关于StructBert、Roberta、Albert、RetroReader的理解

1. StructBert

StructBert的结构还是保持着bert模型的结构，保留了wordpiece，

绝对位置编码等相关信息。

它从预训练任务角度修改了bert，增加了两个预训练任务，同时去掉了NSP

单词结构的预训练任务

它认为好的语言模型可以学习正确顺序

所以它在token输入的时候打乱了词的顺序，但是词的个数不能过多

太多了，难，噪声多，少了，简单，学不到东西，一般取三个。

句子结构的预训练任务

构造了句对三分类任务

•1/3：上下句，label 1

•1/3：上下句逆序，label 2

•1/3：不同文档的句子，label 0

1. Roberta

它是对bert的优化版本，保持结构不变，整体改动不大

模型规模、算力和数据上：

更长的训练时间

更大bacth size

更多的训练数据

它去掉了NSP任务

NSP局限性在于当我们设置了输入的长度时候，如果AB两个句子长度小于它时候，无法使用，容易造成资源浪费，也会造成输入空间的浪费。如果去掉了NSP.A和B可以分别输入更多预料，学习到更多东西。

动态MASK

为了使得每一轮训练的mask不同，增加模型训练数据，学习更多东西，他会每一轮训练喂给模型的时候随机mask。

文本编码

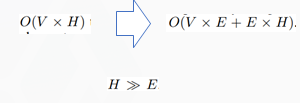
Bert采用字符编码，Roberta采用字节形式的编码，增大了数据量，更多的词表大小。

1. Albert

它是一个轻量级的bert,为了降低模型的参数量,采用了两种参数方式。

矩阵分解

Bert中Embedding维度和hidden的维度大小相同，当我们输入的词汇量大小乘以H就等于词汇量大小乘以E。由于Embedding维度与上下文没有多大关系，H中包含了各种计算attention的，当我们扩大了H维度，E也会增大，所以我们希望E尽可能的小，跟H没有多大关系，将E与H解除绑定信息。所以我们先投影到E一个低维空间，再投影到高维空间，主要是提高H的大小，将模型从窄而深到宽而浅。



参数共享

全连接层层数共享

Attention的参数共享

SOP任务

SOP补偿了一部分因embedding 和 FFN 共享而损失的性能

NSP由于训练时候要同时学习主题预测，和通顺预测，比较难，它

将负样本换成了同一篇文章中的两个逆序的句子，进而消除了主题预测。

N-gram mask

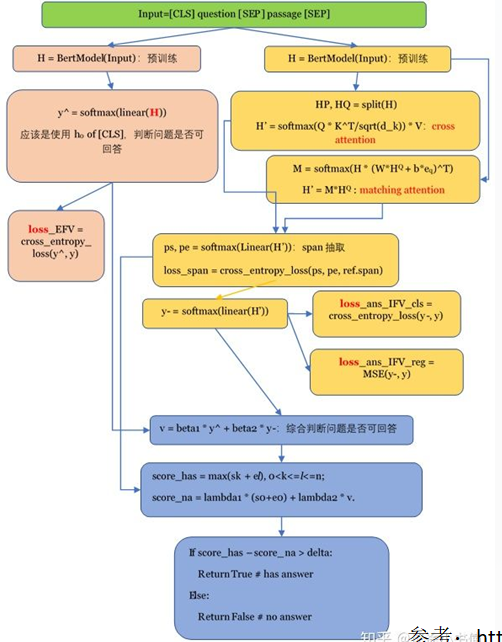
因为词与词之间存在某种关系，当我们预测n-gram片段的mask,它会包含更完整语义信息。n最大取3，其中每一个分的概率由下面公式求得，其中越长概率越小。

1607845746(1)

1. RetroReader

它是一个回顾式阅读器

它的架构的提出是基于前人的各种实验提出的，是一个两阶段的阅读加上答案验证。



整体结构

泛读器

它先是一个Bert模型求得的H向量，然后经过线性softmax得到y^用来回答问题是否可回答

精读器

它也先是一个Bert模型求得的H向量，但是它会将H根据passage和question拆分成HP,HQ，让后用他们来通过cross-attention和Matching Attention得到H’。对H’经过线性和softmax得到ps,pe,然后经过交叉熵损失得到y-。

然后将精度和粗度的结果y-，y^经过一个beta1,beta2组合得到v用来回答问题是否可回答。精度器得到的ps.pe通过求取max来得到有答案的概率。用v和s0,e0来求得无答案的概率，有无答案概率相减与阈值比较就会得到有无答案。