

<https://www.sesameworkshop.org/what-we-do/sesame-streets-50th-anniversary>



Self-Supervised Learning

Hung-yi Lee 李宏毅

死臭酸宅本人

芝麻街

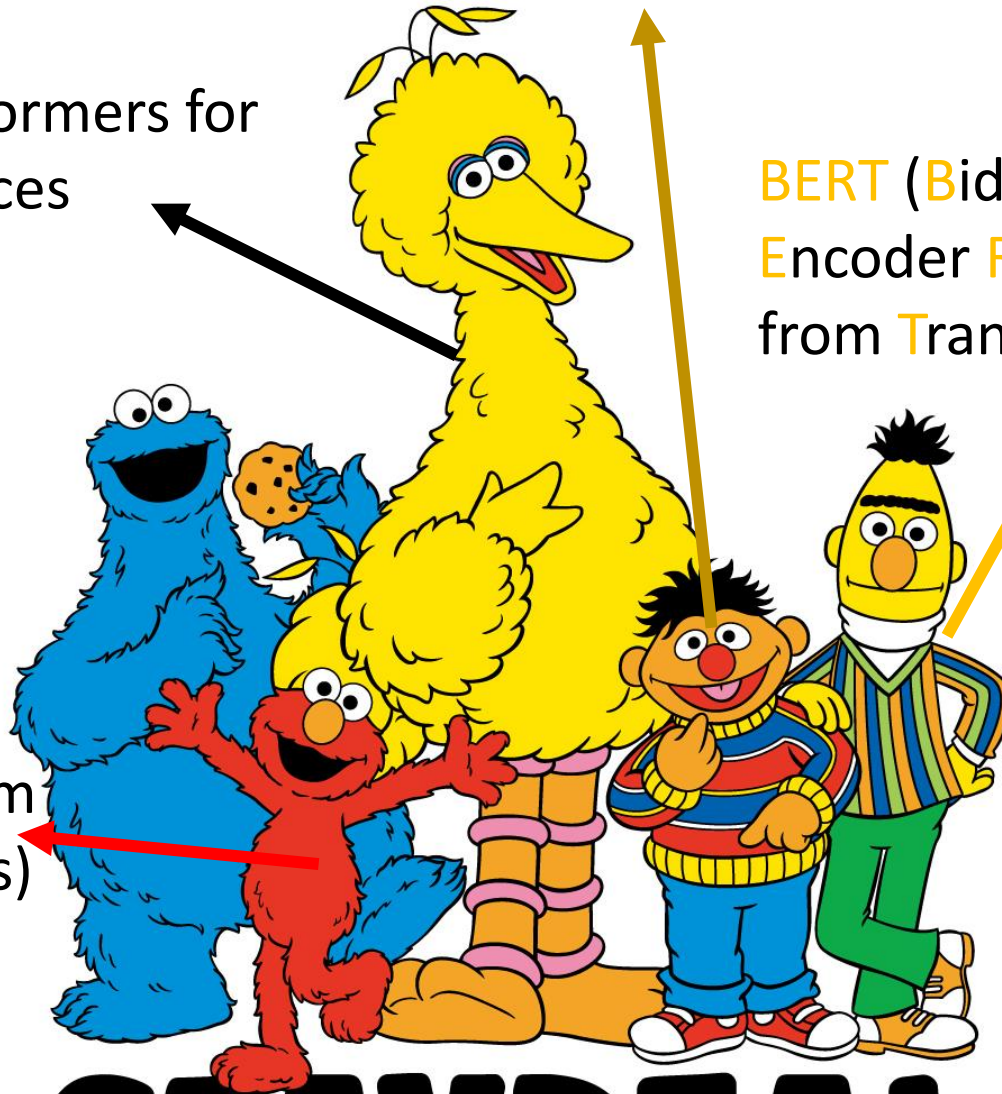


ERNIE (Enhanced Representation
through Knowledge Integration)

Big Bird: Transformers for
Longer Sequences

BERT (Bidirectional
Encoder Representations
from Transformers)

ELMo
(Embeddings from
Language Models)



STAYREAL



BERT

**340M
parameters**



**Bertolt
Hoover**

Source of image:

https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html

GPT-3

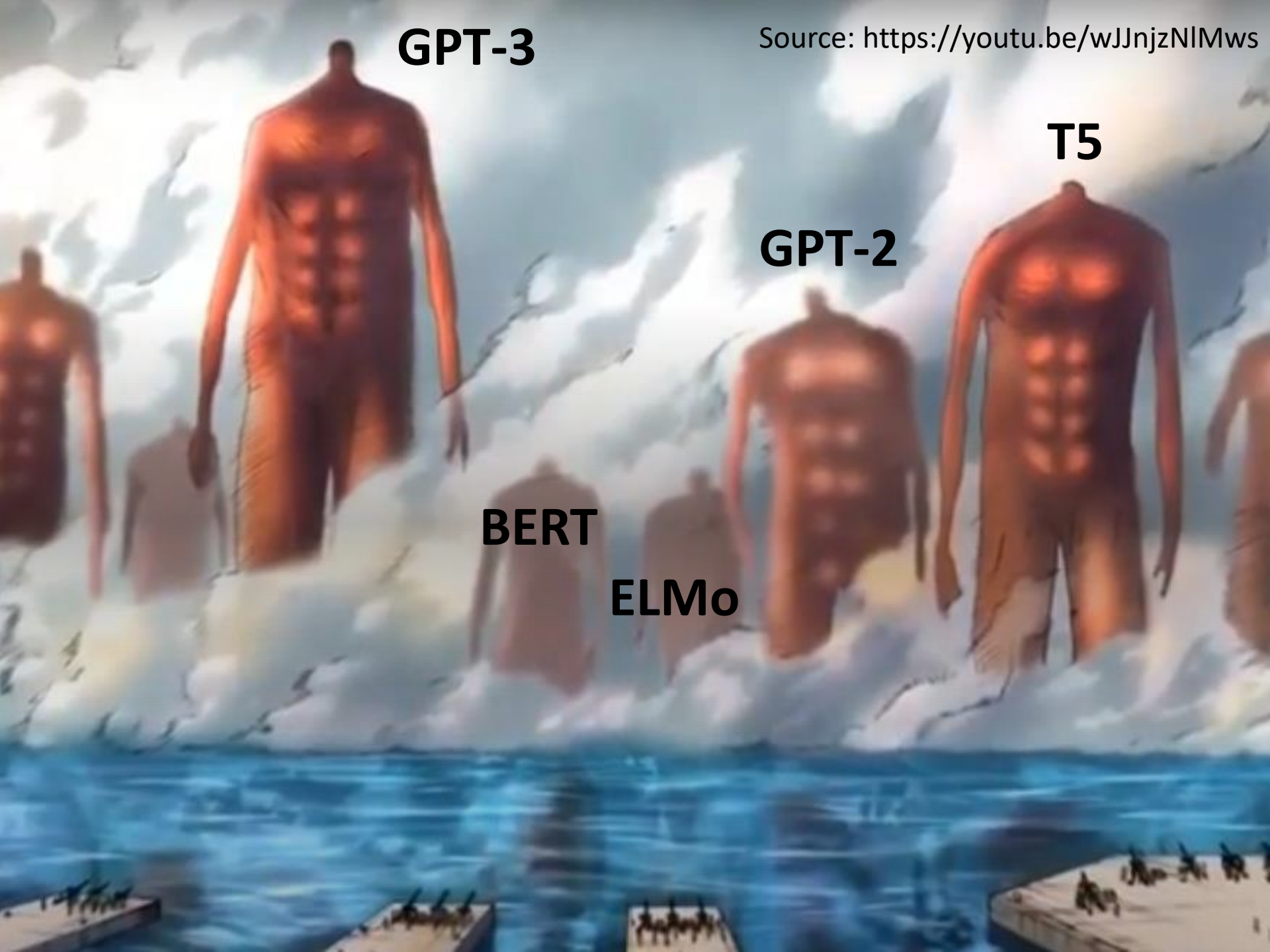
Source: <https://youtu.be/wJJnjzNIMws>

T5

GPT-2

BERT

ELMo

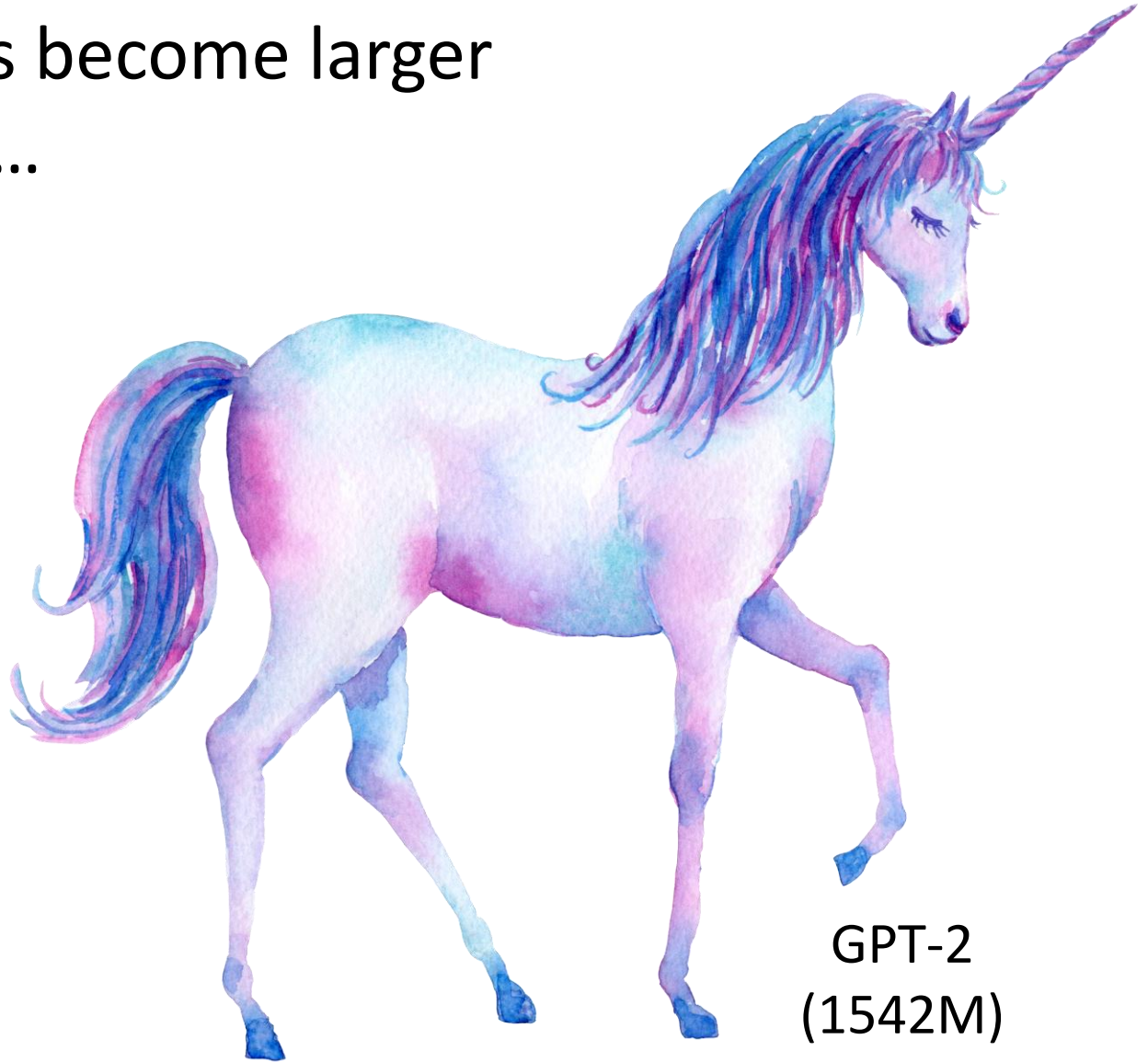


The models become larger
and larger ...

ELMO
(94M)



BERT
(340M)



GPT-2
(1542M)

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

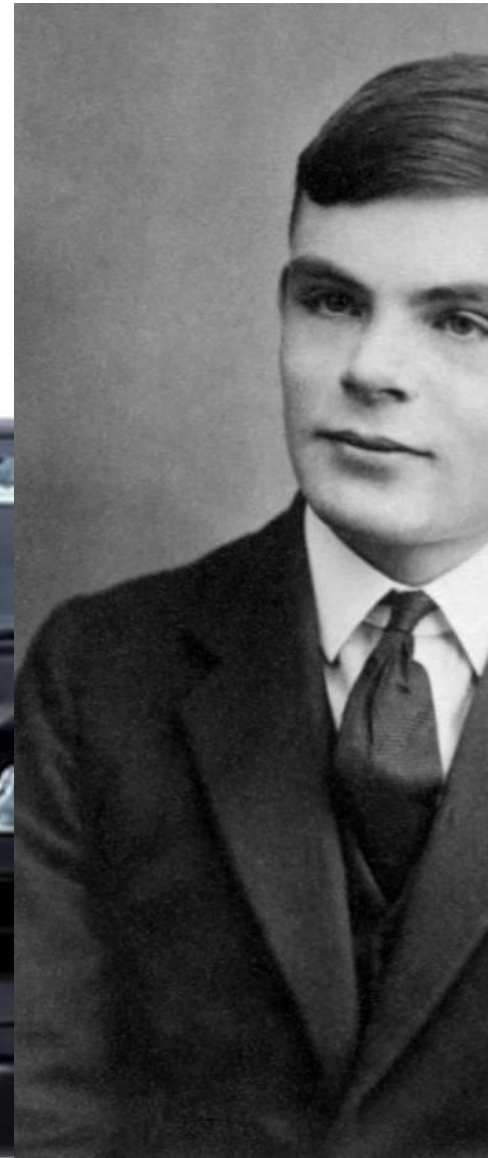
The models become larger
and larger ...

Turing NLG
(17B)

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



TRANSFORMERS
MEGATRON



GPT-2

Megatron (8B)

T5 (11B)



BERT (340M)

GPT-3 (175B)

Switch

Transformer (1.6T)

<https://arxiv.org/abs/2101.03961>



Outline



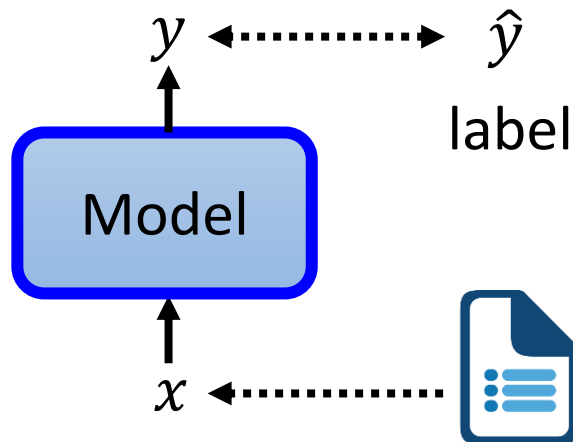
BERT series



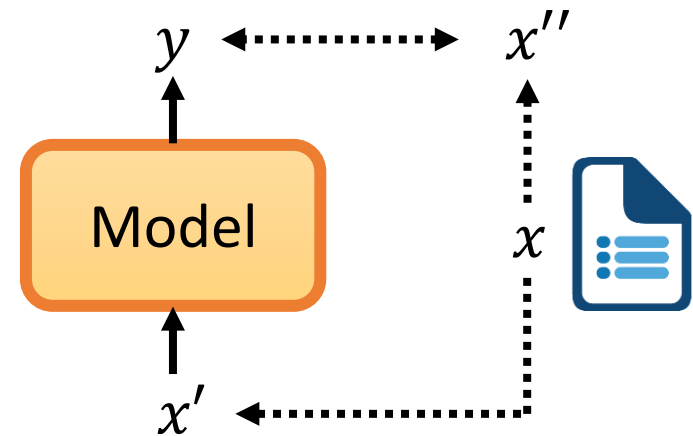
GPT series

Self-supervised Learning

Supervised



Self-supervised



Yann LeCun

2019年4月30日 · 🌐

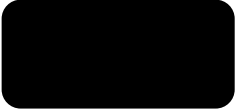

I now call it "self-supervised learning". because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of its 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

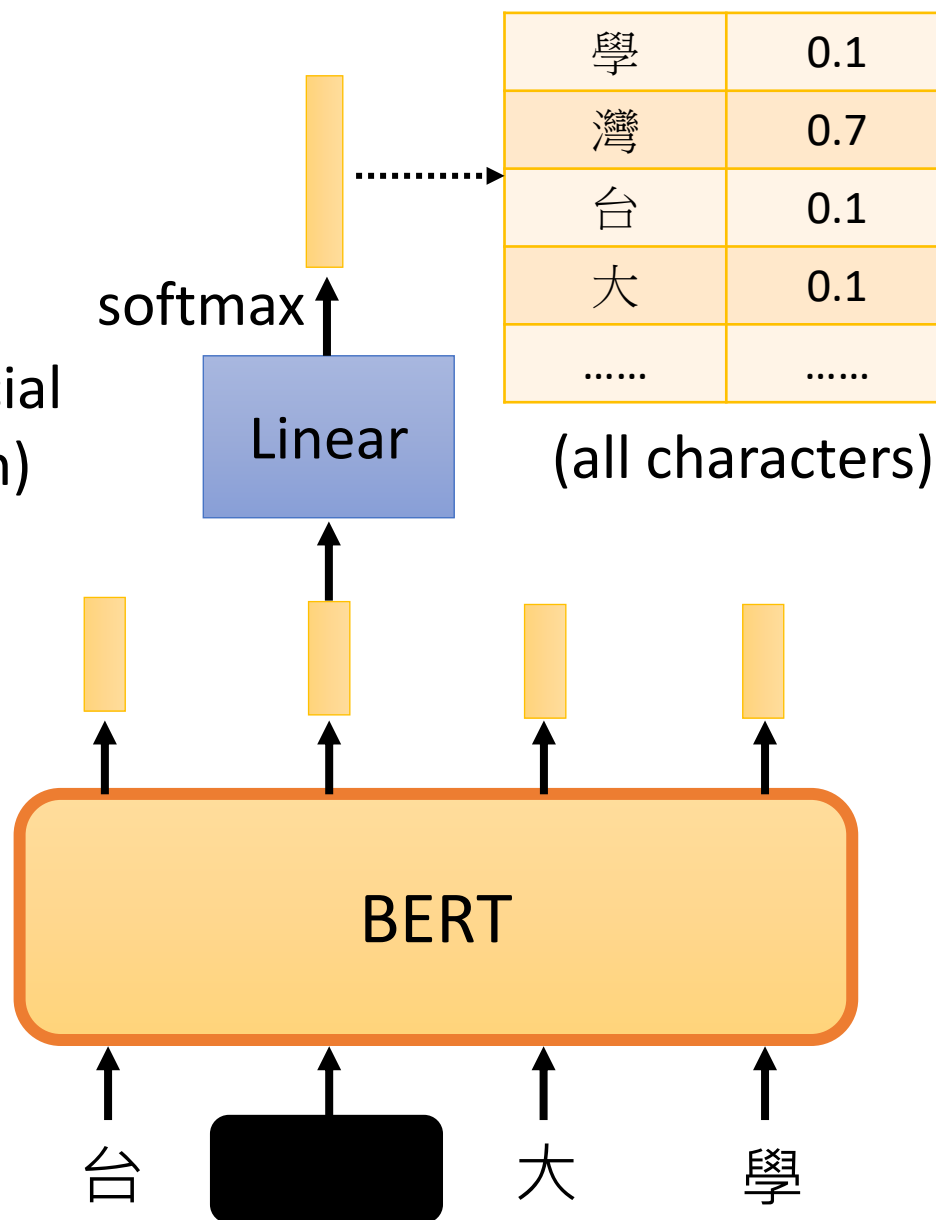
<https://arxiv.org/abs/1810.04805>

 =  (special token)
or

 = 
一、天、大、小 ...

Transformer
Encoder

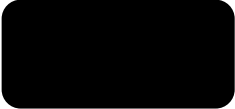

Randomly masking
some tokens



Masking Input

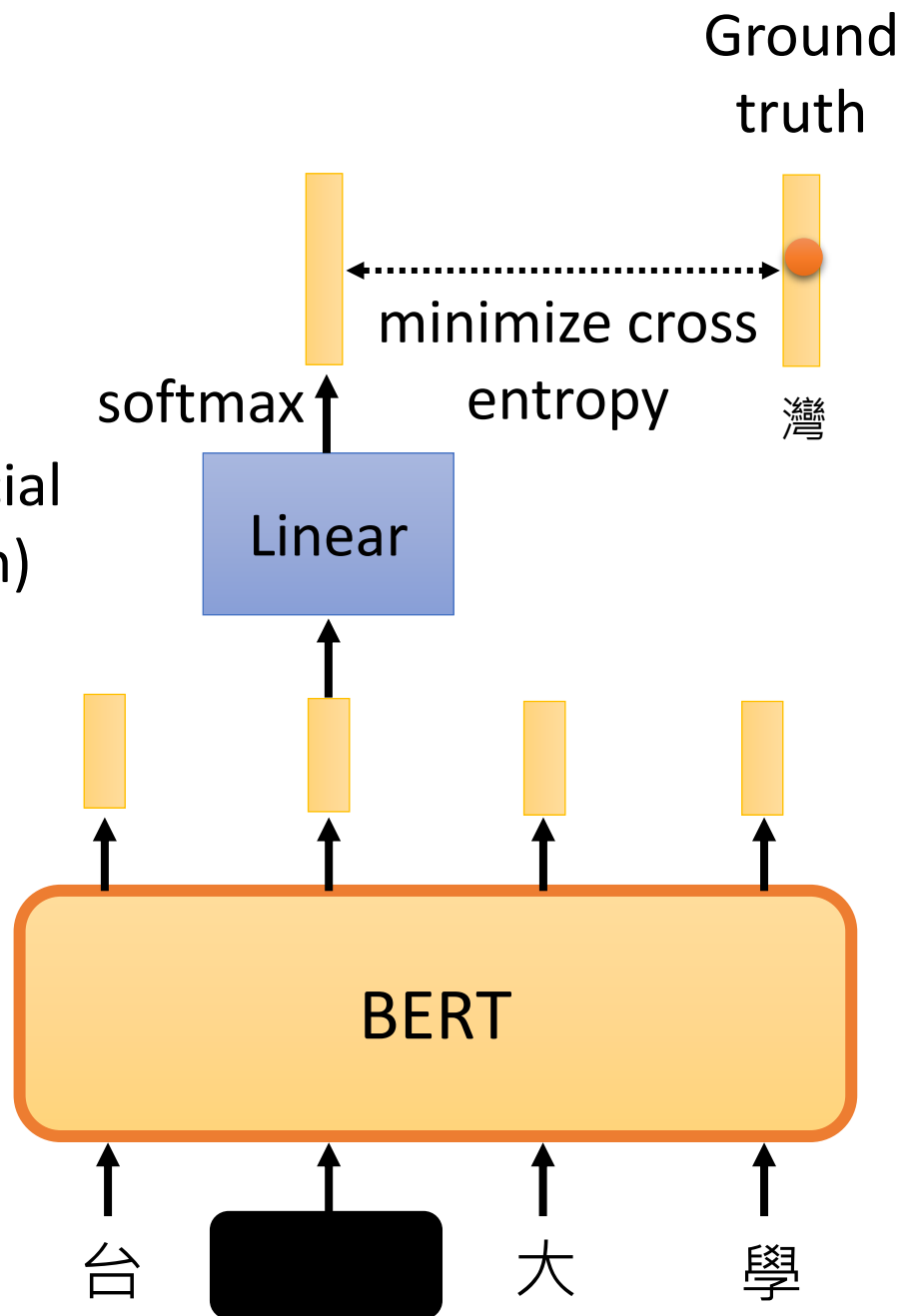
<https://arxiv.org/abs/1810.04805>

 =  (special token)
or

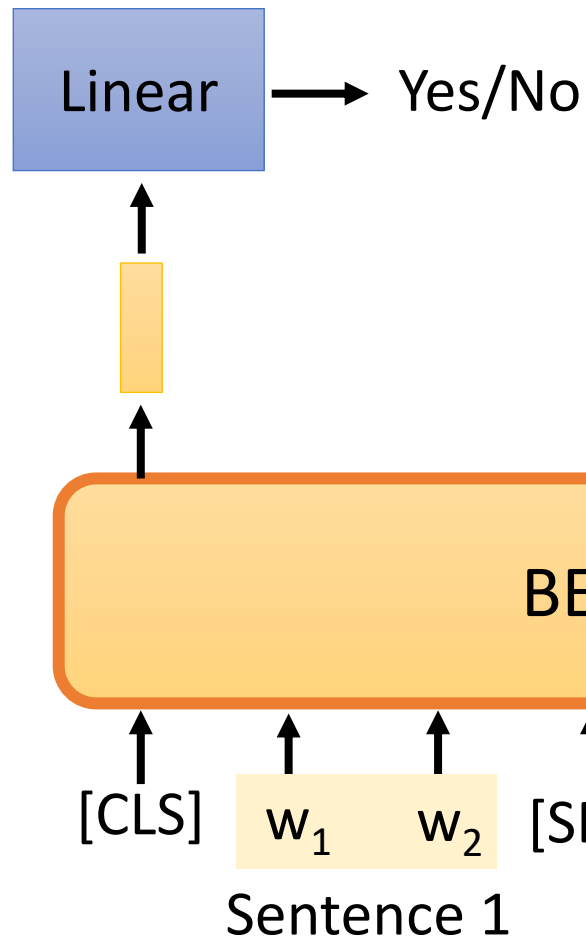
 = 
一、天、大、小 ...

Transformer
Encoder

Randomly masking
some tokens



Next Sentence Prediction



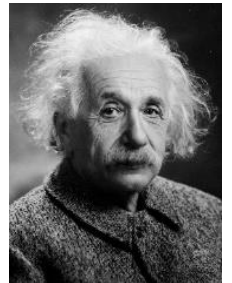
- This approach is not helpful.

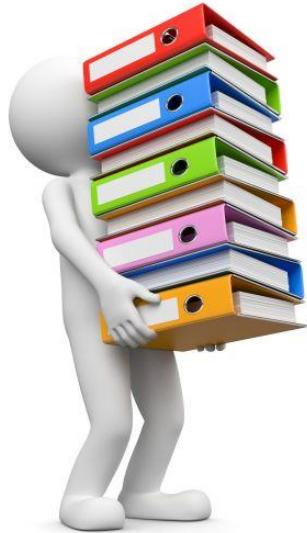
Robustly optimized BERT approach (RoBERTa) <https://arxiv.org/abs/1907.11692>

- **SOP**: Sentence order prediction

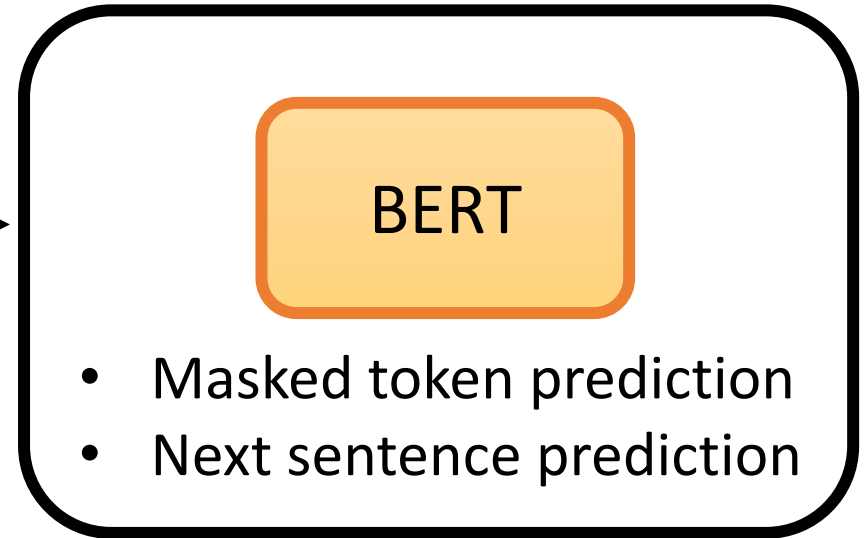
Used in ALBERT

<https://arxiv.org/abs/1909.11942>

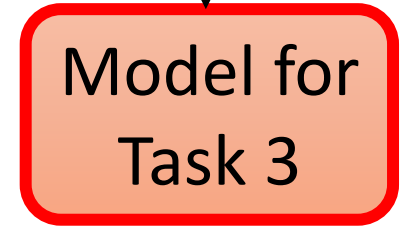
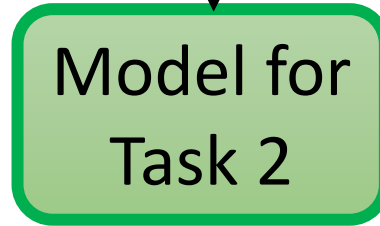
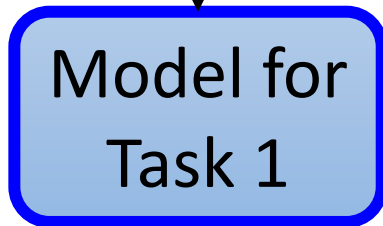




Self-supervised
Learning
Pre-train



Fine-tune



Downstream Tasks

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

GLUE

General Language Understanding Evaluation (GLUE)

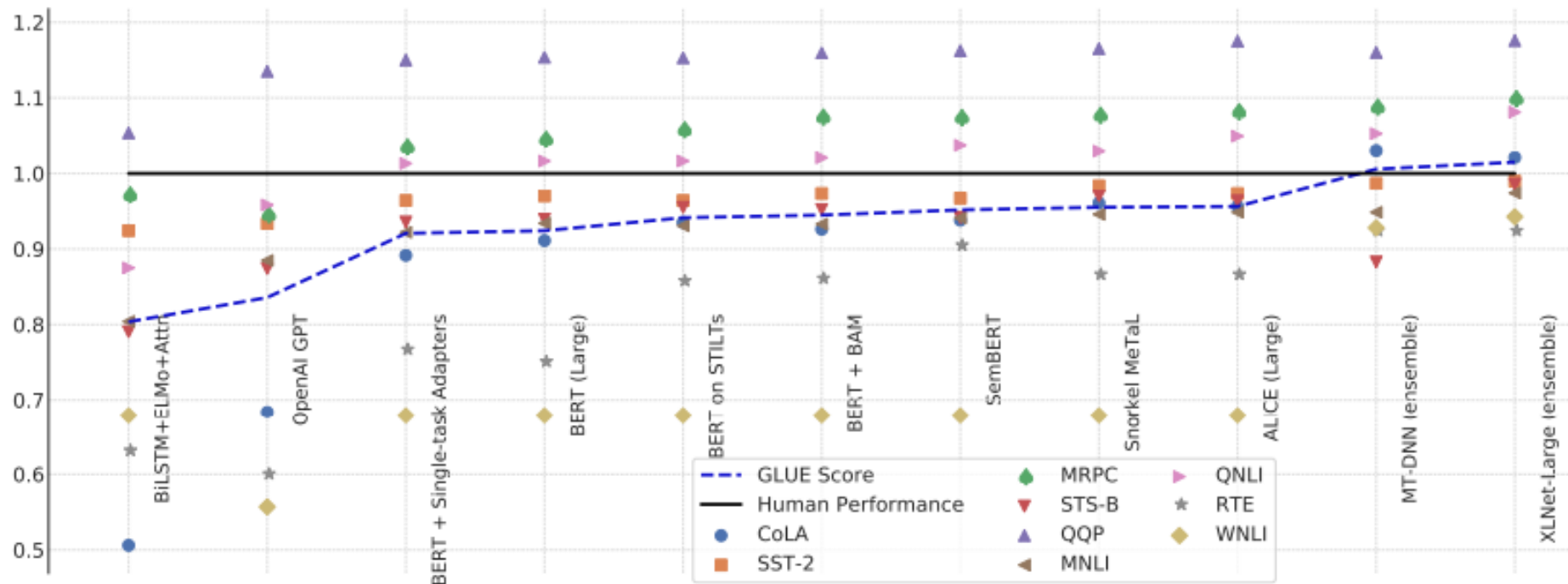
<https://gluebenchmark.com/>

- 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/>)

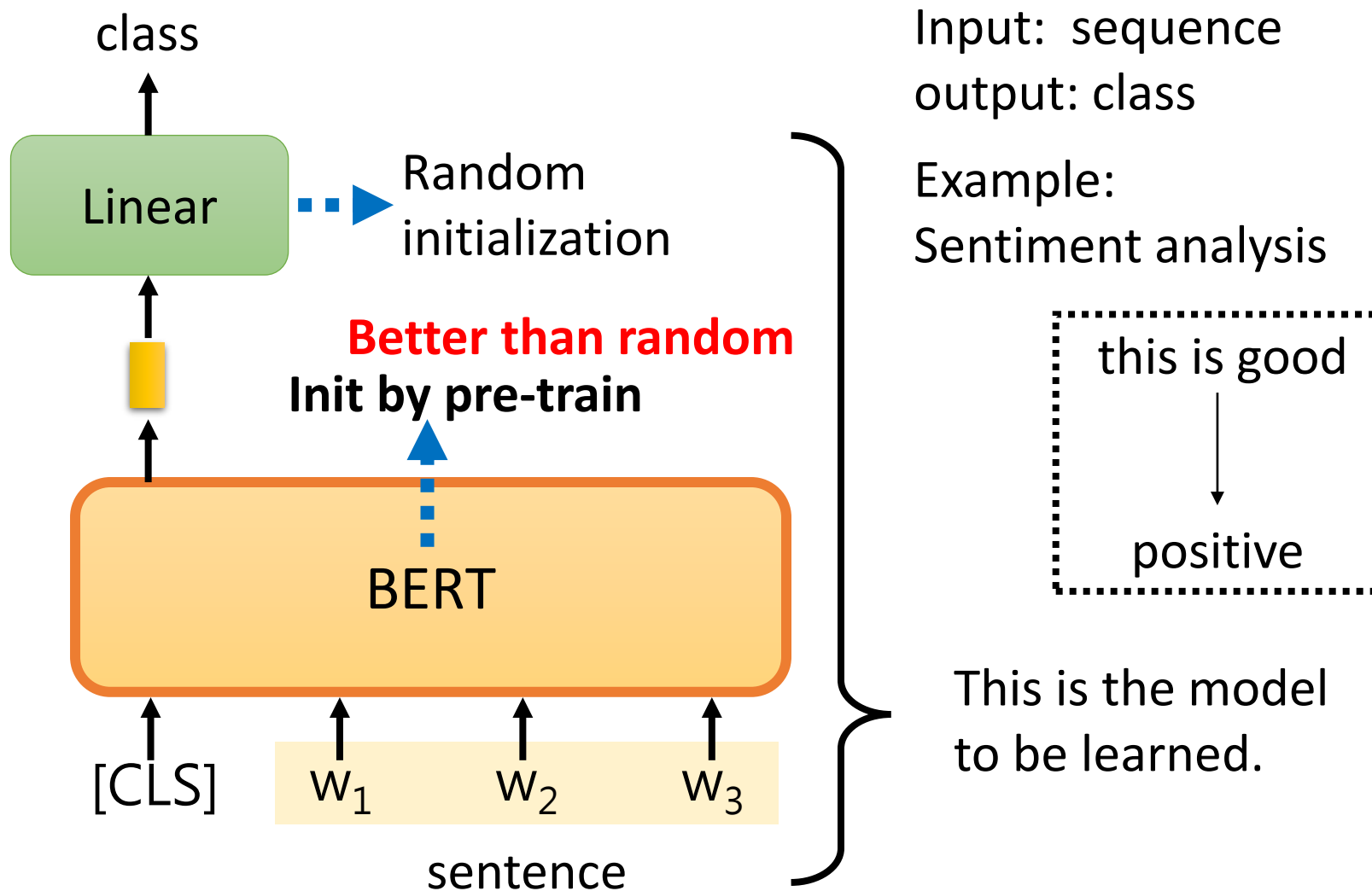
BERT and its Family

- GLUE scores



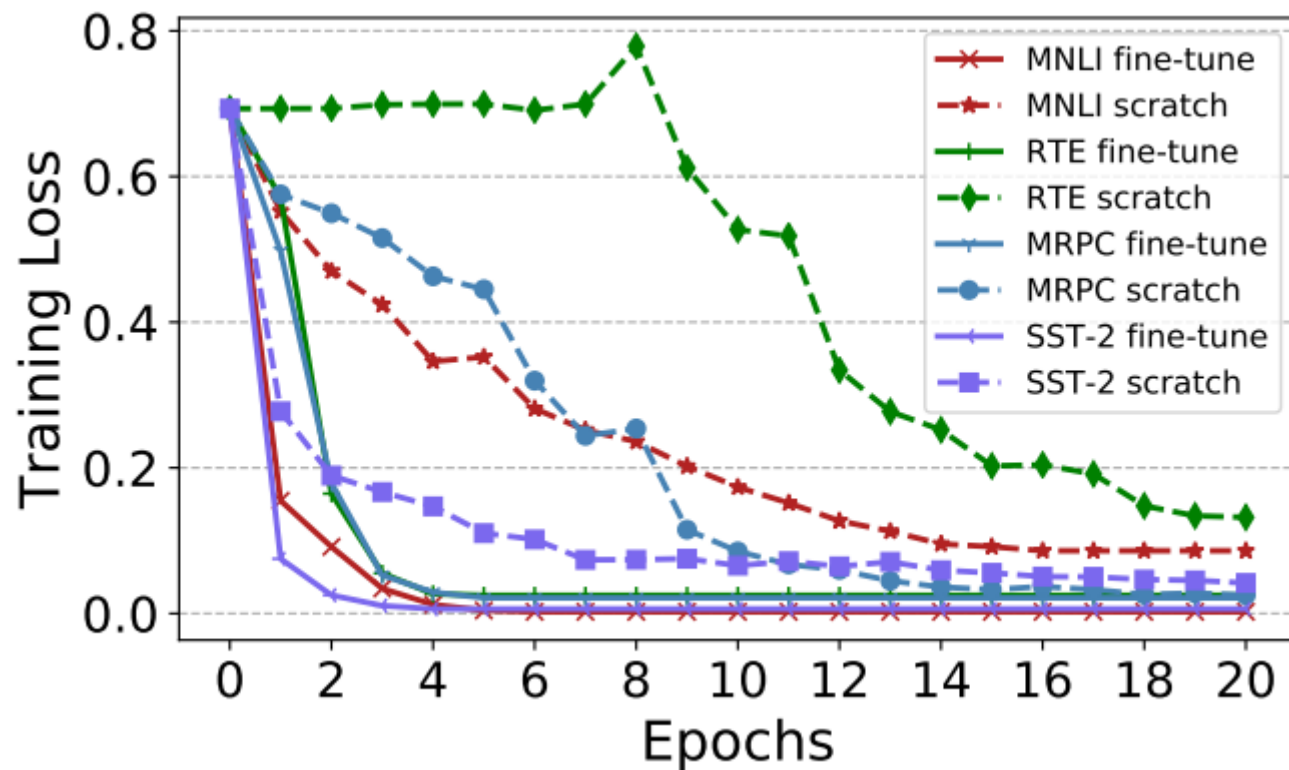
Source of image: <https://arxiv.org/abs/1905.00537>

How to use BERT – Case 1



Pre-train v.s. Random Initialization

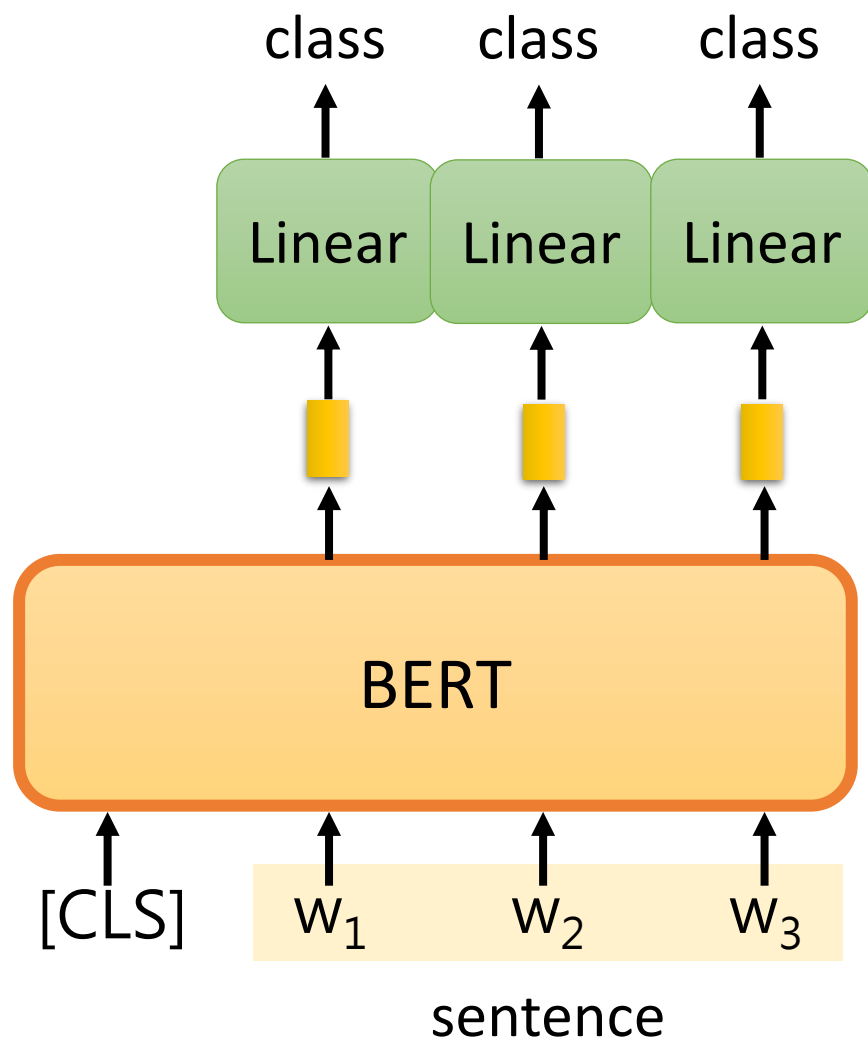
(fine-tune) (scratch)



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

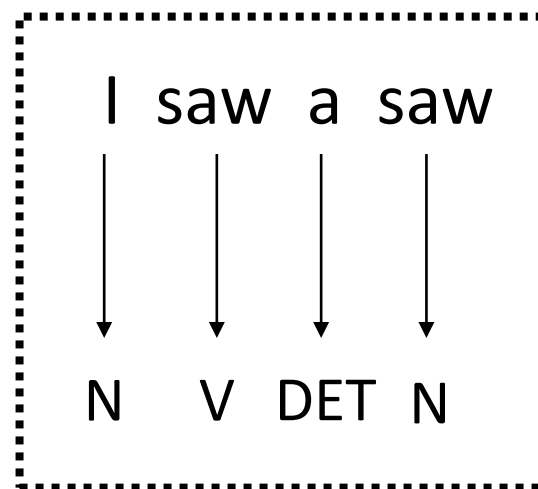


How to use BERT – Case 2



Input: sequence
output: same as input

Example:
POS tagging



How to use BERT – Case 3

Input: two sequences

Output: a class

Example:

Natural Language Inference (NLI)

contradiction

entailment

neutral

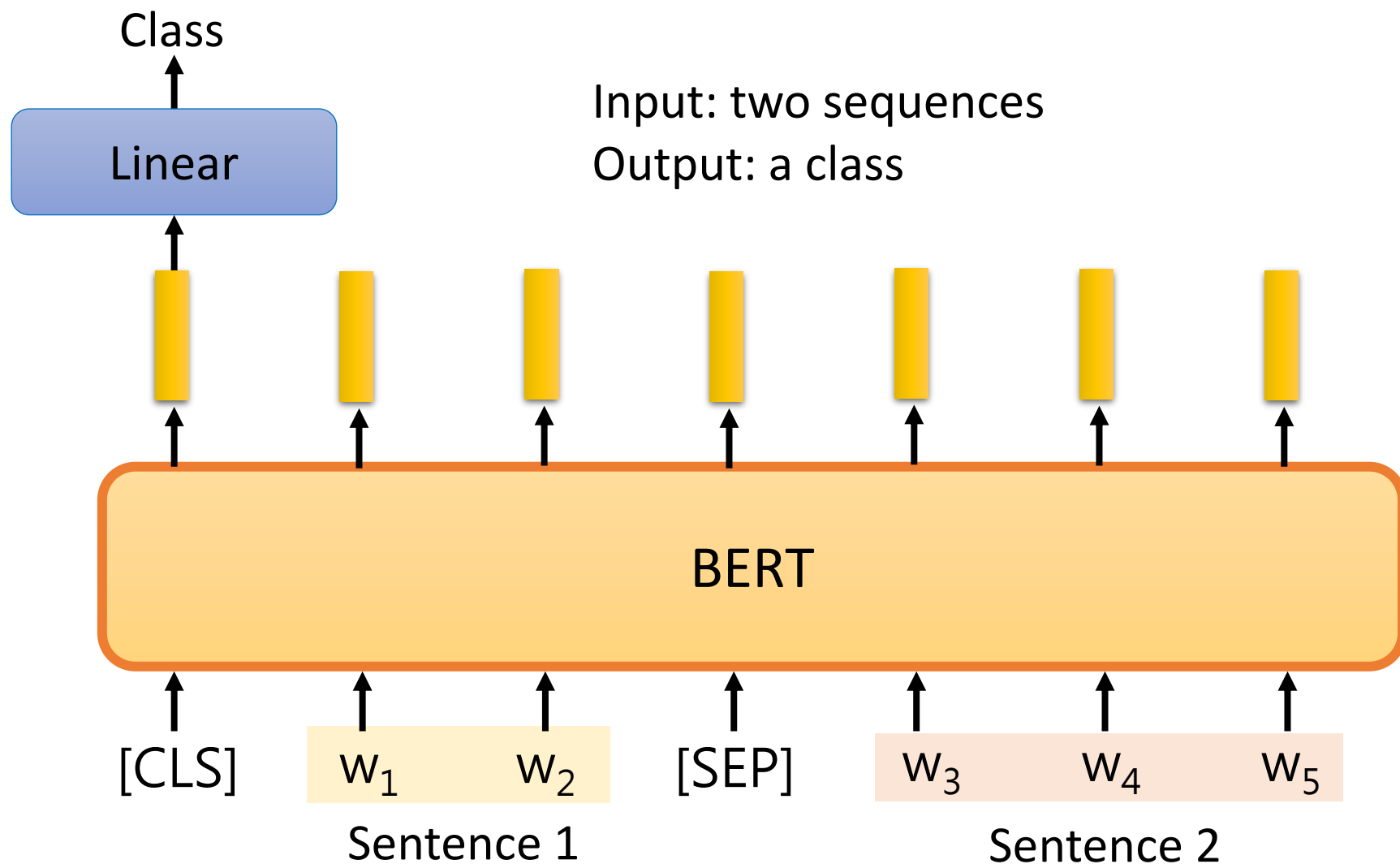
Model

premise: A person on a horse
jumps over a broken down airplane

hypothesis: A person is at a diner.

contradiction

How to use BERT – Case 3

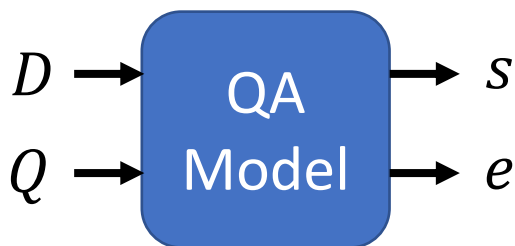


How to use BERT – Case 4

- Extraction-based Question Answering (QA)

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

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



output: two integers (s, e)

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

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, **grau-pel** 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** at **79** locations 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?

grau-pel

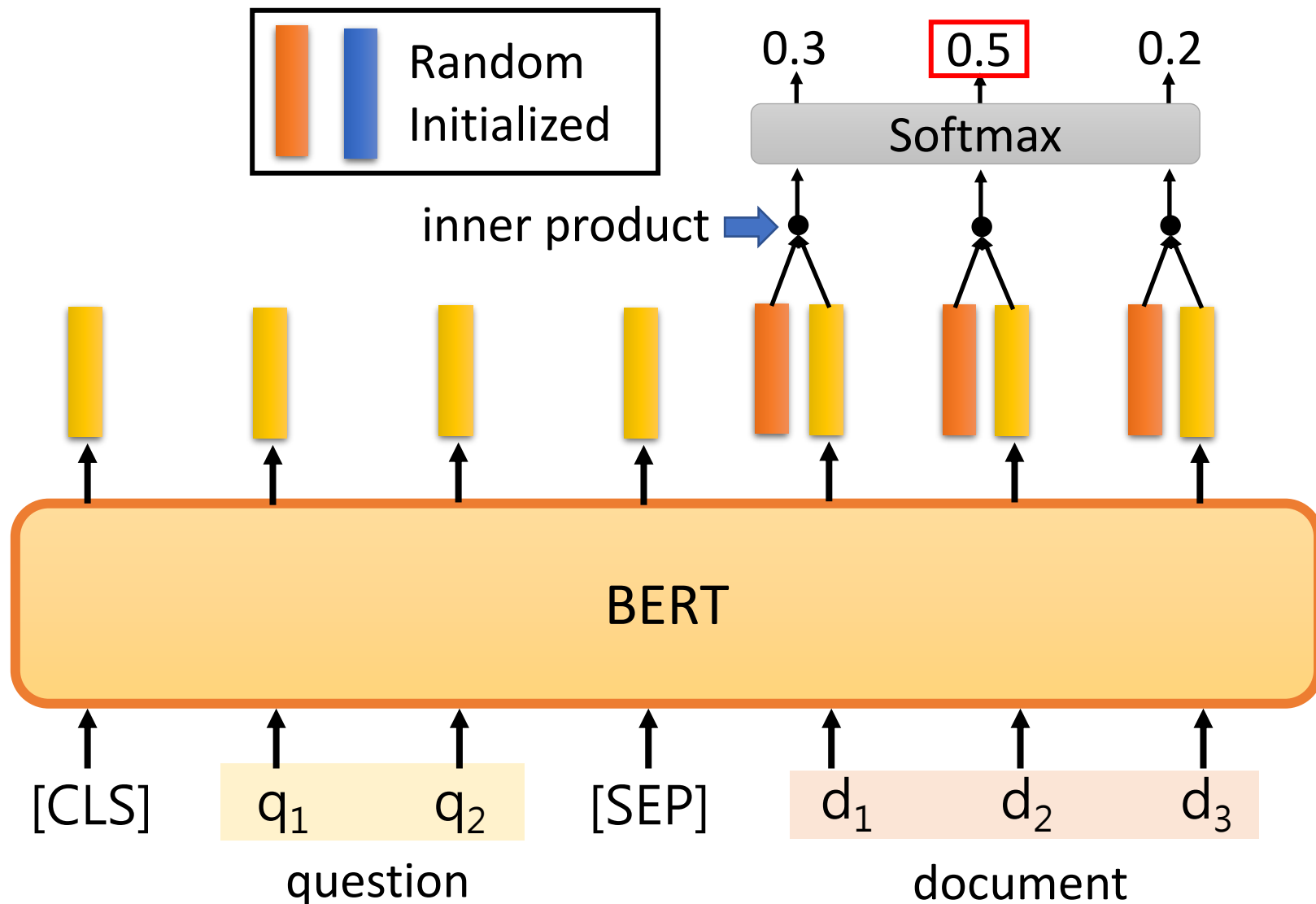
Where do water droplets collide with ice crystals to form precipitation?

within a cloud

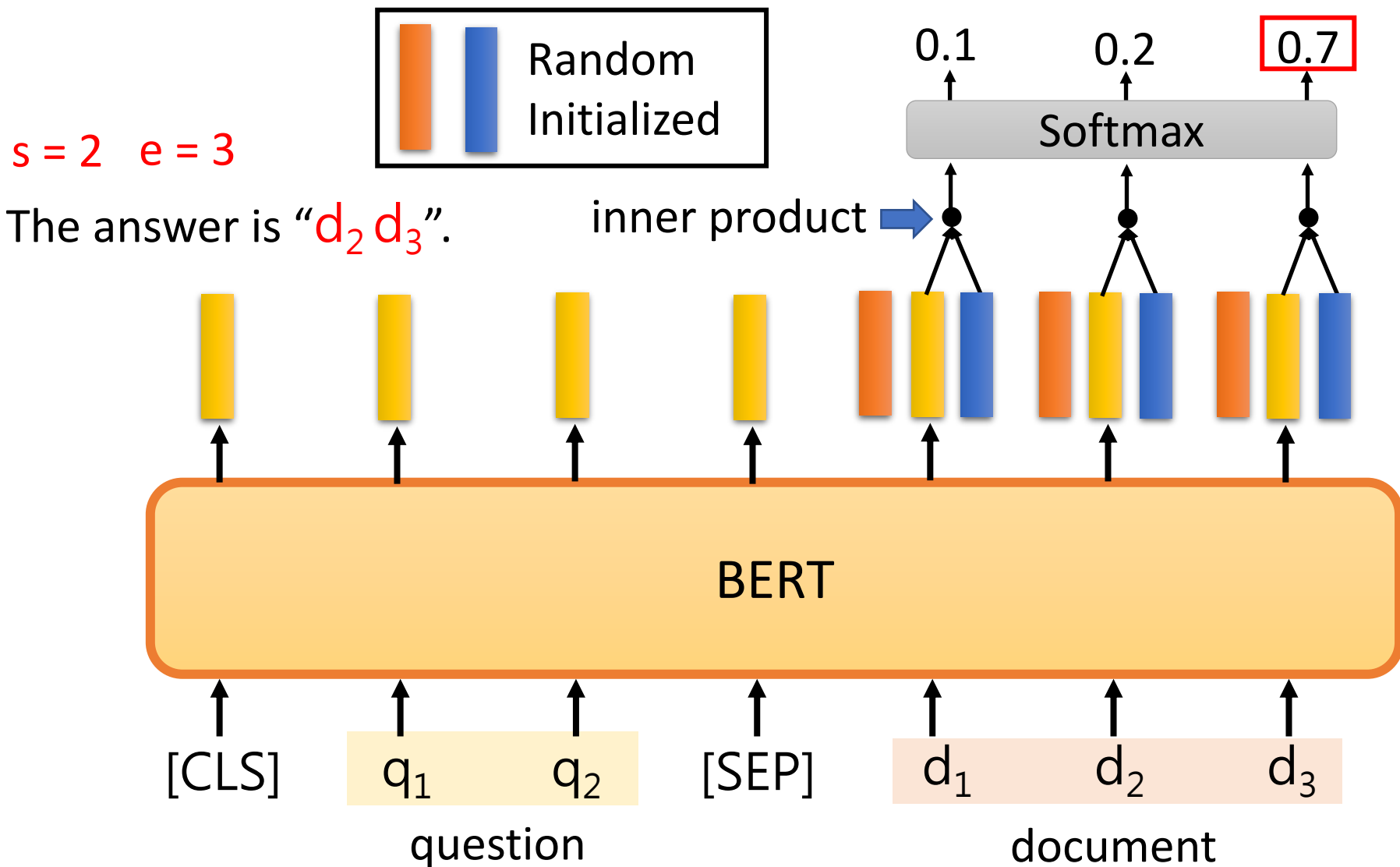
$s = 77, e = 79$

How to use BERT – Case 4

$s = 2$



How to use BERT – Case 4



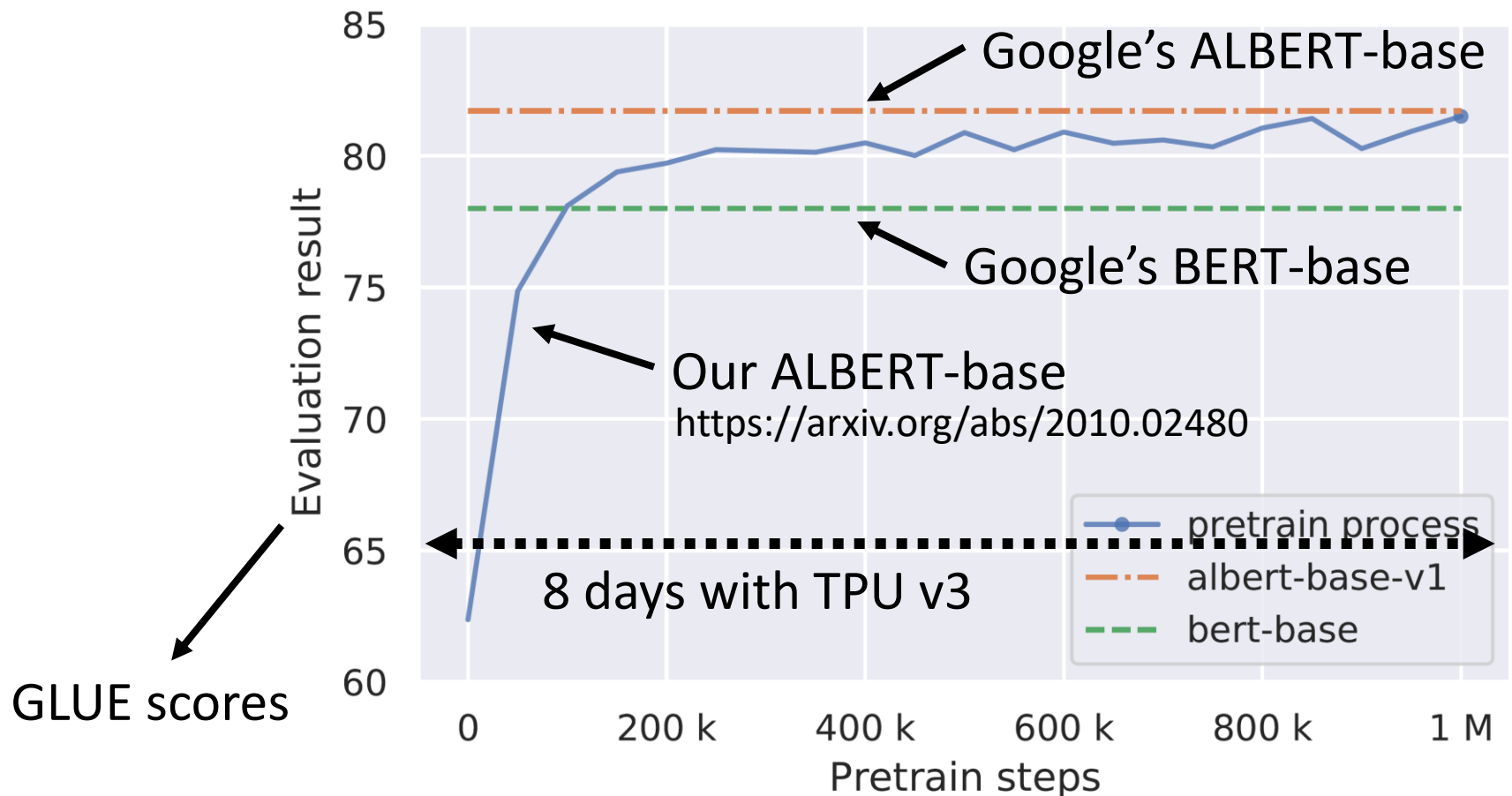
*That's
all!*



Training BERT is challenging!

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

3000 times of **Harry Potter** series



BERT Embryology (胚胎學)

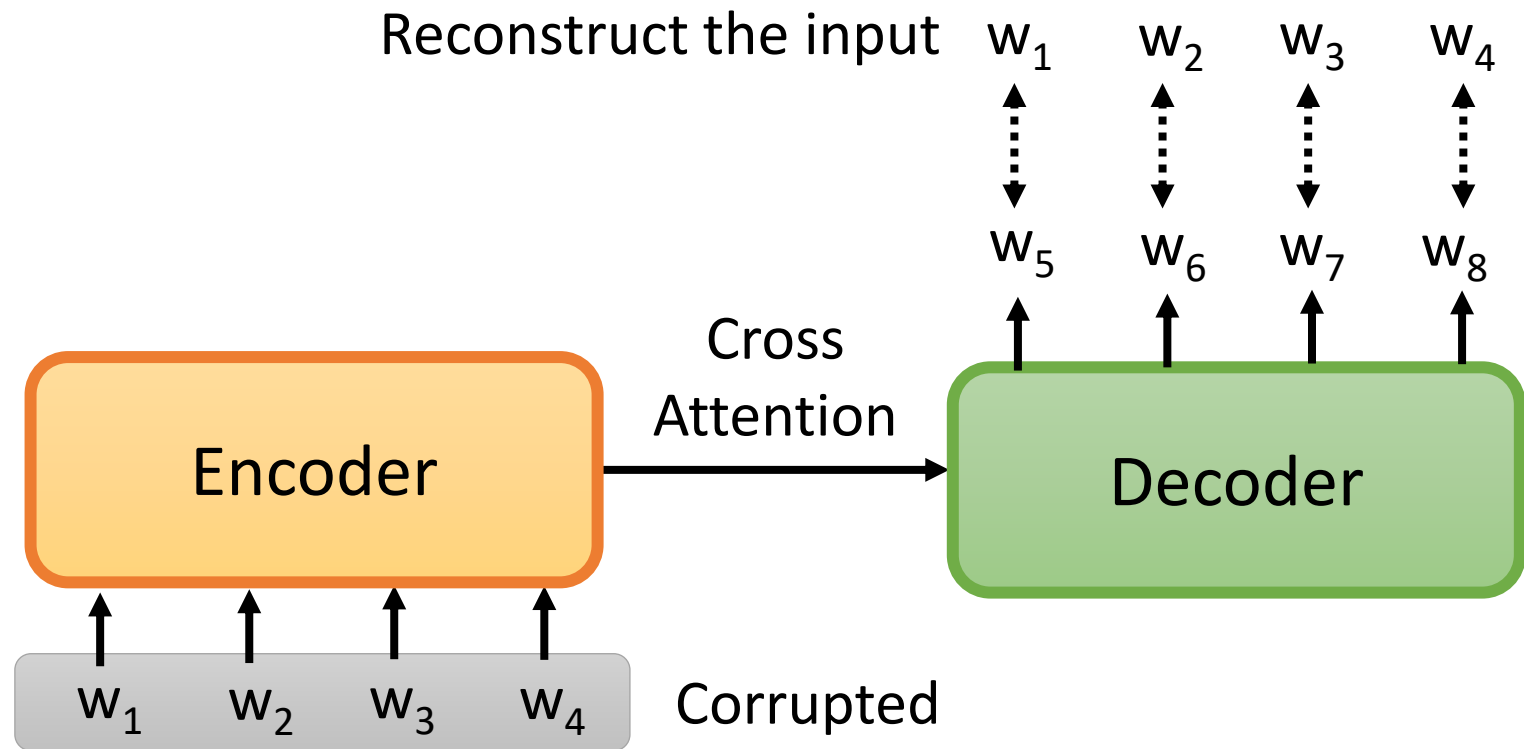
<https://arxiv.org/abs/2010.02480>



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

The answer is counterintuitive!

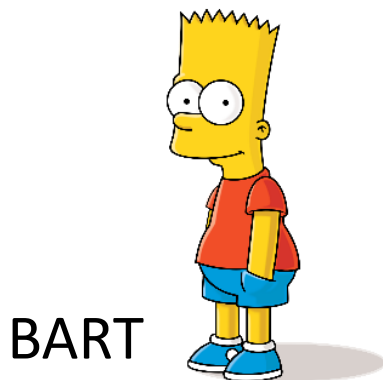
Pre-training a seq2seq model



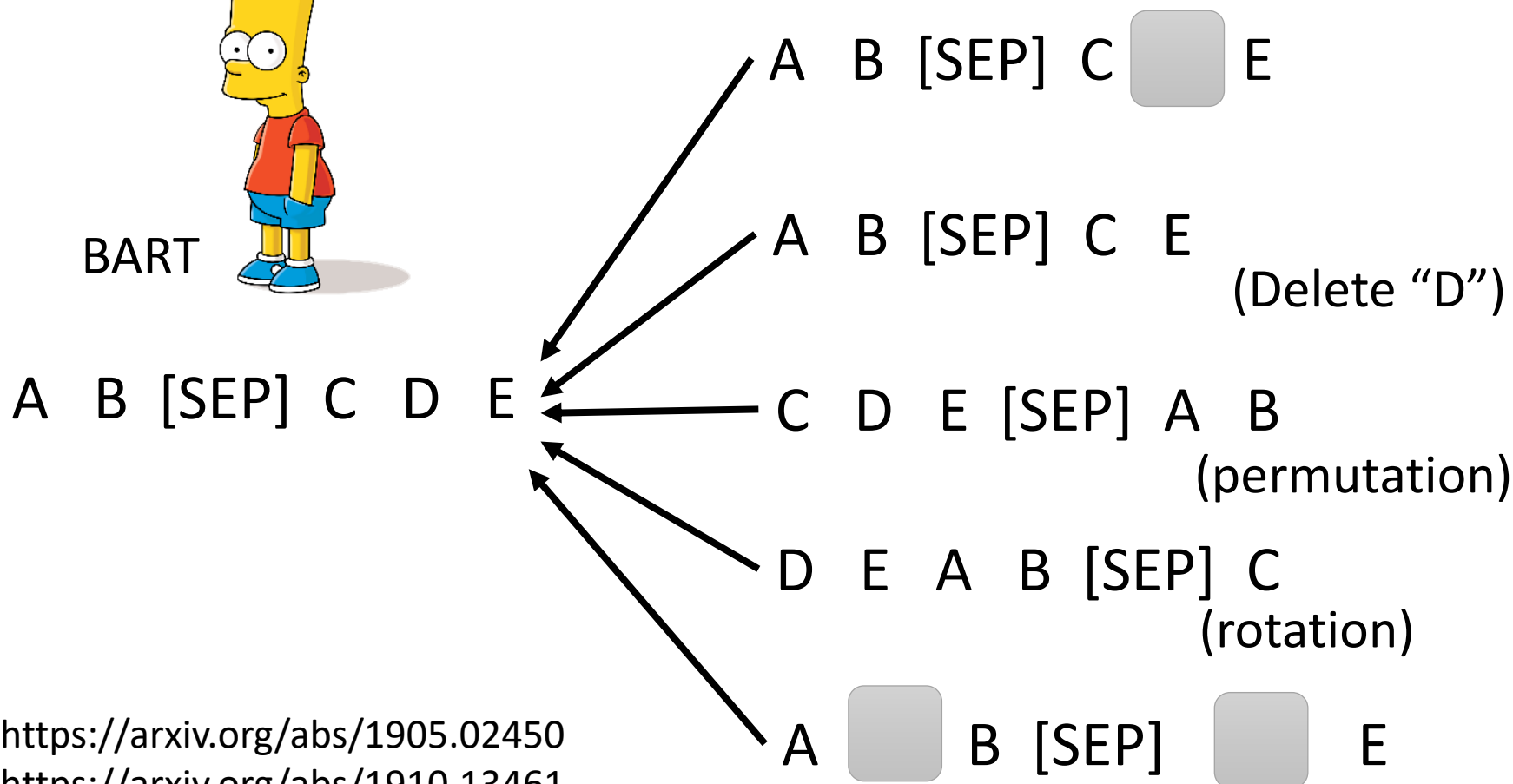
MASS / BART



MASS



BART



<https://arxiv.org/abs/1905.02450>

<https://arxiv.org/abs/1910.13461>

Text Infilling

T5 – Comparison

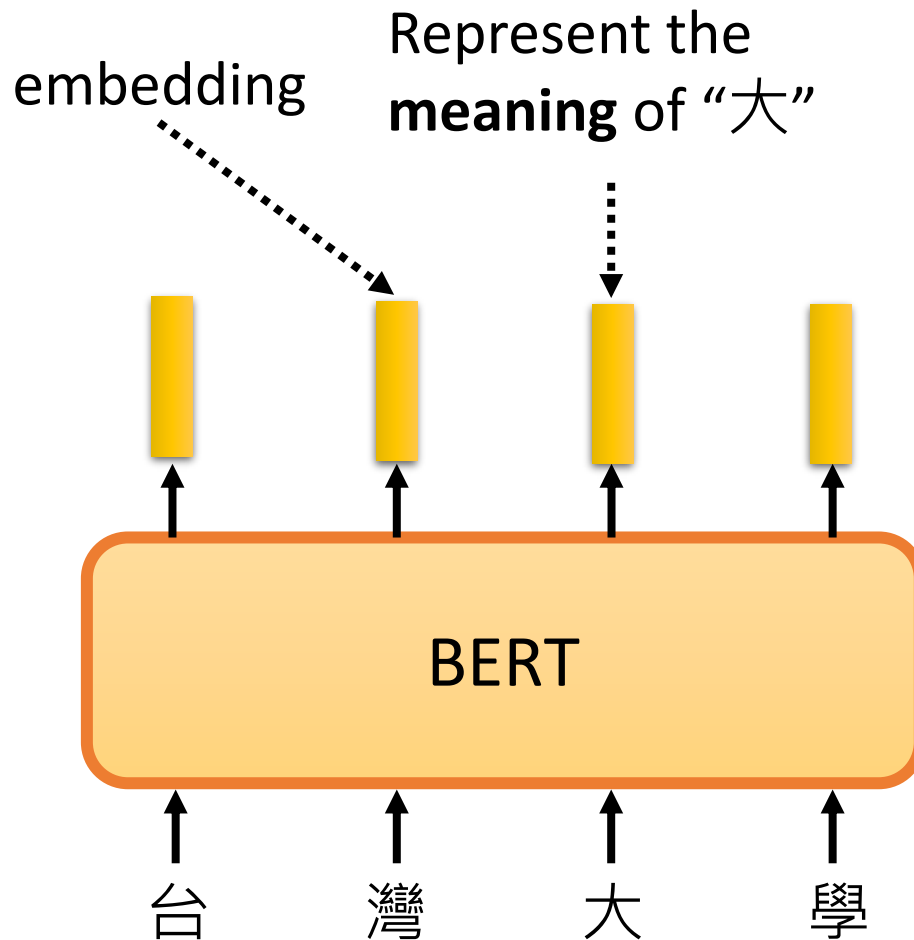
- Transfer Text-to-Text Transformer (T5)
- Colossal Clean Crawled Corpus (C4)



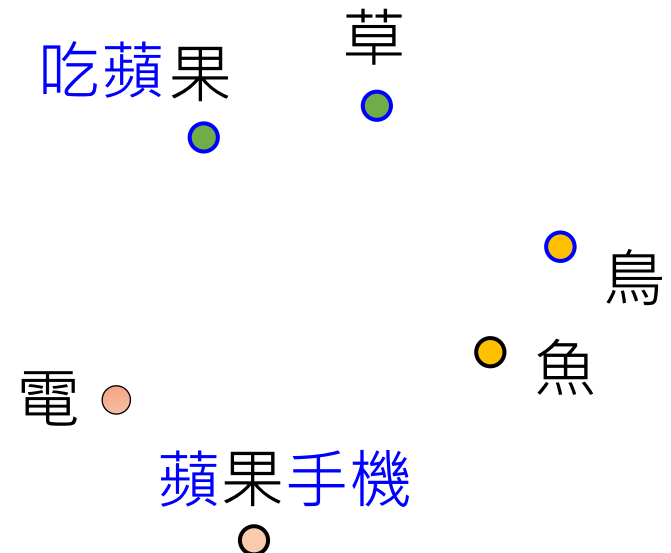
Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	(original text)
Deshuffling	party me for your to . las	
I.i.d. noise, mask tokens	Thank you <M> <M> me to	
I.i.d. noise, replace spans	Thank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

High-level approaches	Corruption strategies	Corruption rate	Corrupted span length
Language modeling	Mask	10%	2
BERT-style	Replace spans	15%	3
Deshuffling	Drop	25%	5
		50%	10

Why does BERT work?

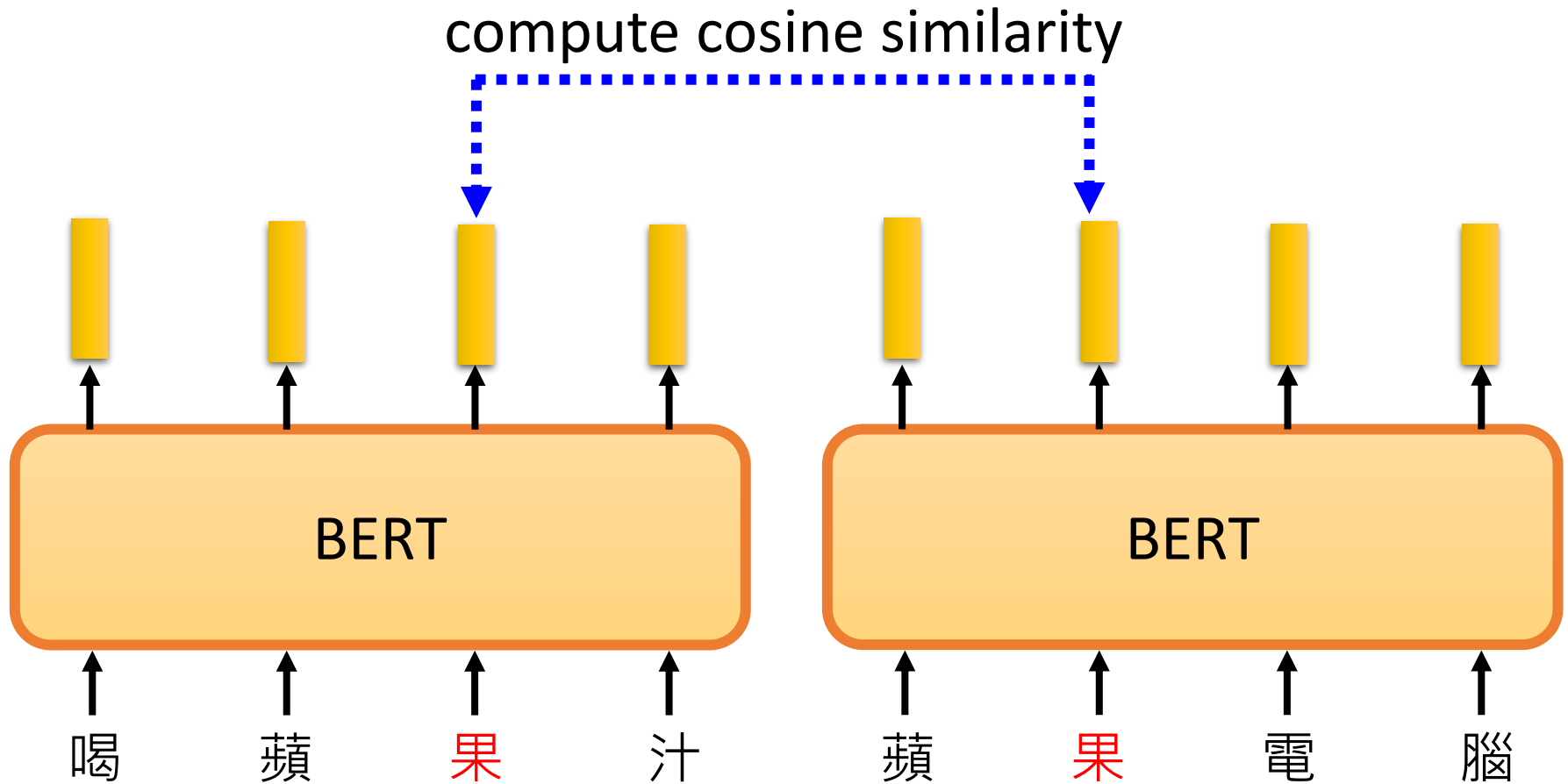


The tokens with similar meaning have similar embedding.

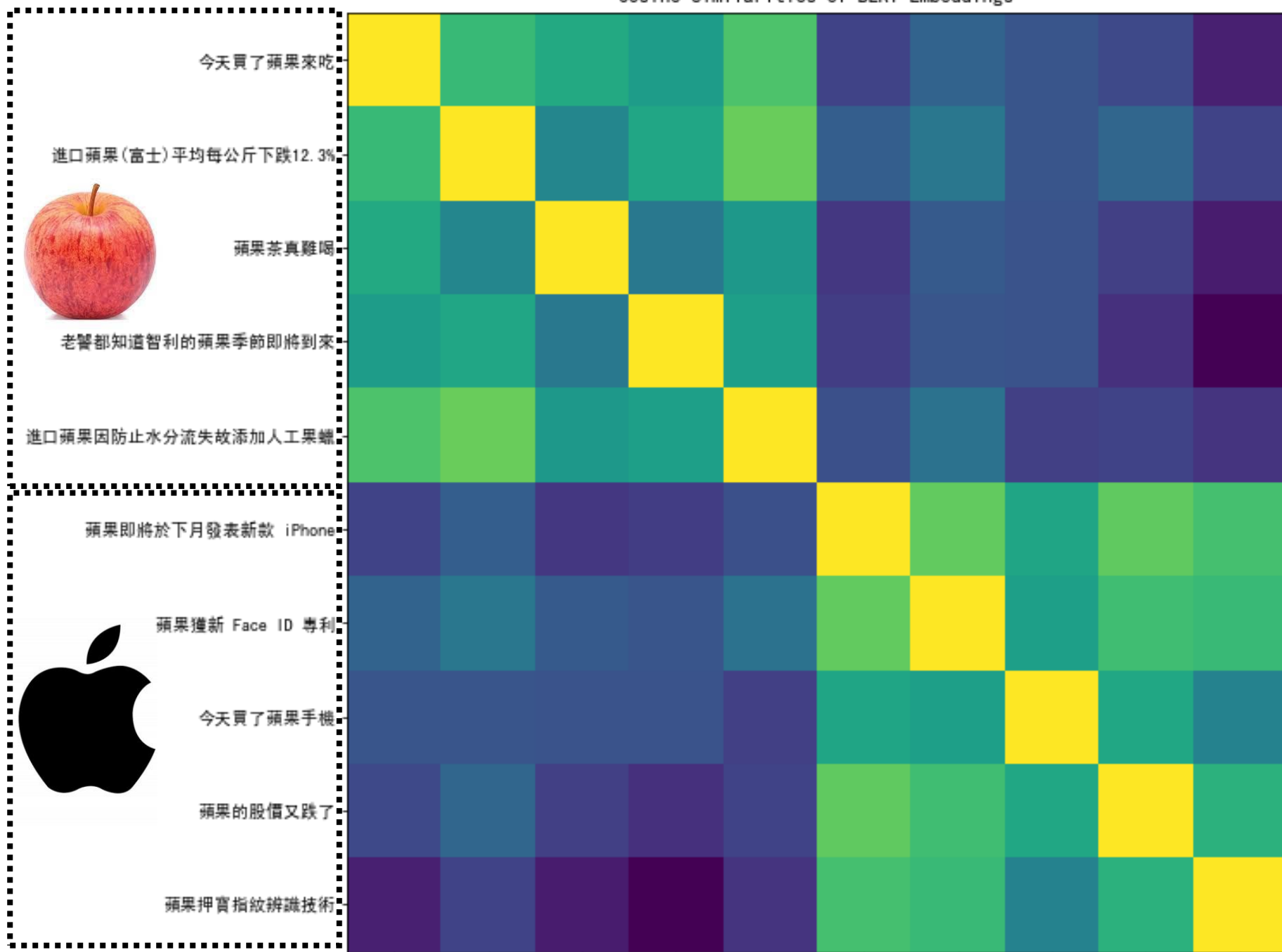


Context is considered.

Why does BERT work?



Cosine Similarities of BERT Embeddings



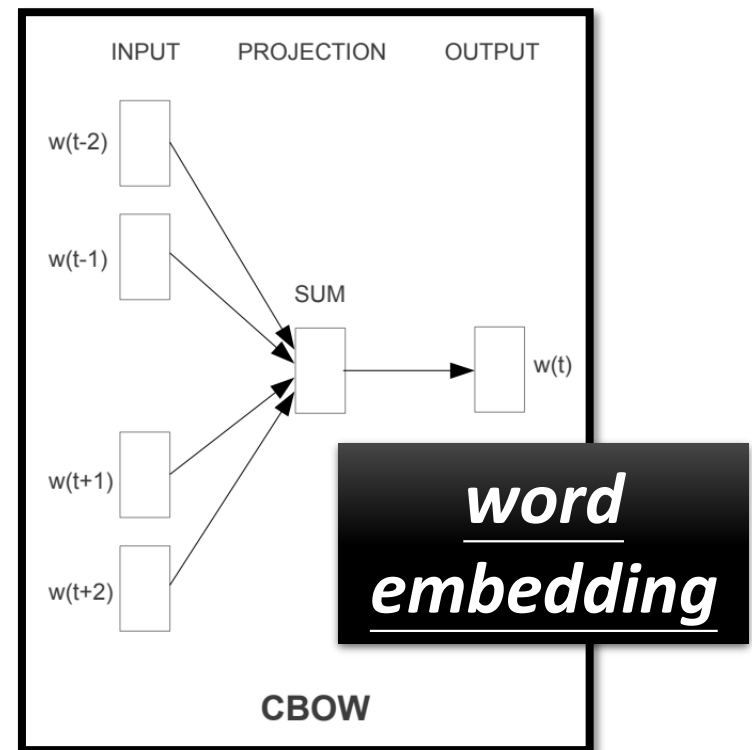
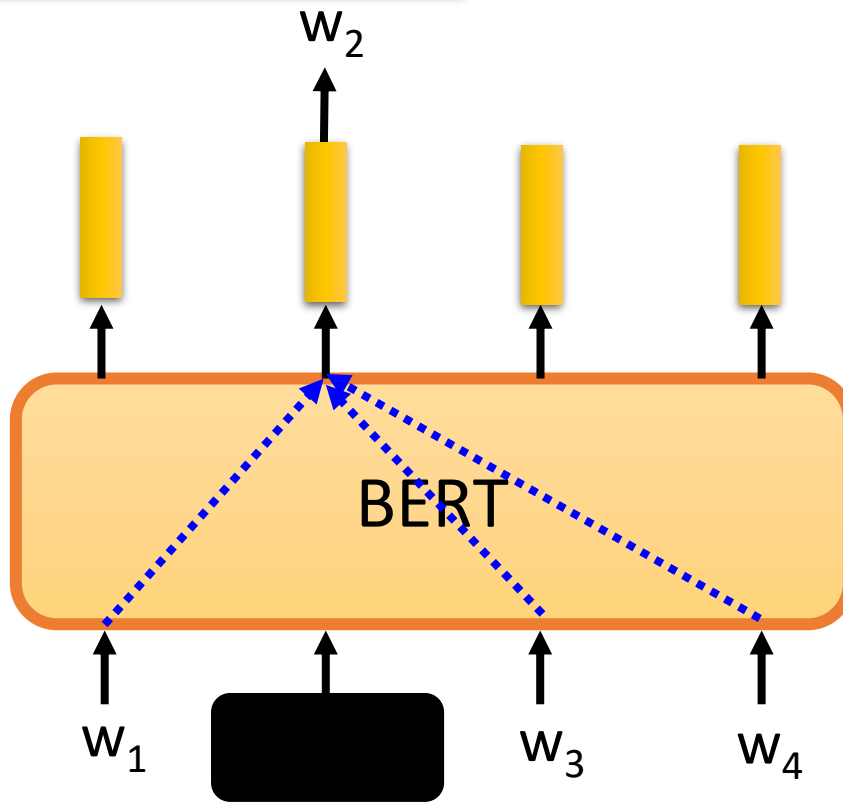
Why does BERT work?

**Contextualized
word embedding**

You shall know a word by
the company it keeps

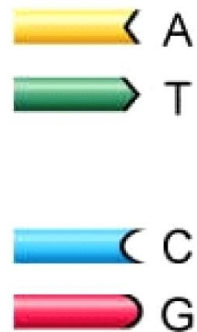
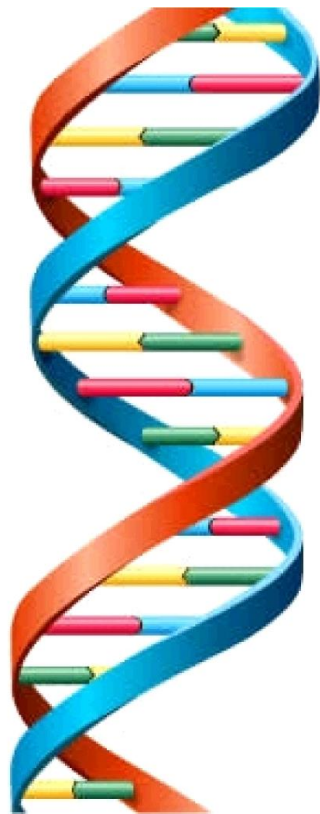


John Rupert Firth



Why does BERT work?

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



EI	CCAGCTGCATCACAGGAGGCCAGCG
EI	AGACCCGCCGGGAGGCGGAGGACCG
IE	AACGTGGCCTCCTTGTGCCCTTCCCC
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
IE	CCTGATCTGGGTCTCCCCTCCCACCC
IE	AGCCCTCAACCCTTCTGTCTCACCTT
IE	CCACTCAGCCAGGCCCTTCTTCTCCT
N	CTGTGTTCAACCACATCAAGCGCCGGC
N	GTGTTACCGAGGGGCATTTCTAACAGT
N	TCTGAGCTCTGCATTTGTCTATTCTCC

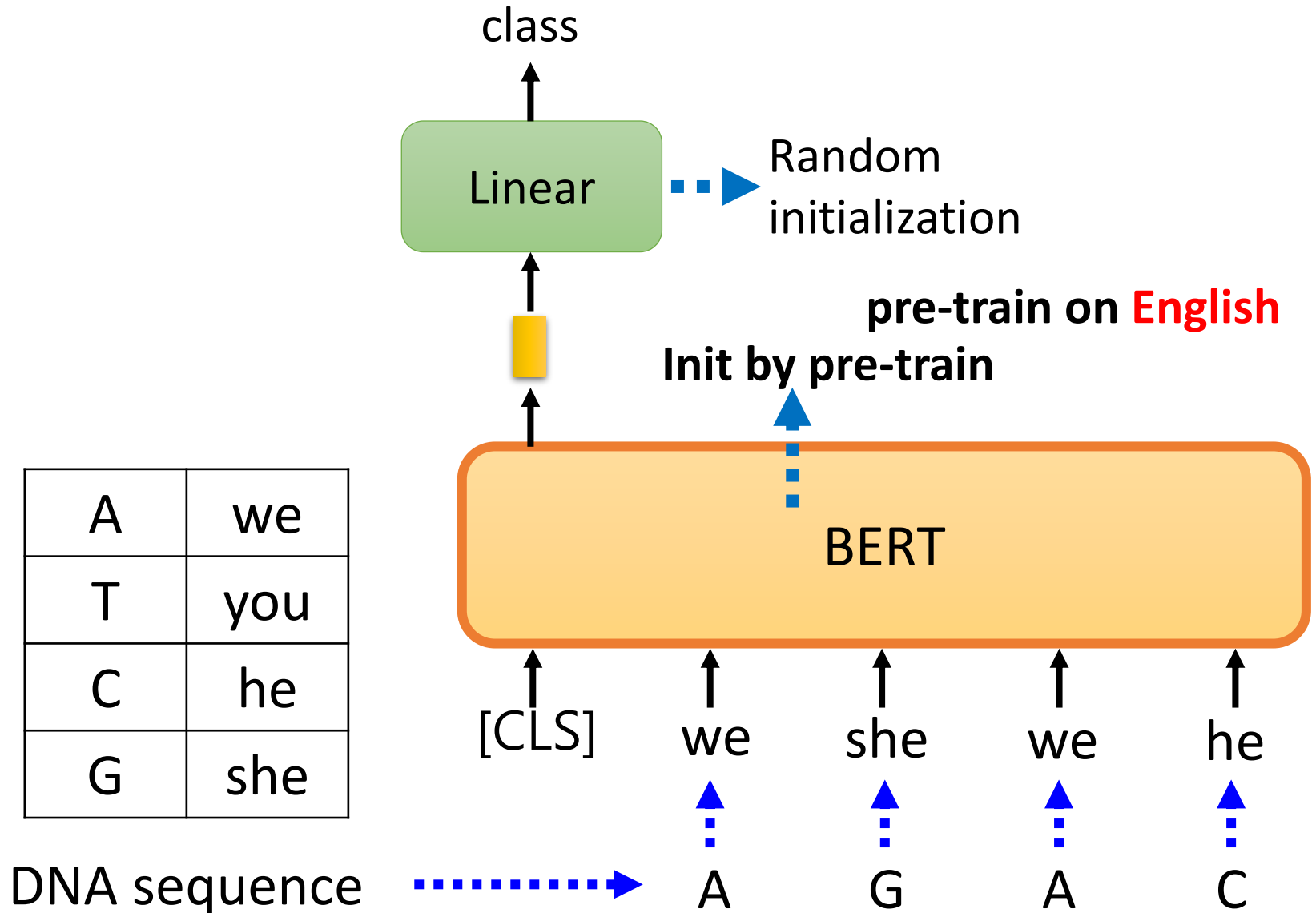
class

DNA sequence

Why does BERT work?

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

This work is done by 高瑋聰



Why does BERT work?

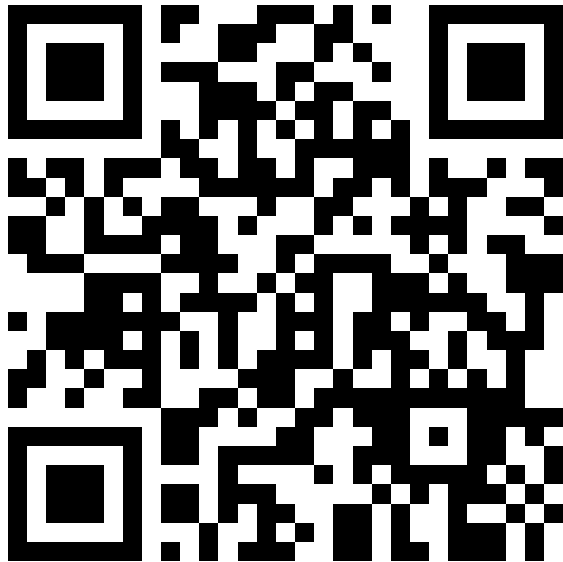
- Applying BERT to **protein, DNA, music classification**

	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



To Learn More

BERT (Part 1)



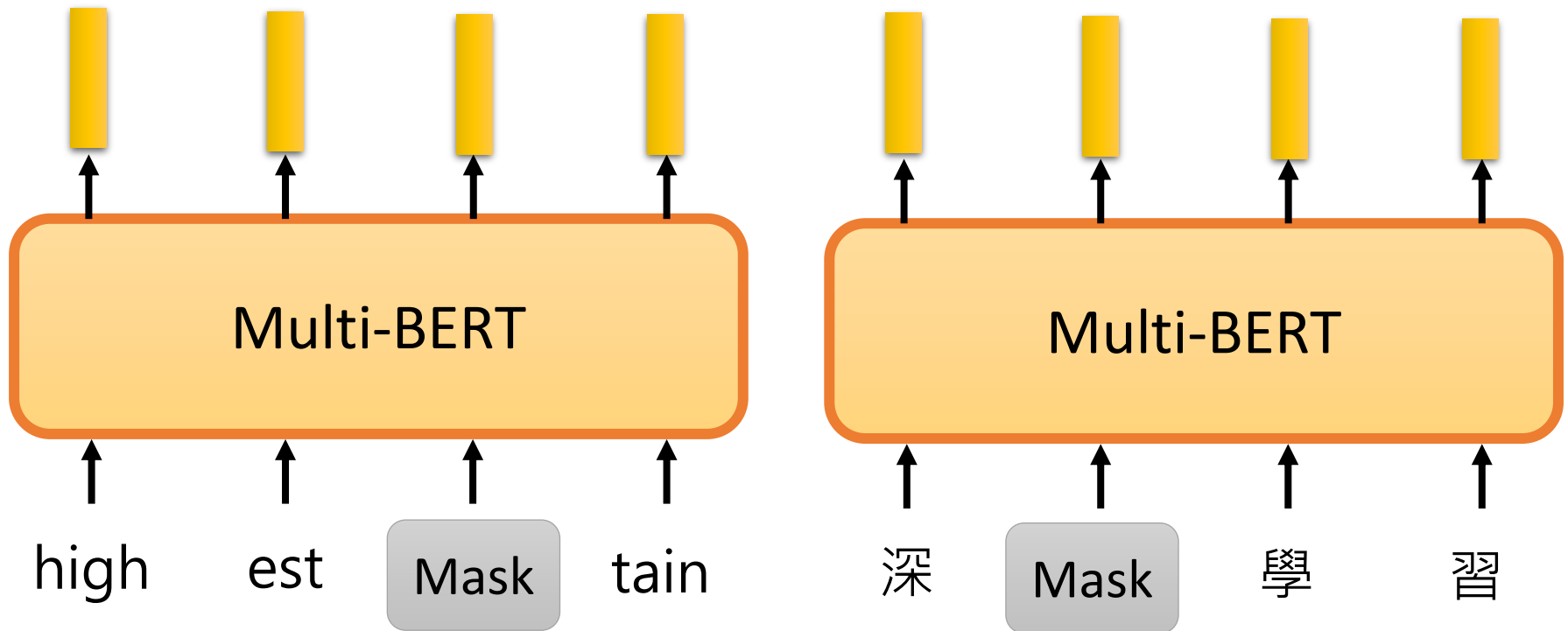
https://youtu.be/1_gRK9ElQpc

BERT (Part 2)



<https://youtu.be/Bywo7m6ySlk>

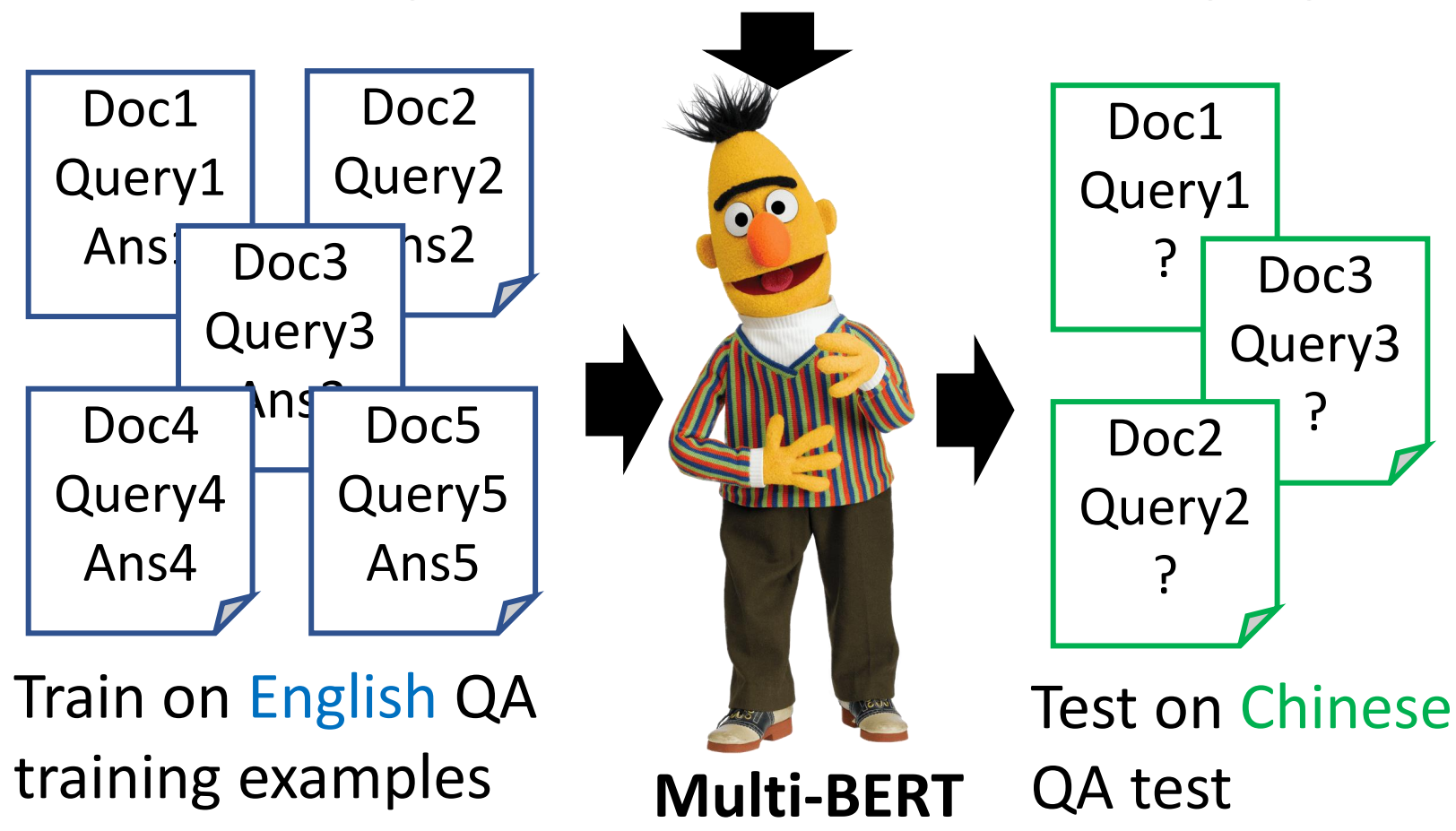
Multi-lingual BERT



Training a BERT model by many different languages.

Zero-shot Reading Comprehension

Training on the sentences of 104 languages



Zero-shot Reading Comprehension

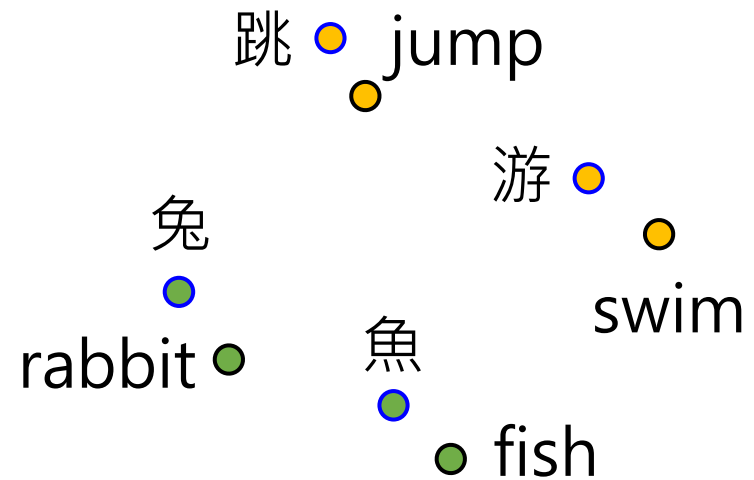
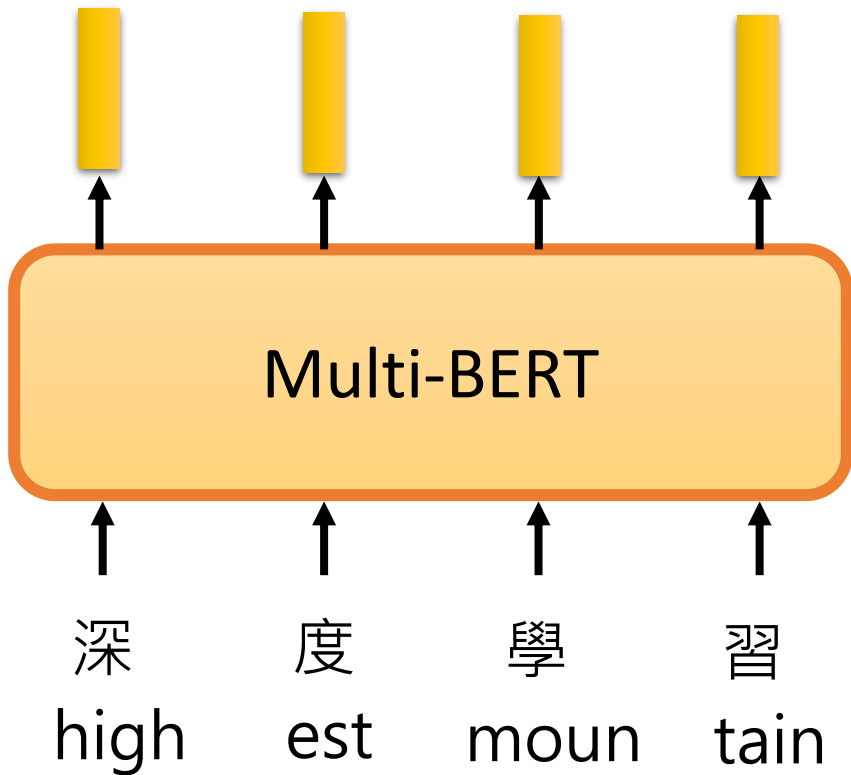
- English: SQuAD, Chinese: DRCD

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%

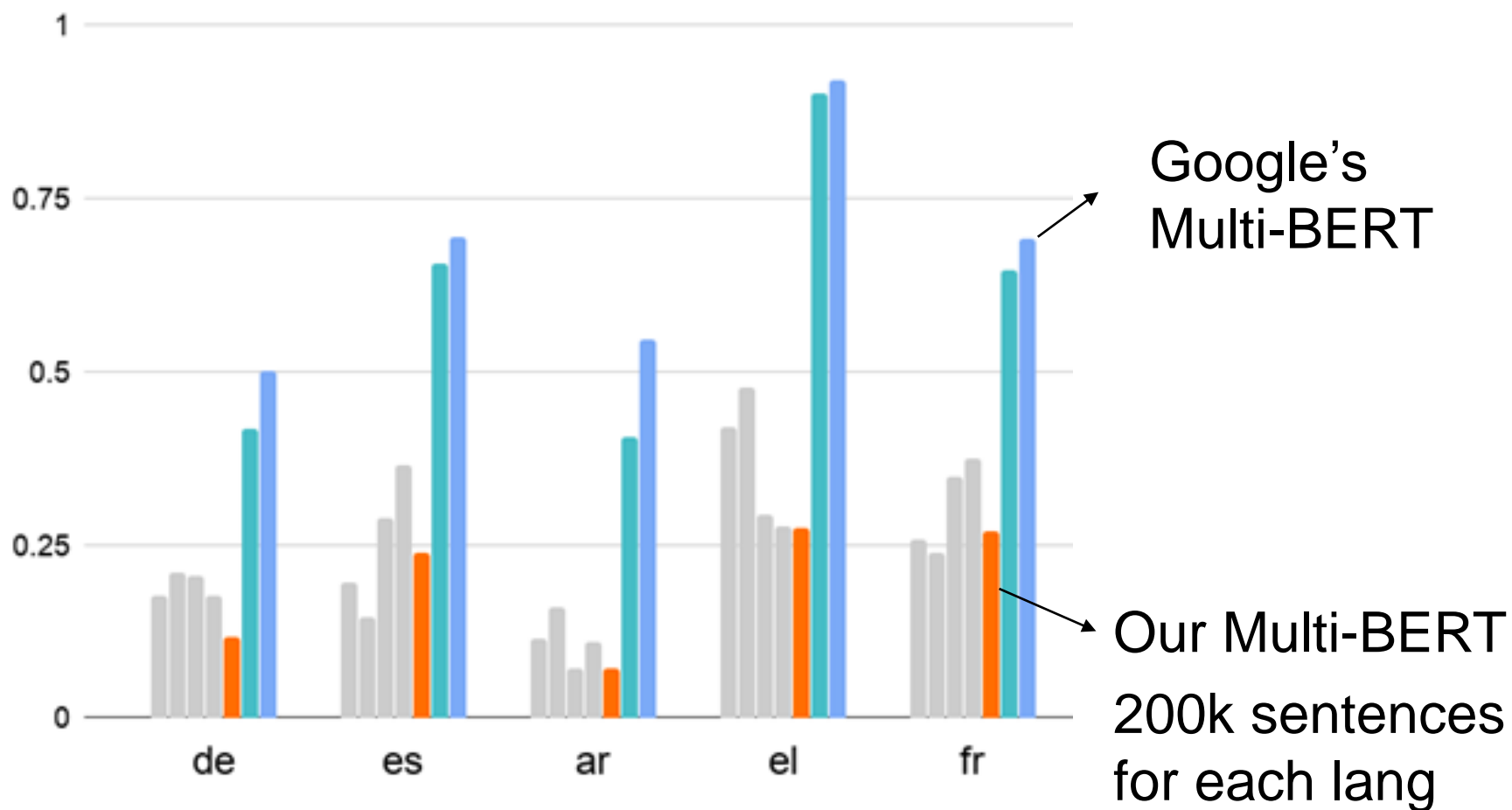
This work is done by 劉記良、許宗嫻
<https://arxiv.org/abs/1909.09587>

Cross-lingual Alignment?



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?

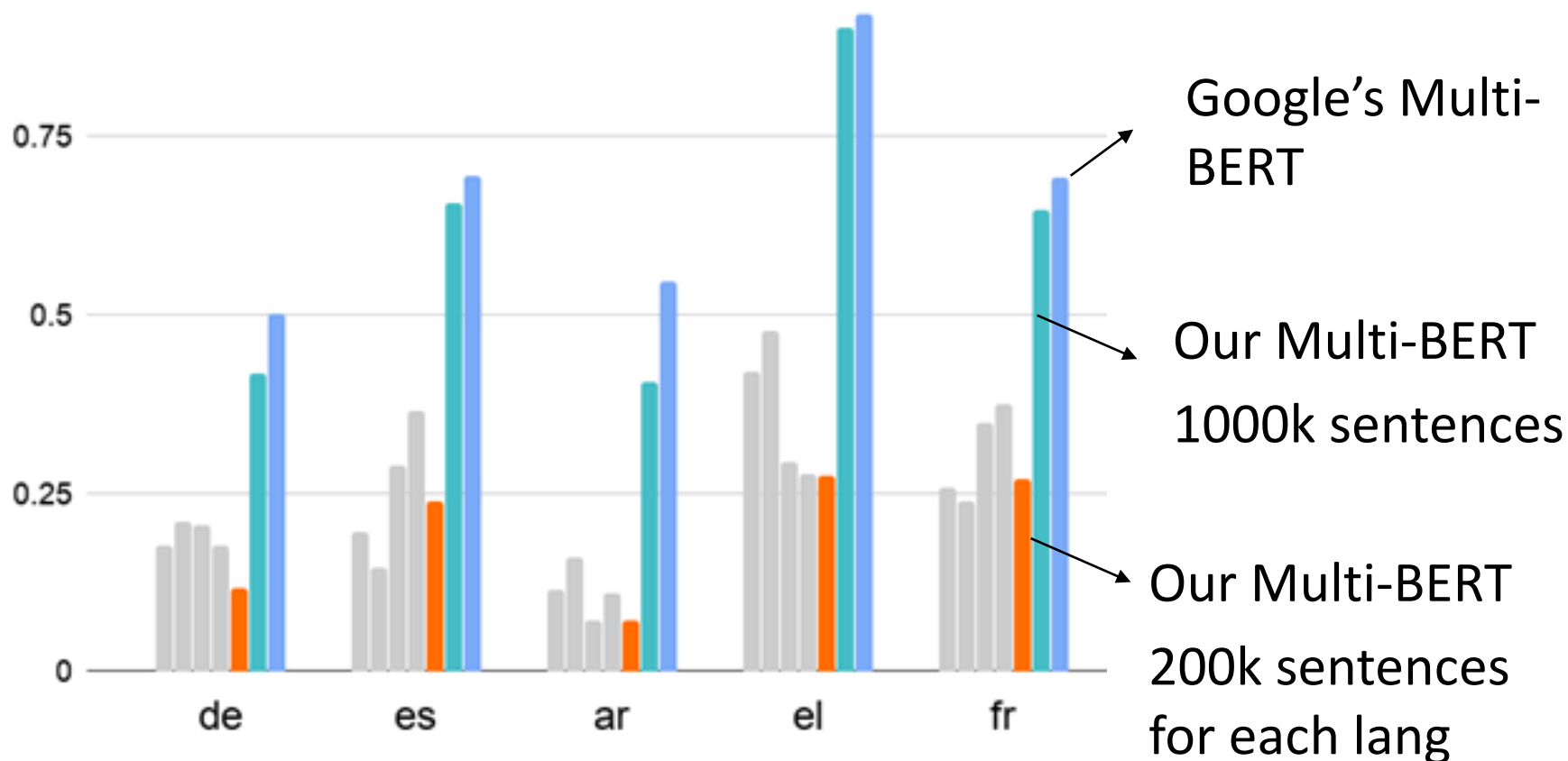
The training is also challenging ...



Mean Reciprocal Rank (MRR):

Higher MRR, better alignment

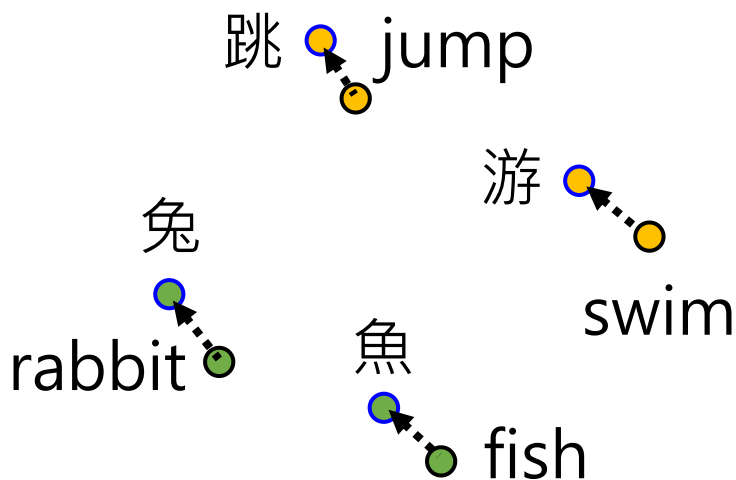
1 — The amount of training data is critical for alignment.



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

投影片來源: 許宗嫻同學碩士口試投影片

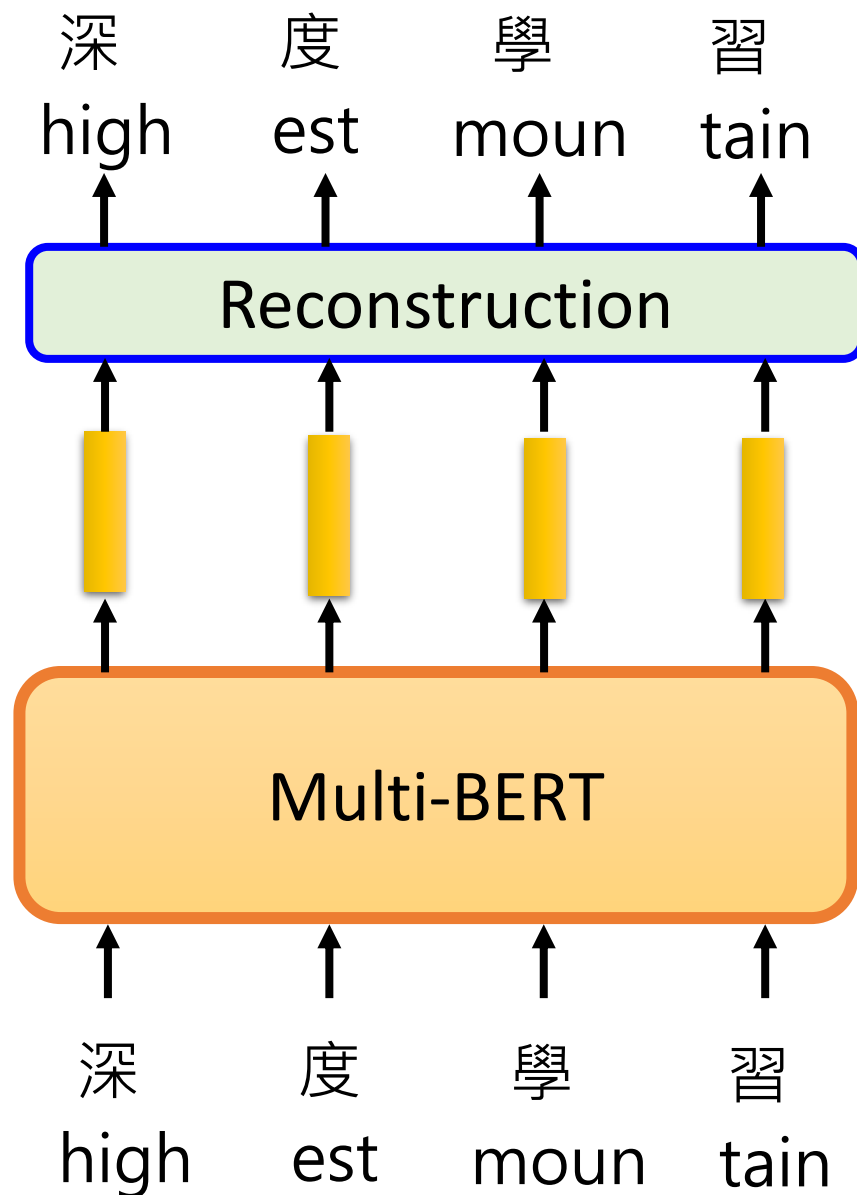
Weird???



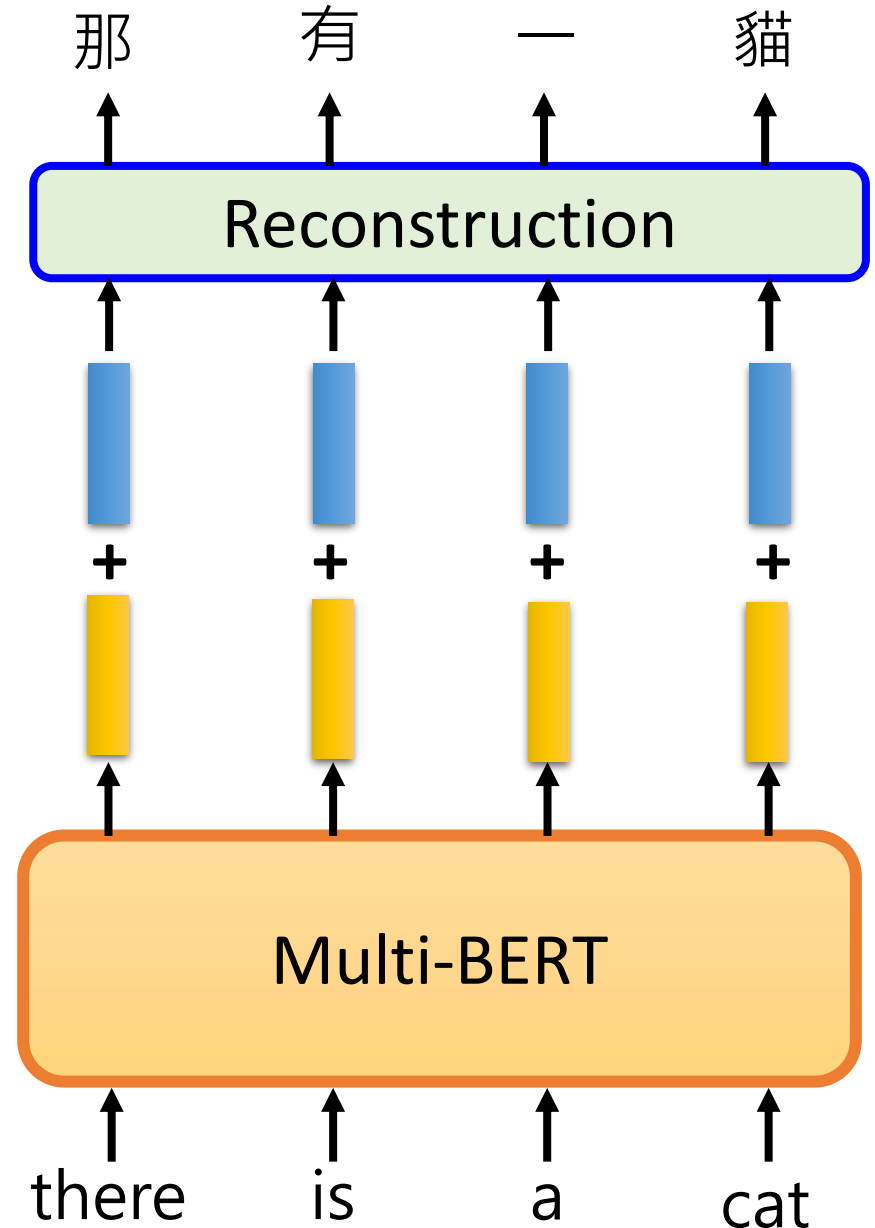
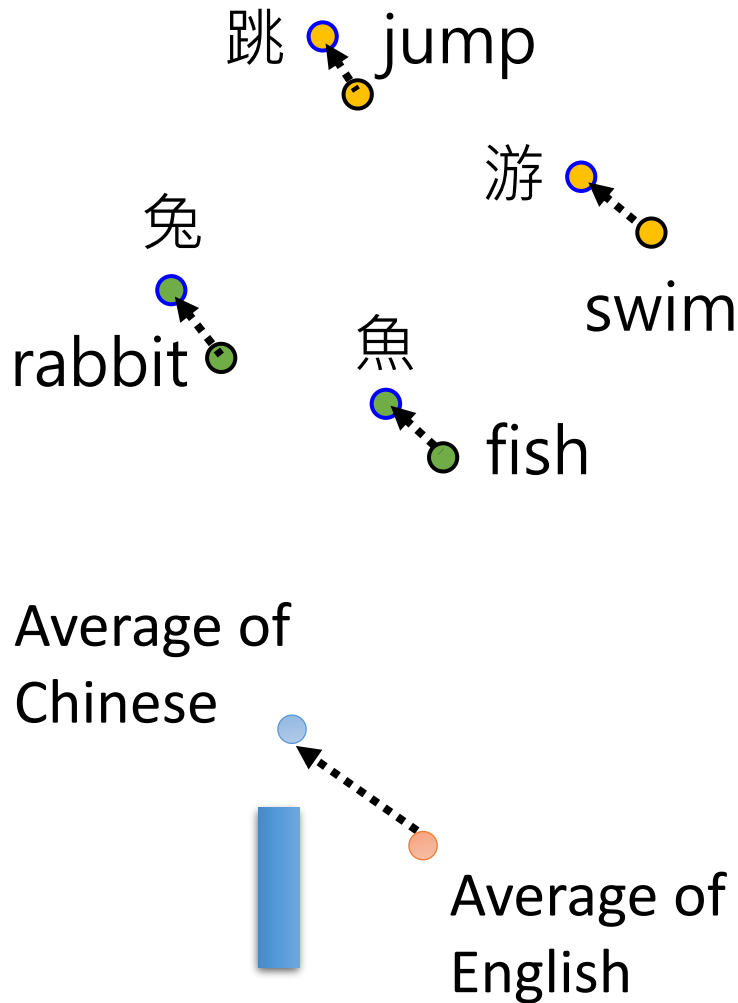
If the embedding is
language independent ...

How to correctly
reconstruct?

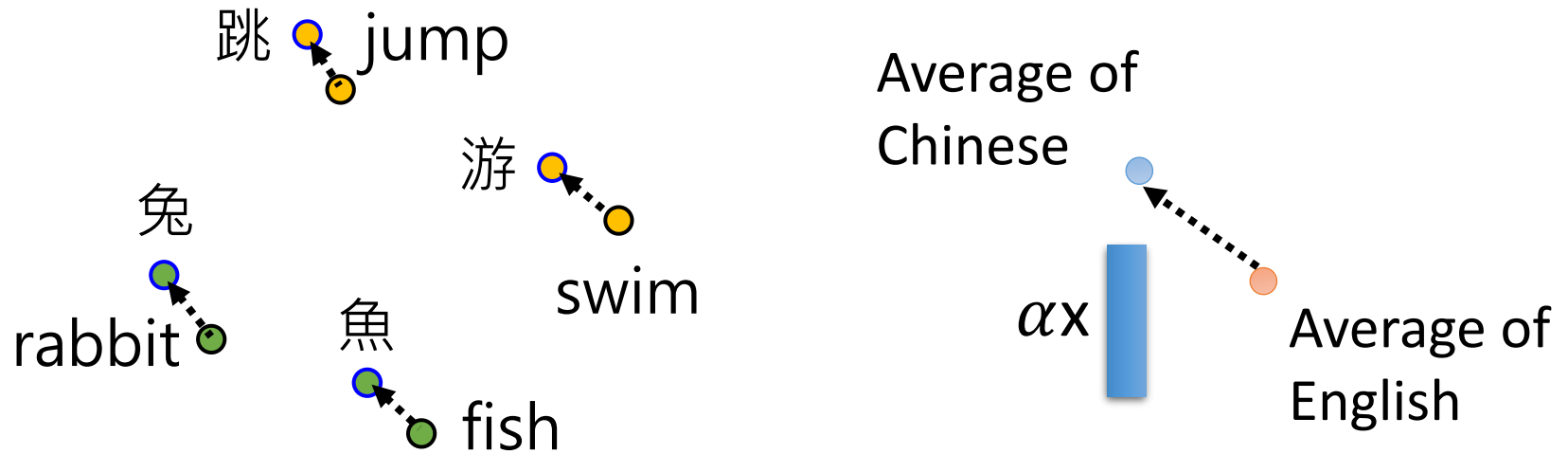
There must be language
information.



Where is Language?



If this is true ...



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)	能帮助我的女孩在小镇的另一边。没有人能帮助我。。
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 😊

Outline

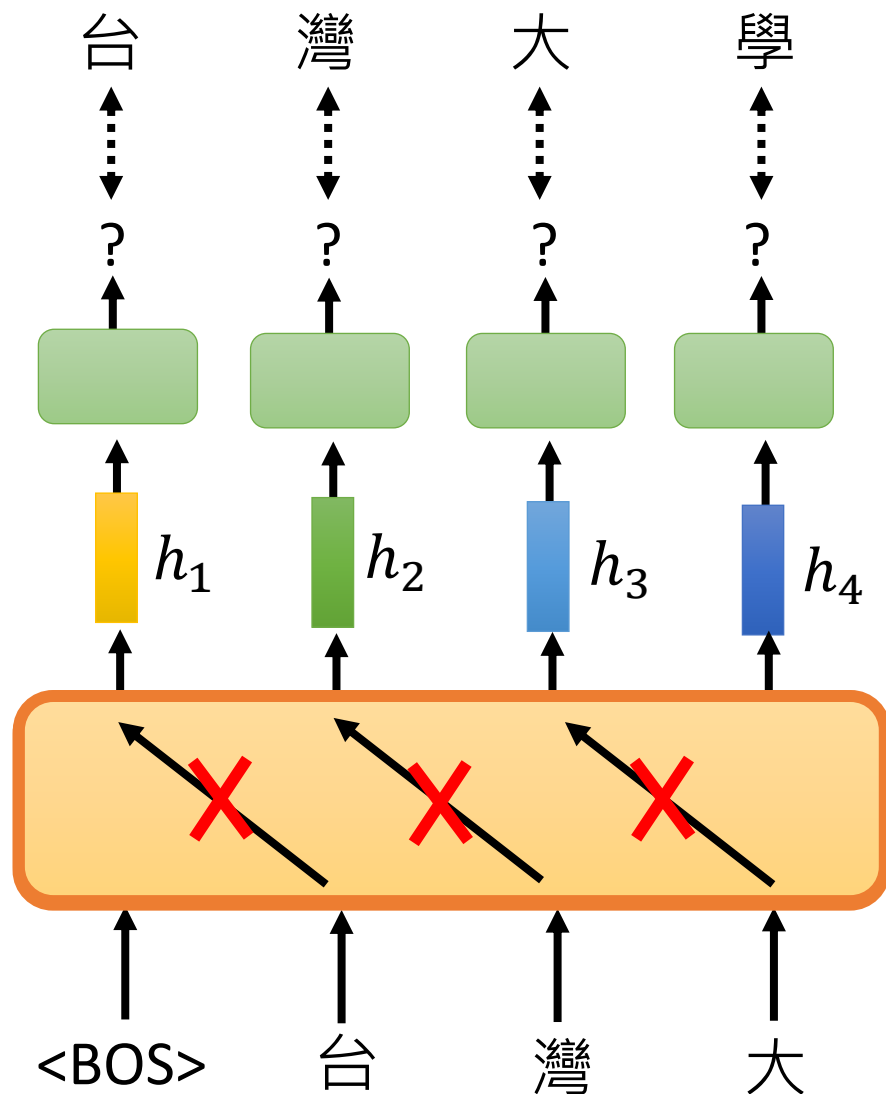


BERT series

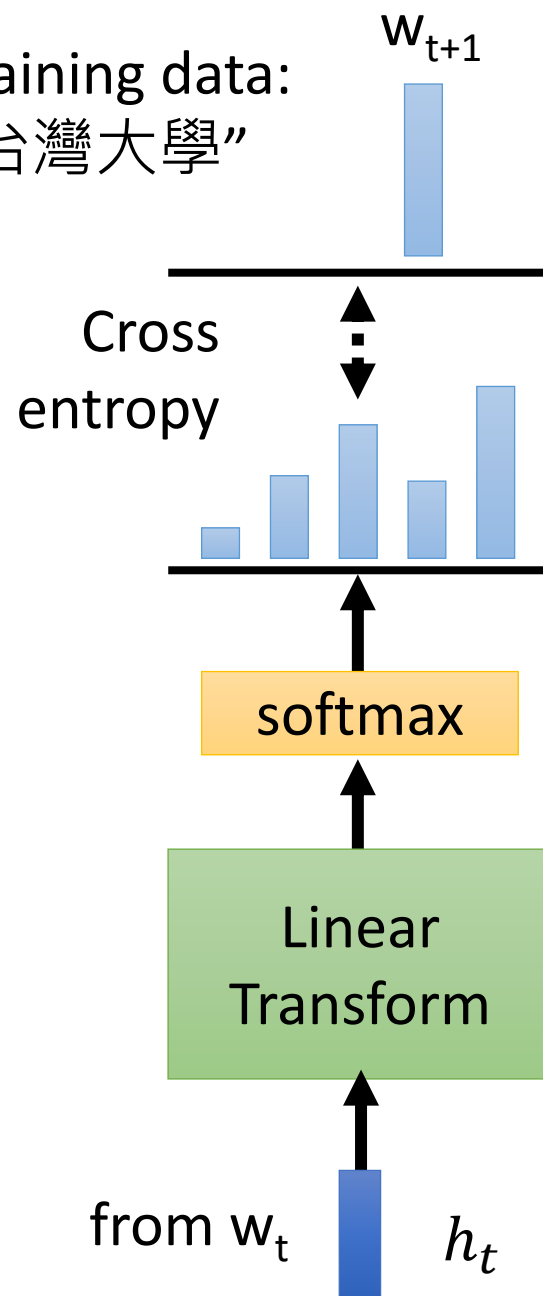


GPT series

Predict Next Token



Training data:
“台灣大學”



Predict Next Token

They can do generation.



EM PROMPT
-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. 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.

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。

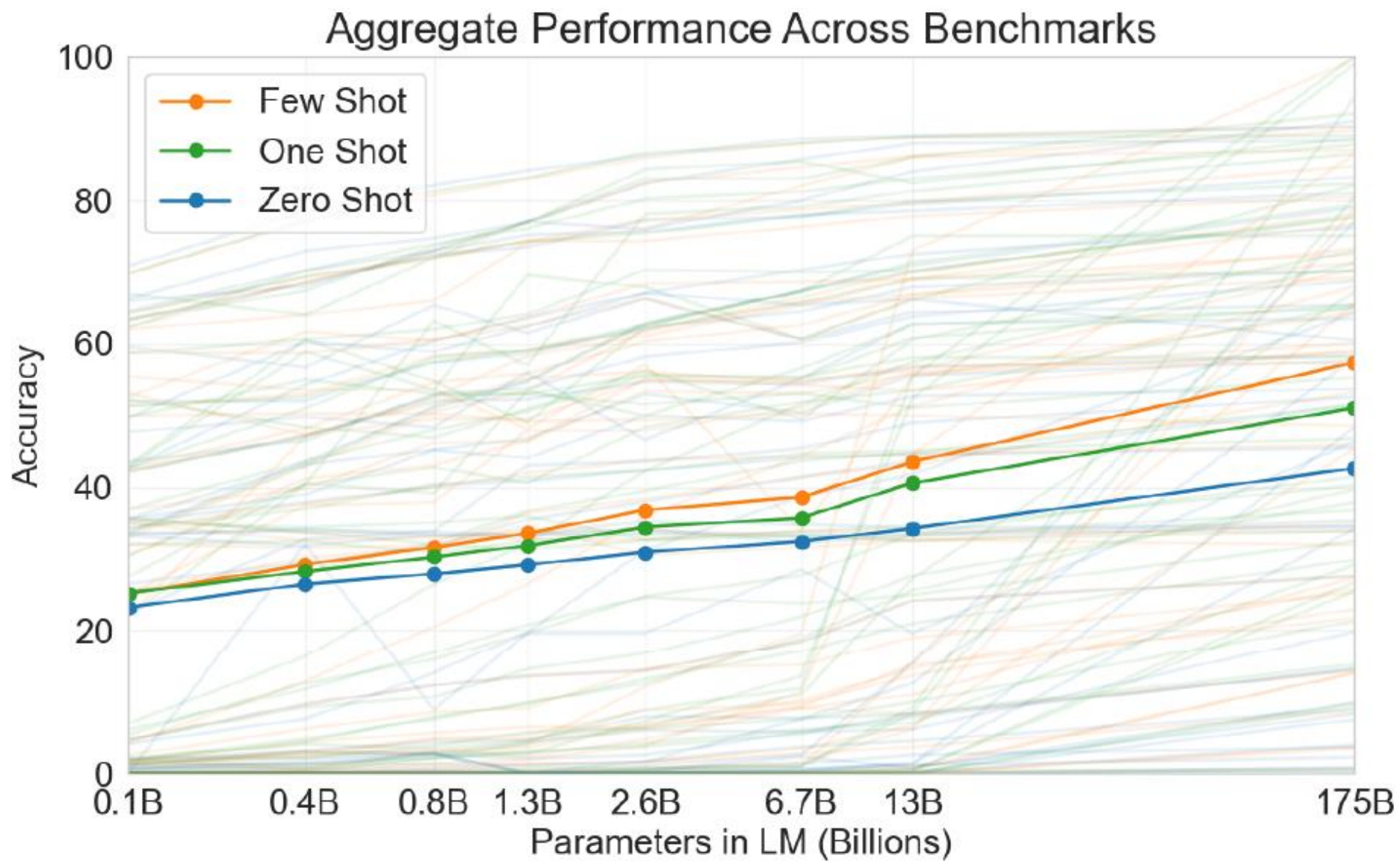
A few example

“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

“In-context” Learning



Average of **42** tasks

To learn more



<https://youtu.be/DOG1L9lvsDY>

Beyond Text

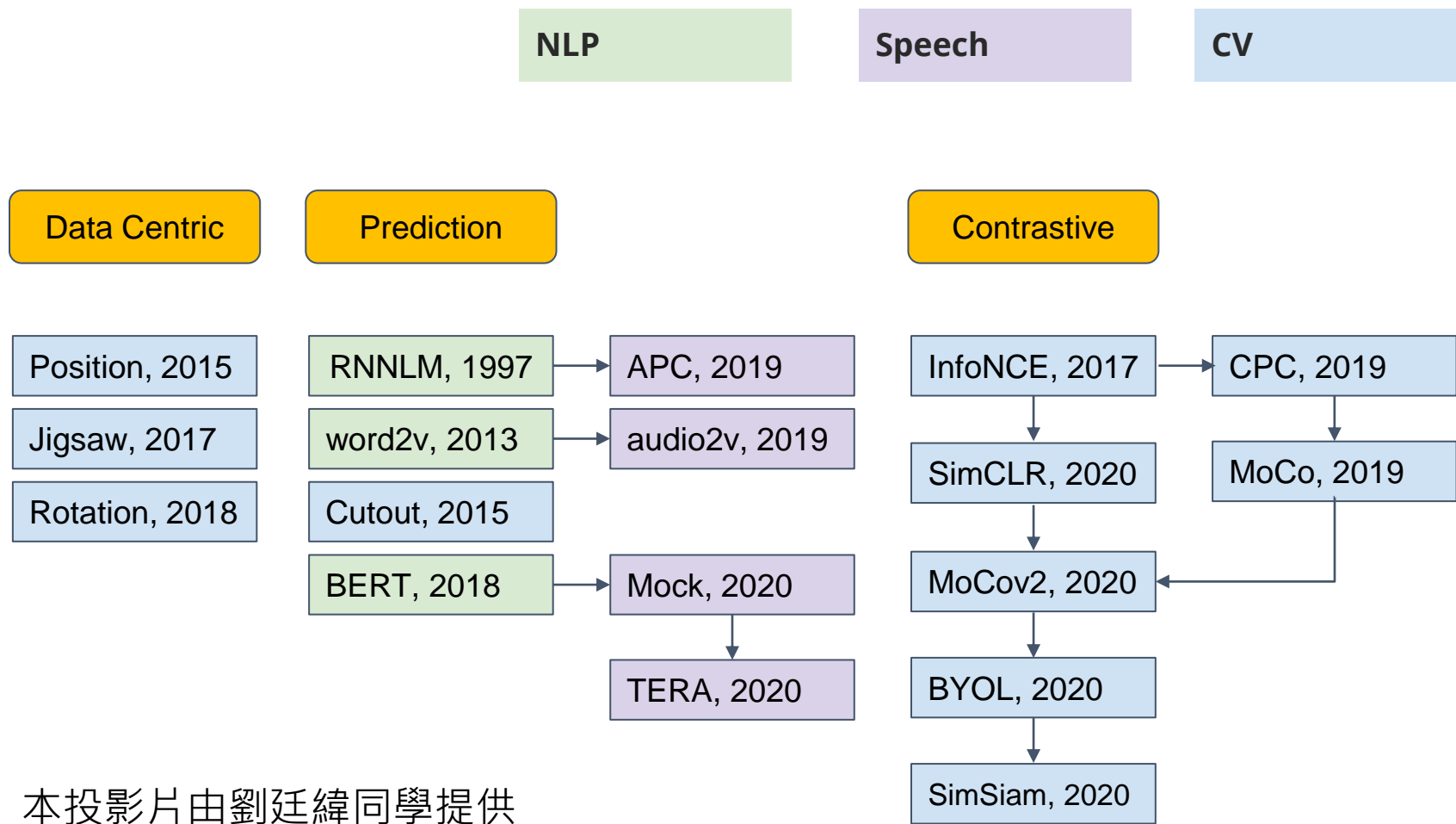


Image - SimCLR

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

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

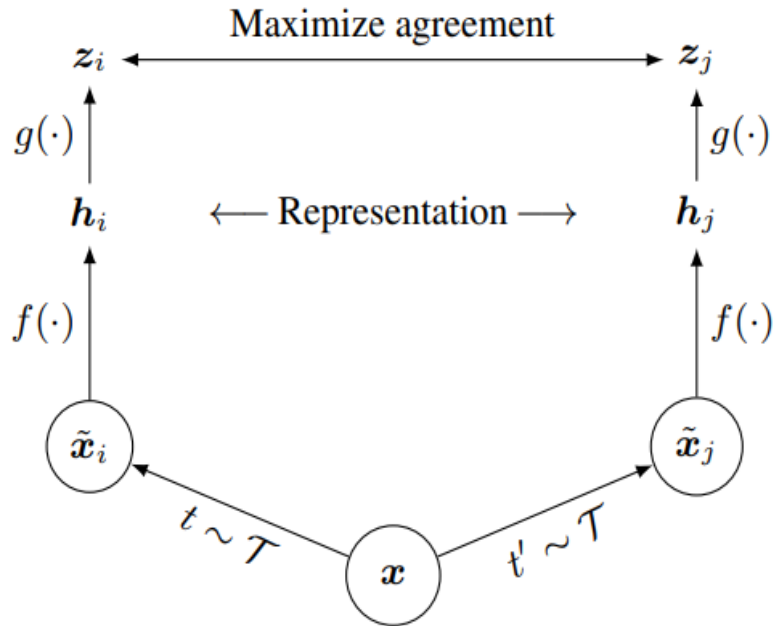
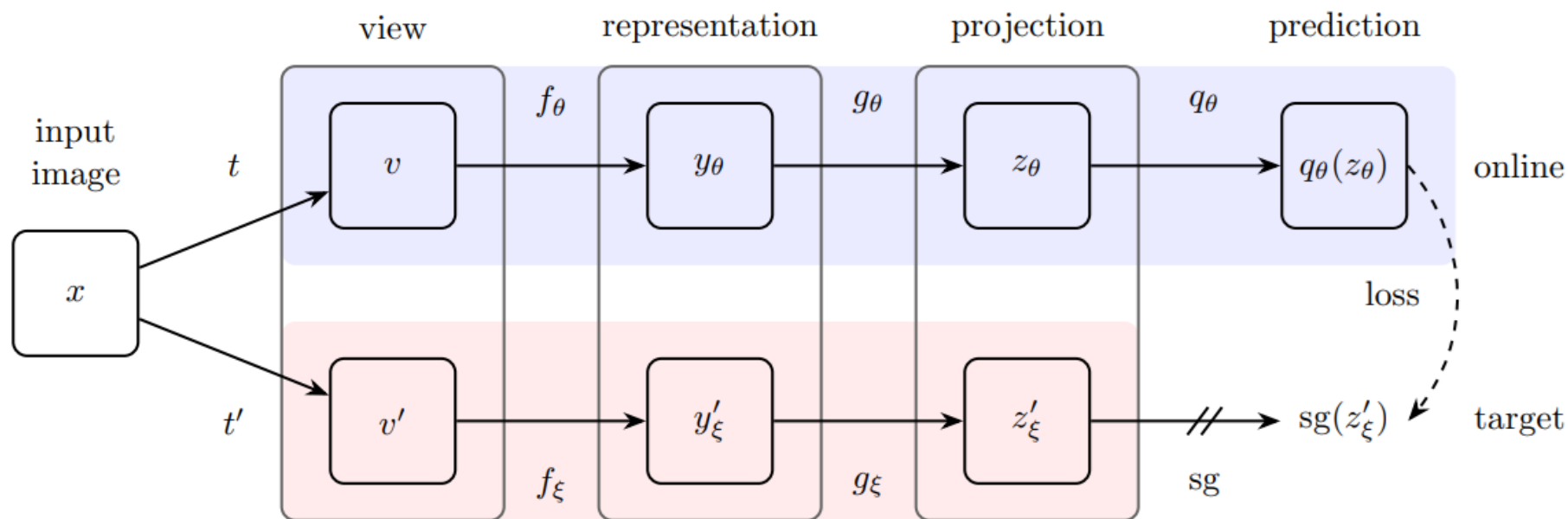


Image - BYOL

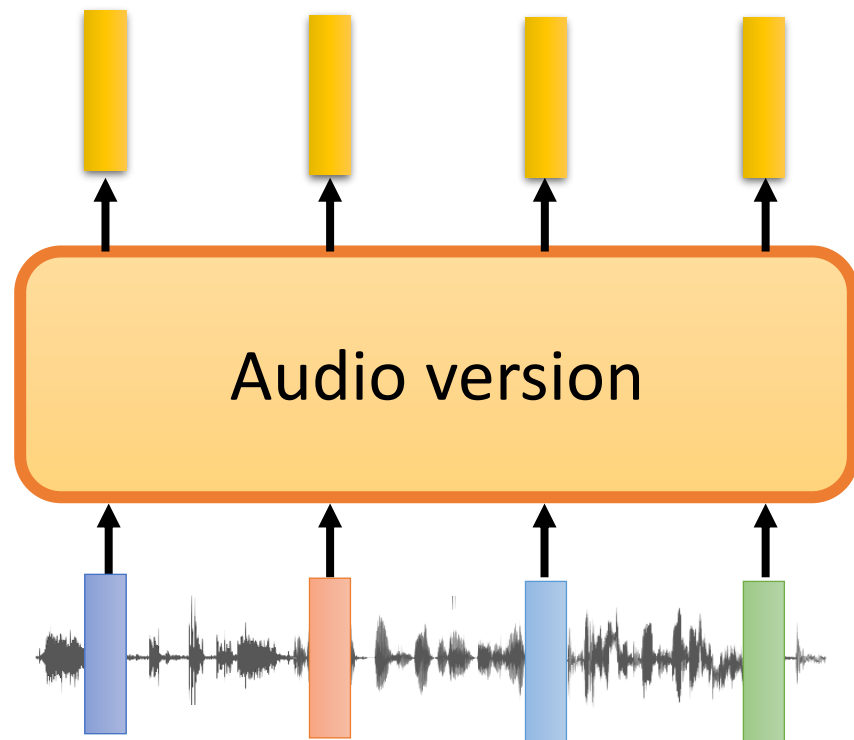
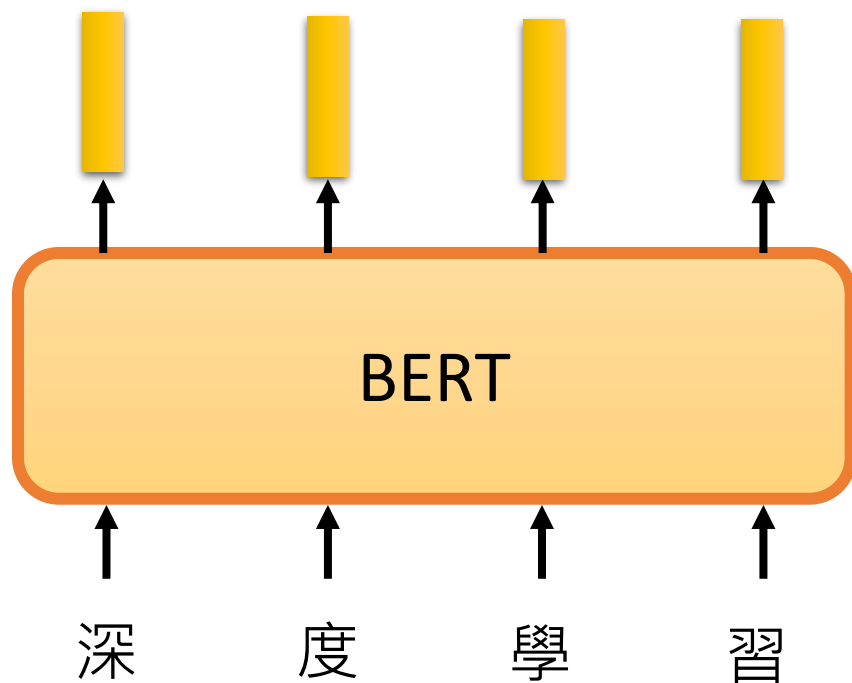
Bootstrap your own latent:

A new approach to self-supervised Learning

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

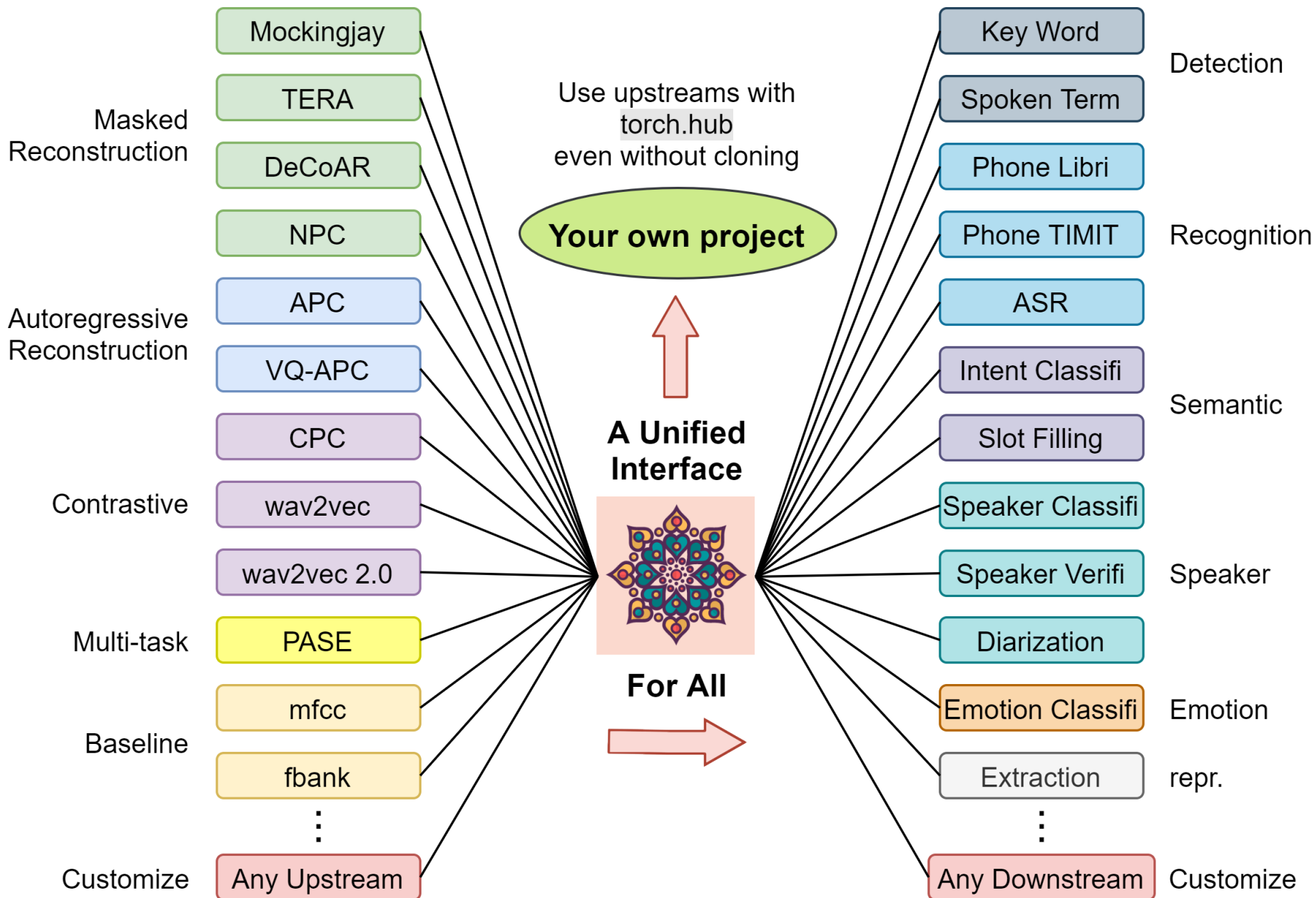


Speech



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>



Appendix

(a joke)

Predict Next Token

They can do generation.



Keaton Patti  @KeatonPatti · 2019年8月13日

I forced a bot to watch over 1,000 hours of Batman movies and then asked it to write a Batman movie of its own. Here is the first page.

BATMAN

INT. TRADITIONAL BATCAVE

BATMAN stands next to his batmobile.
He's sometimes Bruce Wayne sometime

BATMAN

This is now a safe city.
punched a penguin into p

ALFRED, Batman's loyal batler, carr

ALFRED

Eat a dinner, Mattress W

An explosion explodes. THE JOKER ar
Joker is a clown but insane. Two-Pe

BATMAN

No! It is Two-Face and O
They hate me for being a

Batman throws Alfred at Two-Face. I
a coin. Alfred lands heads up which

BATMAN (CONT'D)

It is just you and I, the
Bat versus clown. Moral :

THE JOKER

I am such a freak. Society i
You drink water, I drink ana

BATMAN

I drink bats just like a bat

Batman looks around for his parents, b
This makes him have anger. He fires a
deflects it with his sick sense of hum

THE JOKER

I have never followed a rule
is my rule. Do you follow? I

BATMAN

Alfred, give birth to Robin.

Alfred begins the process since it is
has a present in his hand. He juggles

THE JOKER

Happy batday, Birthman.

Batman opens the present since he's a
coupon for new parents, but is expired

4,165

5.4萬

14.3萬



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.

BATMAN

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

律師

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!!!



Keaton Patti ✓

@KeatonPatti

I forced a bot to watch over 1,000 hours
of Olive Garden commercials and then

ask
con
pag



Keaton Patti ✓

@KeatonPatti

I forced a bot to watch over 1,000
episodes of Jerry Springer and then

asked
Here



Keaton Patti ✓

@KeatonPatti

I forced a bot to watch over 1,000 hours
of the Saw movies and then asked it to
write a Saw movie of its own. Here is the
first page.