NLP的深度学习模型和方法

★ Transformer

Transformer模型是由谷歌在2017年6月发表的论文Attention Is All Your Neea中提出的。这是一种称为seq2seq的模型，在机器翻译等应用中使用广泛。传统的seq2seq模型通常采用RNN,一般在网络结构中会用到Encoder和Decoder，要想提升效果，可以通过注意力(Attention）机制连接Encoder和Decoder。研究表明，如果使用RNN作为Encoder和Decoder，则存在两个问题：一是RNN的递归依赖难以并行化，早期版本的谷歌翻译系统（Google’s neural machine translation system,GNMT）需要96块GPU并行训练一周，而RNN无法提供这方面的支持；二是缺乏对全局语义信息的理解，尤其是在长时记忆、层级化语义表达两方面捉襟见肘。

Transformer模型摈弃了RNN，提出一种全新的并且更简单的网络结构，只需要Attention机制就能解决seq2seq的问题，并且能够一步到位获取全局语义信息。Transformer在机器翻译任务上的表现超过了RNN、CNN,其最大优点是可以高效地并行化。

NLP的深度学习模型和方法

★ Transformer

Transformer的核心是Attention机制。

(1)在编码当前词时，充分考虑上下文的信息。相比ELMO,Attention机制的独到之处是为不同的上下文分配不同的权重。例如，The bird didn’t fly because it was hurt by the cat,如果采用RNN或者LSTM作为编解码器，就是平等对待上下文的，因此不容易理解“it"是指代bird;而Attention机制会给bird分配较高的权重，这样就可以模拟人脑的Attention机制，从而准确地识别出“it”的含义。

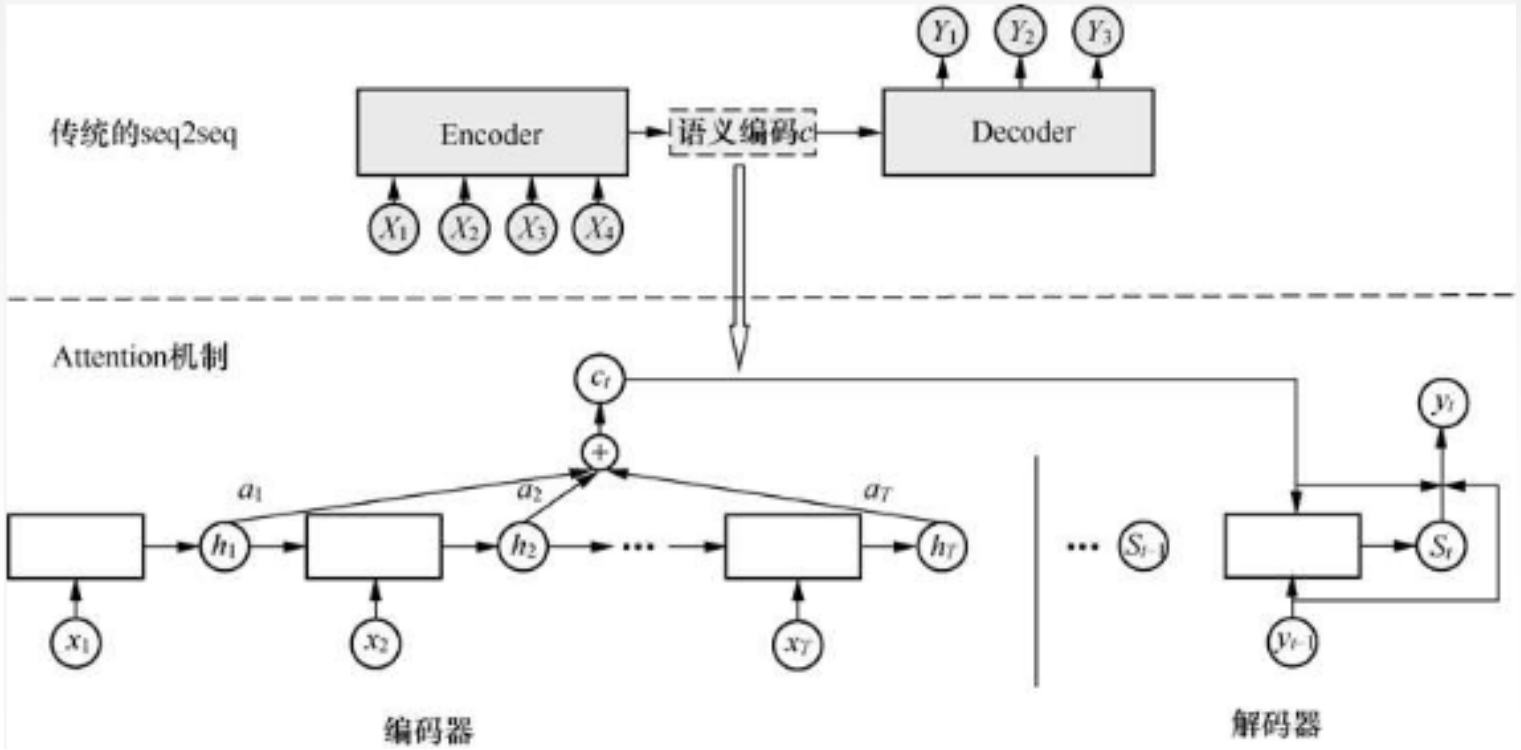
(2)在具体实现时，Transformer又加入了Self-Attention和Multi-Head Attention,通过多组权重参数优化上下文对当前词的影响，进一步提升了语义理解能力。

(3)除了在Encoder和Decoder加入Attention机制外，训练过程中，Decoder在每个时间步中还有单独一个Attention是从Encoder输入的，帮助当前词获取当前需要关注的重点内容。

NLP的深度学习模型和方法

★ Transformer

Attention机制的原理如图所示。传统的Seq2seq结构中，输入编码为一个定长语义编码，然后通过这个编码再生成对应的输出序列。针对这个问题，Bengio率先提出Attention机制，并因此获得2019年的图灵奖。区别在于，Encoder的输出不是一个语义向量，而是一个语义向量的序列，在解码阶段会有选择地从向量序列中选择一个子集，至于这个子集怎么选取，子集元素占比多少，这些都是Attention机制要解决的问题。



NLP的深度学习模型和方法

★ Transformer

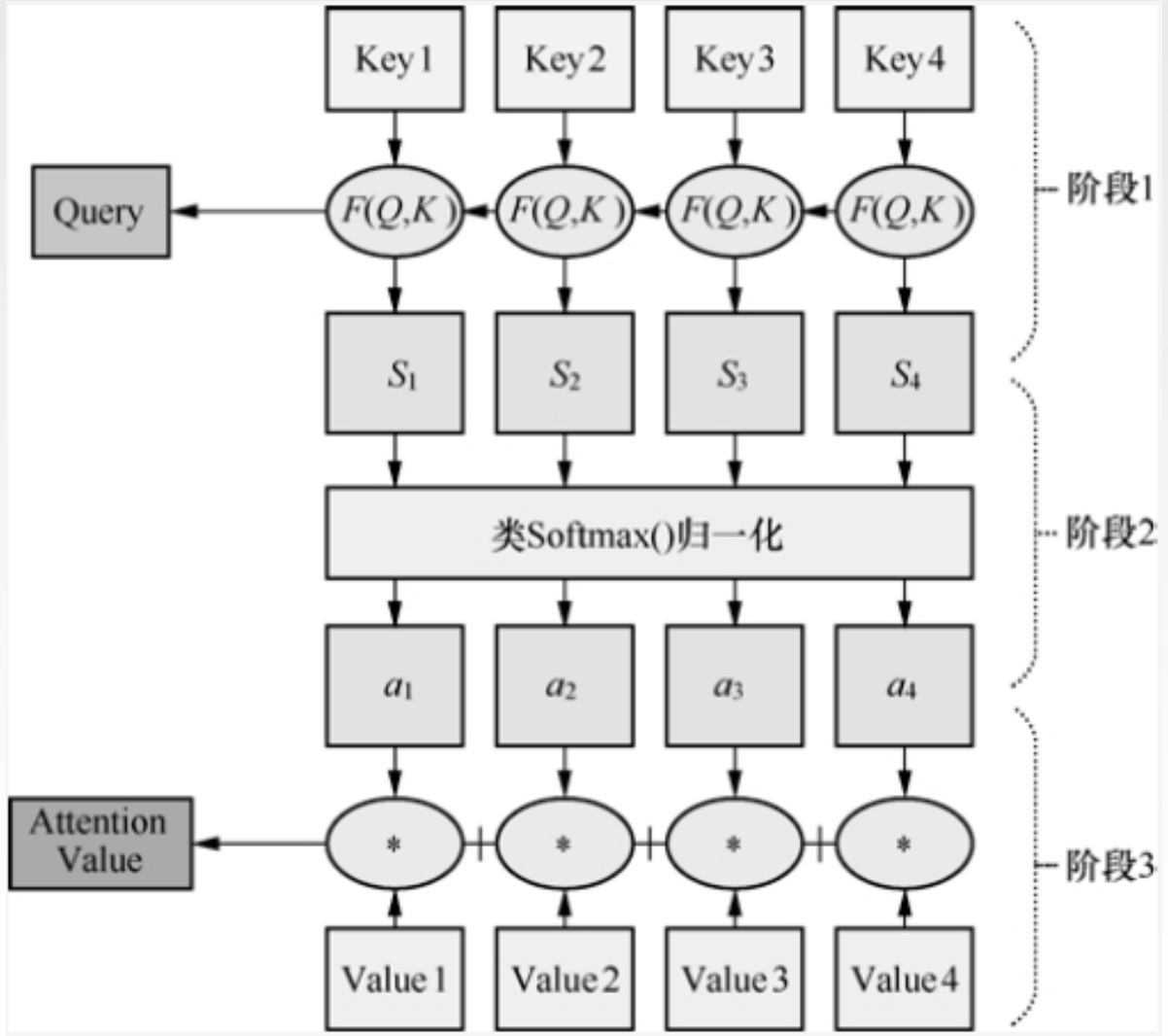
Attention机制本质上可以被描述为一个查询（Query）到一序列对（键Key/值Value）的映射过程。在计算Attention时，主要分为三步，如图所示。

第一步，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。

第二步，使用一个Softmax()函数对这些权重进行归一化。

第三步，将权重和相应的键值Value进行加权求和，得到最后的Attention。

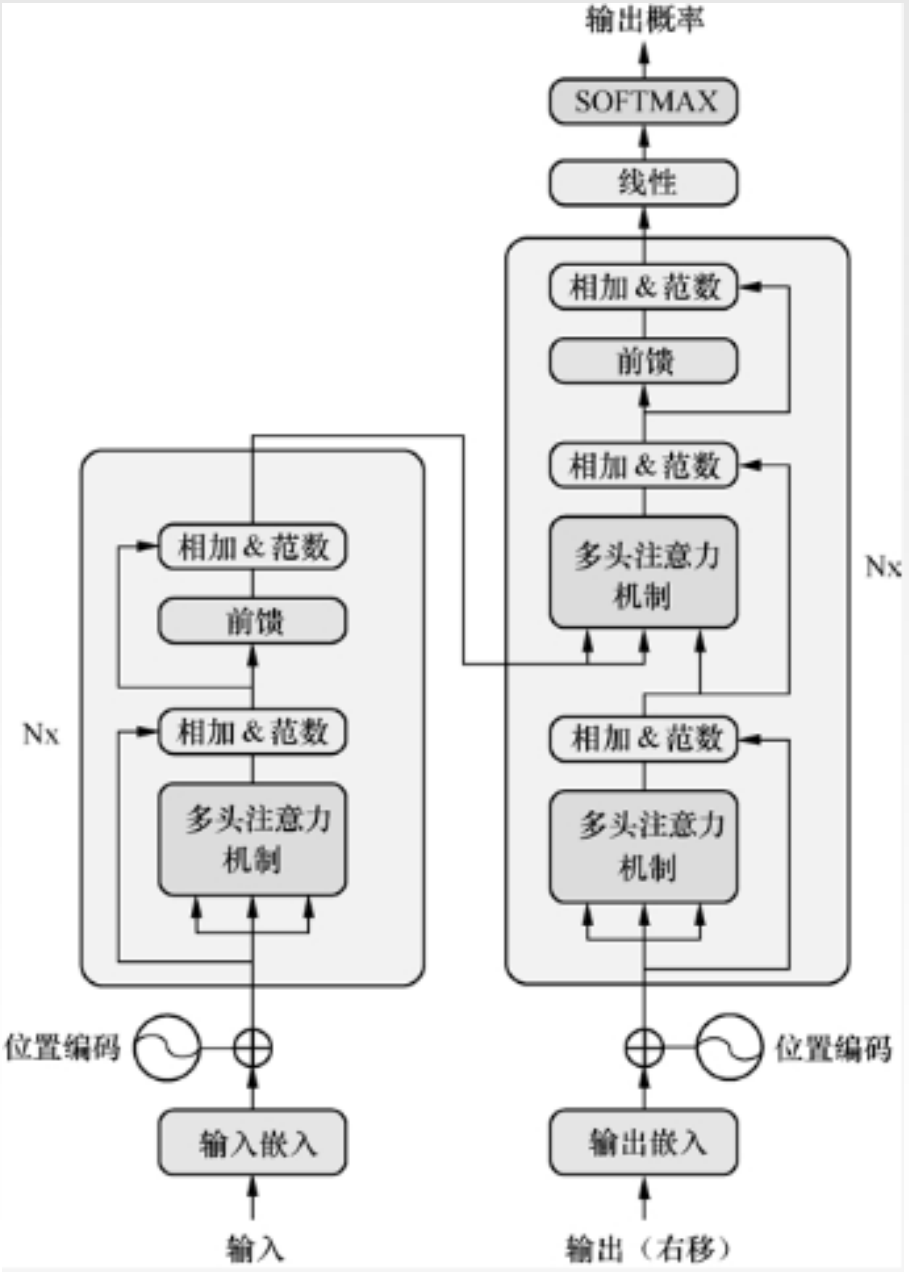
目前，在NLP研究中，Key和Value常常是同一个，即Key=Value。



NLP的深度学习模型和方法

★ Transformer

一个典型的Transformer模型的结构如图所示。左边的结构代表编码器，采用了N=6的重复结构，包含一个Multi-Head Attention和一个Position-wise feed-forward（一次线性变换后用ReLU激活，然后再线性变换）。右边的结构代表解码器，最下面是输出序列的tokens,在翻译任务中就是目标语言的词表，并且第一个Multi-Head Attention是带有Mask的，以消除右侧单词对当前单词Attention的影响，左边的Encoder编码后的输出将会插入右边Decoder的每一层，即Key和Value。



NLP的深度学习模型和方法

★ Transformer

Transformer相比RNN、LSTM等传统递归模型具有如下优点。

(1) 完全的并行计算。Transformer的Attention和feed-forward均可以并行计算，而LSTM则依赖上一时刻，必须串行。

(2)减少对长时记忆的依赖。利用self-attention将每个字之间的距离缩短为1，大大缓解了长距离依赖问题。

(3)提高网络深度。由于大大缓解了长距离依赖梯度衰减问题，Transformer网络可以很深，基于Transformer的网络可以做到20多层，而LSTM一般只有2~4层；根据深度学习的基本思想，网络越深，高阶特征提取能力越强，模型性能越好。

(4)真正的双向网络。Transformer可以同时融合前后位置的信息，而双向LSTM只是简单地将两个方向的结果相加，严格来说，双向LSTM仍然是单向的。

(5) 可解释性强。完全基于Attention的Transformer，可以表达字与字之间的相关关系,可解释性更强。