## 主要改进

1. 对词向量矩阵进行分解，减少参数数量
2. 参数共享，对Encoder内部的全连接层和Attention层都进行参数共享
3. NSP问题新的训练方式
4. 去除Dropout的尝试

## 词向量矩阵分解

1. 假如词表大小为N，词向量维度为H，同时H也是Attention层输出的隐藏层变量的维度，那么词向量矩阵维度就是(N,H)
2. 论文中提及的最优维度H在1024左右
3. 分解方法：
   1. 将词映射到一个低纬度的空间，就得到维度比较小的词向量E(E<H)，这时词向量矩阵维度就是(N,E)
   2. 同时设置一个转换矩阵(E,H)，用于将低维空间中的词向量映射到高维空间
   3. 最终输入到模型中的词向量就是
4. 参数个数对比：
   1. 原来参数个数：N\*H
   2. 分解后参数个数：N\*E + E\*H

## 参数共享

1. Transformer组成主要是Attention层和FeedForward层，已经有论文研究了在这两方面进行参数共享的效果，ALBERT默认是对这两种层都进行参数共享。
2. 实验结果显示，参数共享有利于稳定网络中的参数，加快收敛

## NSP训练方式

1. BERT中使用输出中的NSP位来判断两个句子是否相关，并且计算相应loss，与masked language modeling (MLM) loss进行融合作为最终的loss
2. BERT训练Next-sentence prediction (NSP)时的正样本是来自于同一篇稳定中连续的两个句子，负样本是来自不同文档中的句子
3. BERT训练NSP主要用于解决下游的主题预测与关系一致性预测问题，但是效果并不理想。作者认为是因为仅使用NSP来解决这两种问题太过困难。
4. ALBERT中正样本是来自于同一篇稳定中连续的两个句子，负样本是将正样本中句子顺序颠倒，此时只用其来学习句子关系的一致性，而不学习主题相关性，降低其学习难度，叫做sentence-order prediction(SOP)。
5. 作者认为，主题预测问题相对于关系一致性预测问题更加简单，只学习SOP也能解决主题预测问题。

## 去除Dropout

作者发现在训练100W步之后模型仍然没有过拟合，于是去除了Dropout，下游任务的效果还有了一定的提升。