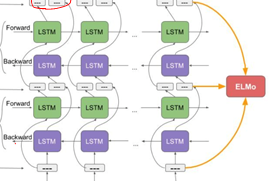
Bert的心得理解

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以及词的多重特性等问题，从而引入了下面诸多模型

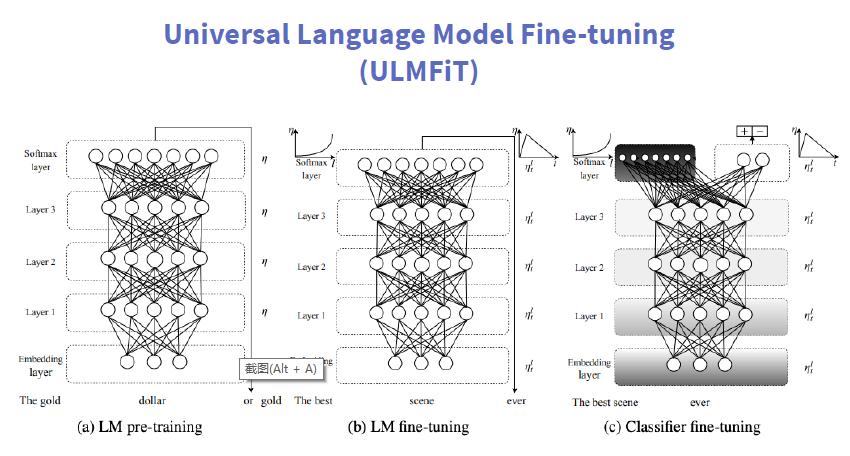
Elmo



结构

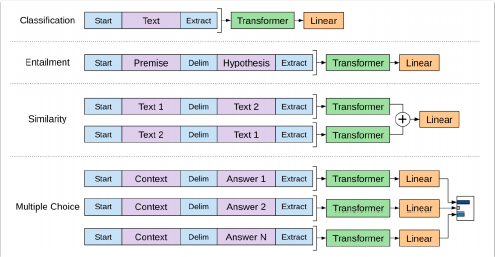
它是一种双向的LSTM结构，同时也由于双向会造成信息暴露的问题，LSTM采用的是串行，所以训练成本增大了。

ULMFiT



它引入了一种预训练的思想，为下面的Bert奠定了坚实的基础。首先它先从大量的预料集中学习一些重要的信息，然后针对任务的预料集进行微调的，最后将模型应用我们要做的任务，最后再进行微调。

GPT



结构

它是一种预训练的Decoder的语言模型，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。

Bert

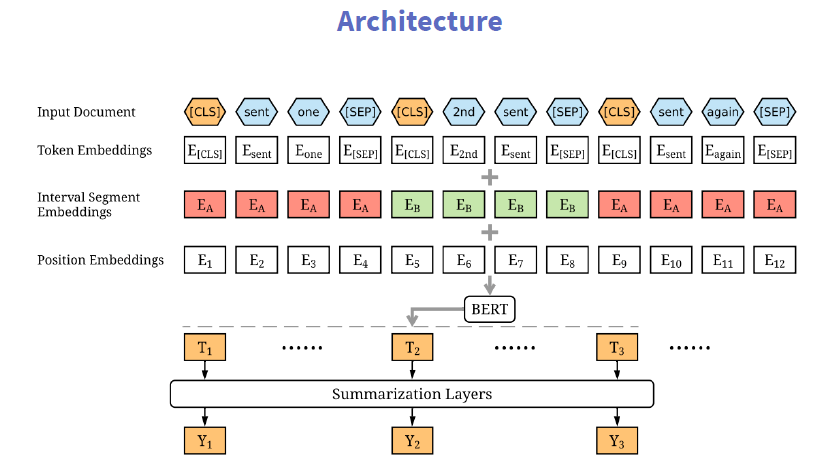
它是双向的Transformer结构预训练模型，它是无标签的训练的，能够依赖上下文构建向量表示。它的目的是为了处理大量的无标签的文本，一词多义，考虑全局的内容，长距离依赖，迁移学习等思想，它采用MASK的思想来学习词的含义，通过随机mask，然后预测mask来学习词的含义的预训练任务。与此同时，采用的15%的token mask，为了防止任务过于简单，增加噪音的目的，在15%的里面又分别进行了80%、10%、10%的处理。它采用了wordpieces代替了words。一方面是为了减少词汇表的大小，增大每个词的表达含义，同时也能防止OOV问题，但是它的问题在于采用这种分词方式mask的部分过于容易被预测，后面采用的是整个词的Masking.

它的第二个预训练任务是NSP,通过预测下一个句子是否为正确的下一句来学习句子的特性。

它的输入包括token,segment，positionEmbedding,句子与句子之间采用sep分割，CLS作为起始标记。在经过bert的结构后，cls位就能代表整句的信息，这是因为下游任务中，我们拿cls得到的信息与对应标签进行loss处理 ，然后反向传播进行的微调参数的过程。

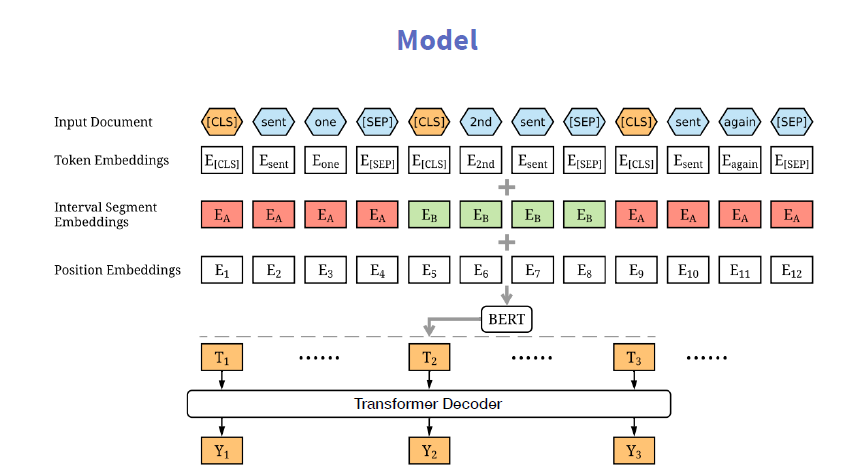
但是由于它的输入长度最高为512，当我们超过512位输入，要么截断操作，会造成一部分信息的丢失，为了解决这个问题，我们可以有一方面可以采用一些压缩的方法从输入文本中选择一部分重要信息作为输入，例如句头，末尾等等。同时也可采用Longformer模型，它里面采用了4096token代替了512位。也可以采用将长的句子切成一段一段，然后将每段都进行训练，得到表达方式，然后进行拼接。也可以采用Reformer结构替代transformer

它可以用做BertSumExt和BertSumAbs等任务



BertSumExt结构

它将bert得到的信息后面接一个Summarization Layers



BertSumAbs结构

它是将Bert得到的信息后面接一个Decoder的部分。