Albert

Albert 是 Google 发布的轻量级BERT模型，使用减少参数的技术，允许大规模的配置，克服以前的内存限制。

BERT(Bidirectional Encoder Representations fromTransformers)：

使用MLM预训练模型进行深度双向训练

1.预训练
在未标注的数据集上进行预训练,使用与训练参数初始化BERT模型
2.微调
用来自下游的标记数据对所有参数进行微调。每个下游数据都有单独的微调模型，但他们使用相同的预训练参数初始化的

原理

Albert 基于BERT的改进

1. **因式分解 (Factorized embedding parameterization)**

> 在BERT模型中，WordPiece Emebeding的大小和hidden size的大小一致。hidden size,指的是transformer的encoder中的hidden size。即E $\equiv$ H对于模型来时WordPiece Emebeding是学习上下文无关的表示，而hidden size则学习的是上下文有关的表示，后者明显更加复杂。因此正常情况下H应该是大于E的。但是随着H变大，正常情况下词汇表V是非常大的，如果E随着H增大，embedding matrix就会非常大。文中提出了对Encoder层进行因式分解，打破了E和H之间的关系。具体的做法是将embedding matrix分解成了两个矩阵V*E和*EH，这样当H非常大的时候可以有效地降低embedding matrix的计算量。在代码中的体现则是在Encoder层中增加了embedding_hidden_mapping_in层，即增加了E*H的矩阵。使得Embedding层和Encoder层解除绑定。

2. **跨层参数共享 (Cross-layer parameter sharing)**

3. **句间一致loss (Inter-sentence coherence loss)**

> 用SOP替换NOP
> NOP：下一句预测， 正样本=上下相邻的2个句子，负样本=随机2个句子
> SOP：句子顺序预测，正样本=正常顺序的2个相邻句子，负样本=调换顺序的2个相邻句子
> NOP任务过于简单，只要模型发现两个句子的主题不一样就行了，所以SOP预测任务能够让模型学习到更多的信息

ref

https://github.com/google-research/albert

Albert paper

BERT paper