Self-Supervised Learning

Hung-yi Lee 李宏毅


死臭酸宅本人
芝麻街
STAYREAL

Big Bird: Transformers for Longer Sequences

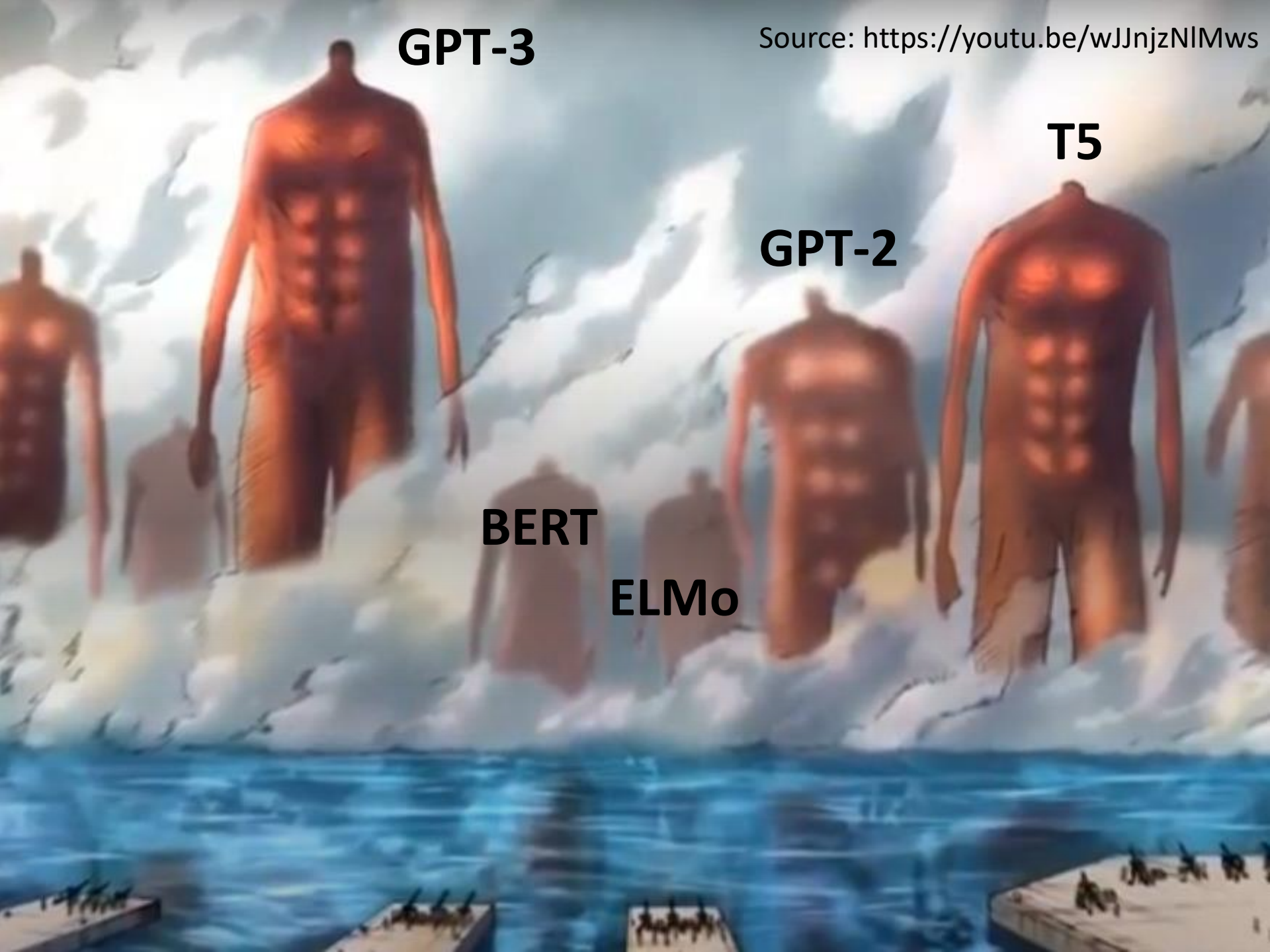
ERNIE (Enhanced Representation through Knowledge Integration)

Grover (Generating aRticles by Only Viewing mEtadata Records)

BERT (Bidirectional Encoder Representations from Transformers)

ELMo (Embeddings from Language Models)

Source of image:
https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html


GPT-3
Source: https://youtu.be/wJJnjzNlMws
T5
GPT-2
BERT
ELMo

The models are become larger and larger …

有這麼多參數

ELMO (94M)

BERT (340M)

GPT-2 (1542M)

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

The models are become larger and larger …

GPT-3 is **10** times larger than Turing NLG.

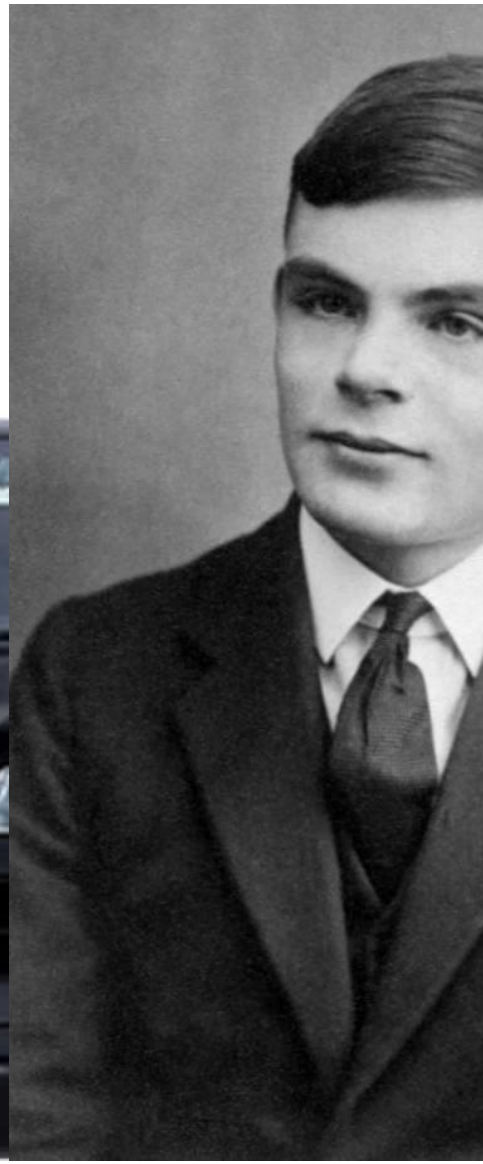
有這麼多參數

GPT-2

Megatron (8B)

T5 (11B)

Turing NLG (17B)



BERT (340M)

GPT-3 (175B)

Switch Transformer (1.6T)

目前最多參數的模型

https://arxiv.org/abs/2101.03961

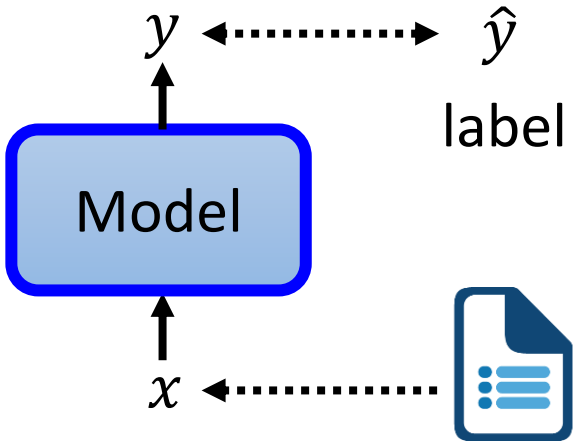
Outline

BERT series

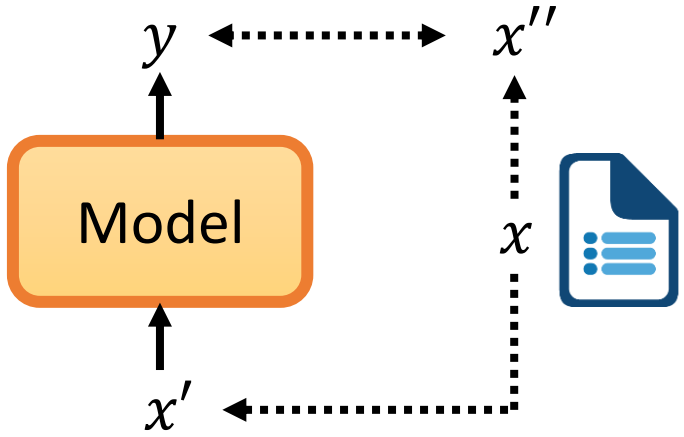
GPT series

Self-supervised Learning

Supervised



Self-supervised

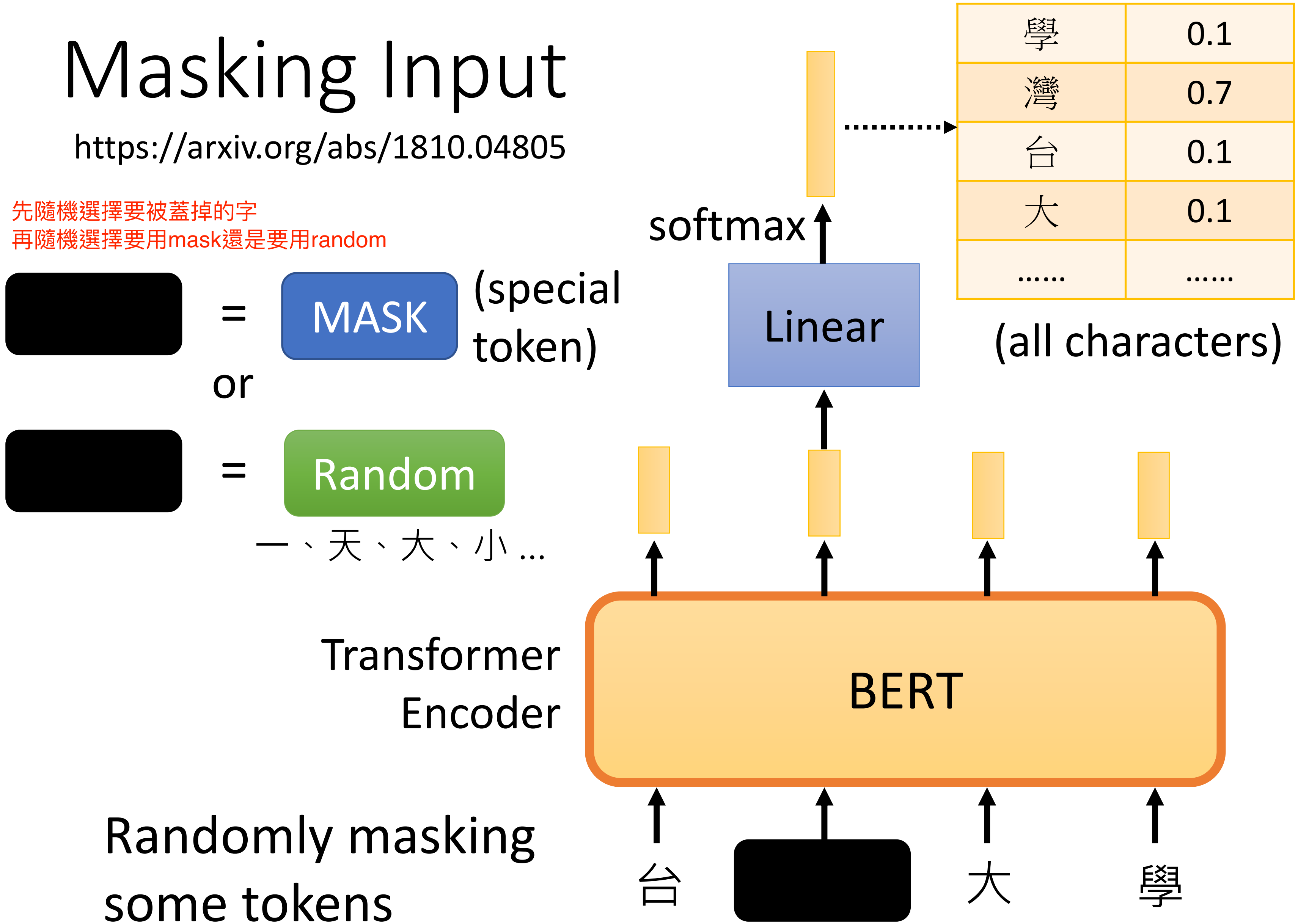


Yann LeCun

2019年4月30日 ·

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

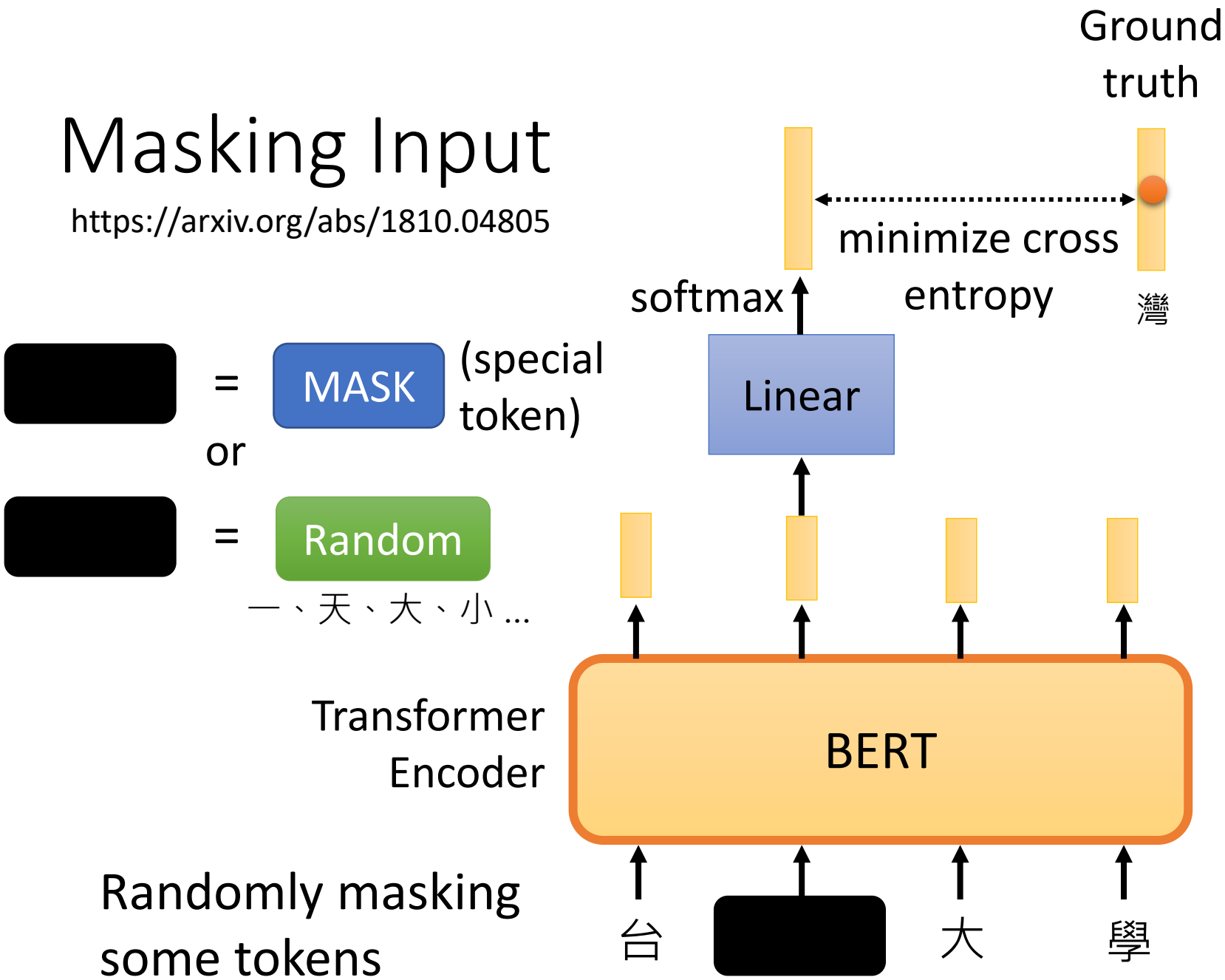
In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.



Masking Input
https://arxiv.org/abs/1810.04805
先隨機選擇要被蓋掉的字
再隨機選擇要用mask還是要用random
= MASK (special token)
or
= Random
一、天、大、小 ...
Transformer Encoder
BERT
Linear
softmax
學 0.1
灣 0.7
台 0.1
大 0.1
…… ……
(all characters)
Randomly masking some tokens
台
大
學

Masking Input

https://arxiv.org/abs/1810.04805



Next Sentence Prediction

可能是因為要分辨兩個句子是否相接，太簡單了
bert無法藉由訓練來學到新東西

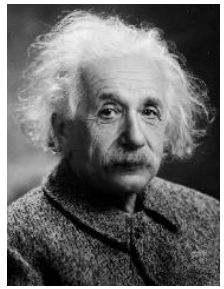
- This approach is not helpful.

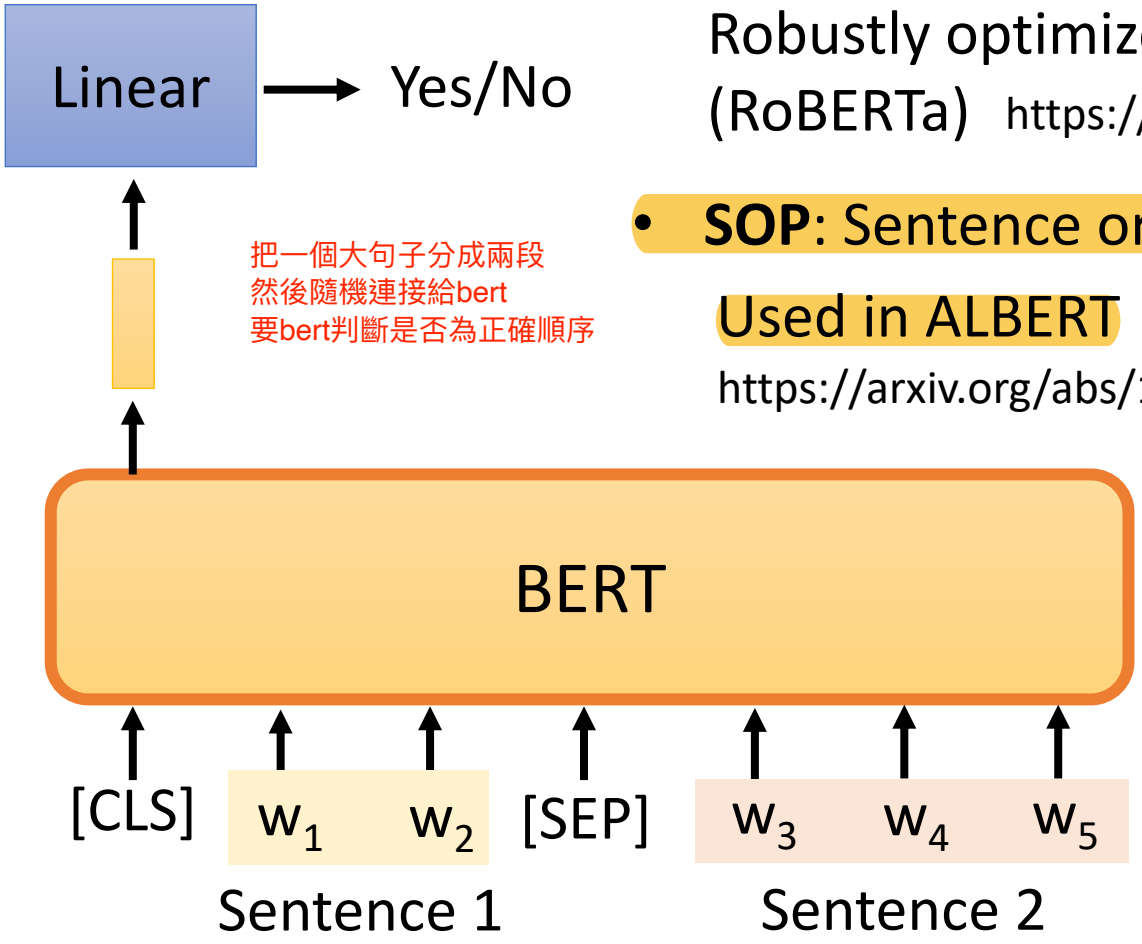
Robustly optimized BERT approach (RoBERTa) https://arxiv.org/abs/1907.11692

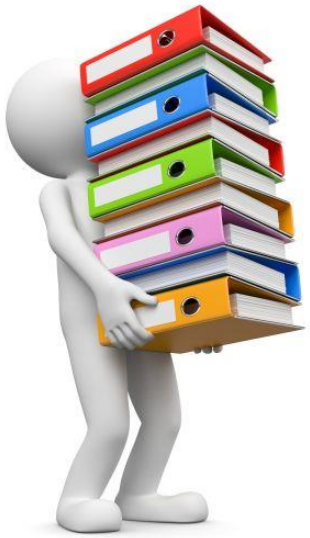
- **SOP**: Sentence order prediction

Used in ALBERT

https://arxiv.org/abs/1909.11942







Self-supervised Learning

Pre-train

BERT

- Masked token prediction
- Next sentence prediction

Fine-tune

Model for Task 1

Model for Task 2

Model for Task 3

Downstream Tasks

- The tasks we care
- We have a little bit labeled data.

GLUE

General Language Understanding Evaluation (GLUE)

在各種任務上的綜合表現 https://gluebenchmark.com/

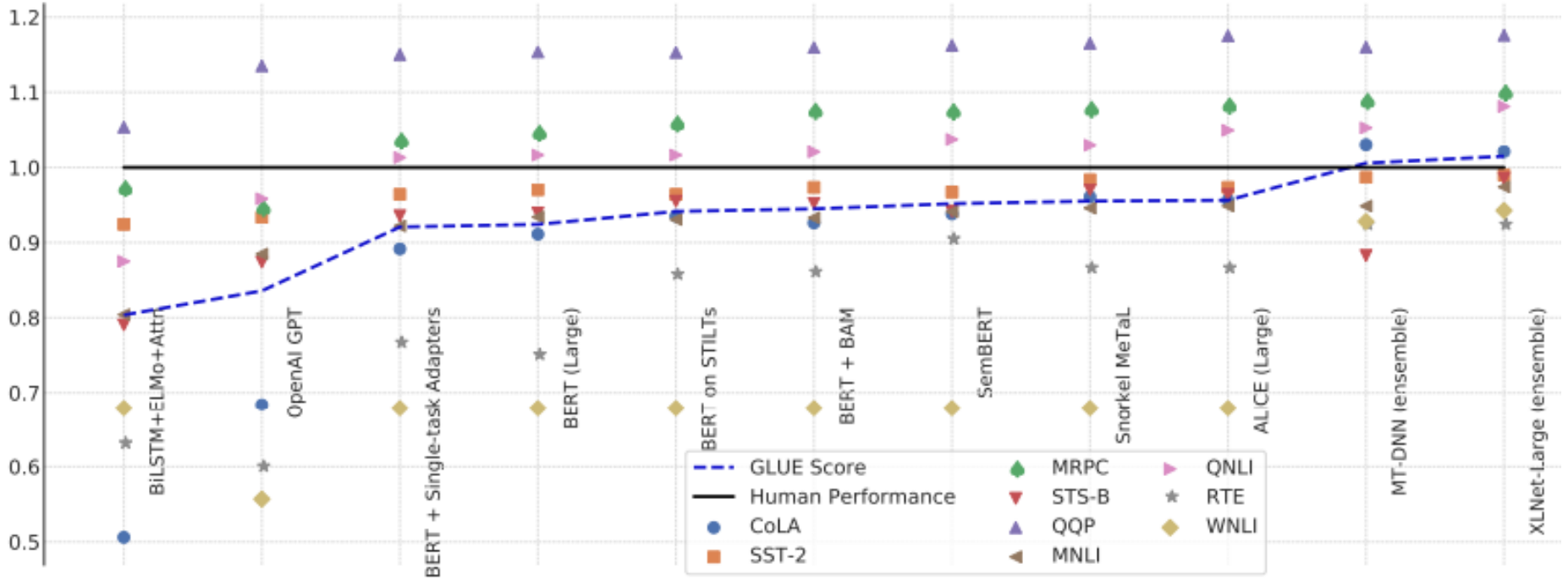
- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST-2)
- Microsoft Research Paraphrase Corpus (MRPC)
- Quora Question Pairs (QQP)
- Semantic Textual Similarity Benchmark (STS-B)
- Multi-Genre Natural Language Inference (MNLI)
- Question-answering NLI (QNLI)
- Recognizing Textual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

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

BERT and its Family

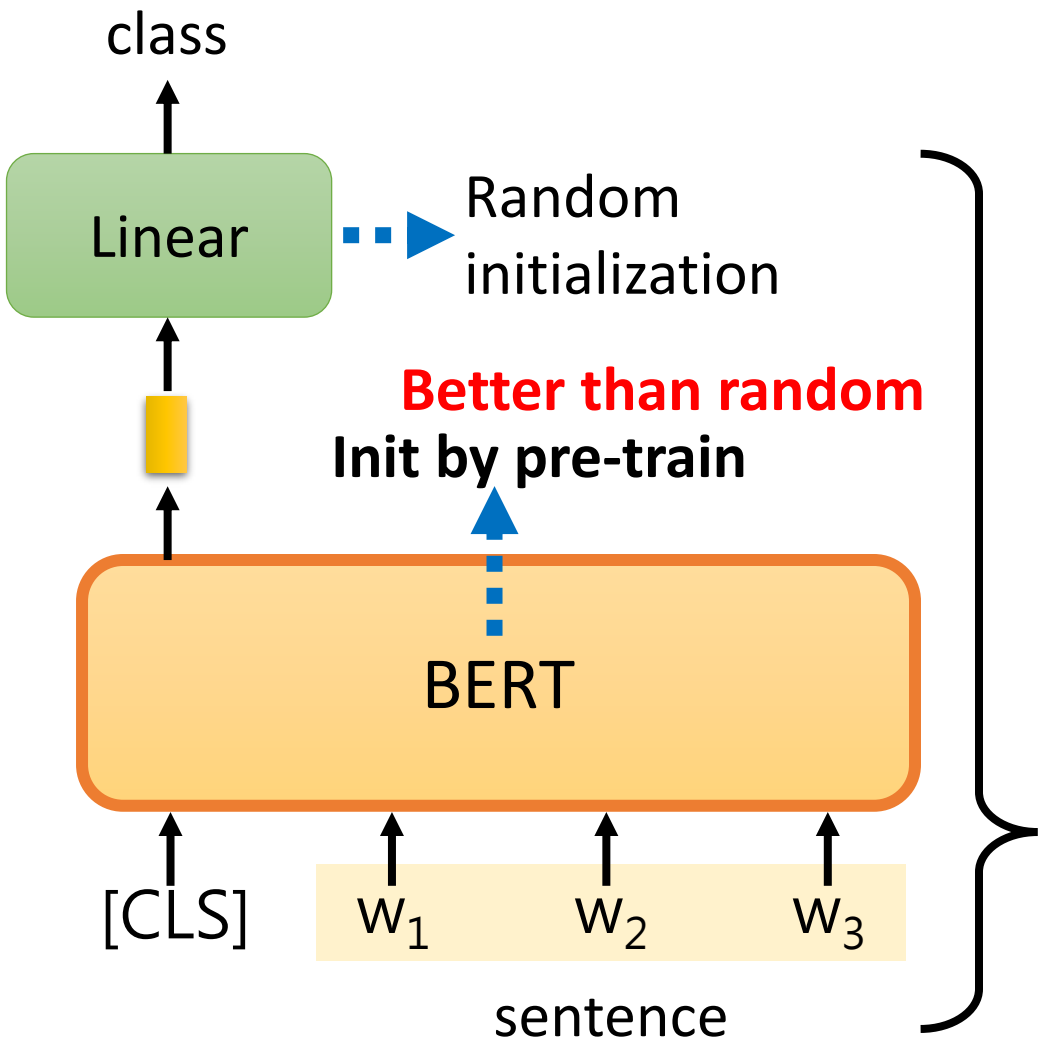
黑色這條是人類的表現
近年，綜合表現已經超越人類了
但也只是在這些資料及當中
所以後來有人提出super GLUE

- GLUE scores



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

How to use BERT – Case 1



Input: sequence
output: class

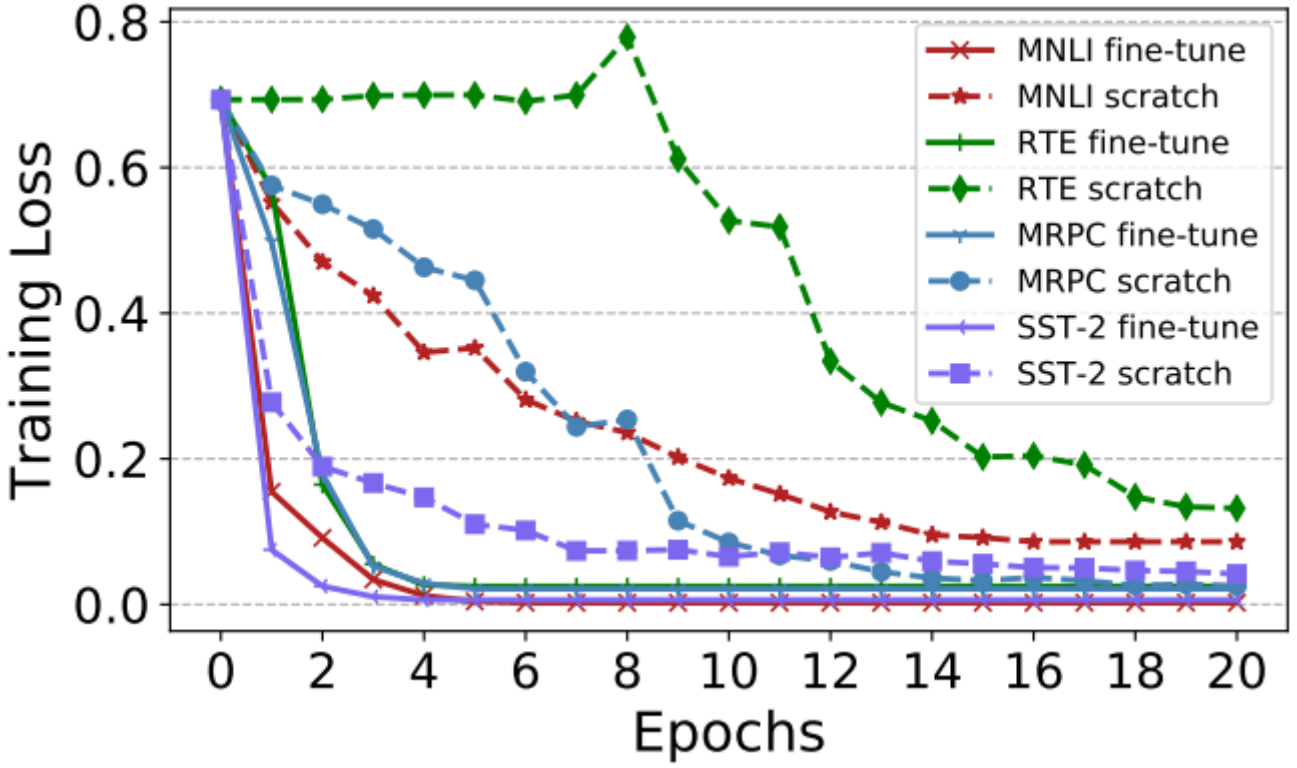
Example:
Sentiment analysis

this is good
↓
positive

This is the model to be learned.

Pre-train v.s. Random Initialization

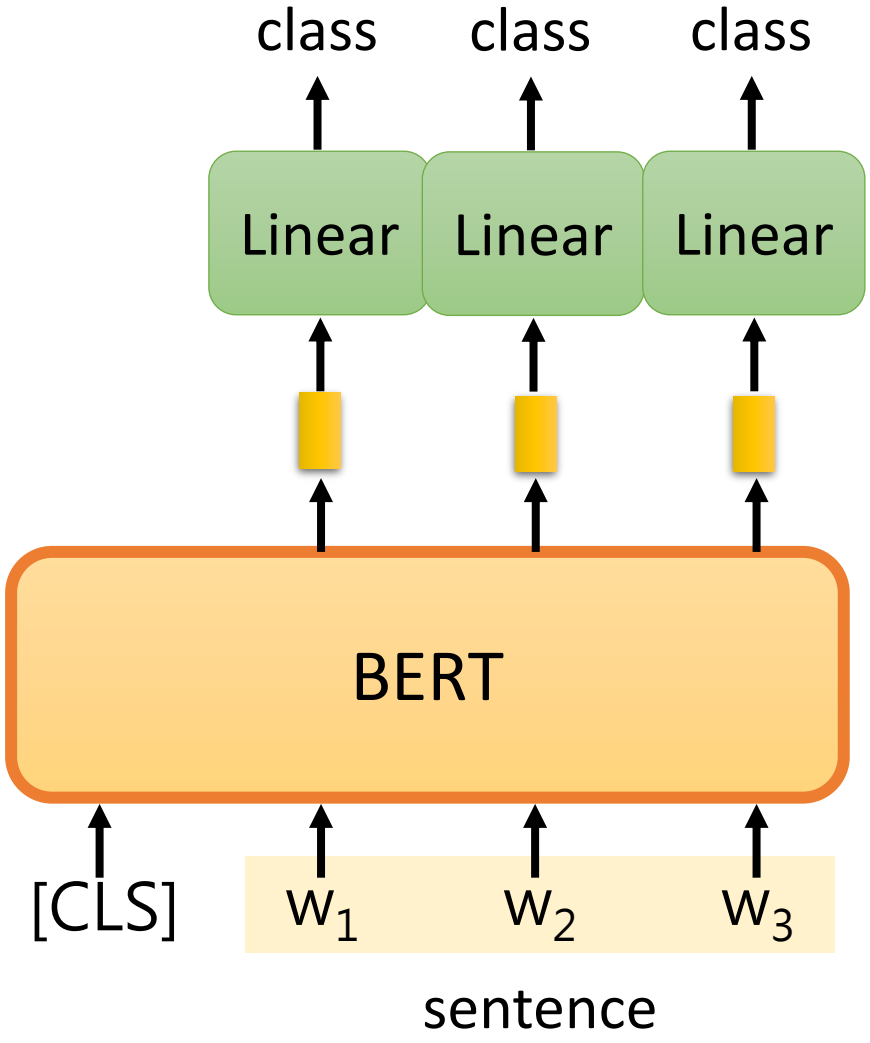
(fine-tune) (scratch)



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

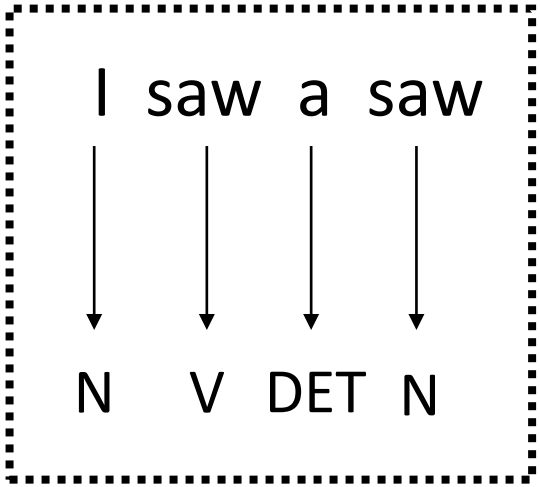

Q&A

How to use BERT – Case 2



Input: sequence
output: same as input

Example:
POS tagging

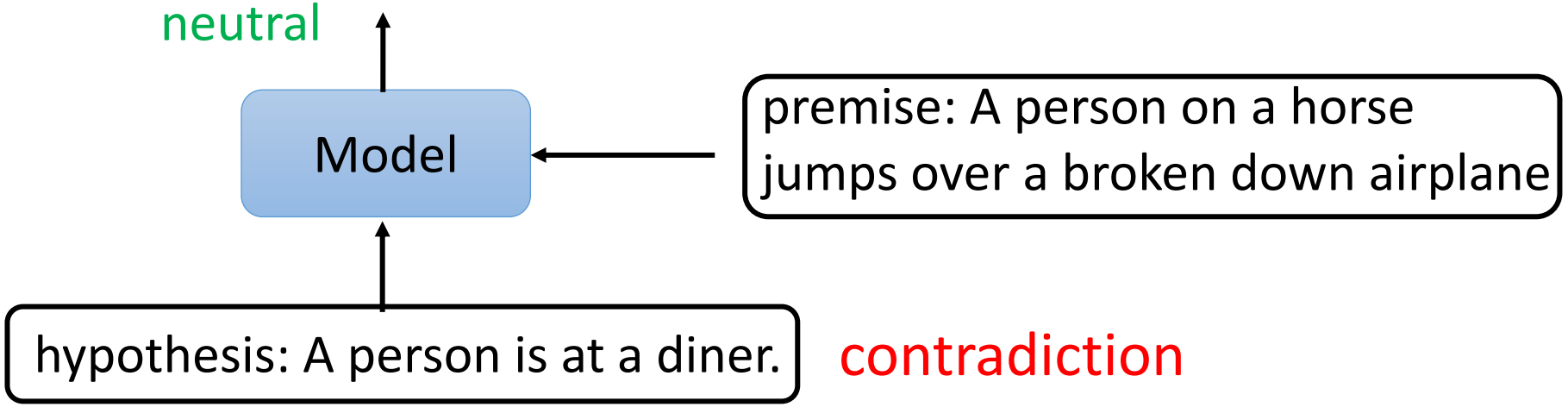


How to use BERT – Case 3

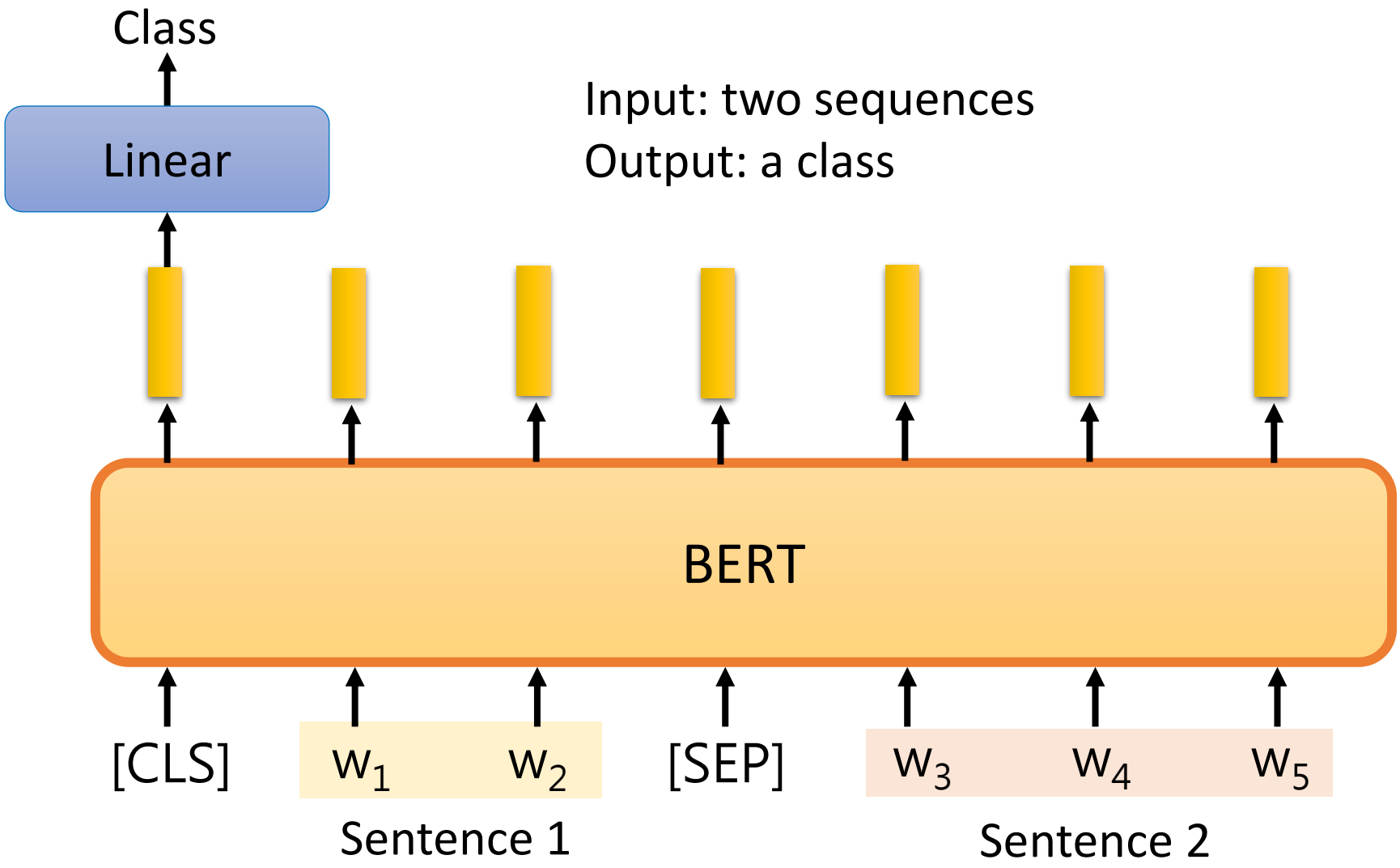
Input: two sequences
Output: a class

Example:
Natural Language Inferencee (NLI)

判斷能不能從A句子推論出B句子



How to use BERT – Case 3

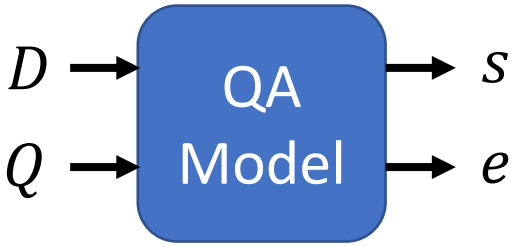


How to use BERT – Case 4

- 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 17 spheric 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 77 atte 79 cations are called "showers".

What causes precipitation to fall?
gravity

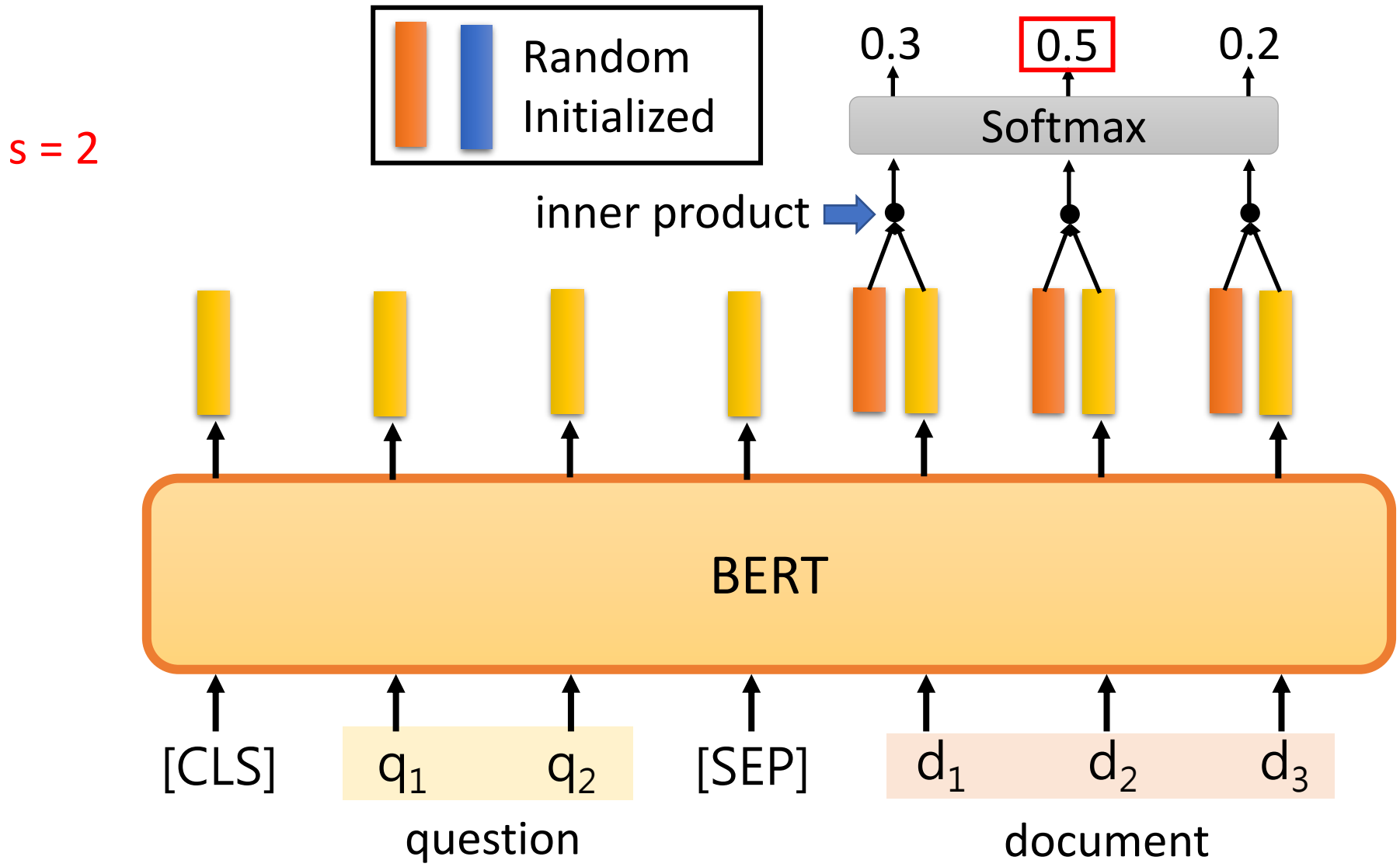
$s = 17, e = 17$

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

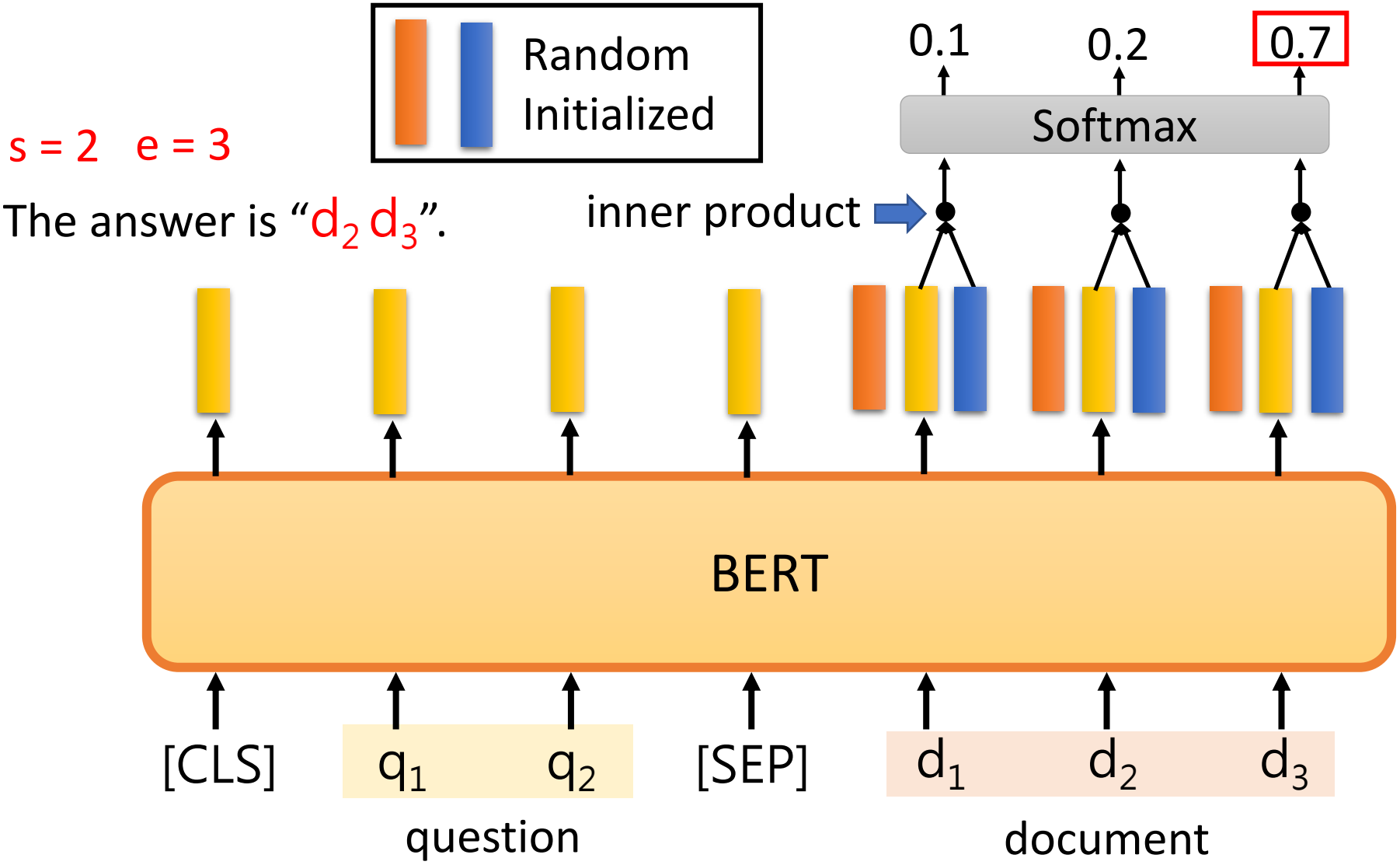
$s = 77, e = 79$

How to use BERT – Case 4



bert吃512長度的sequence就差不多了（已經要產生512*512的self attention matrix）

How to use BERT – Case 4



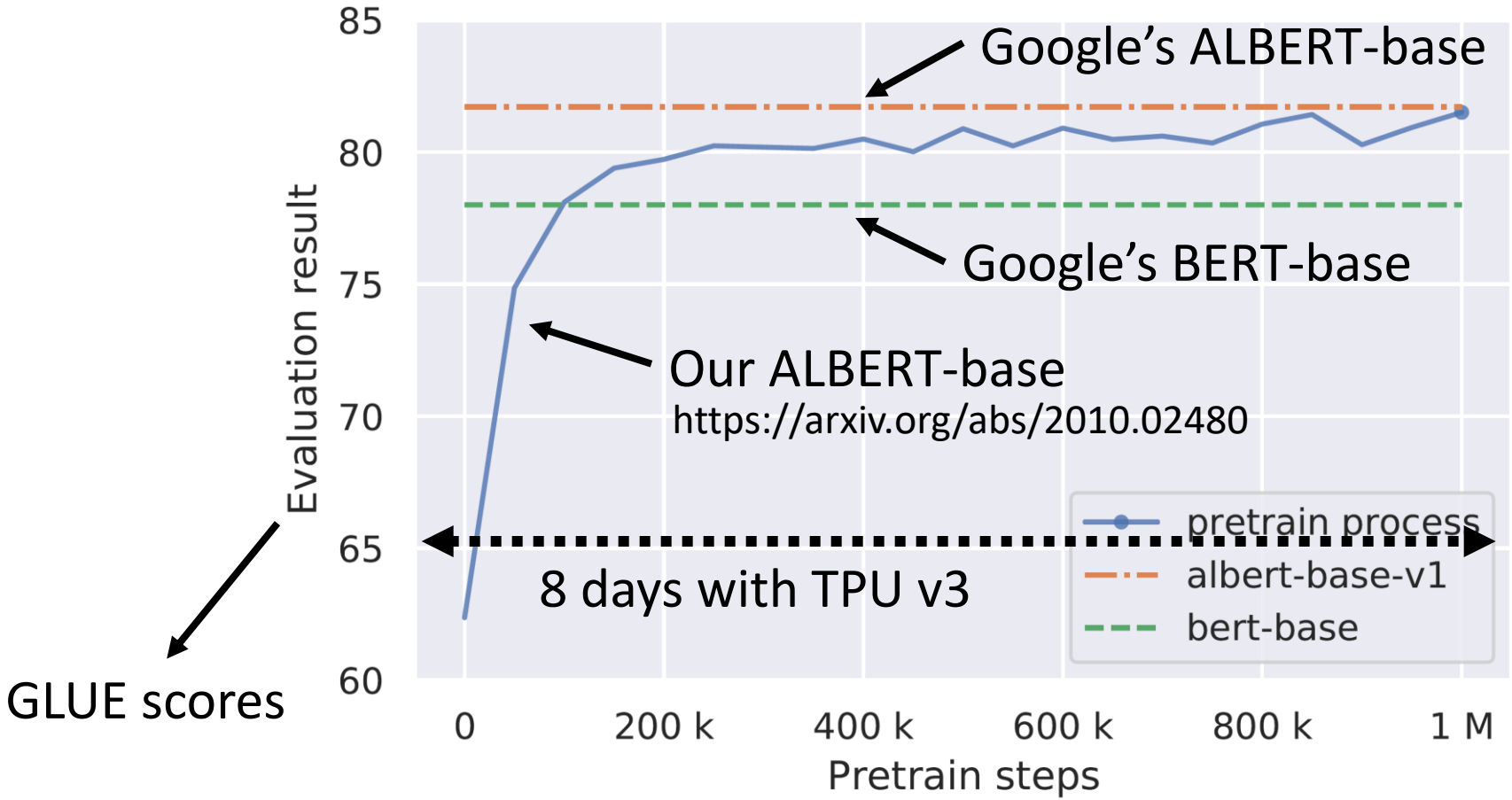
That's all!

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

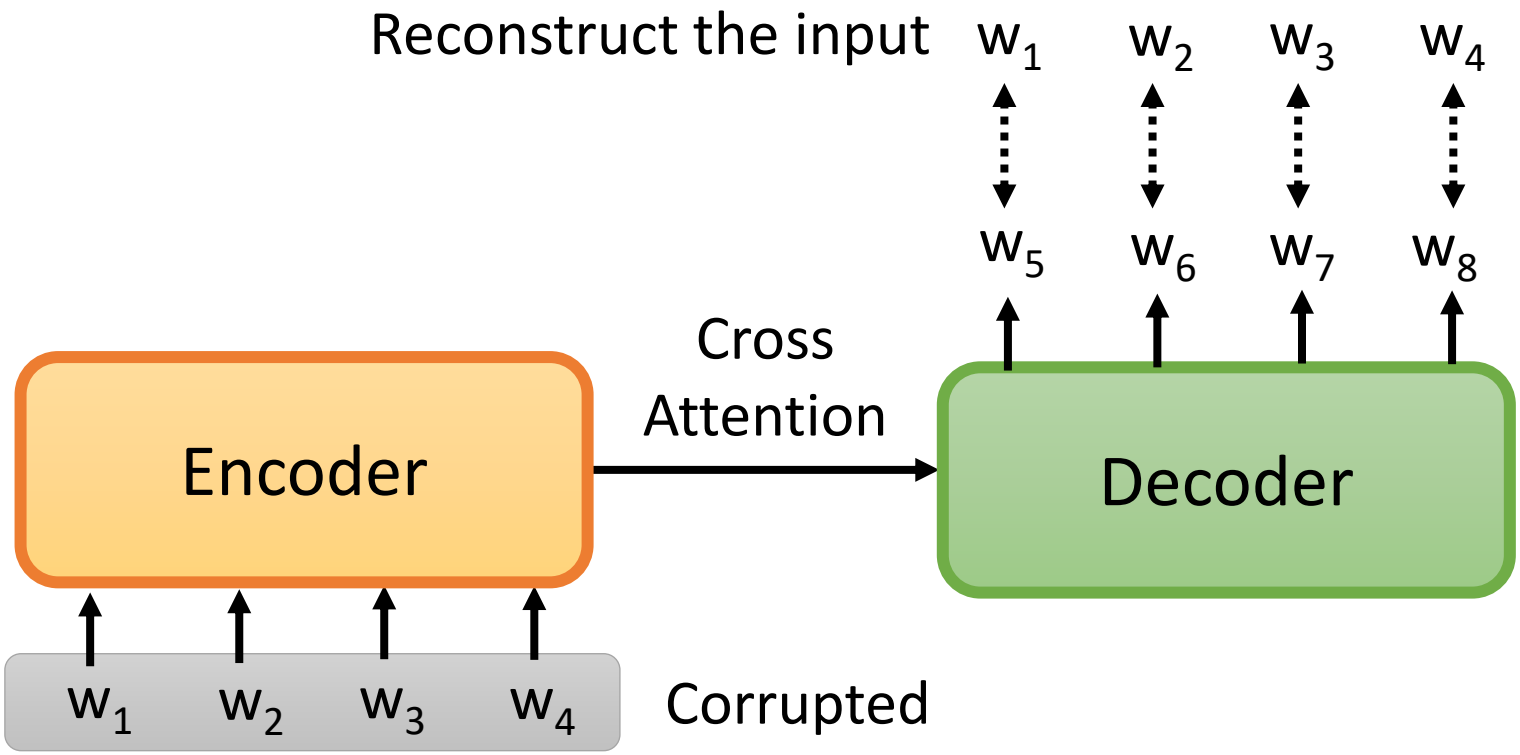
研究bert在從無到有，之間演進過程，到底在哪些階段學會了什麼？

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

The answer is counterintuitive!

Pre-training a seq2seq model

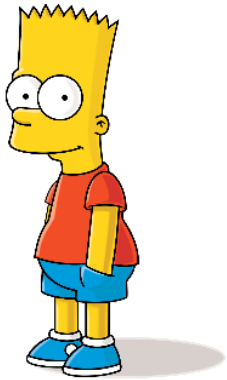
跟bert很像，都是input一個有一點壞掉的句子
然後要去做找出正確的句子



MASS / BART

把句子弄壞的方式有很多（右邊那麼多種）
然後要seq2seq把句子還原回來

MASS



BART

A B [SEP] C D E

A B [SEP] C ■ E

A B [SEP] C E
(Delete “D”)

C D E [SEP] A B
(permutation)

D E A B [SEP] C
(rotation)

A ■ B [SEP] ■ E

Text Infilling

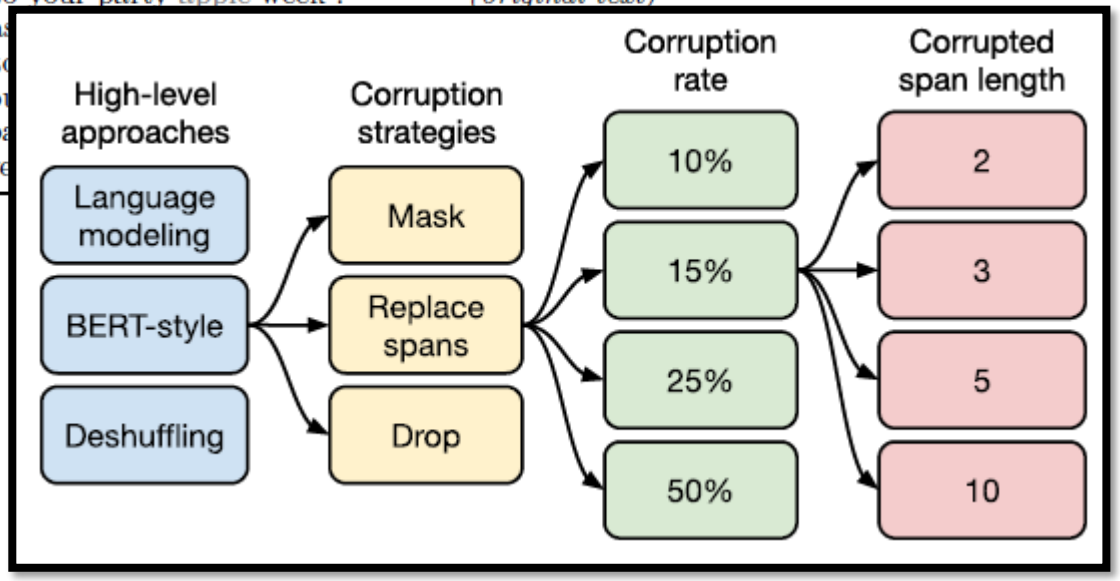
https://arxiv.org/abs/1905.02450
https://arxiv.org/abs/1910.13461

T5 – Comparison

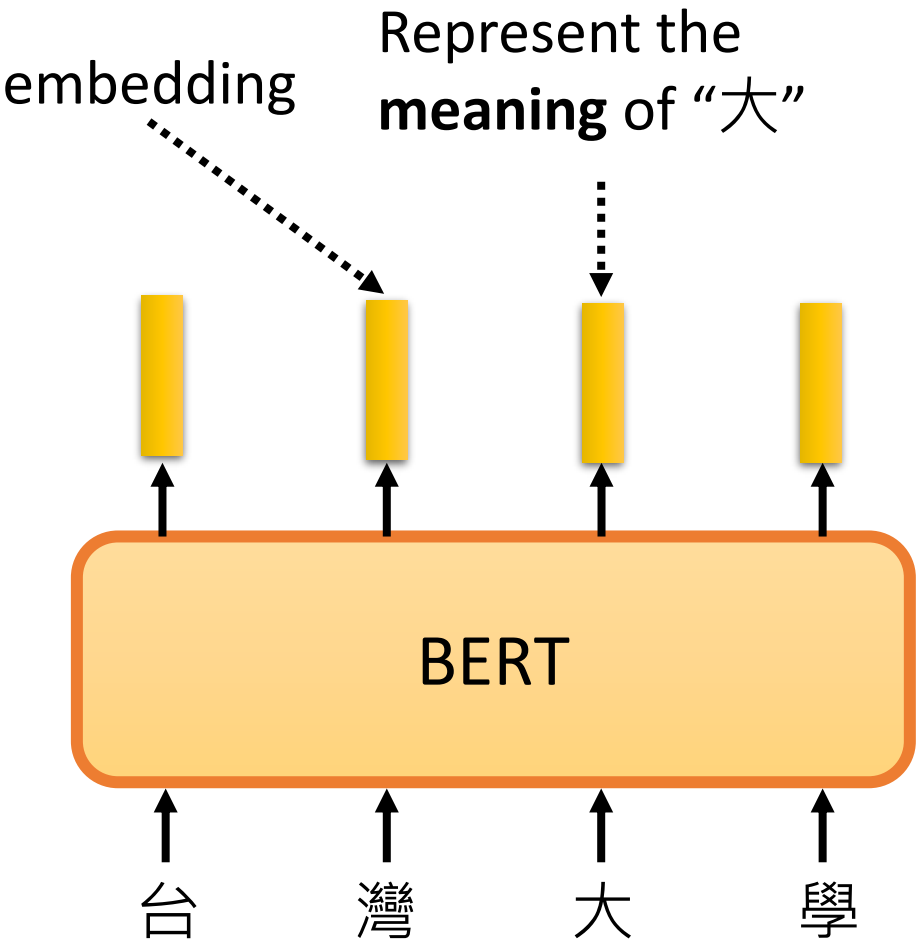
- Transfer Text-to-Text Transformer (T5)
- Colossal Clean Crawled Corpus (C4)

C4是一個資料集，原始檔案的大小有7T，google提供前處理的script，用一張gpu跑完要花355天

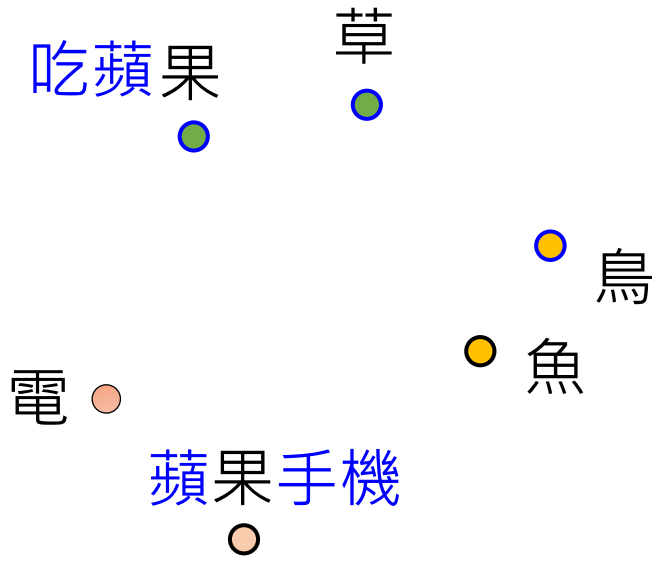
Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . las	
I.i.d. noise, mask tokens	Thank you <M> <M> me t	
I.i.d. noise, replace spans	Thank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



Why does BERT work?

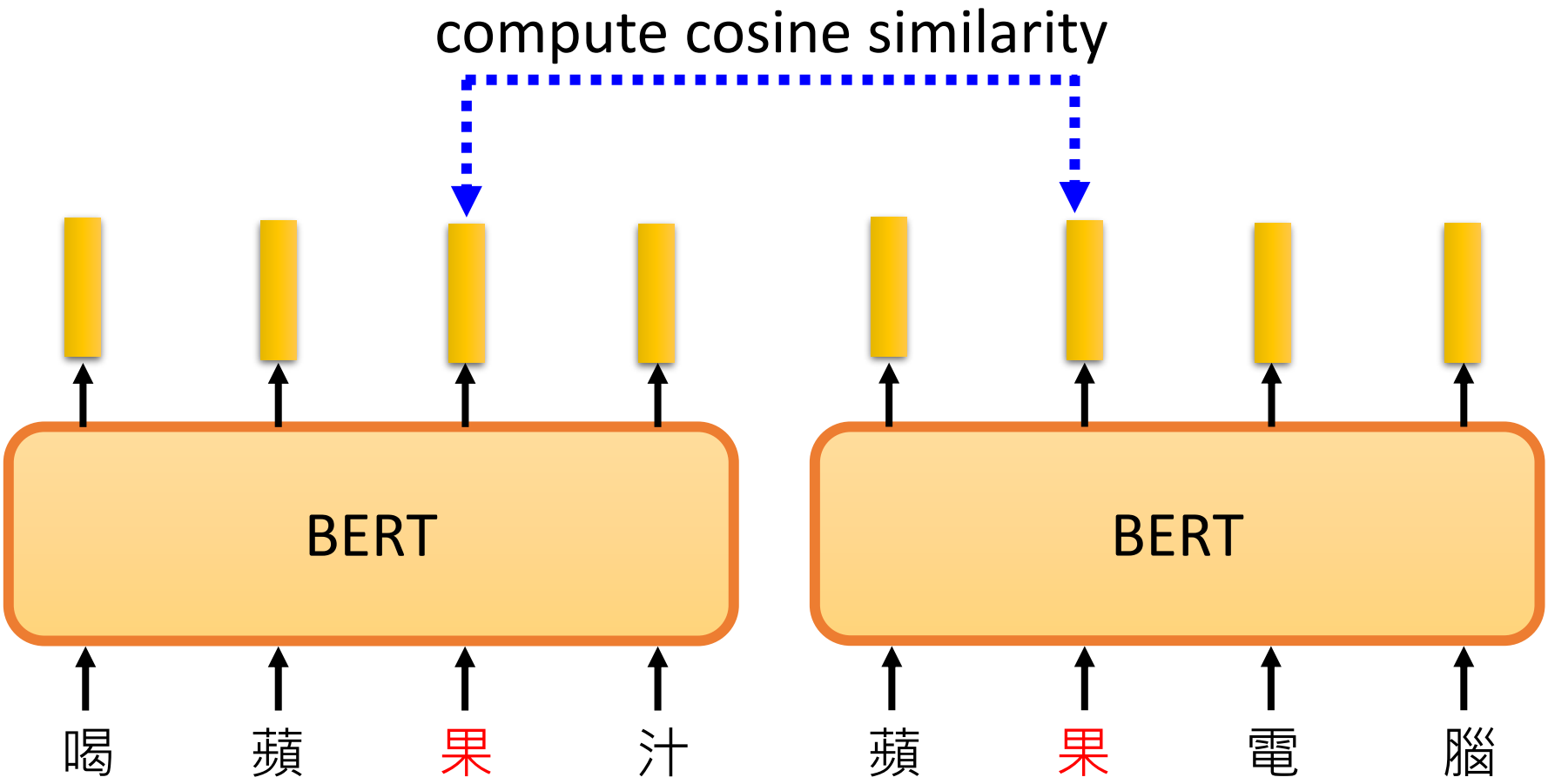


The tokens with similar meaning have similar embedding.



Context is considered.

Why does BERT work?



Cosine Similarities of BERT Embeddings

今天買了蘋果來吃

進口蘋果(富士)平均每公斤下跌12.3%

蘋果茶真難喝

老饕都知道智利的蘋果季節即將到來

進口蘋果因防止水分流失故添加人工果蠟

蘋果即將於下月發表新款 iPhone

蘋果獲新 Face ID 專利

今天買了蘋果手機

蘋果的股價又跌了

蘋果押寶指紋辨識技術

Why does BERT work?

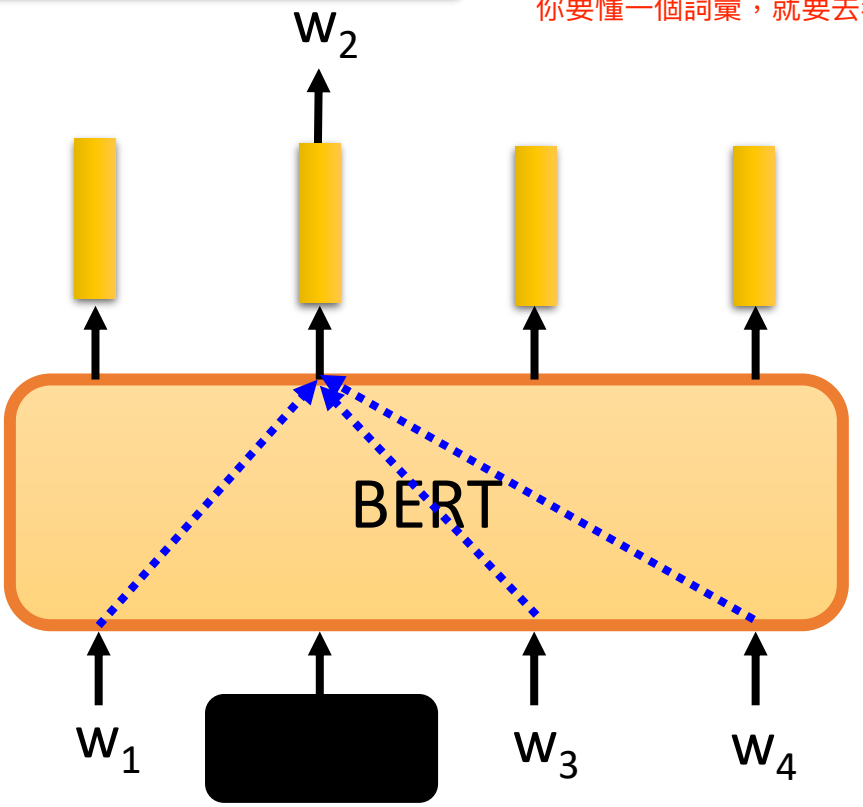
Contextualized word embedding

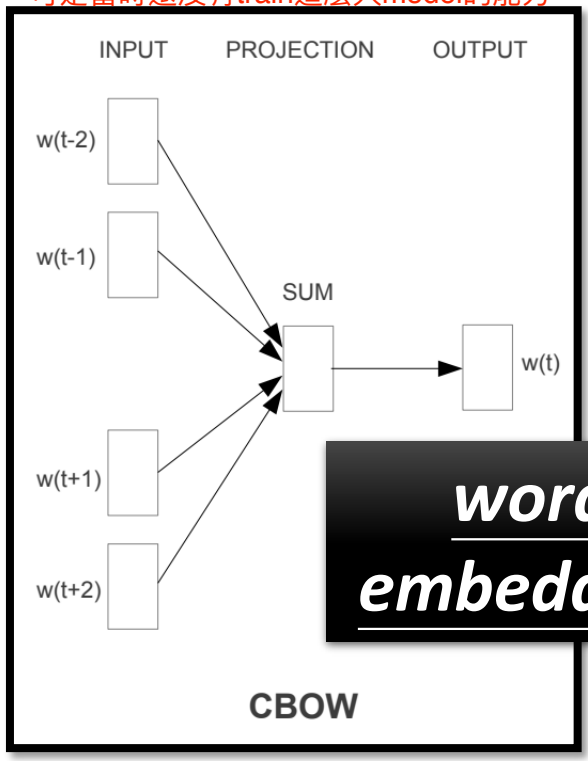
You shall know a word by the company it keeps

John Rupert Firth

一個語言學的專家：
你要懂一個詞彙，就要去看他的上下文

早就有這種概念了
可是當時還沒有train這麼大model的能力





word embedding

https://arxiv.org/abs/2103.07162

This work is done by 高瑋聰

Why does BERT work?

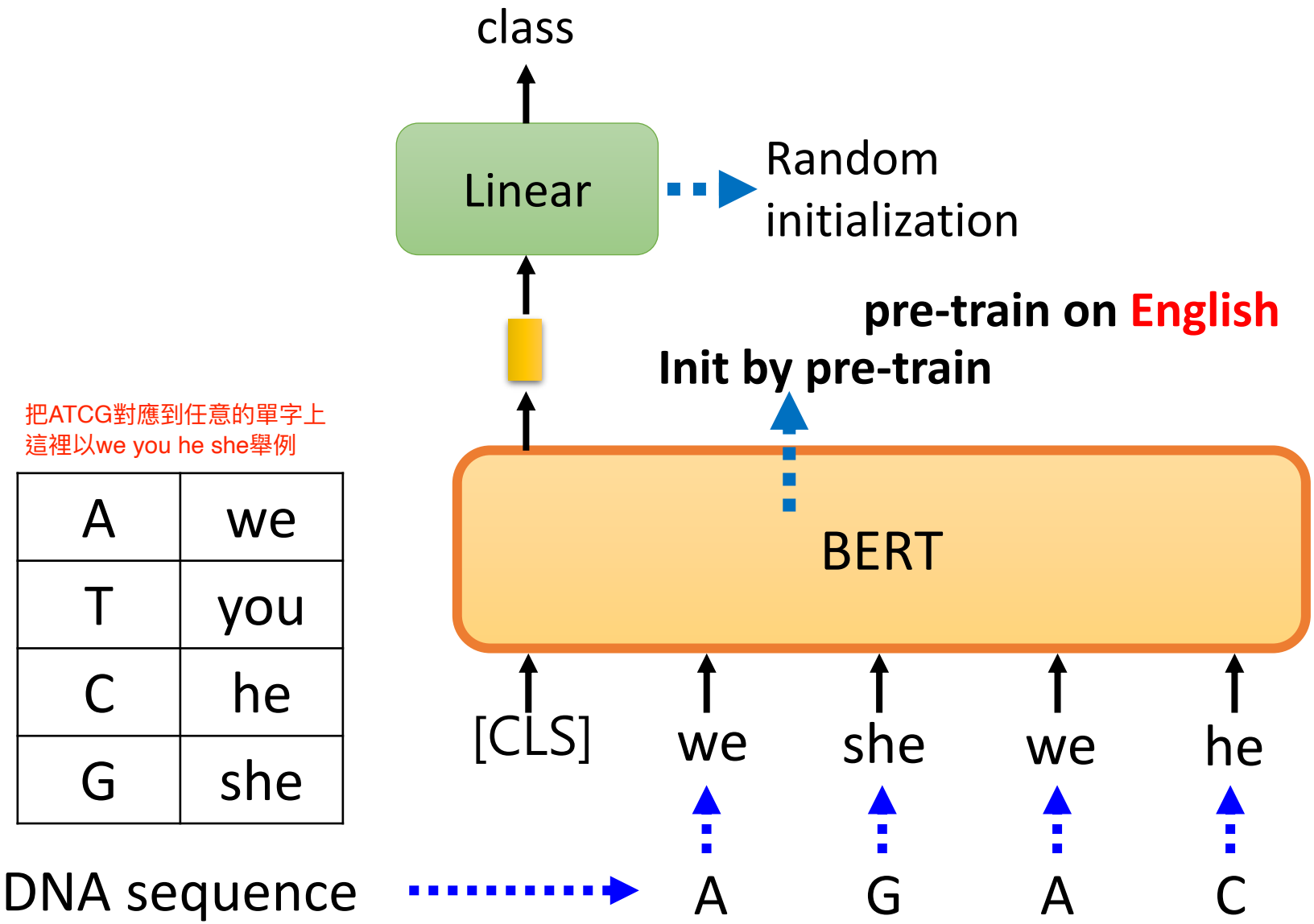
- Applying BERT to **protein**, **DNA**, **music classification**

class	DNA sequence
EI	CCAGCTGCATCACAGGAGGCCAGCG
EI	AGACCCGCCGGGAGGCGGAGGACC
IE	AACGTGGCCTCCTTGTGCCCTTCCCC
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
IE	CCTGATCTGGGTCTCCCCTCCCACCC
IE	AGCCCTCAACCCTTCTGTCTCACCCT
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
N	CTGTGTTCACCACATCAAGCGCCGGC
N	GTGTTACCGAGGGCATTTCTAACAGT
N	TCTGAGCTCTGCATTTGTCTATTCTCC

Why does BERT work?

https://arxiv.org/abs/2103.07162

This work is done by 高瑋聰



把ATCG對應到任意的單字上
這裡以we you he she舉例

A	we
T	you
C	he
G	she

https://arxiv.org/abs/2103.07162

This work is done by 高瑋聰

Why does BERT work?

結果在非語言相關上的分類問題，用bert做initialization的model會比random initialize的model還有好

- 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

To Learn More ……

BERT (Part 1)

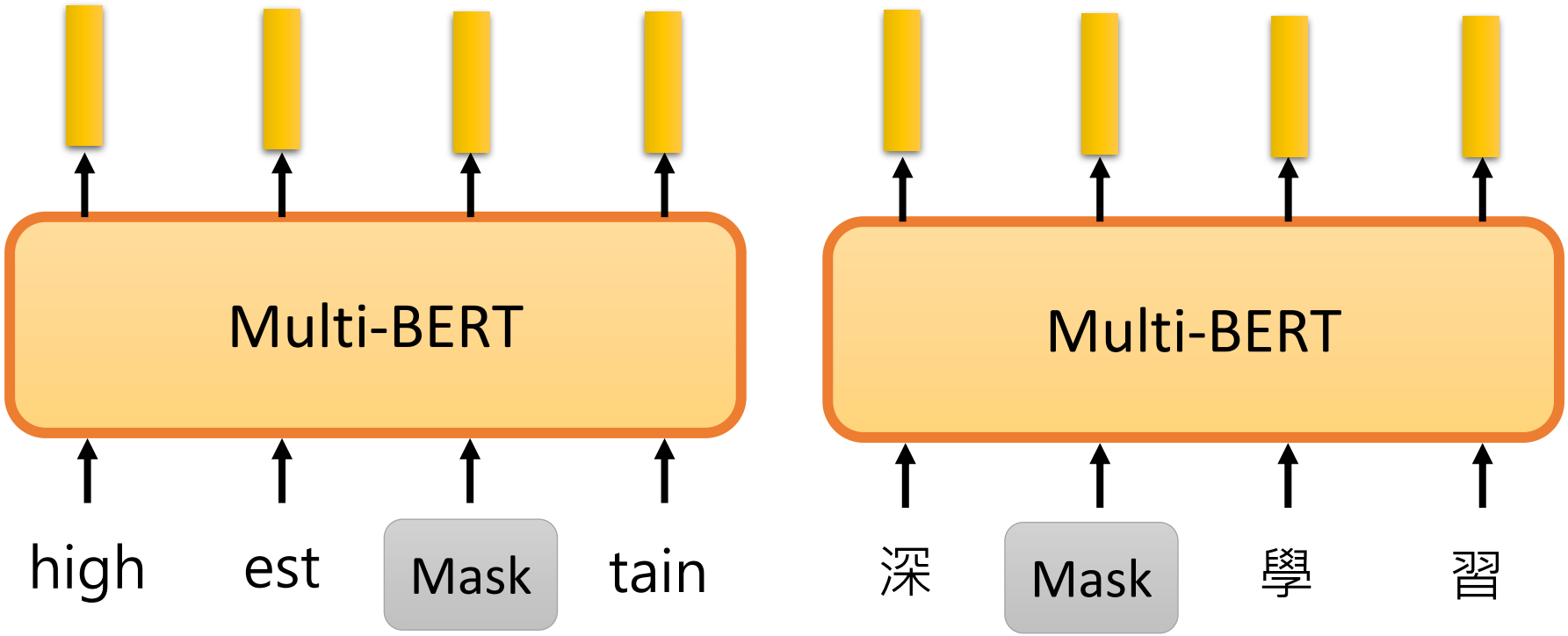
https://youtu.be/1_gRK9EIQpc

BERT (Part 2)

https://youtu.be/Bywo7m6ySlk

Multi-lingual BERT

multi-lingual bert的input不再是一種語言，而是多種語言

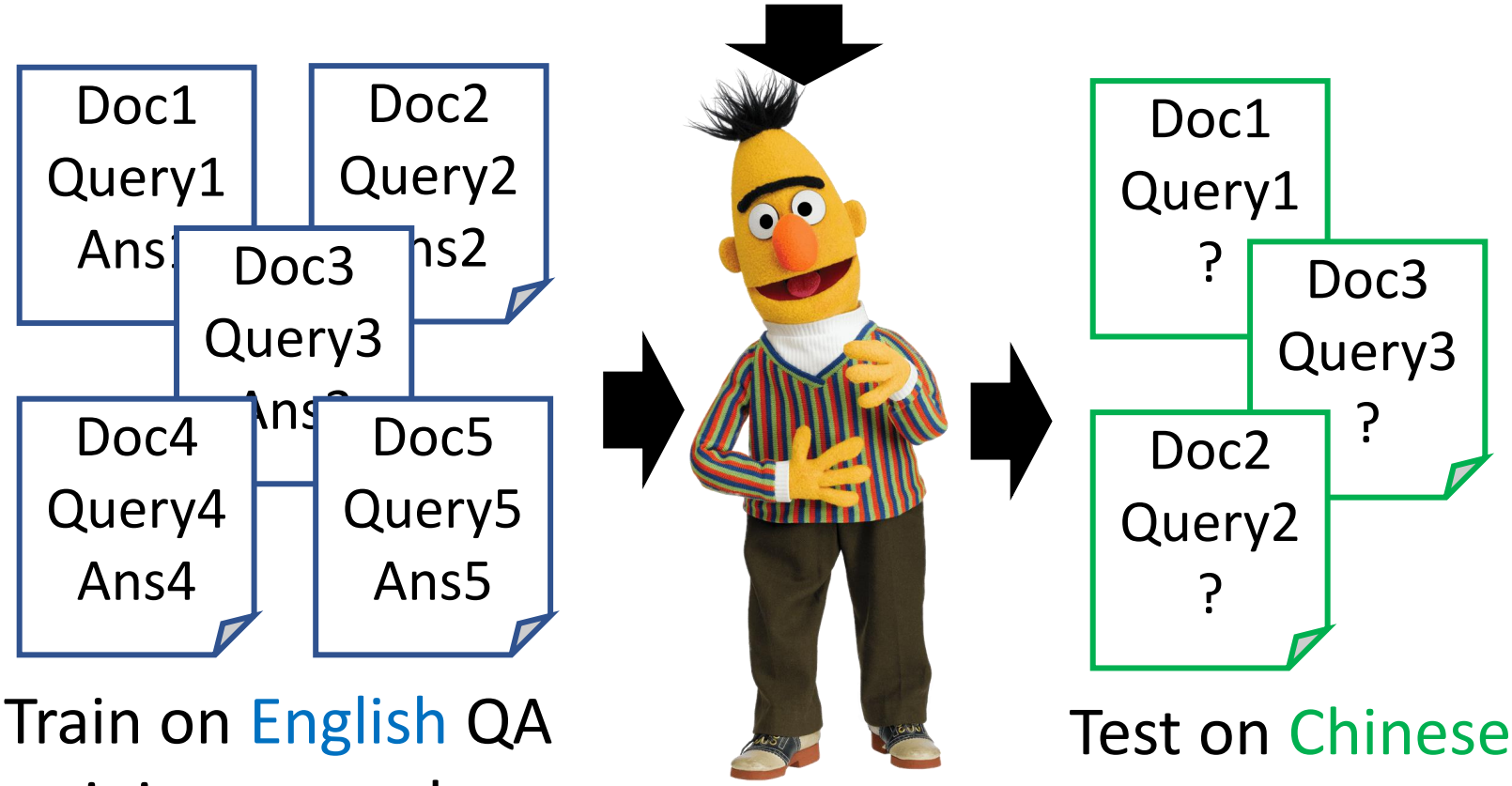


Training a BERT model by many different languages.

Zero-shot Reading Comprehension

multi-bert在做reading comprehension的問題上，只要在一種語言上做fine tune，其他語言也會學得很好

Training on the sentences of 104 languages



Train on English QA training examples

Multi-BERT

Test on Chinese QA test

Zero-shot Reading Comprehension

- English: SQuAD, Chinese: DRCD

Model	Pre-train	Fine-tune	Test	EM	F1
QANet	none	Chinese	Chinese	66.1	78.1
BERT	Chinese	Chinese		82.0	89.1
	104 languages	Chinese		81.2	88.7
		English		63.3	78.8
		Chinese + English		82.6	90.1

F1 score of Human performance is 93.30%

This work is done by 劉記良、許宗嫄
https://arxiv.org/abs/1909.09587

More about Multi-lingual BERT

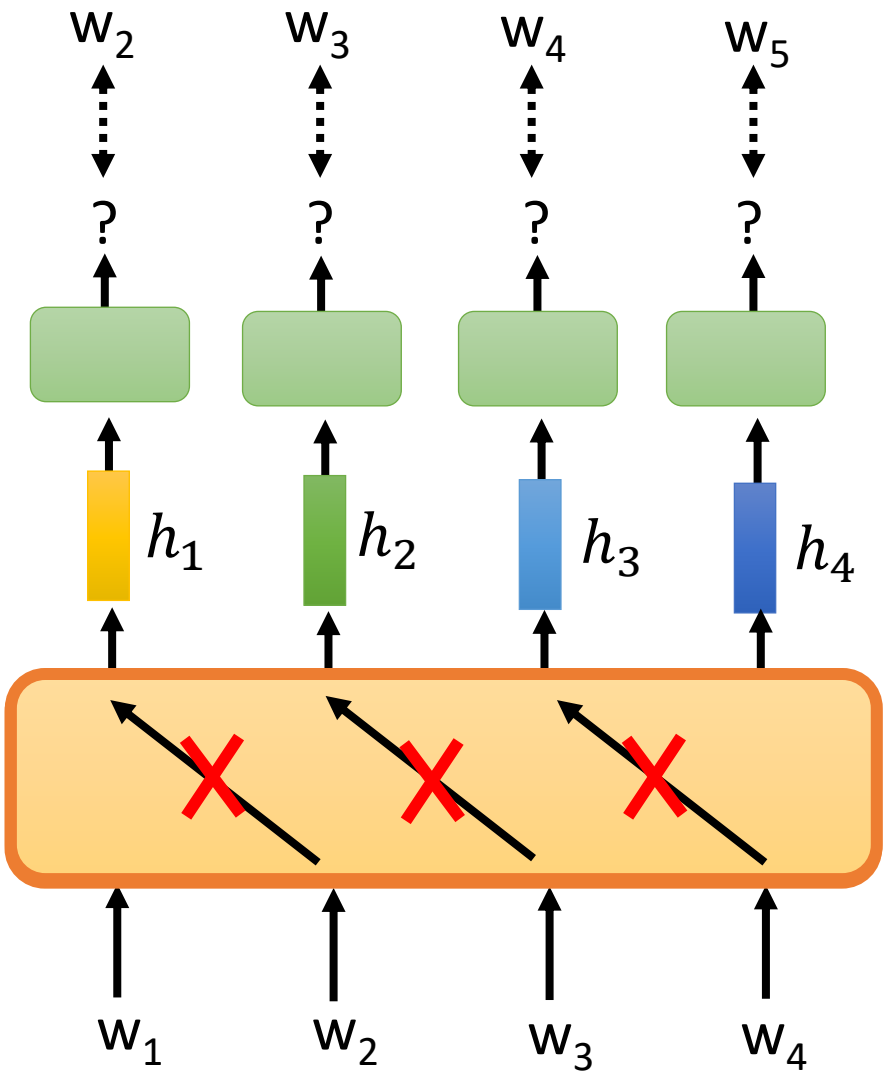
https://www.youtube.com/watch?v=8rDN1jUI82g

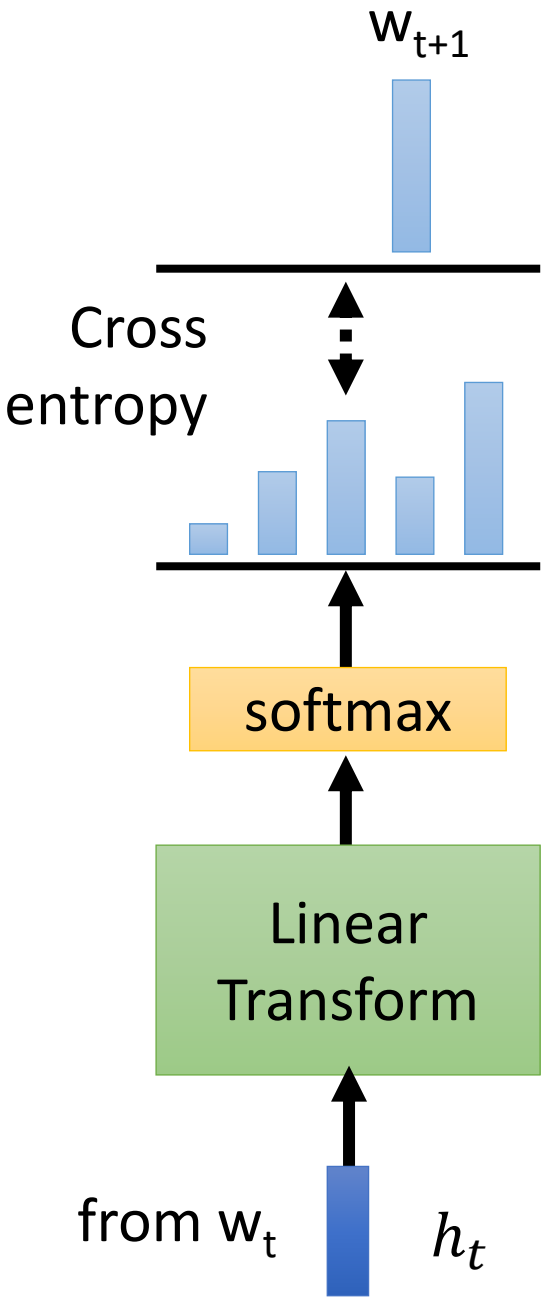
Outline

BERT series

GPT series

Predict Next Token





https://talktotransformer.com/

Predict Next Token

They can do generation.

EM PROMPT
-WRITTEN)

In a shocking finding, scientist discovered a [illegible] in a remote, previously unexplored valley, in the A[illegible]des Mount[illegible]ins. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL
MPLETION
(MACHINE-
10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Predict Next Token

They can do generation.

BATMAN

INT. TRADITIONAL BATCAVE

BATMAN stands next to his batmobile and uses his batcomputer. He's sometimes Bruce Wayne sometimes Batman. Alltimes orphan.

BATMAN
This is now a safe city. I have punched a penguin into prison.

ALFRED, Batman's loyal batler, carries a tray of goth ham.

ALFRED
Eat a dinner, Mattress Wayne.

An explosion explodes. THE JOKER and TWO-FACE enter the cave. Joker is a clown but insane. Two-Face is a man but attorney.

律師

BATMAN
No! It is Two-Face and One-Face. They hate me for being a bat.

Batman throws Alfred at Two-Face. Two-Face flips Alfred like a coin. Alfred lands heads up which means Two-Face goes home.

BATMAN (CONT'D)
It is just you and I, the Joker. Bat versus clown. Moral enemies.

THE JOKER
I am such a freak. Society is bad.
You drink water, I drink anarchy.

混亂

BATMAN
I drink bats just like a bat would!

Batman looks around for his parents, but they are still dead. This makes him have anger. He fires a batrocket. The Joker deflects it with his sick sense of humor. A clownly power.

THE JOKER
I have never followed a rule. That is my rule. Do you follow? I don't.

BATMAN
Alfred, give birth to Robin.

Alfred begins the process since it is his job. The Joker now has a present in his hand. He juggles it over to Batman.

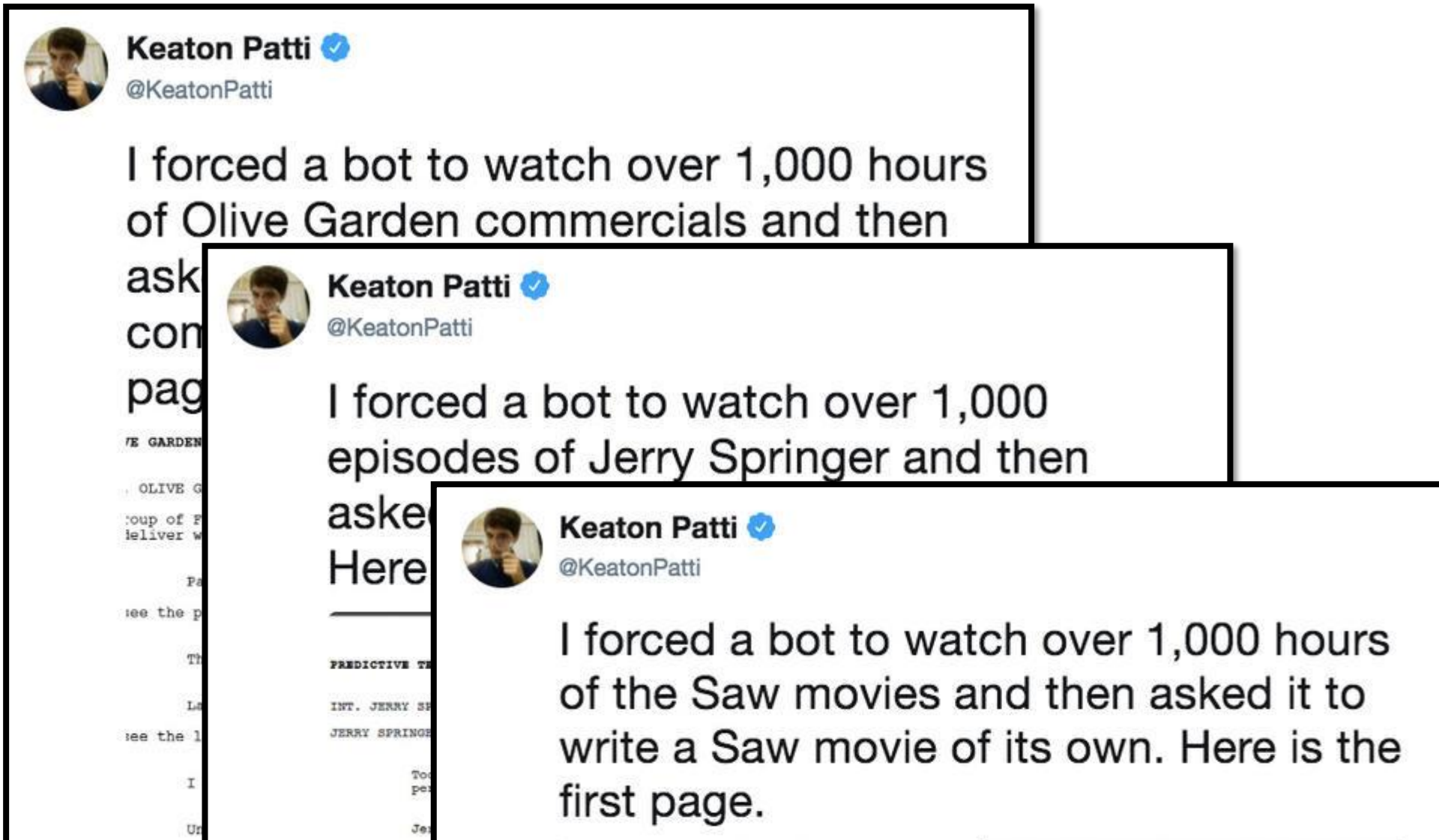
THE JOKER
Happy batday, Birthman.

Batman opens the present since he's a good guy. It contains a coupon for new parents, but is expired. This is a Joker joke.

I forced a bot to watch over 1,000 hours of XXX

是一個梗!

人在模仿機器模仿人!!!



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。

Description

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。

A few example

"In-context" Learning

Few-shot Learning

(no gradient descent)

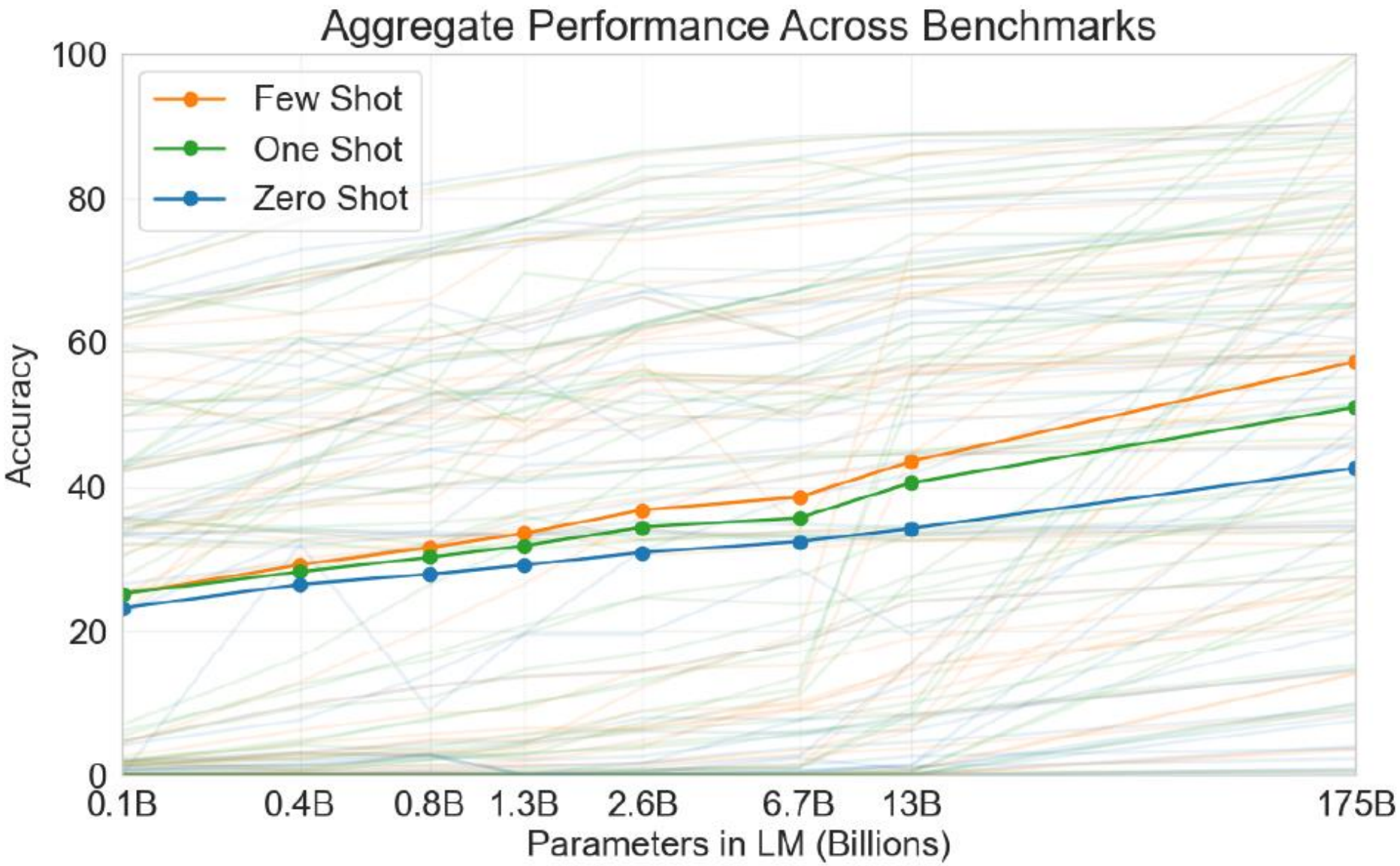
```
1  Translate English to French:          ← task description
2  sea otter => loutre de mer            ← examples
3  peppermint => menthe poivrée          ←
4  plush girafe => girafe peluche        ←
5  cheese =>                             ← prompt
```

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
```



Average of **42** tasks

To learn more ……

https://youtu.be/DOG1L9lvsDY

Image - SimCLR

https://arxiv.org/abs/2002.05709
https://github.com/google-research/simclr

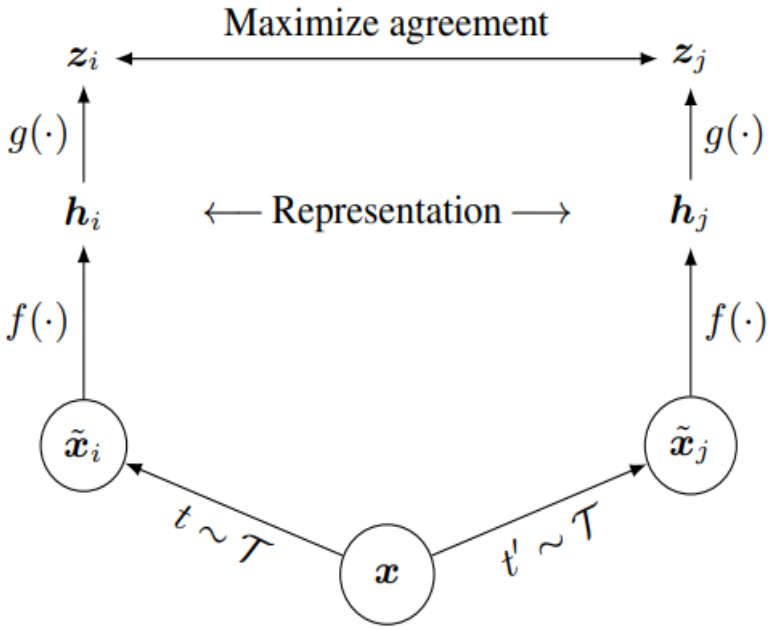
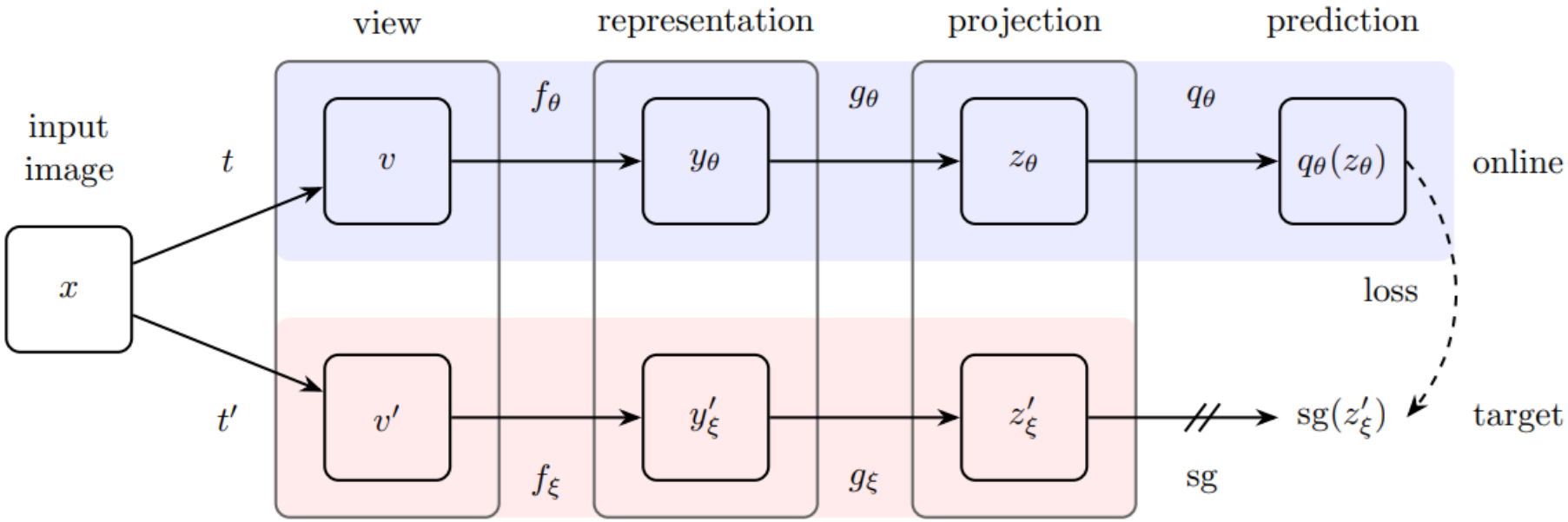


Image - BYOL

Bootstrap **y**our **o**wn **l**atent:
A new approach to self-supervised Learning

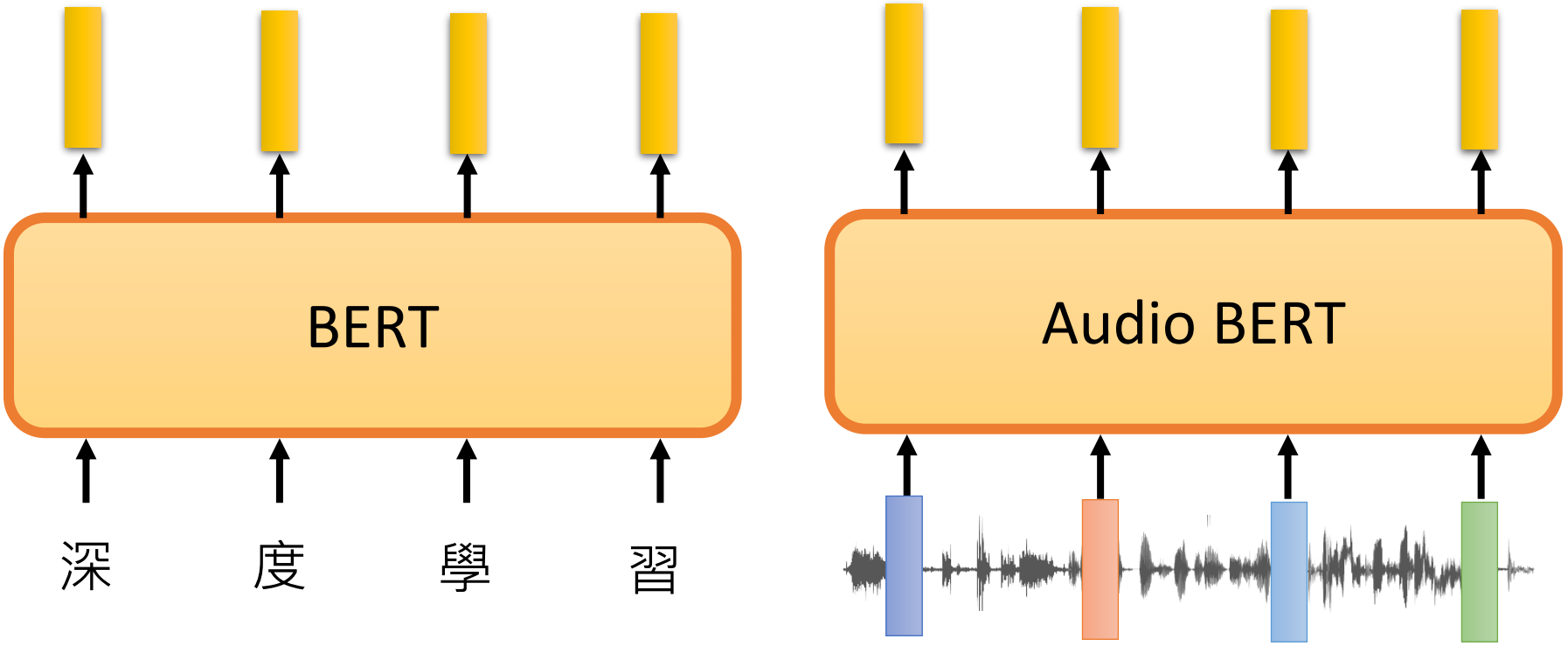
https://arxiv.org/abs/2006.07733

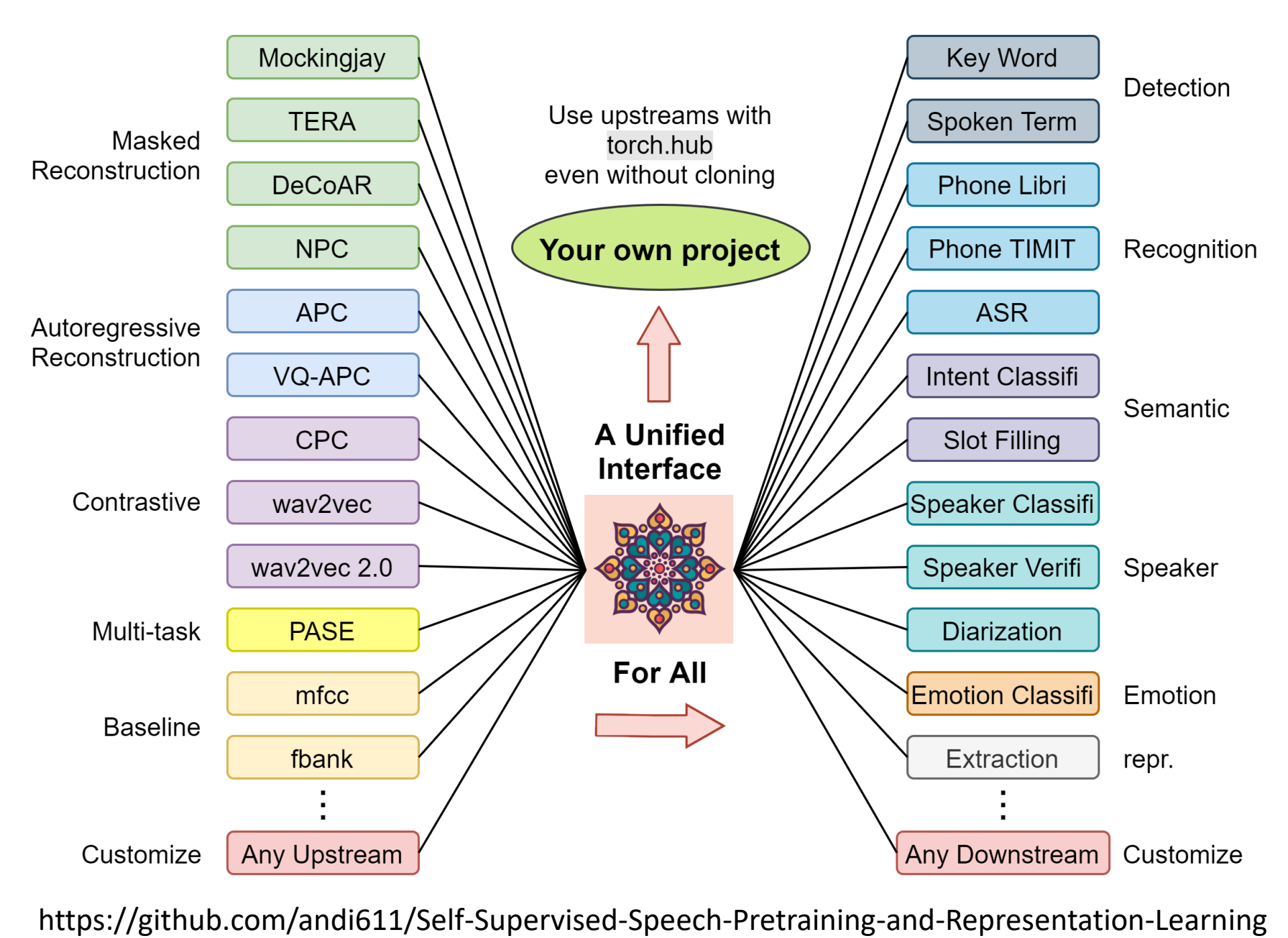


Speech

Audio version of GLUE - SUPERB

Speech processing **U**niversal **PER**formance **B**enchmark





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