

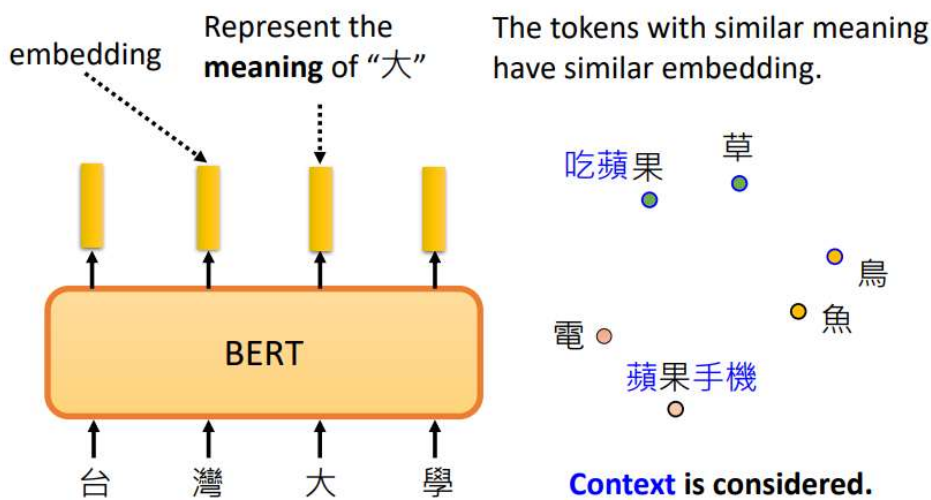
# 自督導式學習 (Self-supervised Learning) (下)

Create at 2022/06/26

- 自督導式學習 (Self-supervised Learning) (下)
  - BERT的奇聞軼事
    - Multi-lingual BERT
  - GPT 的野望
- 上課資源：
  1. 自督導式學習 (Self-supervised Learning) (三) – BERT的奇聞軼事  
(<https://www.youtube.com/watch?v=ExXA05i8DEQ>)
  2. 自督導式學習 (Self-supervised Learning) (四) – GPT的野望 ([https://www.youtube.com/watch?v=WY\\_E0Sd4K80](https://www.youtube.com/watch?v=WY_E0Sd4K80))

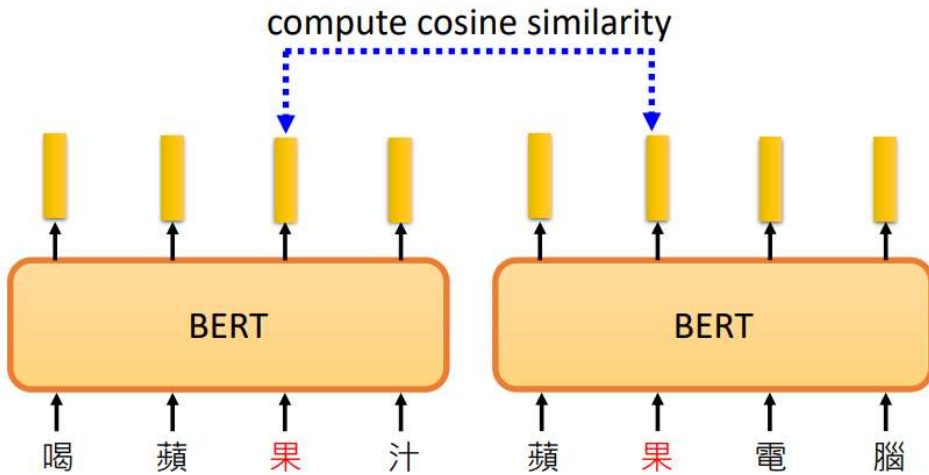
## BERT的奇聞軼事

### Why does BERT work?

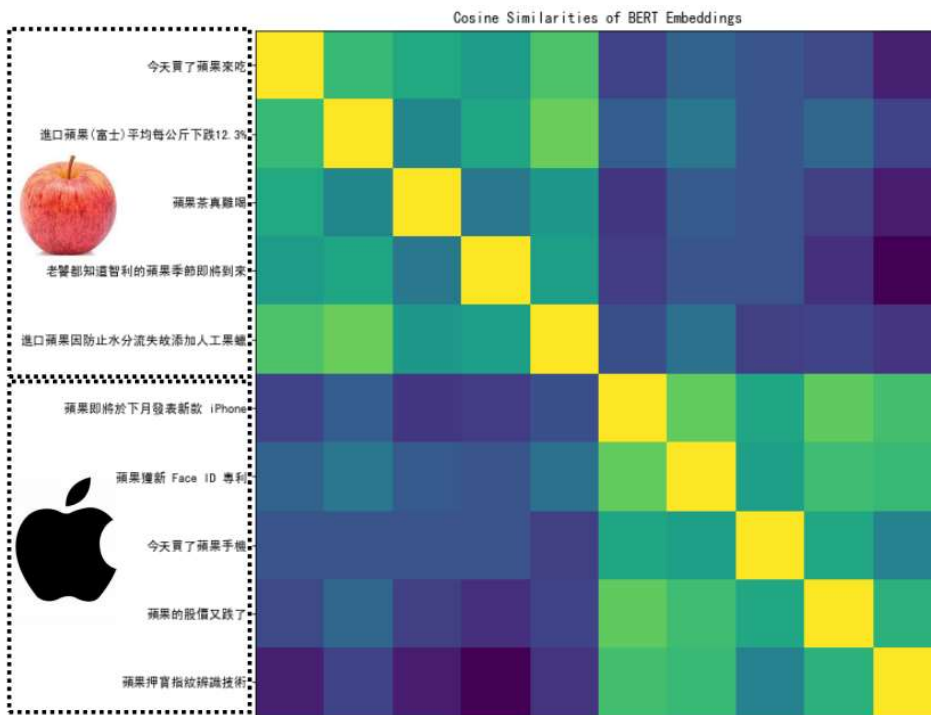


- 為甚麼 BERT 有用呢？
  - 輸入一段文字之後產生一段向量稱它為 embedding
  - 這些向量代表輸入的字的意思
  - 意思越相近，向量就越像

# Why does BERT work?



- 把兩個句子都丟到 BERT 裡面，接著分別計算兩個“果”之間的 cosine similarity



- 這邊有 10 個“果”，兩兩之間計算相似度
- 前五個“果”相似度比較高，後五個“果”相似度比較高

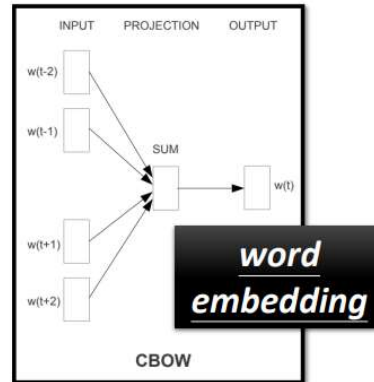
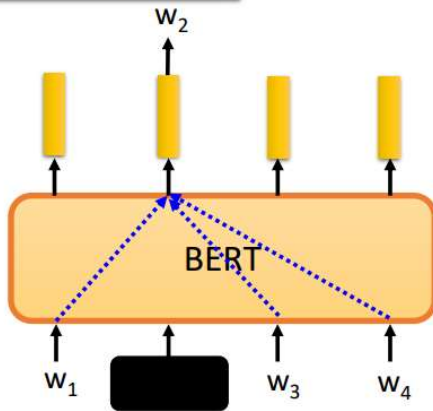
## Why does BERT work?

**Contextualized word embedding**

You shall know a word by the company it keeps



John Rupert Firth



- BERT 輸出的每一個向量，就代表了那個字的意思
- BERT 在學填空題的過程中，學會了每個中文字的意思，所以可以在接下來的任務做的更好
- 要知道一個詞彙的意思取決於它的上下文

<https://arxiv.org/abs/2103.07162>

This work is done by 高瑋聰

## Why does BERT work?

- Applying BERT to **protein, DNA, music classification**



A  
T  
C  
G

EI	CCAGCTGCATCACAGGAGGCCAGCC
EI	AGACCCGCCGGGAGGCGGAGGACC
IE	AACGTGGCCTCCTTGTGCCCTTCCCC
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
IE	CCTGATCTGGGTCTCCCCTCCACCC
IE	AGCCCTCAACCCTTCTGTCTCACCT
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
N	CTGTGTTACACCATCAAGCGCCGGG
N	GTGTTACCGAGGGCATTCTAACAGT
N	TCTGAGCTCTGCATTGTCTATTCTCC

class

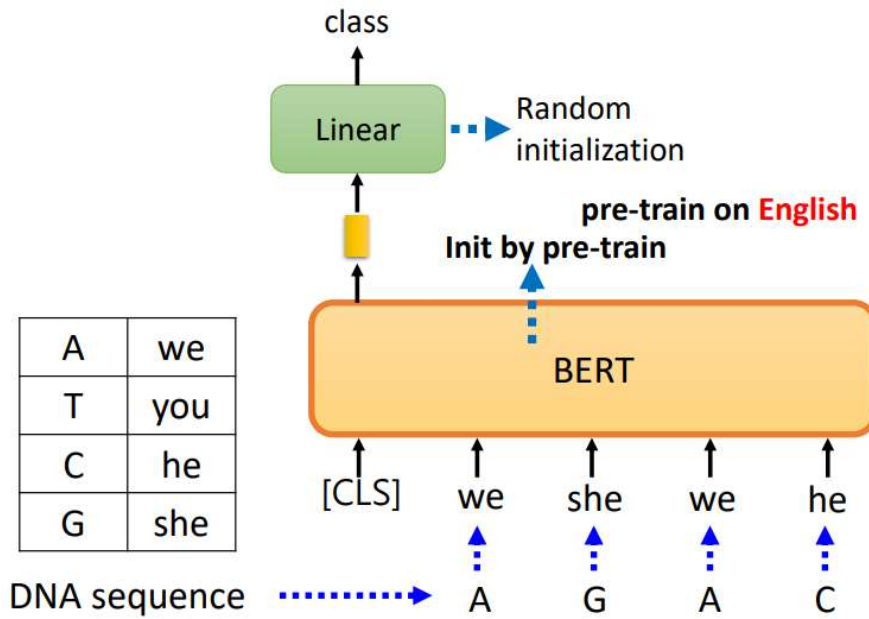
DNA sequence

- 把訓練在文字上的 BERT，拿來做蛋白質的分類、DNA 的分類、音樂的分類
- 以 DNA 的分類為例

## Why does BERT work?

<https://arxiv.org/abs/2103.07162>

This work is done by 高瑋聰



- DNA 用 A、T、C、G 來表示
- 現在要把 BERT 用在 DNA 的分類上
- 把 A、T、C、G 分別對應到隨便一個英文的詞彙

<https://arxiv.org/abs/2103.07162>

This work is done by 高瑋聰

## Why does BERT work?

- Applying BERT to **protein, DNA, music classification**

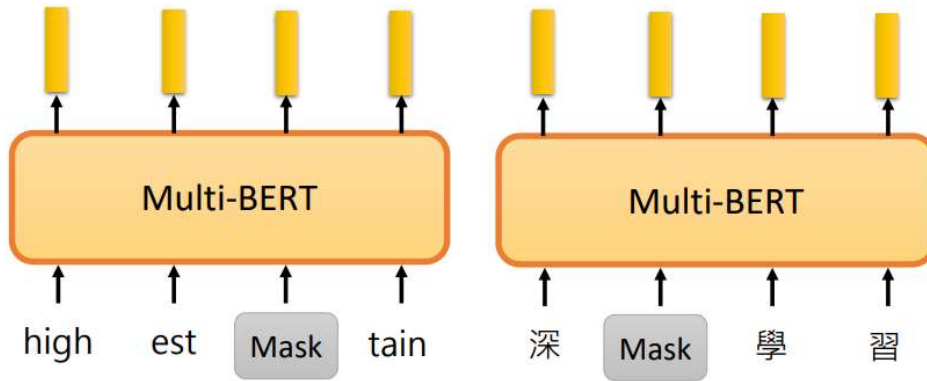
	Protein			DNA				Music
	localization	stability	fluorescence	H3	H4	H3K9ac	Splice	composer
specific	69.0	76.0	63.0	87.3	87.3	79.1	94.1	-
BERT	64.8	74.5	63.7	83.0	86.2	78.3	97.5	55.2
re-emb	63.3	75.4	37.3	78.5	83.7	76.3	95.6	55.2
rand	58.6	65.8	27.5	75.6	66.5	72.8	95	36



- BERT 的能力不完全來自於它看得懂文章

## Multi-lingual BERT

# Multi-lingual BERT

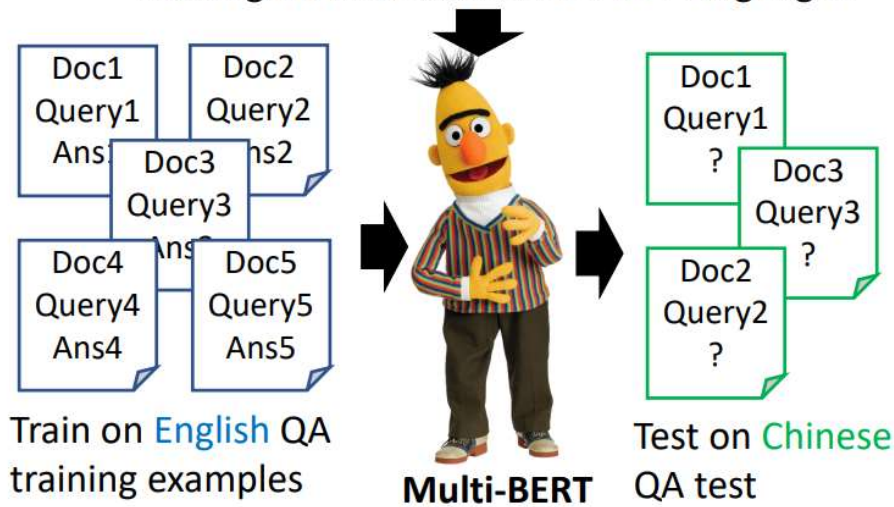


Training a BERT model by many different languages.

- 用不同的語言給他做填空

## Zero-shot Reading Comprehension

Training on the sentences of 104 languages



- 拿英文的 QA 資料做訓練，它就會自動學會做中文的 QA



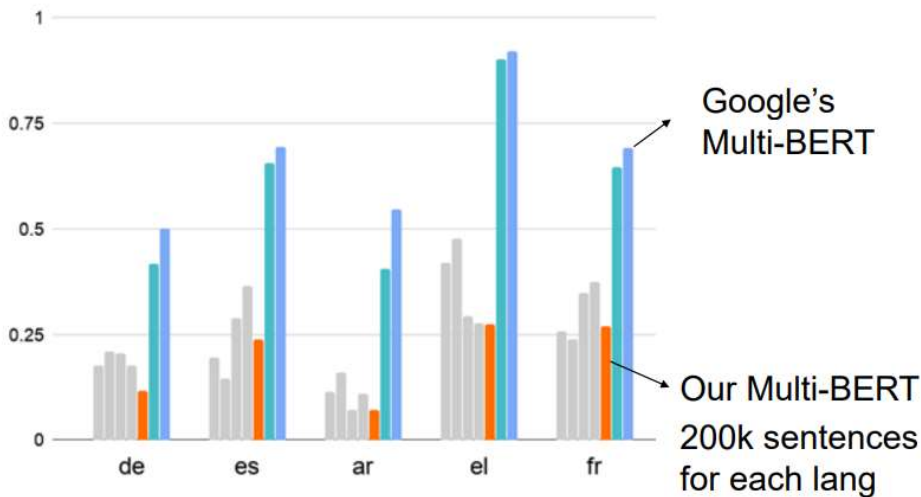
- English: SQuAD, Chinese: DRCD

F1 score of Human performance is 93.30%

- 也許對 Multi-BERT 而言，不同語言之間沒有什麼差異

## Mean Reciprocal Rank (MRR):

Higher MRR, better alignment


<https://arxiv.org/abs/2010.10938>

投影片來源: 許宗嫻同學碩士口試投影片

How about 1000k?

- MRR 的值越高，代表兩個不同語言它們的 align 越好 (向量越接近)
- 藍色的線是 google 試出來的 104 種語言的 Multi-BERT 得到的 MRR
- 橘色的是我們自己訓練的 Multi-BERT 使用 200k sentences

## The training is also challenging ...

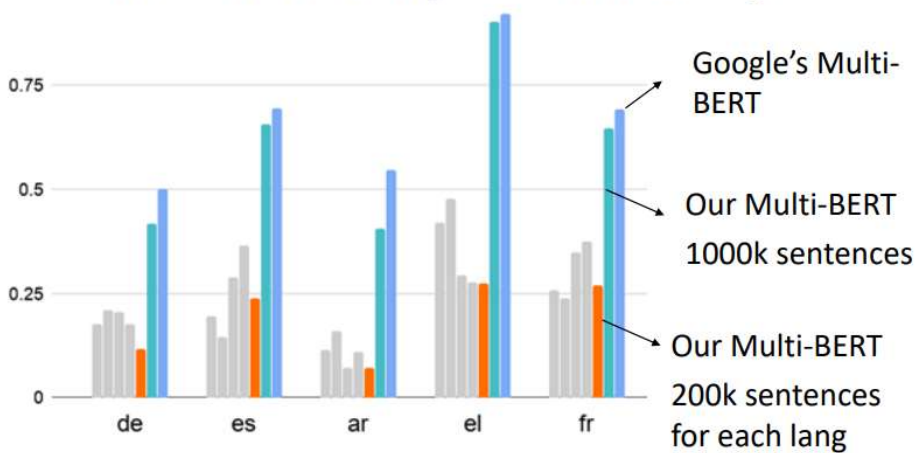


- 資料變成 100k sentences

## Mean Reciprocal Rank (MRR):

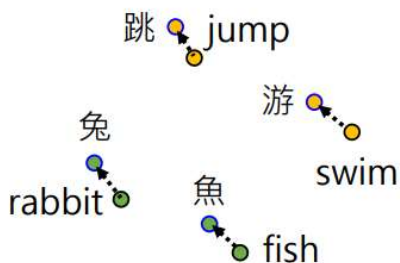
Higher MRR, better alignment

1 — The amount of training data is critical for alignment.


<https://arxiv.org/abs/2010.10938>

投影片來源: 許宗嫻同學碩士口試投影片

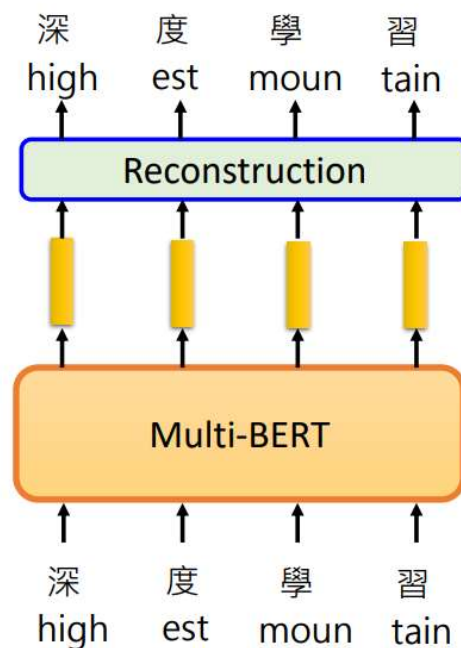
- 綠色的線是資料量改成 1000k 的 MRR

<https://arxiv.org/abs/2010.10041>
**Weird???**

If the embedding is  
language independent ...

How to correctly  
reconstruct?

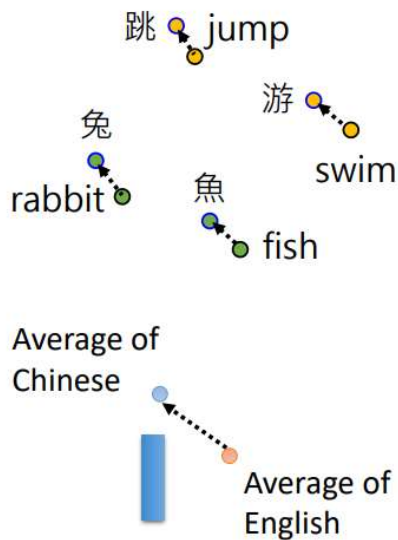
There must be language  
information.



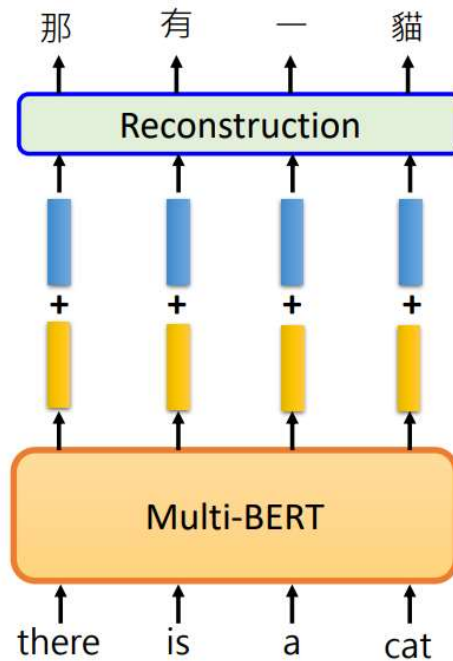
- 它知道語言的資訊



## Where is Language?



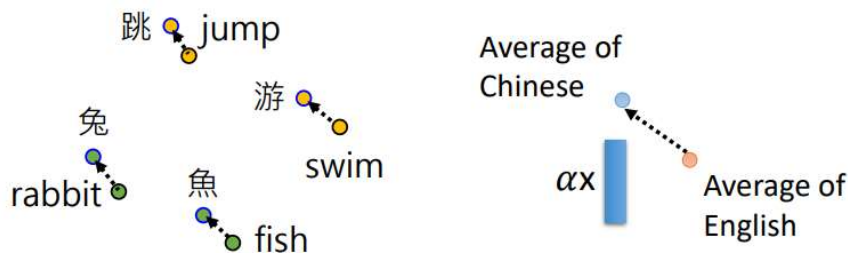
This work is done by 劉記良、許宗嫻、莊永松



- 把所有英文的詞彙都丟到 Multi-BERT，把英文的 embedded 計算平均，再計算所有中文的 embedded 平均
- 兩者相減就是中文跟英文之間的差距 (藍色向量)
- 給 Multi-BERT 一句英文得到 embedded，把這個 embedded 加上藍色的向量，最後的向量對 Multi-BERT 來說就變成了中文的句子

## If this is true ...

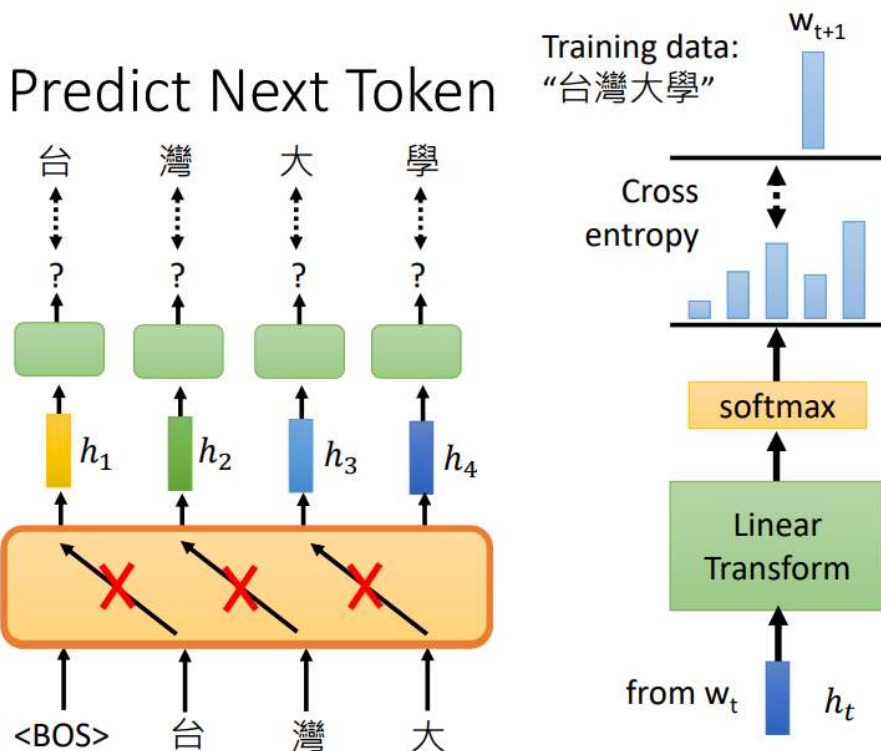
This work is done by 劉記良、許宗嫻、莊永松  
<https://arxiv.org/abs/2010.10041>



Input (en)	The girl that can help me is all the way across town. There is no one who can help me.
Ground Truth (zh)	能幫助我的女孩在小镇的另一边。没有人能幫助我。
en→zh, $\alpha = 1$	孩。 can 来我是all the way across 市。 There 是无人人 can help 我。
en→zh, $\alpha = 2$	孩的的家我是这个人的市。他是他人人的到我。
en→zh, $\alpha = 3$	。的的的他的是个的的。：他是他人，的。他。

Unsupervised token-level translation 😊

## GPT 的野望

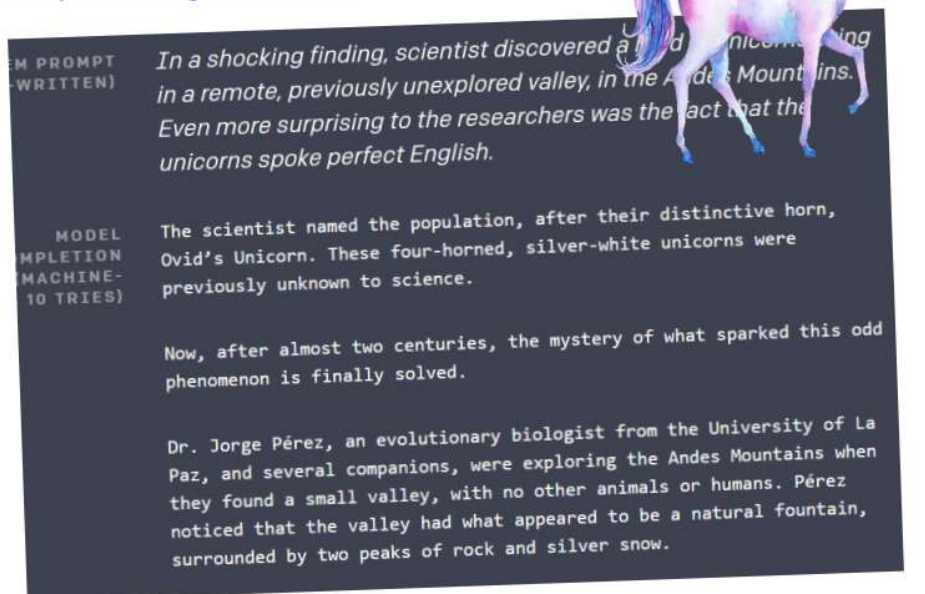


- GPT 就是改一下我們在做 self-supervised learning 的時候要模型做的任務
- GPT 要做的任務是預測接下來會出現的 token 是什麼
- GPT 模型像是 transformer 的 decoder

<https://talktotransformer.com/>

## Predict Next Token

They can do generation.



- <https://app.inferkit.com/demo> (<https://app.inferkit.com/demo>)

# How to use GPT?

## 第一部份：詞彙和結構

本部份共 15 題，每題含一個空格。請就試題冊上 A、B、C、D 四個選項中選出最適合題意的字或詞，標示在答案紙上。

例：

It's eight o'clock now. Sue \_\_\_\_\_ in her bedroom.

- A. study
- B. studies
- C. studied
- D. is studying

正確答案為 D，請在答案紙上塗黑作答。

## Description

## A few example

- GPT 是如何運作？
  - 先給問題一個描述
  - 再給一個範例

### "Few-shot" Learning

(no gradient descent)

1	Translate English to French:	← task description
2	sea otter => loutre de mer	← examples
3	peppermint => menthe poivrée	←
4	plush girafe => girafe peluche	←
5	cheese => .....	← prompt

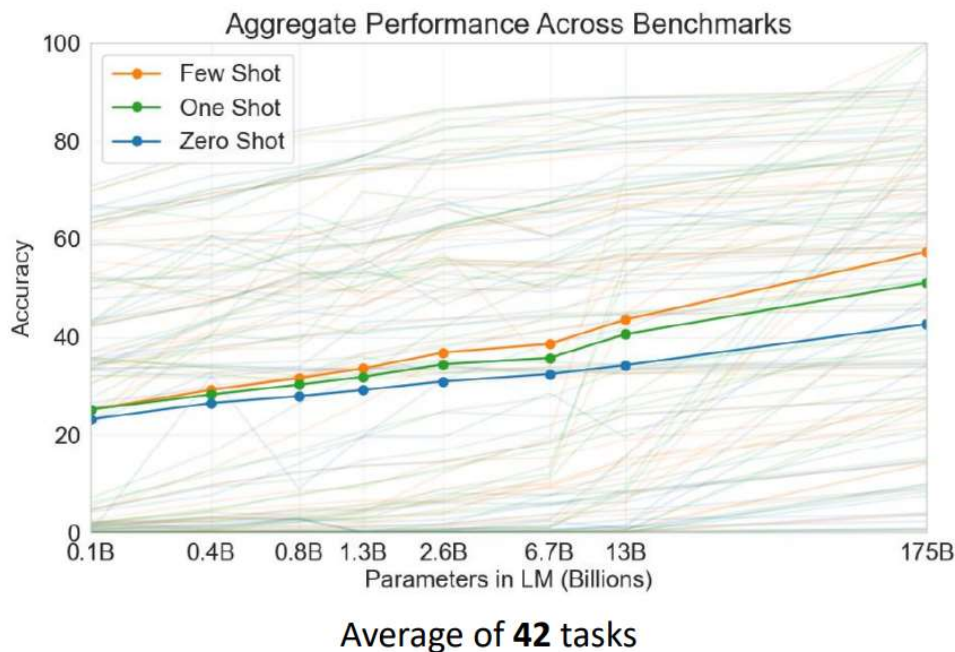
### "One-shot" Learning

1	Translate English to French:	← task description
2	sea otter => loutre de mer	← example
3	cheese => .....	← prompt

### "Zero-shot" Learning

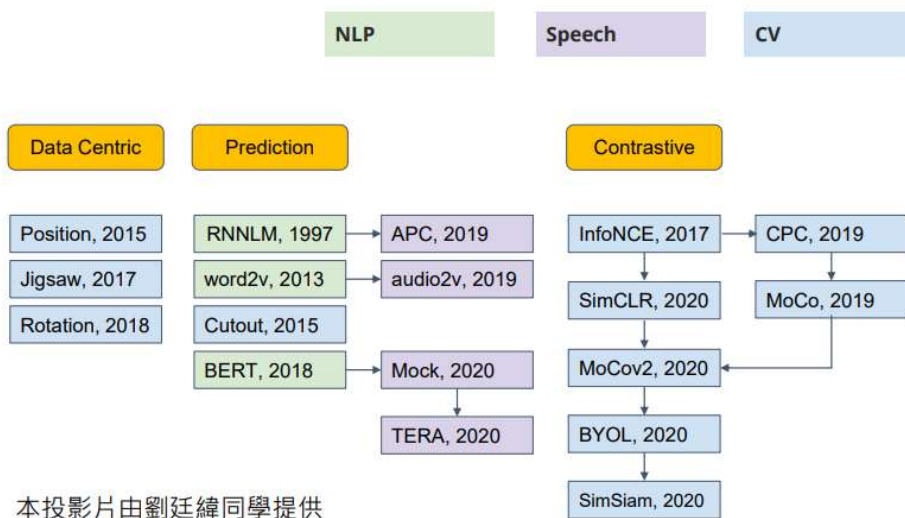
1	Translate English to French:	← task description
2	cheese => .....	← prompt

- Few-shot Learning：給一點例子
- In-context Learning (不是一般的 learning，它連 gradient descent 都沒有做)
- One-shot Learning：給一個例子
- Zero-shot Learning：完全不給例子



- 三條線是 42 個任務的平均正確率

## Beyond Text

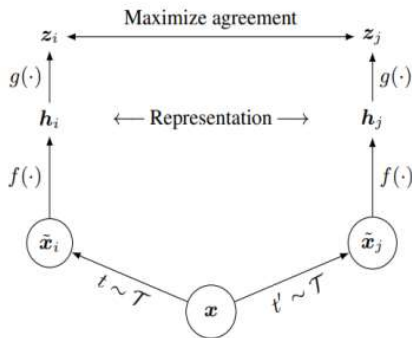


- 在語音、影像都可以用 self-supervised learning 的技術

# Image - SimCLR

<https://arxiv.org/abs/2002.05709>

<https://github.com/google-research/simclr>



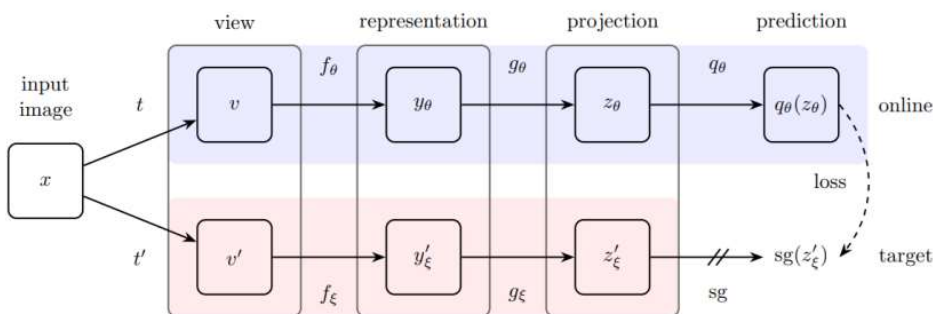
- Image - SimCLR

# Image - BYOL

**Bootstrap your own latent:**

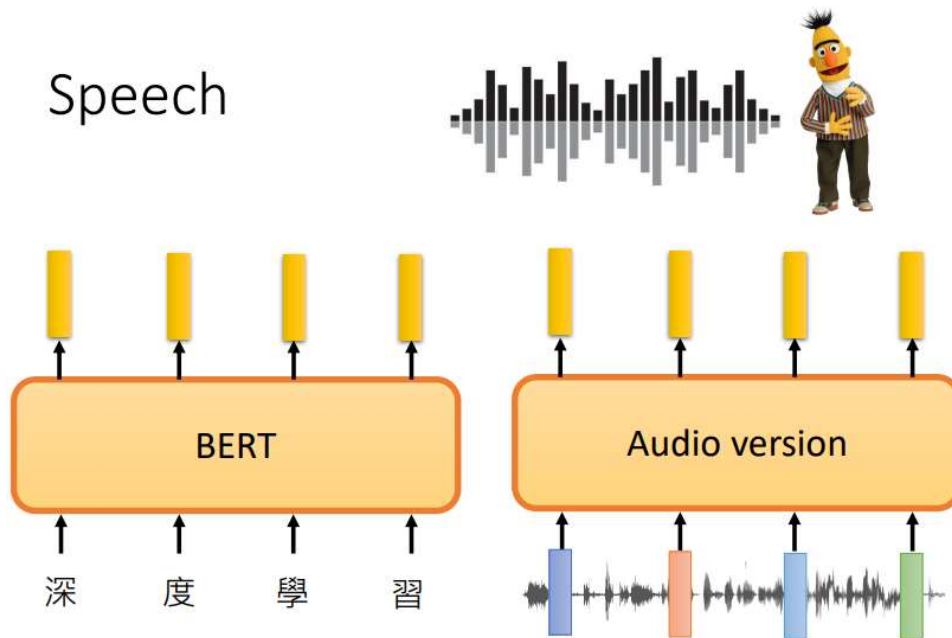
A new approach to self-supervised Learning

<https://arxiv.org/abs/2006.07733>



- Image - BYOL



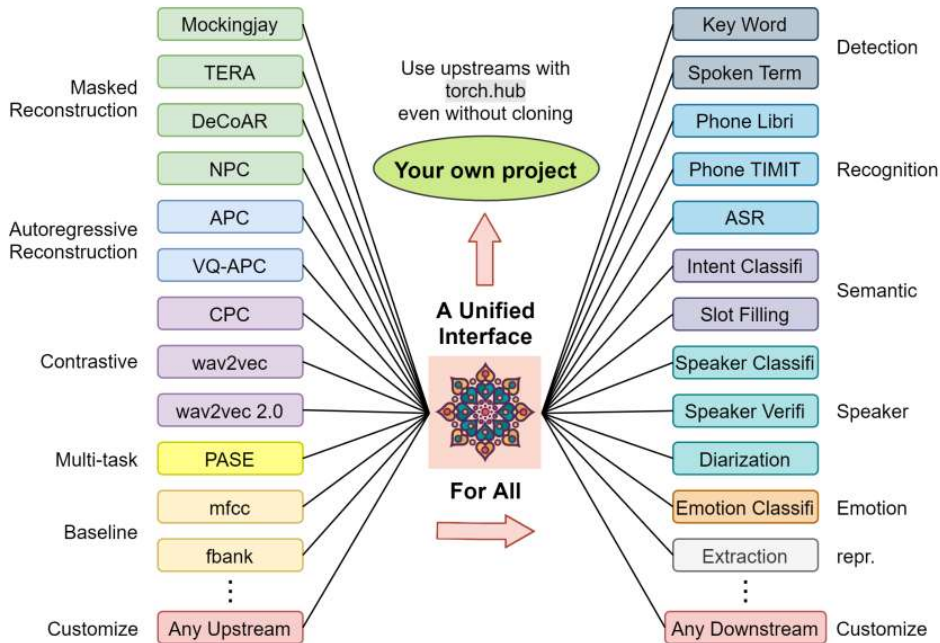


- 也可以做語音版的 GPT、BERT

## Speech GLUE - SUPERB

- Speech processing **Universal PERFORMANCE Benchmark**
  - Will be available soon
- **Downstream:** Benchmark with 10+ tasks
  - The models need to know how to process content, speaker, emotion, and even semantics.
- **Toolkit:** A flexible and modularized framework for self-supervised speech models.
  - <https://github.com/s3prl/s3prl>

- 在語音上還沒有類似 GLUE 基準的資料庫
- **SUPERB**



<https://github.com/andi611/Self-Supervised-Speech-Pretraining-and-Representation-Learning>

- Toolkit 裡面包含各式各樣 self-supervised learning 模型
- 以及這些 self-supervised learning 模型可以做的各式各樣的下游的任務

- self-supervised learning 技術，不只能用在文字上，在影像上、語音上都仍然有非常大的空間可以使用 self-supervised learning 的技術

tags: 2022 李宏毅\_機器學習