BERT

①特点：(1)深而窄，中间层1024，共12层
2) MLM

② mask掉15%的词，但用三种
1) 80%、直接mask
2) 10%、随机替换 因为会mask掉的词
3) 10%、保持不变 f-7期间从未看到。

③ NSP：判断语句对是否连续。

④ 模型架构
(1) 使用Transformer原始架构
(2) 输入表示：
- token embedding
- segment embedding
- position embedding：可学习的

添加[CLS]，用于分类任务

(5)关键创新
1) 应用Cloze任务，去噪自编码
这来预测被mask的词。
2) 缺点、1.没有见过被mask掉的词
2.每个batch只预测了15% token.
表明需要更多的步骤才能收敛

3) NSP 下一句预测

(4)优点：① 双向
② Cloze
③ NSP+MLM
④ 通用：输入层输出层相同
⑤ 微调成本小

缺点：① 随机mask
② mask实际预测没有过多影响
③ 收敛慢
④ 消耗巨大

5) 局限：
1) 依赖于具体任务本身
2) 句子预测匹配
3) 深层语义任务
4) 输入长度不是很长

Q 为什么缩放点积

A 维度d较高有较大方差，softmax梯度较小，缩放点积缓解这一问题

Q：点积模型与加性模型

A：复杂度差不多，但点积可以用矩阵乘积，效率更高

Q muti Head 为什么有效

A：更多特征表示空间，类似CNN多通道选择，先通过切头再分别进行scale Dot，可以使维度d不过大，同时缩小attention mask矩阵

Q FFN

A：FFN中：ReLU主要非线性单元

Q weight tying

A：减少参数量

Q GELU

A：随机正则，乘以一个mask，mask依赖于输入（输入）
mask，m服从正态分布
$GELU(x) = xP(X \leq x) = x\Phi(x)$

Q GELU 优点

A：缺乏随机因素 RELU只有0，1，

Q 优点 BERT

A①双向 ②并行化