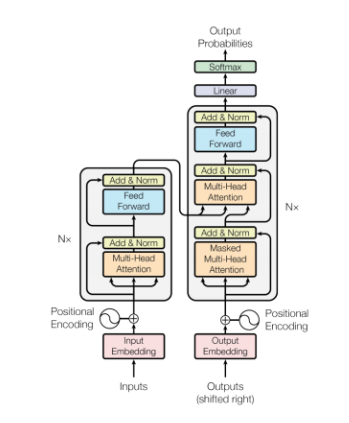
Transformer和Bert以及作业的心得

1. Transfromer

由于word2Vec是一种局部的词向量的生成模型，虽然Glove引入了全局的统计信息，但是它们都没法很好解决一词多义的问题，一词多义它依赖于上下文信息，这就引入了Transfromer。



整体架构

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。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。

Mult-head Attention

它是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。Transfromer中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。

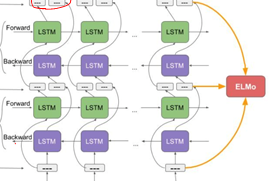
缺点

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。

2、ElMo

它可以很好捕捉语义与语法信息

它的区分一词多义

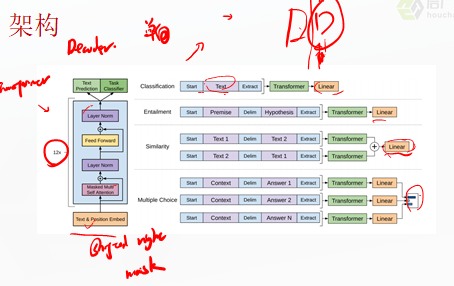


它采用两个单向的LSTM,因为双向会造成信息暴露的问题，由于LSTM

采用的是串行，所以训练成本增大了。LSTM对长距离的提取特征不如Transforner

3、GPT

它的优点在于通过半监督方法处理语言理解任务。为了解决需要大标注数据，通过未标记的数据语言的模型+少量标注数据。



它是单向的Transformer预训练模型。

1. Bert

它是双向的Transformer结构预训练模型，它是无标签的训练的，能够依赖上下文构建向量表示。

它跟GPT,ELMo区别

它是双向的Encoder模型，GPT是单向的，Elmo是融合了从左到右和从右到左独立的LSTM模型去匹配下游任务，除此之外，Bert和GPT同时是fine-tuning方法，而Elmo只是基于特征的方法

Bert输入包括token,segment，positionEmbedding,句子与句子之间采用sep分割，CLS作为起始标记。

它采用两个任务MASKed LM（词级别）和NSP（句级别）来作为预训练任务。MASKed LM会随机用15%的词遮罩，这就会造成与下游任务的不匹配情况。为了推理句子之间的关系，它50%的句子是下一个句子，其他的来自语料库的随机句子。

Transformer的自我注意机制允许Bert通过交换适当的输入和输出模拟下游任务（decoder中的多头attention的输入部分来源于encoder）。当我们预训练好的模型，将所有参数一起转移到下游任务，然后通过少量的数据进行微调。这就是Fine-tuning。

关于作业

知道怎么通过Bert获取一些信息，但是对于拿来与下游任务进行fine-tuning还是有些陌生，这需要多看一下项目学习，同时如何微调的经验也是需要积累的。