**Self-supervised Learning**

图示

描述已自动生成

一篇文章x，将文章分为两部分，一部分是x’作为文章的输入，一部分x’’,作为label。

Bert其实就是Transformer的Encoder，作用就是输入一排向量，输出一排向量。

**第一步：Masking Input （训练方式）**

将输入的文字，随机挑选位置进行遮盖。

方式1: 用特殊符号将文字替换 – special token

方式2: 随即把某一个字换为另外一个字

图示

描述已自动生成

“湾”被替换为mask，mask传入BERT获得一个向量，传入一个模型（Linear）中，获得输出，与ground truth进行比较，完成训练。

**第二步：Next Sentence Prediction**

图示

描述已自动生成

加入一个[SEP]，判断S1和S2这两个句子是否可以被连接在一起。

**如何使用BERT**

图示

描述已自动生成

通过Masked token prediction只学会了如何做填空题。但给BERT一些其他的标注的资料，可以做很多下游的任务。 --- Fine tune 微调

总结：先pre-train获得一个bert, 再做fine tune。

**Case1:**

Input: sequence

Output: class

Eg. 分辨一个句子是positive还是negative

图示

描述已自动生成

只有[CLS] 部分是随机初始化参数，别的都是用pre-train的参数。

**Case2:**

Input: sequence  
Output: sequence

Eg. 分辨单词的词性

图示

描述已自动生成

**Case3:**

Input: two sequence

Output: class

Eg. 有前提和假设，判断前提能否推出假设

日程表

描述已自动生成

**Case4:**

QA: 问答系统 – 答案会出现在文章里面

图示

描述已自动生成

输出的s, e是int，直接带着这两个数字去文章里截取这两个数字代表的文字之间的文字，那就是答案。

图形用户界面, 文本, 应用程序

描述已自动生成

训练过程：

日程表

中度可信度描述已自动生成

需要从头开始训练（随机初始化参数）的两个，一个是用于判断答案开始的位置s, 另一个好似判断答案结束的位置e。

**为什么BERT有效果**

图示

描述已自动生成

虽然同一个字，但向量并不相同。

意思词性相近的字，他们之间的距离会越短。

**GPT系列模型**

BERT模型做的任务是填空题和连接句子。

GPT模型做的任务是：预测接下来出现的token是什么

图示

描述已自动生成

模型结构类似Transformer的Decoder，也就是做masked attention

训练过程

图形用户界面, 文本, 应用程序

描述已自动生成