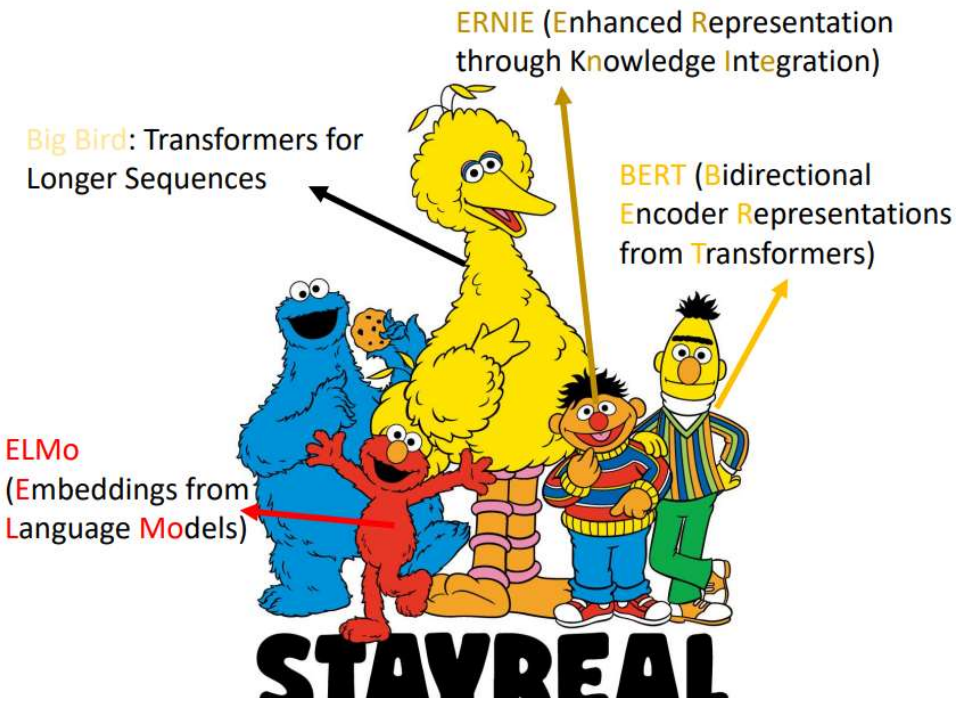
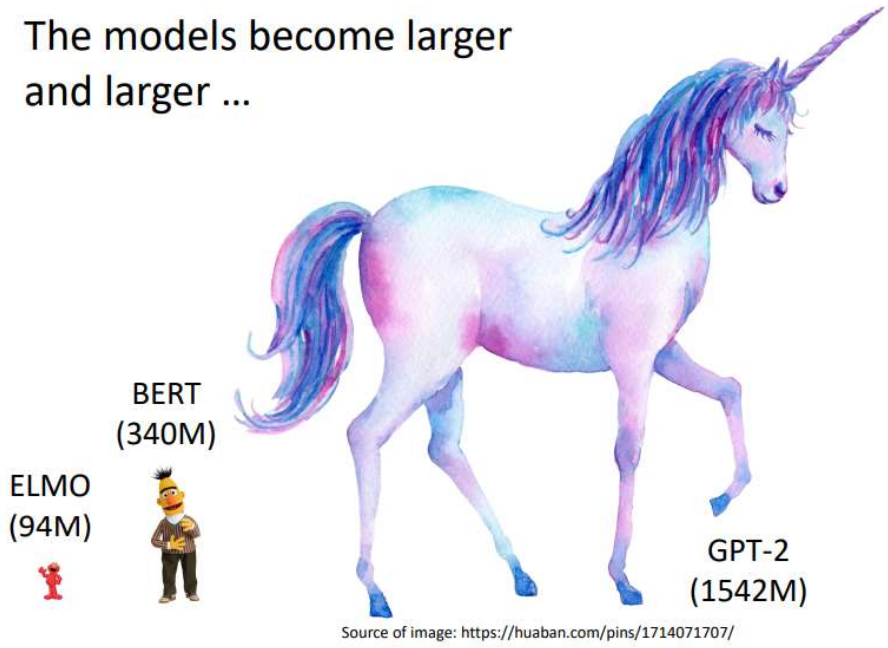
自督導式學習 (Self-supervised Learning) (上)

Create at 2022/06/25

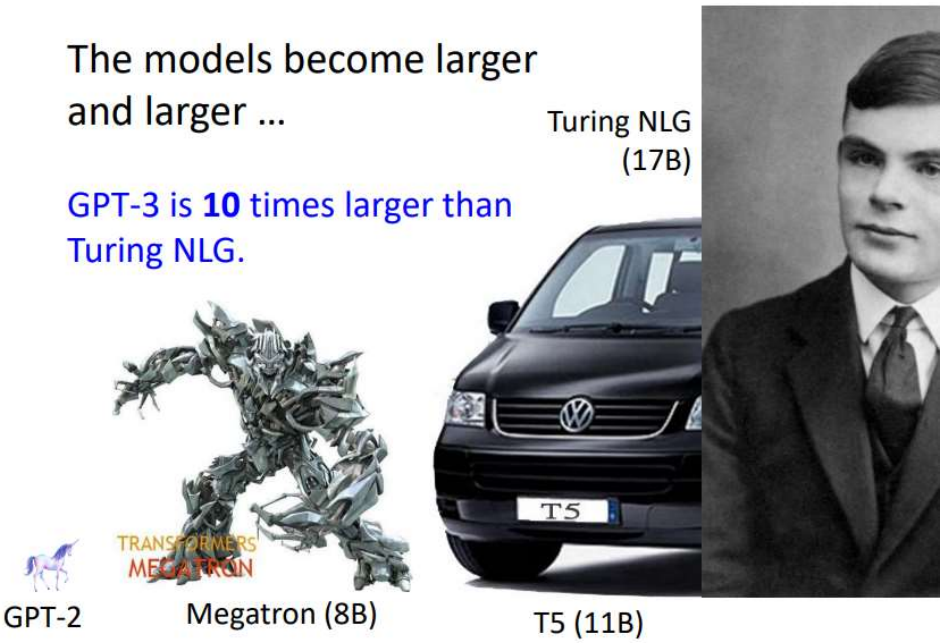
- 上課資源：
  1. 自督導式學習 (Self-supervised Learning) (一) – 芝麻街與進擊的巨人 (https://www.youtube.com/watch?v=e422eIoJ0W4)
  2. 自督導式學習 (Self-supervised Learning) (二) – BERT簡介 (https://www.youtube.com/watch?v=gh0hewYkjgo)
- 參考資料：
  1. BERT and its family - Introduction and Fine-tune (https://www.youtube.com/watch?v=1_gRK9EIQpc)
  2. BERT and its family - ELMo, BERT, GPT, XLNet, MASS, BART, UniLM, ELECTRA, and more (https://www.youtube.com/watch?v=Bywo7m6ySIk)
  3. 來自獵人暗黑大陸的模型 GPT-3 (https://www.youtube.com/watch?v=DOG1L9IvsDY)

芝麻街與進擊的巨人



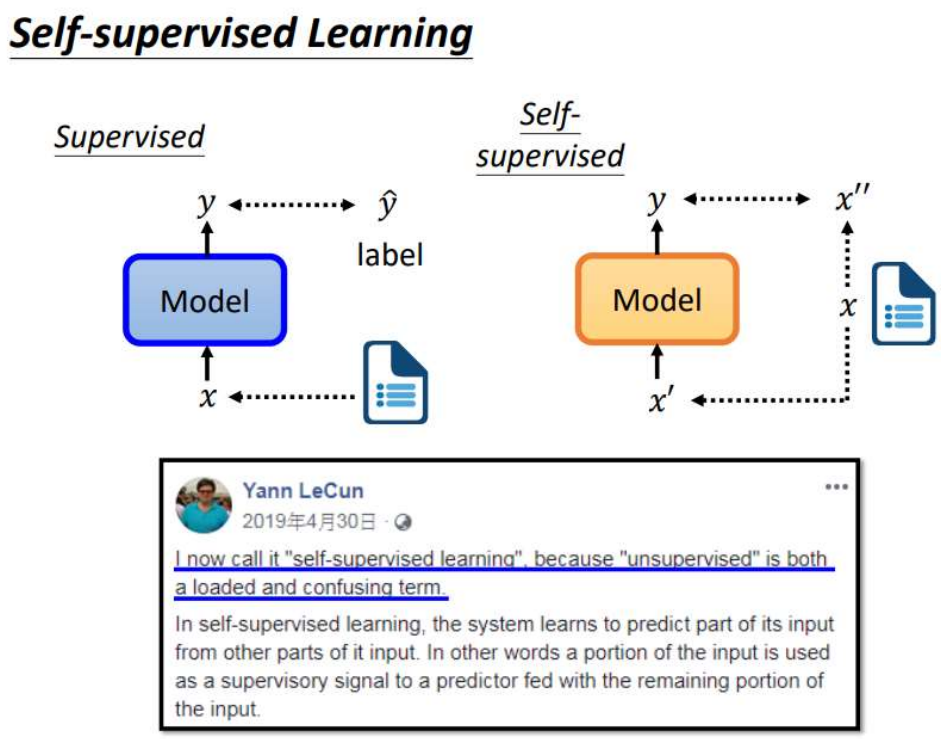


Source of image: https://huaban.com/pins/1714071707/

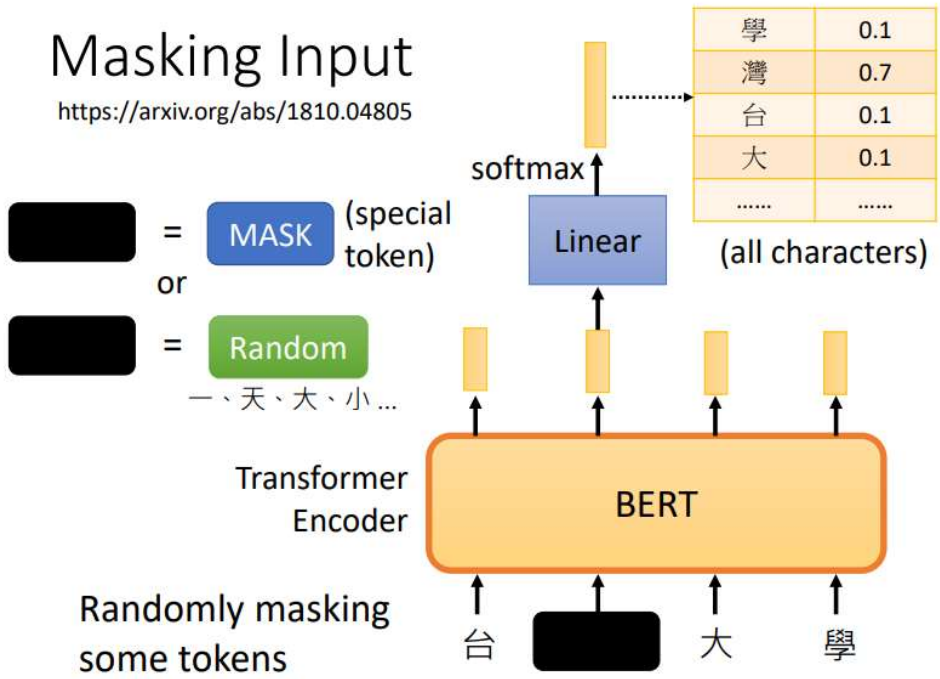


- 這麼巨大的模型到底都在做什麼呢？

BERT簡介

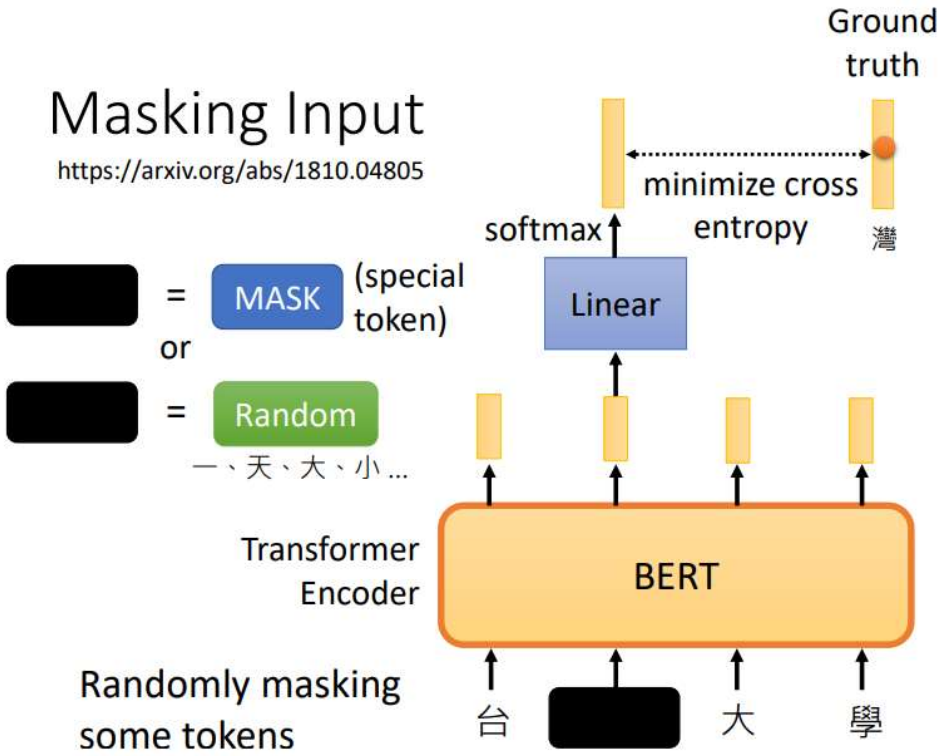


- Supervised Learning
  - 有一個 Model，輸入 x 輸出 y，需要 label $\hat{y}$
- Self-supervised Learning
  - 自己想辦法做 supervised，在沒有 label 的情況下
  - 把 x 分成兩部分 x' 跟 x''
  - 把 x' 輸入到模型裡面，讓模型輸出 y
  - 讓 y 跟 x'' 越接近越好
  - 可以看做是一種 unsupervised learning (沒有 label)

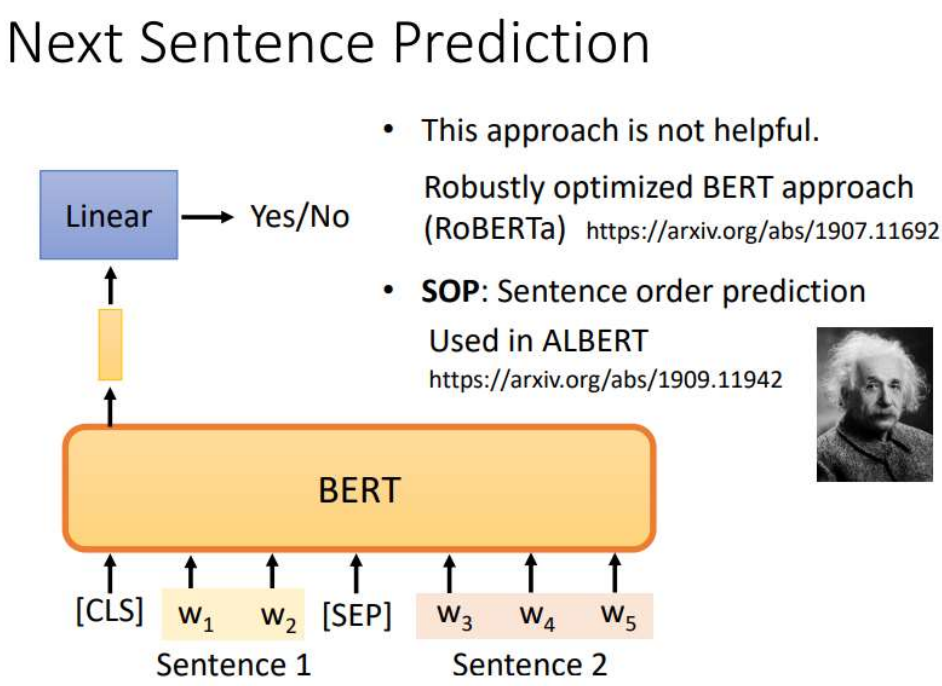


- 直接拿 BERT 來說明這個模型是怎麼運作的
- BERT 這個 model 是一個 Transformer Encoder
  - BERT network 的架構跟 Transformer Encoder 一樣
- BERT 常用在自然語言處理上

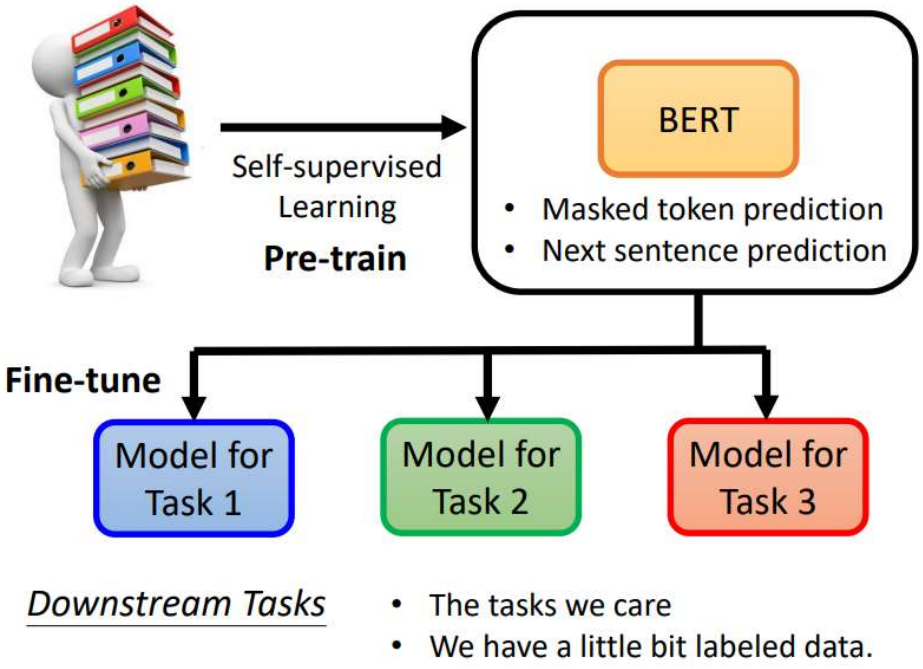
- 把輸入 BERT 的這串文字的其中一些部分，隨機的蓋住
- 蓋住：
  - 把字換成一個特殊的符號 (special token)
  - 把字換成另外一個字
- 蓋住之後一樣是輸入一個 sequence，對應的輸出是另外一個 sequence
- 接著把蓋住的部分所對應的輸出做 linear transform (乘上一個矩陣) 做 softmax，得到輸出一個分佈 (distribution)



- BERT 要學習 softmax 之後的輸出，要跟 Ground truth "灣" 越接近越好，minimize cross entropy



- BERT 在訓練的時候除了做 Masking，同時也會做 Next Sentence Prediction
- 從 dataset 裡面拿兩個句子出來，句子之間會加一個特殊的符號，用來分隔兩個句子，在句子的最前面也會加一個特殊的符號
- 全部丟到 BERT，然後只取第一個輸出，把這個輸出做 linear 做二元的分類問題，判斷後面的句子是不是相接的
  - 如果是相接的，輸出 Yes
  - 如果不是相接的，輸出 No
- 後來發現 Next Sentence Prediction 對於 BERT 接下來想做的事情幫助不大



- BERT 真正學到的是如何做填空
- 可以被用在其他的任務上 (Downstream Tasks)
  - 需要一些被標註的資料
- 具有無限的潛能
- BERT 分化成各式各樣的任務，稱為 **Fine-tune**
- 在 Fine-tune 之前，產生 BERT 的過程稱為 **Pre-train** (Self-supervised Learning 的方法用在 Self-supervised Learning 的模型)

GLUE

General Language Understanding Evaluation (GLUE)

https://gluebenchmark.com/

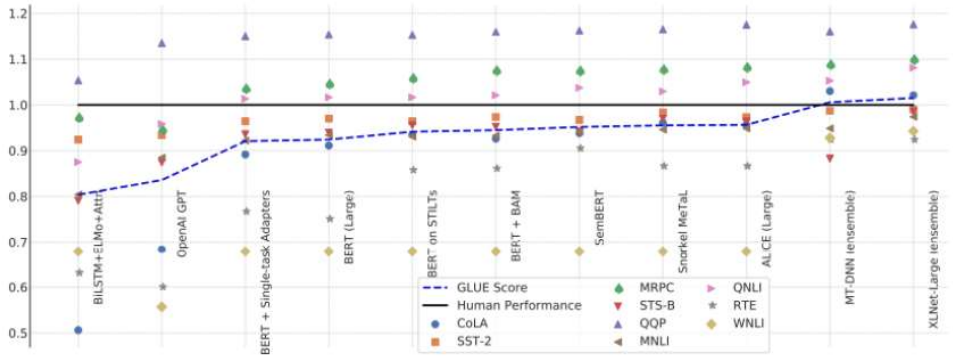
- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST-2)
- Microsoft Research Paraphrase Corpus (MRPC)
- Quora Question Pairs (QQP)
- Semantic Textual Similarity Benchmark (STS-B)
- Multi-Genre Natural Language Inference (MNLI)
- Question-answering NLI (QNLI)
- Recognizing Textual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

GLUE also has Chinese version (https://www.cluebenchmarks.com/)

- 把 BERT 分化去做多種不同的任務，看它在每個任務上得到的不同正確率是多少，再取一個平均值
- 在這種任務集最知名的標竿，稱為 GLUE (General Language Understanding Evaluation)，總共有 9 個任務
- 想知道模型的訓練效果好不好，就會把它分別微調在 9 個任務上，看 9 個任務上正確率的平均，得到一個數值代表 self supervised model 的好壞

BERT and its Family

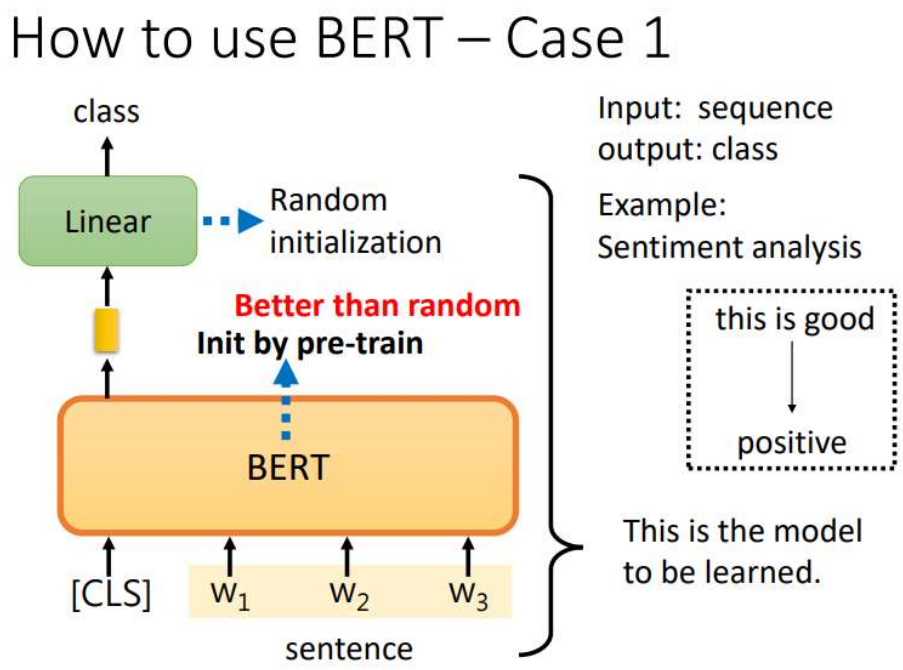
- GLUE scores



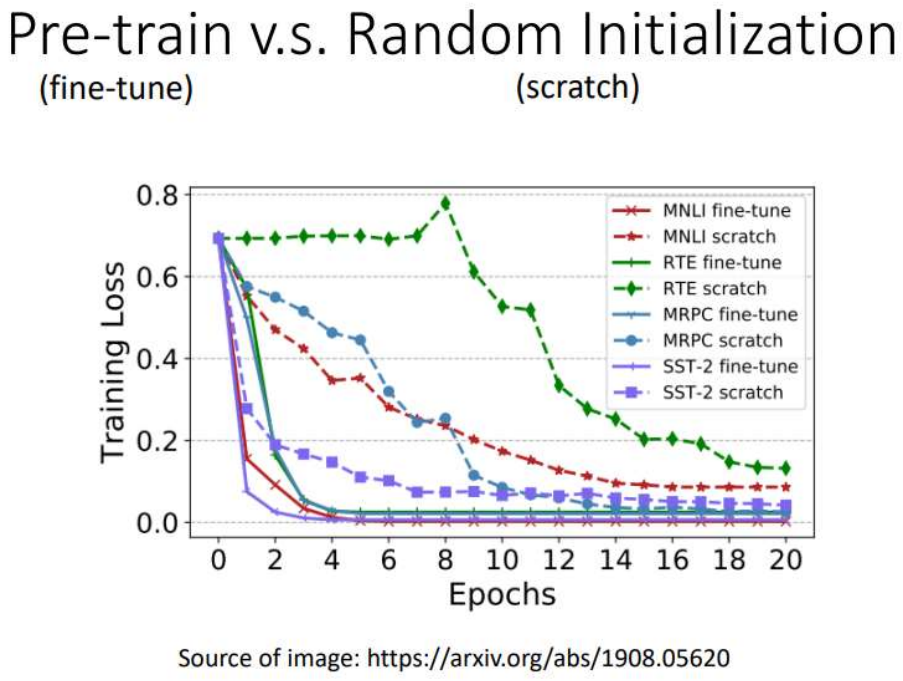
Source of image: https://arxiv.org/abs/1905.00537

- 橫軸：不同的模型
- 縱軸：正確率
- 有了 BERT 的技術之後，機器在自然語言處理的能力上往前邁進了

How to use BERT - Case 1



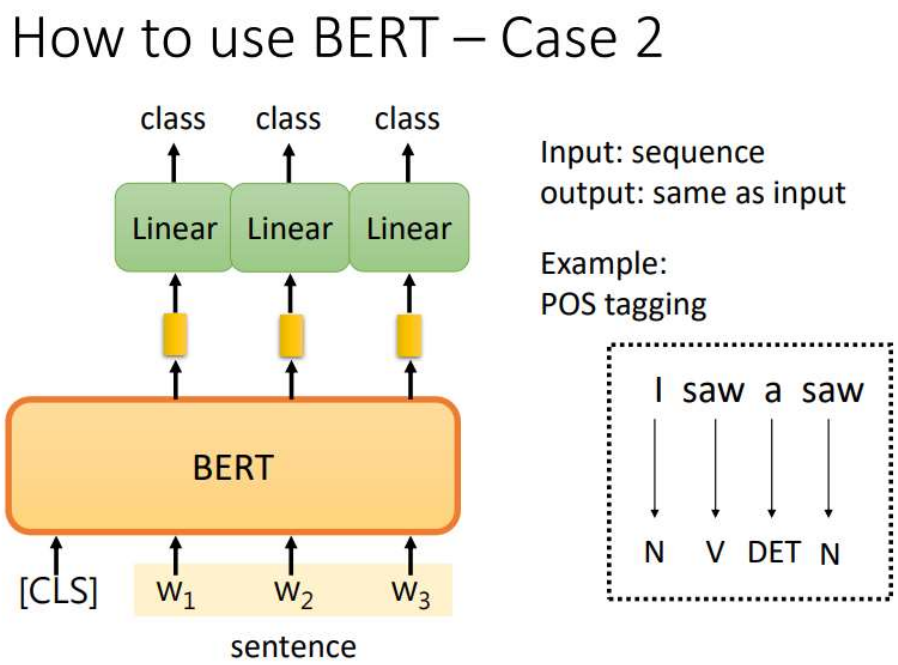
- Input : sequence
- output : class
- 任務 : Sentiment analysis (給機器一個句子，讓機器去判斷是正面的還是負面的)
- 需要提供 BERT 一些標註的資料，才能夠訓練 BERT 模型
- 在訓練的時候會把 BERT 跟 Linear 合起來變成一個 Sentiment classification 模型，並且 Linear 跟 BERT 都會用 gradient descent 去 update
  - 只是現在 **Linear 的參數是隨機初始化的**
  - **BERT 初始的參數是從學習做填空題的 BERT 來的**



Source of image: https://arxiv.org/abs/1908.05620

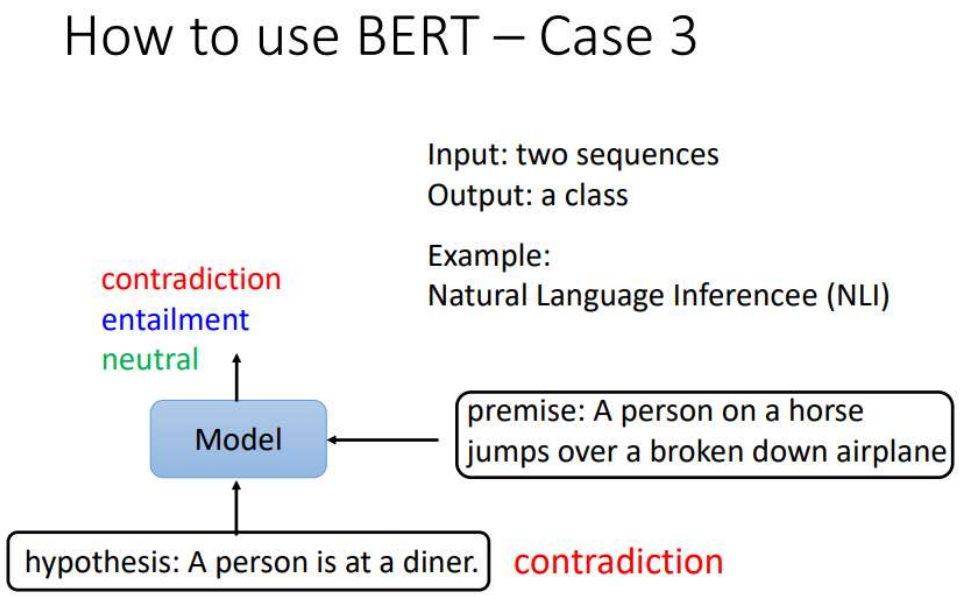
- 橫軸 : Epochs
- 縱軸 : Training Loss
- fine-tune (實線) : BERT 的部分是用學會做填空題的 BERT 的參數來做初始化
- scratch (虛線) : 整個 model 都是隨機初始化的

How to use BERT - Case 2

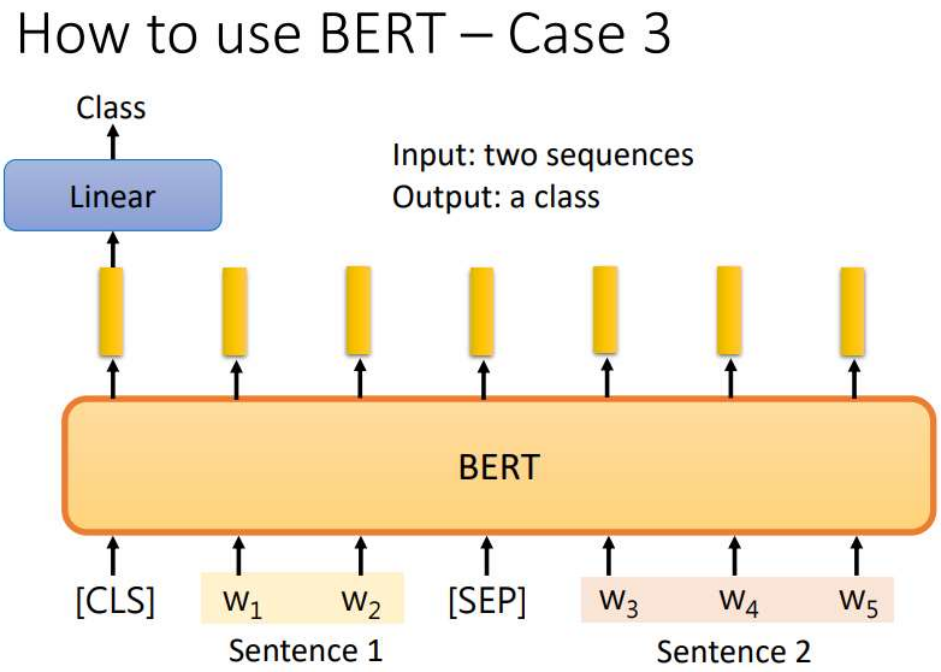


- 輸入 : sequence
- 輸出 : 另外一個 sequence
- 輸入跟輸出的長度相同
- 任務 : POS tagging (詞性標註)
- BERT 的部分是用學會做填空題的 BERT 的參數來做初始化

How to use BERT - Case 3

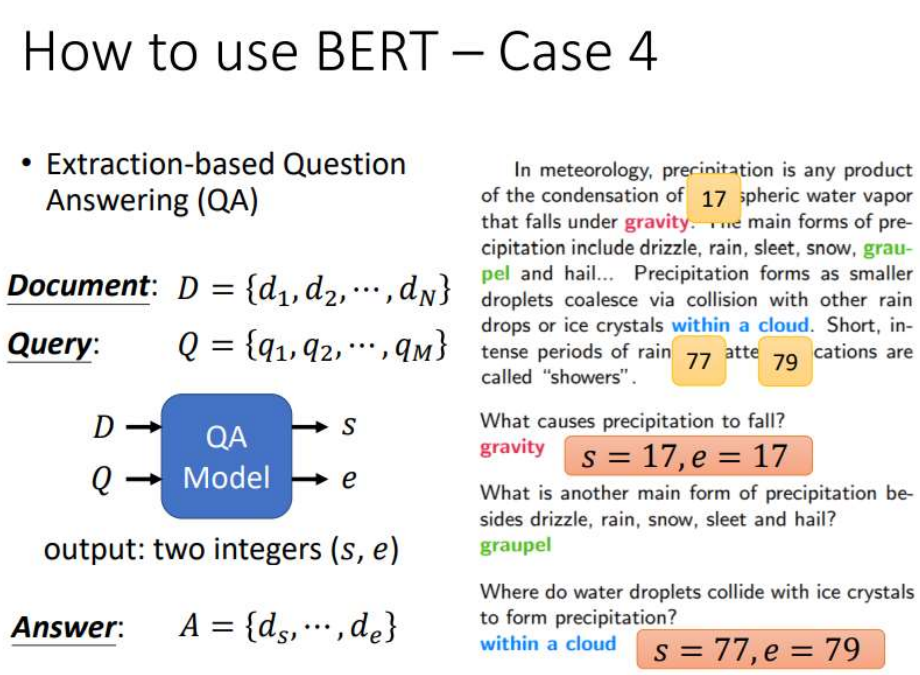


- 輸入：two sequences
- 輸出：一個類別
- 應用：Natural Language Inferencee (NLI)
  - 給機器兩個句子：前提、假設
  - 目的：給機器一個前提一個假設，讓機器產出兩個句子之間的關係 (矛盾、成立、中立)
- 可以改成語音的例子、或是影像的例子

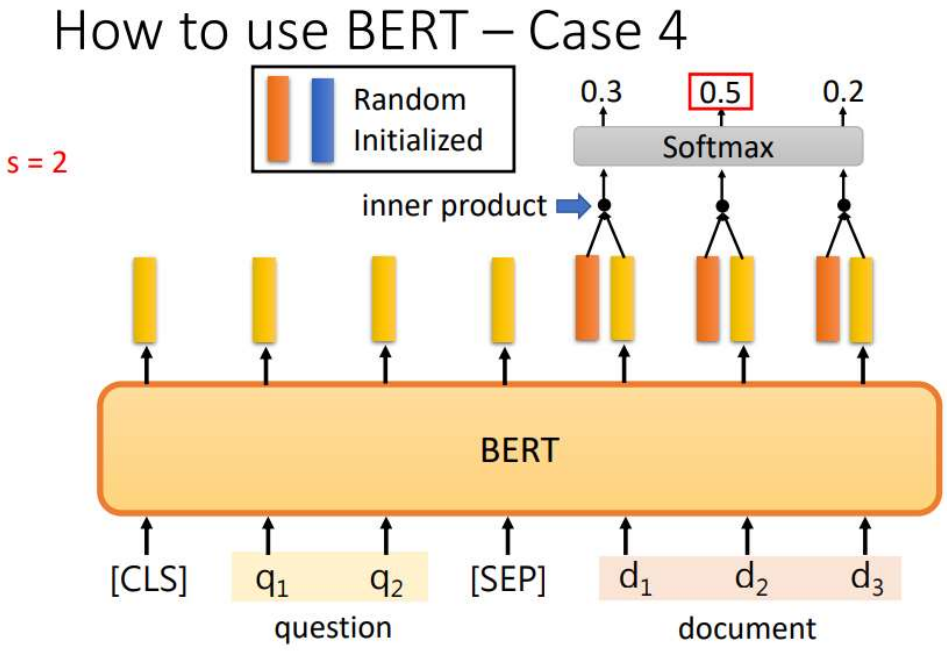


- 只取 CLS 的輸出丟到 Linear Transform 裡面，看輸出應該是甚麼類別
- 一樣需要一些標註的資料
- BERT 的部分是用學會做填空題的 BERT 的參數來做初始化

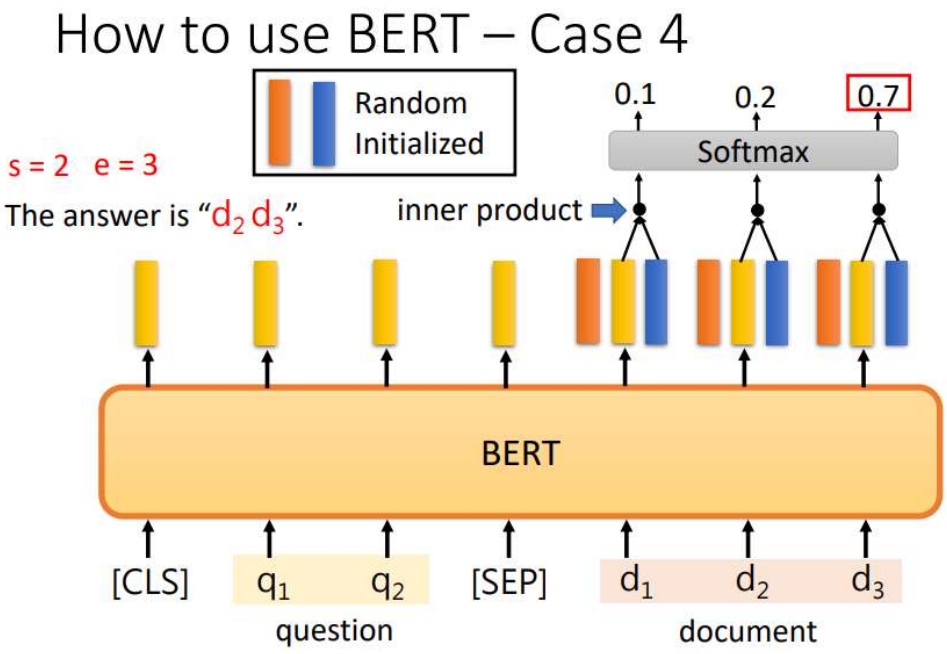
How to use BERT - Case 4



- 作業 7 (問答系統 - Extraction-based QA)
- 給機器一篇文章，問它一個問題，讓機器回答問題
- 限制是：答案在文章裡面的某個片段
- 輸入：文章、問題
- 輸出兩個整數 s、e，(文章的第 s 個字到文章的第 e 個字，就是正確答案)



- 輸入：一個問題、一個文章、問題跟文章之間有個特殊的符號、前面再放一個 [CLS] 的 token
- 把橘色的向量 (代表答案起始的位置) 分別跟文章輸出的每個向量做 inner product，算出三個數值，再去做 softmax 得到三個數值
- d_2 得到的數值最高，所以 s 等於 2



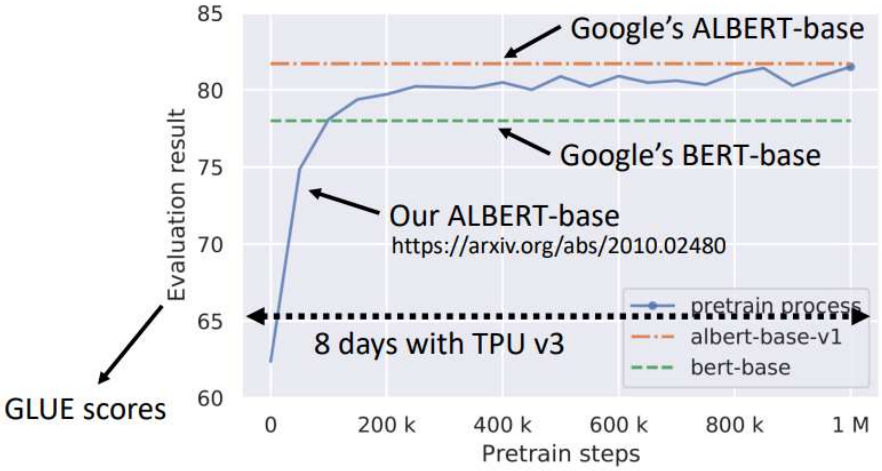
- 藍色的向量 (代表答案結束的位置) ，分別跟文章輸出的每個向量做 inner product，算出三個數值，再去做 softmax 得到三個數值
- d_3 得到的數值最高，所以 e 等於 3

台達電產學合作計畫研究成果
This work is done by 姜成翰

Training BERT is challenging!

Training data has more than **3 billions** of words.

3000 times of **Harry Potter series**

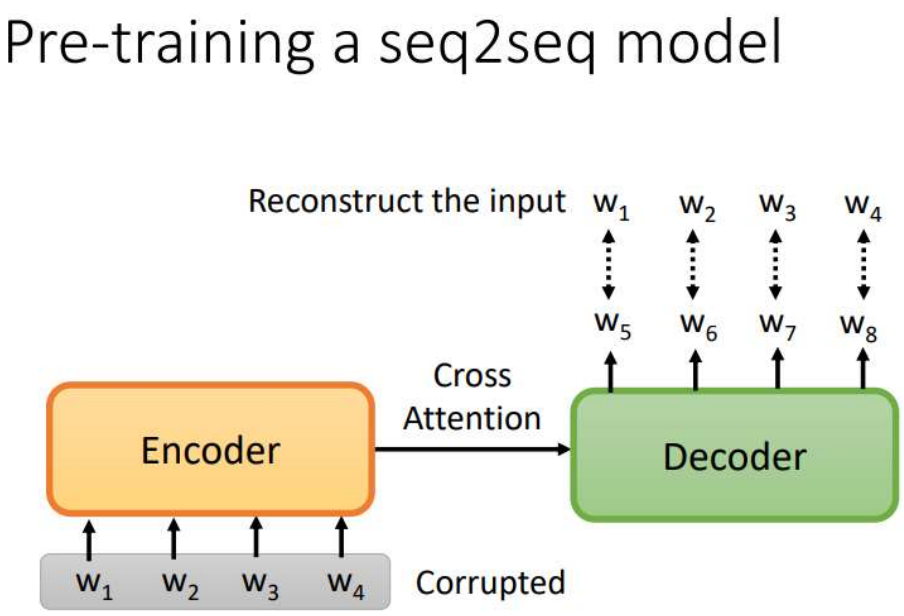


BERT Embryology (胚胎學)

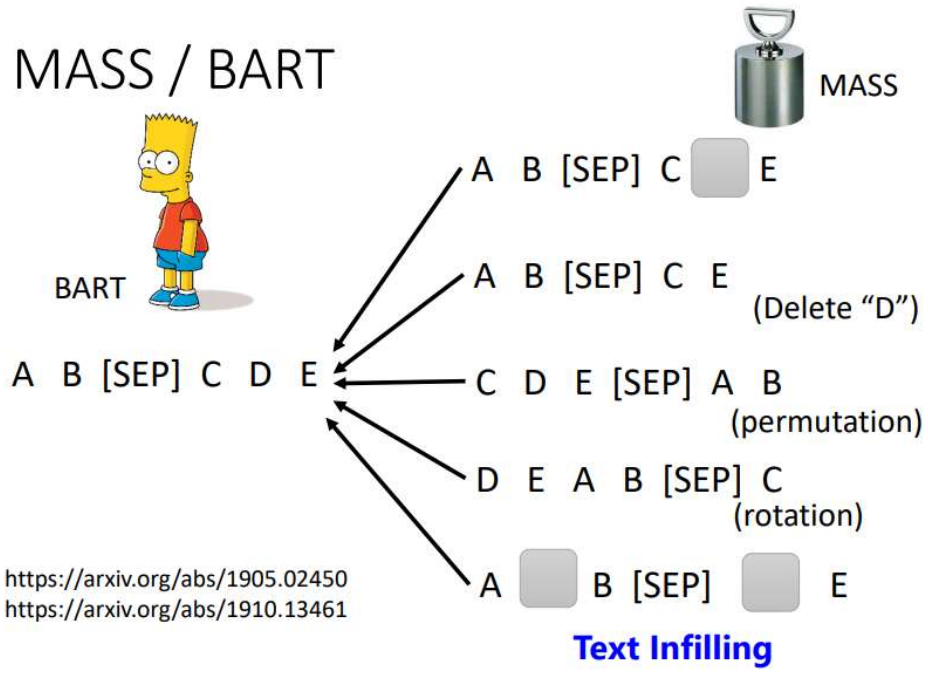
https://arxiv.org/abs/2010.02480

When does BERT know POS tagging, syntactic parsing, semantics?

The answer is counterintuitive!



- 前面提到的任務都沒有包含 seq2seq model
- 如果今天要解的任務是 seq2seq 怎麼辦呢？
  - BERT 只有 pretrain Encoder，但也可以去 pretrain Decoder
  - 有一個 seq2seq 的 transformer 有一個 Encoder、Decoder，輸入一串句子，輸出一串句子，中間用 Cross Attention 連接在一起
    - 把 Encoder 的輸入故意把它弄壞
    - 希望 Decoder 輸出的句子跟它弄壞之前一模一樣
    - Decoder 看到弄壞的結果，要輸出還原弄壞前的結果
    - train 下去，就是 pretrain 一個 seq2seq 的 model

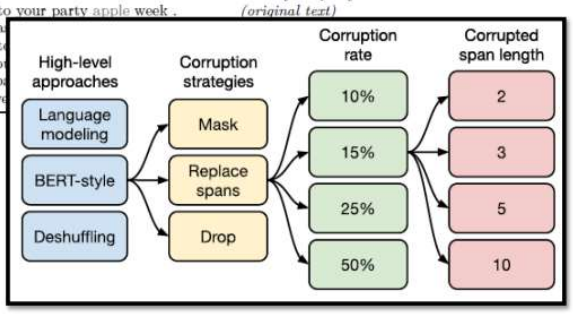


- 很多種弄壞的方法

T5 – Comparison

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

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last	
I.i.d. noise, mask tokens	Thank you <M> <M> me to	
I.i.d. noise, replace spans	Thank you <X> me to yo	
I.i.d. noise, drop tokens	Thank you me to your pa	
Random spans	Thank you <X> to <Y> we	



tags: 2022 李宏毅_機器學習