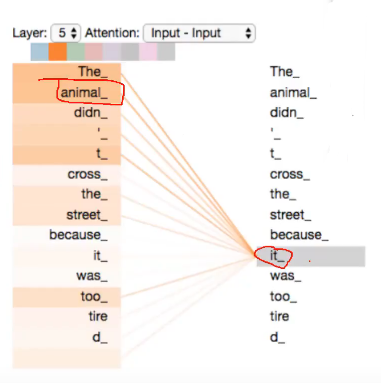
# 传统的RNN的问题



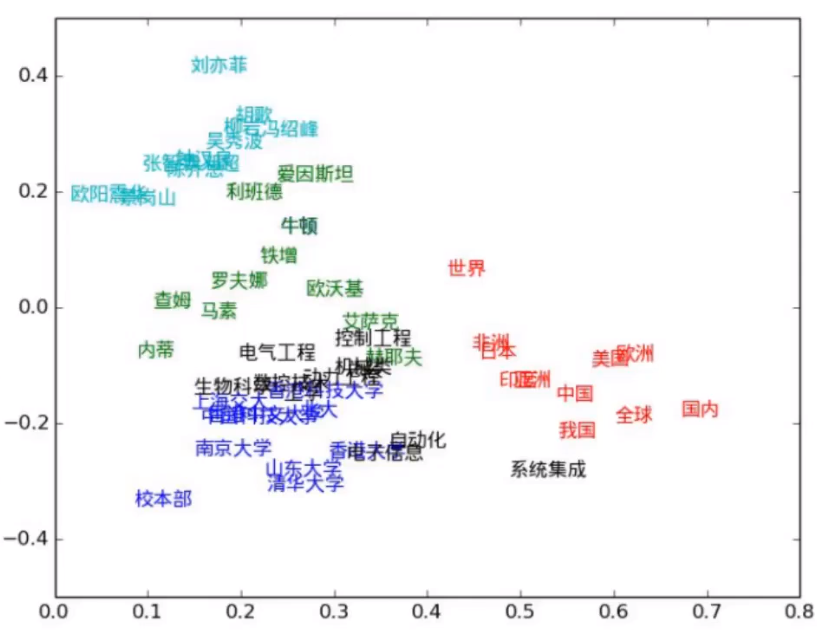
不能并行计算，每个片段在输入之后还要兼顾上一个单元的输出。

使用Self-Attention机制来进行并行计算

Attention的意思就是对于某些特定词/图形片段进行着重处理，关注一些更有价值的信息。



传统的word2



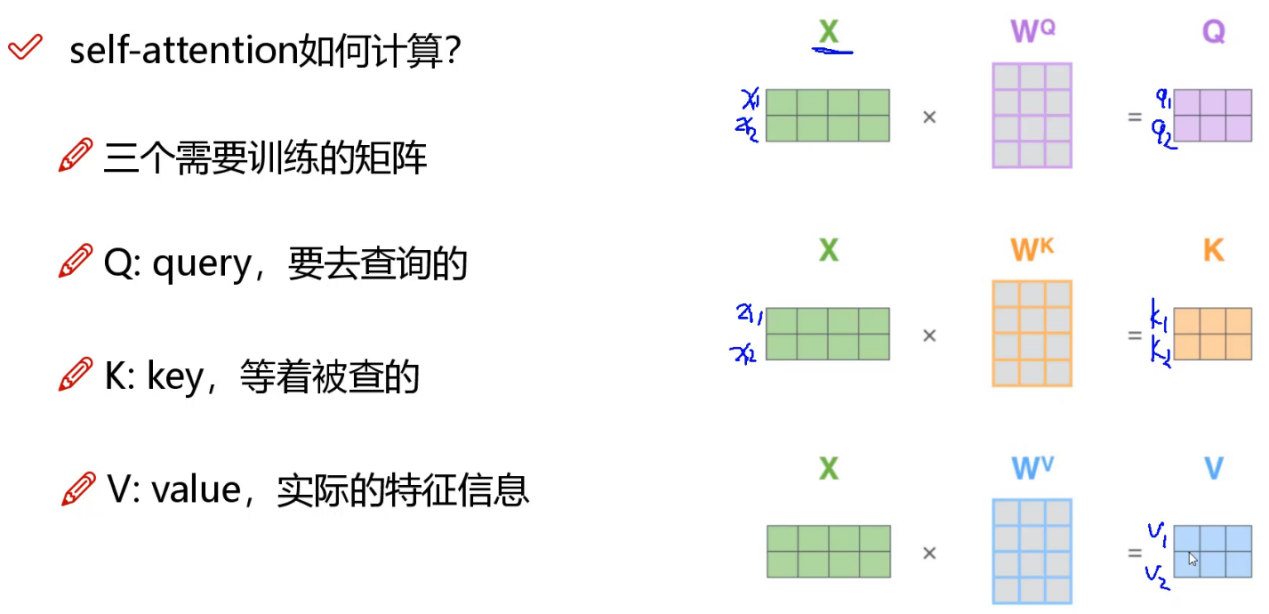
传统向量的问题：

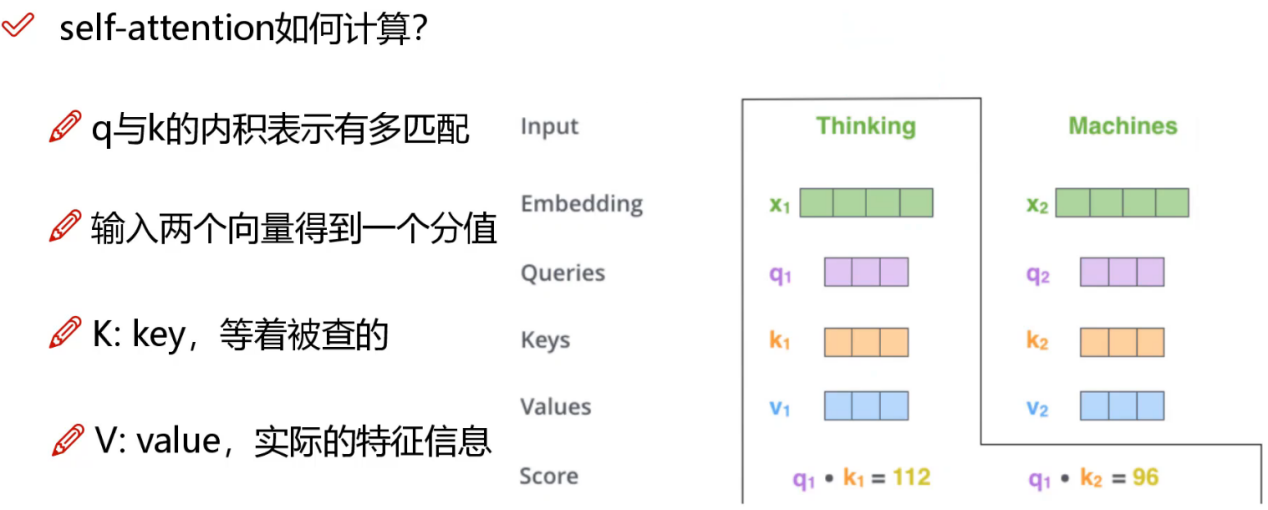
①预训练好的词向量就永久不变了，解决不了一词多意的情况。

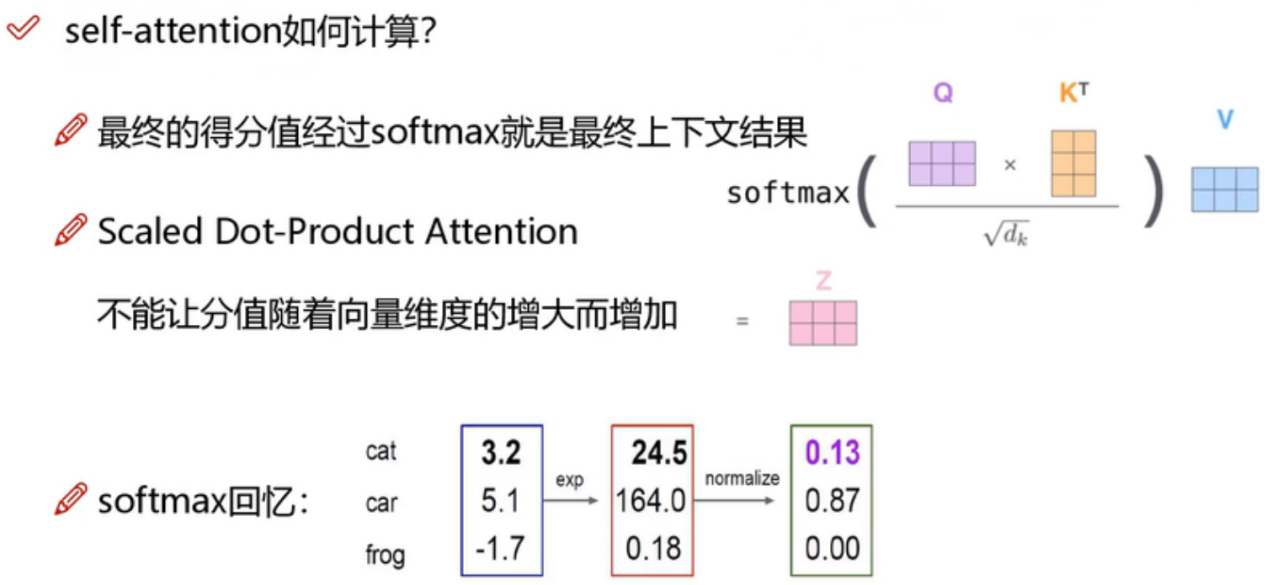
②不同语境中相同的词表达不便。

Bert的词向量与上下文有关联，并且可以作为Word2Vec的转换矩阵迁移到某一任务中

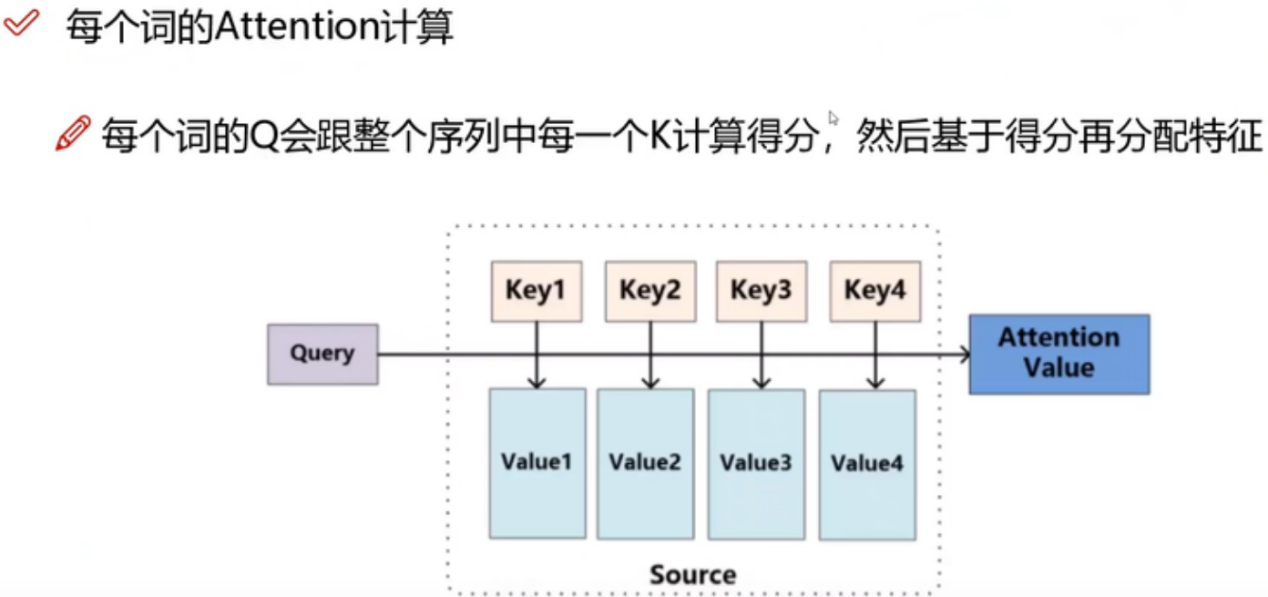
### Self-Attention的计算







最终获得某个词与它另外的每一个词的关联程度。

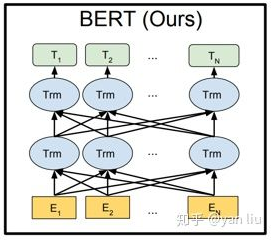


# BERT

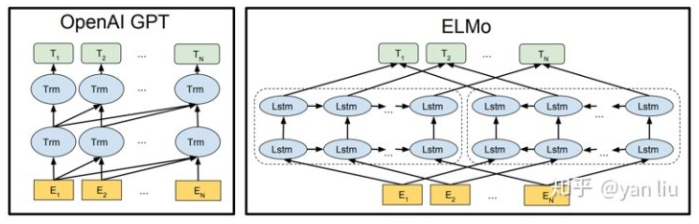
原文摘要  
我们介绍一个叫做BERT的新的语言表示模型，它可以理解为转换器的双向解码表示。不像最近的一些语言表示模型，

BERT被设计成通过在所由层联合左边和右边的上下文来做深度双向表示的预训练。预训练的BERT表示可以通过仅仅

一个额外的输出层的微调，在很多诸如问答，语义推断等自然语言处理任务上达到state-of-the-art的结果,而且不需要为特定任务在结构上做很大的改动。

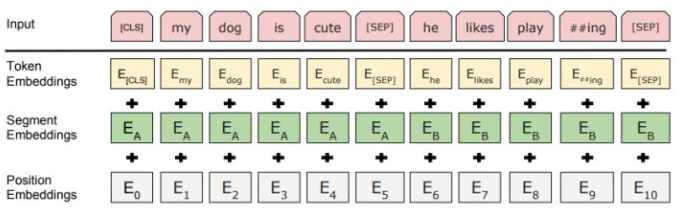


Bert的网络结构



另外两种模型的网络结构

GPT是只向右，ELMo是左右结合，Bert是同时左右



初期输入格式

位置嵌入（Position Embedding）：位置嵌入是指将单词的位置信息编码成特征向量，位置嵌入是向模型中引入单词位置关系的至关重要的一环

用于区分两个句子，例如B是否是A的下文（对话场景，问答场景等）。对于句子对，第一个句子的特征值是0，第二个句子的特征值是1

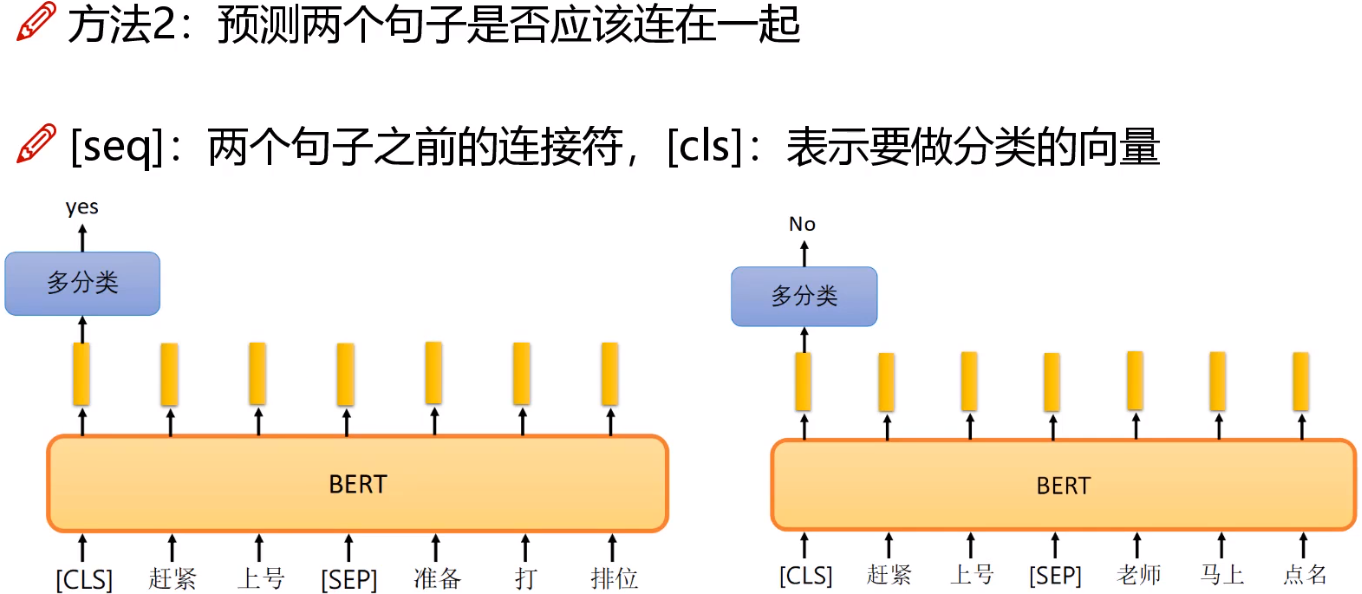


80%：my dog is hairy -> my dog is [mask]

10%：my dog is hairy -> my dog is apple

10%：my dog is hairy -> my dog is hairy

不是100%mask是为了在后期再遇到不认识的词的时候不出现错误，还有避免“记忆”的情况发生



CLS标签将于其它所有词进行Self-Attention的计算，得到整个句子的特征向量。

BERT能做到这一点得益于Transformer中Attention机制将任意位置的两个单词的距离转换成了1

（传统的rnn编码的时候，最后一个词和第一个词之间的距离为整个句长，这种长距离编码导致rnn很难发现他们的依赖关系。

Transformer中的self-attention使得每个单词两两之间都可以进行dot-production，这样更容易克服任意长度的词依赖。）

输入：其它人先做一个词（句）向量，利用Bert模型优化这个词（句）向量

测试近义句或近义词，根据上下文辨别准确的意思