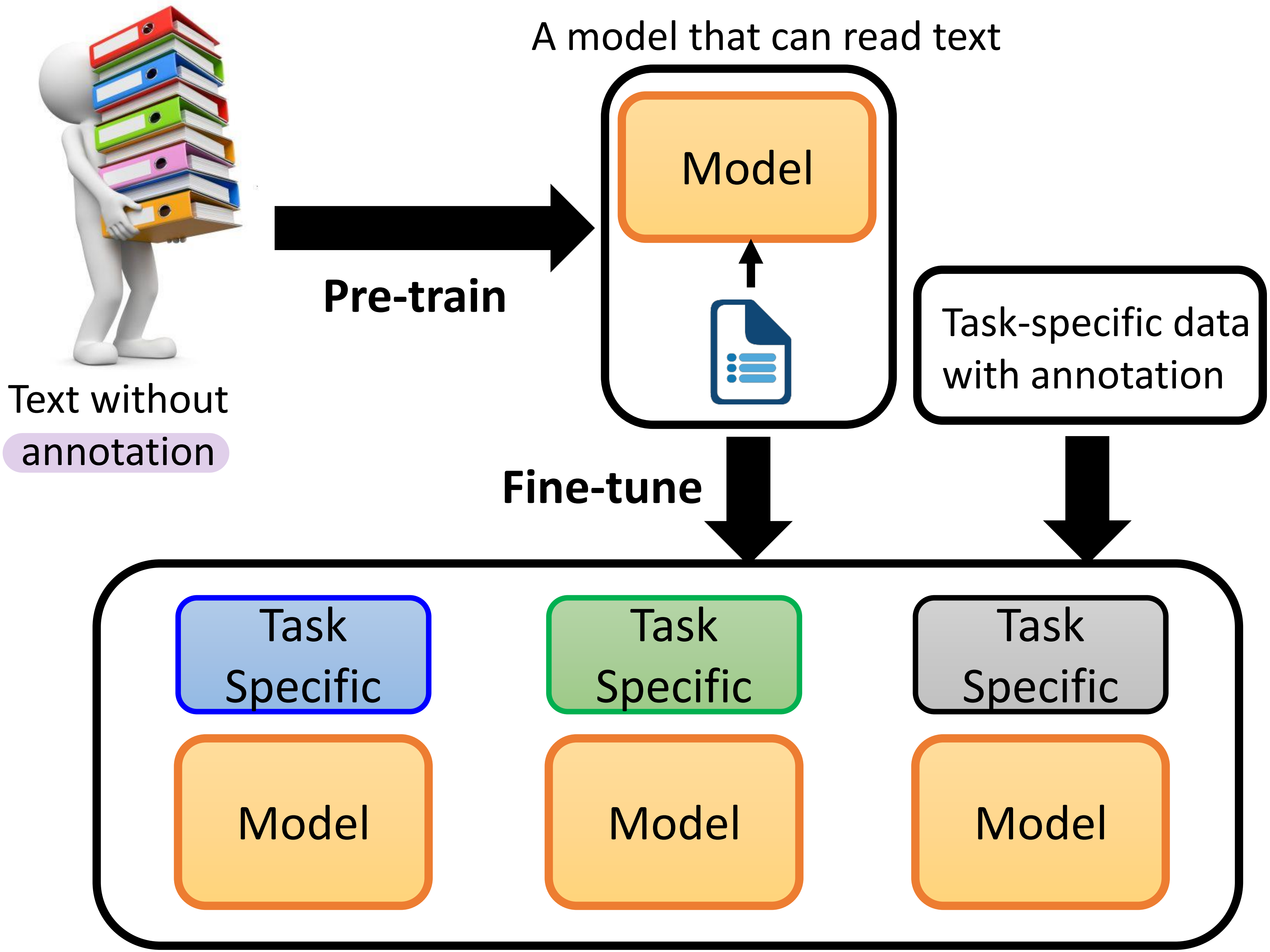
# BERT and its family

Hung-yi Lee 李宏毅



A model that can read text
Model
Text without
annotation
Pre-train
Task-specific data
with annotation
Fine-tune
Task
Specific
Model
Task
Specific
Model
Task
Specific
Model

# Outline

What is pre-train model

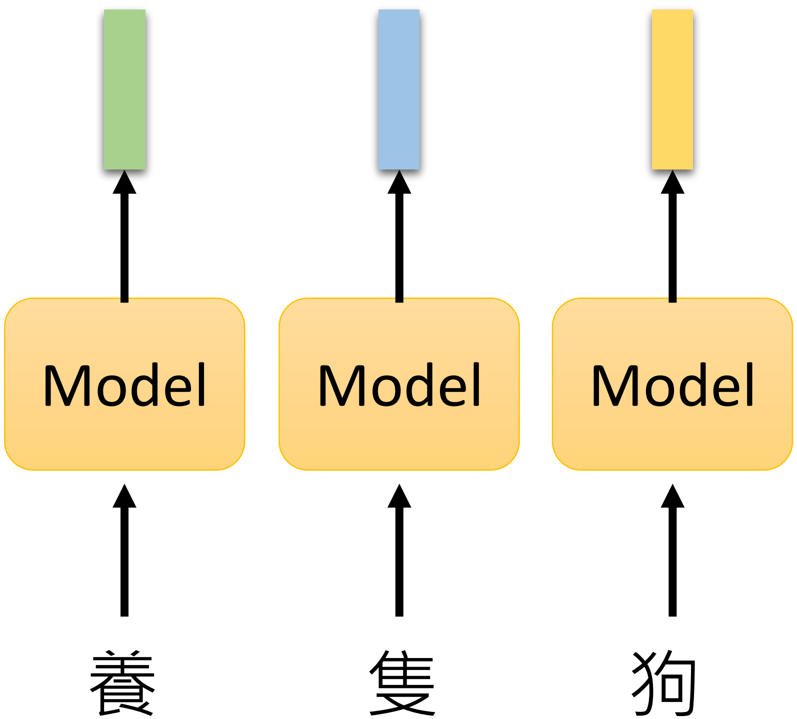
How to fine-tune

How to pre-train

# Pre-train Model

# Pre-train Model

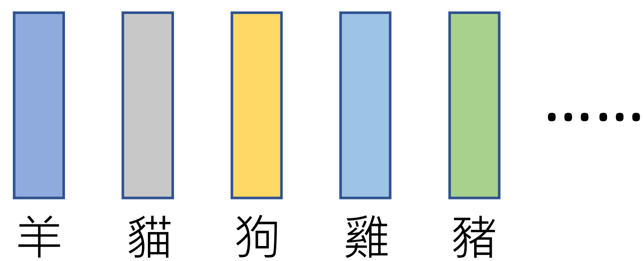
Represent each token by a embedding vector




The token with the same type has the same embedding.

Simply a table look-up

一个token，一个向量，而不考虑上下文信息。

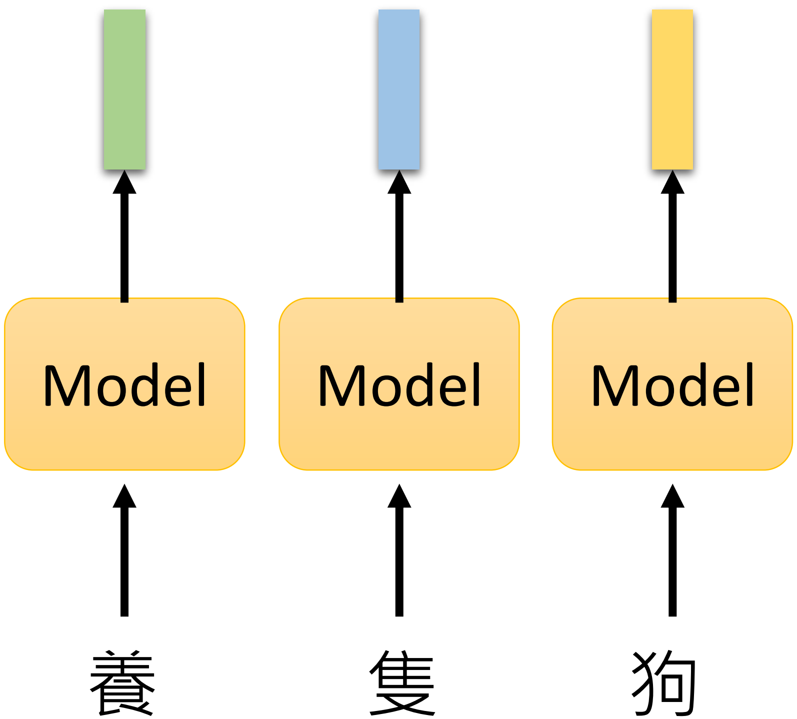



Word2vec [Mikolov, et al., NIPS’13]

Glove [Pennington, et al., EMNLP’14]

# Pre-train Model

Represent each token by a embedding vector

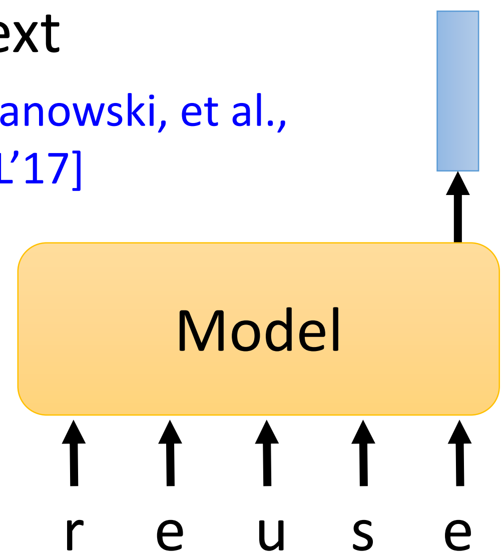



The token with the same type has the same embedding.

English word as token ...

FastText

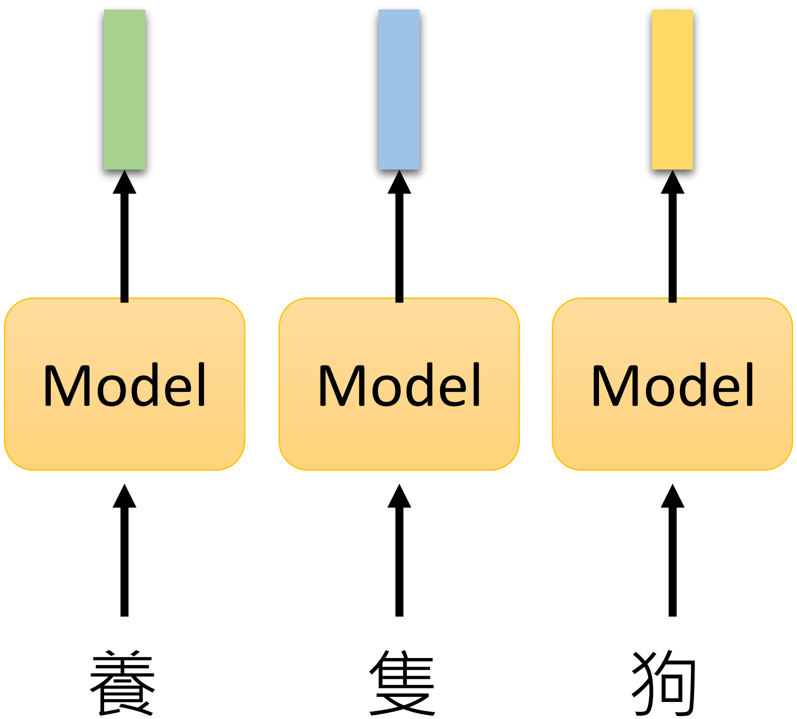
[Bojanowski, et al., TACL’17]




对于英文，输入character，输出向量，根据词根词缀，可以给出未见过的单词的embedding

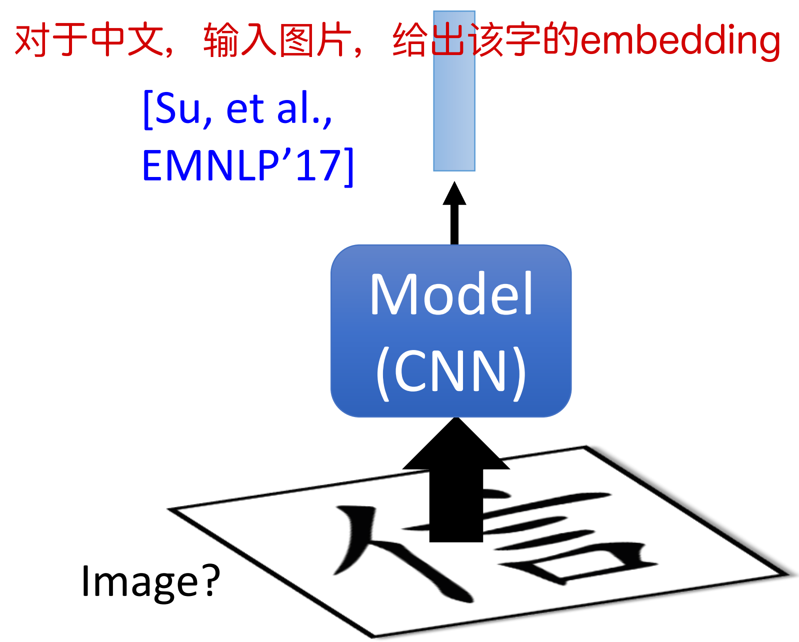
# Pre-train Model

Represent each token by a embedding vector



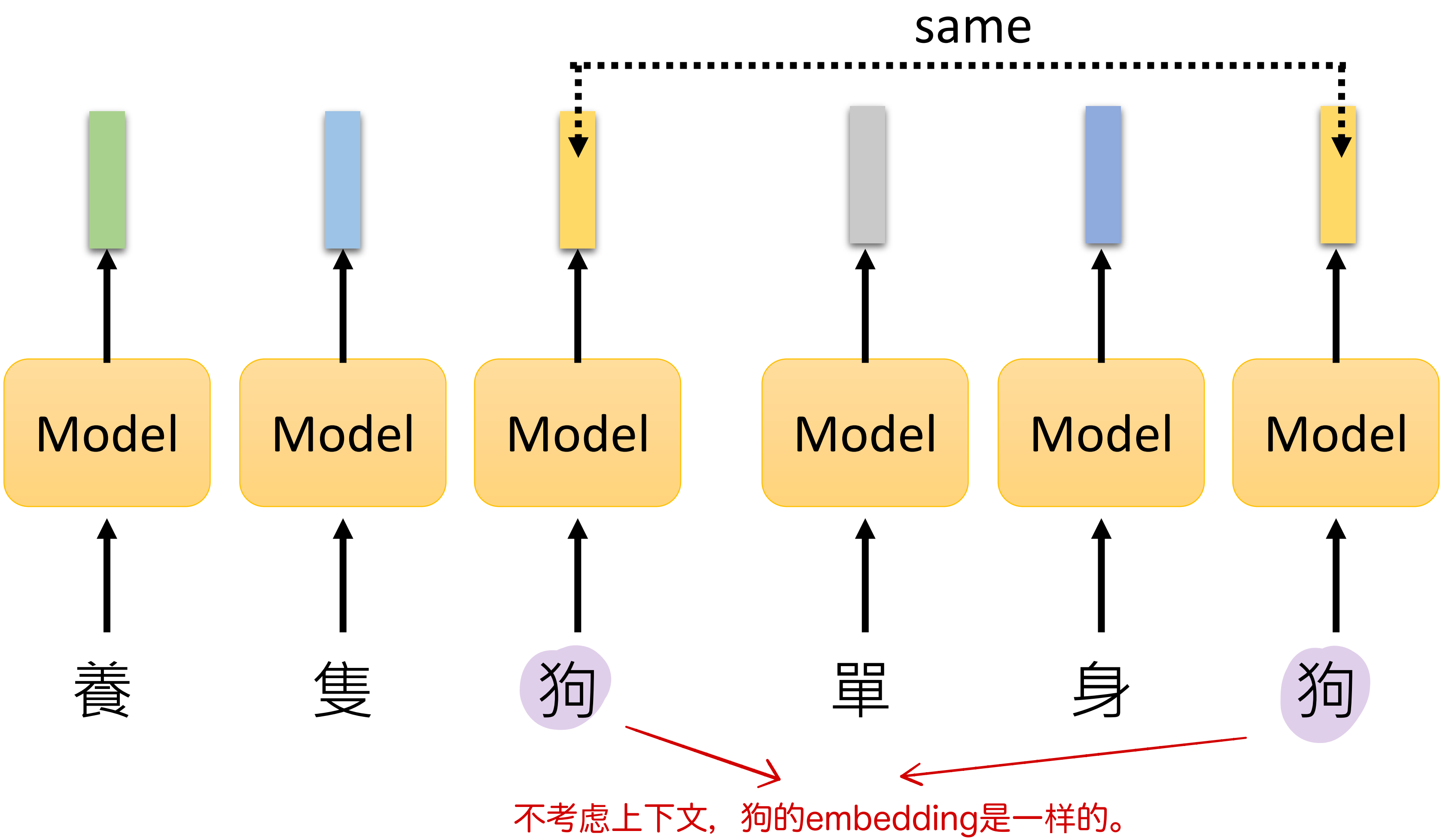

The token with the same type has the same embedding.

Chinese character as token ...



# Pre-train Model

Represent each token by a embedding vector



# Pre-tra

Represent e


Model Model

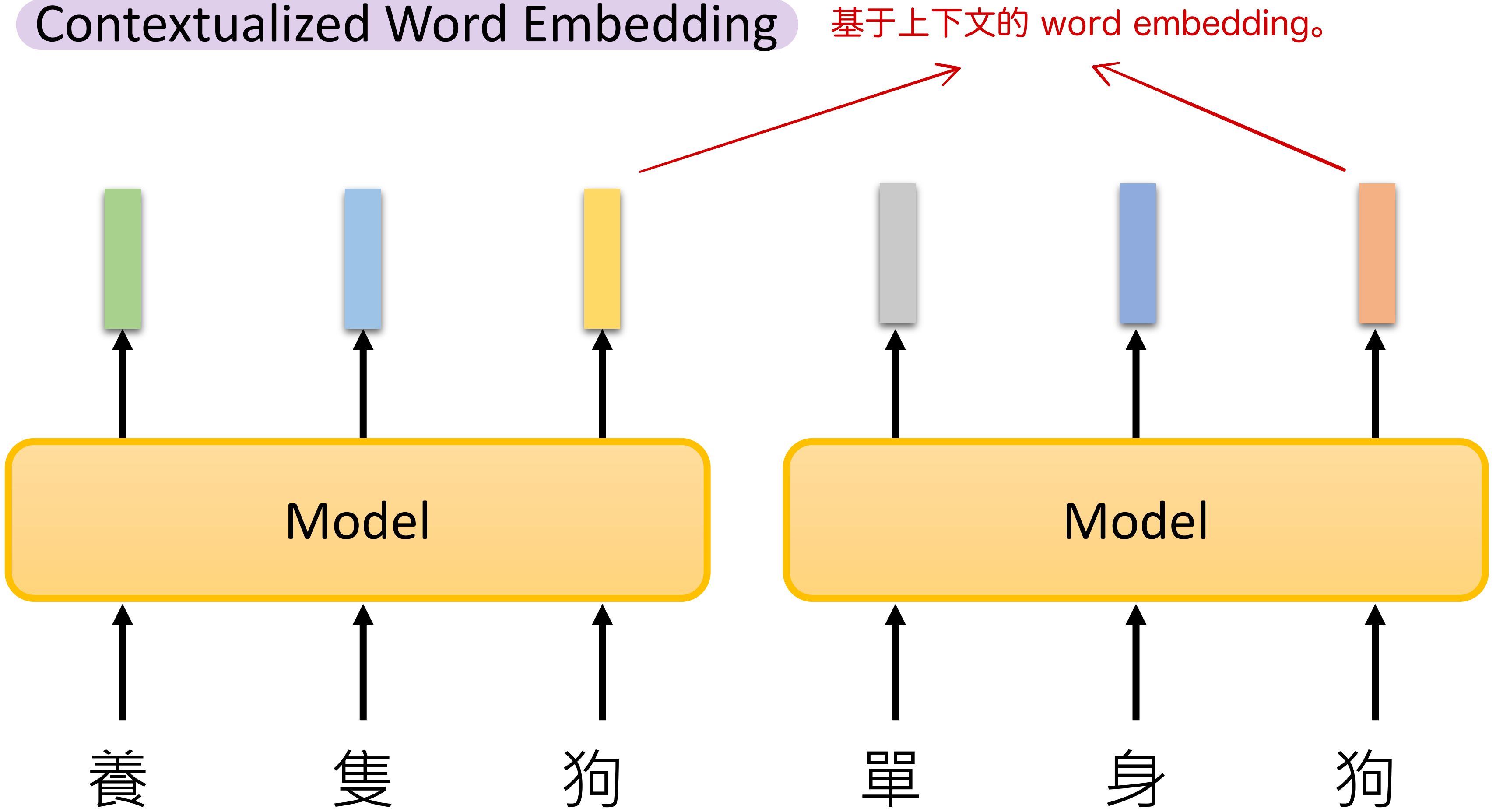
养 隻 狗 單 身 狗

但实际上，二者的含义是不一样的。

different

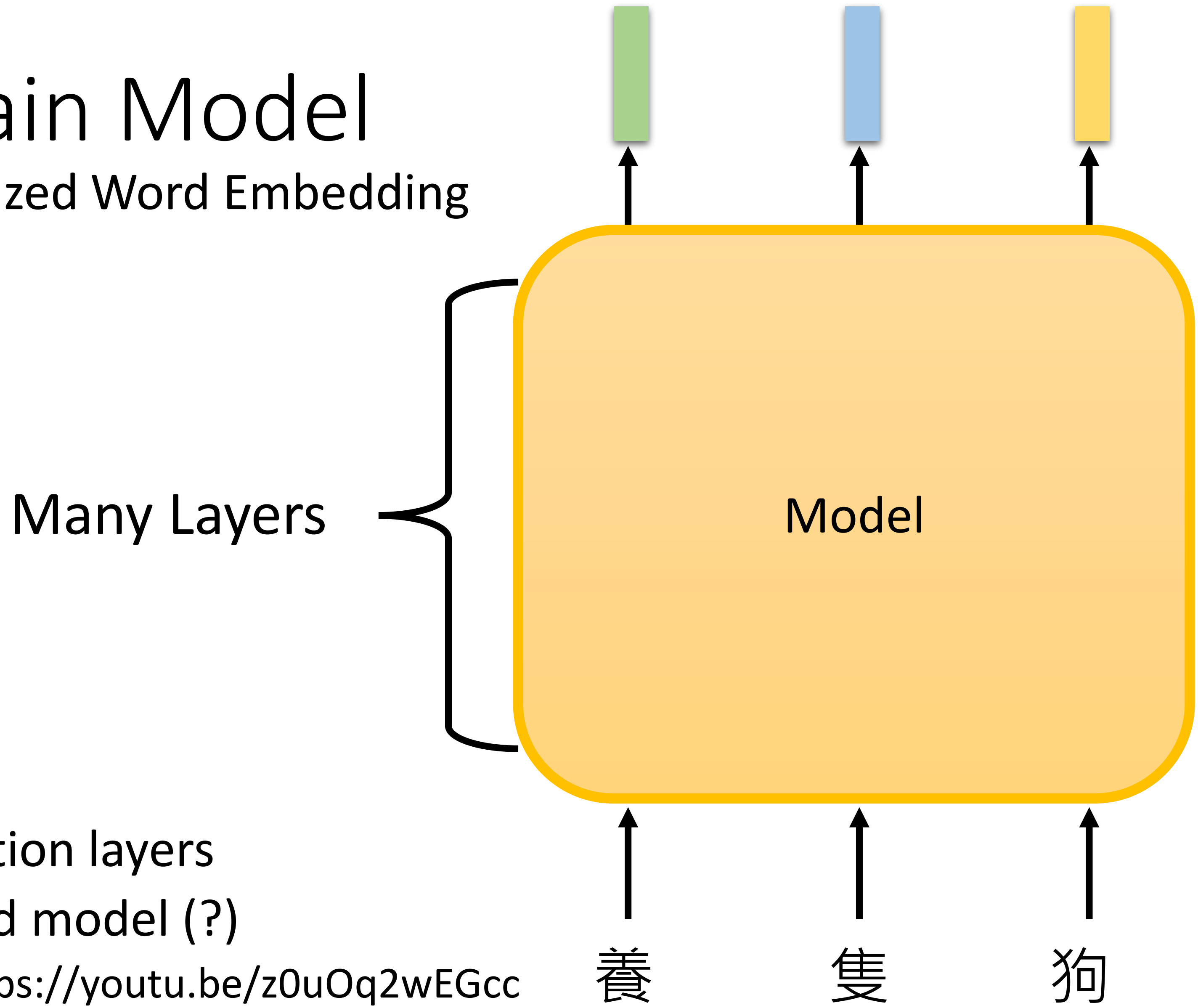
# Pre-train Model

Contextualized Word Embedding



# Pre-train Model

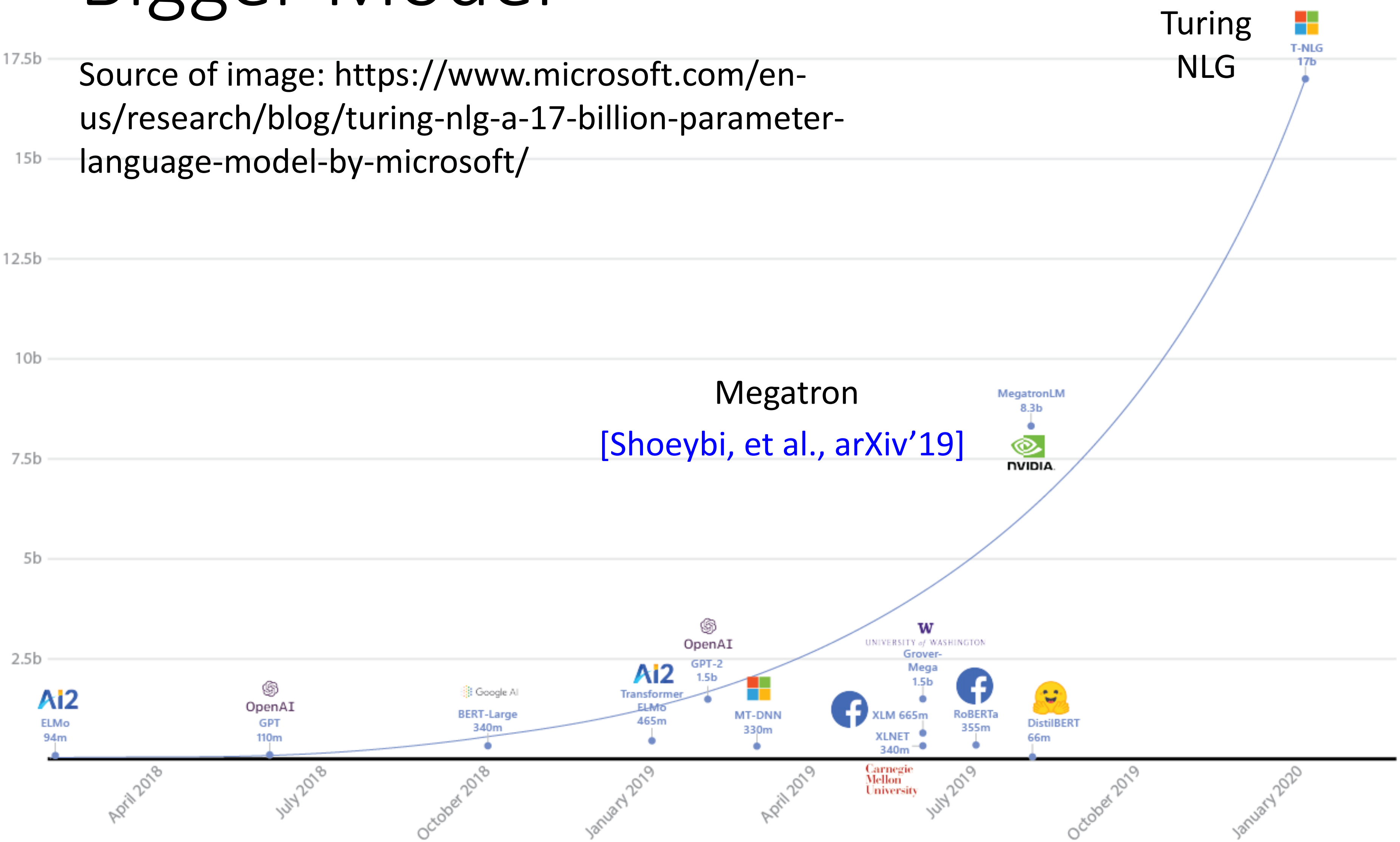
Contextualized Word Embedding



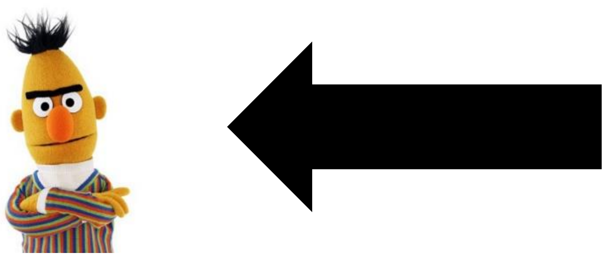

- LSTM
- Self-attention layers
- Tree-based model (?)
    - Ref: https://youtu.be/z0uOq2wEGcc

# Bigger Model 模型越来越大，参数量越来越多

Source of image: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/



# Smaller Model



Distill BERT
[Sanh, et al., NeurIPS workshop'19]

Tiny BERT [Jian, et al., arXiv'19]

Mobile BERT [Sun, et al., ACL'20]

Q8BERT
[Zafrir, et al., NeurIPS workshop 2019]

ALBERT [Lan, et al., ICLR'20]

# Smaller Model

网络压缩

- Network Compression

Ref: https://youtu.be/dPp8rCAnU_A

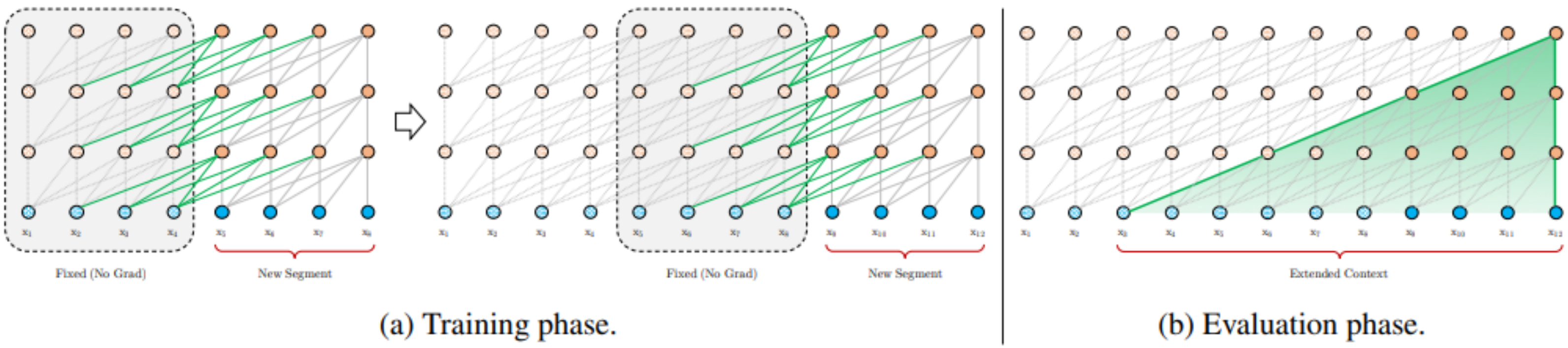
  - Network Pruning
  - Knowledge Distillation
  - Parameter Quantization
  - Architecture Design

All of them have been tried.

Excellent reference: http://mitchgordon.me/machine/learning/2019/11/18/all-the-ways-to-compress-BERT.html

# Network Architecture

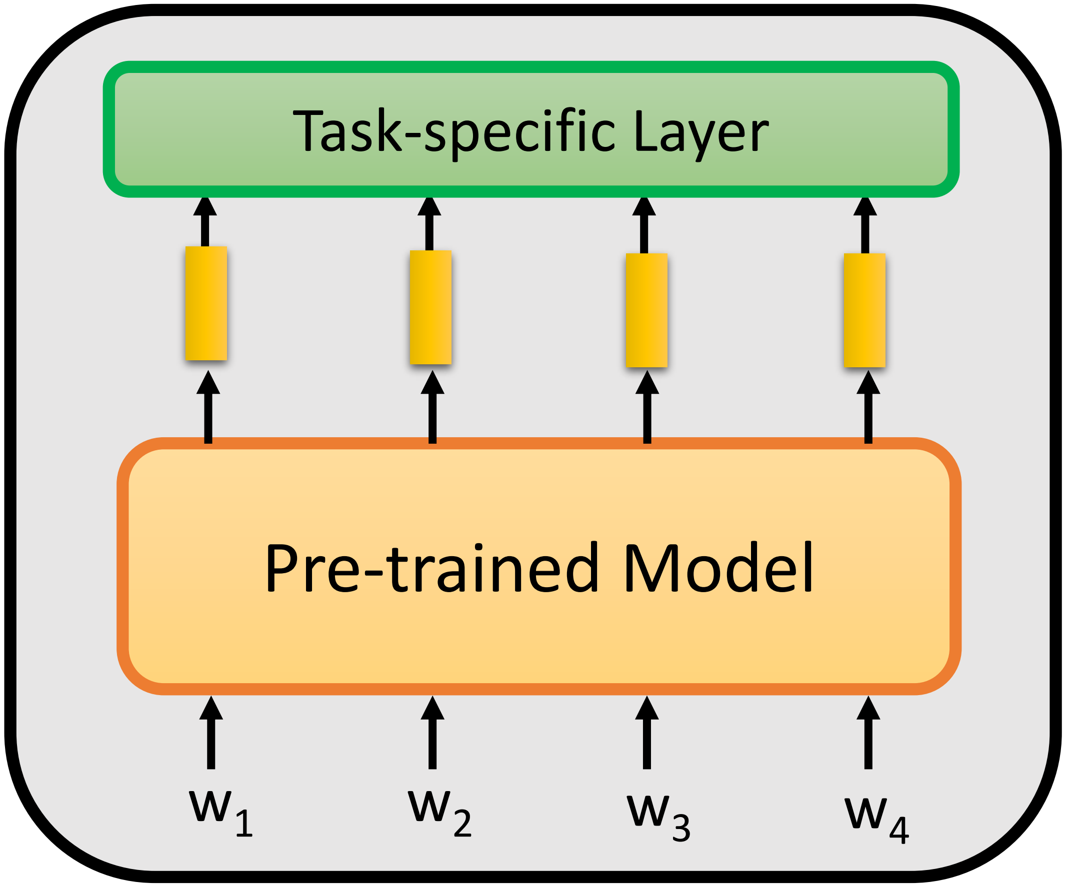
- Transformer-XL: Segment-Level Recurrence with State Reuse [Dai, et al., ACL'19]




- Reformer [Kitaev, et al., ICLR'20]
- Longformer [Beltagy, et al., arXiv'20]

Reduce the complexity of self-attention

# How to fine-tune




For a specific NLP task

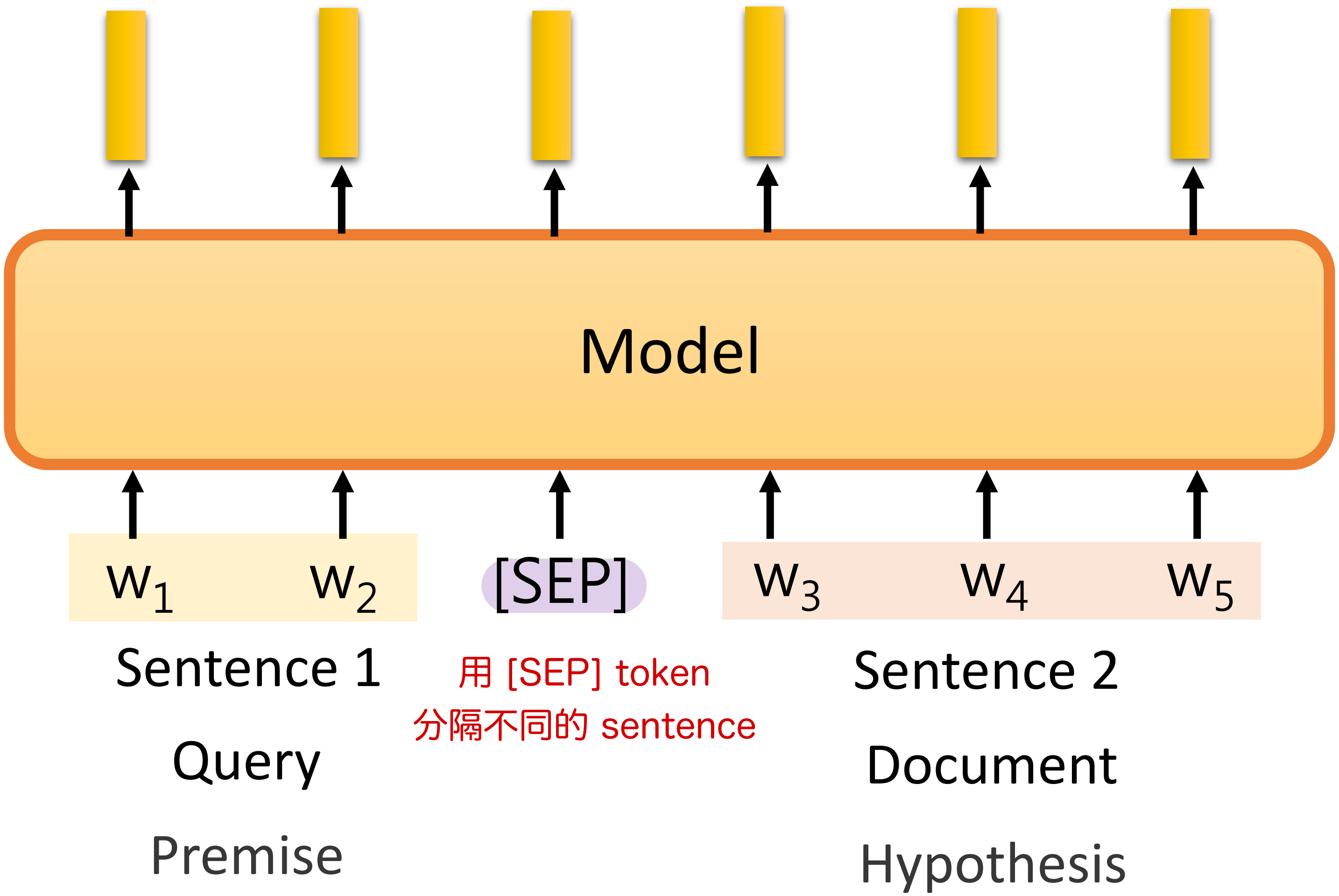
# NLP tasks

Input
- one sentence
- multiple sentences

Output
- one class
- class for each token
- copy from input
- general sequence

# Input

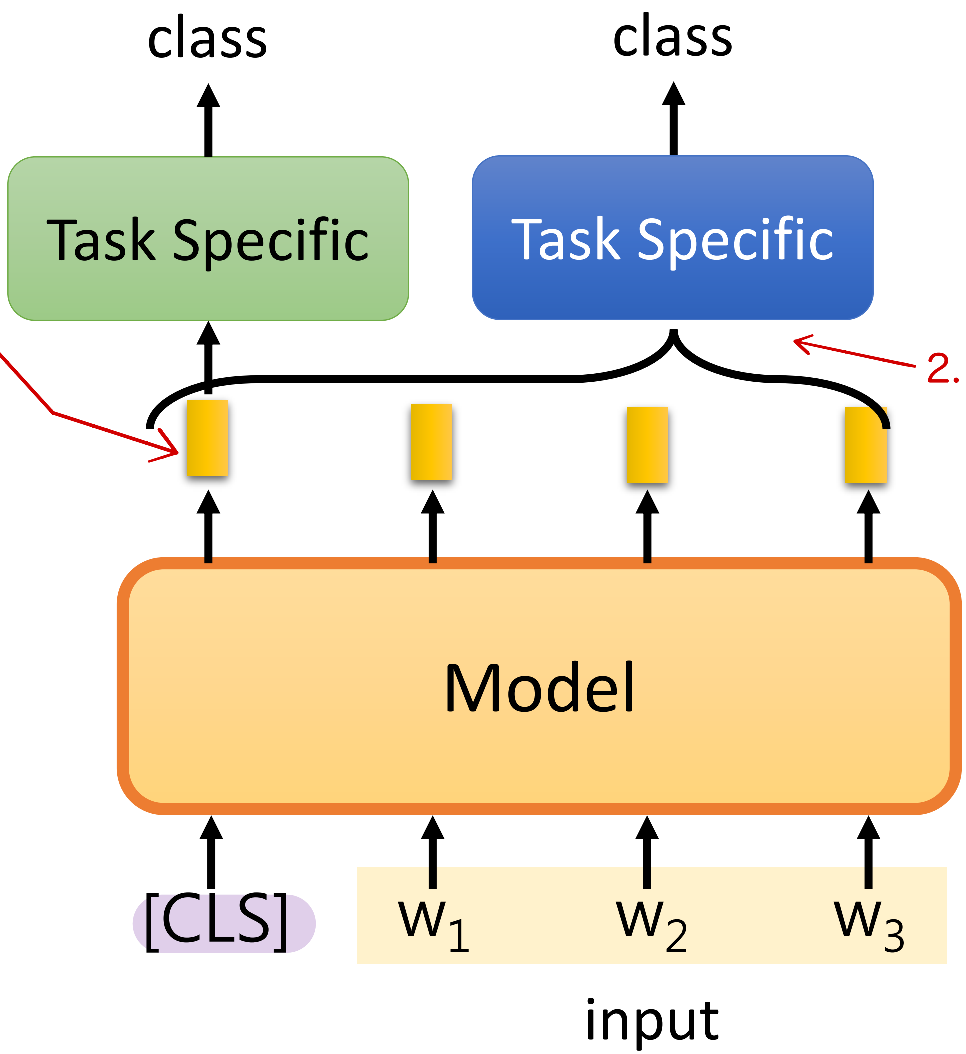
- one sentence
- multiple sentences



1. model 输出一个跟整个句子有关的 embedding，
将这个 embedding 输入到 task specific，得到输出类别。

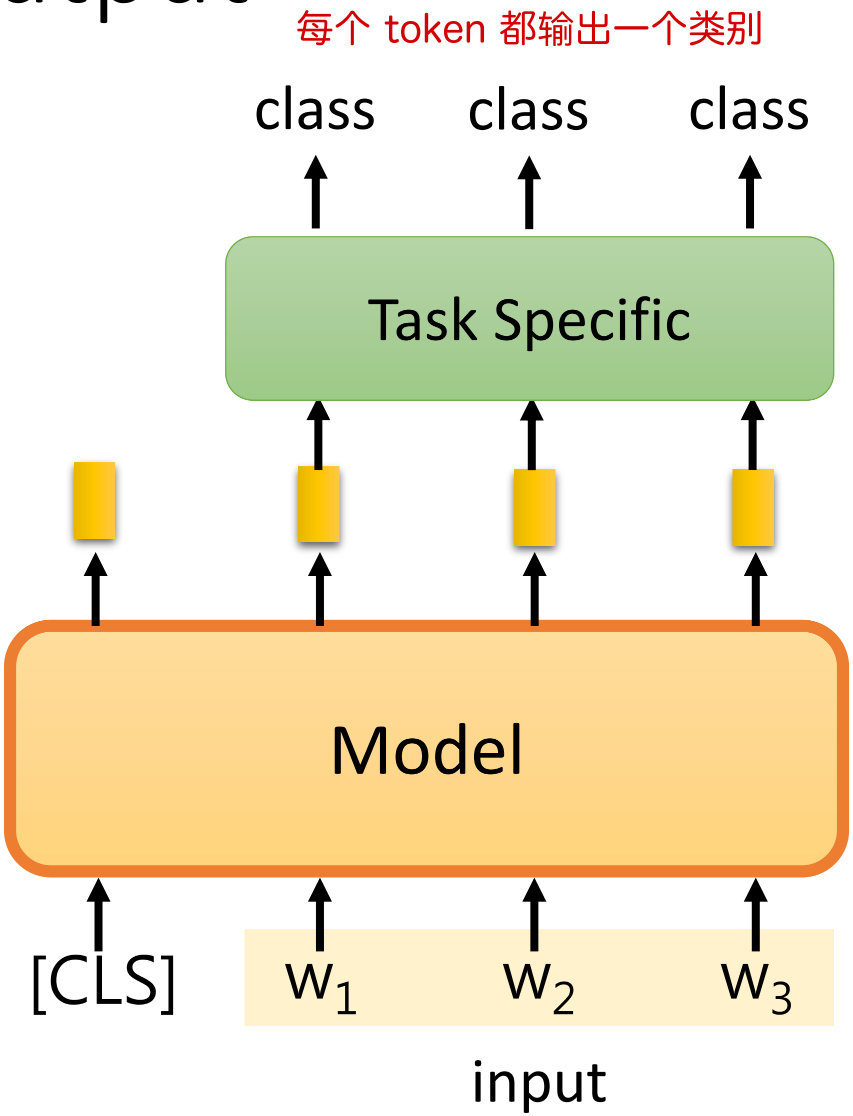
# Output

- $\boxed{\text{one class}}$ 输出单个类别
- class for each token
- copy from input
- general sequence




2. model 得到所有 token 的 embedding，
将其全部输入到 task specific 中，
输出类别。

# Output




- one class
- class for each token
- copy from input
- general sequence

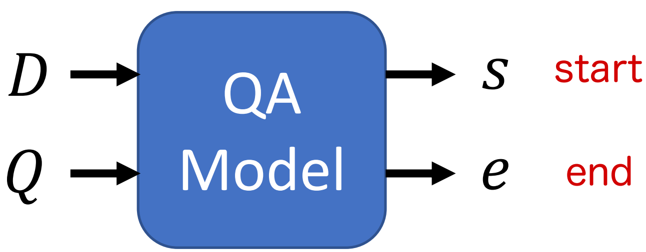
# Output

- one class
- class for each token
- copy from input
- general sequence

- Extraction-based 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***: $A = \{d_s, \cdots, d_e\}$

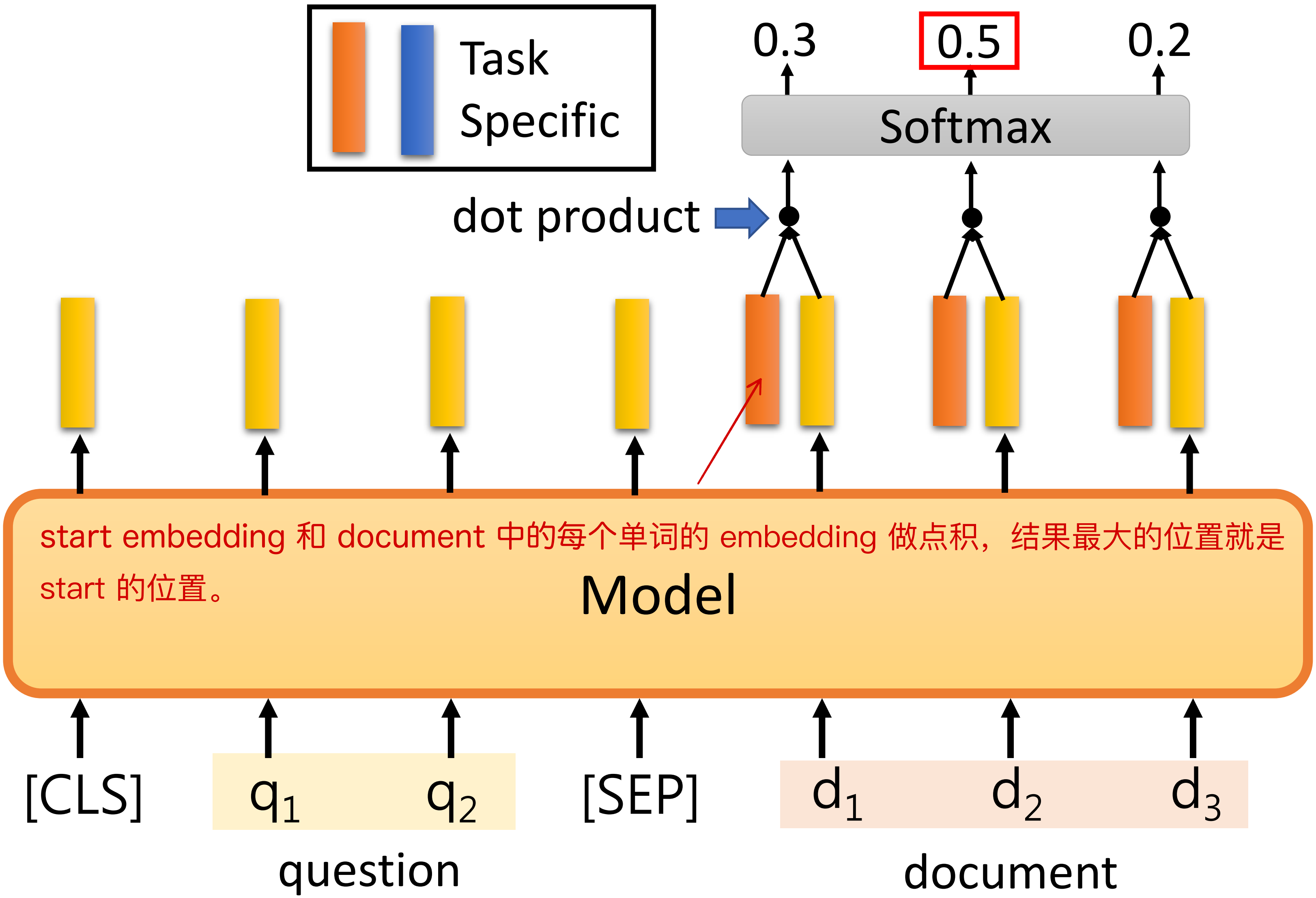
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Preci 77 on 79 as smaller droplets coalesce via ion other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

Where do water droplets collide with ice crystals to form precipitation?
within a cloud $s = 77, e = 79$

# Copy from Input (BERT)（举例）

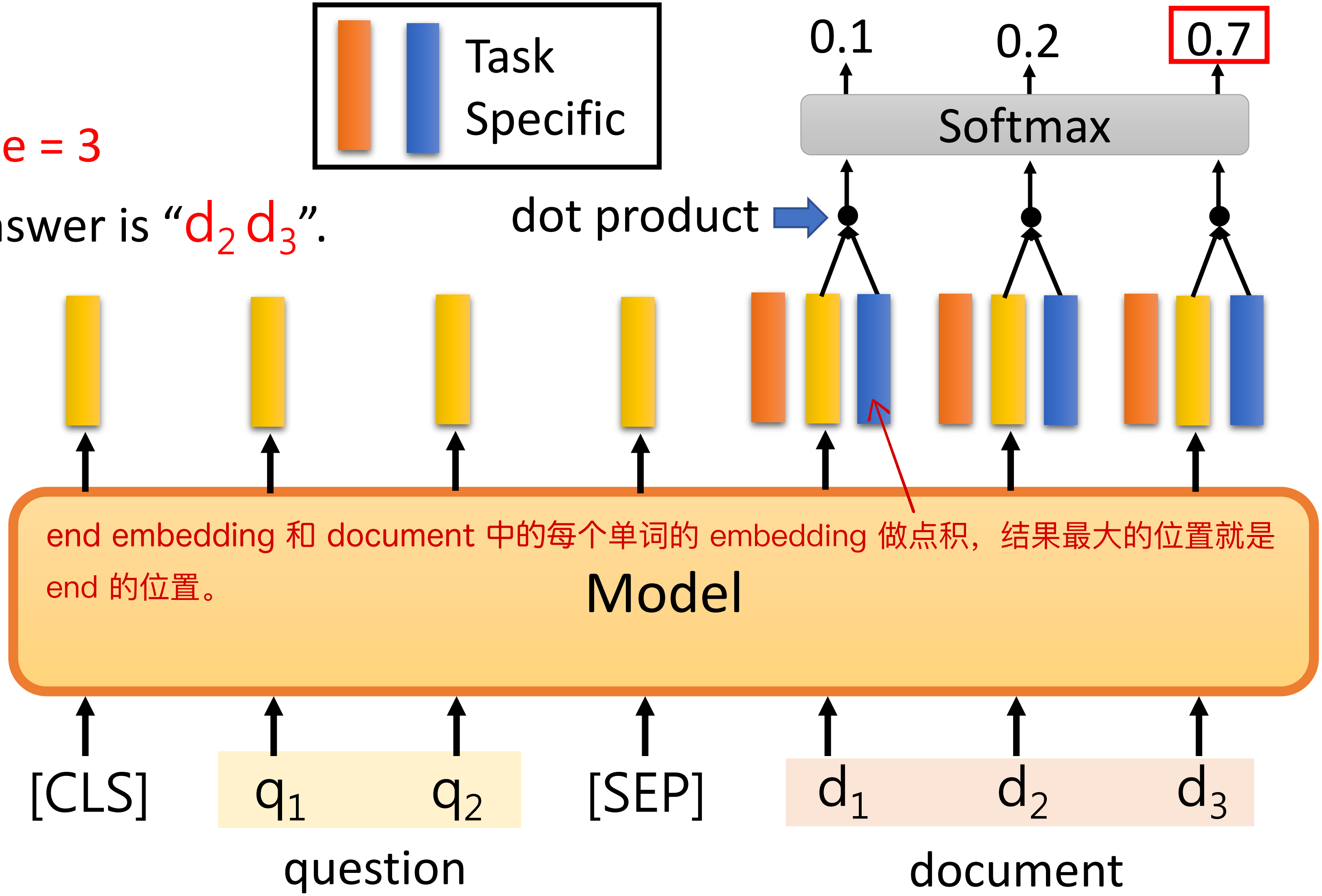
s = 2



# Copy from Input (BERT)

s = 2   e = 3

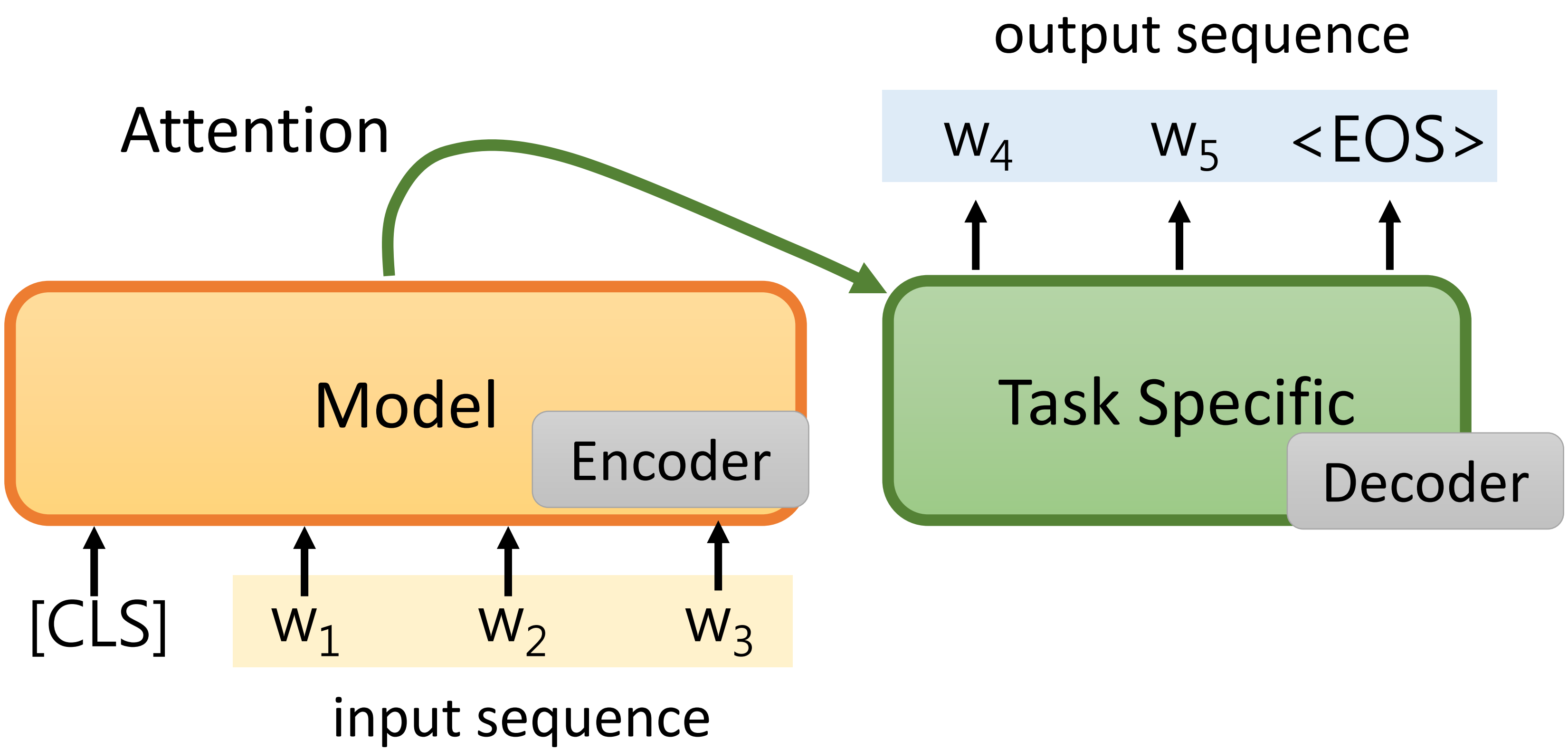
The answer is “$d_2$ $d_3$”.



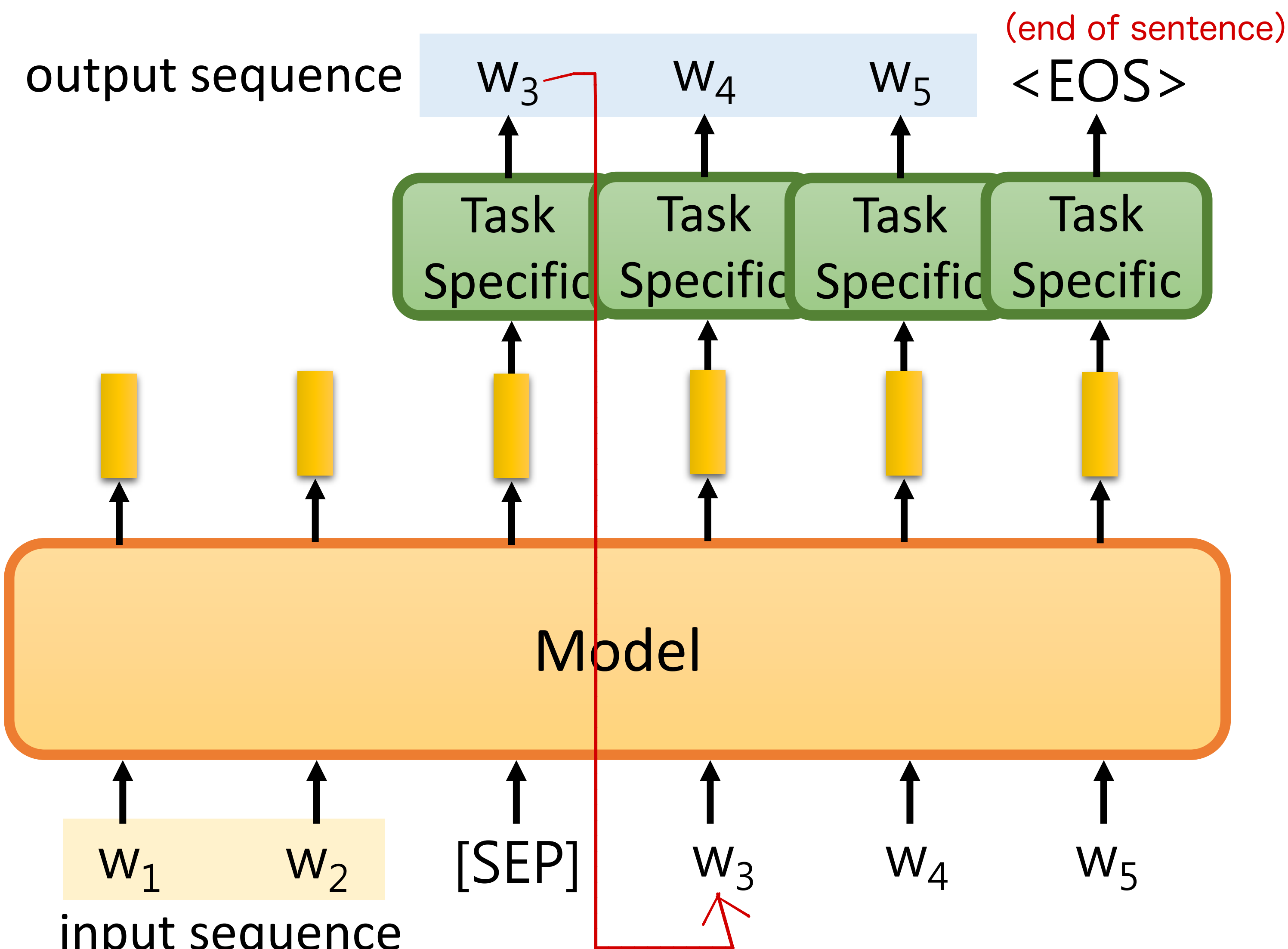
# Output – General Sequence (v1)

输入一个 sequence，输出一个sequence

- Seq2seq model

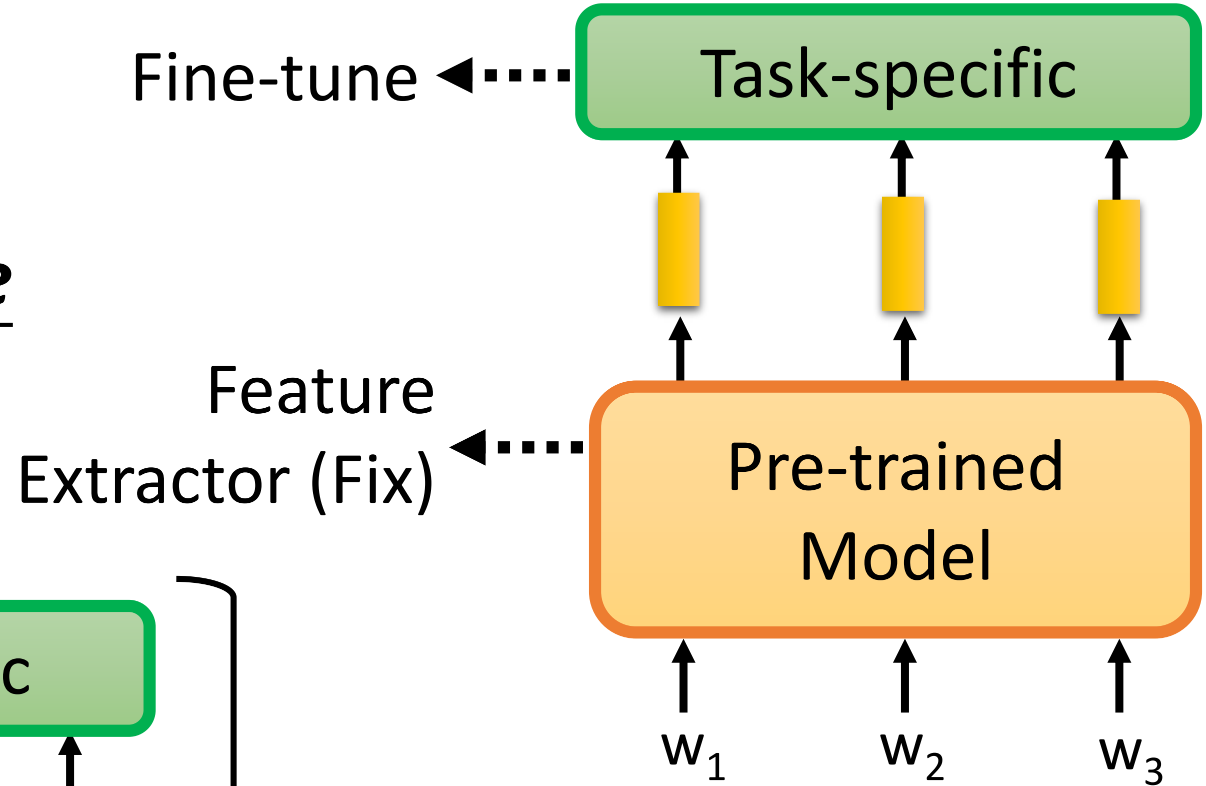


# Output – General Sequence (v2)

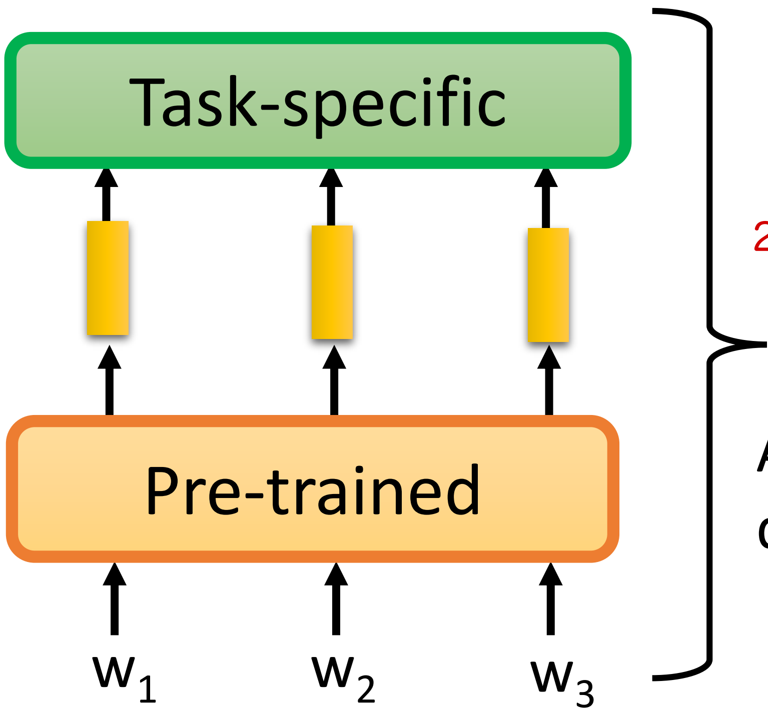


# ***How to fine-tune***

1. pre-train model 固定，只对 task–specific 进行 fine–tune。




2. 将 pre-train model 和 task-specific 串联起来一起进行 fine-tune。（效果更好）

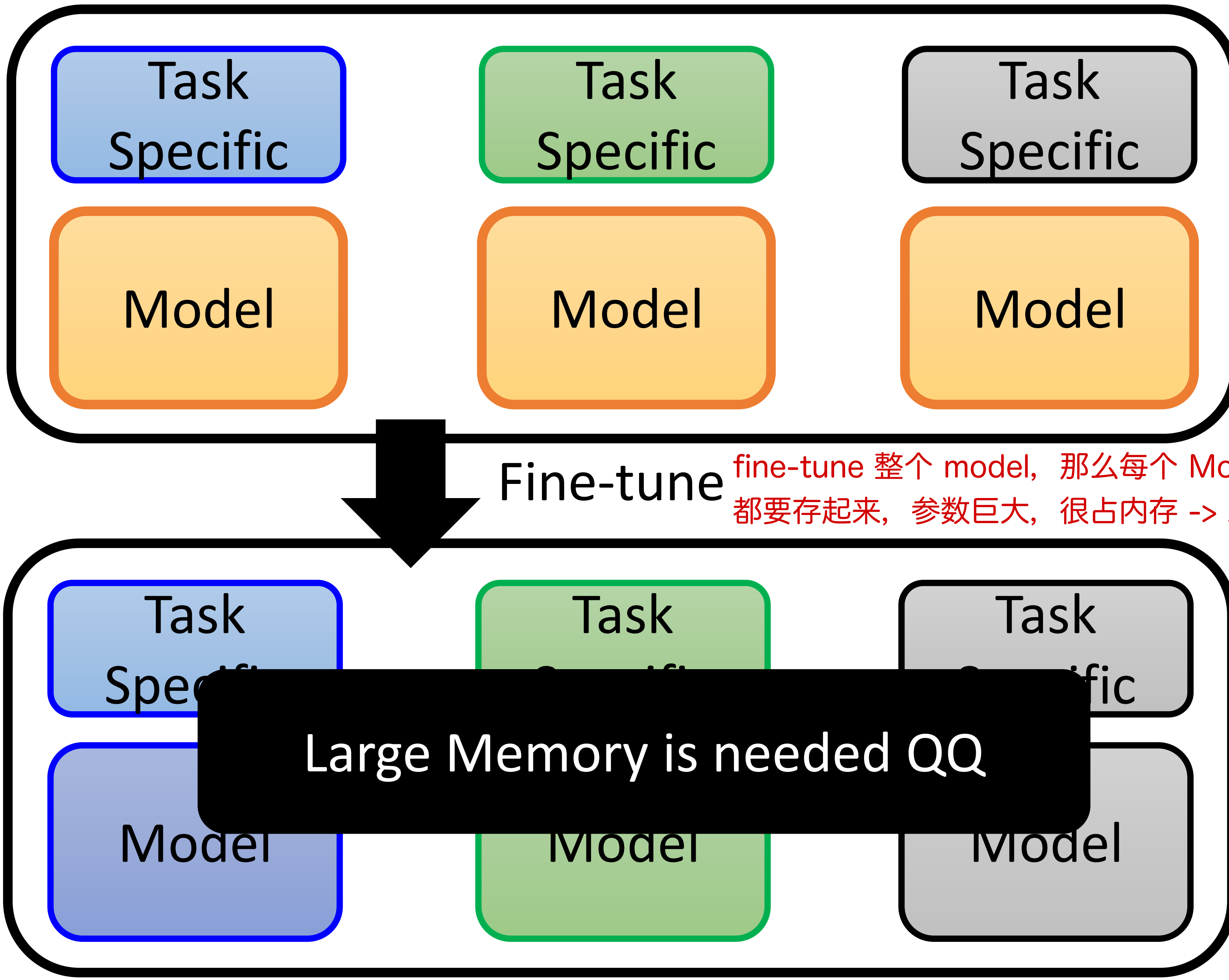



Fine-tune

A gigantic model for down-stream tasks

## ***Adaptor***

[Stickland, et al., ICML’19] [Houlsby, et al., ICML’19]

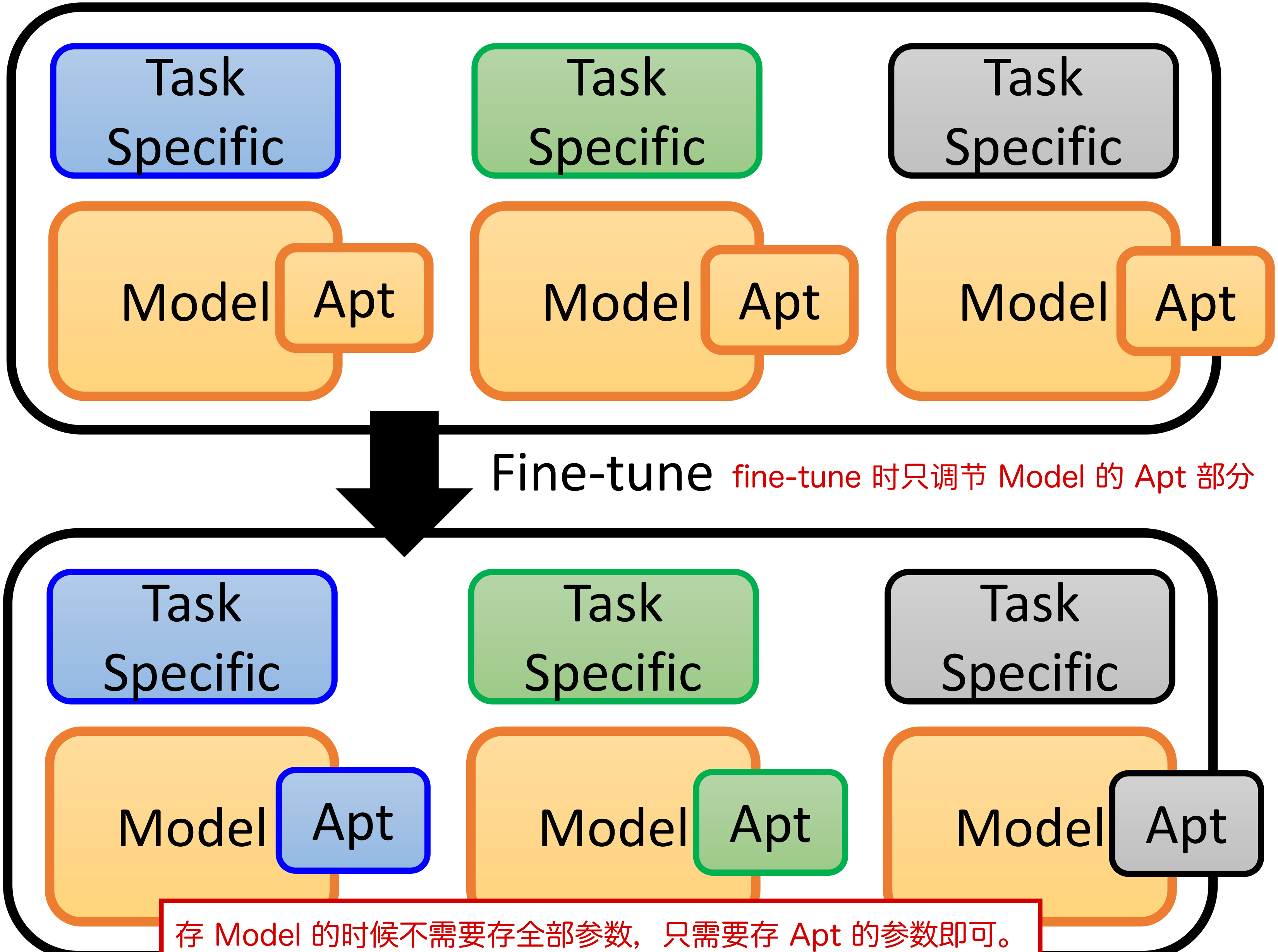



fine-tune 整个 model，那么每个 Model 最后都要存起来，参数巨大，很占内存 -> Adaptor

## *Adaptor*

只调 pre-train model 的一部分 -> 在 pre-train model 中加入一些 layer（Apt）

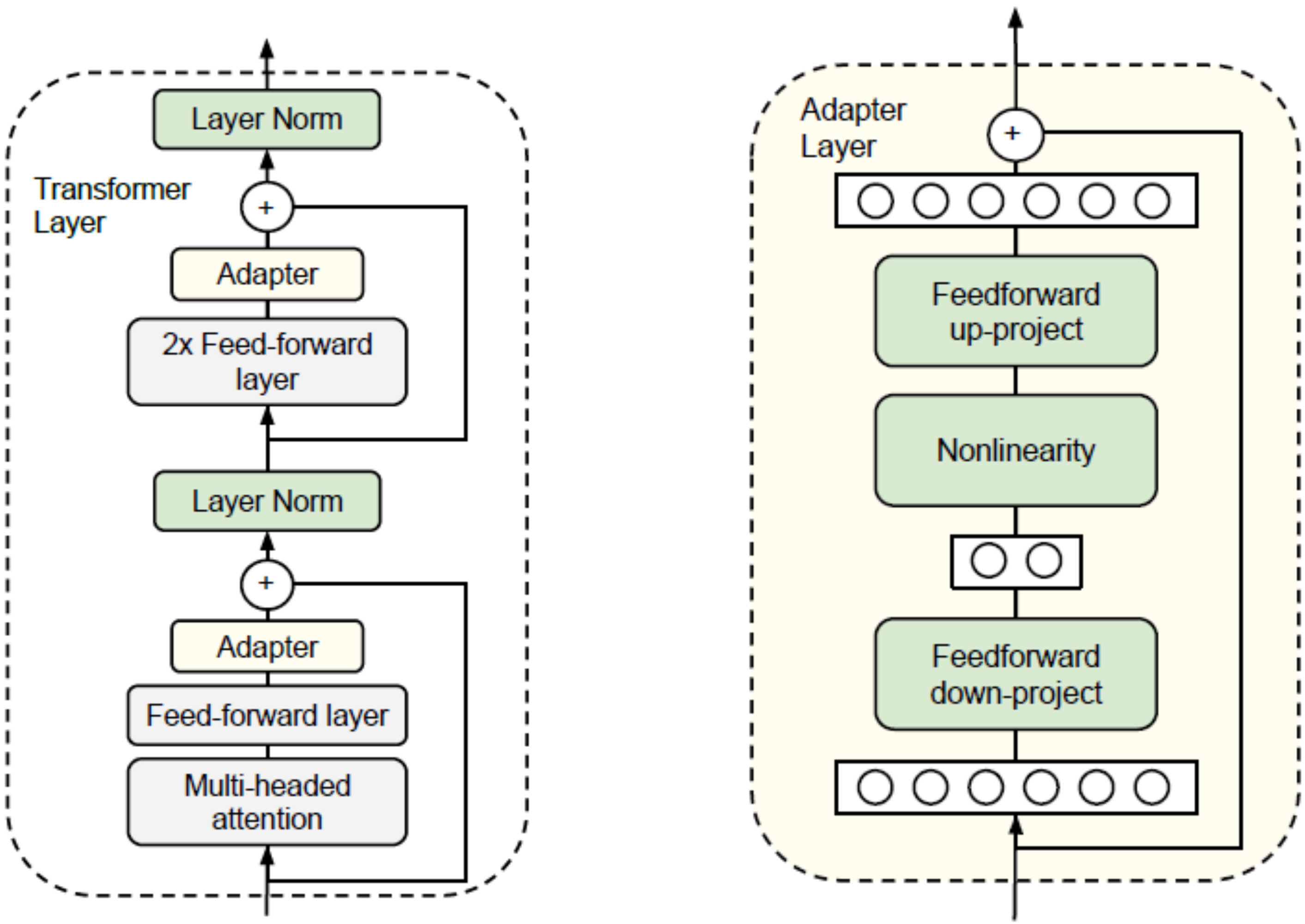
[Stickland, et al., ICML'19] [Houlsby, et al., ICML'19]




fine-tune 时只调节 Model 的 Apt 部分

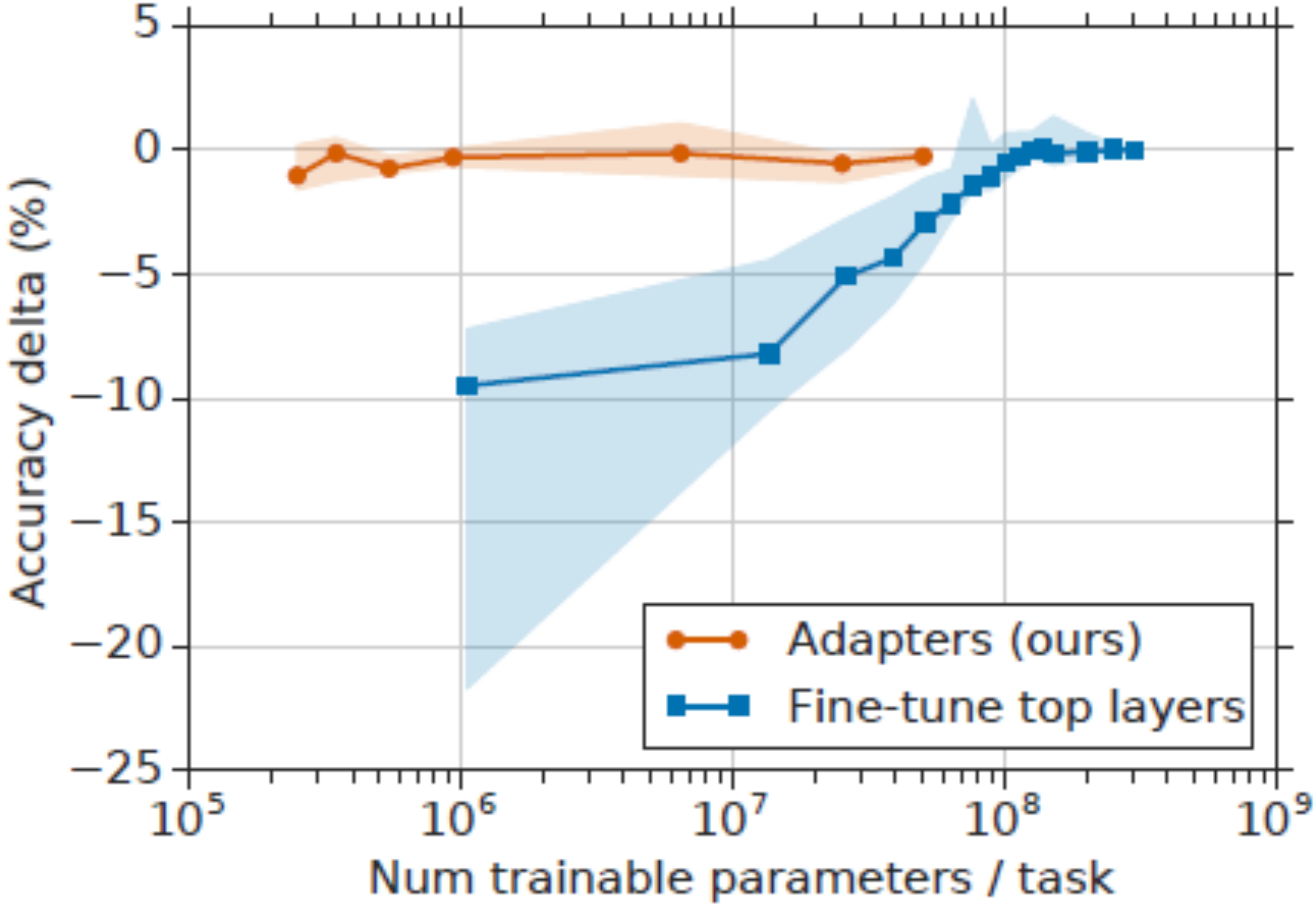
存 Model 的时候不需要存全部参数，只需要存 Apt 的参数即可。

现有的一些 Adaptor 的结构




Source of image: https://arxiv.org/abs/1902.00751

[Houlsby, et al., ICML'19]

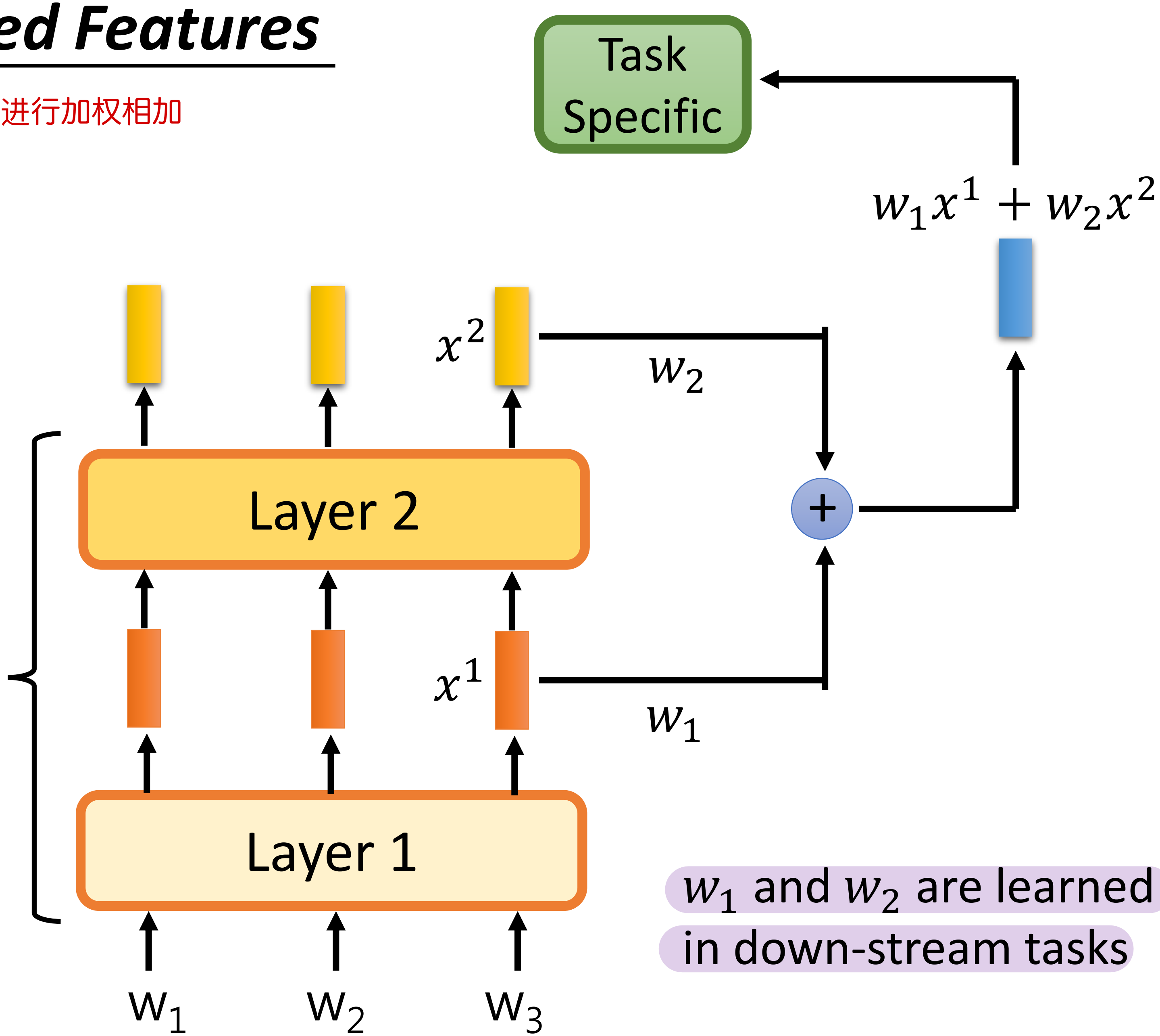



Source of image: https://arxiv.org/abs/1902.00751

[Houlsby, et al., ICML'19]

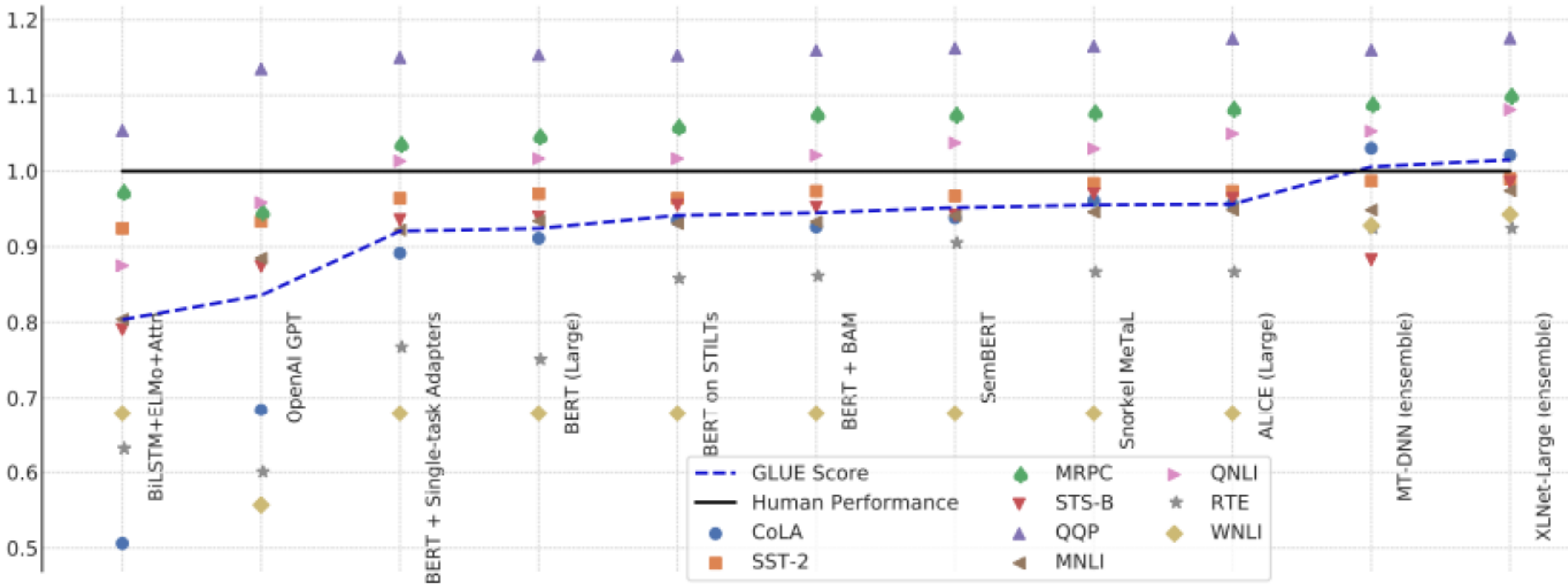
# ***Weighted Features***

对特征进行加权相加



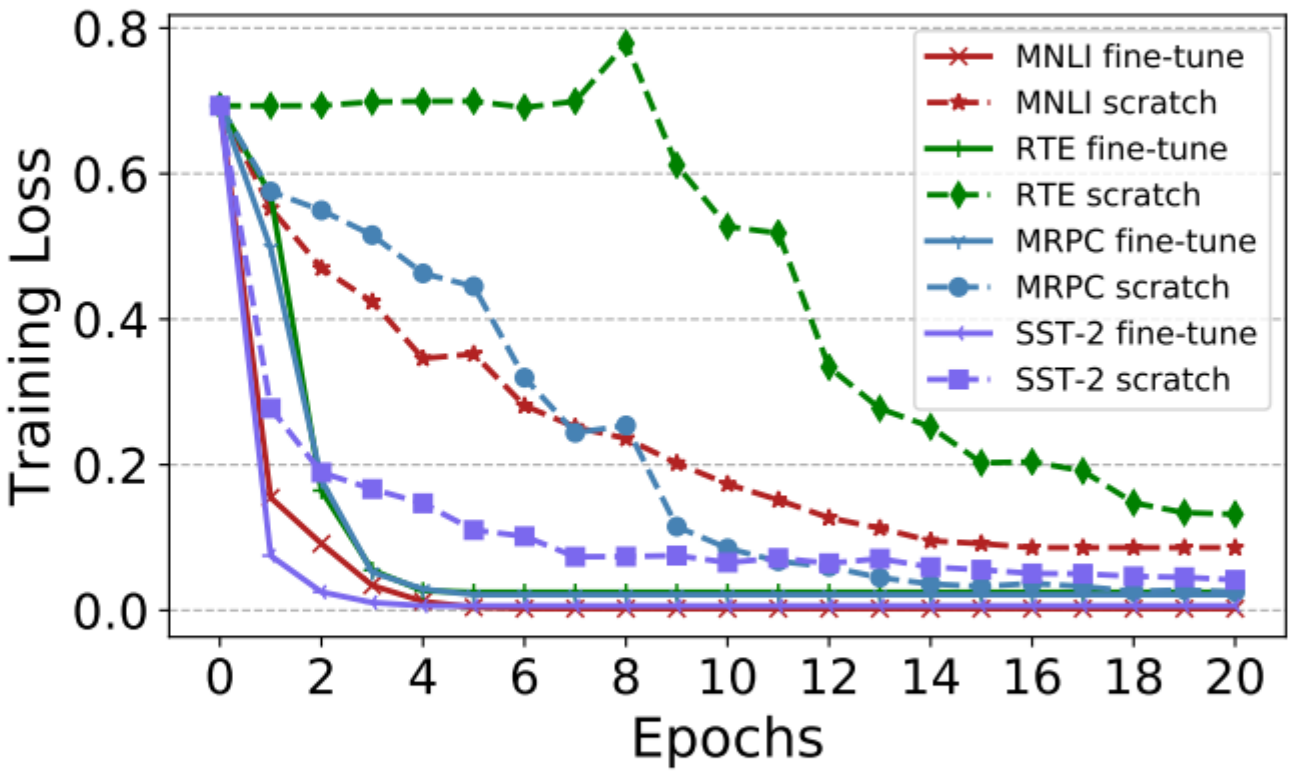
# Why Pre-train Models?

- GLUE scores 检测一个模型 in general 了解人类语言的能力




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

# Why Fine-tune?

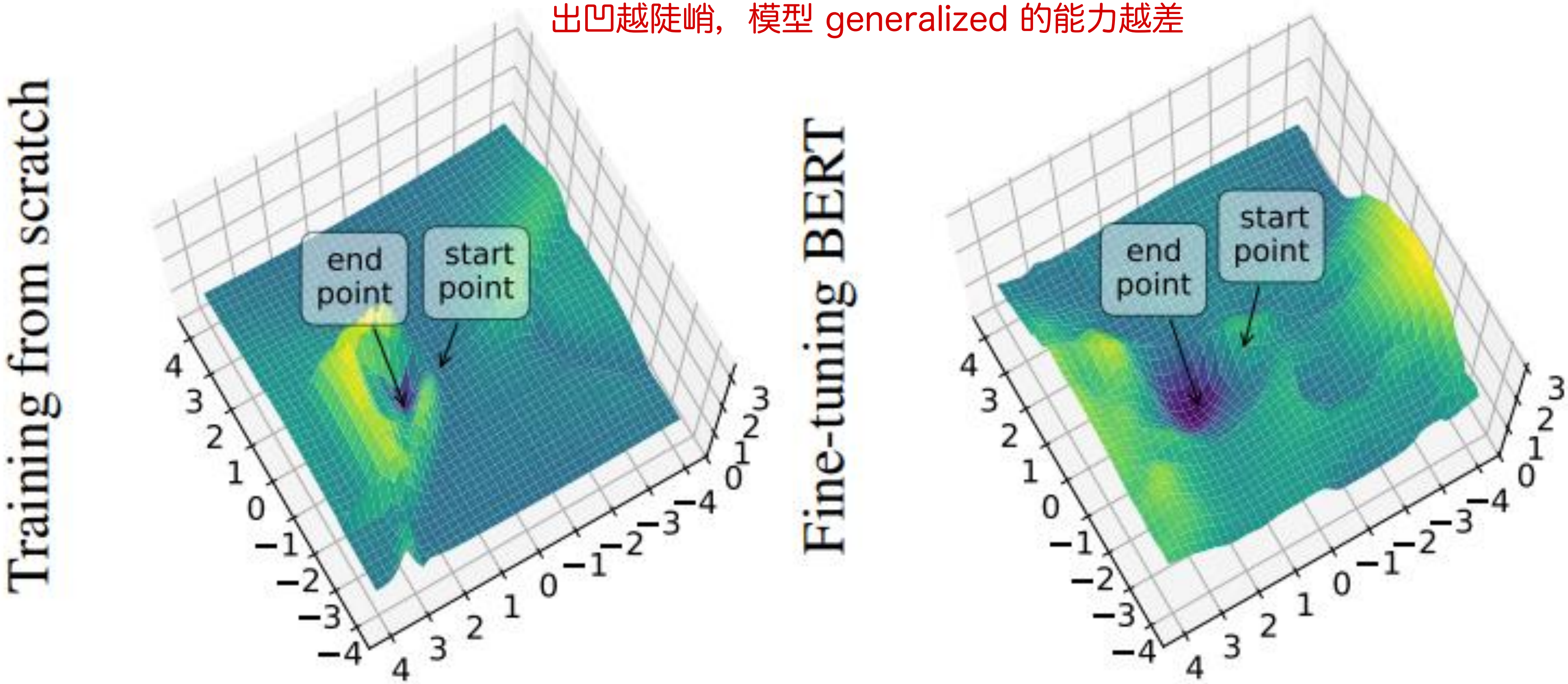



[Hao, et al., EMNLP'19] Source of image: https://arxiv.org/abs/1908.05620

# Why Fine-tune?

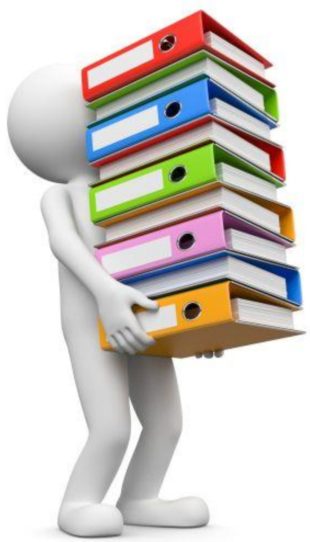
How to generate the figures below?
https://youtu.be/XysGHdNOTbg

出凹越陡峭，模型 generalized 的能力越差

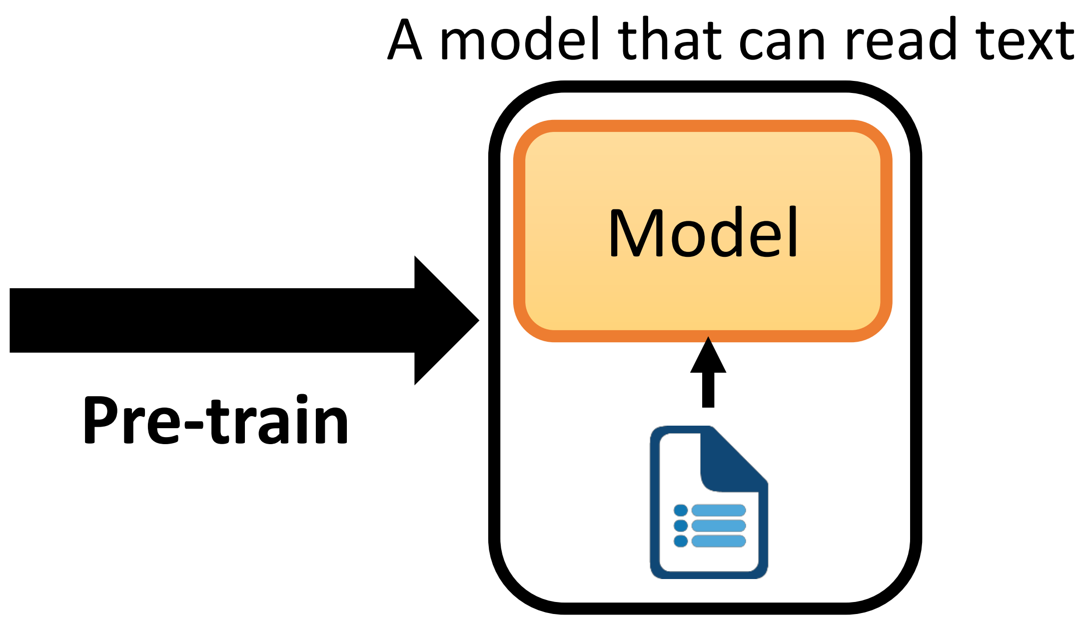



[Hao, et al., EMNLP’19] Source of image: https://arxiv.org/abs/1908.05620

# How to Pre-train



Text without annotation

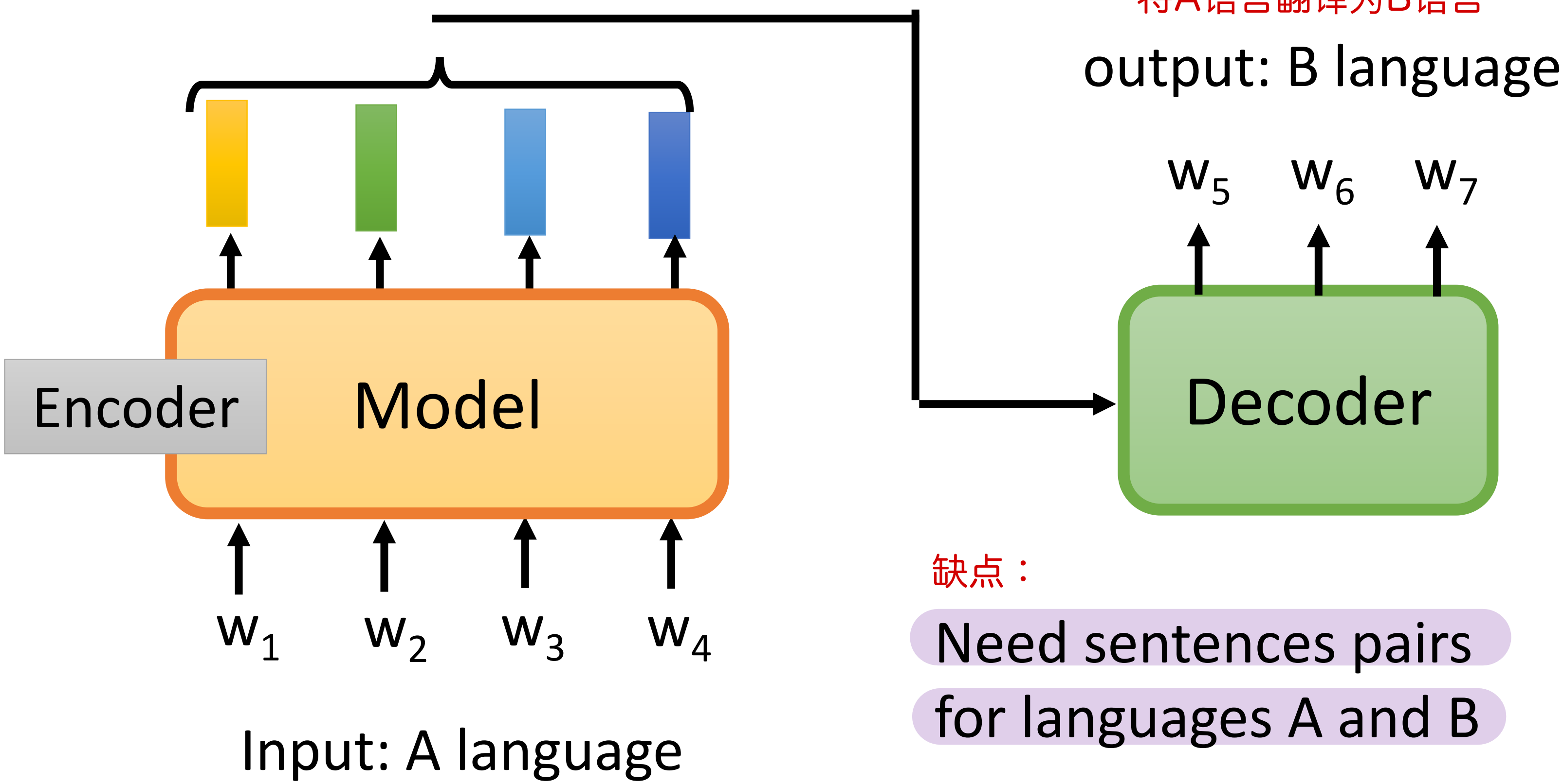


通过翻译来进行pre-train：优点：可以考虑到每个token的意义

# Pre-training by Translation

- Context Vector (CoVe) 可以考虑上下文信息

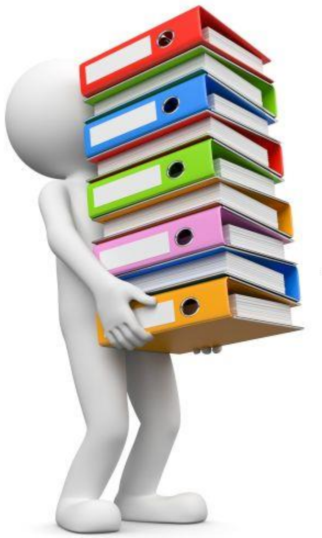
将A语言翻译为B语言




缺点：

Need sentences pairs for languages A and B

# ***Self-supervised Learning*** 自监督学习

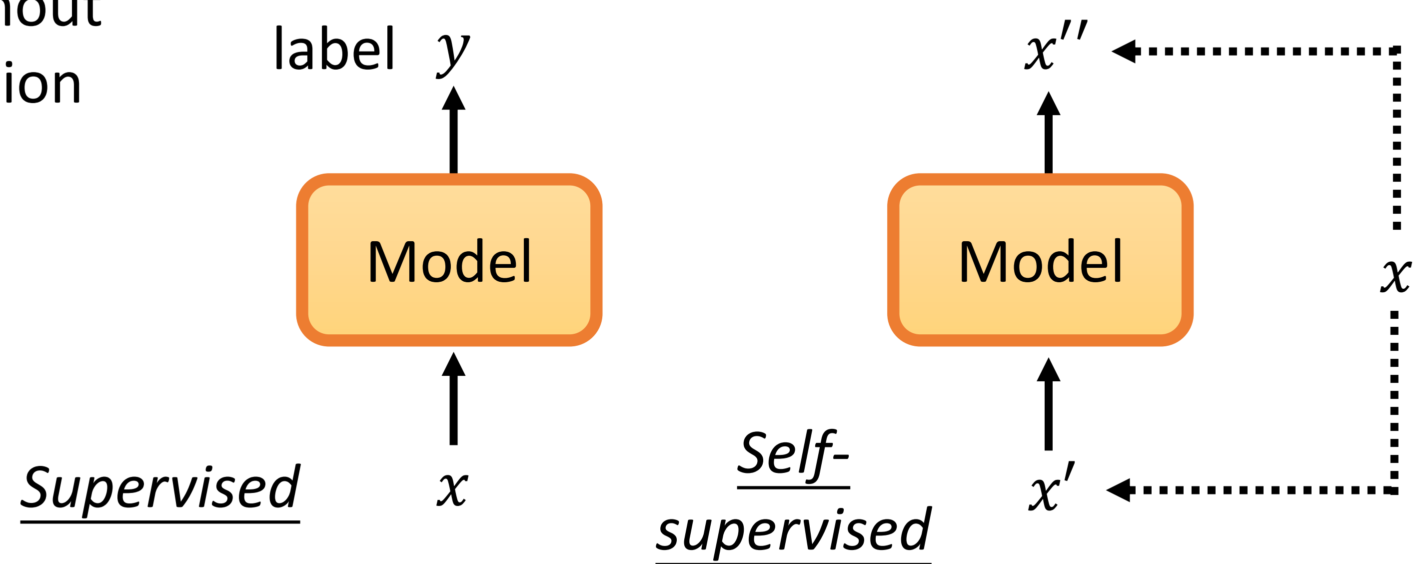


Text without annotation

**Yann LeCun**
2019年4月30日 ·

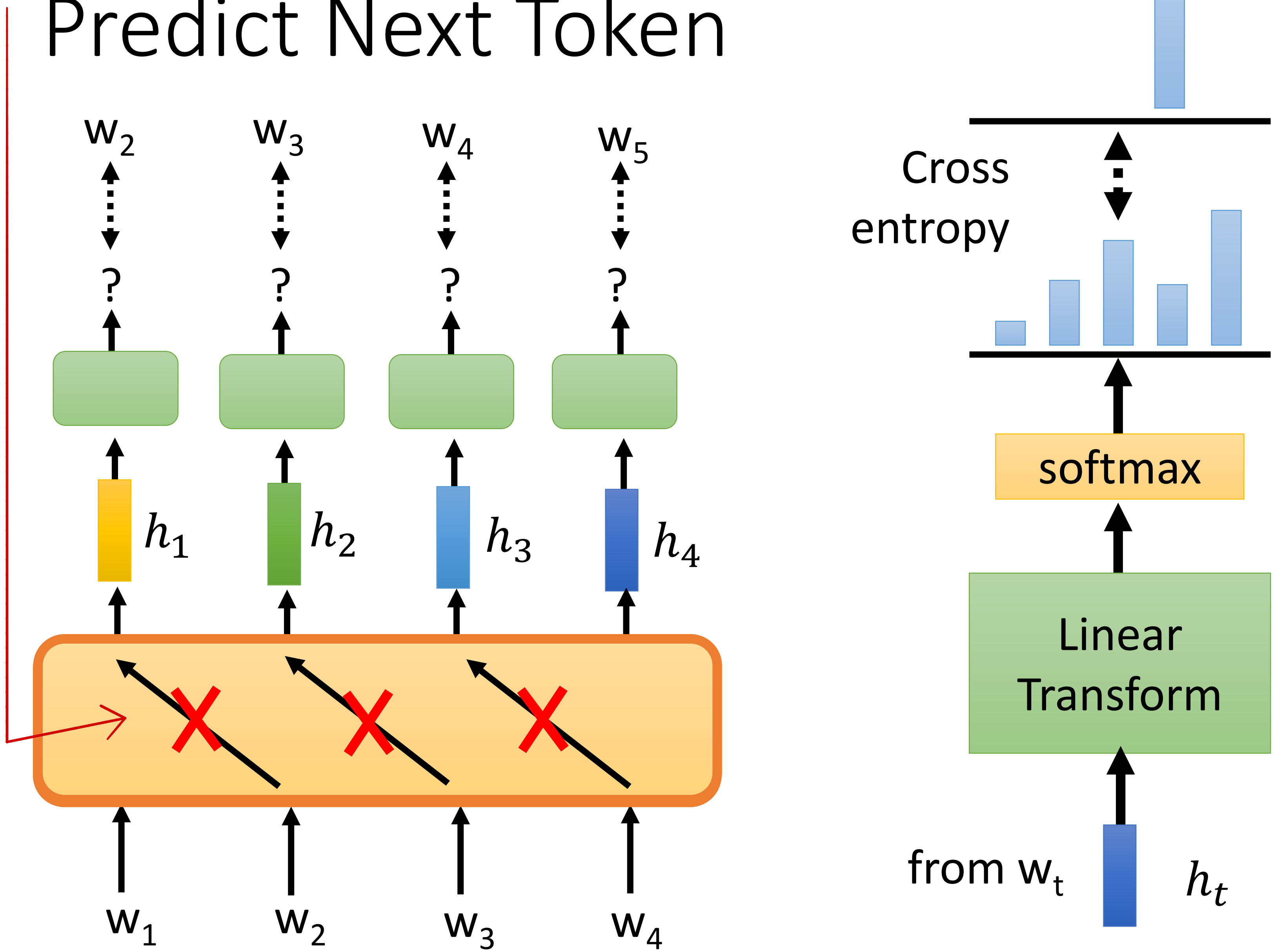
I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

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.

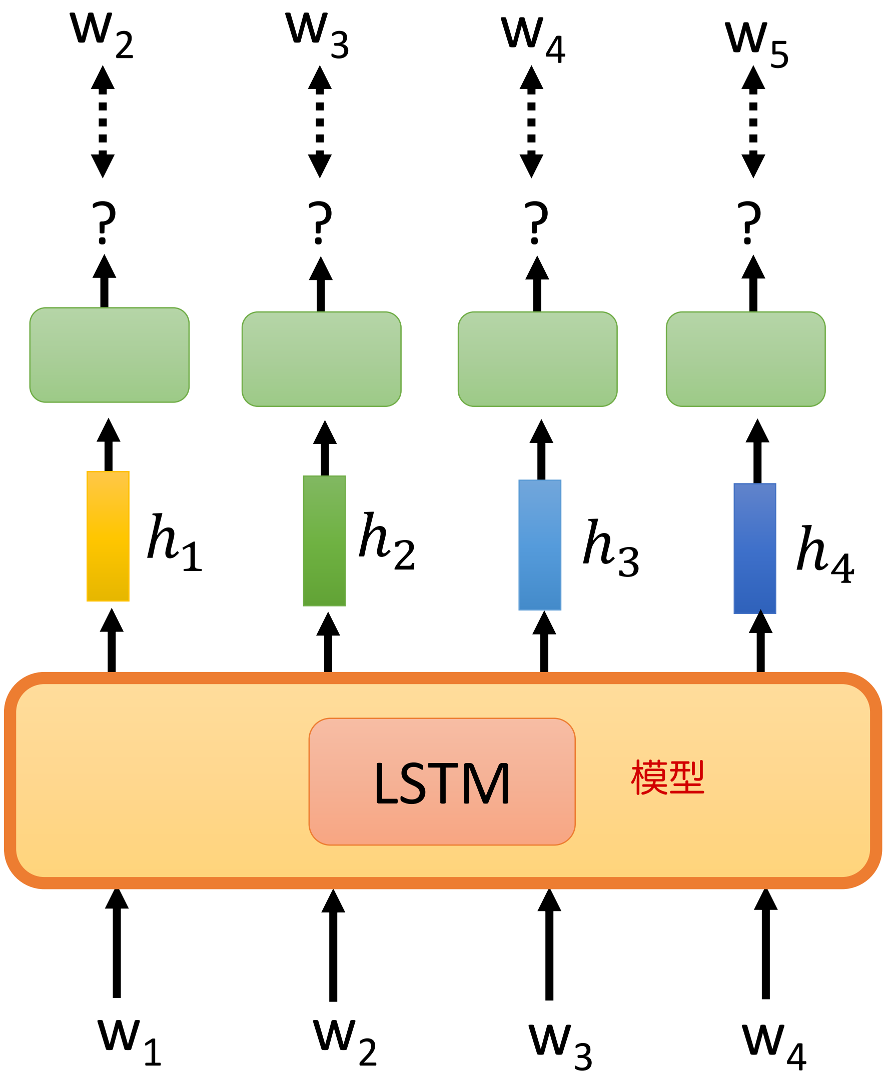


不能将 w1,w2,w3,w4 全部都输入进去，然后预测 w2,w3,w4,w5，因为模型可能学偏，
就是觉得输入的下一个就是答案。
也就是说，不能让模型提前看到答案，要不然他就可能会找规律，而不是提取特征。

# Predict Next Token



# Predict Next Token




This is exactly how we train language models (LM).

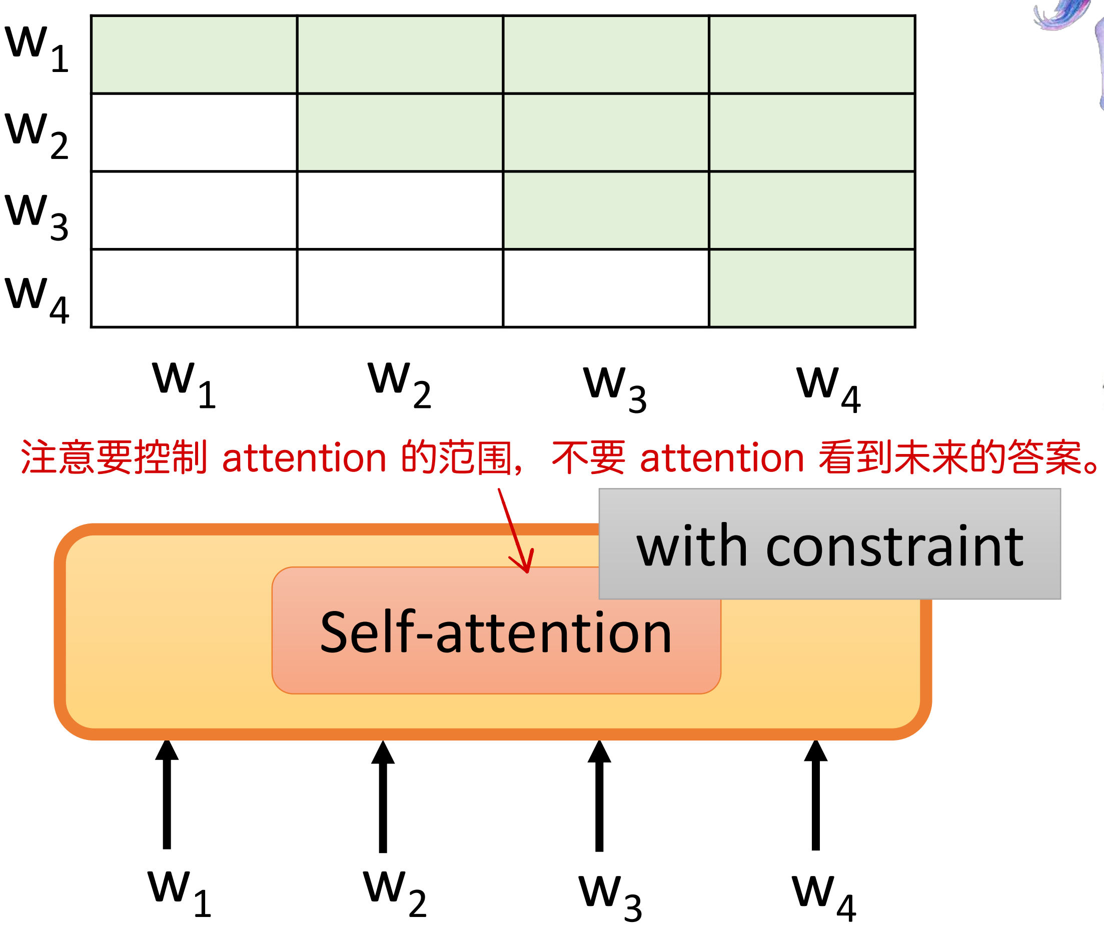
Universal Language Model Fine-tuning (ULMFiT)

[Howard, et al., ACL'18]

ELMo

[Peters, et al., NAACL'18]

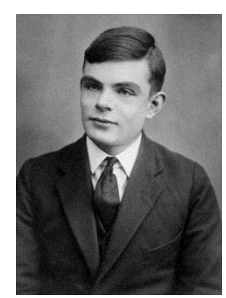
# Predict Next Token




GPT
[Alec, et al., 2018]

GPT-2
[Alec, et al., 2019]

Megatron
[Shoeybi, et al., arXiv'19]



Turing NLG

https://talktotransformer.com/

# Predict Next Token

## They can do generation.

EM PROMPT
-WRITTEN)

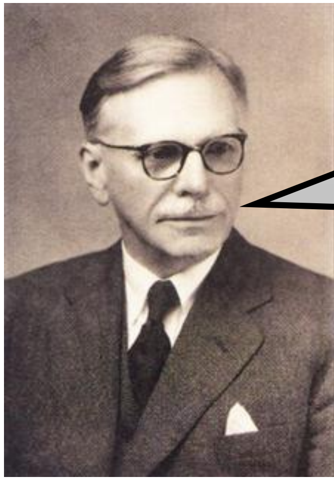
*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

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.

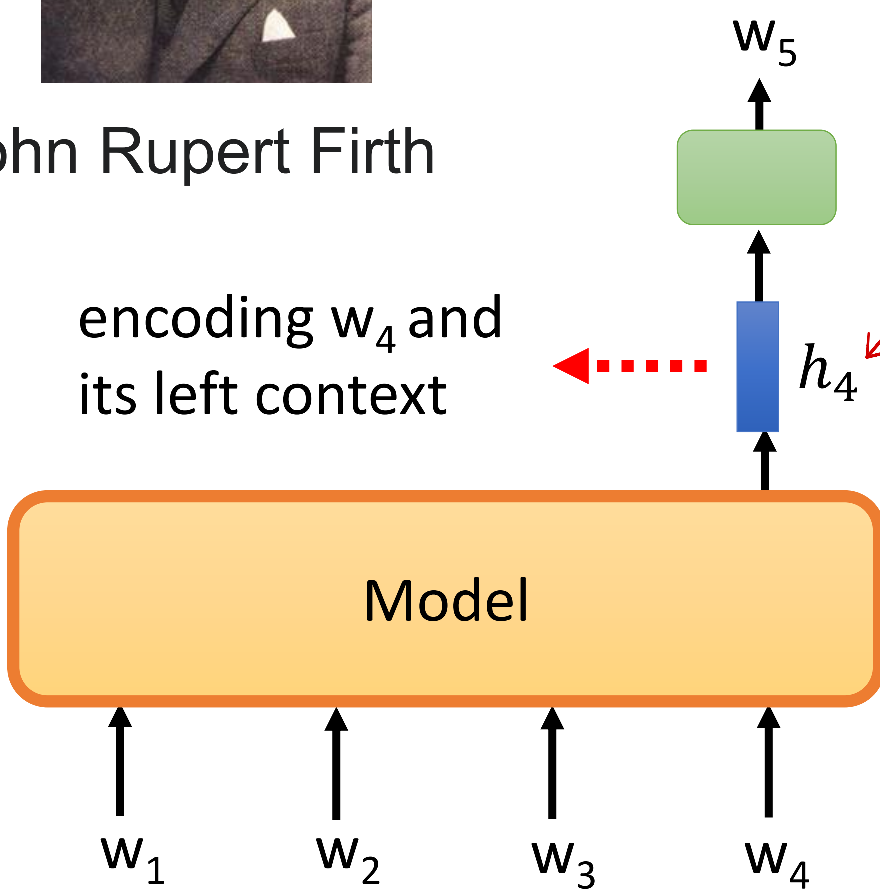
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock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



You shall know a word by the company it keeps

要了解一个单词，
就要知道他的邻居。

John Rupert Firth



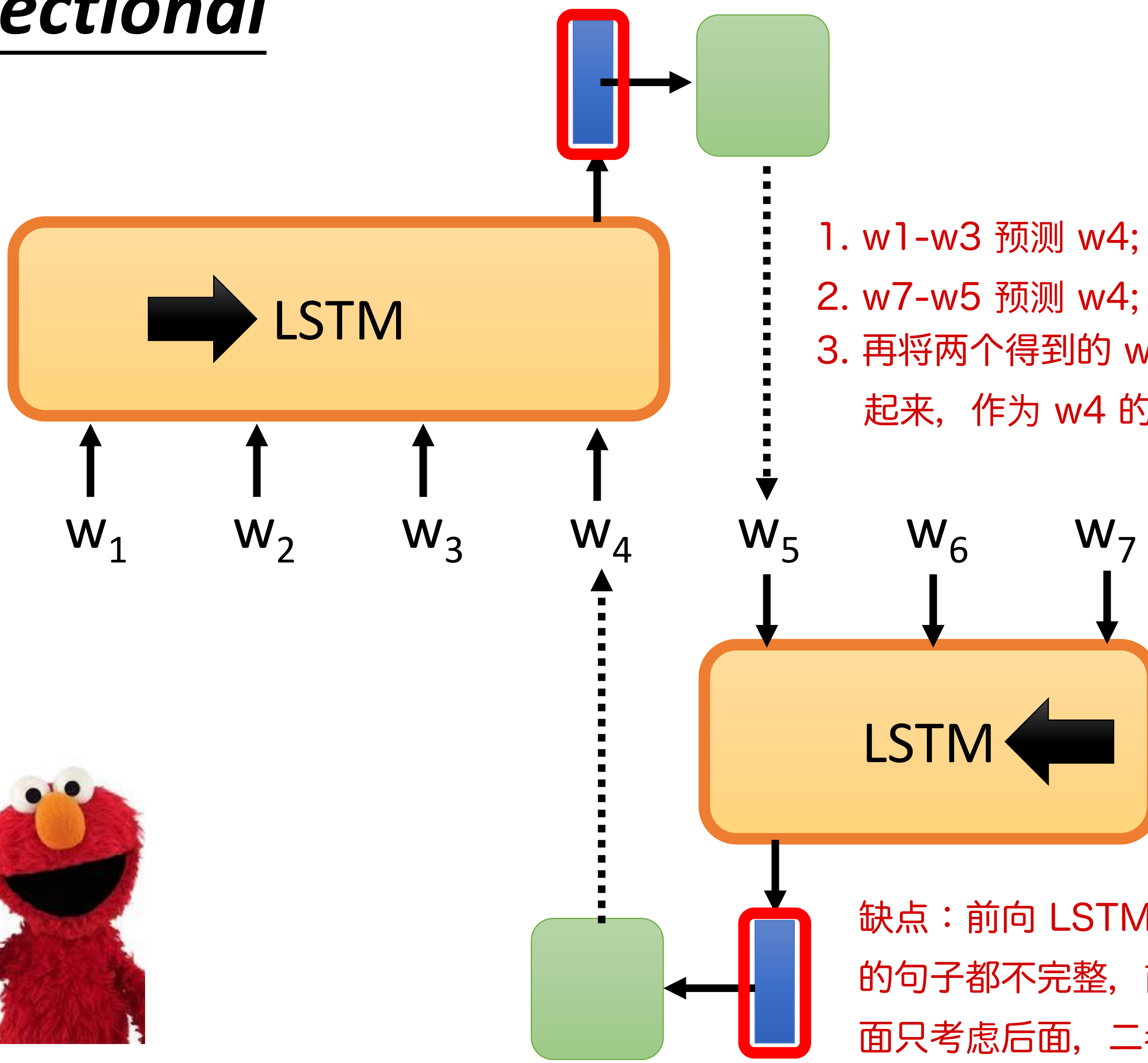

h4包含了 w1 w2 w3 w4 的所有信息，
然后再去预测 w5。

How about the right context!?

这样做只包含了左边的信息，
右边的信息怎么办呢？

# *Predict Next Token - Bidirectional*

双向考虑




1. w1-w3 预测 w4;
2. w7-w5 预测 w4;
3. 再将两个得到的 w4 的 emb concat 起来，作为 w4 的 emb.

缺点：前向 LSTM 和后向 LSTM 看到的句子都不完整，前面只考虑前面，后面只考虑后面，二者没有交互。

ELMO

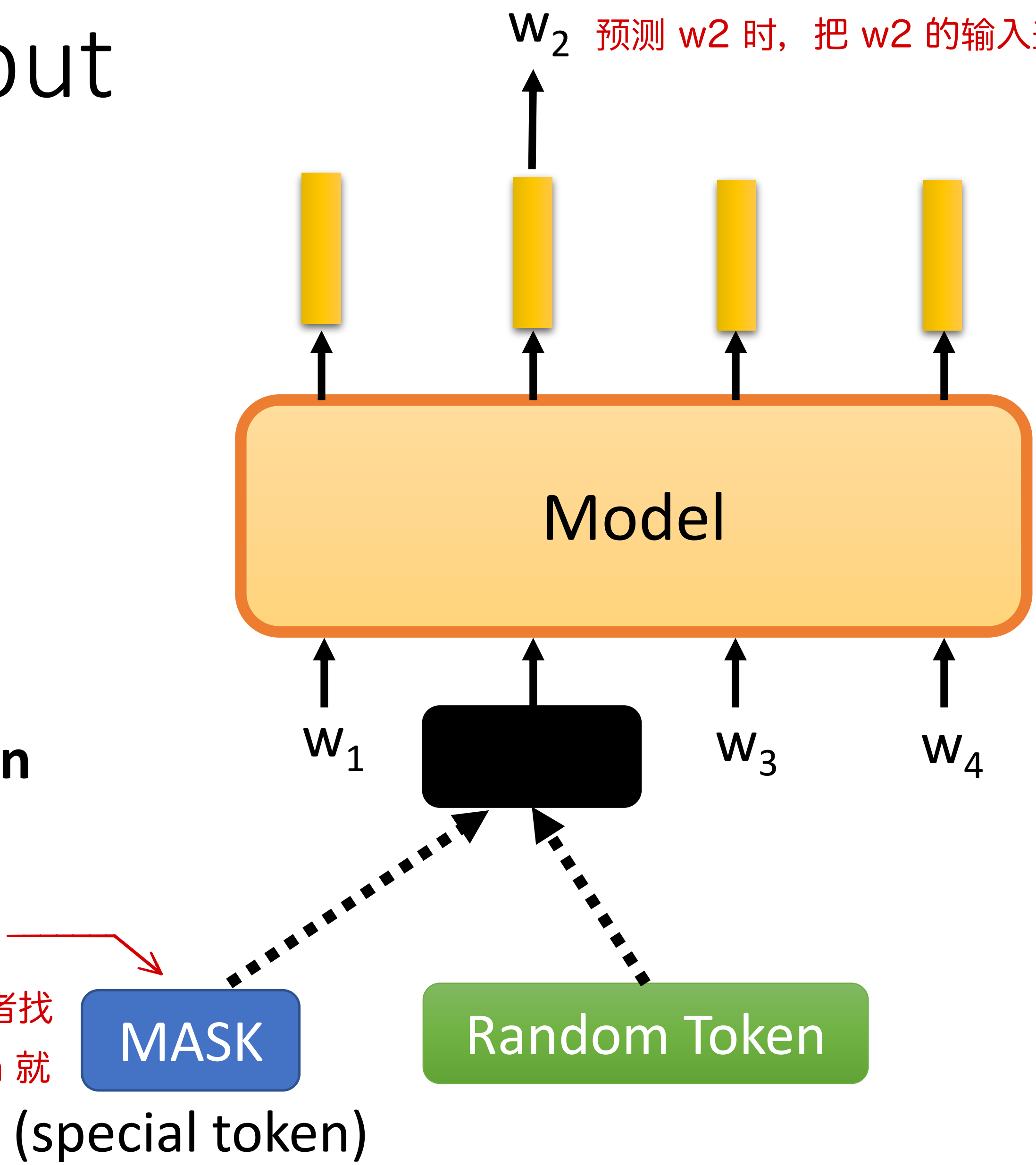
# Masking Input

BERT

[Devlin, et al., NAACL’19]

Transformer

**(no limitation on self-attention)**



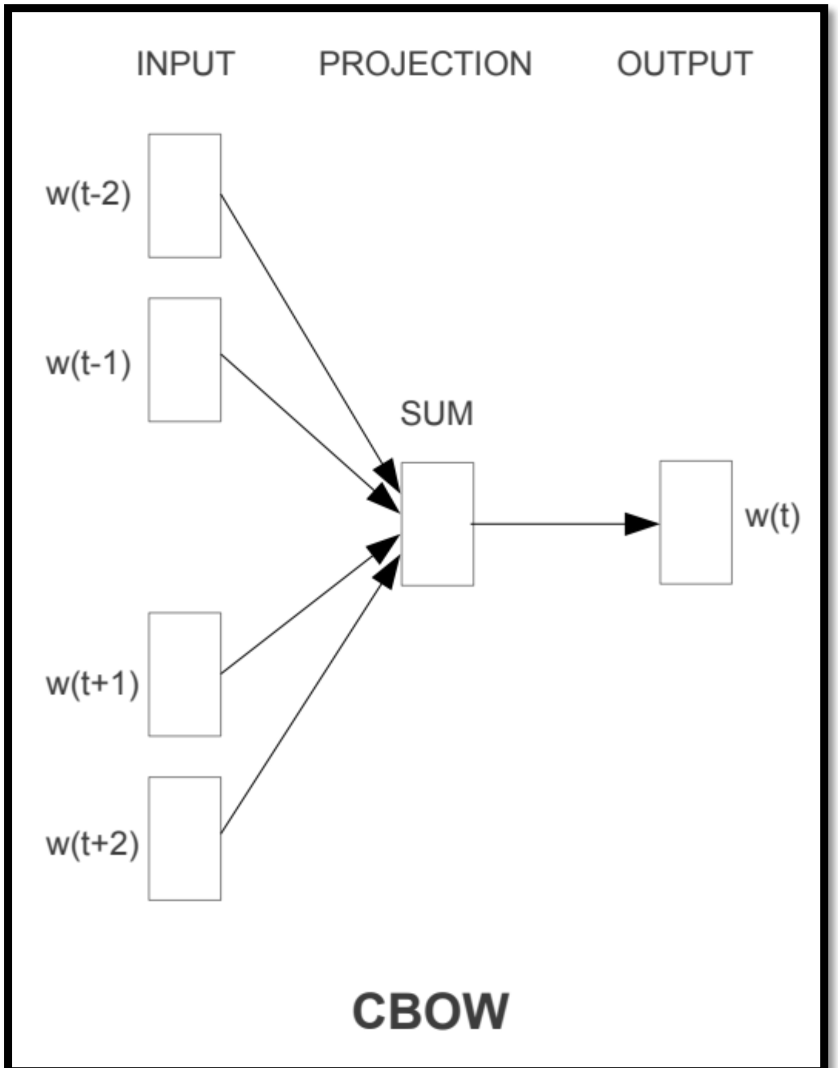

预测 w2 时，把 w2 的输入盖住。

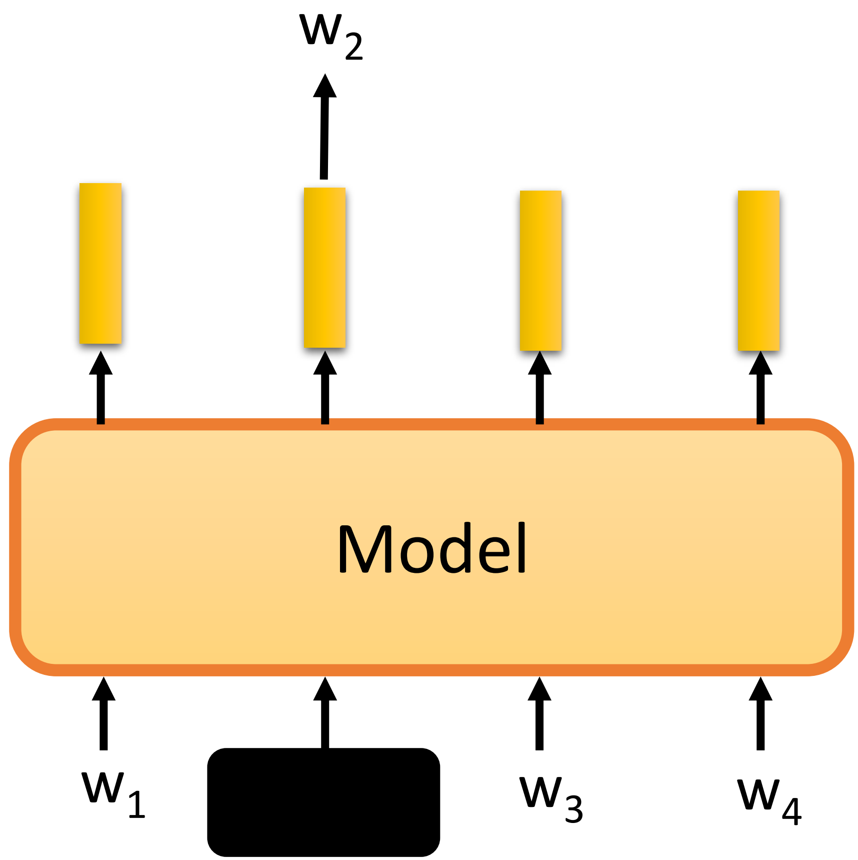
用 mask 或者 随机 token 把 w2 盖住。
反正就是不让 model 提前知道答案，或者找到其他的规律。比如 w1 的下一个 token 就是答案之类的这种规律。

1. CBOW 有固定的 window，只能看左右固定个数的 token，而 Bert 往左往右 想看多少看多少。

2. CBOW 的网络结构比较简单，而 Bert 的网络结构很复杂。

# Masking Input




Using context to predict the missing token

# Masking Input

Is random masking good enough?

mask 是随机的，没有 long-term dependency

- Whole Word Masking (WWM) [Cui, et al., arXiv'19]

**[Original Sentence]**
使用语言模型来预测下一个词的probability。
**[Original Sentence with CWS]** 分词
使用 语言 **模型** 来 **预测** 下 一个 词 的 **probability** 。

**[Original BERT Input]** 盖掉字
使 用 语 言 **[MASK] 型** 来 **[MASK] 测** 下 一 个 词 的 **pro [MASK] ##lity** 。
**[Whold Word Masking Input]** 盖掉整个词
使 用 语 言 **[MASK] [MASK]** 来 **[MASK] [MASK]** 下 一 个 词 的 **[MASK] [MASK SK]** 。

Source of image: https://arxiv.org/abs/1906.08101

盖掉 phrase 盖掉实体词

- Phrase-level & Entity-level
  [Sun, et al., ACL'19]

Enhanced Representation through Knowledge Integration (ERNIE)

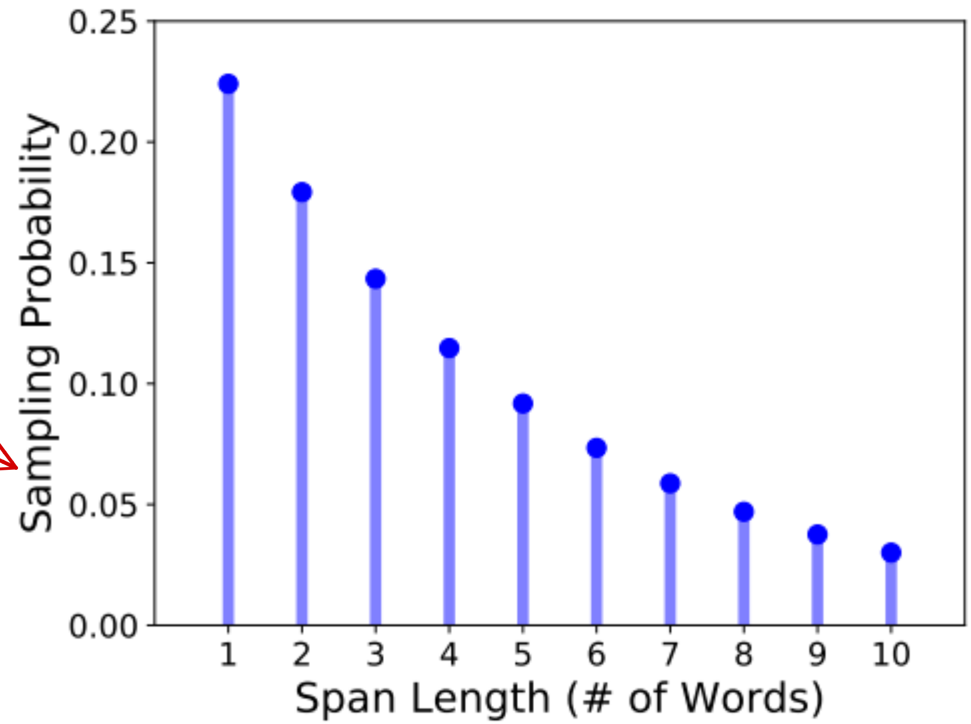
Source of image: https://arxiv.org/abs/1907.10529

# SpanBert

[Joshi, et al., TACL'20]

一次盖住一排 token

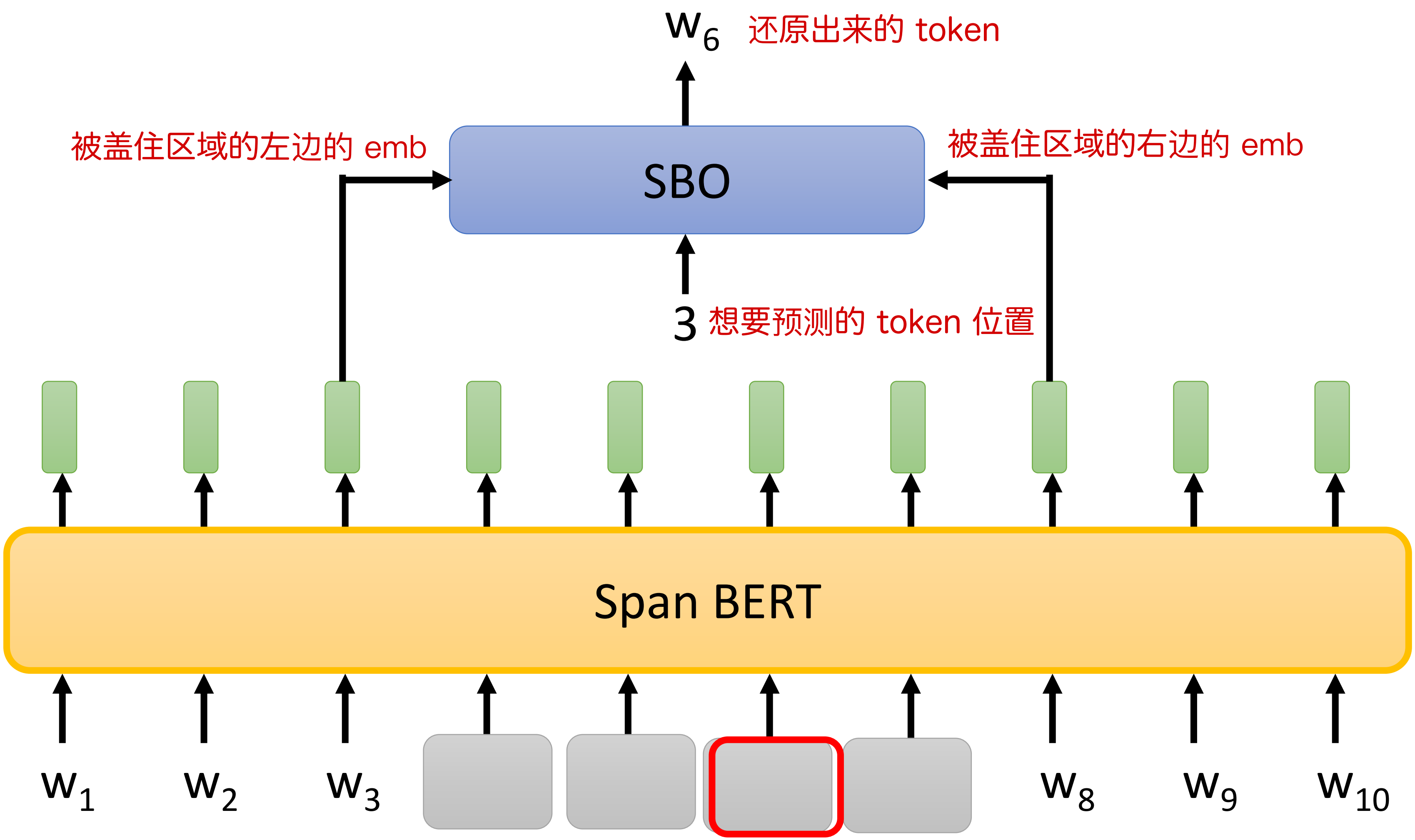
采样概率




| | SQuAD 2.0 | NewsQA | TriviaQA | Coreference | MNLI-m | QNLI | GLUE (Avg) |
|---|---|---|---|---|---|---|---|
| Subword Tokens | 83.8 | 72.0 | 76.3 | **77.7** | 86.7 | 92.5 | 83.2 |
| Whole Words | 84.3 | 72.8 | 77.1 | 76.6 | 86.3 | 92.8 | 82.9 |
| Named Entities | 84.8 | 72.7 | 78.7 | 75.6 | 86.0 | 93.1 | 83.2 |
| Noun Phrases | 85.0 | **73.0** | 77.7 | 76.7 | 86.5 | 93.2 | **83.5** |
| Geometric Spans | **85.4** | **73.0** | **78.8** | 76.4 | **87.0** | **93.3** | 83.4 |

# SpanBert – Span Boundary Objective (SBO)

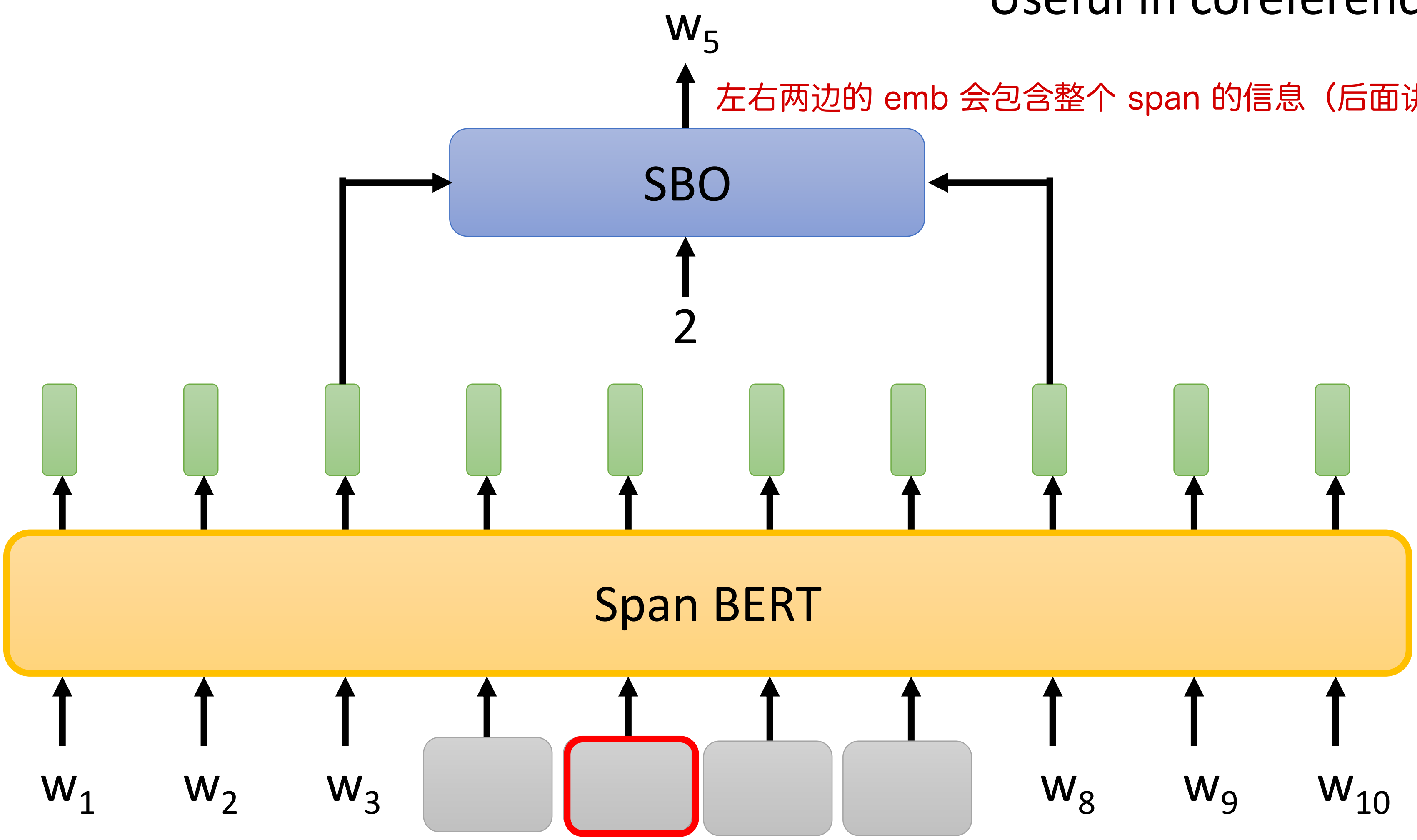
提出的一个新的训练方法



# SpanBert –
# Span Boundary Objective (SBO)

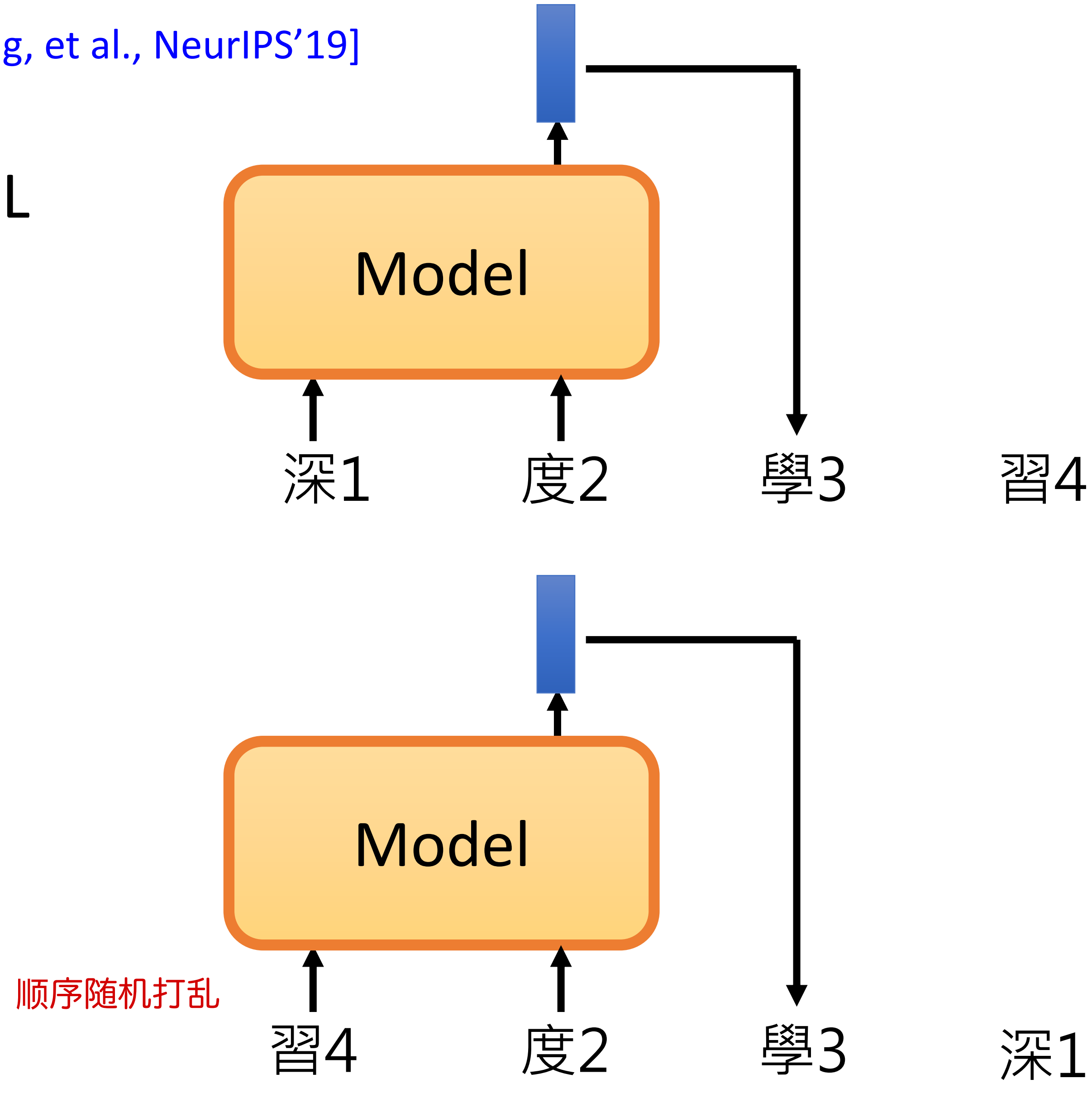
Useful in coreference?

左右两边的 emb 会包含整个 span 的信息（后面讲）



# XLNet [Yang, et al., NeurIPS'19]

Transformer-XL



# XLNet [Yang, et al., NeurIPS'19]

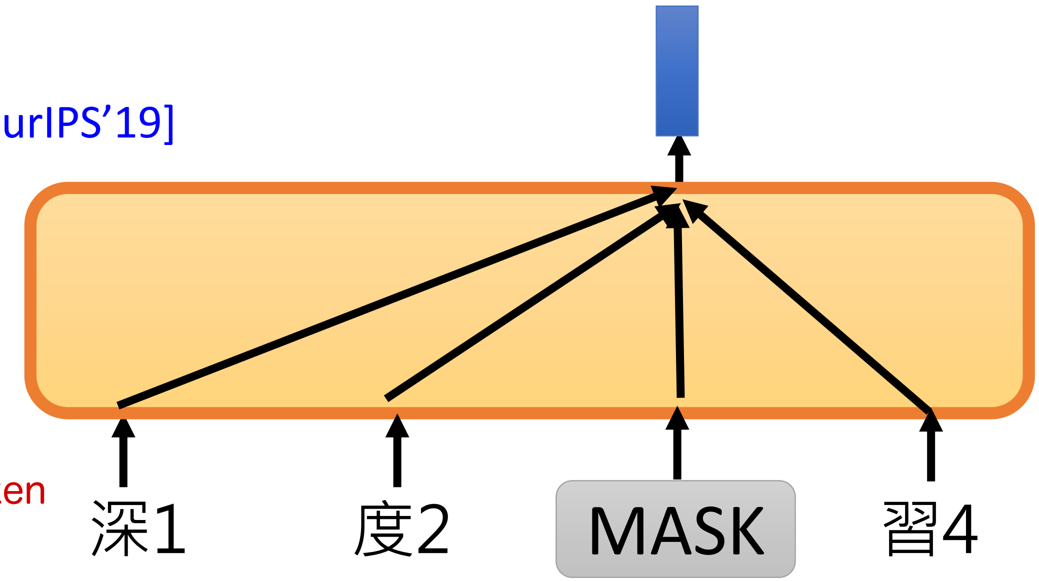
（其他内容，自己看论文）

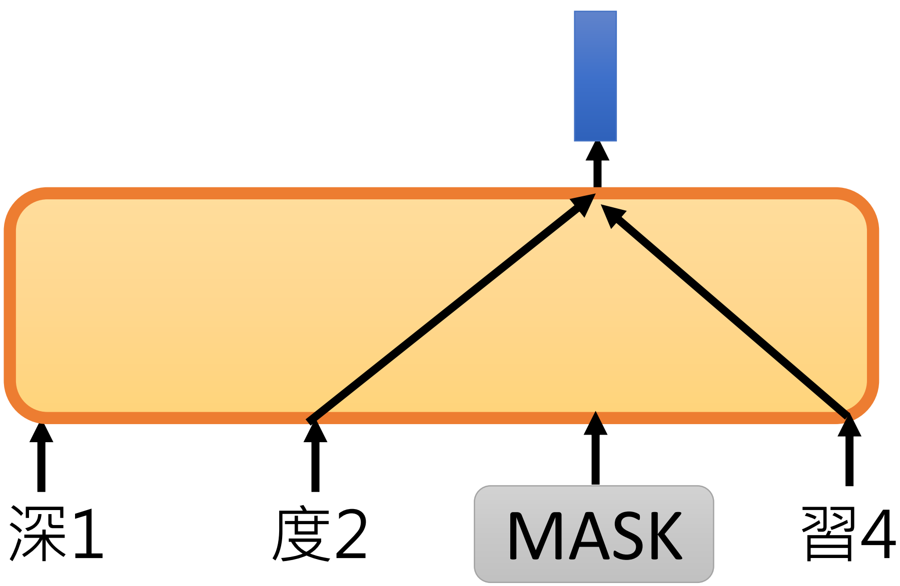
Transformer-XL

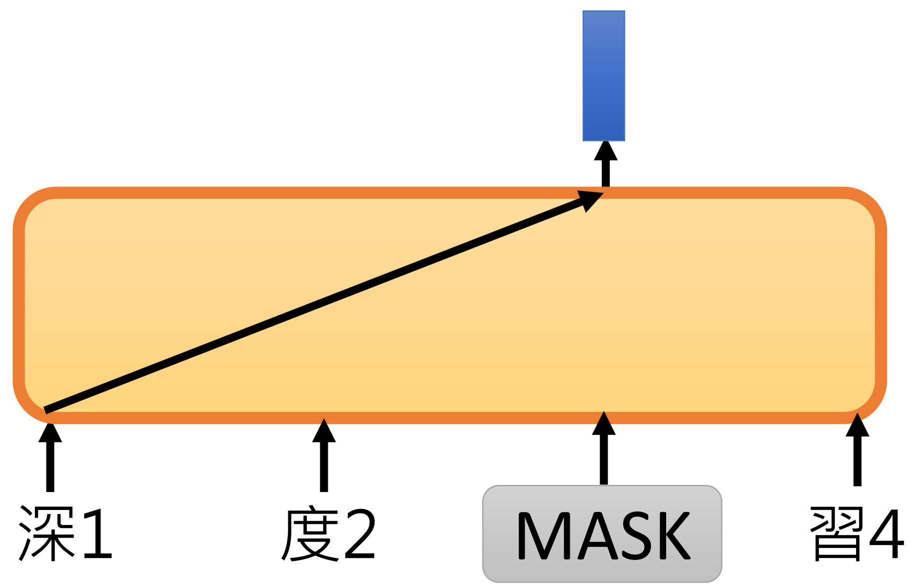
不给 model 看 mask token

只给 position information

只告诉他预测的词的位置，而不给 mask token



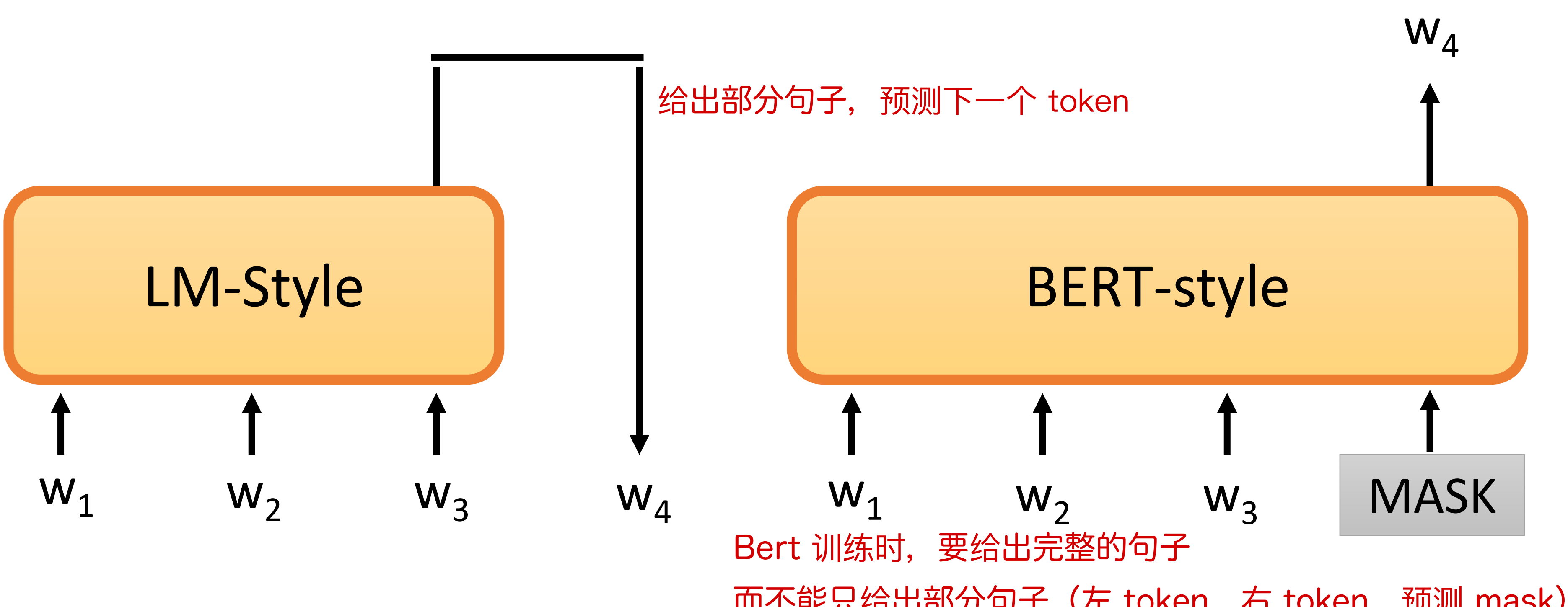



句子生成

# BERT cannot talk?

Limited to autoregressive model (non-autoregressive next time)

Given partial sequence, predict the next token



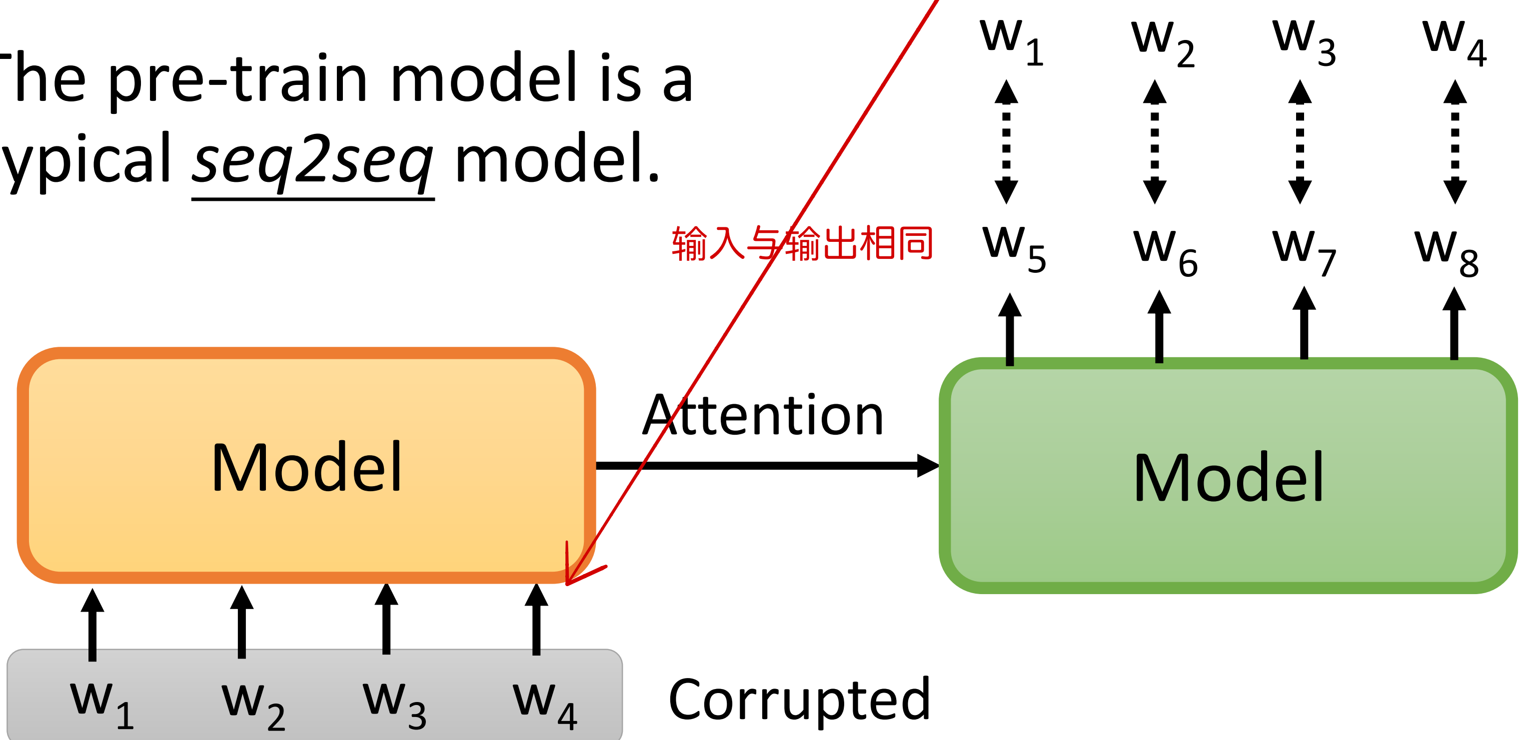

What LM born for

Never seen partial sequence

# MASS / BART

- The pre-train model is a typical *seq2seq* model.

Reconstruct the input




要对输入进行一些破坏，否则 model 直接把输入输出就好了，什么也学不到。

如何进行破坏

MAsked Sequence to Sequence pre-training (MASS) [Song, et al., ICML’19]

Bidirectional and Auto-Regressive Transformers (BART) [Lewis, et al., arXiv’19]

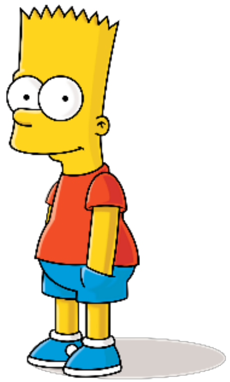
输入破坏

# Input Corruption

MASS

把一些部分 mask 起来

A B [SEP] C [ ] E



BART

A B [SEP] C E 删除 (Delete “D”)

A B [SEP] C D E

C D E [SEP] A B 乱序 (permutation)

D E A B [SEP] C 旋转 (rotation)

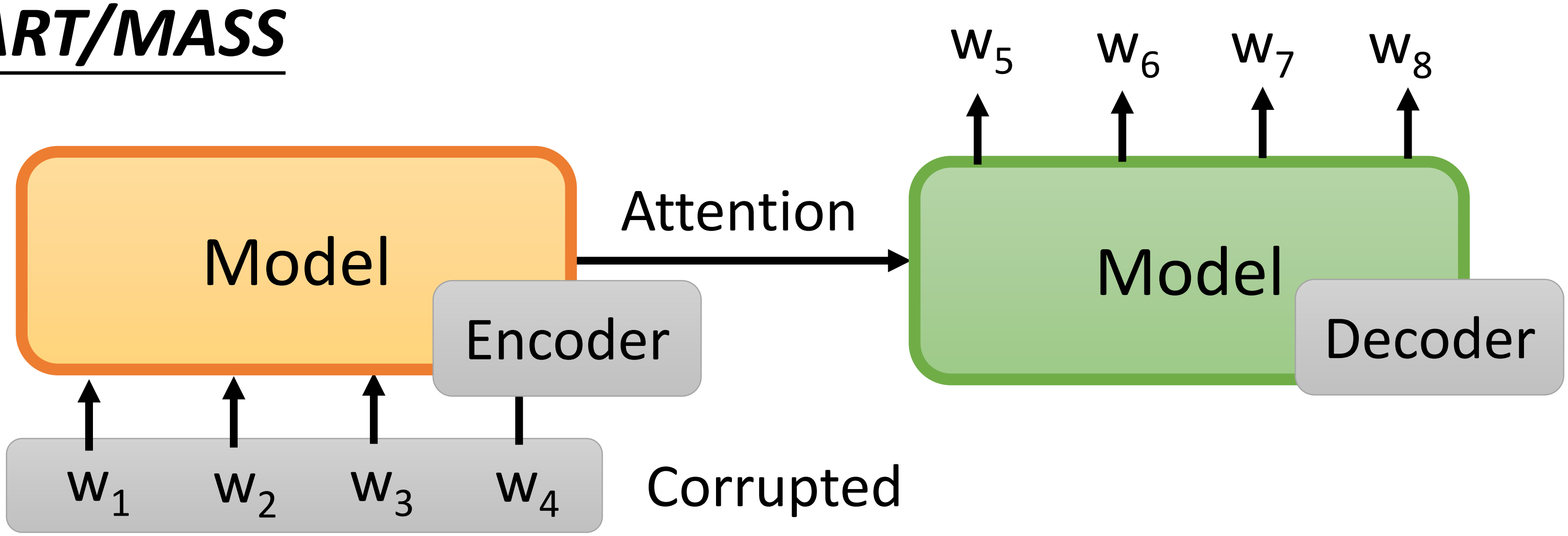
- Permutation / Rotation do not perform well.
- Text Infilling is consistently good.

A [ ] B [SEP] [ ] E

随机插入 mask，用 mask 遮挡 tokens

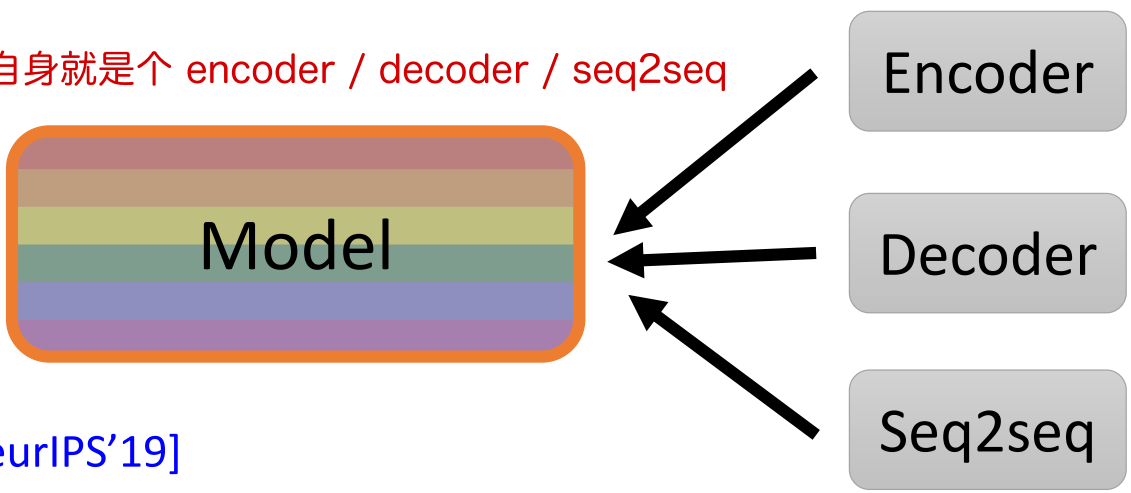
**Text Infilling**

## BART/MASS



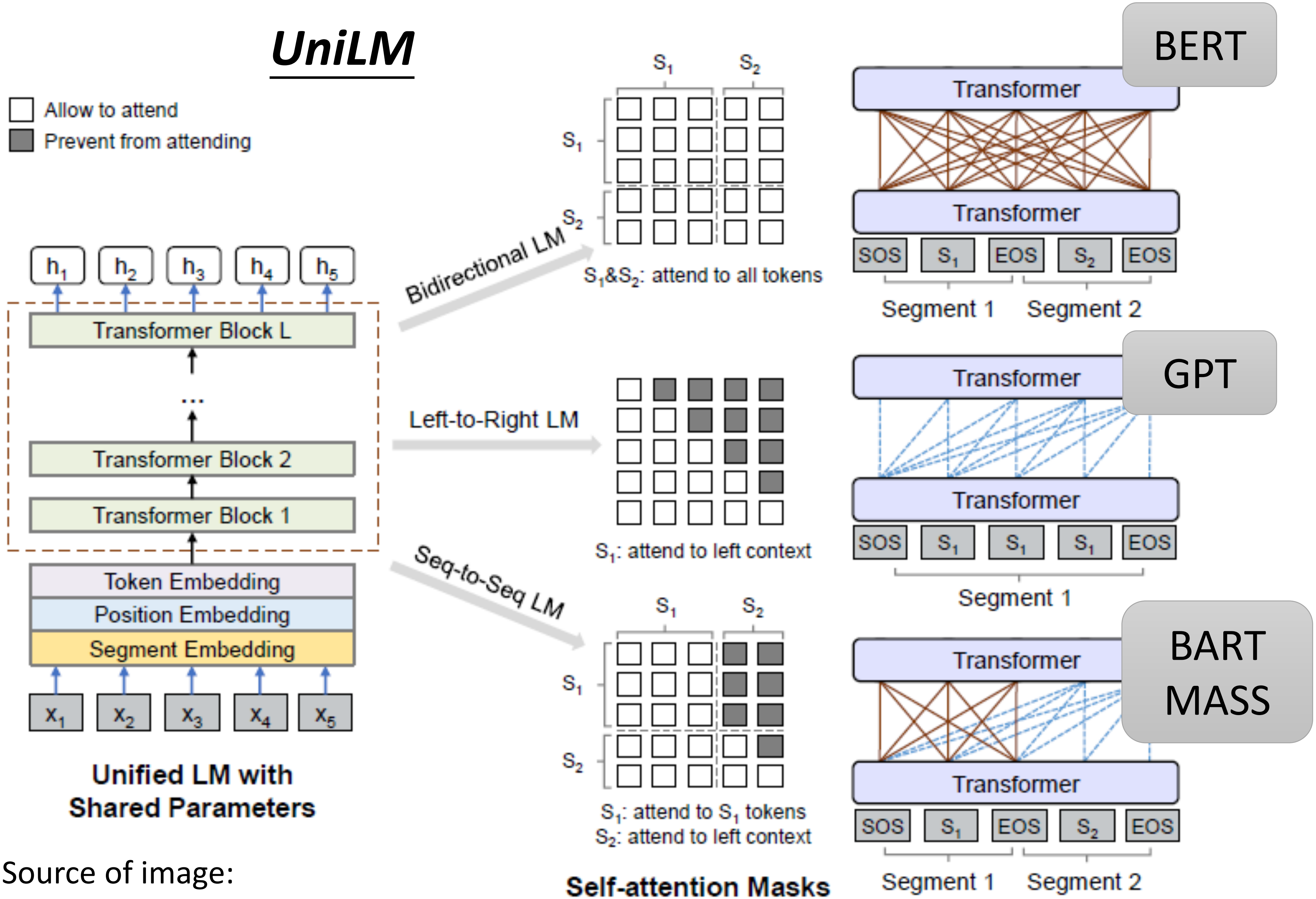

## UniLM

自身就是个 encoder / decoder / seq2seq




[Dong, et al., NeurIPS’19]

# *UniLM*



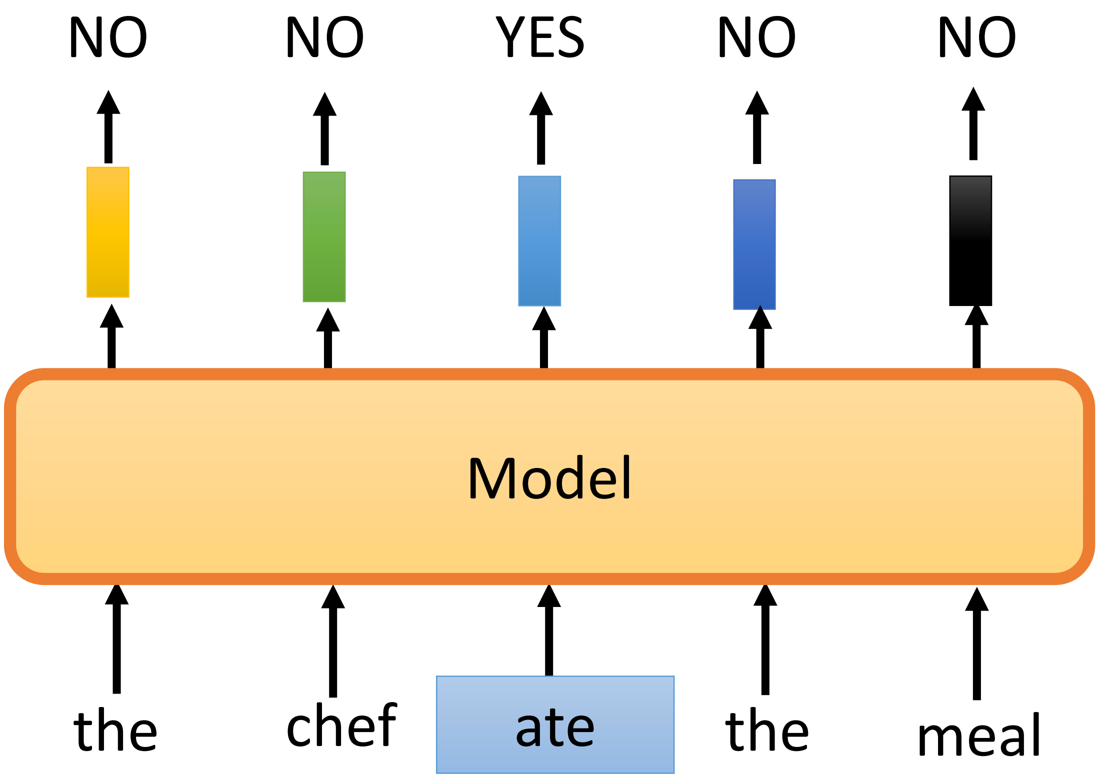

Source of image:
https://arxiv.org/pdf/1905.03197.pdf

# Replace or Not?

Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)

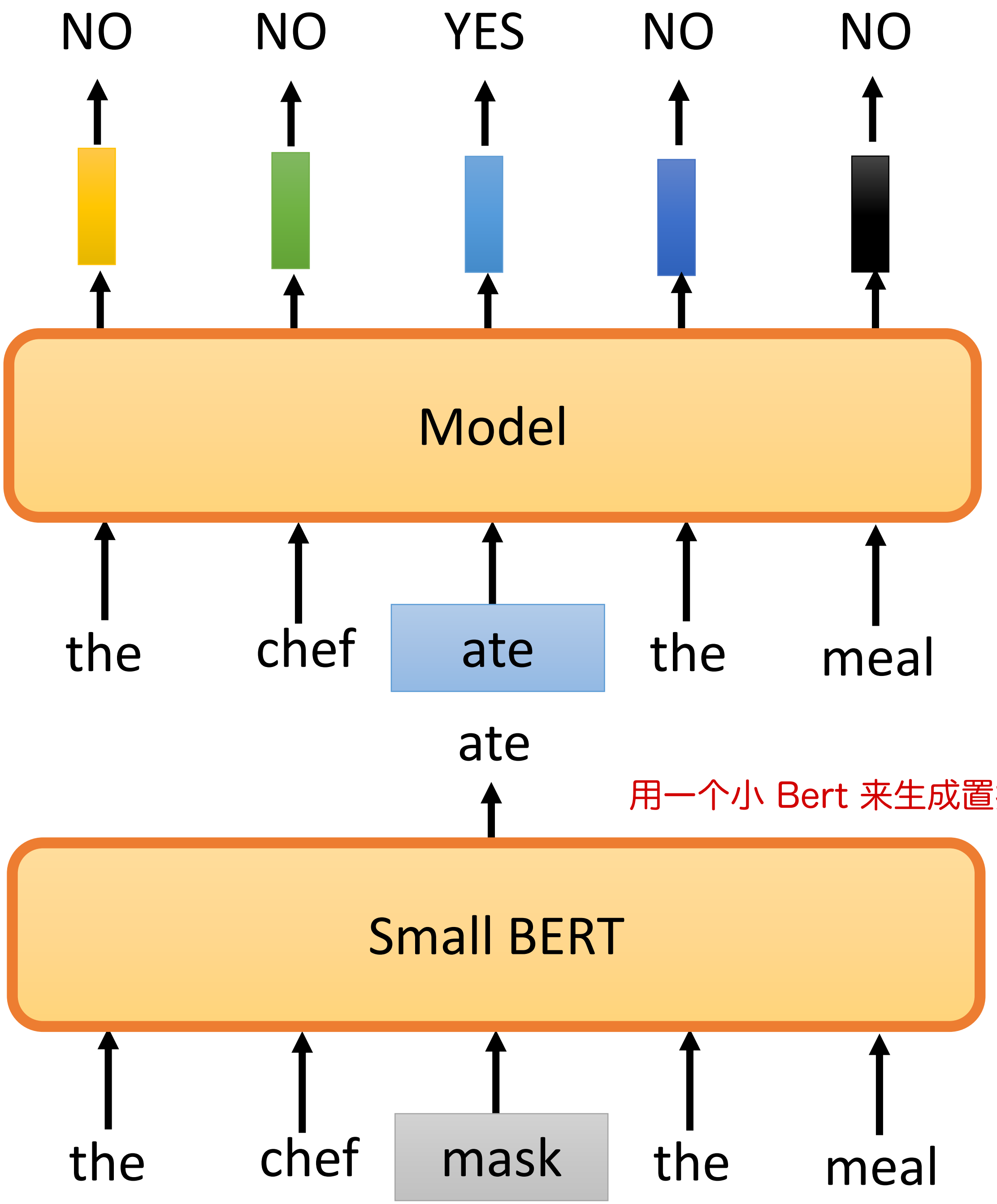
把一个句子中的某个词替换掉，判断哪个词是被换掉的

ELECTRA



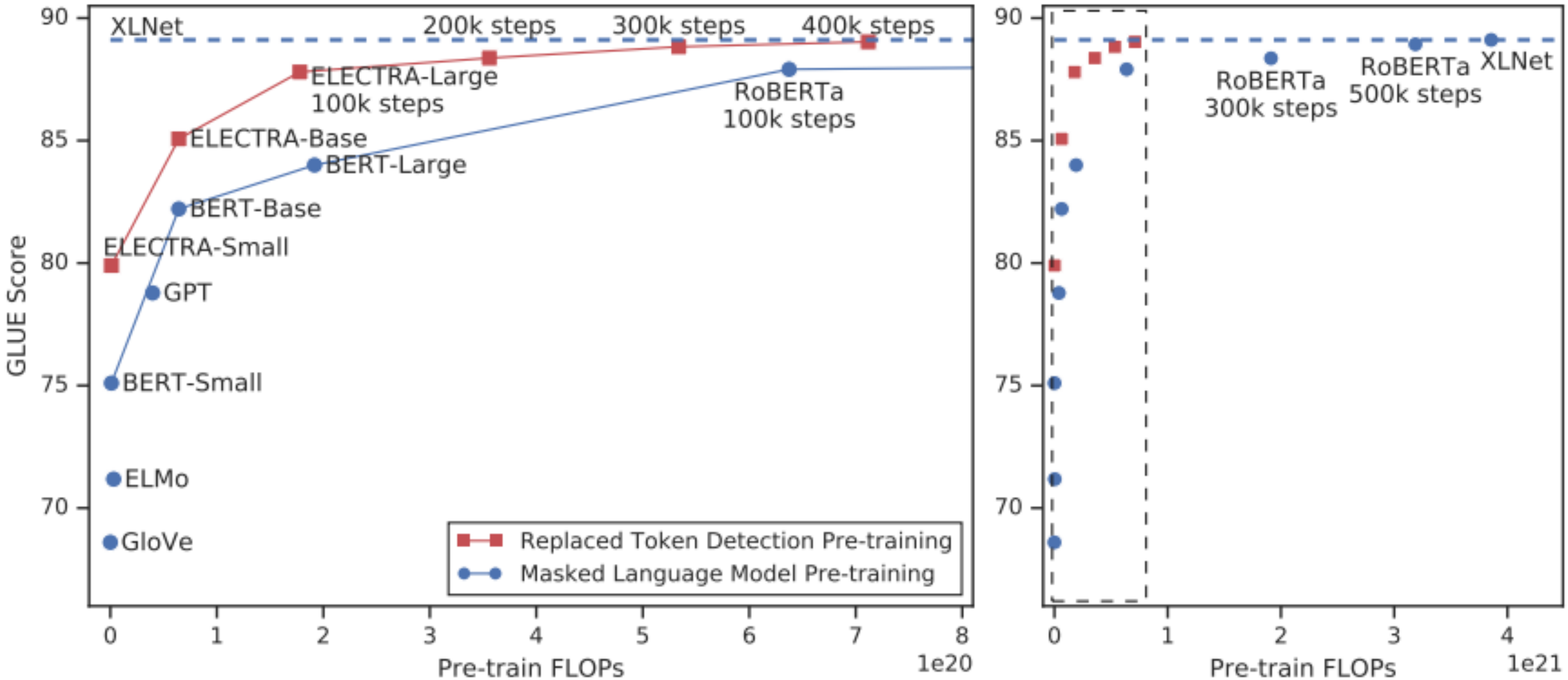

Predicting yes/not is easier than reconstruction.

Every output position is used.



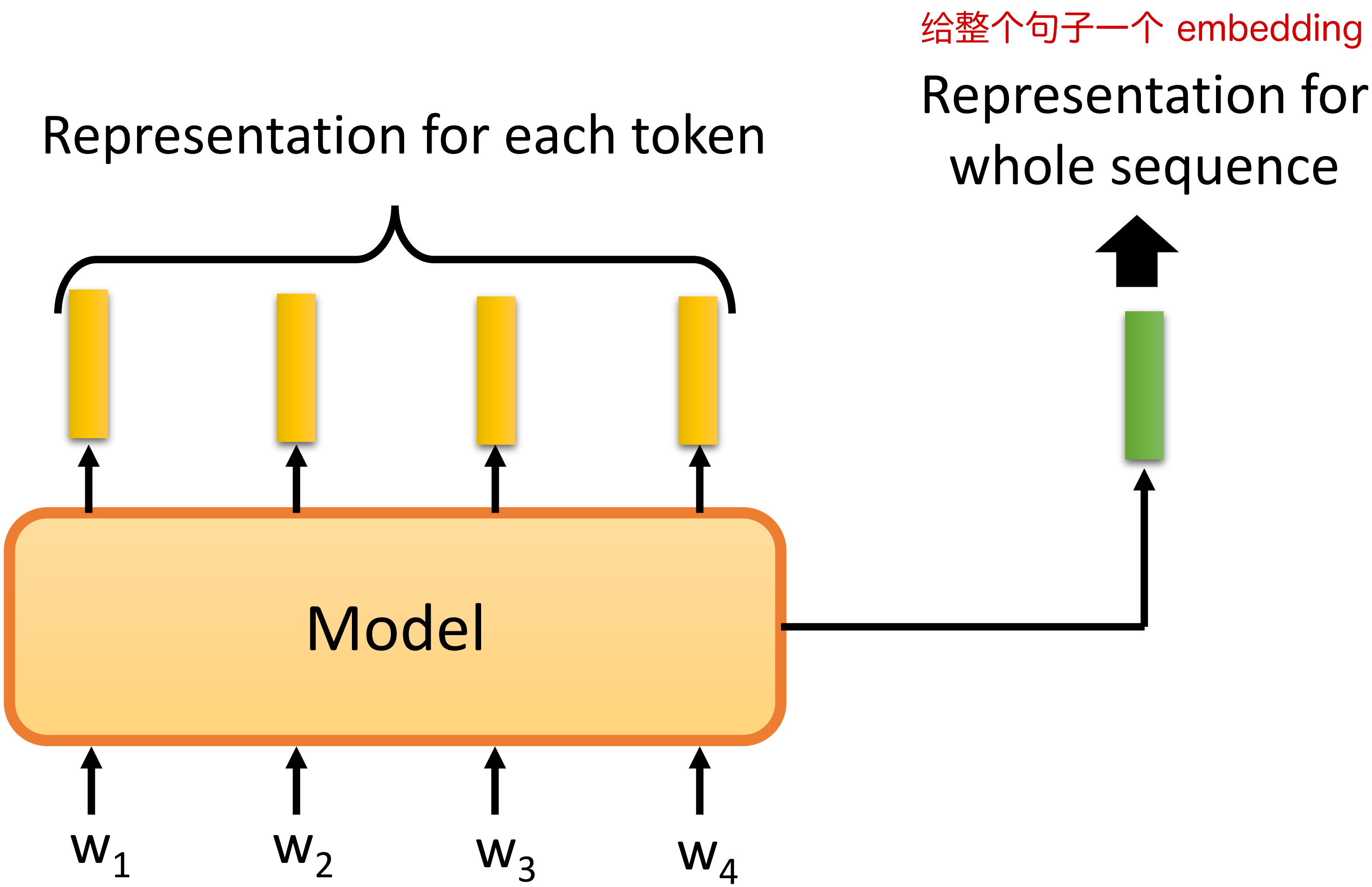

用一个小 Bert 来生成置换的词，要不然太好猜了。

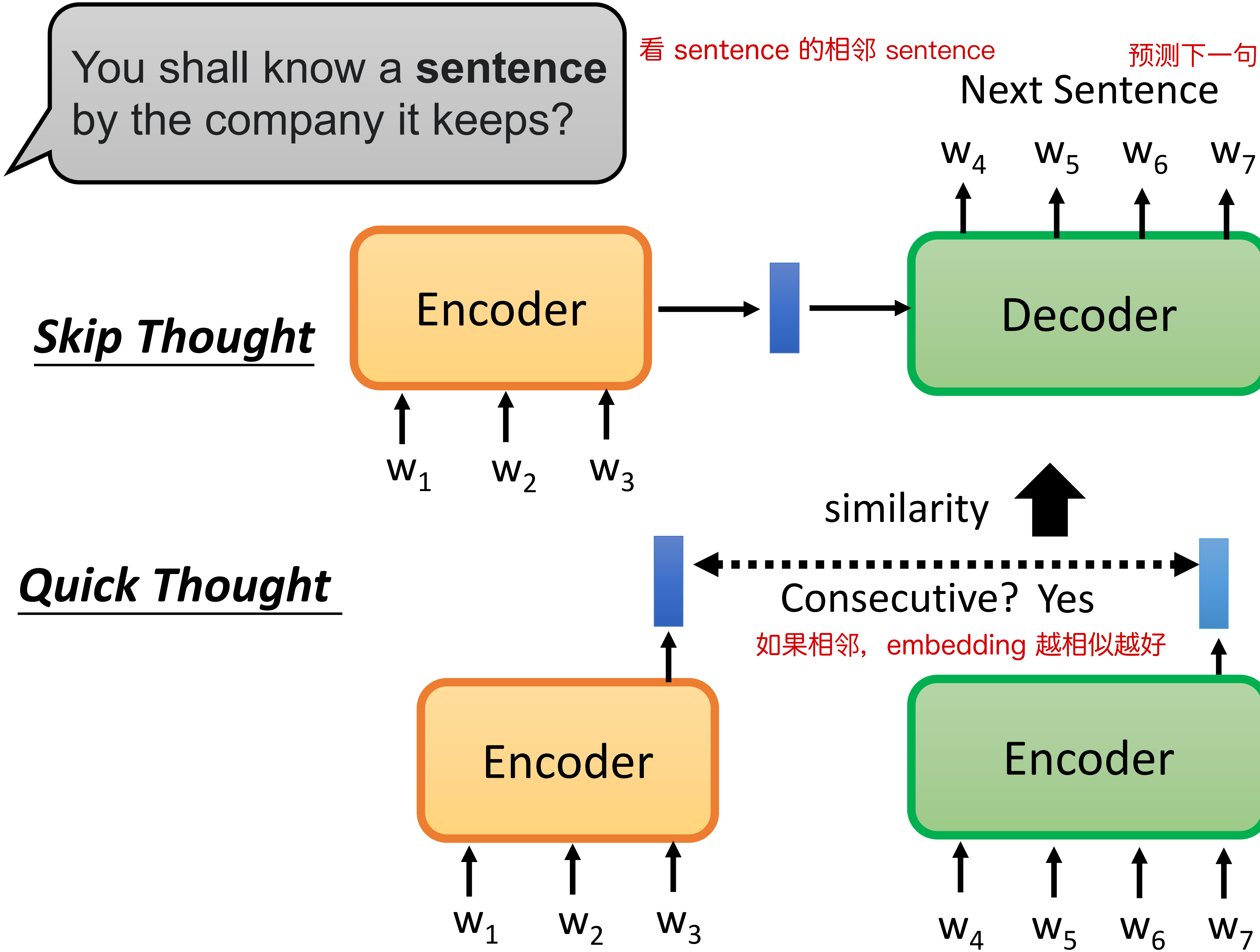
Note: This is not GAN.



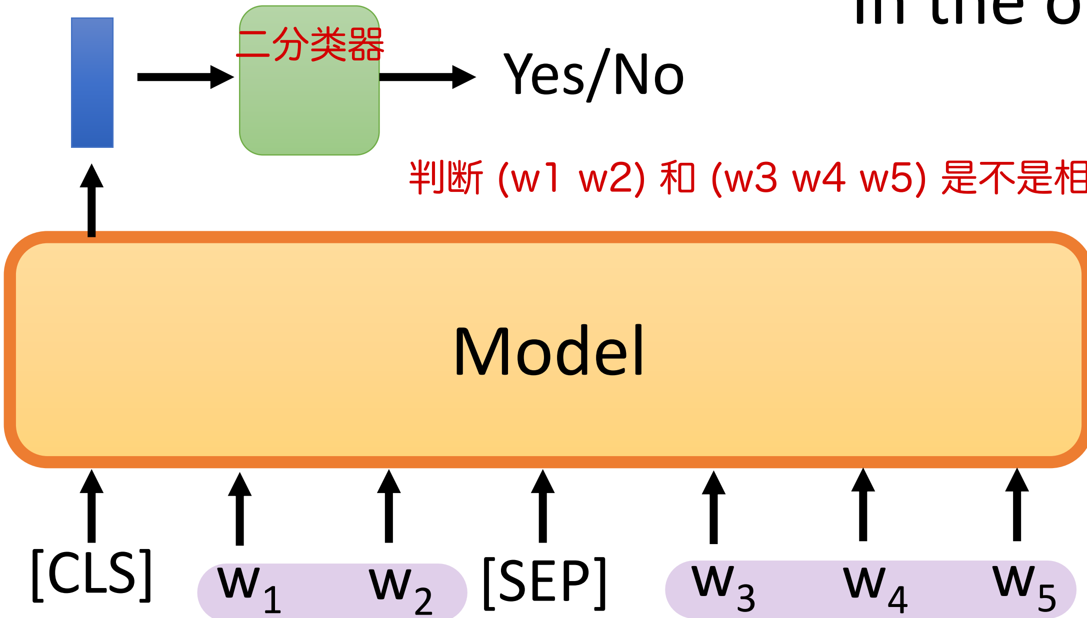

Source of image: https://arxiv.org/abs/2003.10555

# Sentence Level




You shall know a **sentence** by the company it keeps?
看 sentence 的相邻 sentence
预测下一句
Next Sentence
w4 w5 w6 w7
Skip Thought
Encoder
Decoder
w1 w2 w3
similarity
Quick Thought
Consecutive? Yes
如果相邻，embedding 越相似越好
Encoder
Encoder
w1 w2 w3
w4 w5 w6 w7

# In the original BERT, .....




**NSP**: Next sentence prediction

（效果不好，没什么用）

Robustly optimized BERT approach (RoBERTa)

[Liu, et al., arXiv'19]

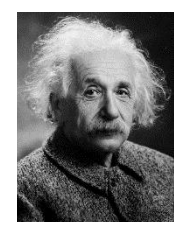
（句子顺序预测）

**SOP**: Sentence order prediction

(w1w2)(w3w4w5) - 输出 yes （正序）

(w3w4w5)(w1w2) - 输出 no （反序）

Used in ALBERT



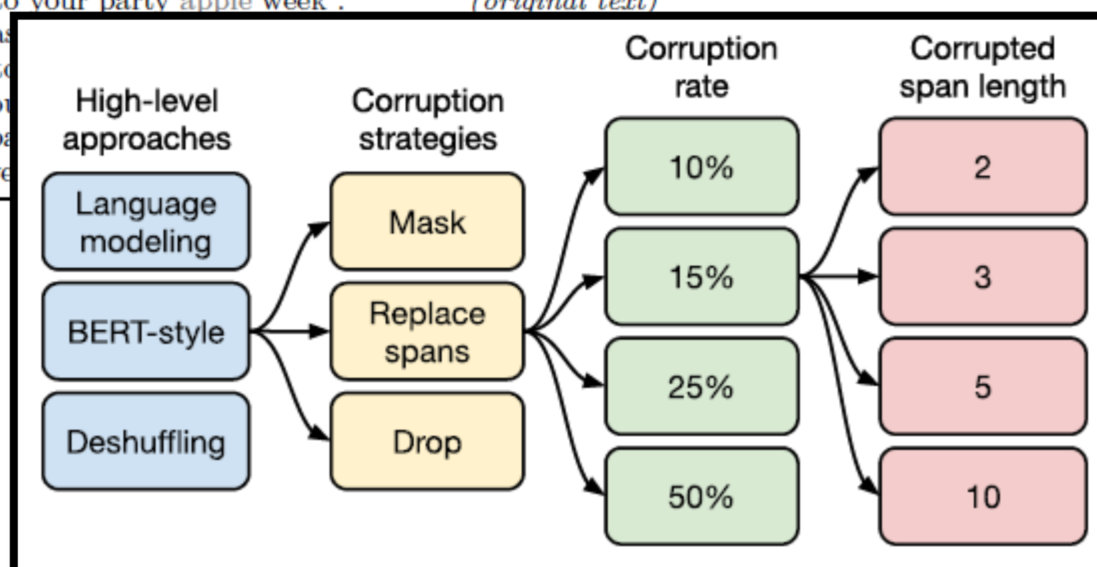
structBERT (Alice) [Want, et al., ICLR'20]

# T5 – Comparison [Raffel, et al., arXiv’19]

里面比较了很多 pre-train 的方法

- 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 you | |
| I.i.d. noise, drop tokens | Thank you me to your pa | |
| Random spans | Thank you <X> to <Y> we | |



# Knowledge

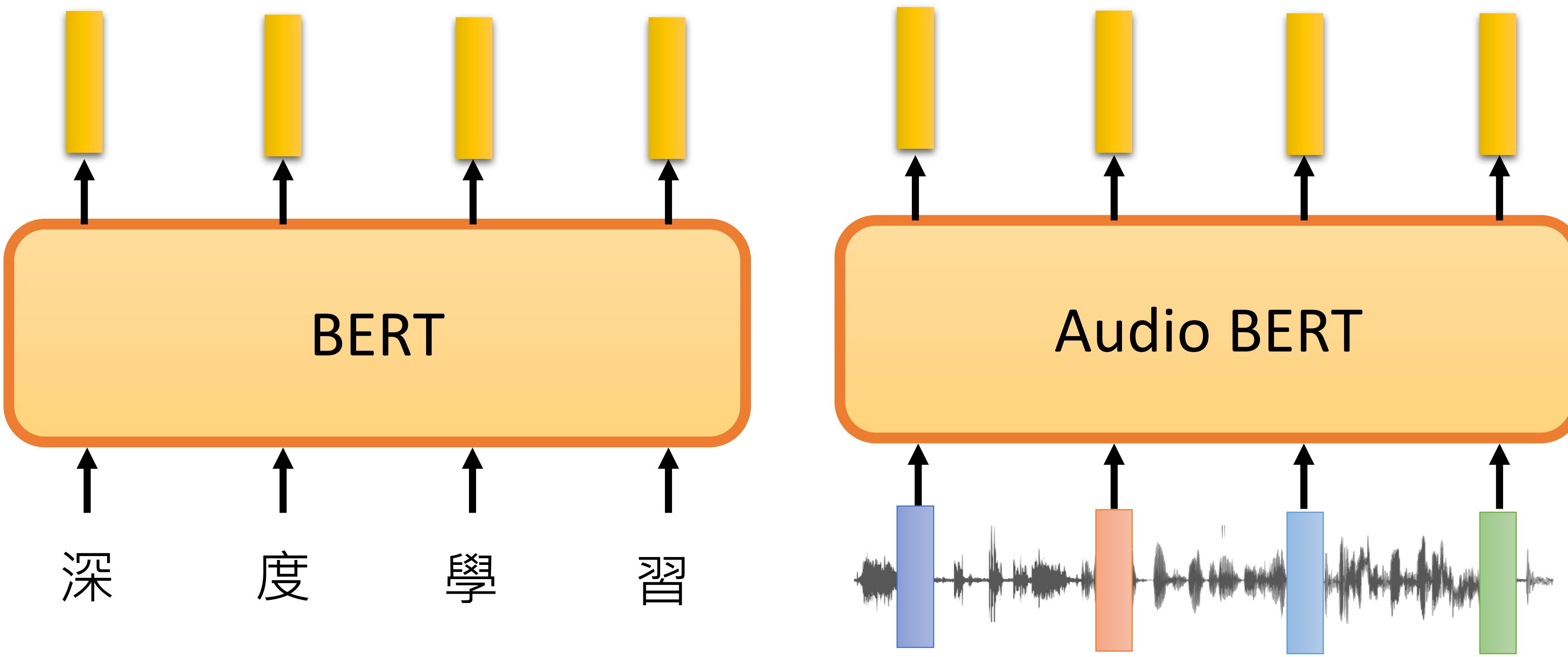
This is another story ……

- **E**nhanced **L**anguage **R**epresentatio**N** with **I**nformative **E**ntities (ERNIE)

+


# Audio BERT

This is another story ……



# Reference


- [Lewis, et al., arXiv'19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, arXiv, 2019
- [Raffel, et al., arXiv'19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, arXiv, 2019
- [Joshi, et al., TACL'20] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy, SpanBERT: Improving Pre-training by Representing and Predicting Spans, TACL, 2020
- [Song, et al., ICML'19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu, MASS: Masked Sequence to Sequence Pre-training for Language Generation, ICML, 2019
- [Zafrir, et al., NeurIPS workshop 2019] Ofir Zafrir, Guy Boudoukh, Peter Izsak, Moshe Wasserblat, Q8BERT: Quantized 8Bit BERT, NeurIPS workshop 2019

# Reference


- [Houlsby, et al., ICML'19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly, Parameter-Efficient Transfer Learning for NLP, ICML, 2019
- [Hao, et al., EMNLP'19] Yaru Hao, Li Dong, Furu Wei, Ke Xu, Visualizing and Understanding the Effectiveness of BERT, EMNLP, 2019
- [Liu, et al., arXiv'19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv, 2019
- [Sanh, et al., NeurIPS workshop's] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, NeurIPS workshop, 2019
- [Jian, et al., arXiv'19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu, TinyBERT: Distilling BERT for Natural Language Understanding, arXiv, 19

# Reference


- [Shoeybi, et al., arXiv'19]Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, Bryan Catanzaro, Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, arXiv, 19
- [Lan, et al., ICLR'20]Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, ICLR, 2020
- [Kitaev, et al., ICLR'20] Nikita Kitaev, Lukasz Kaiser, Anselm Levskaya, Reformer: The Efficient Transformer, ICLR, 2020
- [Beltagy, et al., arXiv'20] Iz Beltagy, Matthew E. Peters, Arman Cohan, Longformer: The Long-Document Transformer, arXiv, 2020
- [Dai, et al., ACL'19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov, Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, ACL, 2019
- [Peters, et al., NAACL'18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations, NAACL, 2018

# Reference


- [Sanh, et al., NeurIPS workshop's] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, NeurIPS workshop, 2019
- [Jian, et al., arXiv'19] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun Liu, TinyBERT: Distilling BERT for Natural Language Understanding, arXiv, 19
- [Sun, et al., ACL'20] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, Denny Zhou, MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, ACL, 2020
- [Zafrir, et al., NeurIPS workshop 2019] Ofir Zafrir, Guy Boudoukh, Peter Izsak, Moshe Wasserblat, Q8BERT: Quantized 8Bit BERT, NeurIPS workshop 2019
- [Sun, et al., ACL'20] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, Denny Zhou, MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices, ACL, 2020

# Reference


- [Pennington, et al., EMNLP'14] Jeffrey Pennington, Richard Socher, Christopher Manning, Glove: Global Vectors for Word Representation, EMNLP, 2014
- [Mikolov, et al., NIPS'13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, Distributed Representations of Words and Phrases and their Compositionality, NIPS, 2013
- [Bojanowski, et al., TACL'17] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, Enriching Word Vectors with Subword Information, TACL, 2017
- [Su, et al., EMNLP'17] Tzu-Ray Su, Hung-Yi Lee, Learning Chinese Word Representations From Glyphs Of Characters, EMNLP, 2017
- [Liu, et al., ACL'19] Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Multi-Task Deep Neural Networks for Natural Language Understanding, ACL, 2019
- [Stickland, et al., ICML'19] Asa Cooper Stickland, Iain Murray, BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning, ICML, 2019

# Reference


- [Howard, et al., ACL’18] Jeremy Howard, Sebastian Ruder, Universal Language Model Fine-tuning for Text Classification, ACL, 2018
- [Alec, et al., 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, Improving Language Understanding by Generative Pre-Training, 2018
- [Devlin, et al., NAACL’19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL, 2019
- [Alec, et al., 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, Language Models are Unsupervised Multitask Learners, 2019
- [Want, et al., ICLR’20] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, Luo Si, StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding, ICLR, 2020
- [Yang, et al., NeurIPS’19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, NeurIPS, 2019

# Reference


- [Cui, et al., arXiv’19] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, Guoping Hu, Pre-Training with Whole Word Masking for Chinese BERT, arXiv, 2019
- [Sun, et al., ACL’19] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu, ERNIE: Enhanced Representation through Knowledge Integration, ACL, 2019
- [Dong, et al., NeurIPS’19] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon, Unified Language Model Pre-training for Natural Language Understanding and Generation, NeurIPS, 2019