Attention 注意力机制

鲁老师 2023年5月4日 大约 4 分钟 深度学习 注意力机制

Transformer、BERT等模型在NLP领域取得了突破，其模型主要依赖了注意力机制（Attention Mechanism）。注意力Attention机制被应用到越来越多的地方，那么注意力Attention机制的原理和本质到底是什么？

发展历史

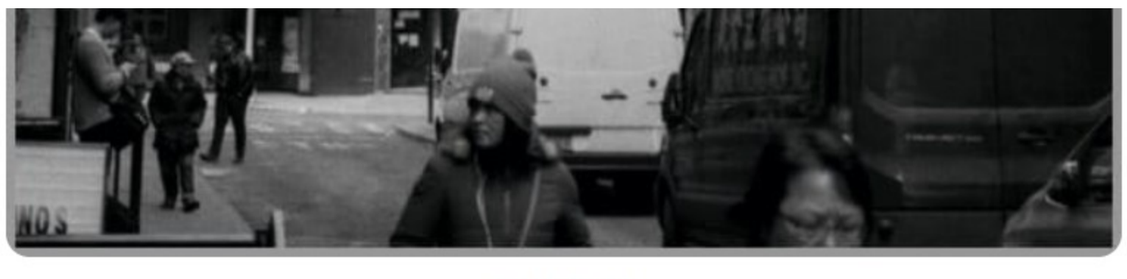
Attention的发展主要经历了两个阶段：

一、2017年前，Attention开始被广泛应用在各类NLP任务上。各种各样的花式Attention被提出，比如用在机器翻译上的Bahdanau Attention等。这一阶段的Attention常常和RNN、CNN结合。

二、2017年之后是Transformer的时代。2017年，Transformer模型被提出，Transformer完全抛弃了RNN结构，突破了RNN无法并行化的缺点。之后BERT使用了Transformer中的Encoder部分。

基本原理

从“Attention”这个名字可以读出，Attention机制主要是对注意力的捕捉。Attention的原理与大脑处理信息有一些相似。比如看到下面这张图，短时间内大脑可能只对图片中的“锦江饭店”有印象，即注意力集中在了“锦江饭店”处。短时间内，大脑可能并没有注意到锦江饭店上面有一串电话号码，下面有几个行人，后面还有“喜运来大酒家”等信息。



原始图片

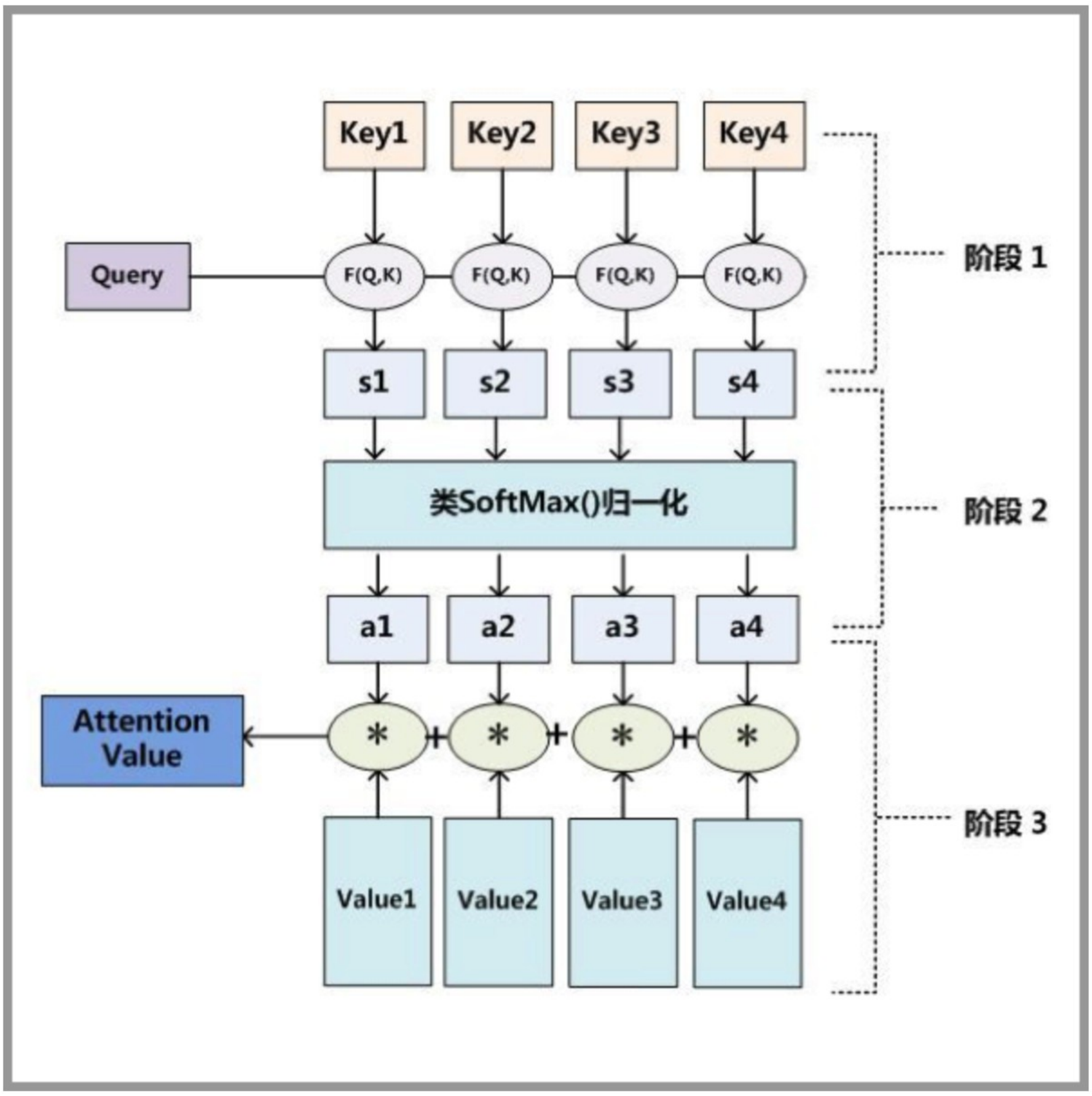
所以，大脑在短时间内处理信息时，主要将图片中最吸引人注意力的部分读出来了，类似下面。

大脑注意力只关注吸引人的部分

Attention机制

Attention的输入由三部分构成： Query、Key和Value。其中， (Key, Value)是具有相互关联的KV对，Query是输入的“问题”，Attention可以将Query转化为与Query最相关的向量表示。

Attention的计算主要分3步，如下图所示。



Attention3步计算过程

第一步：Query和Key进行相似度计算，得到Attention Score；

第二步：对Attention Score进行Softmax归一化，得到权值矩阵；

第三步：权重矩阵与Value进行加权求和计算。

Query、Key和Value的含义是什么呢？我们以刚才大脑读图为例。Value可以理解为人眼视网膜对整张图片信息的原始捕捉，不受“注意力”所影响。我们可以将Value理解为像素级别的信息，那么假设只要一张图片呈现在人眼面前，图片中的像素都会被视网膜捕捉到。Key与Value相关联，Key是图片原始信息所对应的关键性提示信息，比如“锦江饭店”部分是将图片中的原始像素信息抽象为中文文字和牌匾的提示信息。一个中文读者看到这张图片时，读者大脑有意识地向图片获取信息，即发起了一次Query，Query中包含了读者的意图等信息。在一次读图过程中，Query与Key之间计算出Attention Score，得到最具有吸引力的部分，并只对具有吸引力的Value信息进行提取，反馈到大脑中。就像上面的例子中，经过大脑的注意力机制的筛选，一次Query后，大脑只关注“锦江饭店”的牌匾部分。

再以一个搜索引擎的检索为例。使用某个Query去搜索引擎里搜索，搜索引擎里面有好多文章，每个文章的全文可以被理解成Value；文章的关键性信息是标题，可以将标题认为是Key。搜索引擎用Query和那些文章们的标题（Key）进行匹配，看看相似度（计算Attention Score)。我们想得到跟Query相关的知识，于是用这些相似度将检索的文章Value做一个加权和，那么就得到了一个新的信息，新的信息融合了相关性强的文章们，而相关性弱的文章可能被过滤掉。

尽管举了两个例子，但是理解起来还是有些抽象，与具体的代码编写相距甚远，下一篇文章，我们通过Transformer的Self-Attention来详细解读Attention的计算过程。