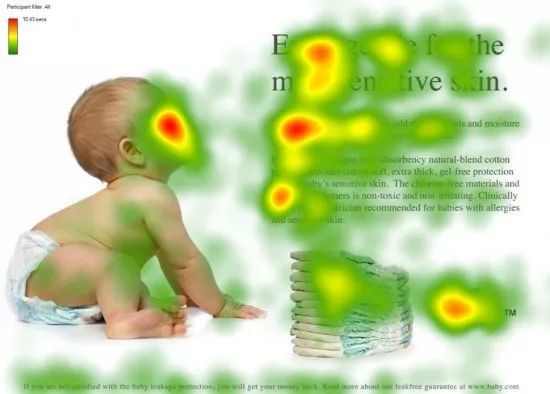
Transformer的研究及其应用

Transformer模型简介

从Attention讲起

人类的视觉注意力

Attention是注意力的意思，字面来看，借鉴了人类的注意力机制。



最为典型的注意力机制是人类的视觉注意力机制，人类视觉可以通过快速全局的图像，获得需要重点关注的最为感兴趣的目标区域，然后对这个区域进行更为细节的观测，获得更为详细的信息。

这种注意力机制也就是说让人类能够从全幅的大量信息精选出值得关注的部分细节信息，更够提高视觉信息处理的效率。

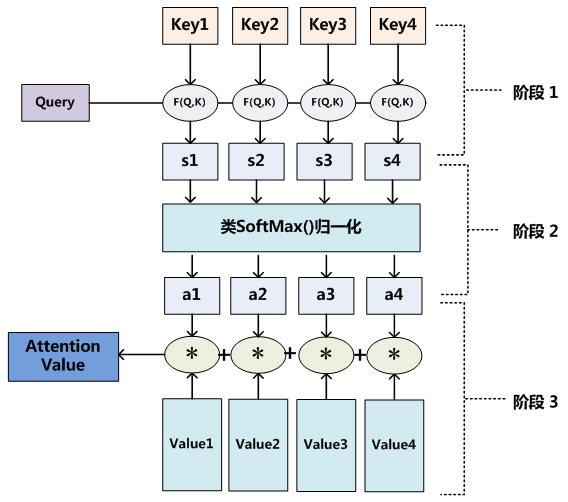
深度学习的注意力信息也和人类选择的视觉注意力信息类似，本质上是从众多信息中选择出较为关键的信息。

Attention的本质思想

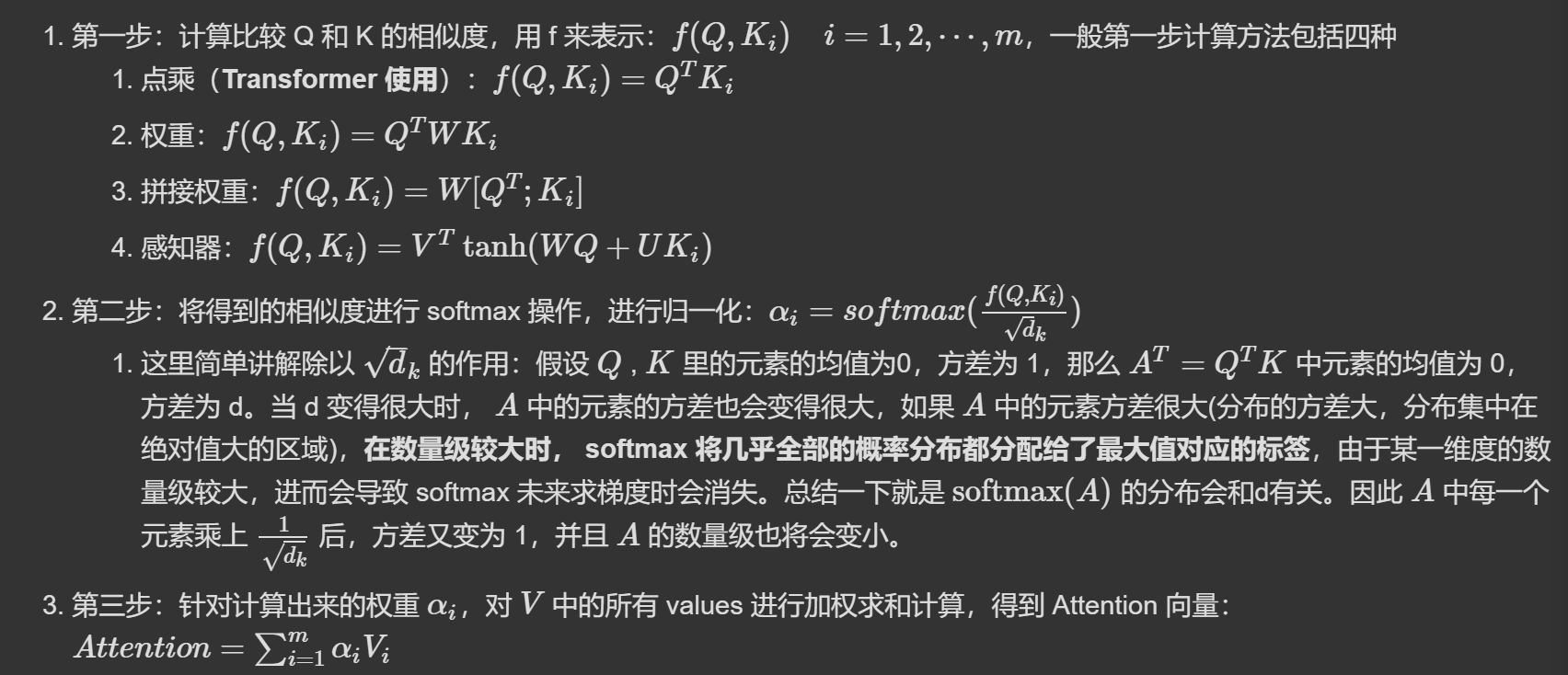
Attention能够从众多信息中找到更为有价值和重要的部分，例如LSTM解决了nlp中序列长距离的依赖问题，但是词汇超过200的时候就会失效，通过attention机制可以较好解决这样的问题。

首先注意力模型从大量信息的values中筛选出重要的信息，这种重要的信息也是相对于query而言， 比如在上文的婴儿图，query就是观察图像的人，也就是说一个注意力模型，有一个 query和一个values， 然后query从values里面筛选出重要的信息

通过query从这个values中筛选出重要的信息，也就是计算出query和values中信息的相关程度

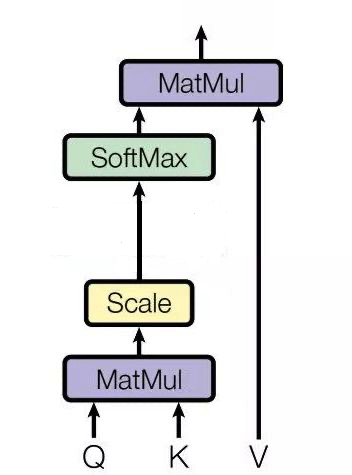


具体来说Attention可以由上图所示，将Query(Q)， 和key-value pairs映射到输出上，其中每一个query， key， value 输出是V中所有values的加权，其中权重是Query和每个 key计算出



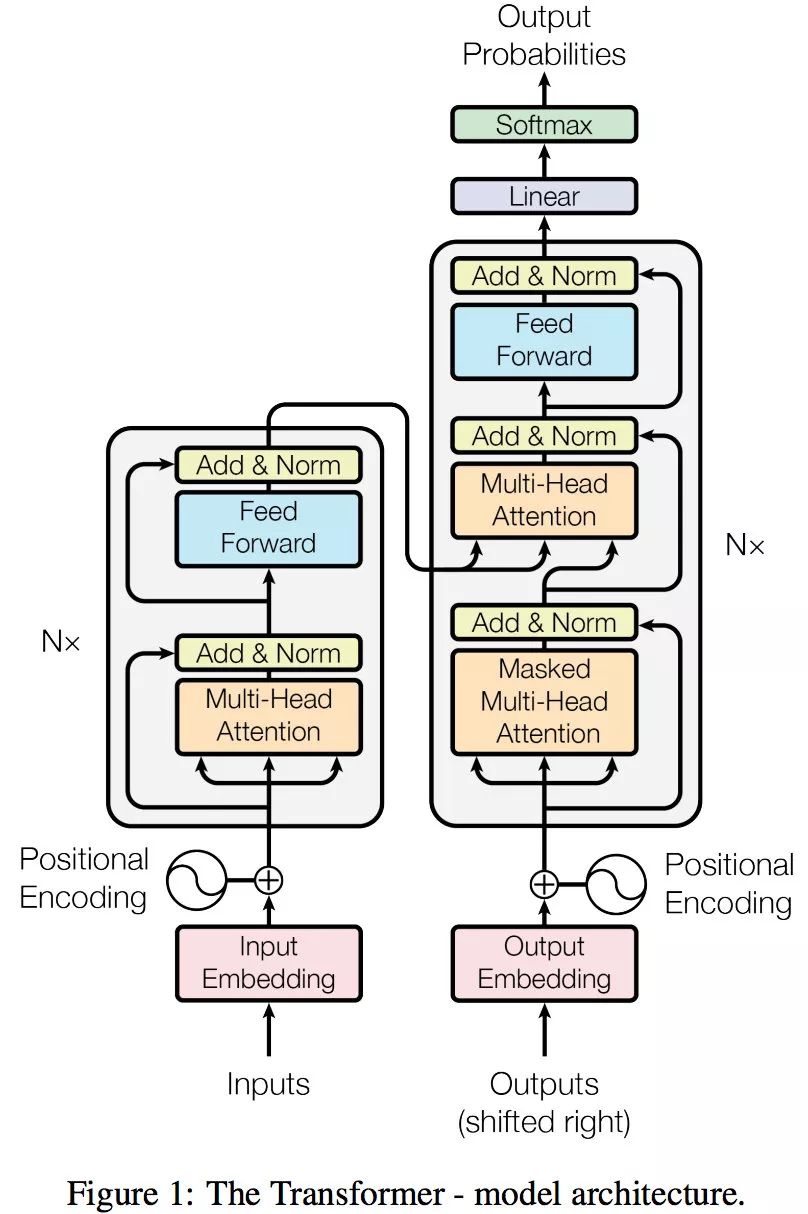
 Self Attention 模型

Self-Attention模型来详细说明如何找到这些重要的信息

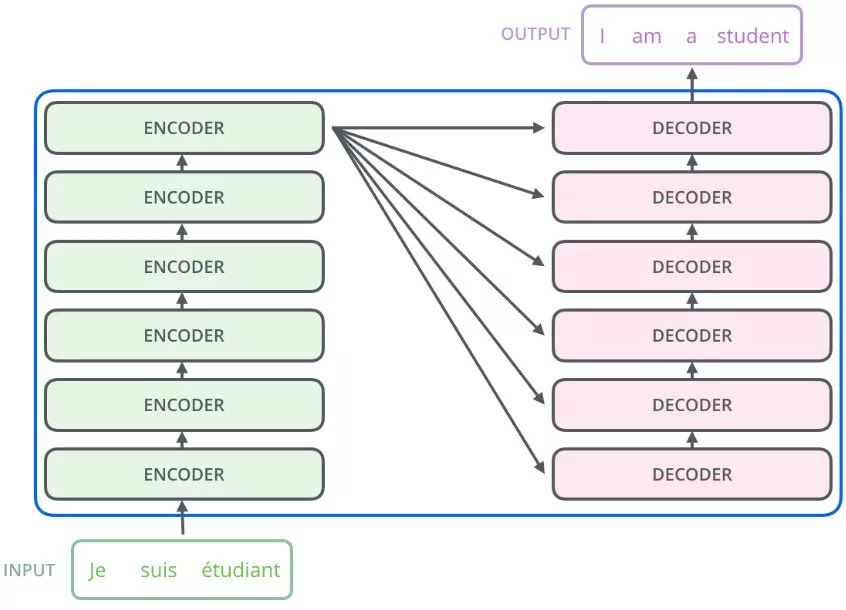


Self Attention模型如上图所示

Transformer



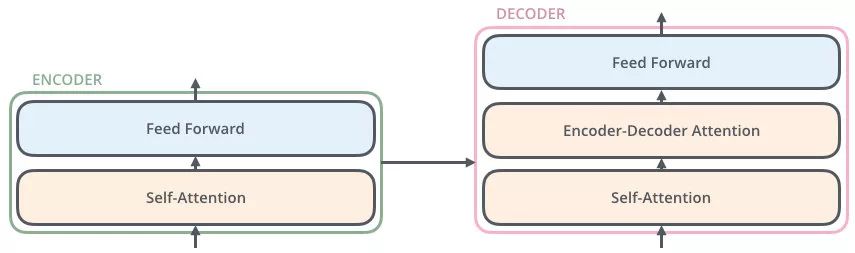
再往细里讲，Transformer 也是一个 Seq2Seq 模型（Encoder-Decoder 框架的模型），左边一个 Encoders 把输入读进去，右边一个 Decoders 得到输出，如下所示：



随之产生的问题是encoder和decoder是如何结合的，实际上里面有n层的encoder，encoder的输出会和Decoder进行结合

具体的某一层如图所示

Encoder

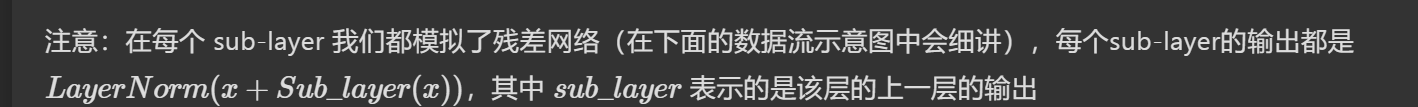


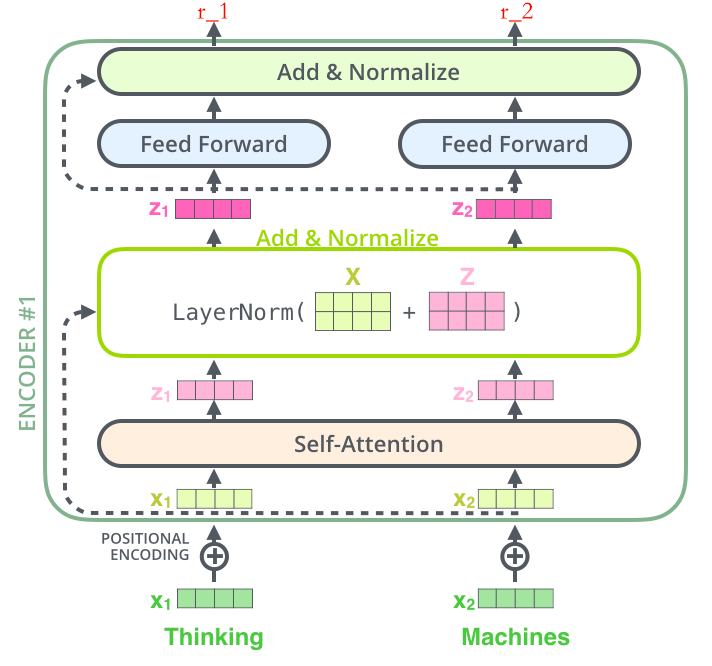
Encoder是6层， 每层包括 2个sub-layers

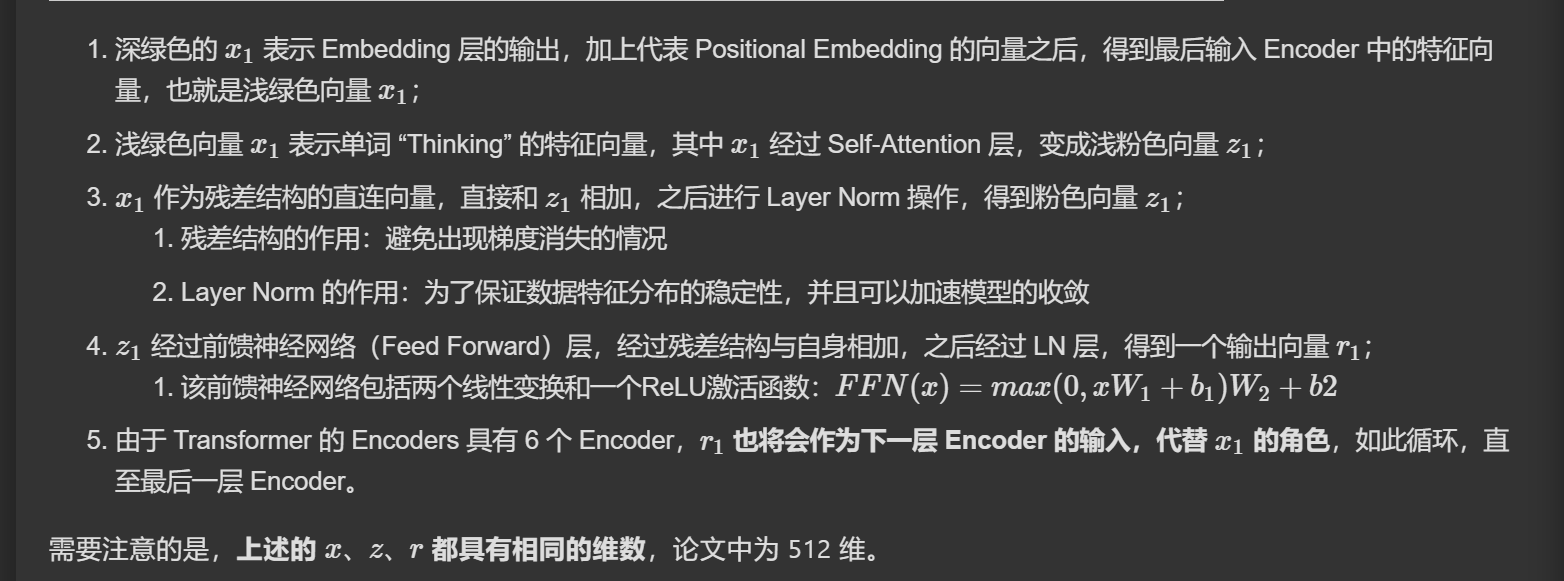
第一个sub-layer是 multi-head self-attention

第二个sub-layer 是简单的Feed Forward

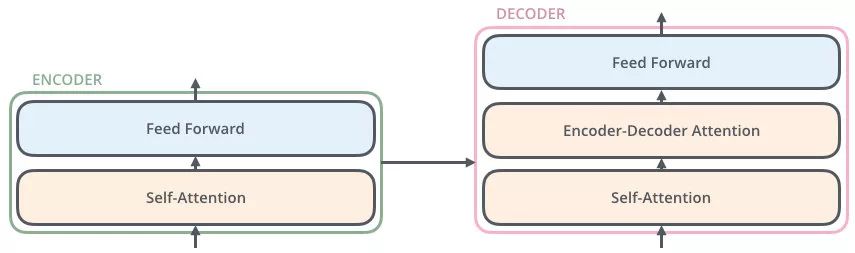
每一个sub-layer都模拟了残差网络







Decoder



Decoder包括3个sub-layer

第一个sub-layer是Masked multi-head self-attention 也是计算的输入的attention

第二个 sub-layer 是Encoder-Decoder attention计算， 对encoder的输入和Decoder的Masked multi-head self-attention的输出进行attention计算

数学推导

相关工作及应用

Transformer 最早提出于