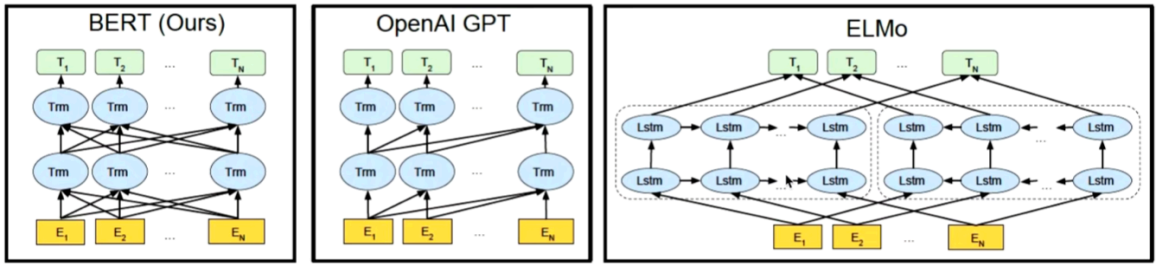
第八节 BERT和GPT（transformer的衍生，集大成者）

我们可以看到BERT、GPT和ELMO是非常类似的，这也是我们要学习前面的基础的原因，如果直接看BERT或者GPT，大家可能大概率是知其然而不知其所以然。



BERT和GPT其实是分别加强了transformer的encoder端（特征提取）和decoder端（生成）

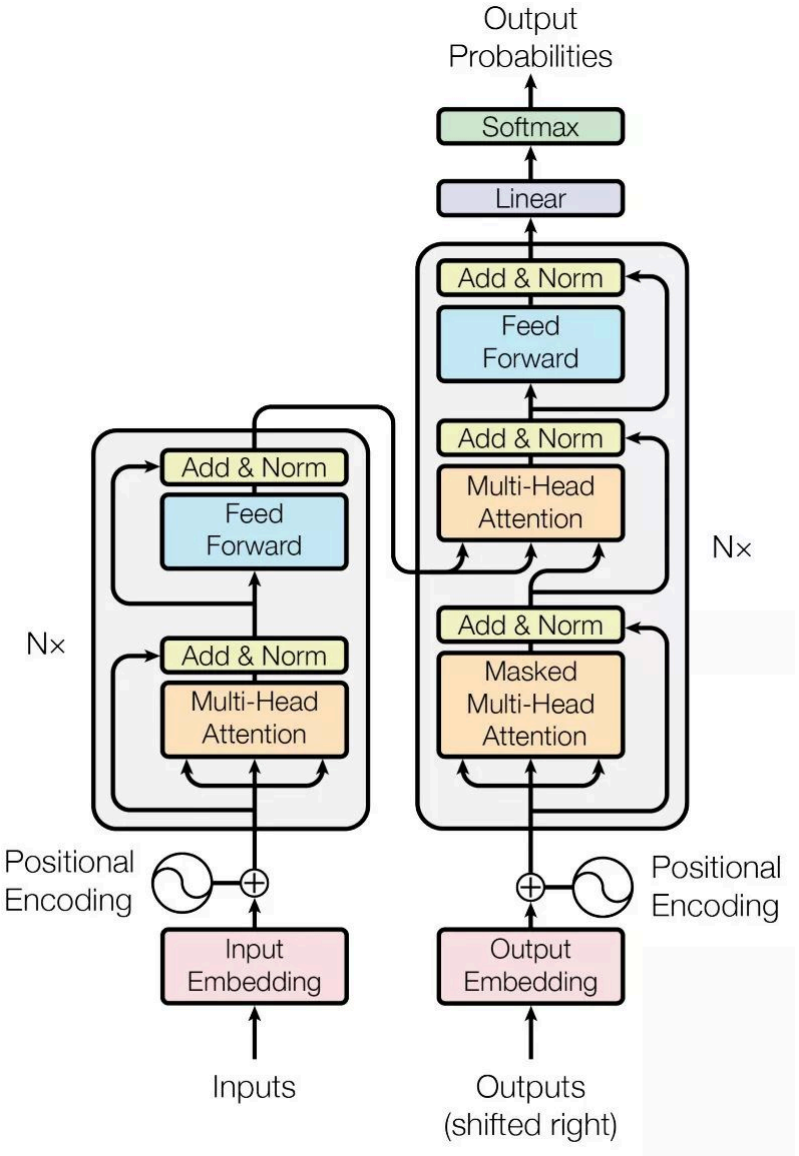


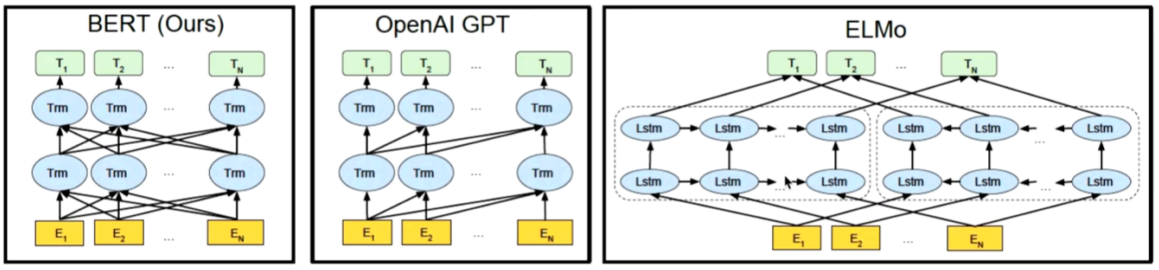
Figure 1: The Transformer - model architecture.

BERT：我在做更强大的编码器

我们之前讲了这么多的词向量模型，它们本质上都在干一件事，用一个更好的向量去表示一个词、一句话甚至一段文章。

那么BERT和ELMO本质上要干的也是同一件事，就是获取更好的词向量，当然BERT的效果要比ELMO好很多了，但是两者的结构我们可以看到是非常类似的。

BERT的大体框架其实只是把ELMO中间的双向LSTM模块换成了更好的transformer当中的encoder端，也就是说BERT其实就是ELMO的加强版，可以更好地提取词、句子甚至文章的特征，最后得到更好的词向量，这样下游任务的表现就会更好（当然BERT实际上做出了很多改进，但是这里不细讲，有兴趣的同学可以去研究一下BERT的论文）



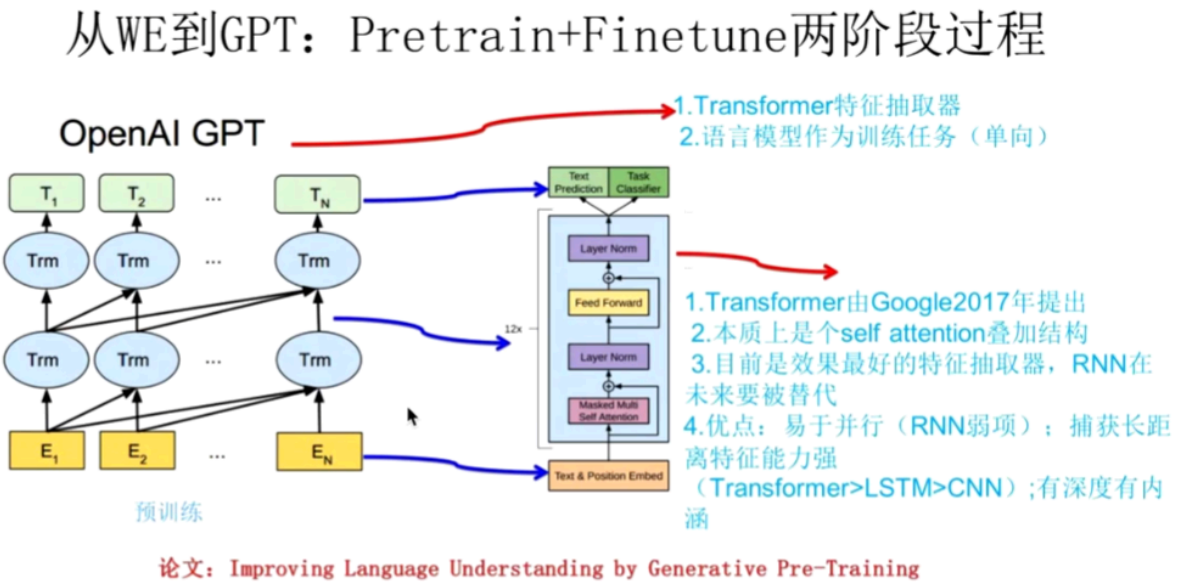
在把LSTM模块换成transformer的encoder之后，它不仅解决了LSTM无法并行和长序列依赖的问题，还真正意义上地解决了一词多义的问题。之前的ELMO其实是有点强行地把对上文的特征和对下文的特征拼在了一块，这其实并没有能够让模型能够同时获取一个词的上下文，有点事后诸葛亮的感觉。

但是transformer不一样，大家回想一下self-attention，里面的运算过程里面其实是可以同时获取一个词的上文和下文信息的，因此BERT其实真正意义上地解决了一词多义的问题。

当然上面图里面的BERT其实还是比较简单的，只有两层，实际上的BERT一般会有很多层，所以BERT可以学习到很多复杂的特征，也就可以更好地去表示一个词、或者一个句子甚至一篇文章。

而且大家可以充分发挥想象力，对于一个词我们可以提取特征，那么对于一张图片、一个图我们是不是一样可以去提取特征啊，这也就是vision-transformer的由来，现在的视觉特征提取器一般都用ViT来做了，也可以说是视觉领域的BERT。

GPT：我要更好地完成生成的任务



讲完BERT之后，大家再看GPT，其实就很简单了。GPT和BERT的结构其实非常类似，但是两者在目的上是不一样的。

首先BERT的出现是为了更好地提取特征，得到词向量。而GPT的出现，是为了更好地去完成生成的任务，也就是例如翻译、人机对话等等。所以我们可以看到，BERT得到的词向量我们是可以拿来做不同的下游任务的，但是GPT这里就限制了这个下游任务，它只能用来做生成类的任务。

并且由于GPT的目的是生成任务，它里面的transformer模块用的就不是encoder端，而是decoder端了。