Transformer的心得理解

Attention

说到Transformer,就不得不先说下Attention，Attention我们可以理解为一种权重的思想，比如说我们某时某刻，有一些重要的事情在大脑中占据着很重要的部分，我们愿意花费多一些精力在上面，所以这就跟Attention的思想差不多，我们愿意为这件事情分配多一些的权重.

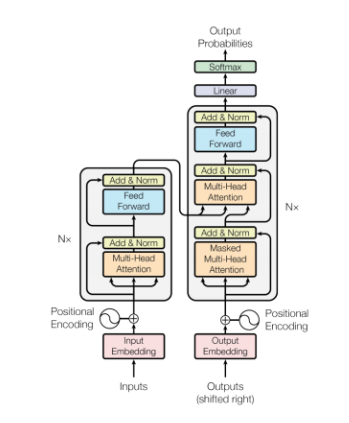
Self-Attention就是对自己的权重分配。它与Attention的区别主要有以下

1. 在神经网络中，attention更多的被应用在输出或者状态层，Self-Attention则更多的关注在input
2. Attention经常应用在连接在两个不同的组件,Self-Attention，则是一个组件
3. SA可以在一个模型中被多次的，独立的使用，AT则经常只是被使用一次
4. SA比较擅长在一个序列当中寻找不同部分的关系，AT更擅长寻找两个序列的关系
5. AT可以连接两种不同的模态，SA则被应用在同一种模态
6. SA结构更加的general,可以用来降维、特征表示、交叉等功能

Muli-head Attention

它是说我们可以从多种角度去思考Attention,也就是我们可以多种特征表示。我们可以通过不同的线性变换对Q,K,V投影，最后将不同的attention拼接起来。代码中采用的将维度512切分成8个64维的计算他们的attention，然后再将它们拼接起来。

Transformer



整体结构

它跟传统的Encoder-Decoder不同，里面引入了self-Attention，同时Decoder部分还引入了Multi-Head attention。为了体现词语之间的顺序关系同时引入了Postional Encoding。

Encoder部分

由6个相同的层，每个层由两个self-Attention和全连接层组成，同时为了防止梯度消失，层之间还采用残差网络和层归一化。（其中层归一化与批归一化不同在于，批归一化是对同一批次数据进行归一化，使其分布一致，但是它对batchsize的大小很敏感，同时需要计算并保存某一层神经网络batch的均值和方差。层归一化则是对特征维度的归一化，它针对的是某一层的所有神经元的归一化，同层神经元输入拥有相同的均值和方差。）输入部分由Embedding和Postional Encoding组成。

Decoder部分

它跟Encoder不同点在于，输入它增加了mask遮罩，这是因为输出我只能通过从前面推测后面的，只允许关注输出序列中较前的位置。

它还多了一个Mult-head Attention。Decoder的Mult-head Attention自带Query矩阵，Key,Value矩阵来自Encoder.

最后将Decoder的输出加上Linear映射到vocab size维度大小的向量，通过softmax将得分转换成对应词概率最大的输出。

优缺点：

1. 层输出可以并行计算，不像RNN需要序列
2. 远距离可以影响彼此输出
3. 可以学习长距离依赖。
4. 不对数据的时间和空间关系做假设。

缺点

如果输入数据有时间或者空间关系，必须加上位置编码，否则模型会看到一堆单词。