5. 自监督模型self-surpervised model

笔记本： 【课】原理-李宏毅 deep leaning
创建时间： 2023/4/23 18:18 **更新时间：** 2023/4/23 23:33
作者： 1256876216@qq.com

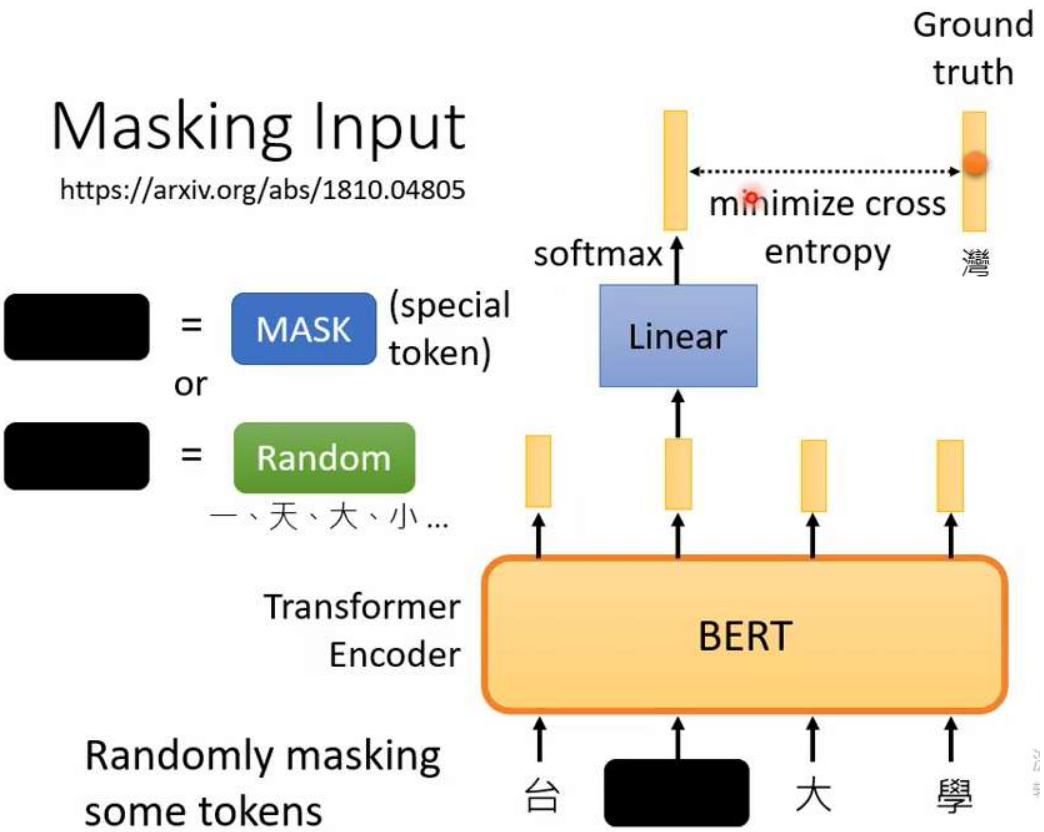
自监督

无标签，将输入分两部分，一部分作为输入，另一部分作为label，让输出接近label

bert

一、Masking Input

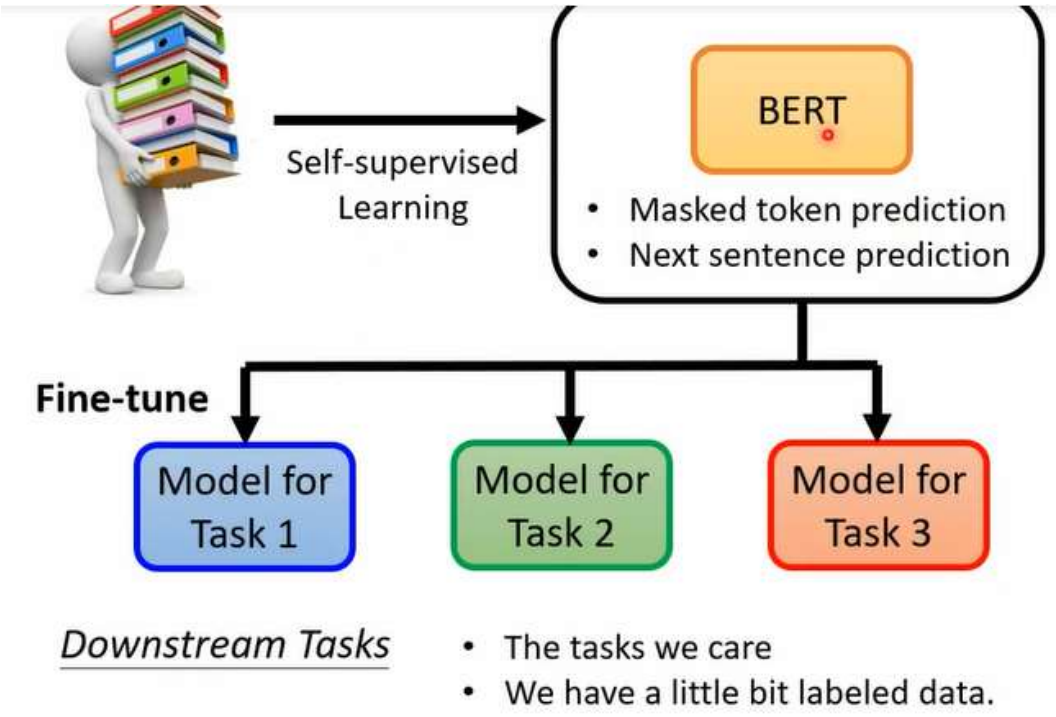
- bert 与Transformer 的Encoder一样，输入向量，输出等长向量
- 应用：处理序列，一般用于文字处理（nlp自然语言处理）、语音、图片
- 处理过程：
    1. 随机决定盖住某些tokens： 使用方法
        - 用特殊符号
        - 或随机的字
    2. 输出另一层，将被盖住的位置进入linear模型 通过softmax（输出多维的概率向量），将得出的结果接近被盖住的字



二、Next Sentence Prediction
判断两个句子，是否应该接在一起

- SOP: Sentence order predient
  - 使用在AlBERT
  - 判断本来一起的两个句子之间的前后顺序分类

三、 应用

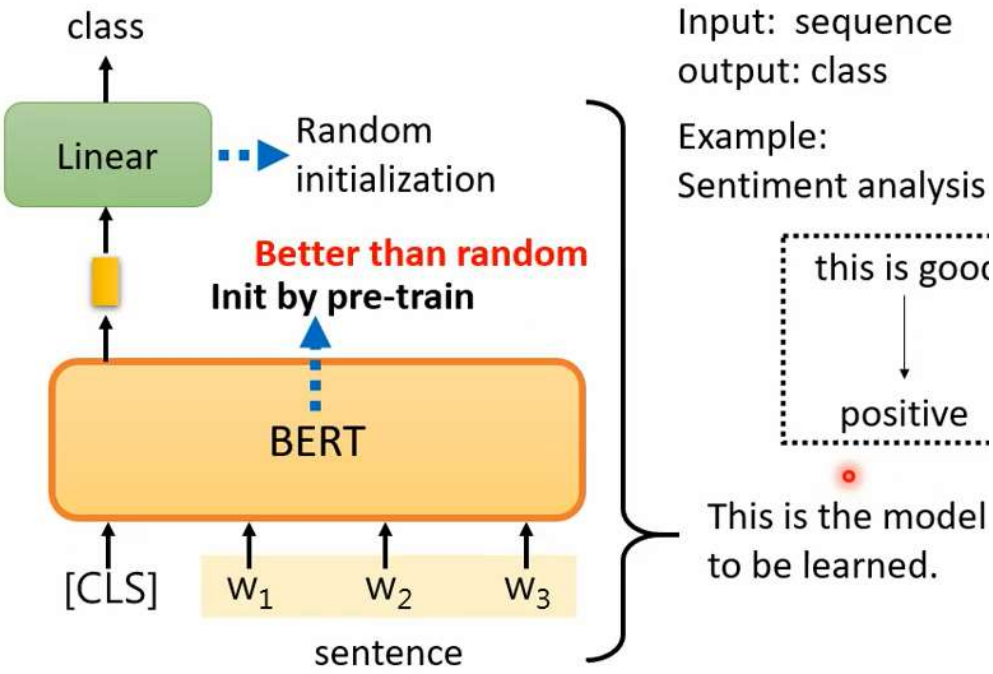


- bert只可以做填空以及下句话的预测，
- 下游任务Downstream：在bert基础上做fine-tune 处理不同的任务
- GLUE：General Language Understanding Evaluation任务级的标准，9个任务。glue也有中文版本 cluebenchmarks

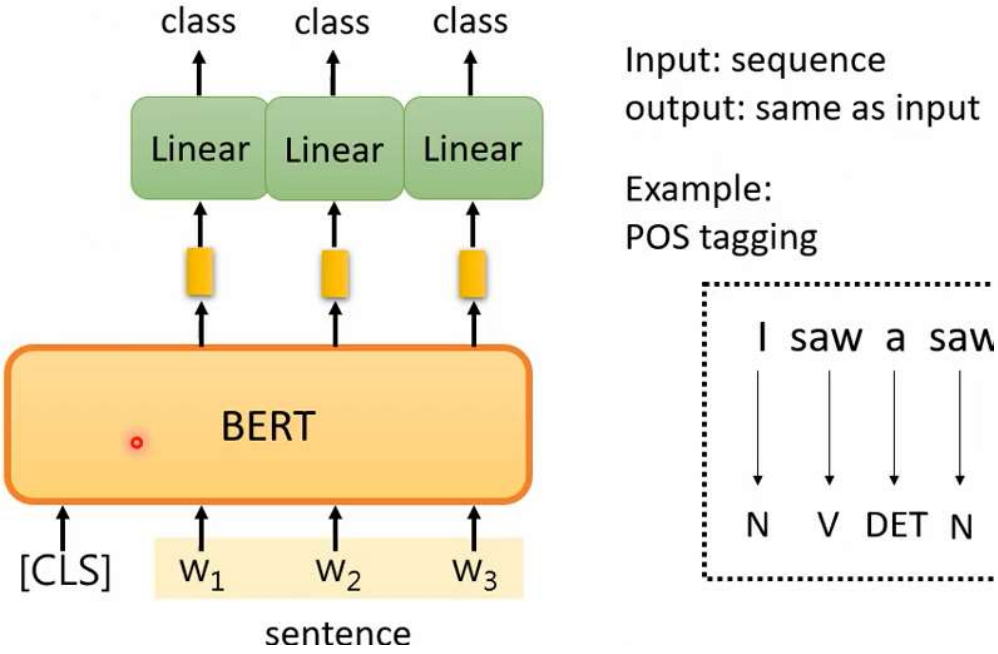
- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST-2)
- Microsoft Research Paraphrase Corpus (MRPC)
- Quora Question Pairs (QQP)
- Semantic Textual Similarity Benchmark (STS-B)
- Multi-Genre Natural Language Inference (MNLI)
- Question-answering NLI (QNLI)
- Recognizing Textual Entailment (RTE)
- Winograd NLI (WNLI)

四、bert使用

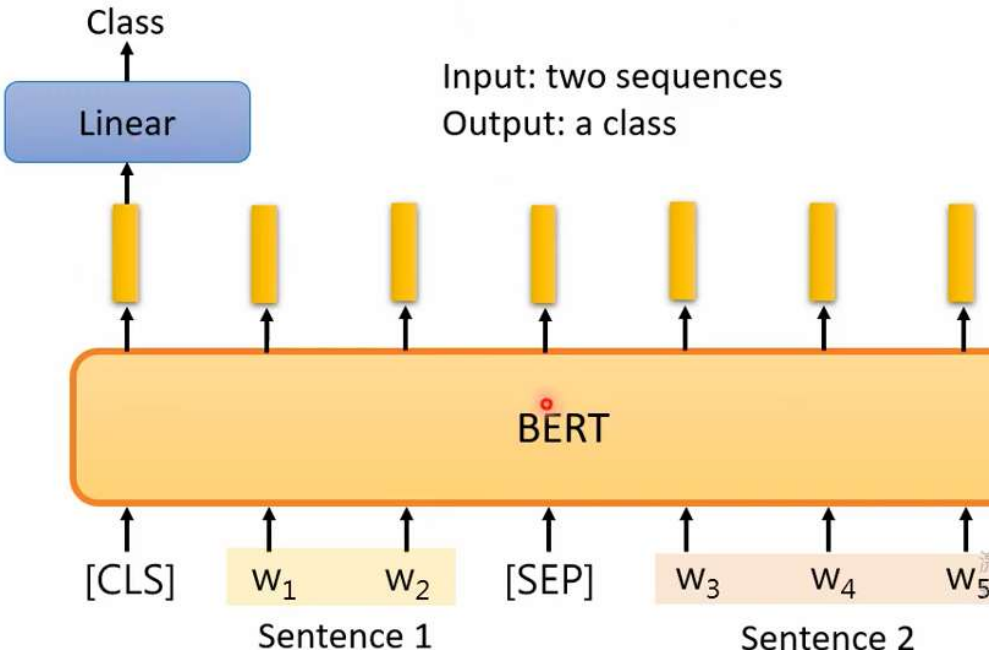
1. 预训练的结果作为初始化参数



2. bert 随机初始化参数



3. 判断两个句子之间的关系

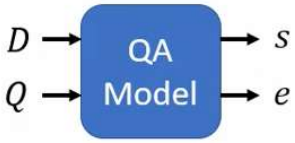


4. 问答系统：输入文章+问题，得到文章中的答案

- Extraction-based Question Answering (QA)

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_M\}$



output: two integers (s, e)

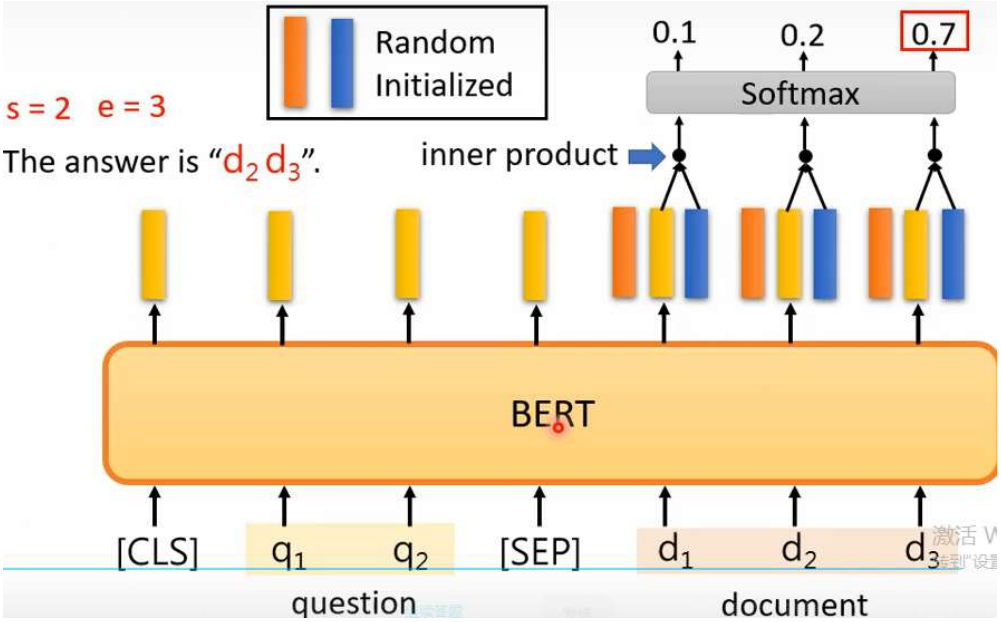
Answer: $A = \{d_s, \cdots, d_e\}$

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud



T5 – Comparison

- Transfer Text-to-Text Transformer (T5)
- Colossal Clean Crawled Corpus (C4)