BERT

李宏毅

Hung-yi Lee

1-of-N Encoding

apple = [1 0 0 0 0]

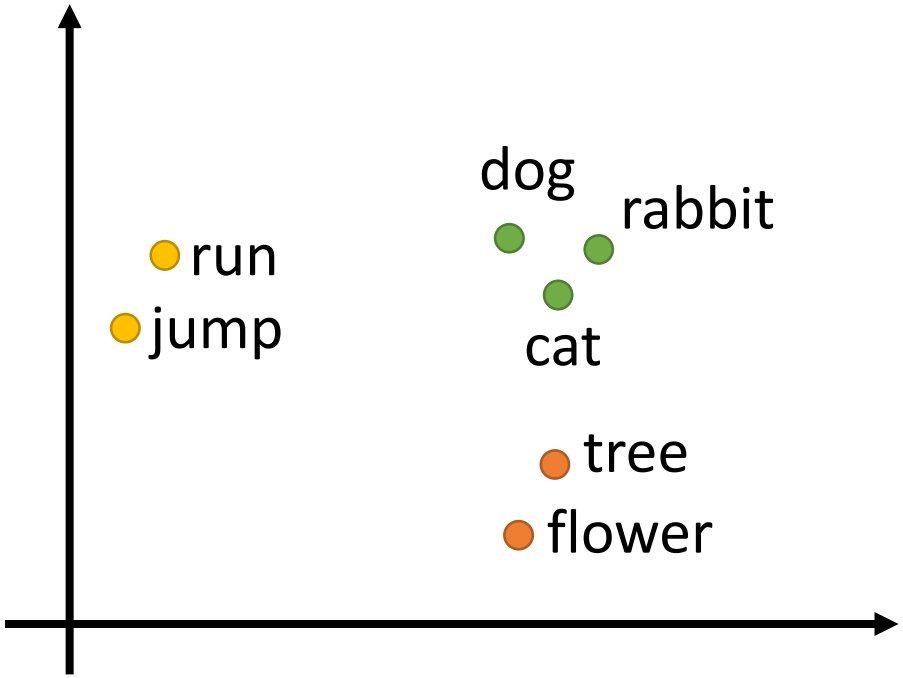
bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

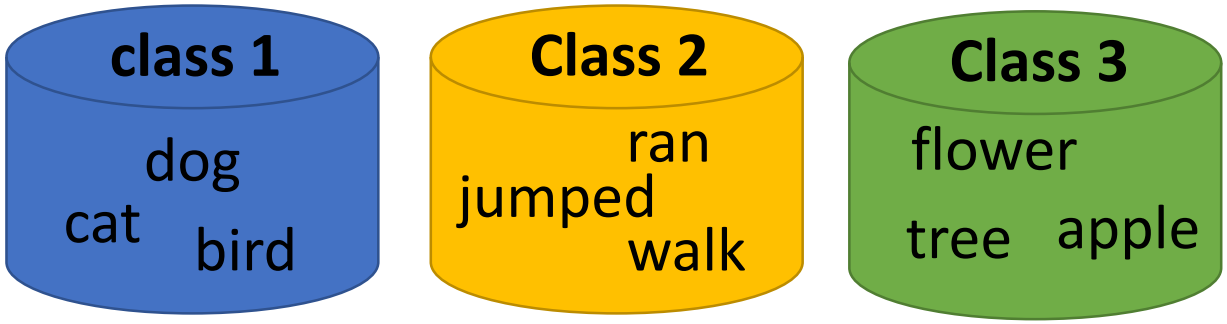
dog = [0 0 0 1 0]

elephant = [0 0 0 0 1]

Word Embedding



Word Class



https://arxiv.org/abs/1902.06006

A word can have multiple senses.

Have you paid that money to the bank yet ?
It is safest to deposit your money in the bank .

The victim was found lying dead on the river bank .
They stood on the river bank to fish.

The hospital has its own blood bank.

The third sense or not?

type token embedding

More Examples

這是
加賀號護衛艦

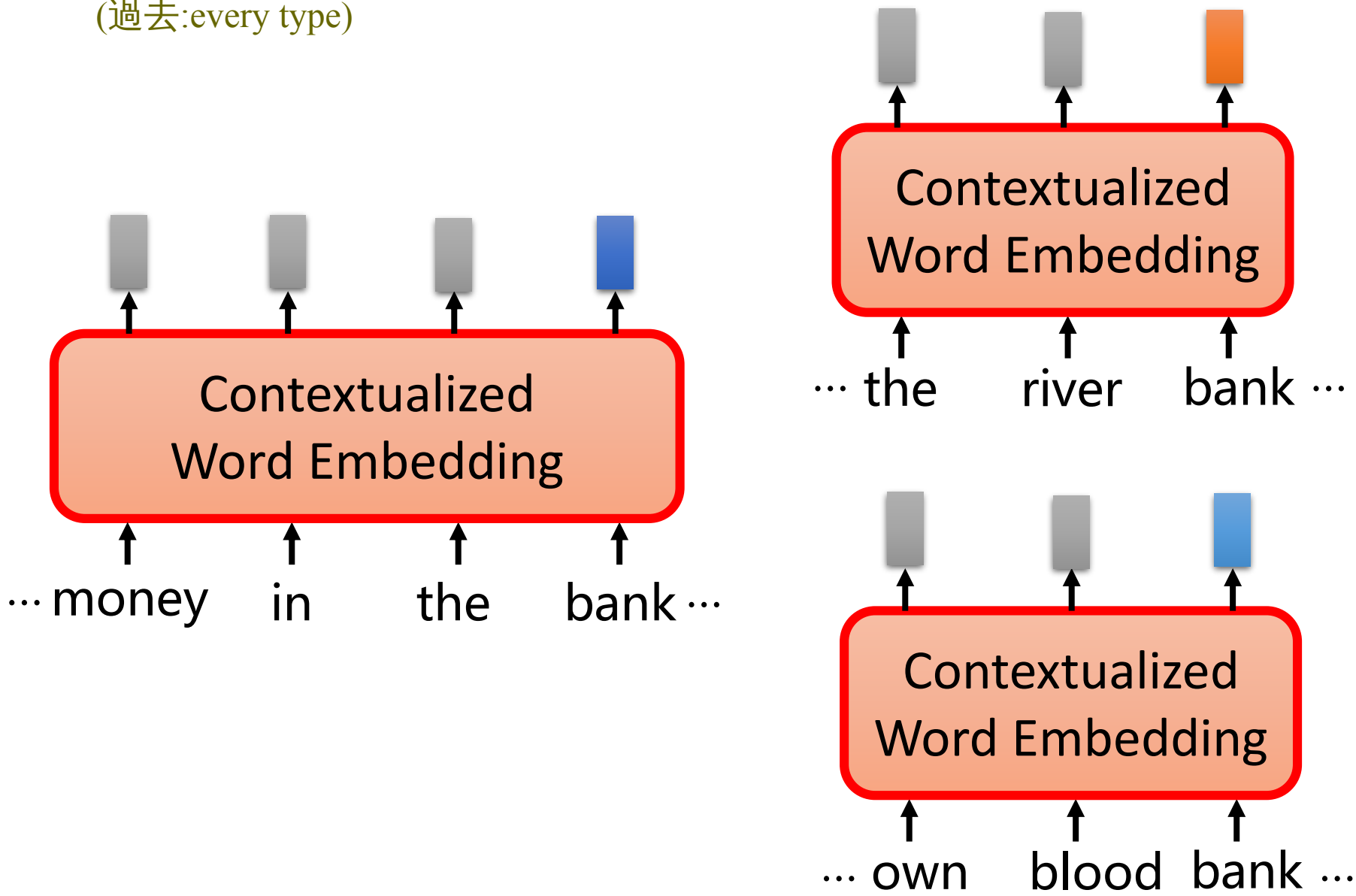
他是尼祿

她也是尼祿

這也是加賀
號護衛艦

Contextualized Word Embedding

every token都有一個embedding even though it has the same word type
(過去:every type)



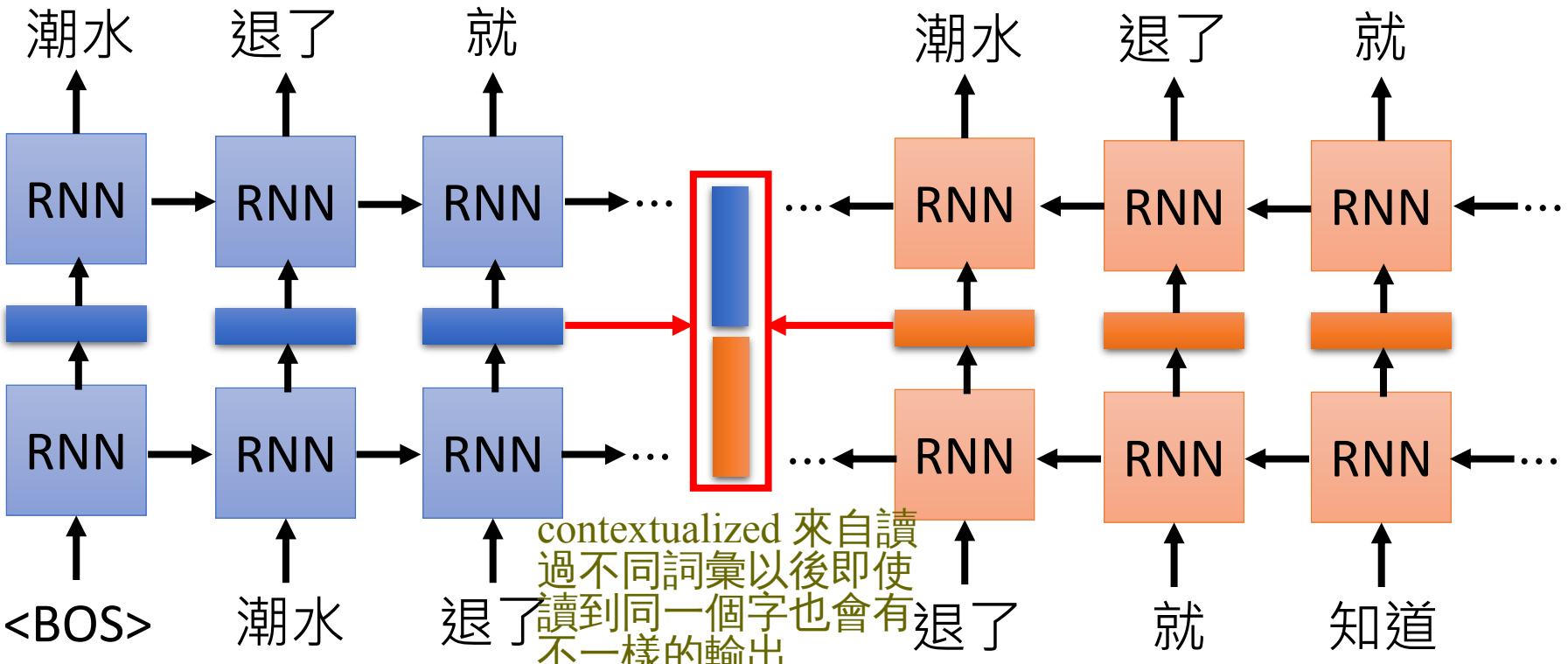
Embeddings from Language Model (ELMO)

https://arxiv.org/abs/1802.05365

- RNN-based language models (trained from lots of sentences)

e.g. given “潮水 退了 就 知道 誰 沒穿 褲子”

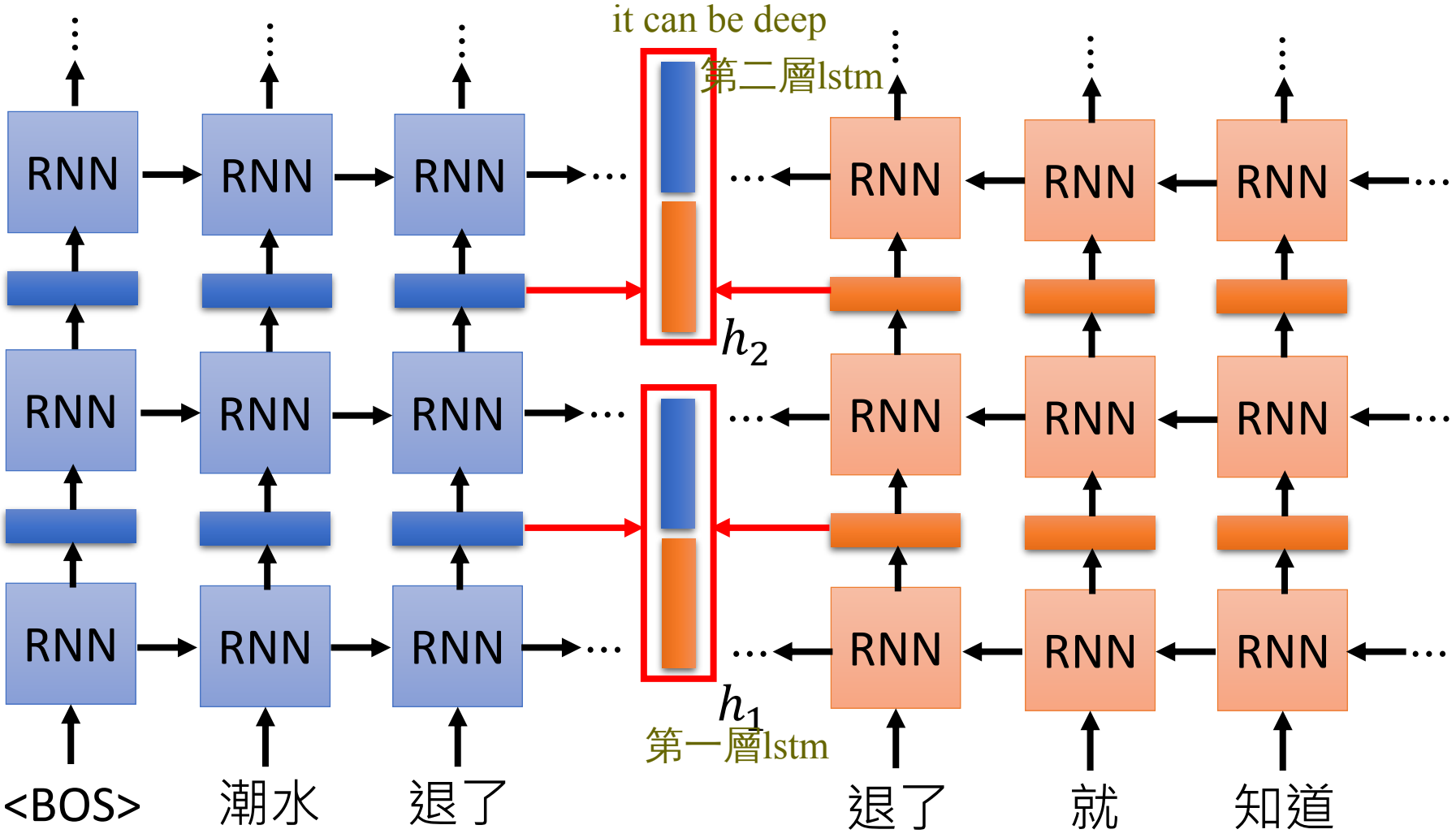
不只考慮前文 還考慮下文



ELMO

Each layer in deep LSTM can generate a latent representation.

Which one should we use???




我全都要

ELMO

type: 字長的樣子
token: 字義

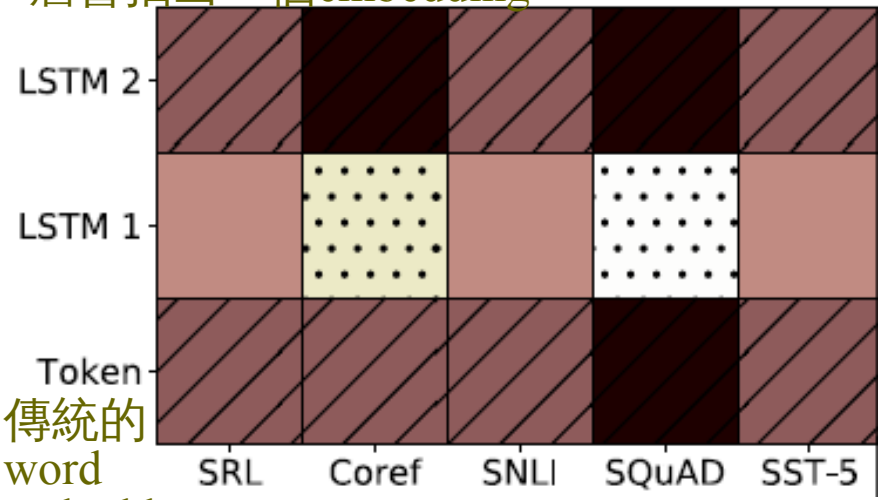
每個embedding加起來做
weighted sum

$= \alpha_1 + \alpha_2$

Learned with the down stream tasks

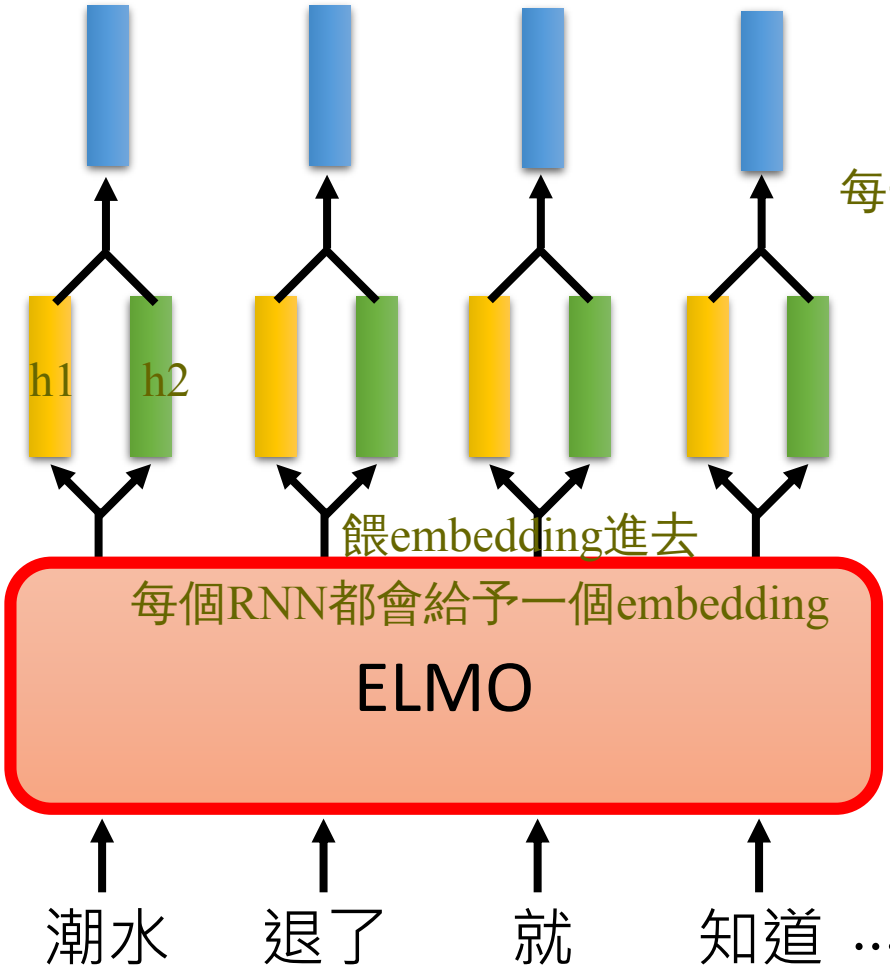
different tasks with different alphas

每一層會抽出一個embedding



傳統的 word embedding

small large

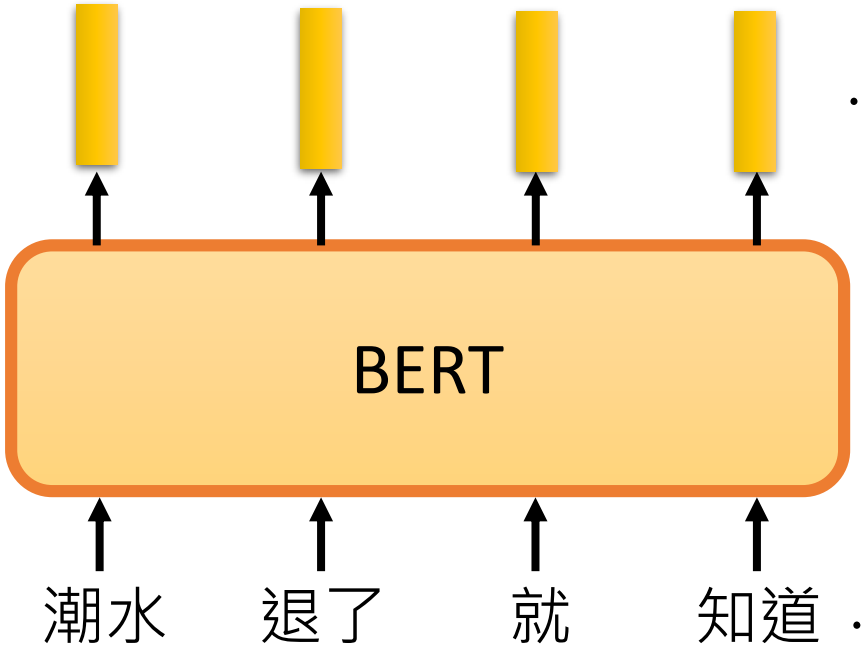


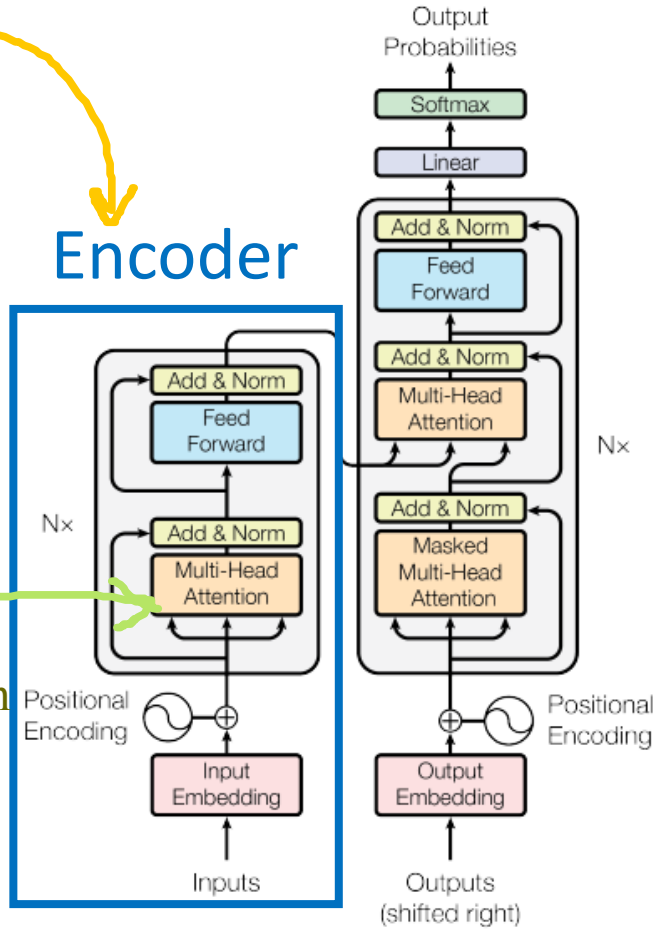
Bidirectional Encoder Representations from Transformers (BERT)

- BERT = Encoder of Transformer

Learned from a large amount of text

without annotation 不需要正確答案(label)





self attention layer

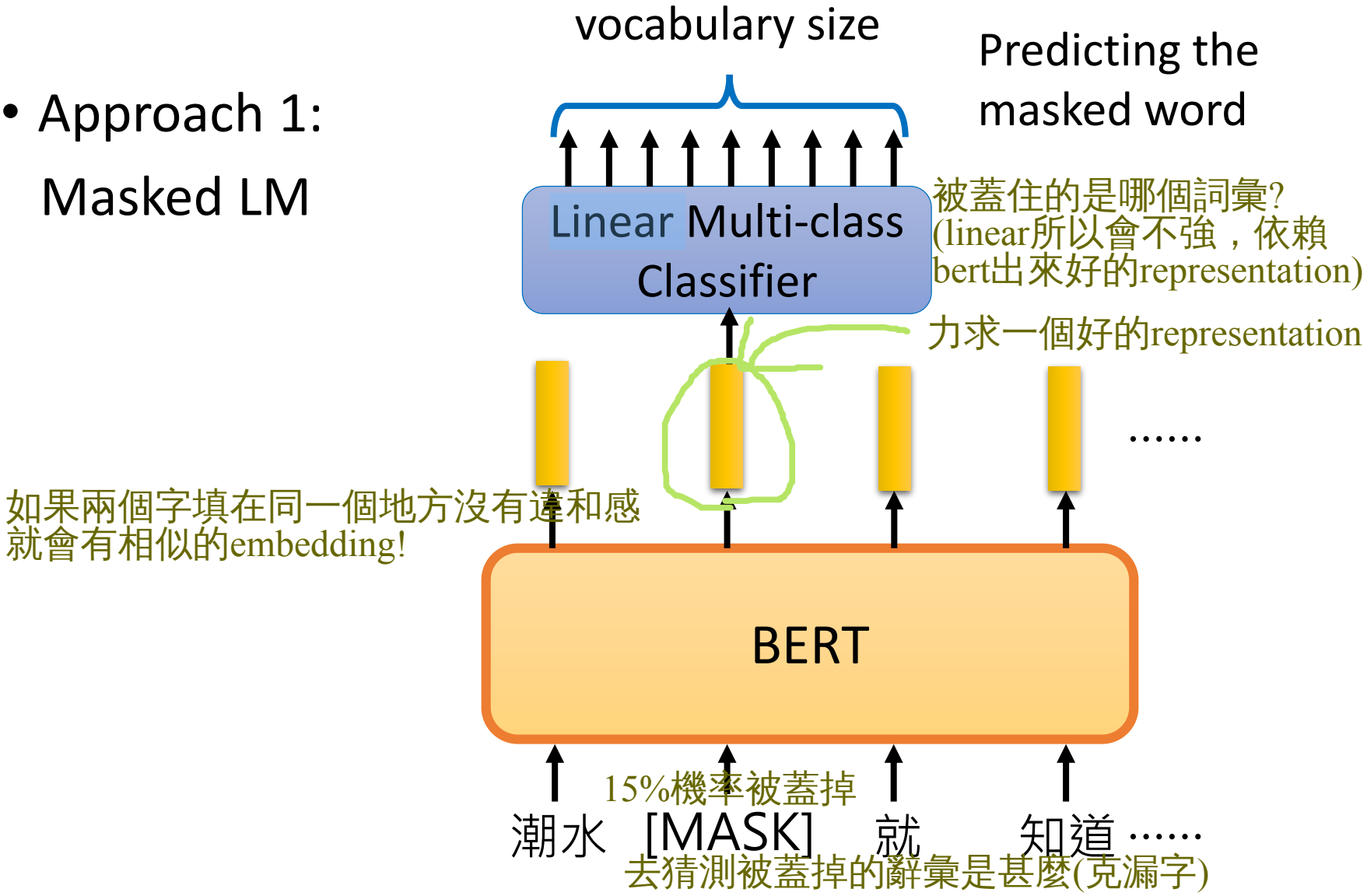
token unit, one hot : 中文的character會比word(詞)比較方便以及合適

pretrain: approach
task:cases

Training of BERT

- Approach 1:

  Masked LM



Training of BERT

是self attention not RNN因此CLS放在句子開頭或結尾都是一樣的(天涯若比鄰)

Approach 2: Next Sentence Prediction

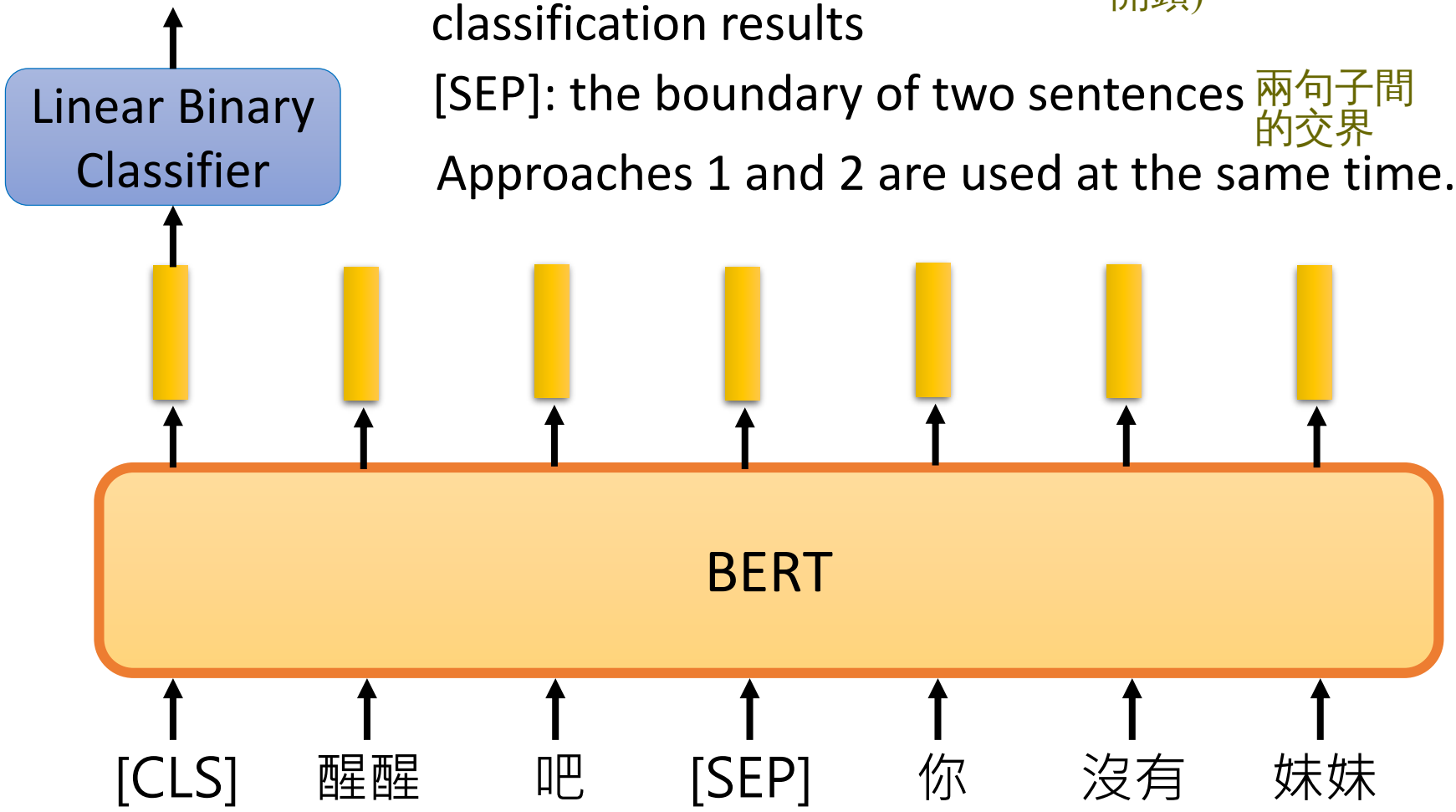
兩個句子是否應該要接再一起? yes

代表在這個位置要做分類(放在兩個句子的開頭)

[CLS]: the position that outputs classification results

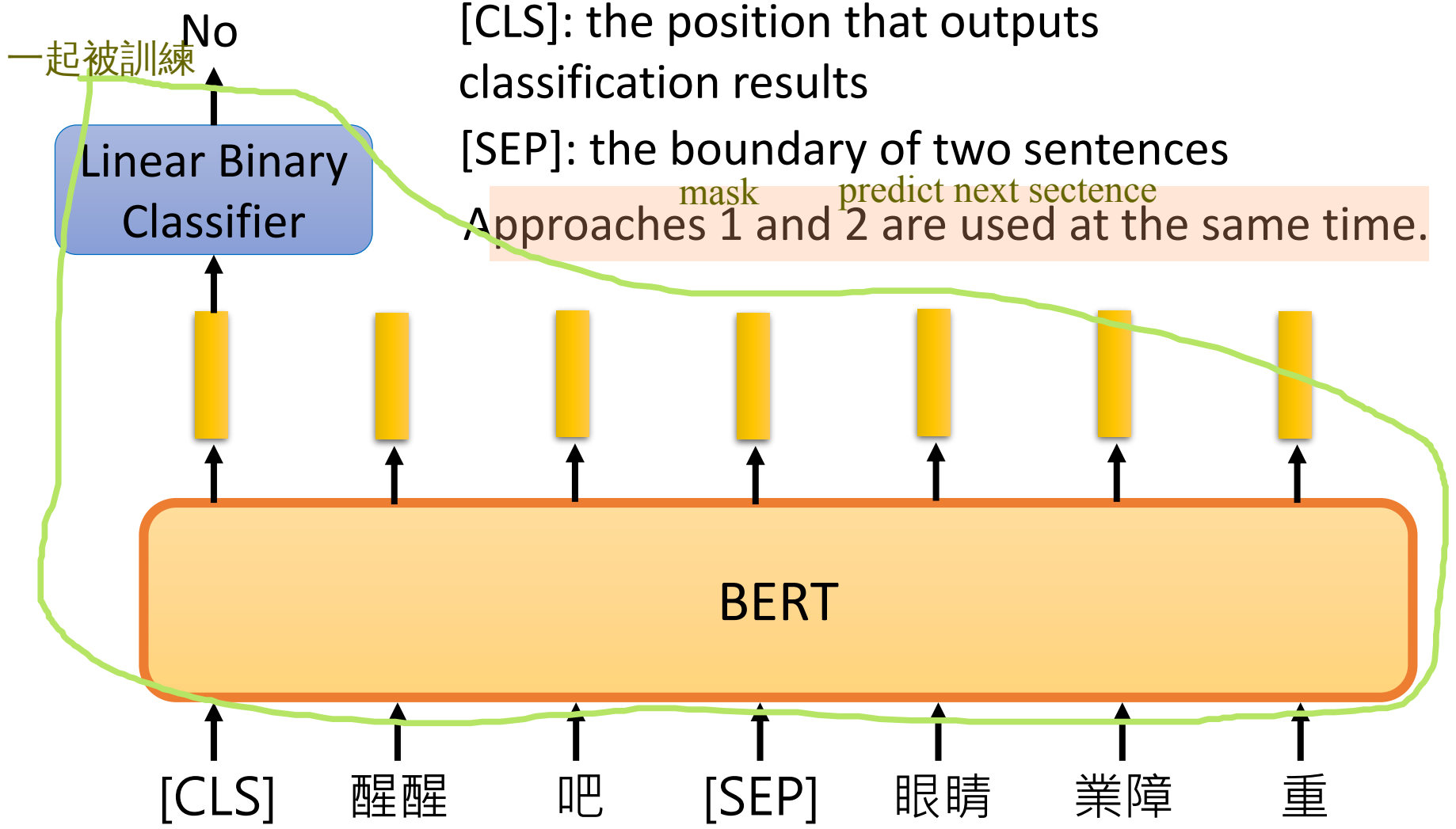
[SEP]: the boundary of two sentences 兩句子間的交界

Approaches 1 and 2 are used at the same time.



Training of BERT

Approach 2: Next Sentence Prediction



[CLS]: the position that outputs classification results

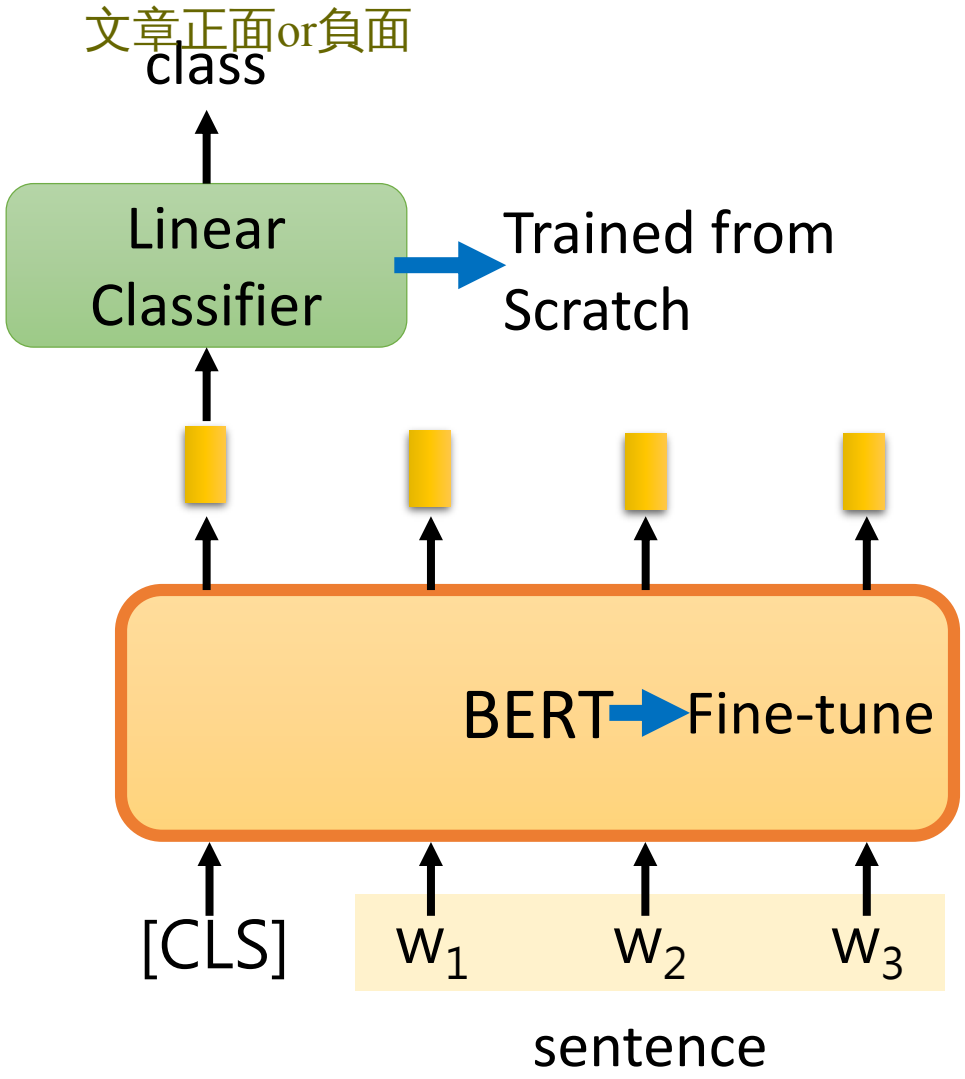
[SEP]: the boundary of two sentences

mask predict next sectence

Approaches 1 and 2 are used at the same time.

https://zhuanlan.zhihu.com/p/46652512

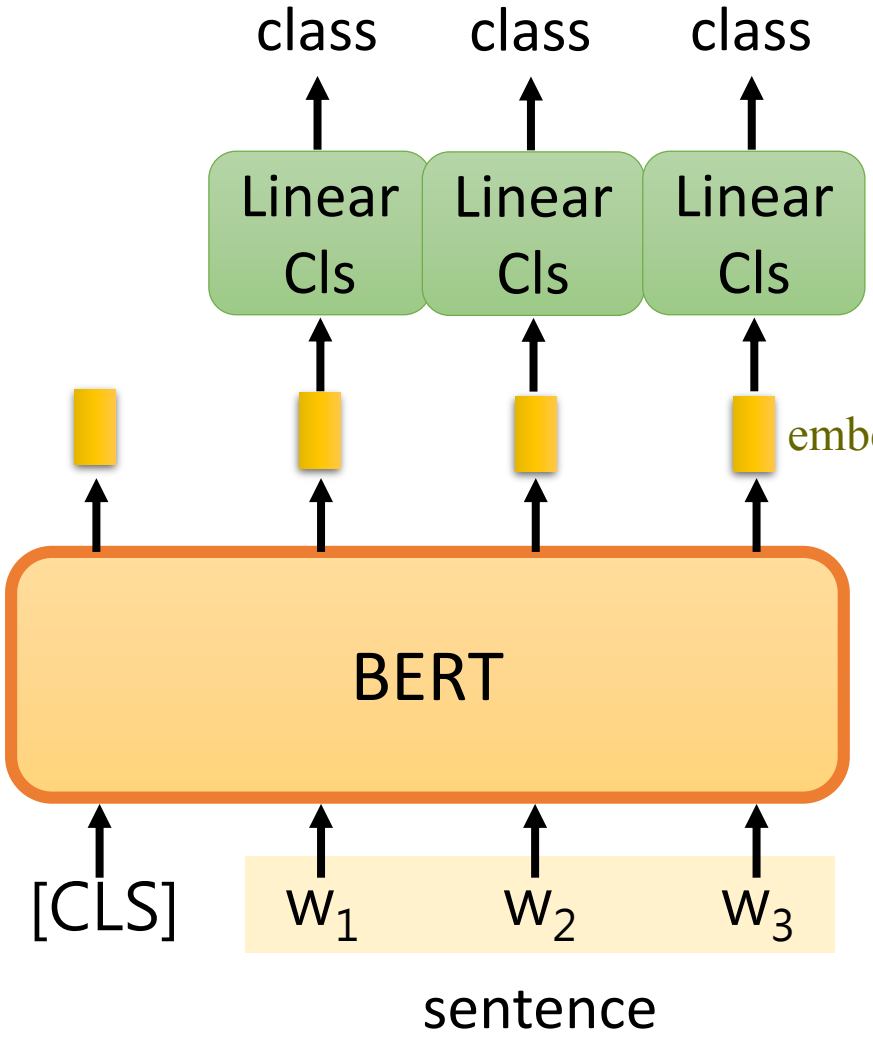
How to use BERT – Case 1



Input: single sentence, output: class

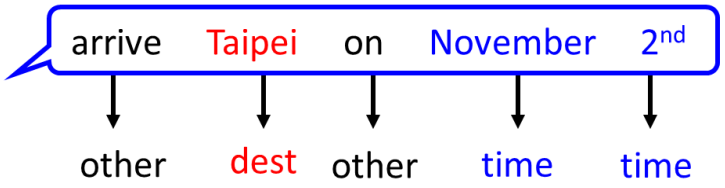
Example:
Sentiment analysis (our HW),
Document Classification

How to use BERT – Case 2

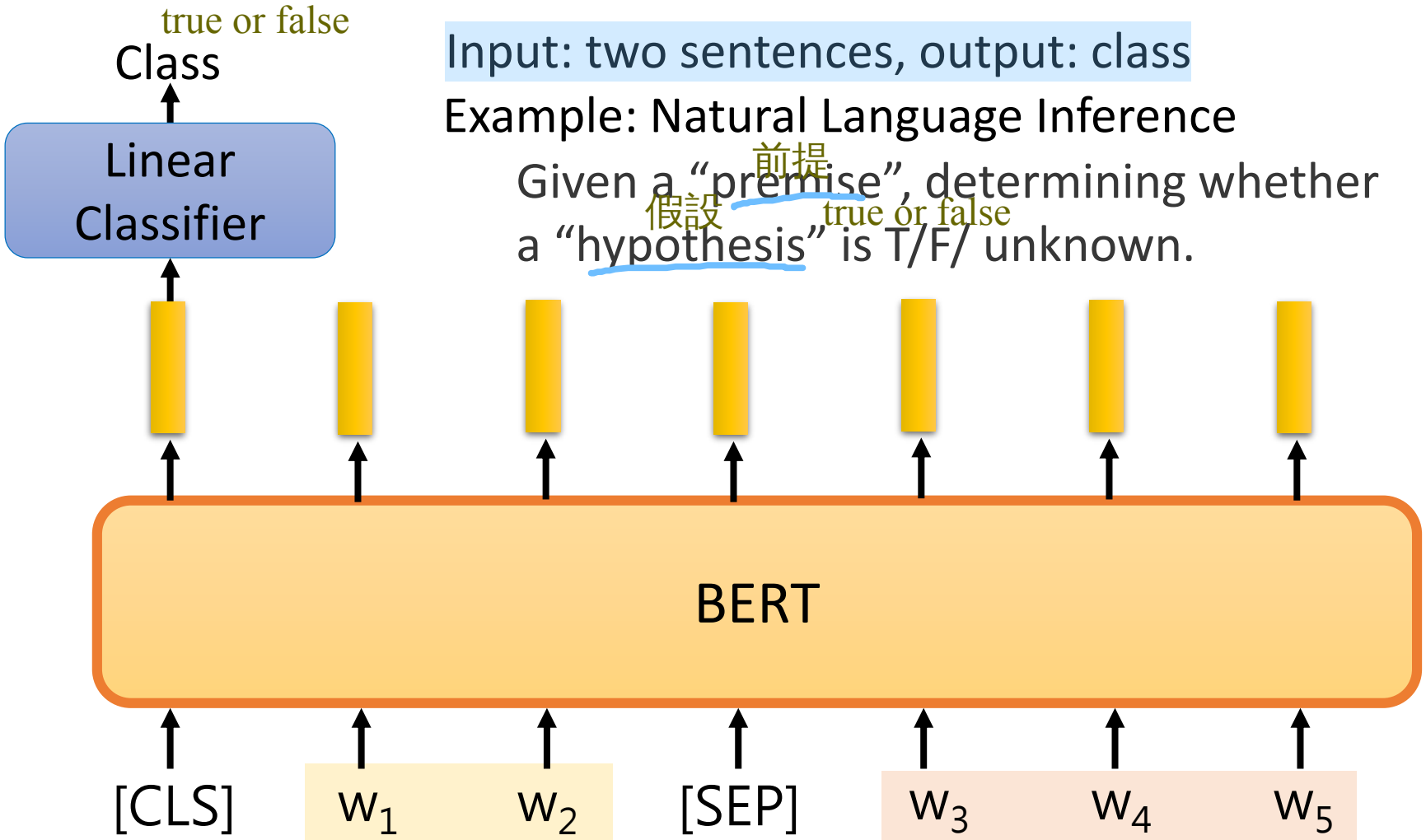


Input: single sentence, output: class of each word

Example: Slot filling



How to use BERT – Case 3



Input: two sentences, output: class

Example: Natural Language Inference

Given a "premise", determining whether a "hypothesis" is T/F/ unknown.

前提 假設 true or false

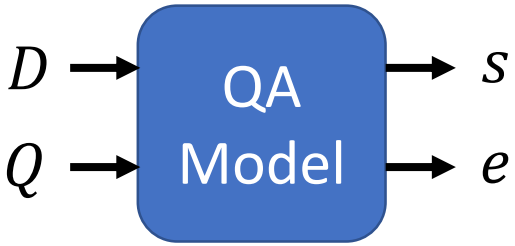
How to use BERT – Case 4

reading comprehension

- Extraction-based Question Answering (QA) (E.g. SQuAD)

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

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



output: two integers (s, e)

第s token到第e個 token

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

In meteorology, precipitation is any product of the condensation of 17spheric 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 77atte 79cations are called "showers".

此例放水:必定可以在文章中找到某字

What causes precipitation to fall?

gravity

$s = 17, e = 17$

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

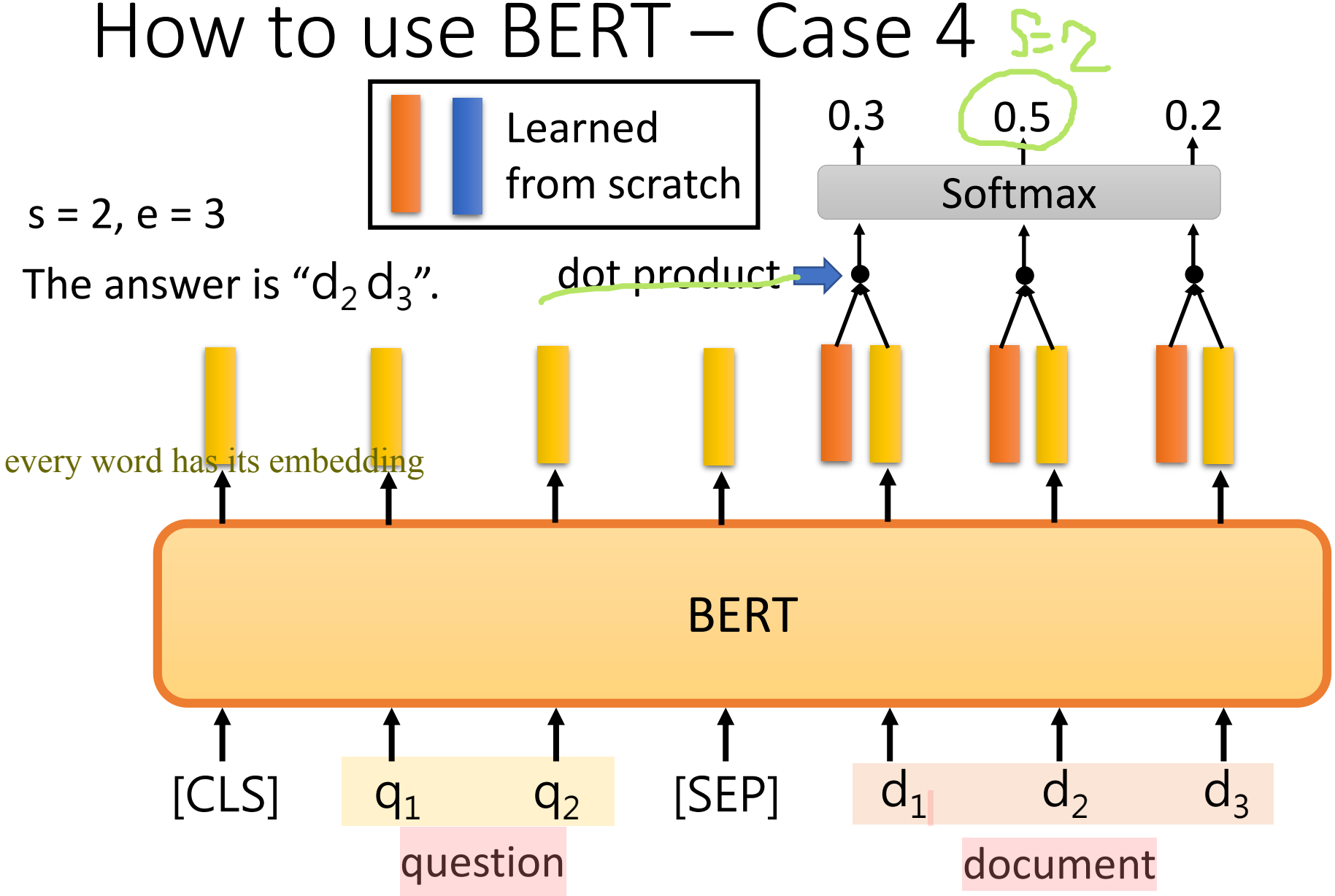
graupel

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

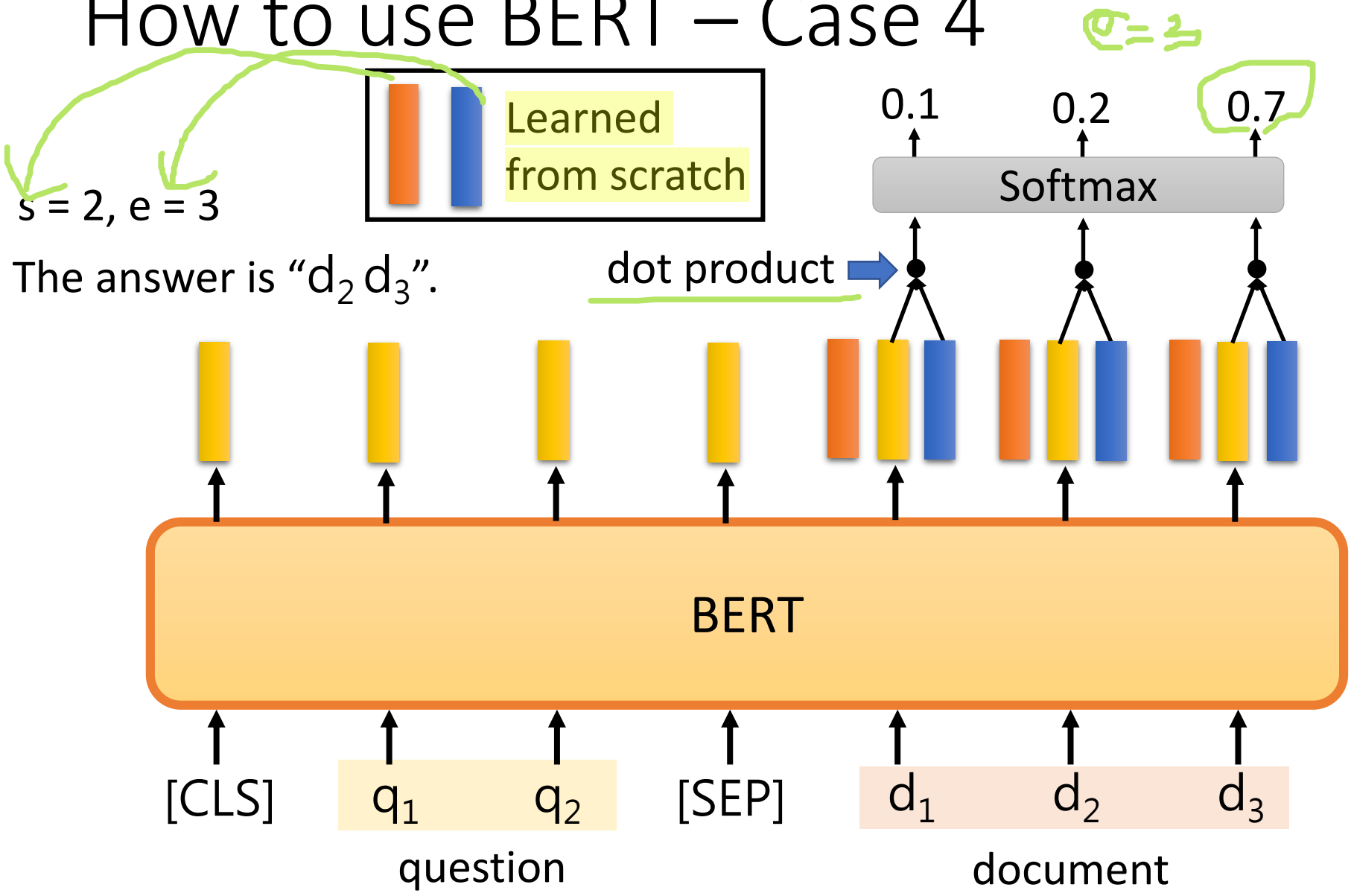
within a cloud

$s = 77, e = 79$

How to use BERT – Case 4



How to use BERT – Case 4



BERT 屠榜 .....

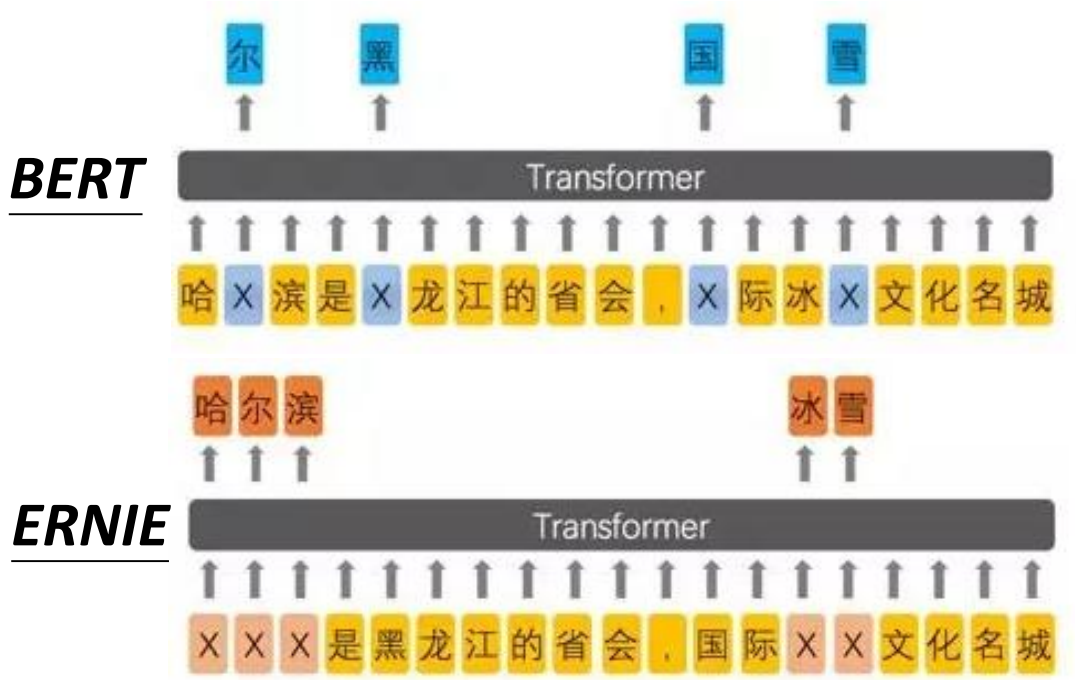
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 May 21, 2019	XLNet (single model) *XLNet Team*	86.346	89.133
5 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886

SQuAD 2.0

Enhanced Representation through Knowledge Integration (ERNIE)

- Designed for Chinese

以中文字詞為單位

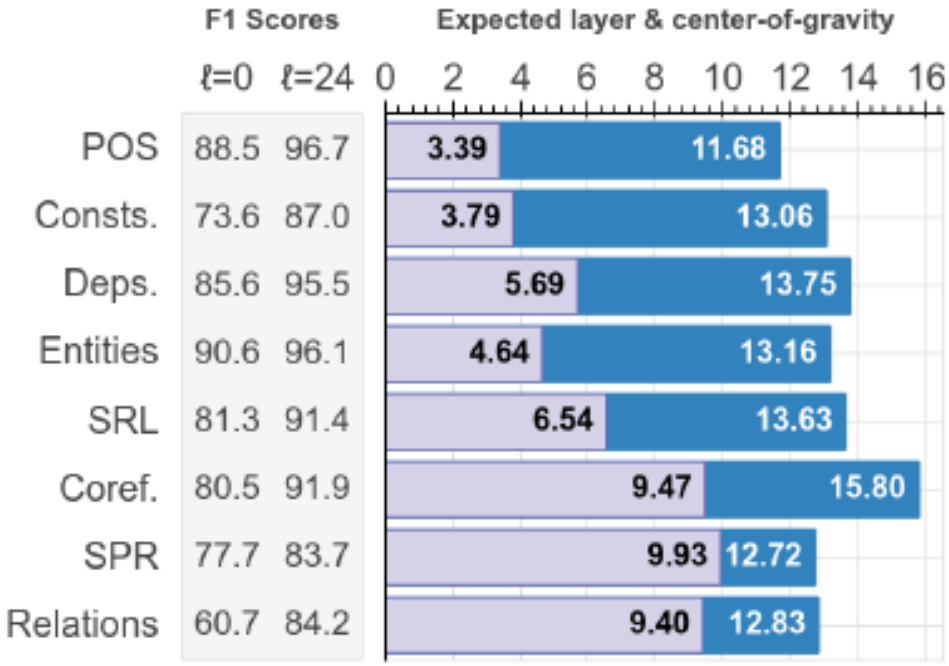


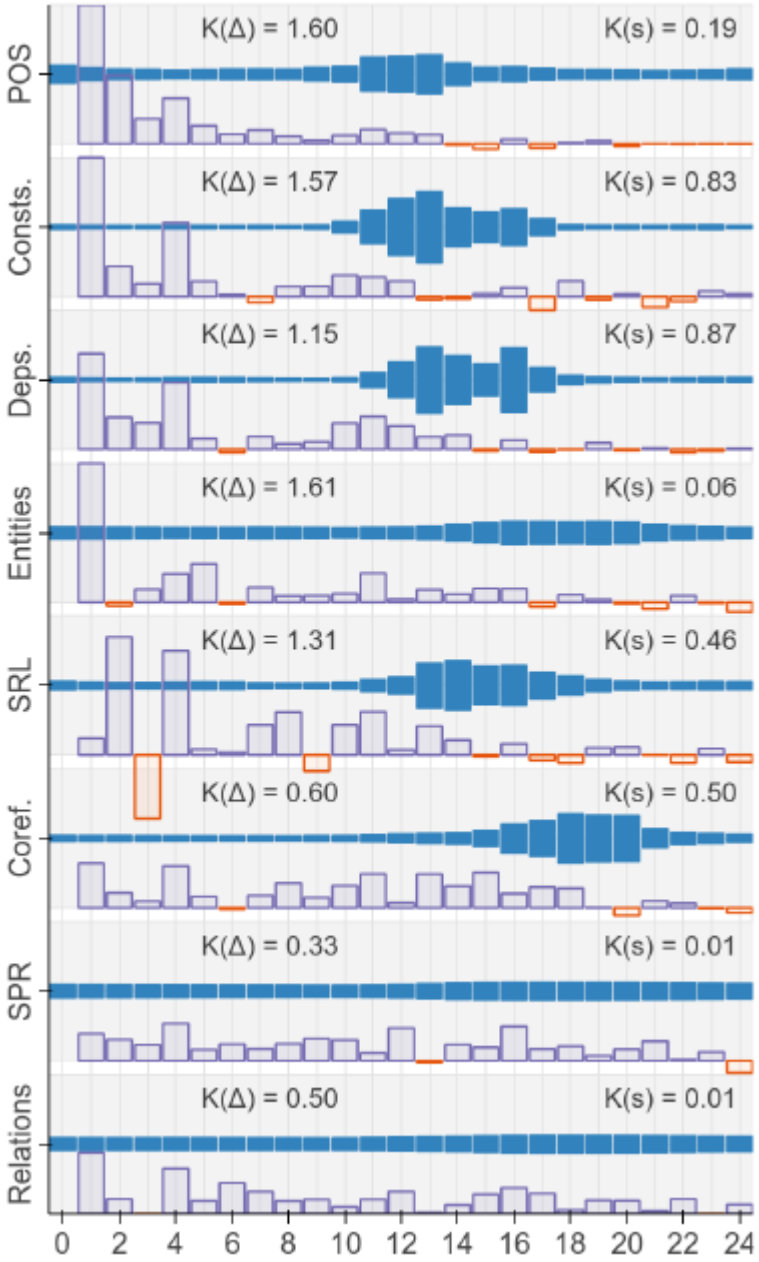
Source of image:
https://zhuanlan.zhihu.com/p/59436589

https://arxiv.org/abs/1904.09223

What does BERT learn?

https://arxiv.org/abs/1905.05950
https://openreview.net/pdf?id=SJzSgnRcKX





https://arxiv.org/abs/1904.09077

Multilingual BERT

Trained on 104 languages

看過中文文章 但未學過分類 zero-shot

Task specific training data for English

En Class 1

En Class 2

En Class 3

Task specific testing data for Chinese

Zh ?

Zh ?

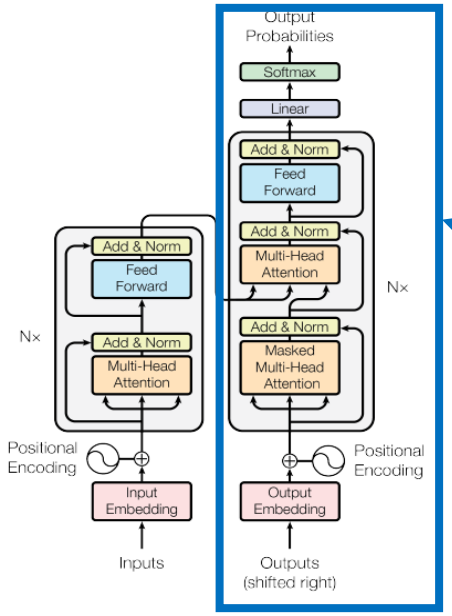
Zh ?

Zh ?

Zh ?

https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Generative Pre-Training (GPT)



Transformer Decoder

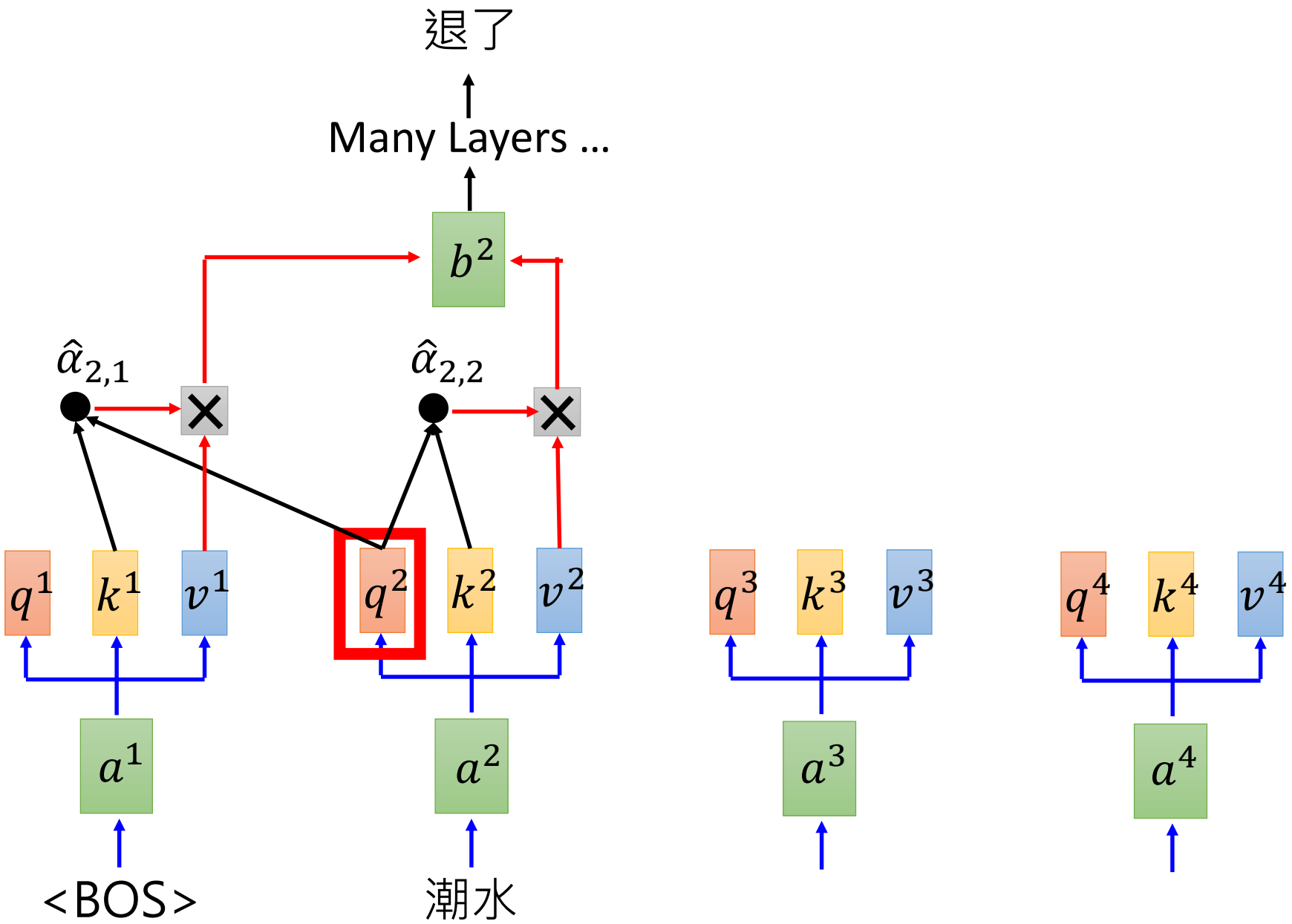
BERT (340M)

ELMO (94M) parameters

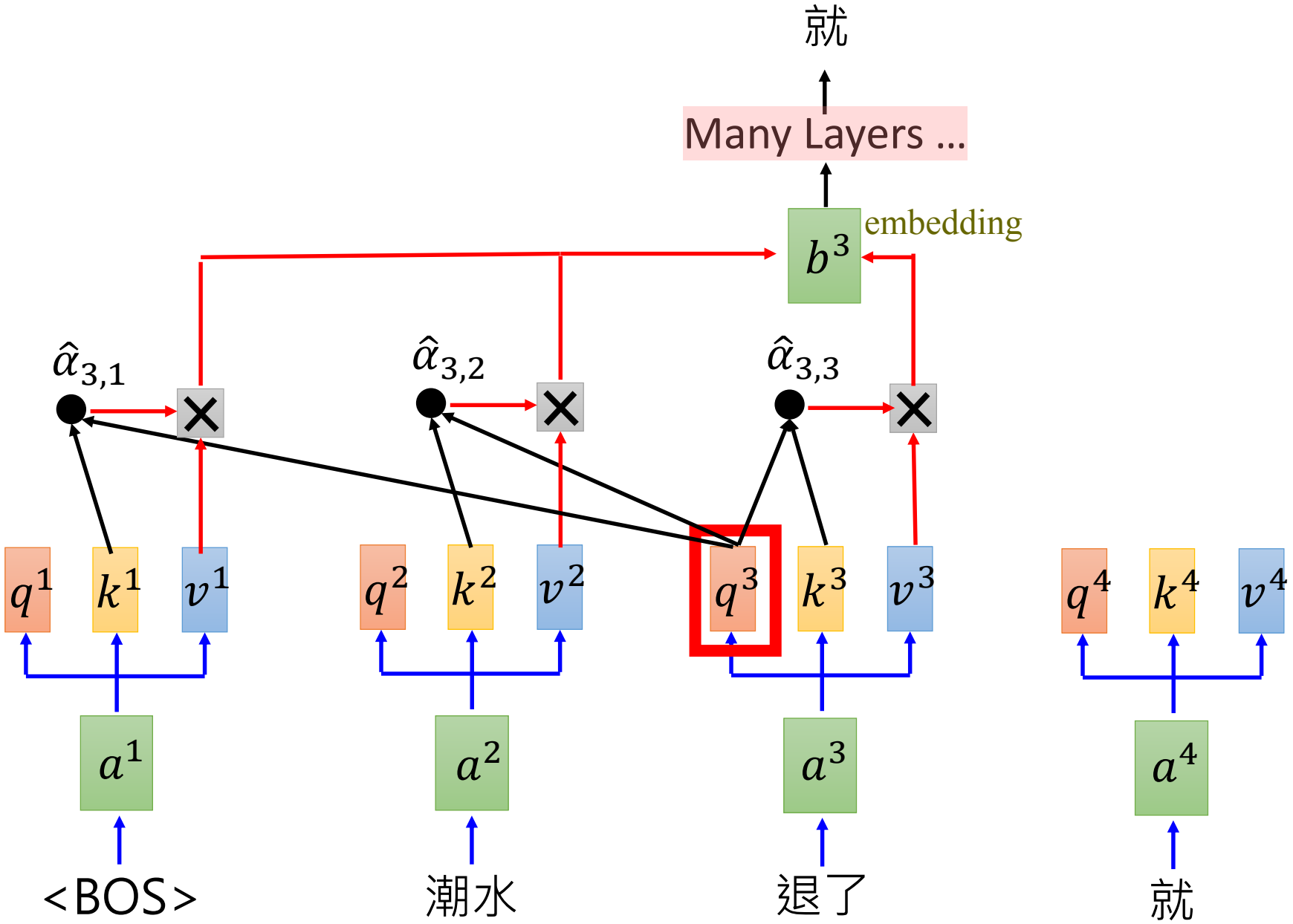
GPT-2 (1542M)

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

Generative Pre-Training (GPT)



Generative Pre-Training (GPT)

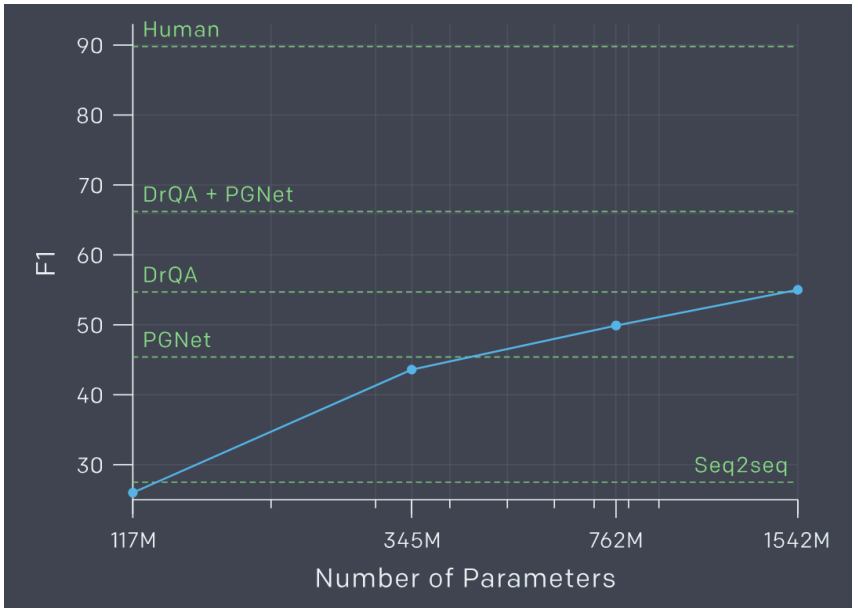


Zero-shot Learning?

GPT2

- ***Reading Comprehension***

$d_1, d_2, \cdots, d_N,$
"Q:", $q_1, q_2, \cdots, q_N,$
"A:"



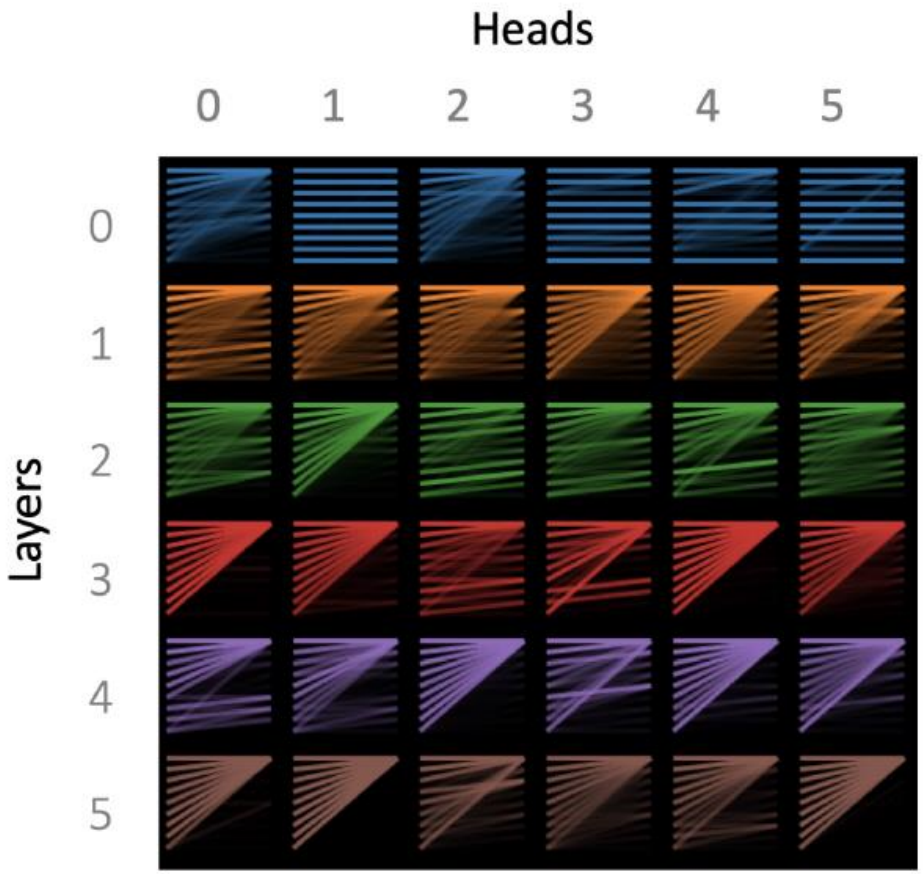
- ***Summarization*** $d_1, d_2, \cdots, d_N$,"TL;DR:"

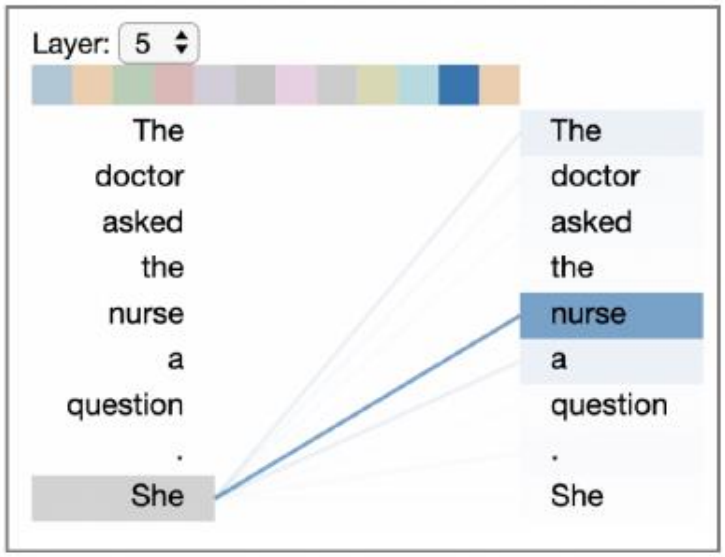
- ***Translation***

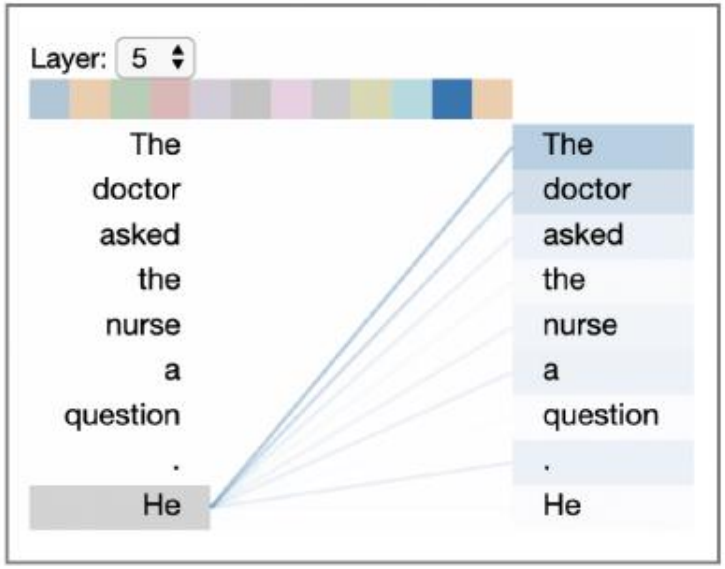
English sentence 1	=	French sentence 1
English sentence 2	=	French sentence 2
English sentence 3	=	

Visualization

https://arxiv.org/abs/1904.02679
(The results below are from GPT-2)







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.

https://talktotransformer.com/

Credit: Greg Durrett

Can BERT speak?

- Unified Language Model Pre-training for Natural Language Understanding and Generation
  - https://arxiv.org/abs/1905.03197
- BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model
  - https://arxiv.org/abs/1902.04094
- Insertion Transformer: Flexible Sequence Generation via Insertion Operations
  - https://arxiv.org/abs/1902.03249
- Insertion-based Decoding with automatically Inferred Generation Order
  - https://arxiv.org/abs/1902.01370