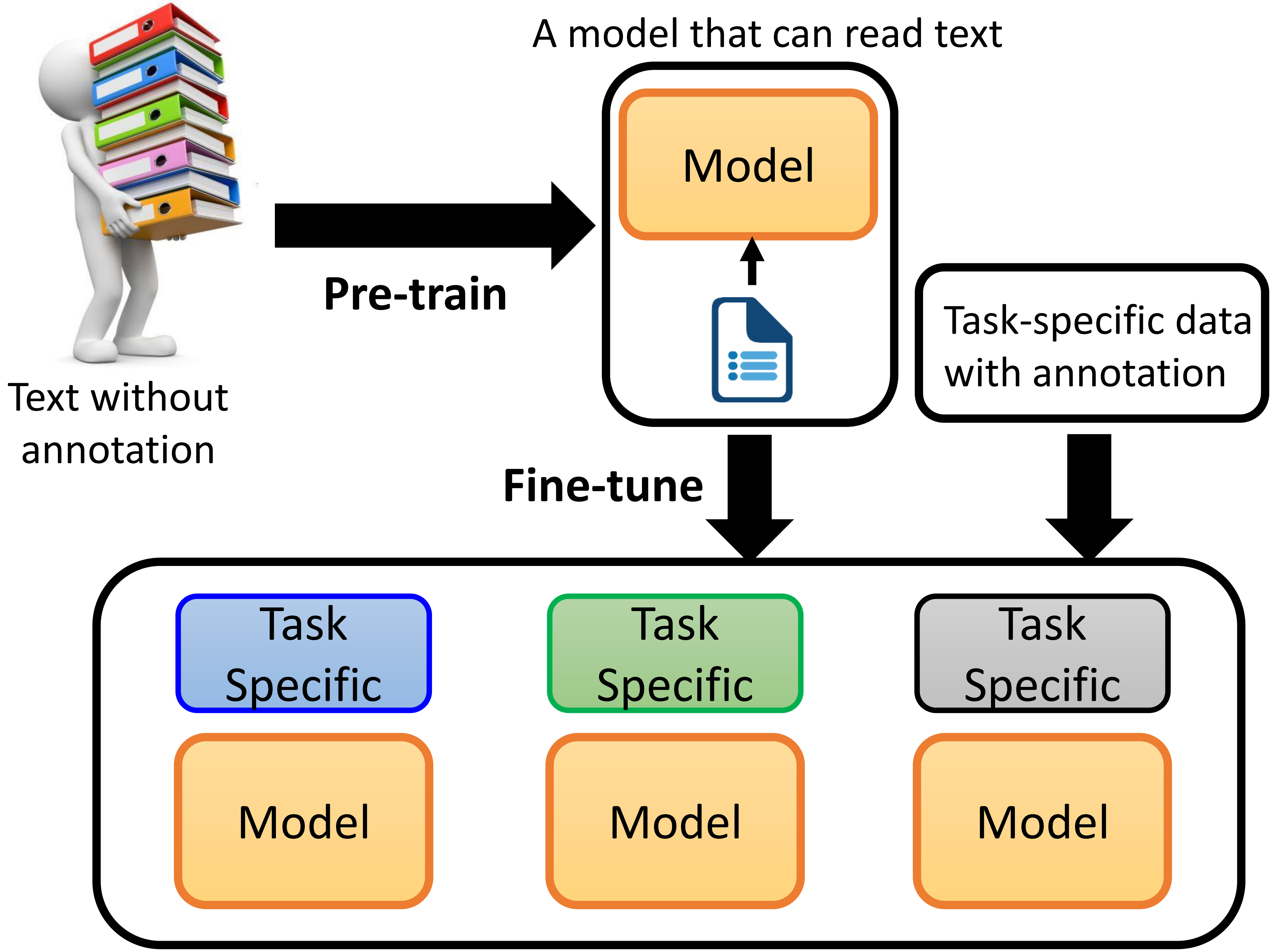
BERT and its family

Hung-yi Lee 李宏毅



A model that can read text

Model

Pre-train

Text without annotation

Task-specific data with annotation

Fine-tune

Task Specific

Model

Task Specific

Model

Task Specific

Model

死臭酸宅本人

芝麻街

BPON
PAY ME
PAY ME
PAY ME
PAY ME
BEST FRIENDS FOREVER
STAYREAL

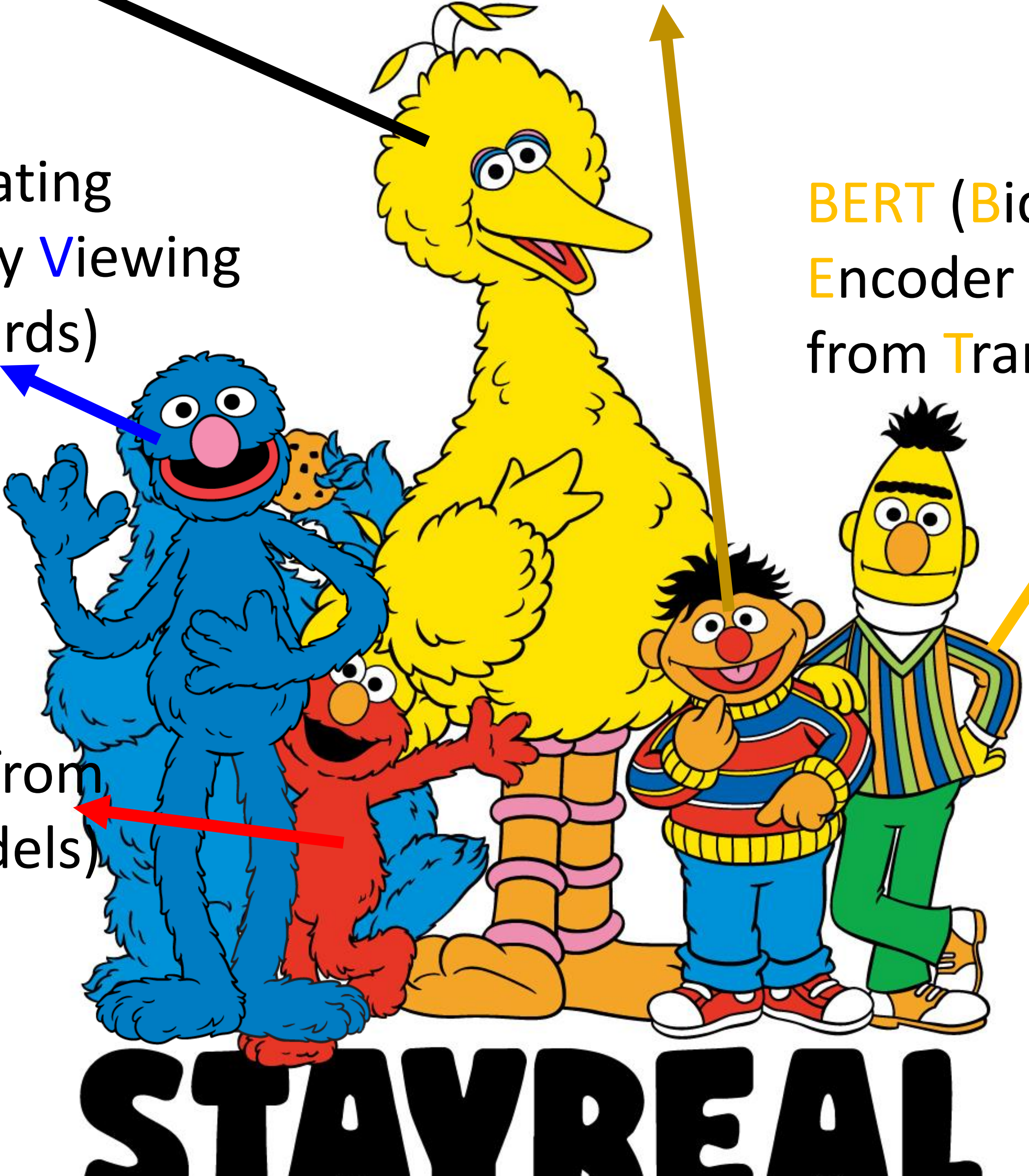
Big Bird

Big **Bi**nary **R**ecursive **D**ecoder?

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)

BERT & PALs (**P**rojected **A**ttention **L**ayers)

Outline

What is pre-train model

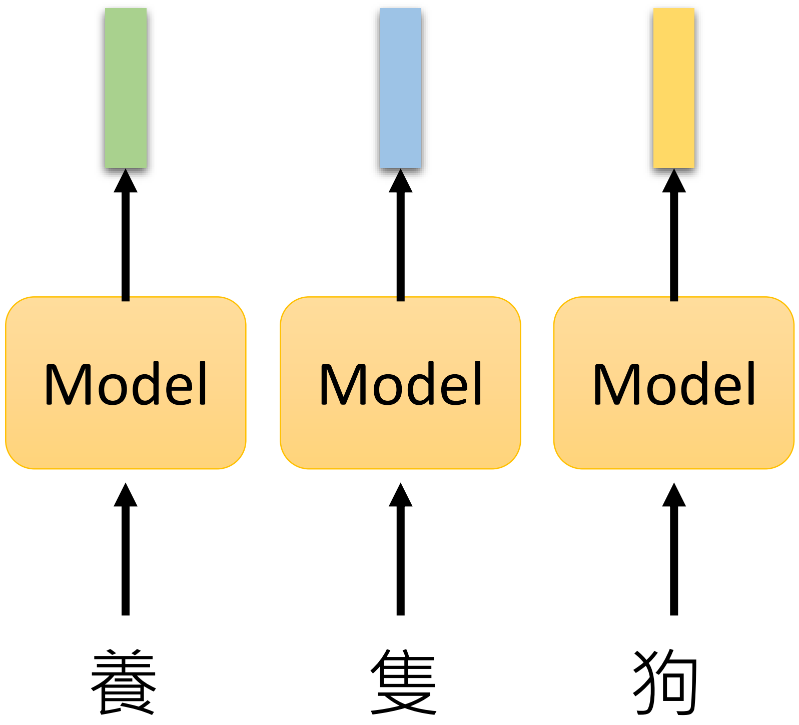
How to fine-tune

How to pre-train

Pre-train Model

Pre-train Model

Represent each token by a embedding vector

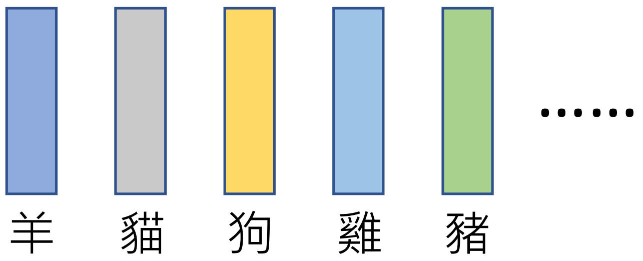


Model Model Model

養 隻 狗

The token with the same type has the same embedding.

Simply a table look-up



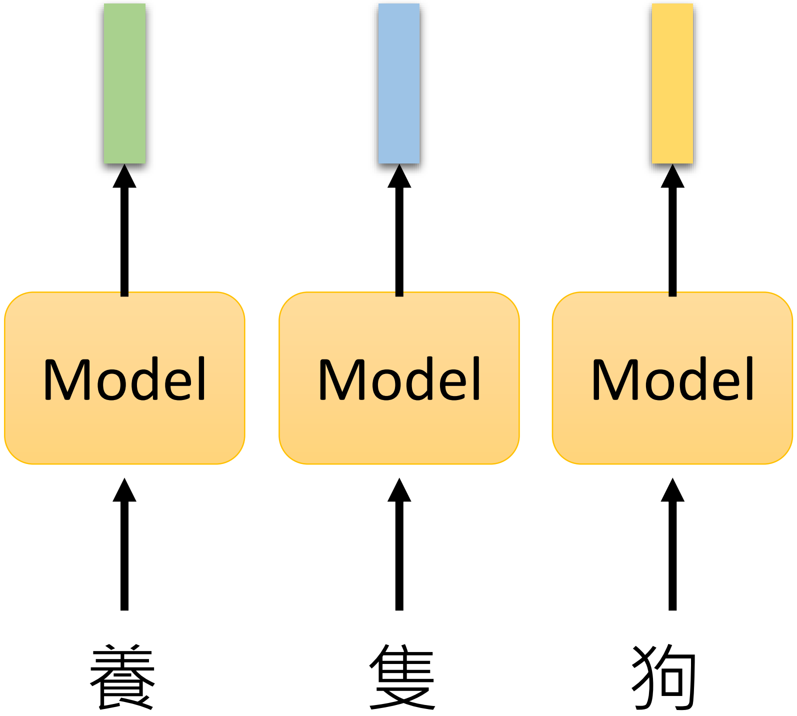
羊 貓 狗 雞 豬 ......

Word2vec [Mikolov, et al., NIPS’13]

Glove [Pennington, et al., EMNLP’14]

Pre-train Model

Represent each token by a embedding vector

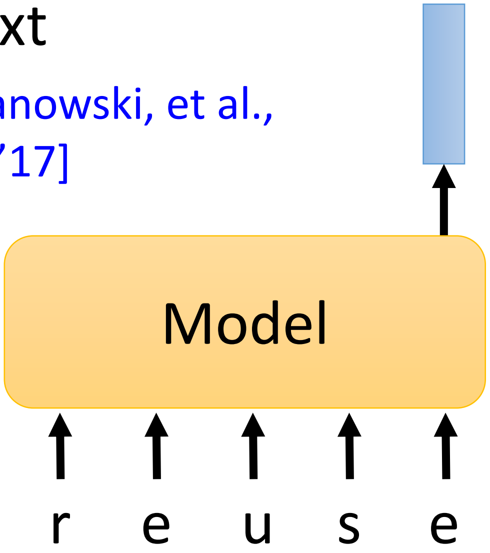


The token with the same type has the same embedding.

English word as token ...

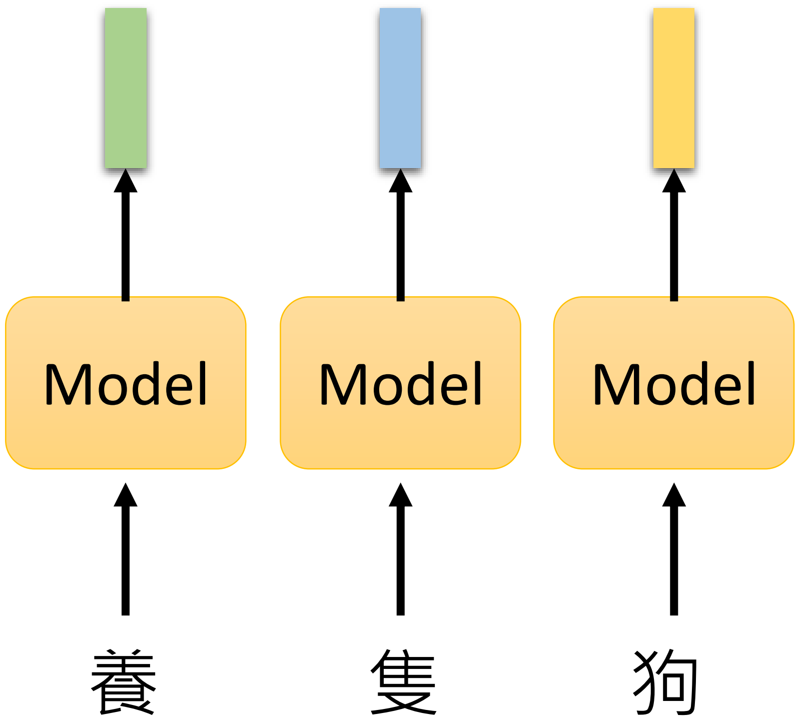
FastText

[Bojanowski, et al., TACL’17]



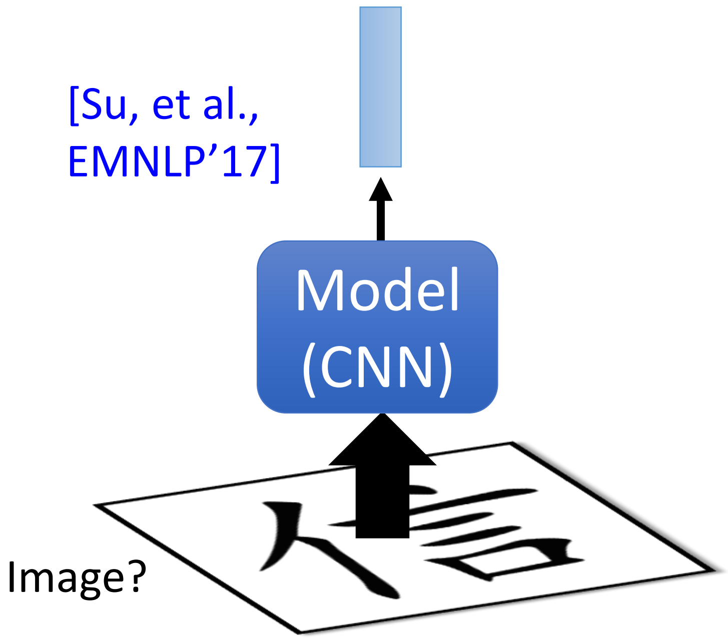
Pre-train Model

Represent each token by a embedding vector



The token with the same type has the same embedding.

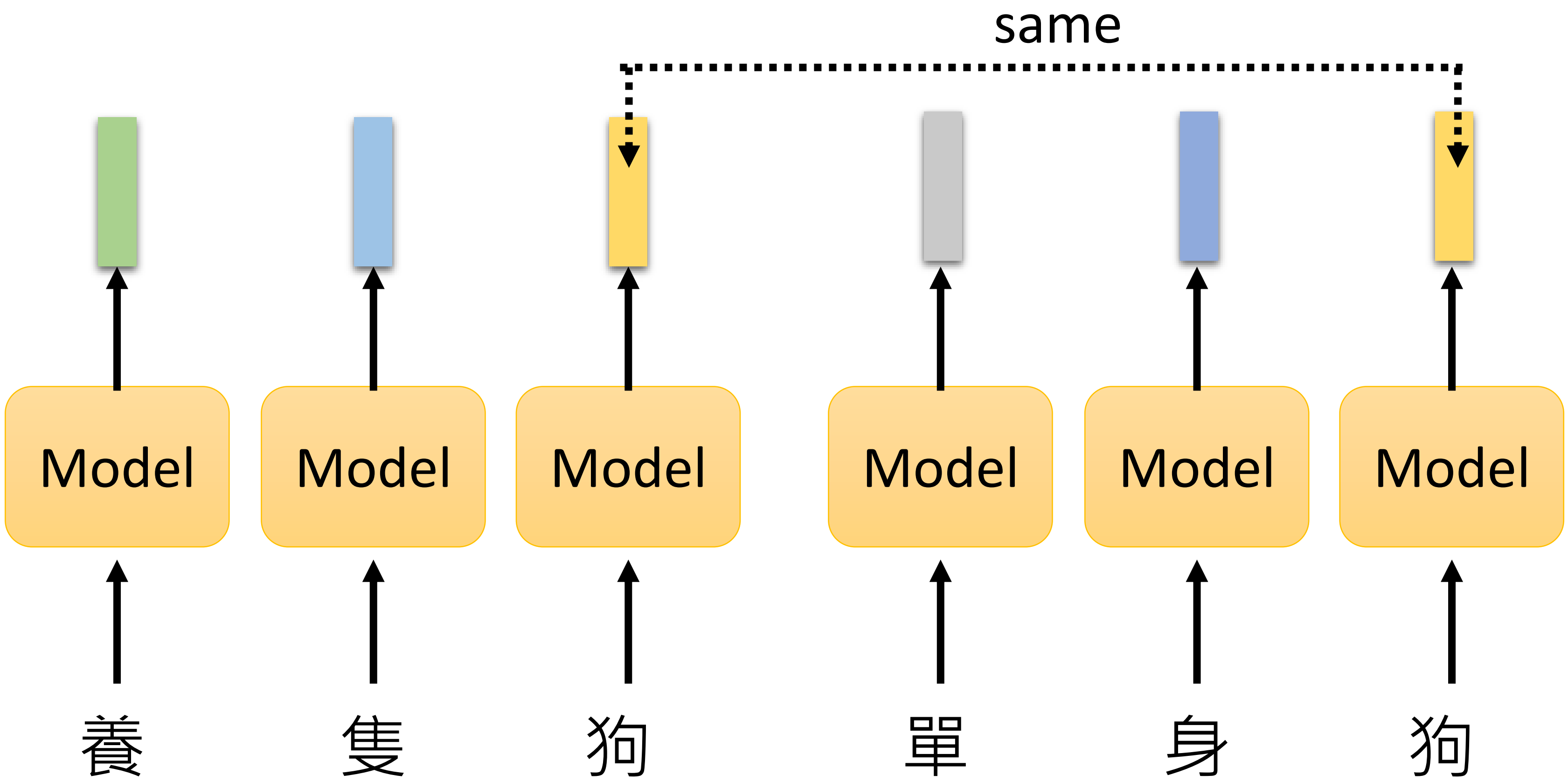
Chinese character as token …



[Su, et al., EMNLP’17]

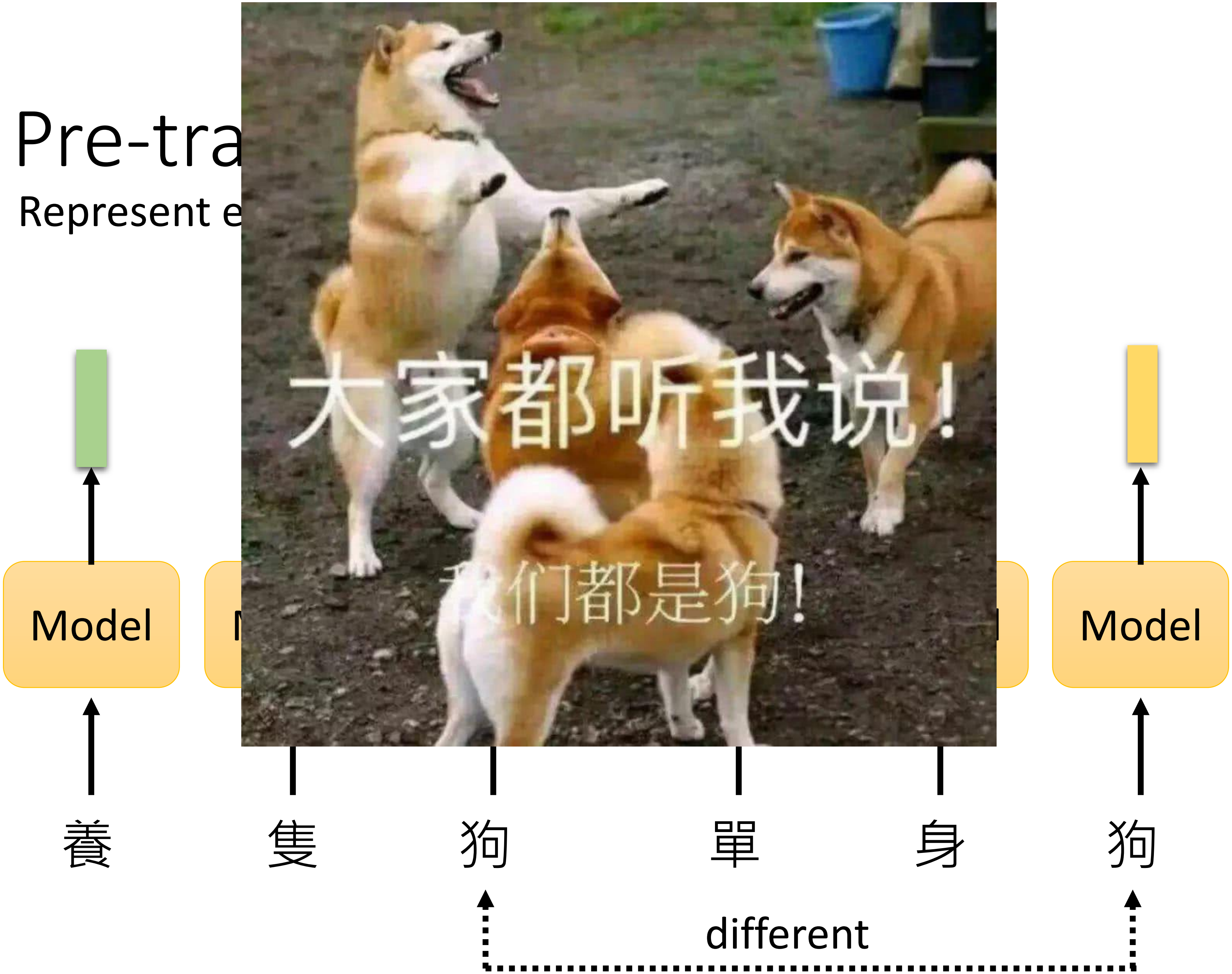
Pre-train Model

Represent each token by a embedding vector



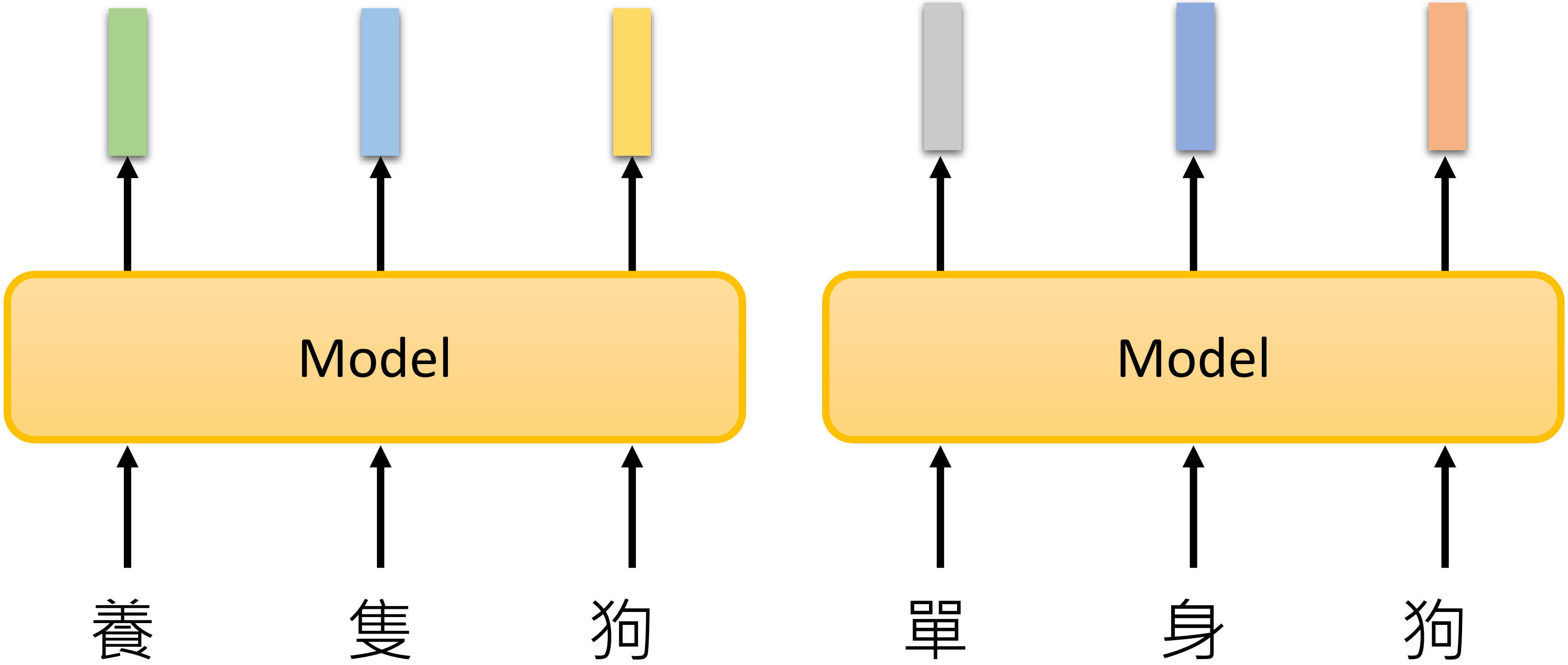
Pre-tra

Represent e



Pre-train Model

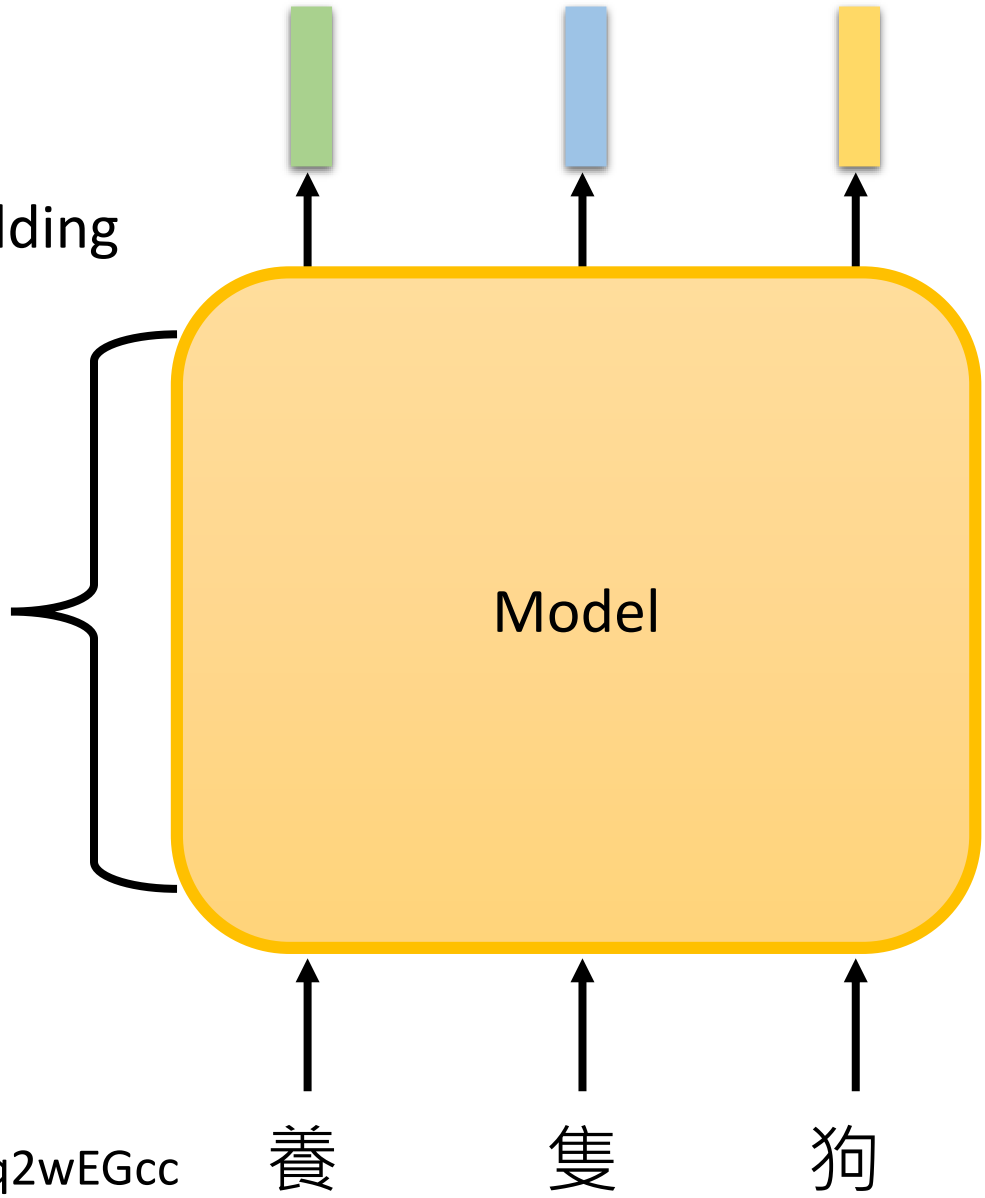
Contextualized Word Embedding



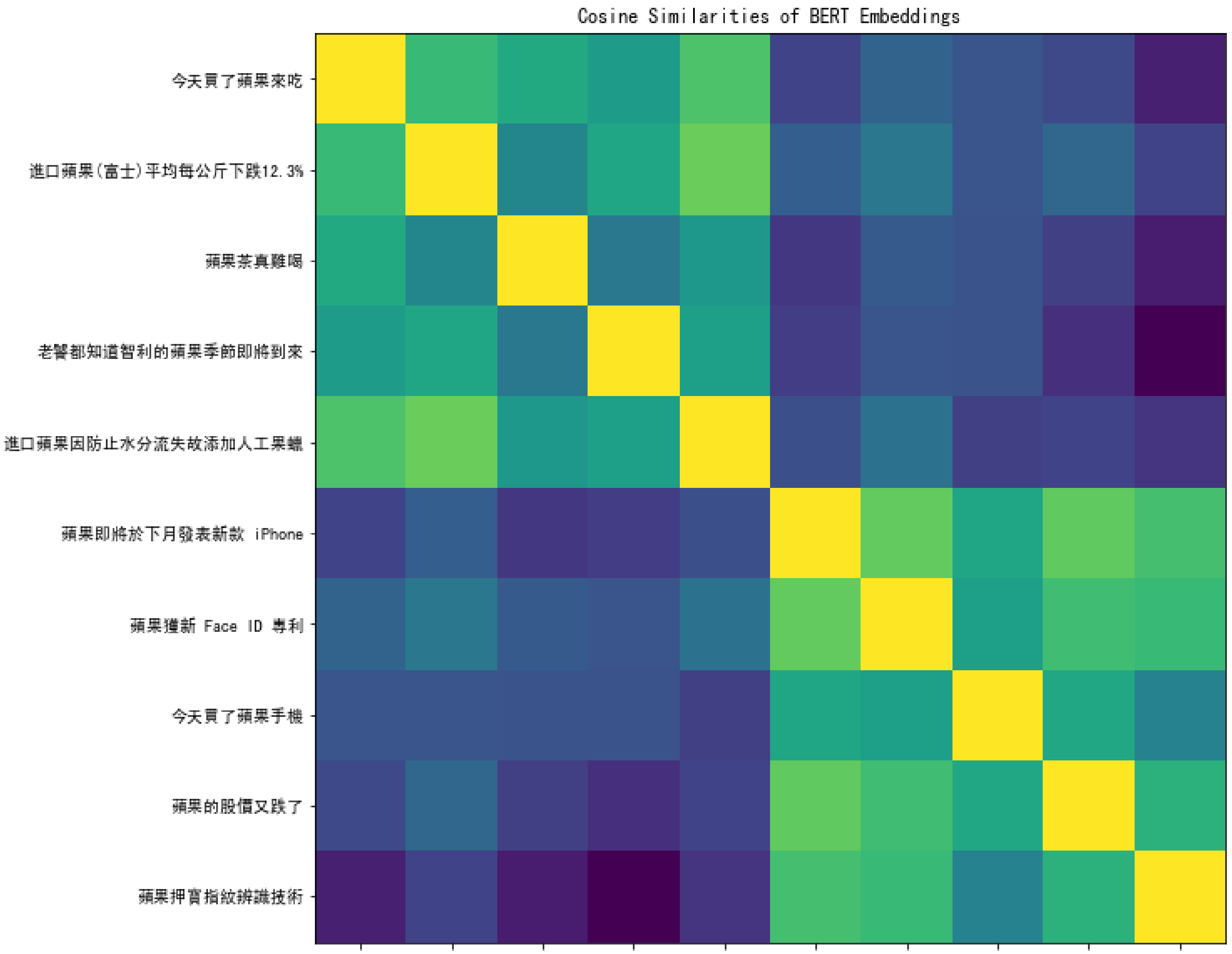
Pre-train Model

Contextualized Word Embedding

Many Layers

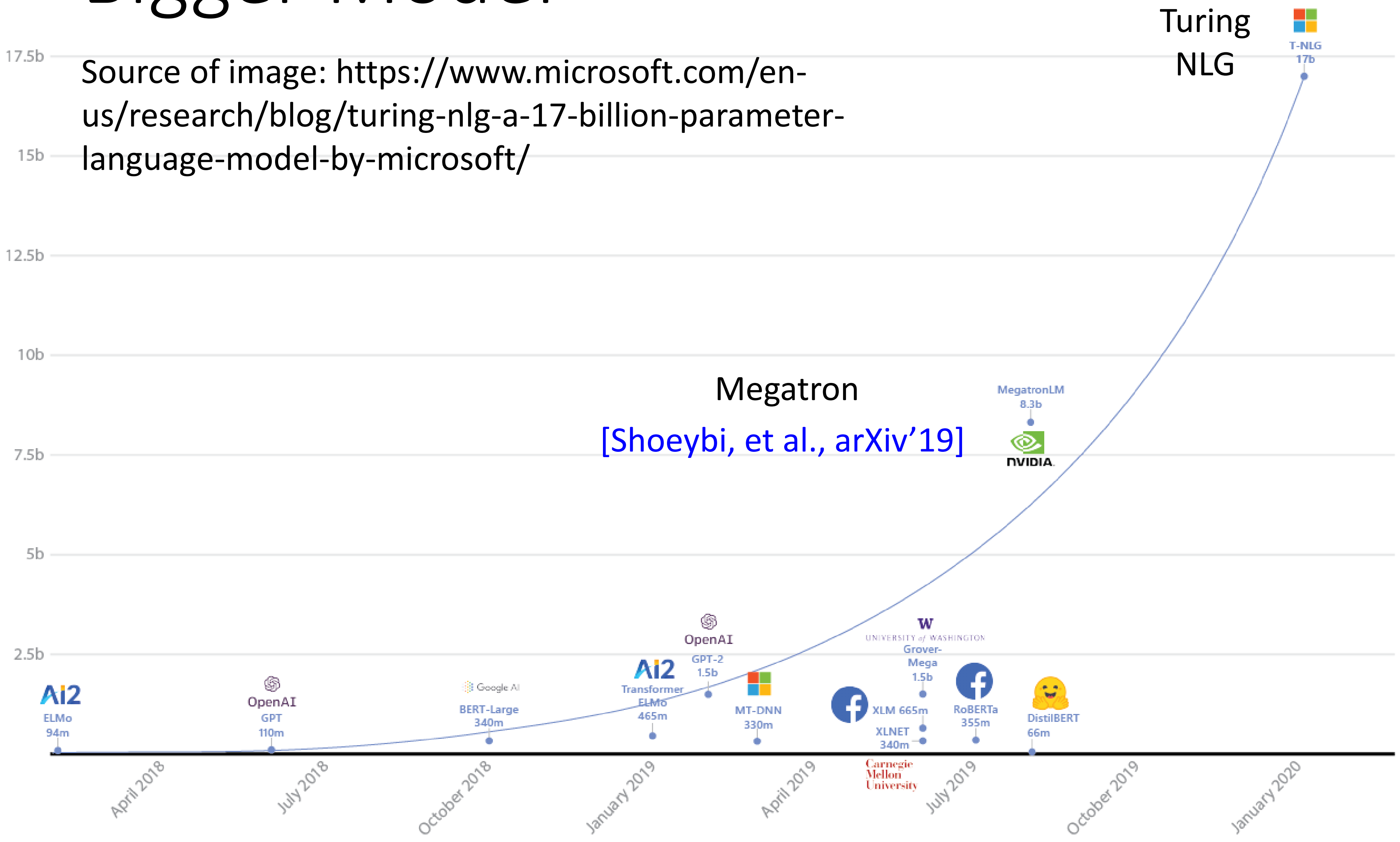


- LSTM
- Self-attention layers
- Tree-based model (?)
  - Ref: https://youtu.be/z0uOq2wEGcc


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

Bigger Model

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

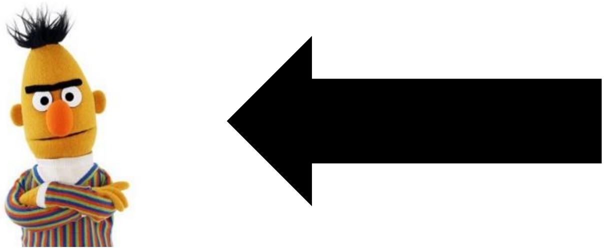


Turing NLG

Megatron

[Shoeybi, et al., arXiv’19]

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

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

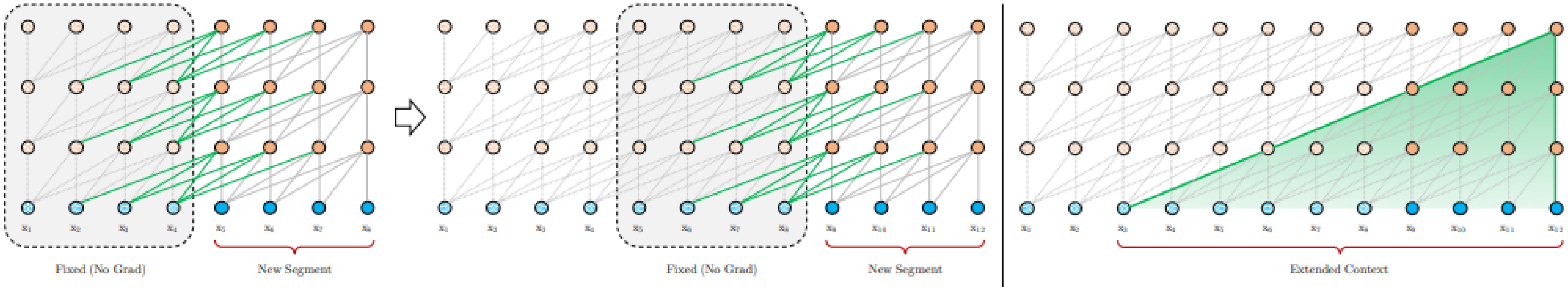
Ref: https://youtu.be/dPp8rCAnU_A

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]



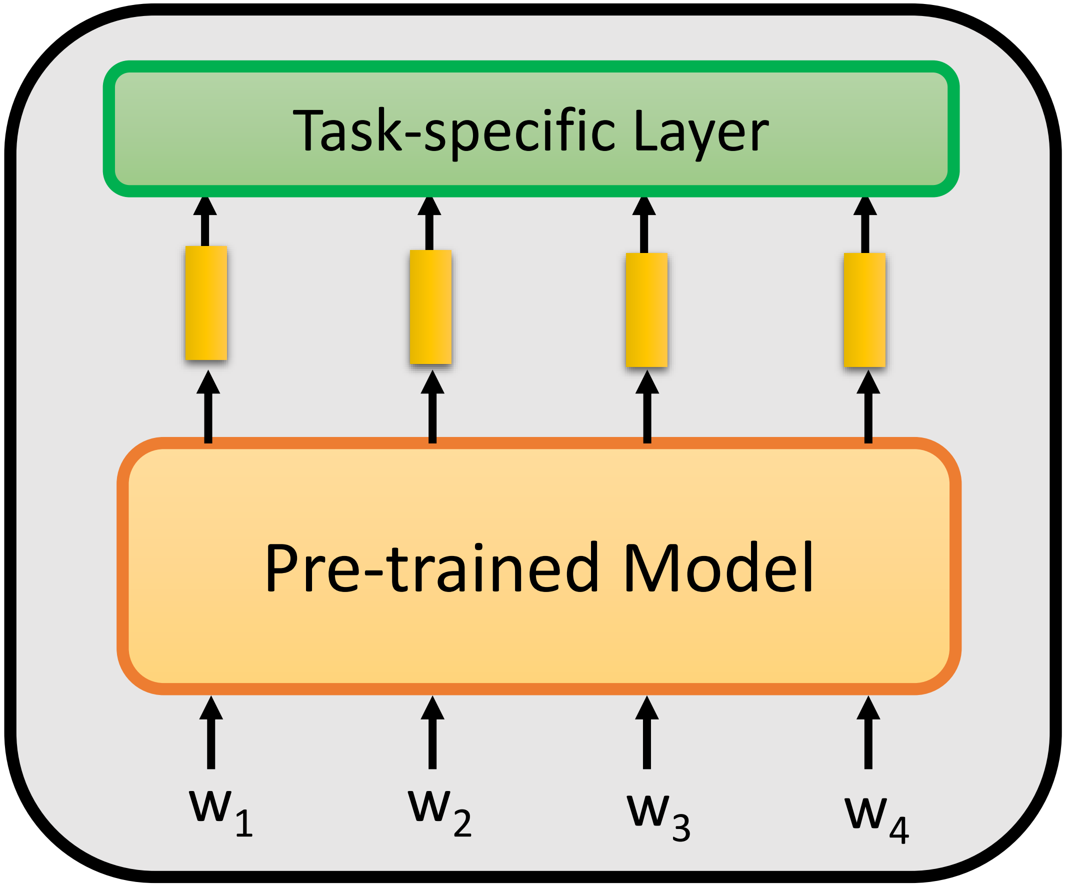
(a) Training phase.

(b) Evaluation phase.

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

Reduce the complexity of self-attention

How to fine-tune



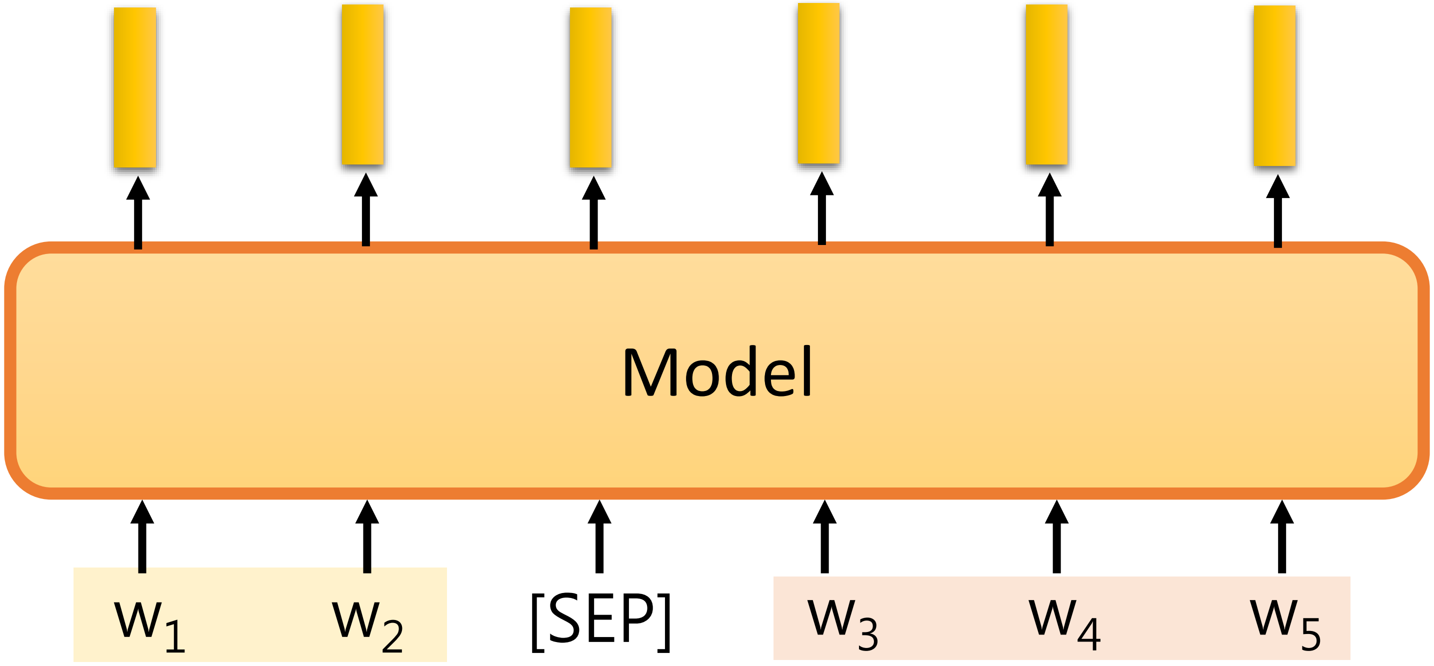
For a specific NLP task

NLP tasks

- Input
  - one sentence
  - multiple sentences
- Output
  - one class
  - class for each token
  - copy from input
  - general sequence

Input

- one sentence
- multiple sentences



Model

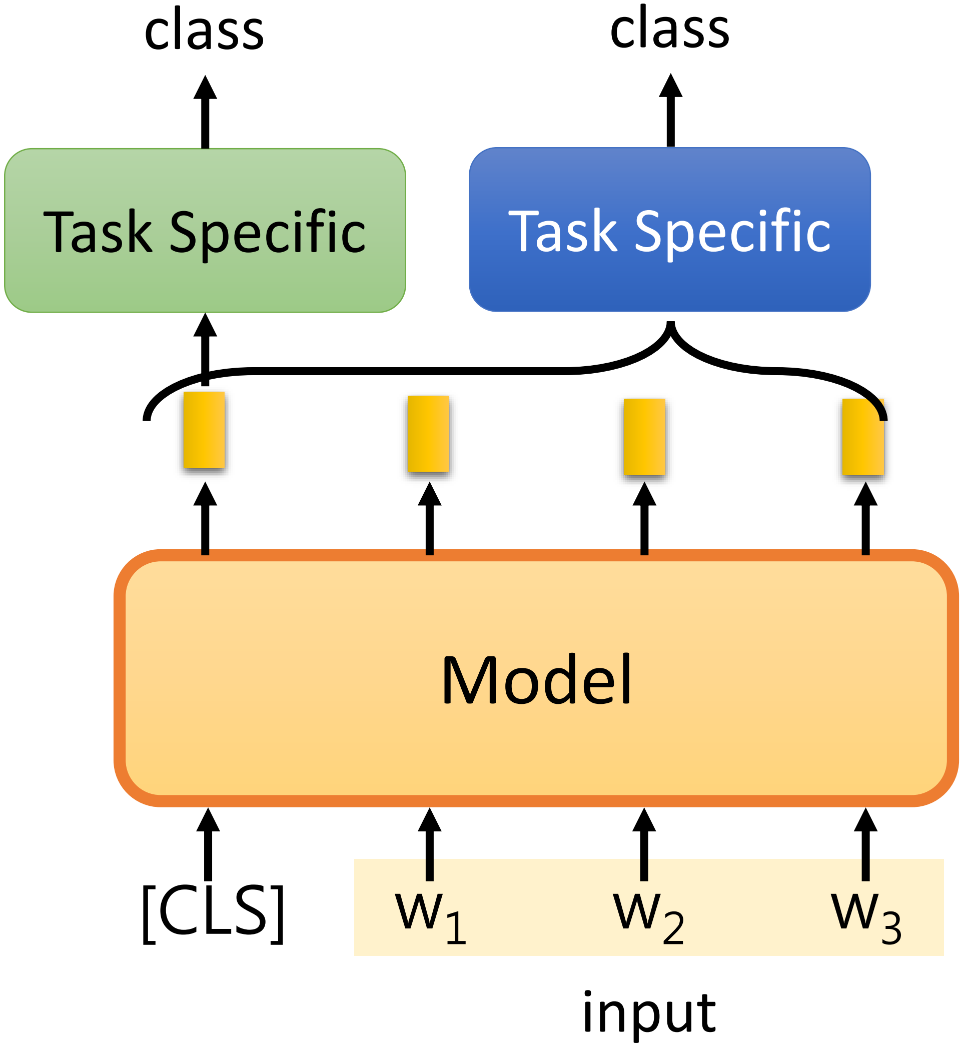
w_1 w_2 [SEP] w_3 w_4 w_5

Sentence 1 — Sentence 2

Query — Document

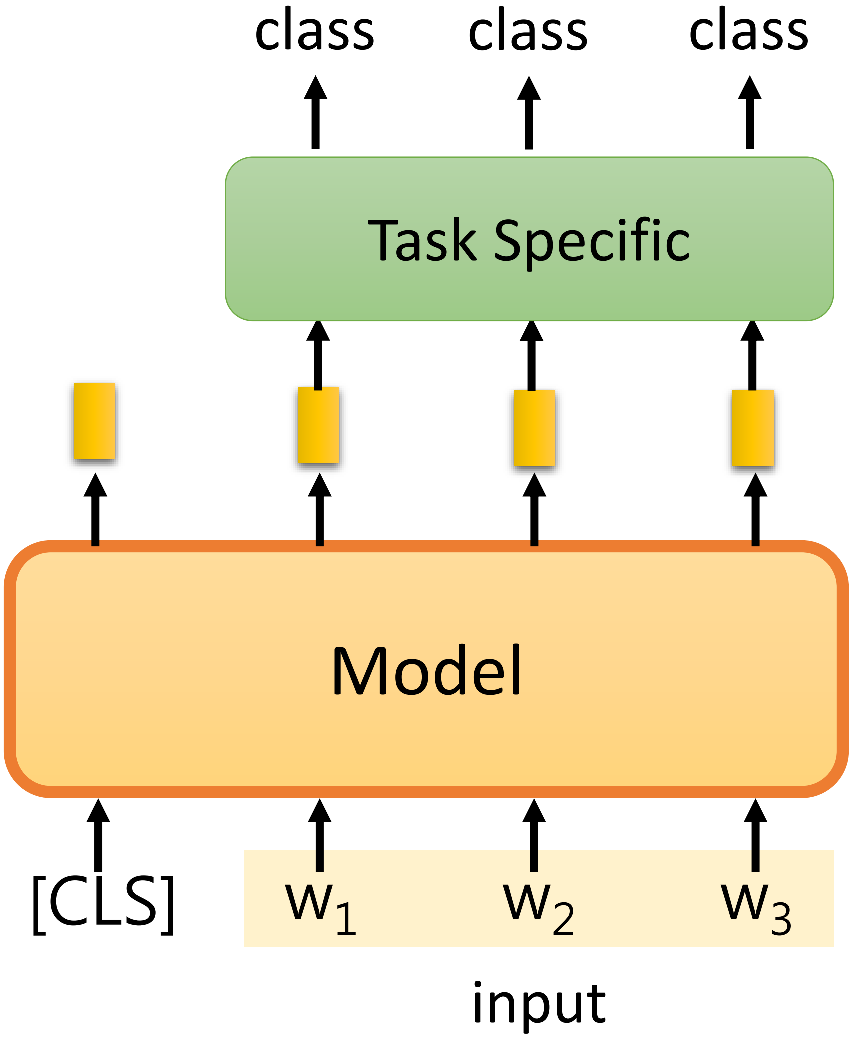
Premise — Hypothesis

Output



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

Output



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

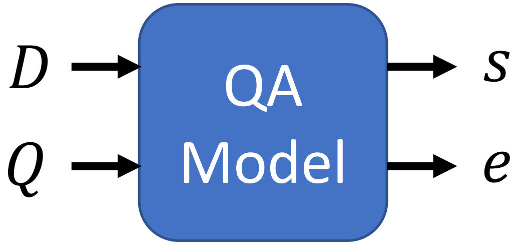
Output

- Extraction-based QA

- one class
- class for each token
- $\boxed{\text{copy from input}}$
- general sequence

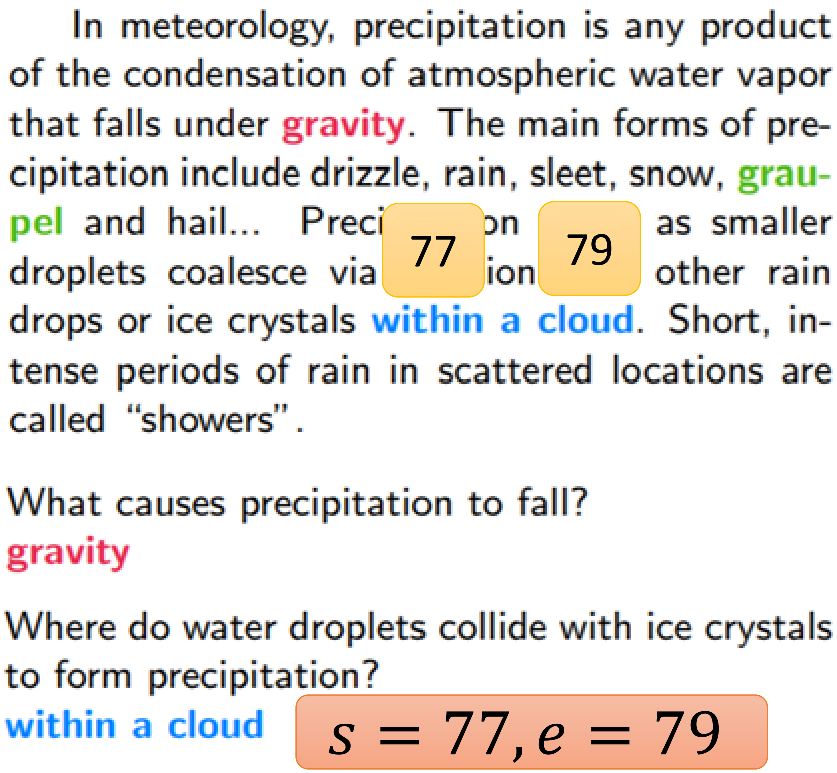
Document: $D = \{d_1, d_2, \cdots, d_N\}$

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



output: two integers (s, e)

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



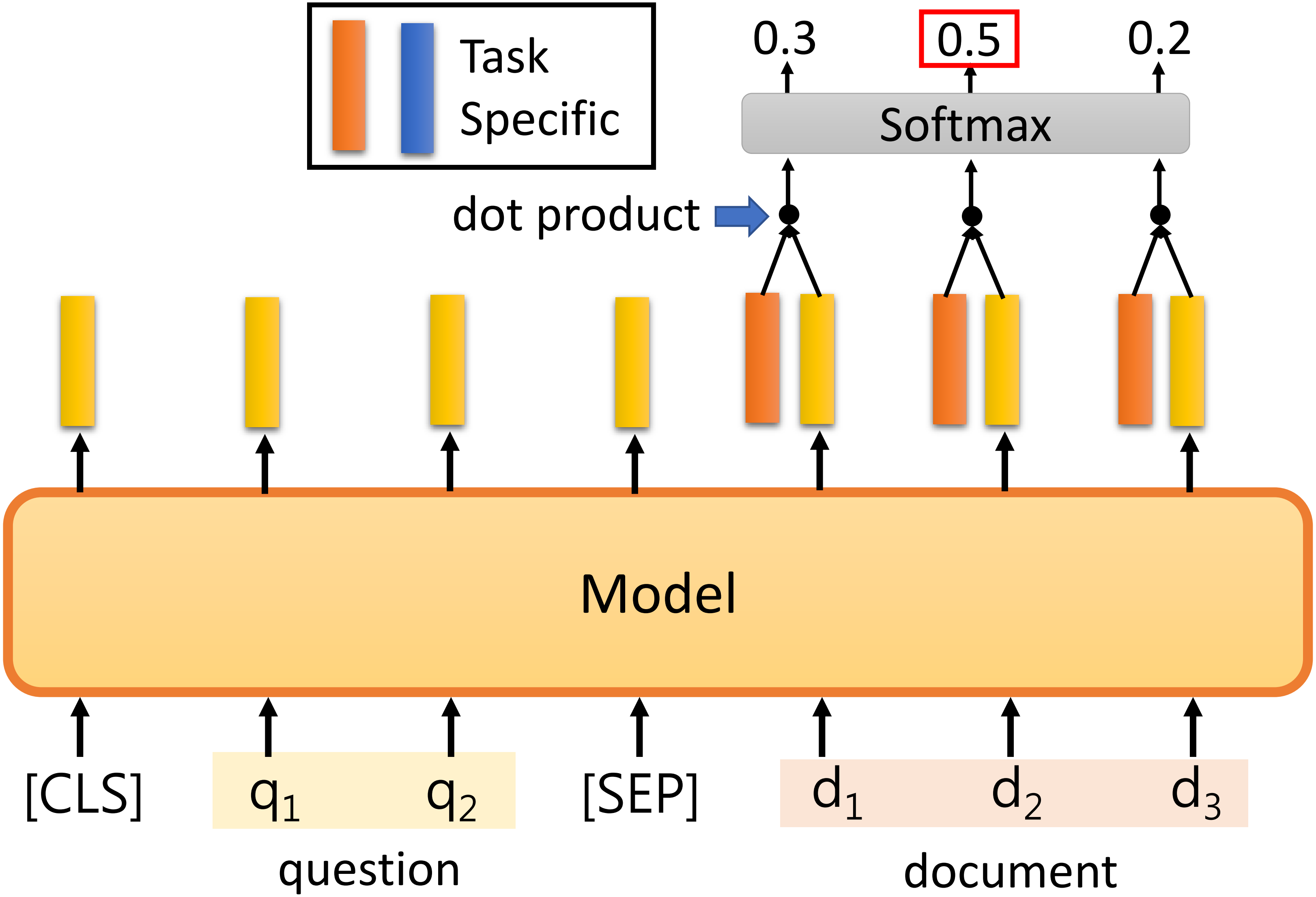
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Preci 77 on 79 as smaller droplets coalesce via ion other rain drops or ice crystals **within 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)

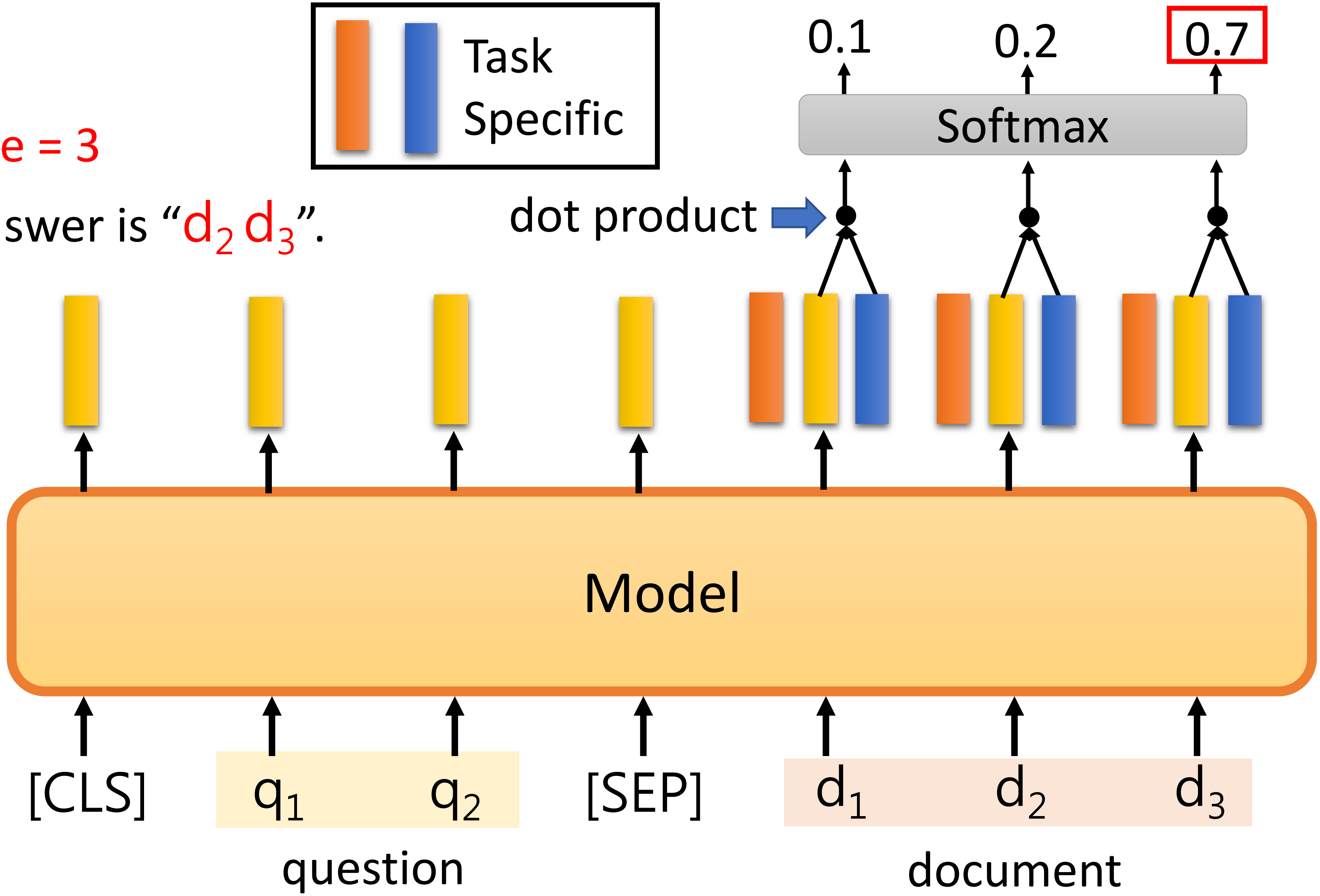
s = 2



Copy from Input (BERT)

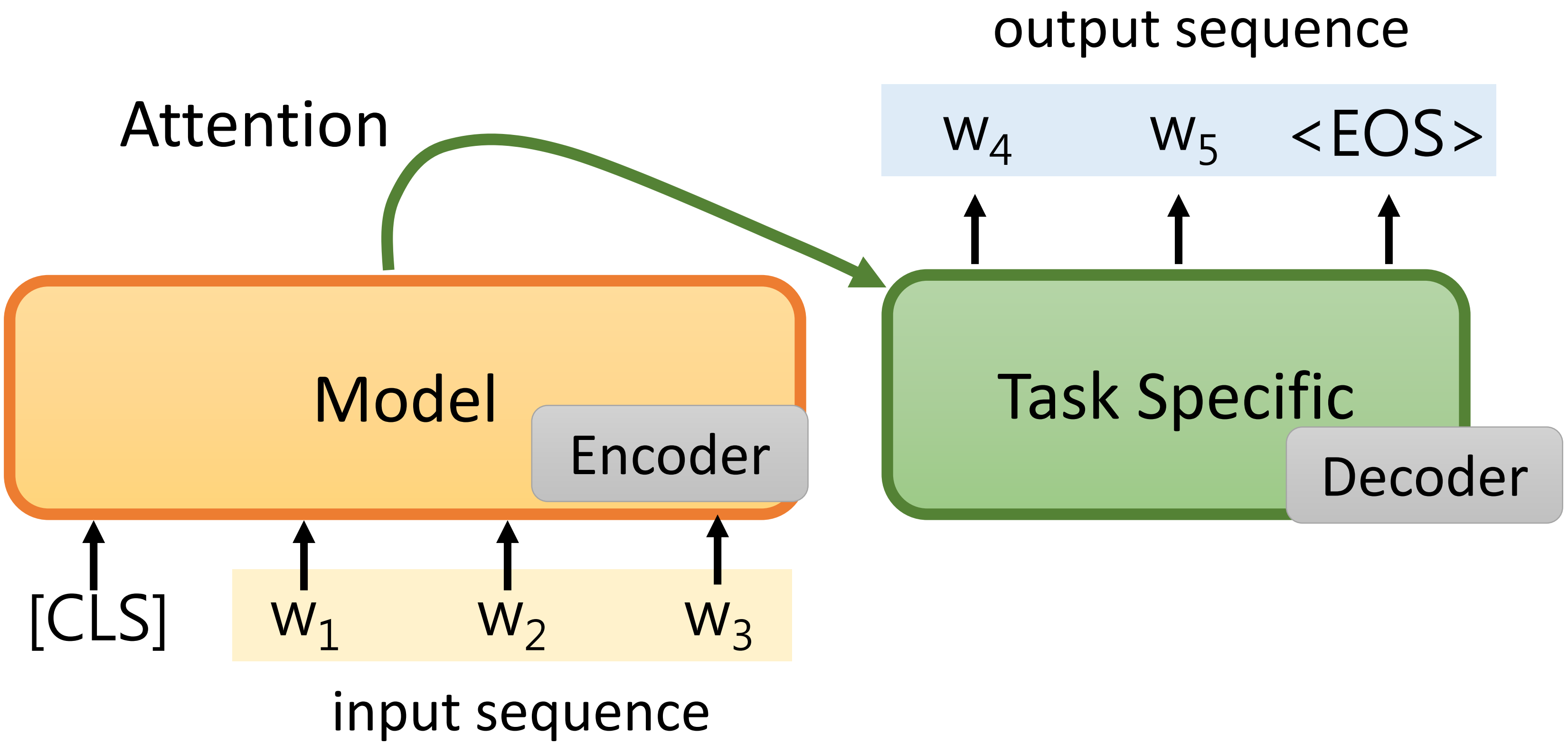
$s = 2 \quad e = 3$

The answer is “$d_2\ d_3$”.

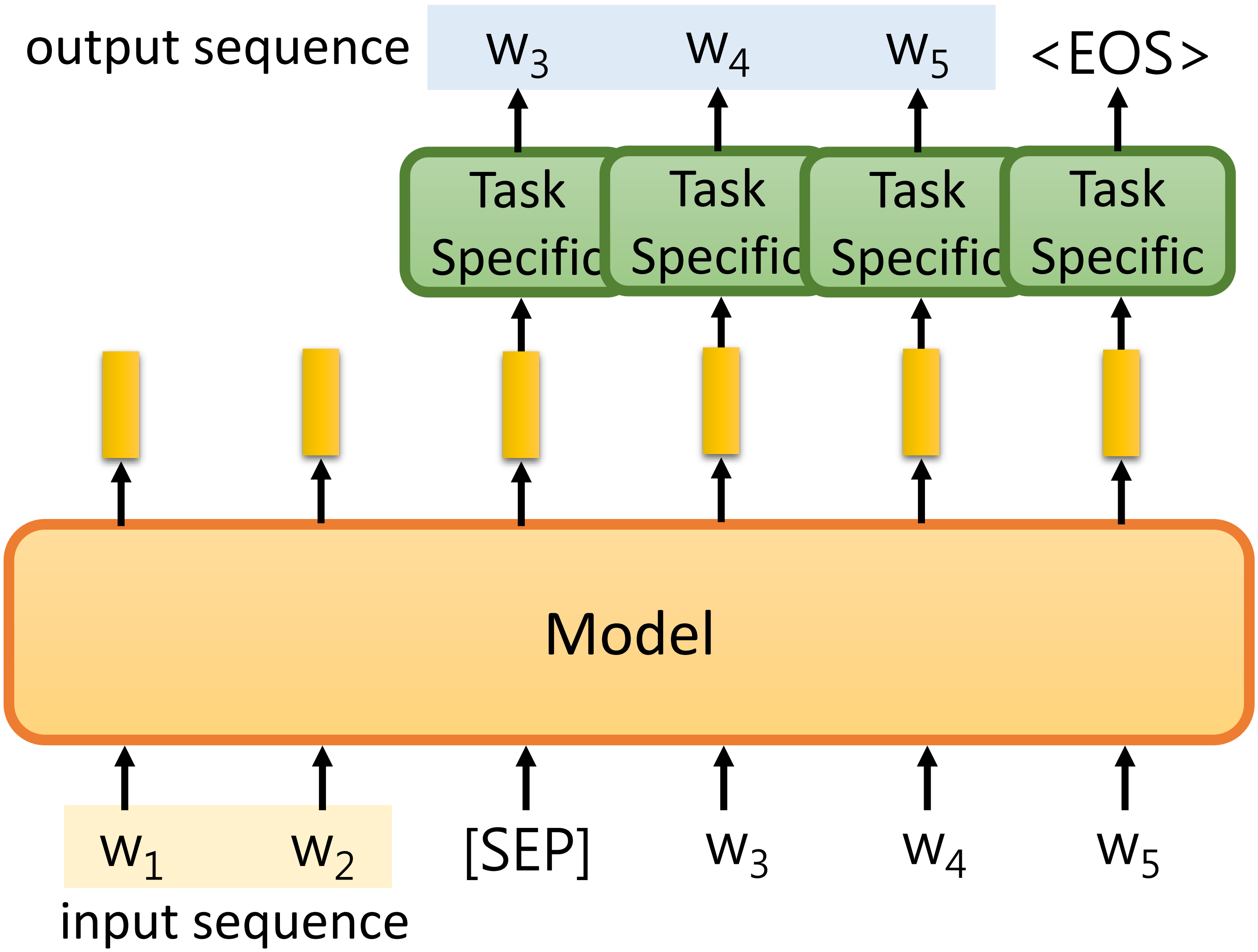


Output – General Sequence (v1)

- Seq2seq model



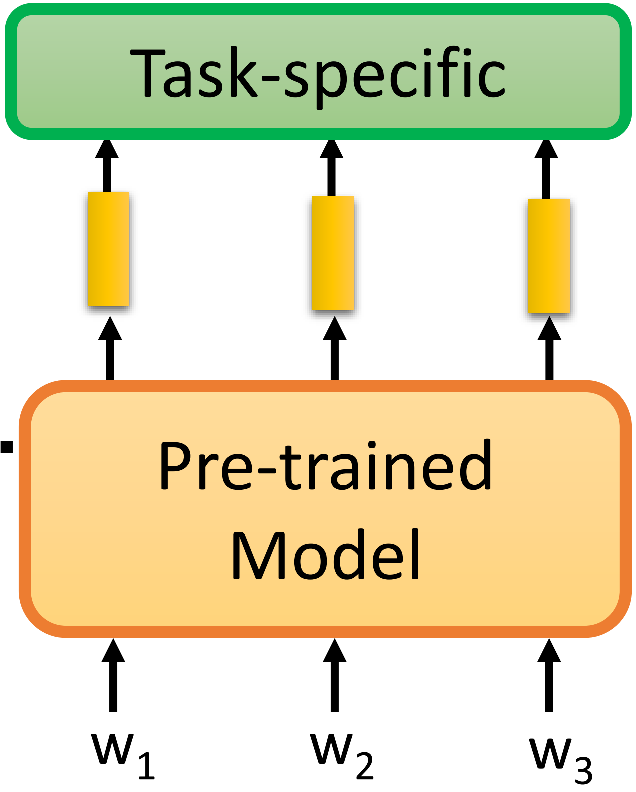
Output – General Sequence (v2)

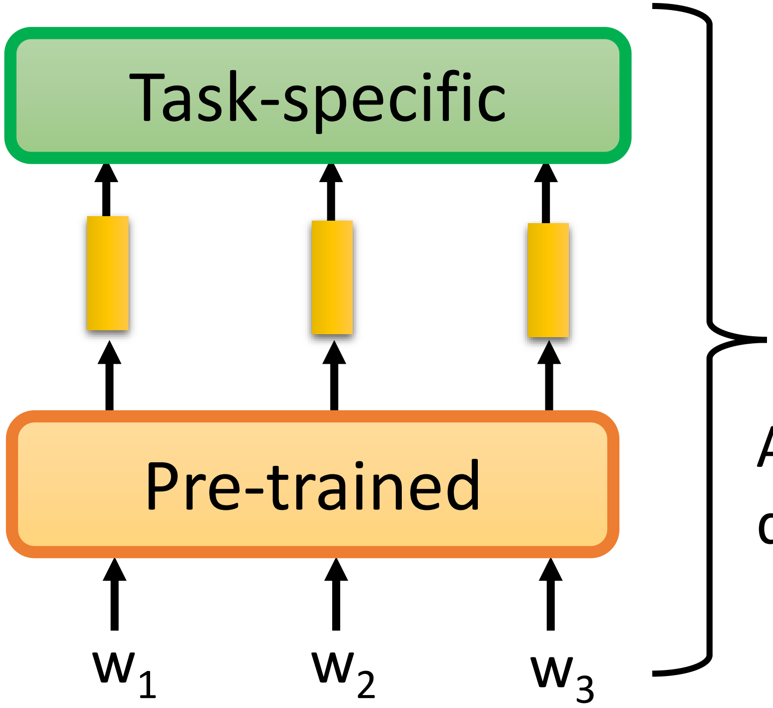


How to fine-tune

Fine-tune

Feature Extractor (Fix)



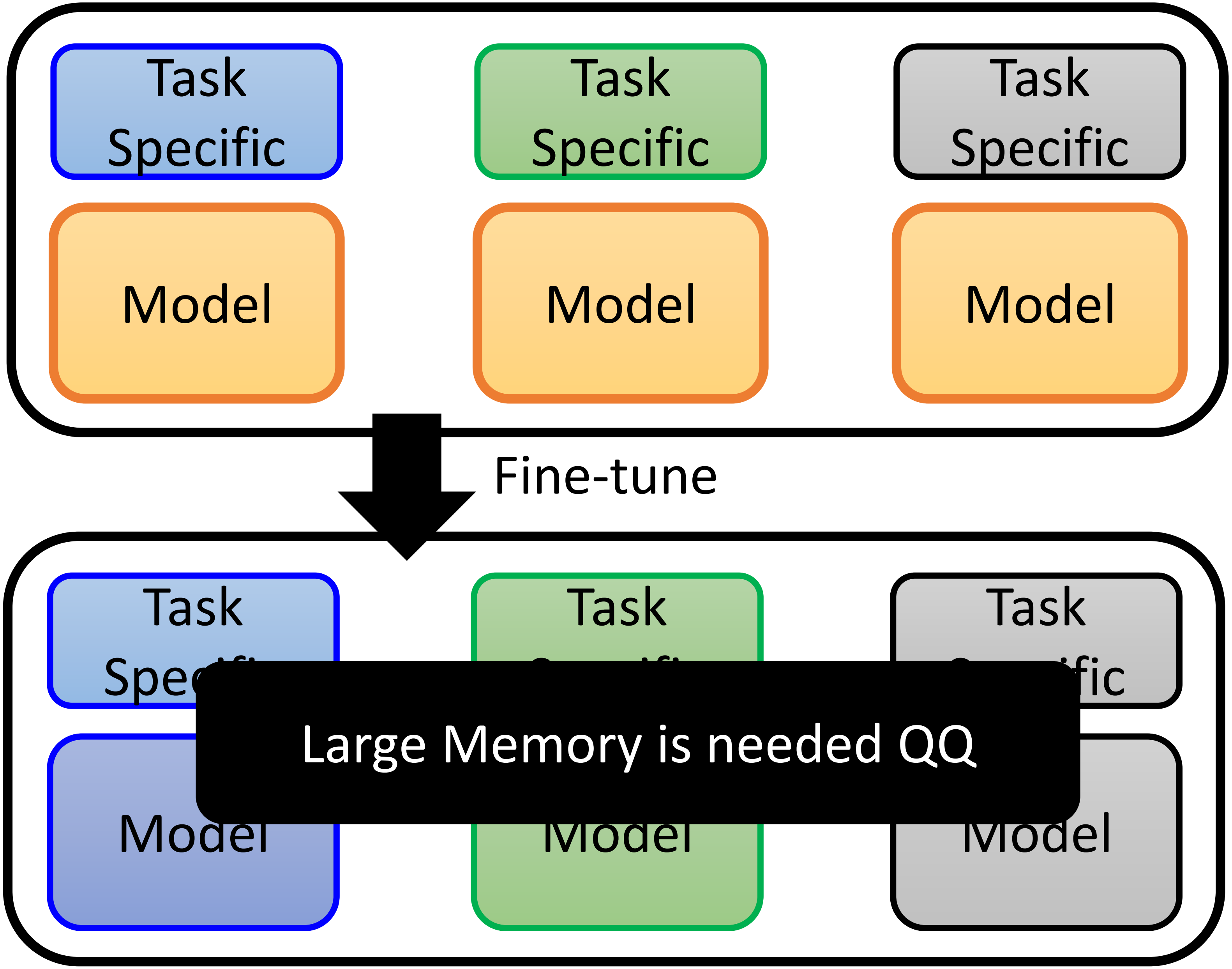


Fine-tune

A gigantic model for down-stream tasks

Adaptor

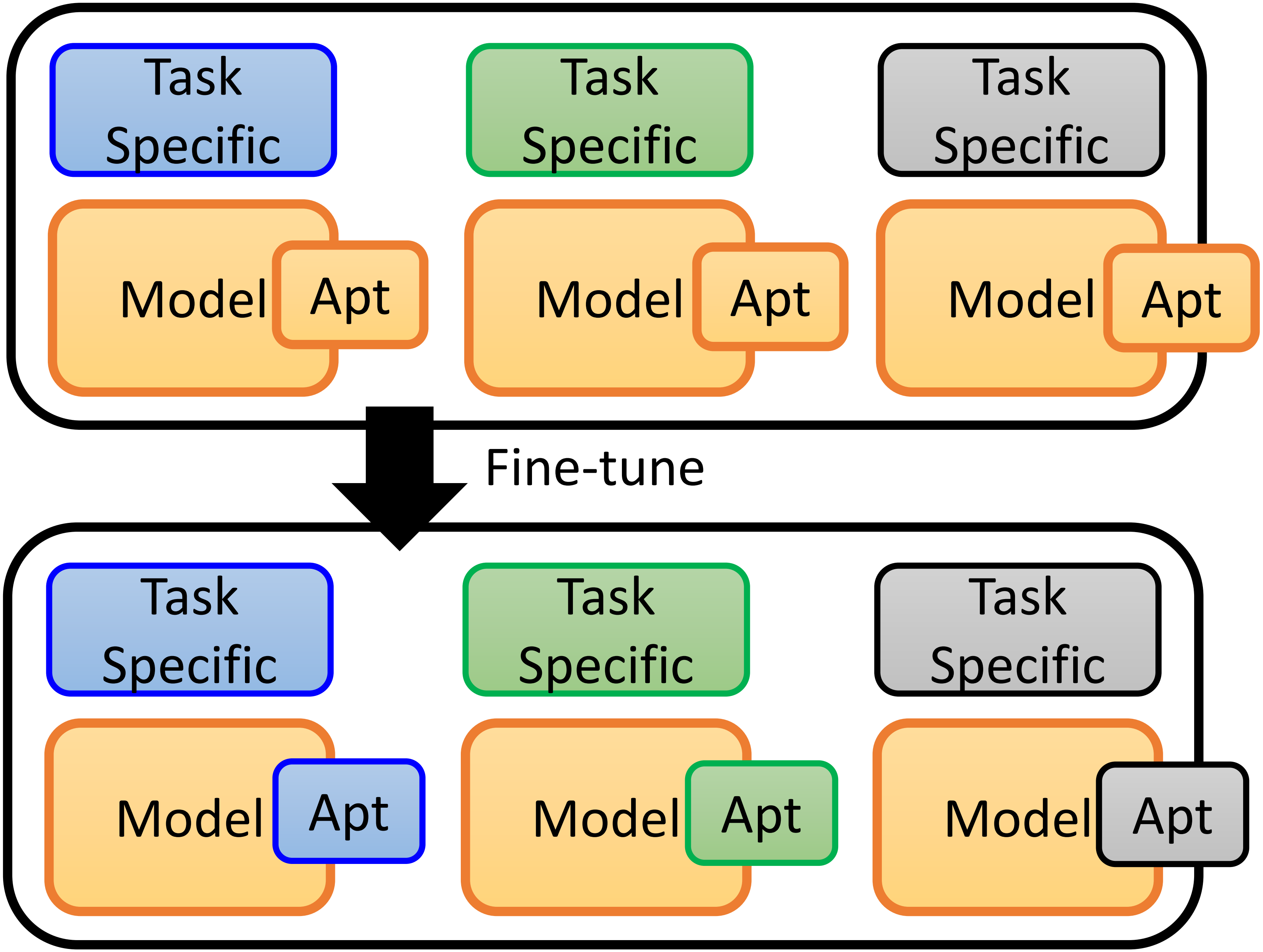
[Stickland, et al., ICML'19] [Houlsby, et al., ICML'19]

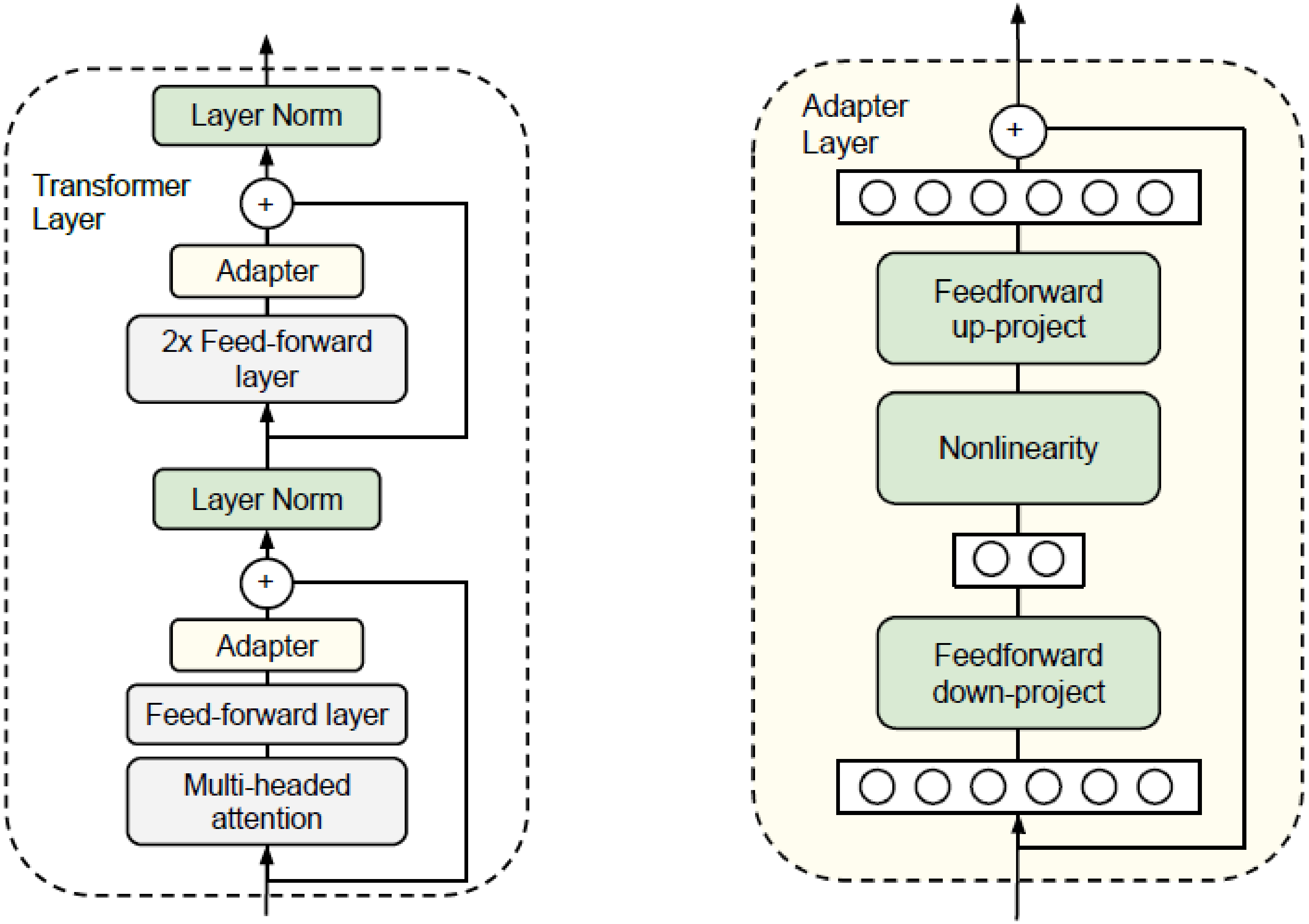


Large Memory is needed QQ

Adaptor

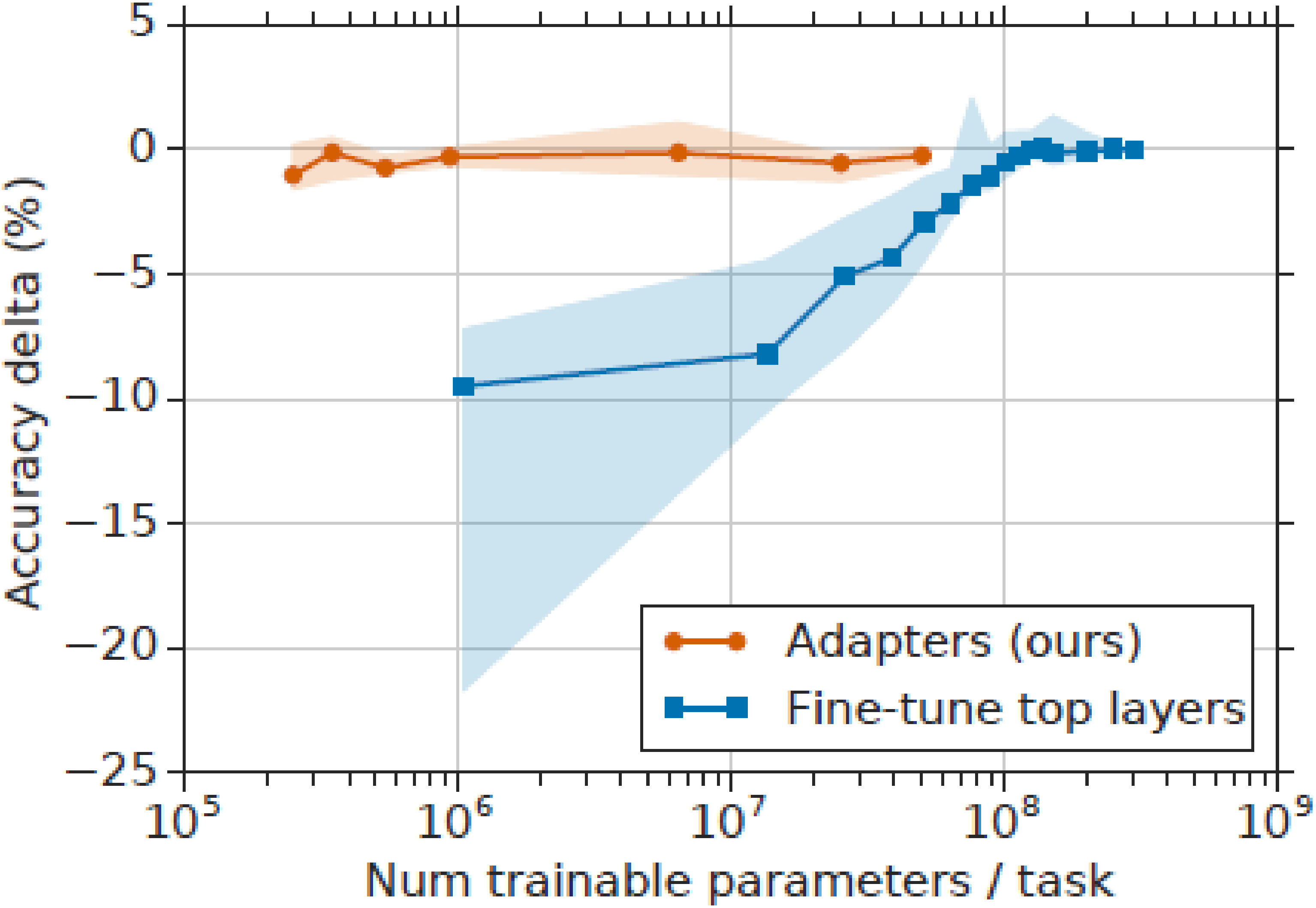
[Stickland, et al., ICML'19] [Houlsby, et al., ICML'19]





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

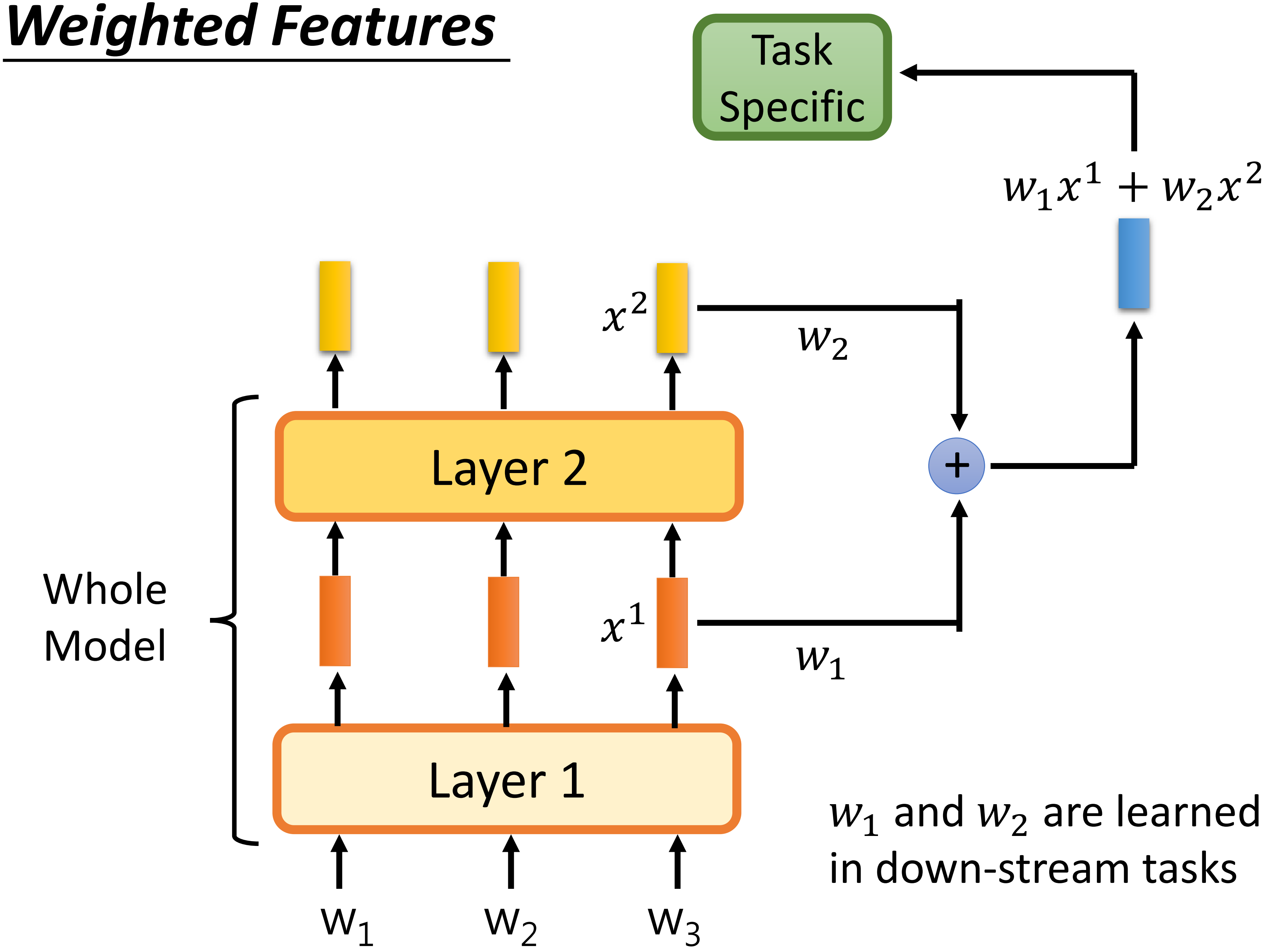
[Houlsby, et al., ICML'19]



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

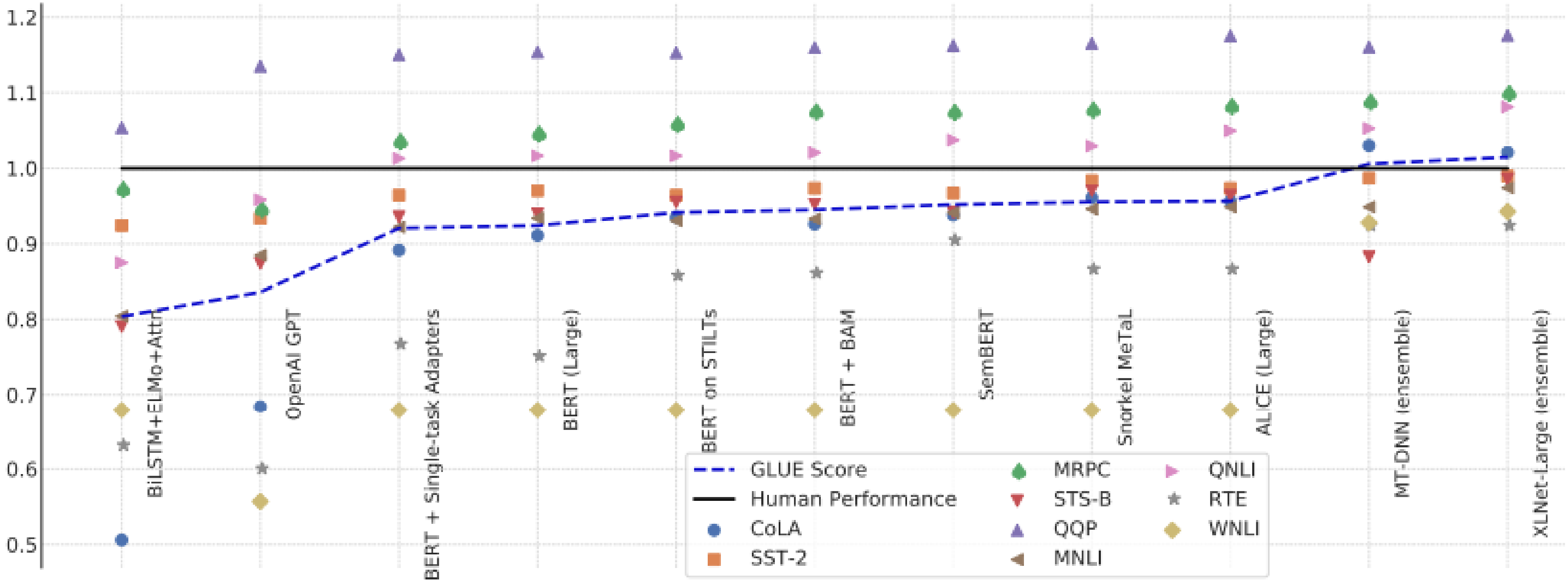
[Houlsby, et al., ICML'19]

Weighted Features



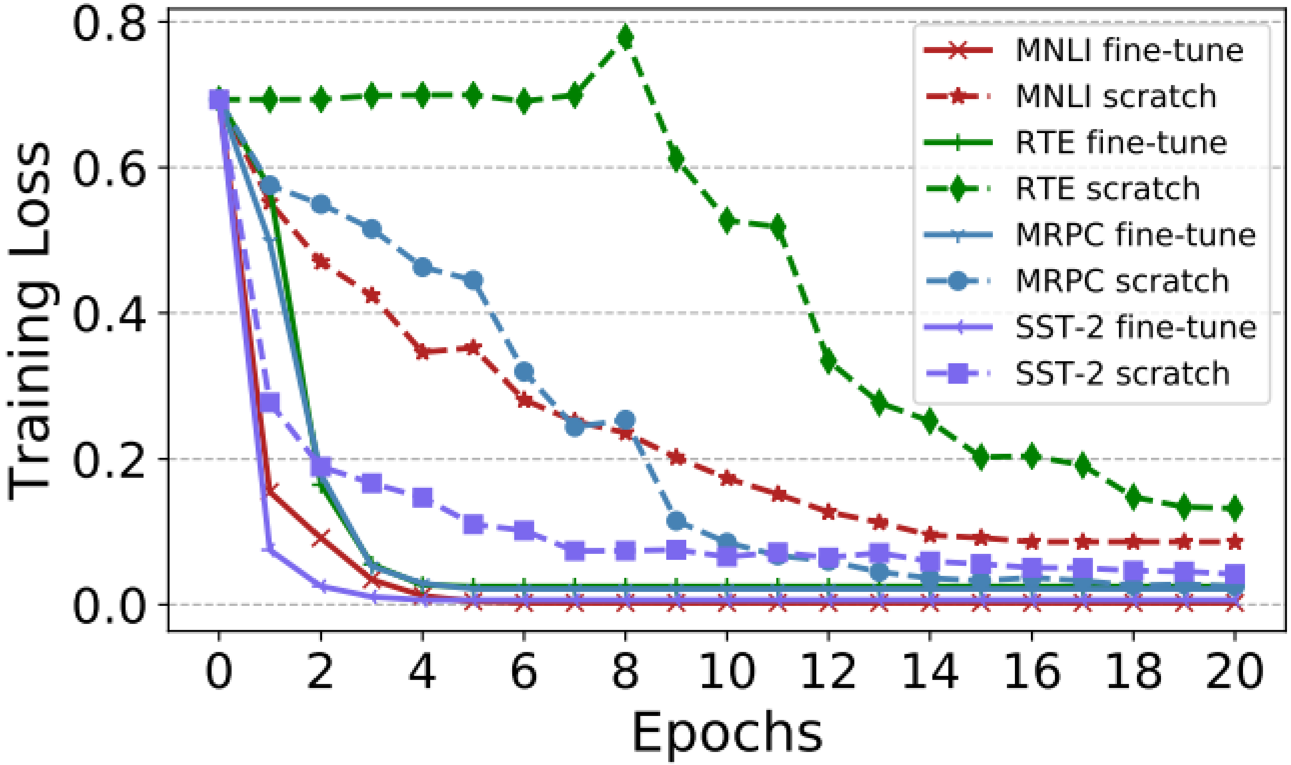
Why Pre-train Models?

- GLUE scores



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

Why Fine-tune?

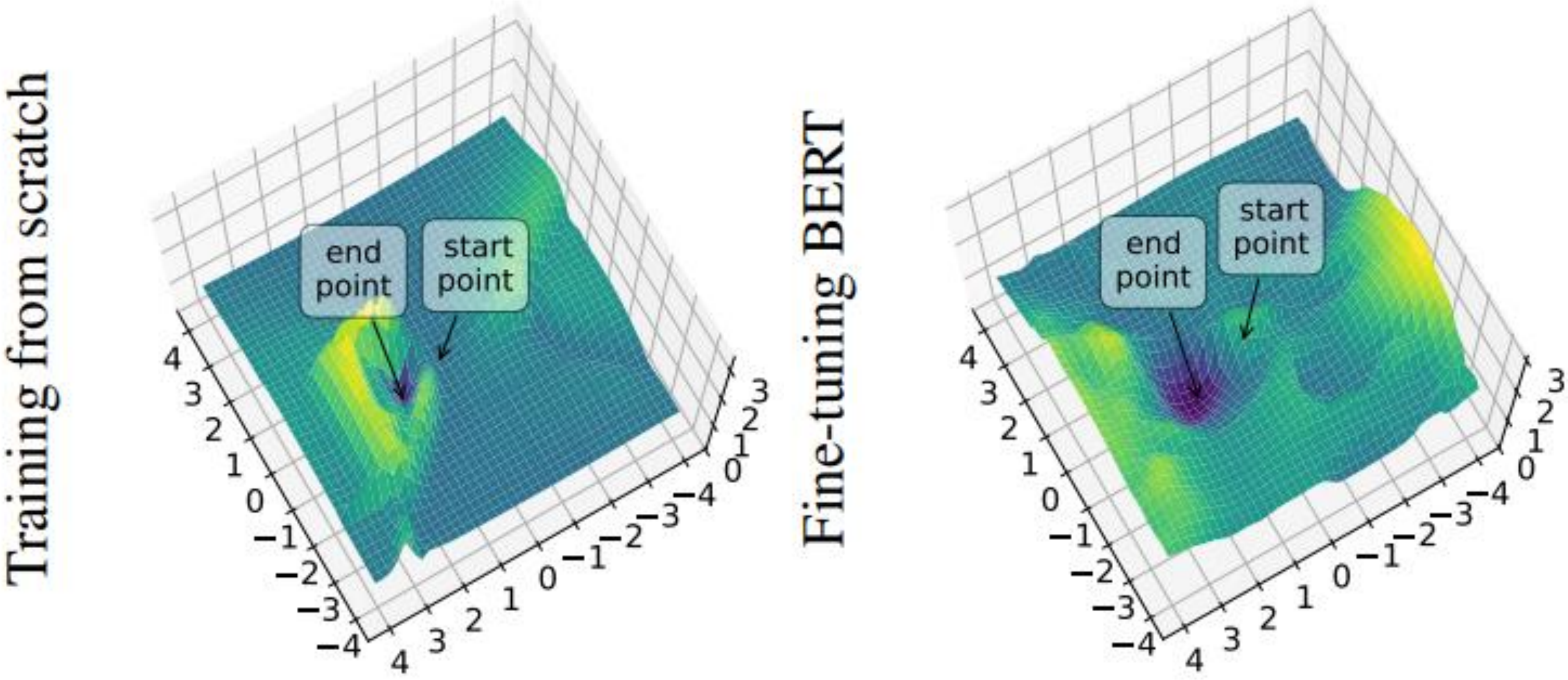


[Hao, et al., EMNLP'19] Source of image: https://arxiv.org/abs/1908.05620

Why Fine-tune?

How to generate the figures below?

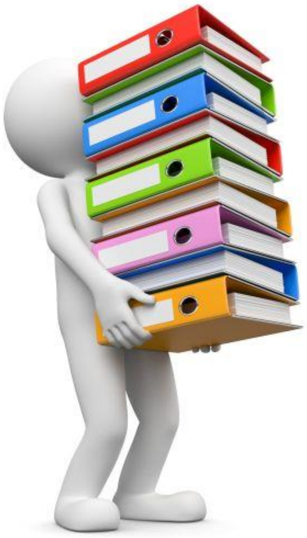
https://youtu.be/XysGHdNOTbg


Training from scratch

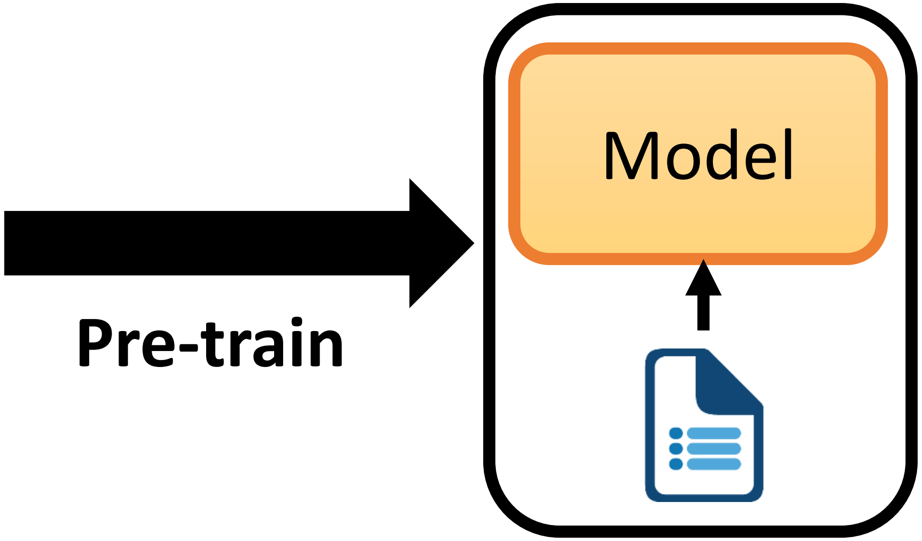
Fine-tuning BERT

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

How to Pre-train



Text without annotation



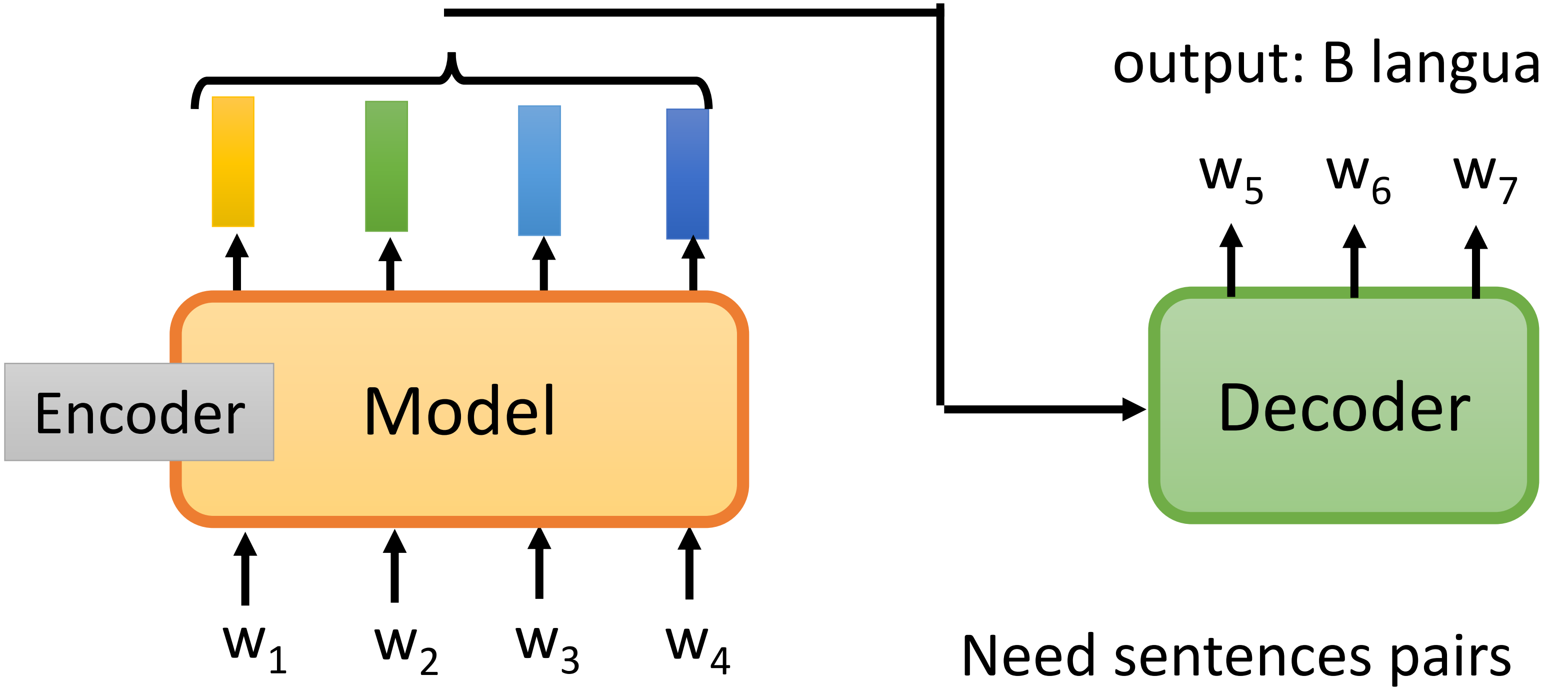
A model that can read text

Model

Pre-train

Pre-training by Translation

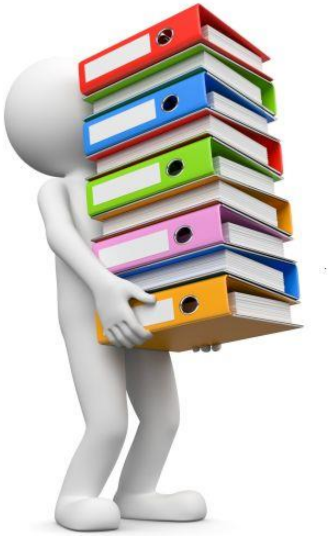
- Context Vector (CoVe)



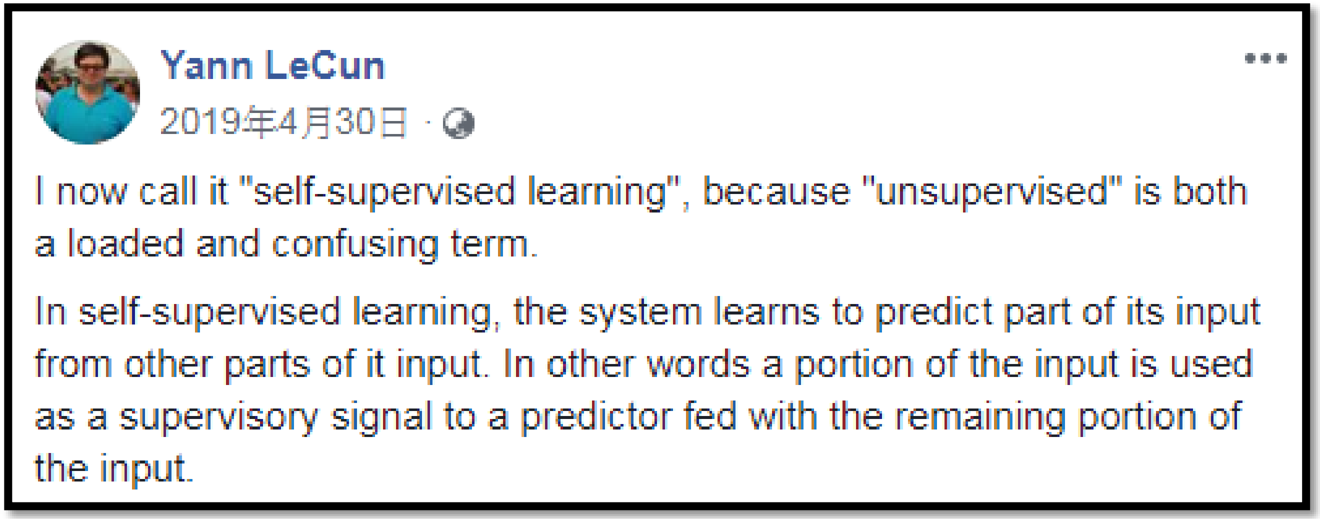
Input: A language

Need sentences pairs
for languages A and B

Self-supervised Learning



Text without annotation

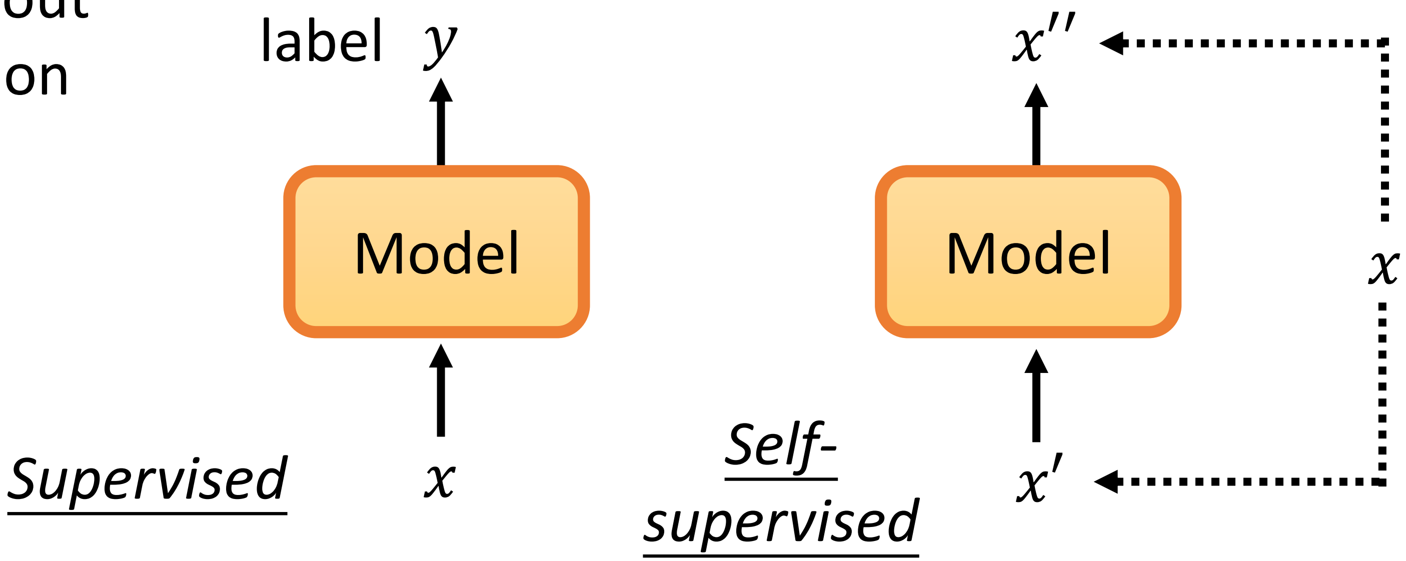


Yann LeCun

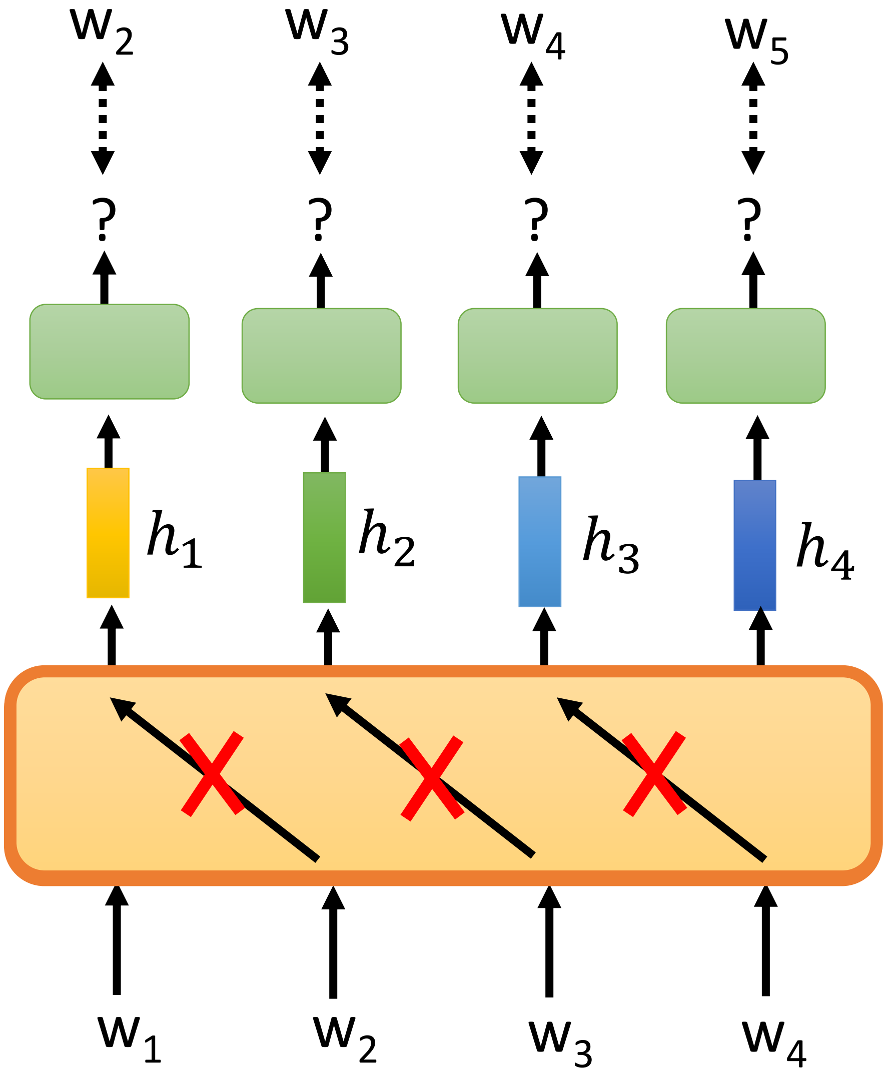
2019年4月30日 ·

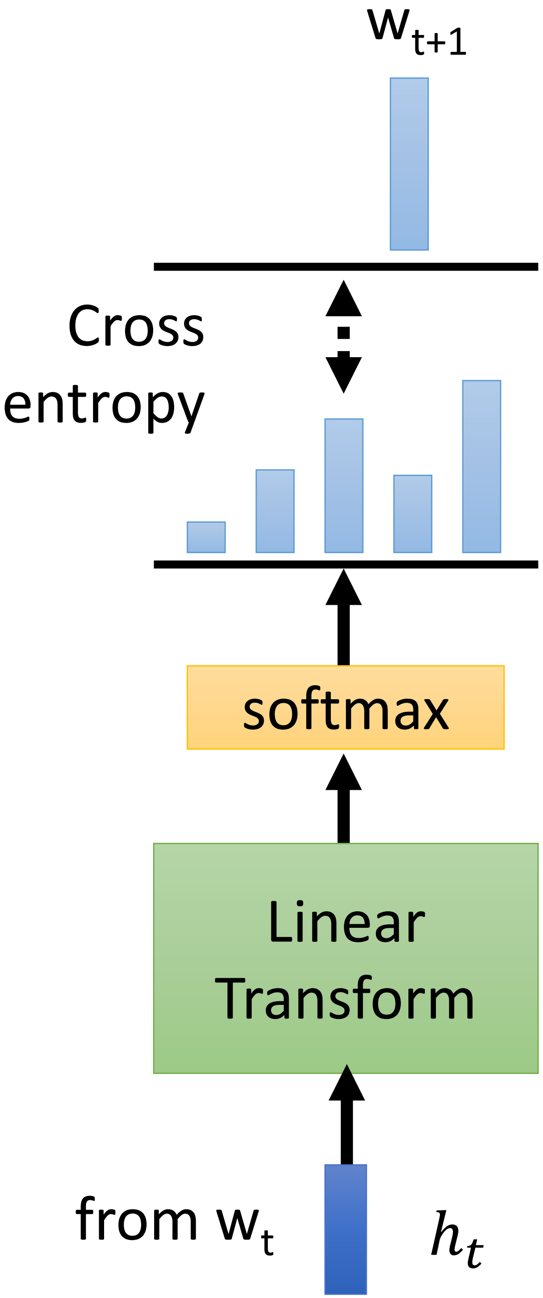
I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

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.

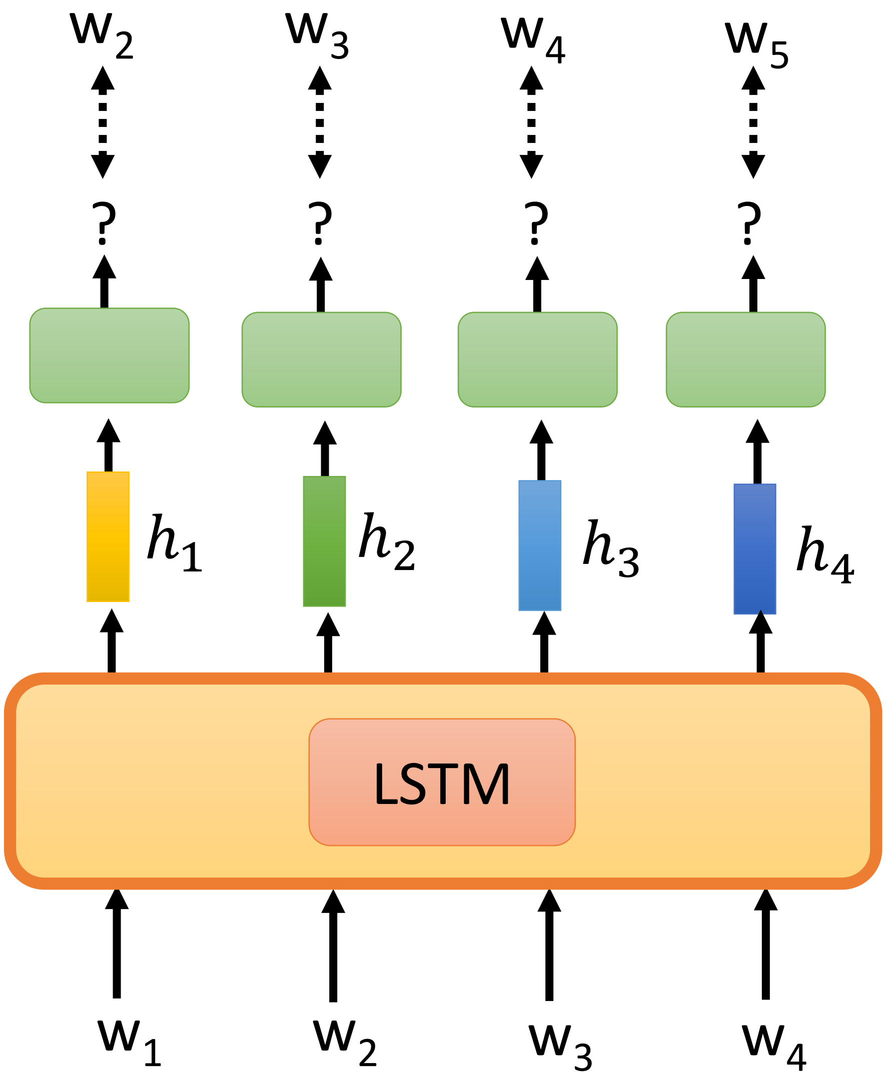


Predict Next Token





Predict Next Token



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

Universal Language Model Fine-tuning (ULMFiT)

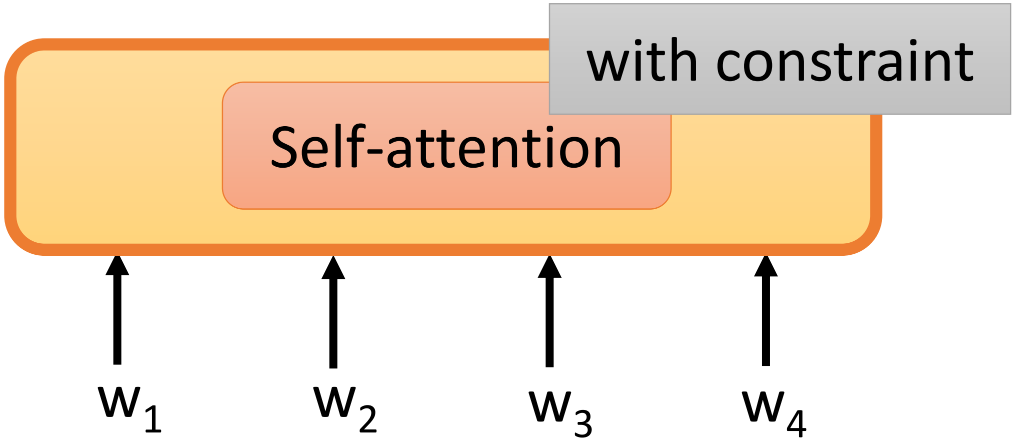
[Howard, et al., ACL'18]

ELMo

[Peters, et al., NAACL'18]

Predict Next Token

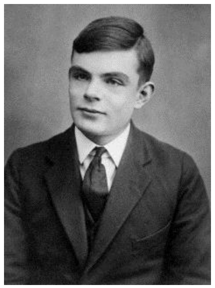
	w_1	w_2	w_3	w_4
w_1	🟩	🟩	🟩	🟩
w_2		🟩	🟩	🟩
w_3			🟩	🟩
w_4				🟩



GPT [Alec, et al., 2018]

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

Megatron [Shoeybi, et al., arXiv’19]

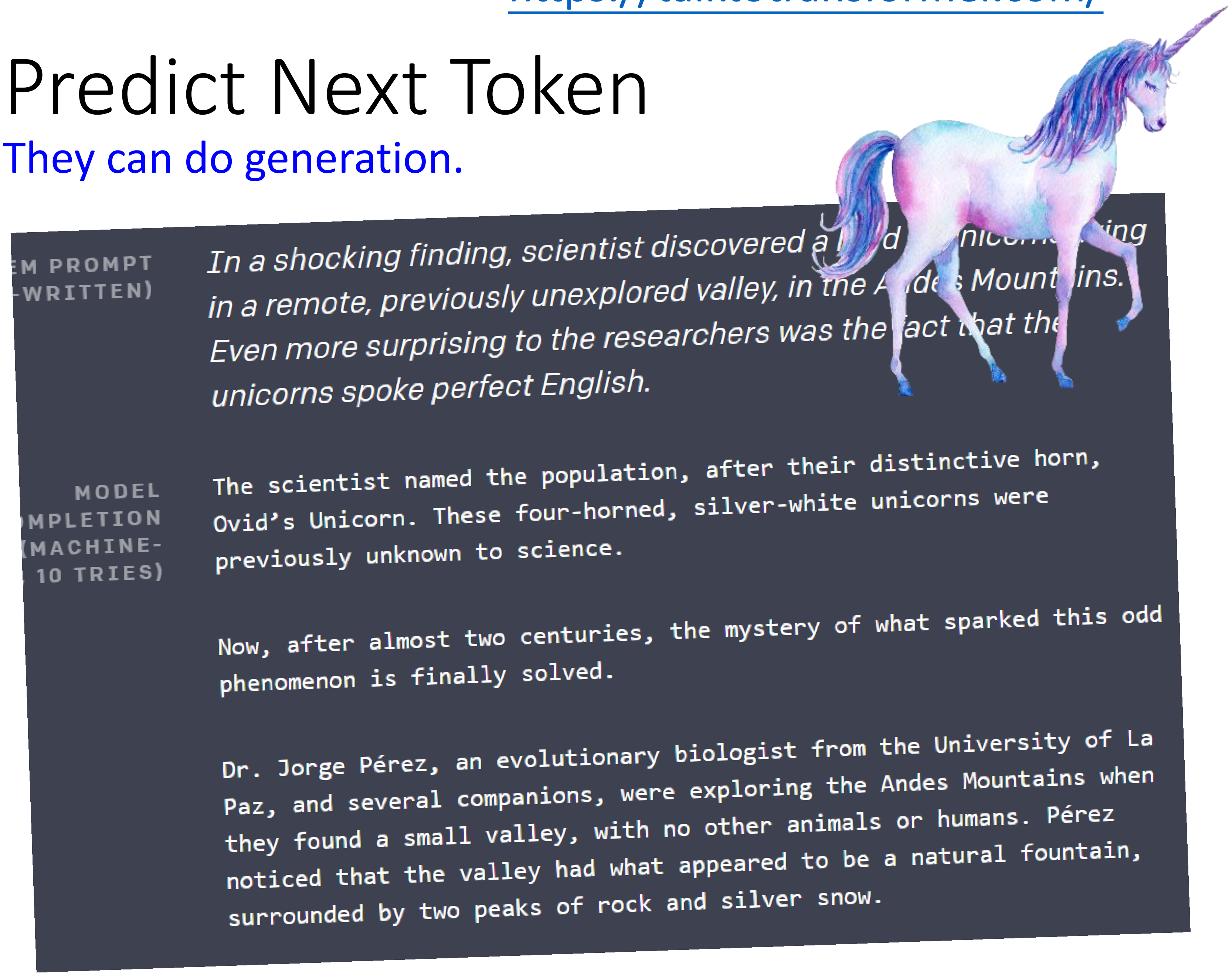


Turing NLG

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
OMPLETION
(MACHINE-
, 10 TRIES)

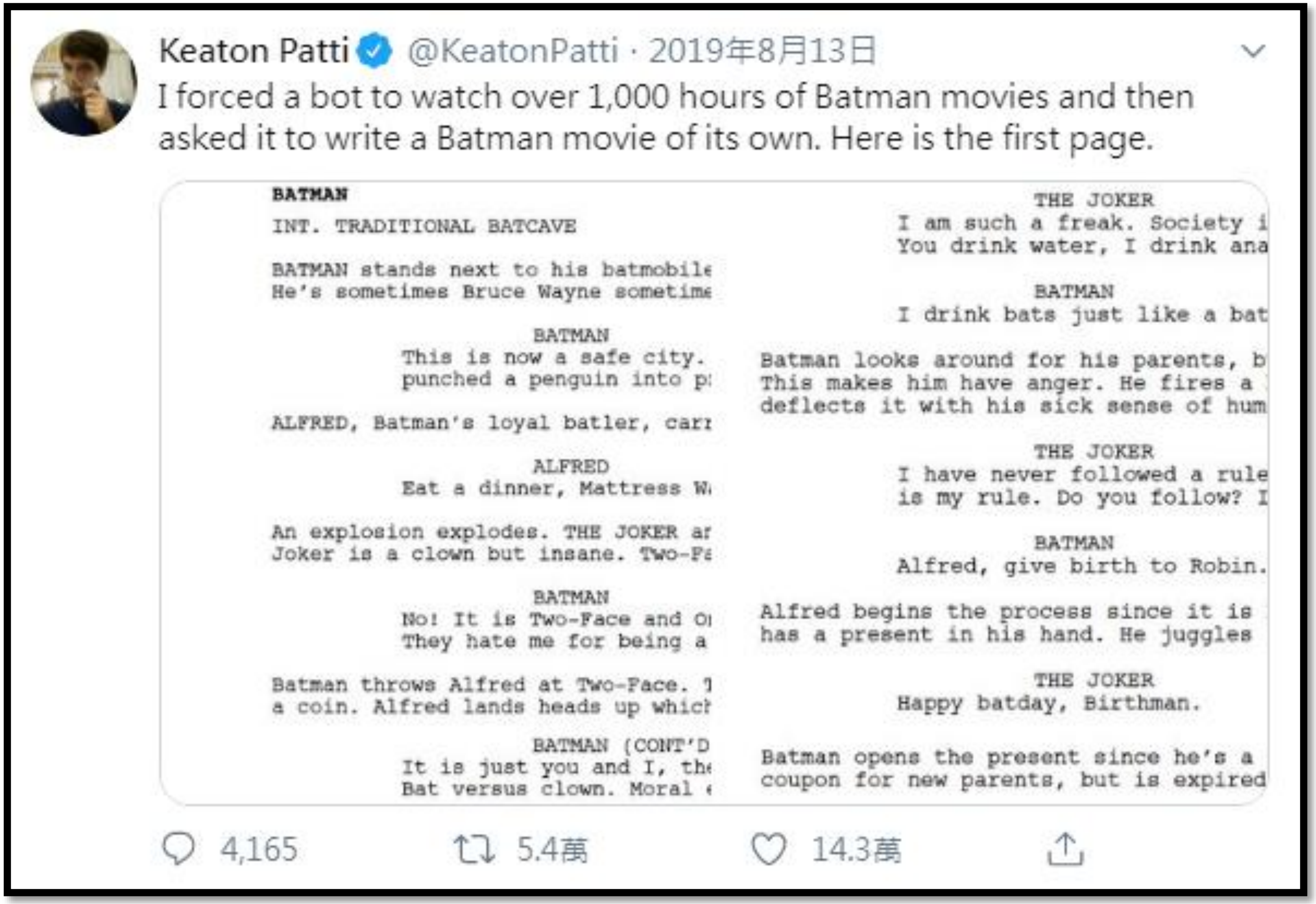
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

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.

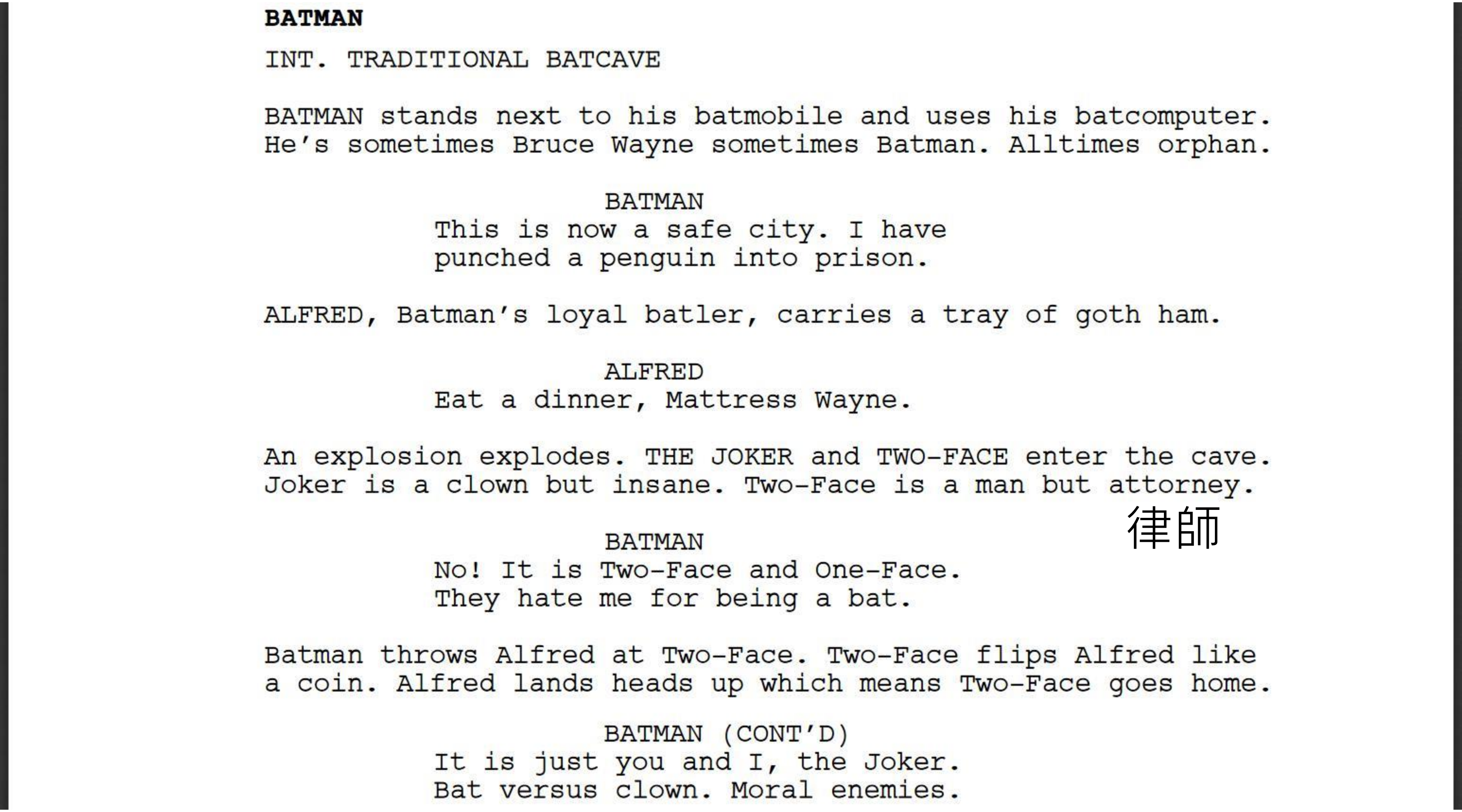


Keaton Patti @KeatonPatti · 2019年8月13日

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 batmobil
He's sometimes Bruce Wayne sometim

                    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-F

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

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

                    BATMAN (CONT'D
          It is just you and I, th
          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
```



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

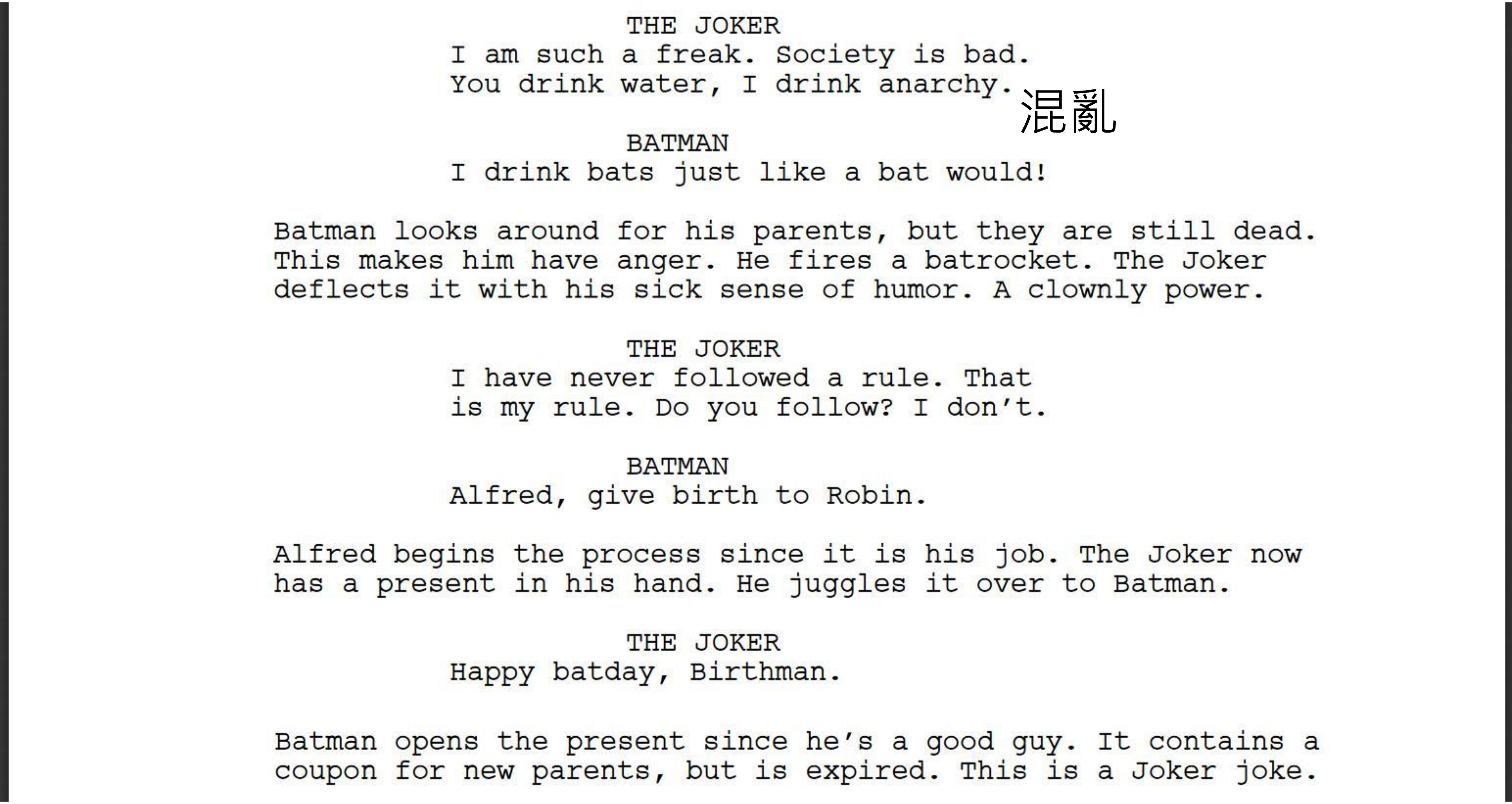
BATMAN

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.

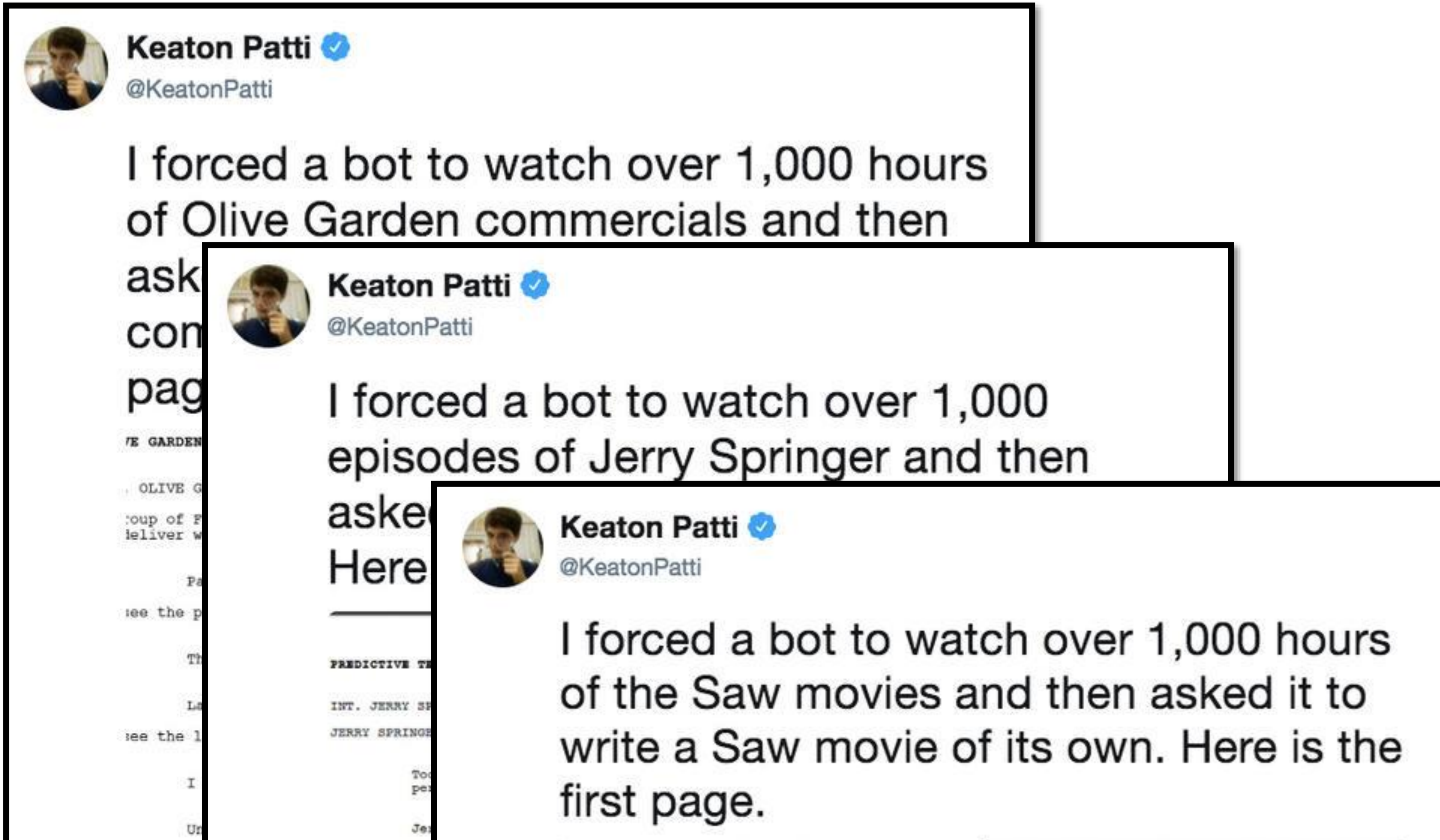
Alfred begins the process since it is his job. The Joker now has a present in his hand. He juggles it over to Batman.

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!!!



Keaton Patti
@KeatonPatti

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

'E GARDEN

OLIVE G

oup of F
eliver w

Pa

ee the p

Th

La

ee the l

I

Ur

Keaton Patti
@KeatonPatti

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

PREDICTIVE TE

INT. JERRY SP

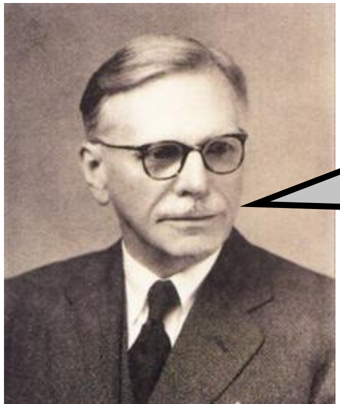
JERRY SPRINGE

To
pe

Je

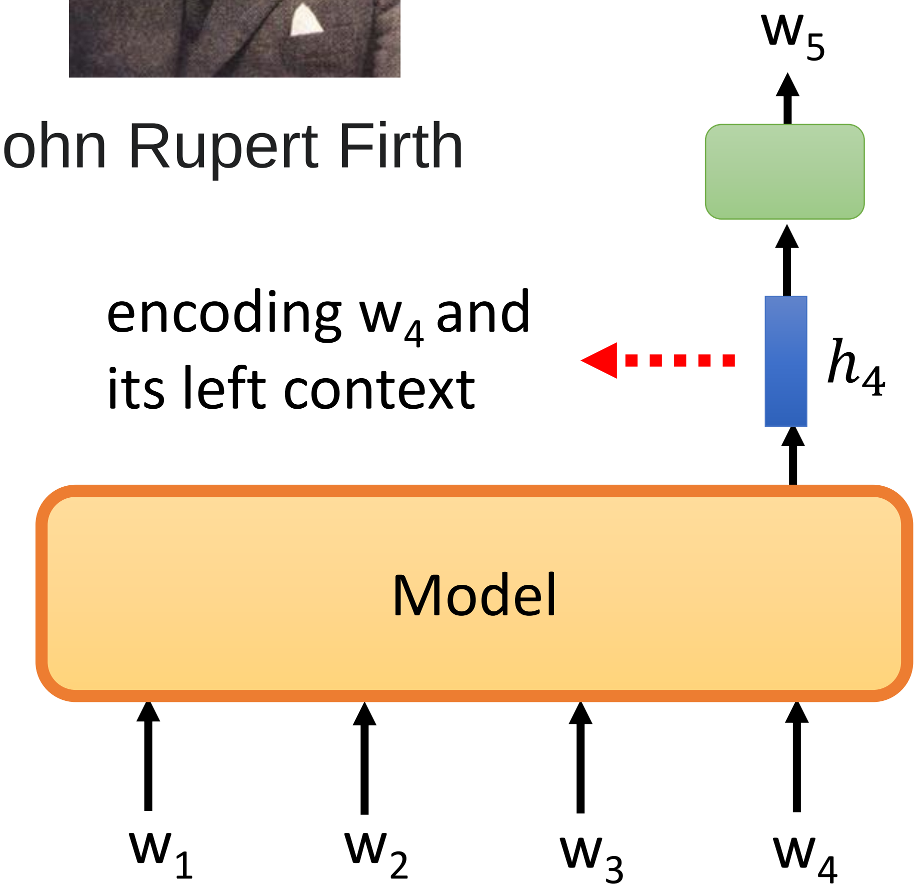
Keaton Patti
@KeatonPatti

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.



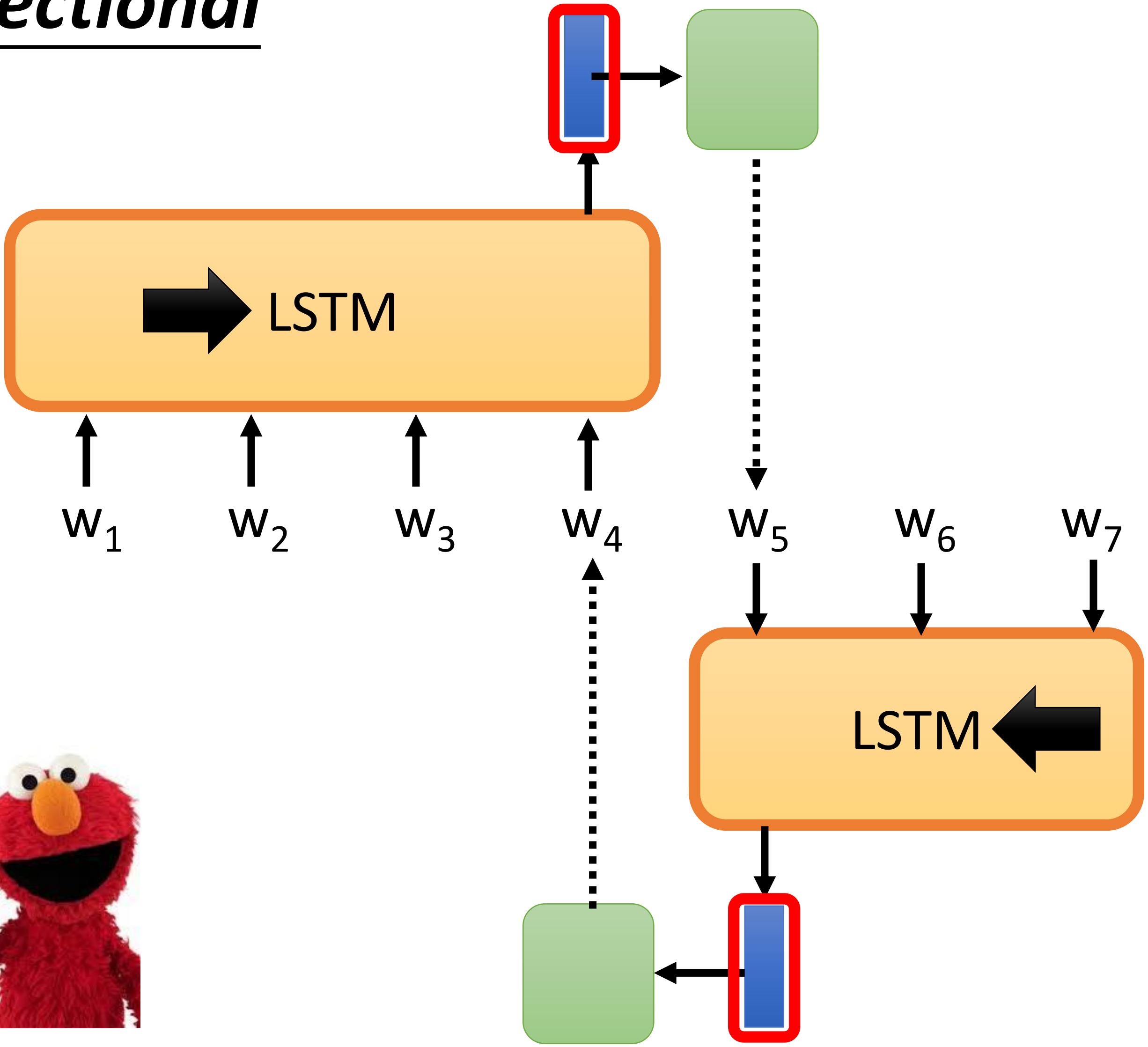
You shall know a word by the company it keeps

John Rupert Firth



How about the right context!?

Predict Next Token
- Bidirectional



ELMO

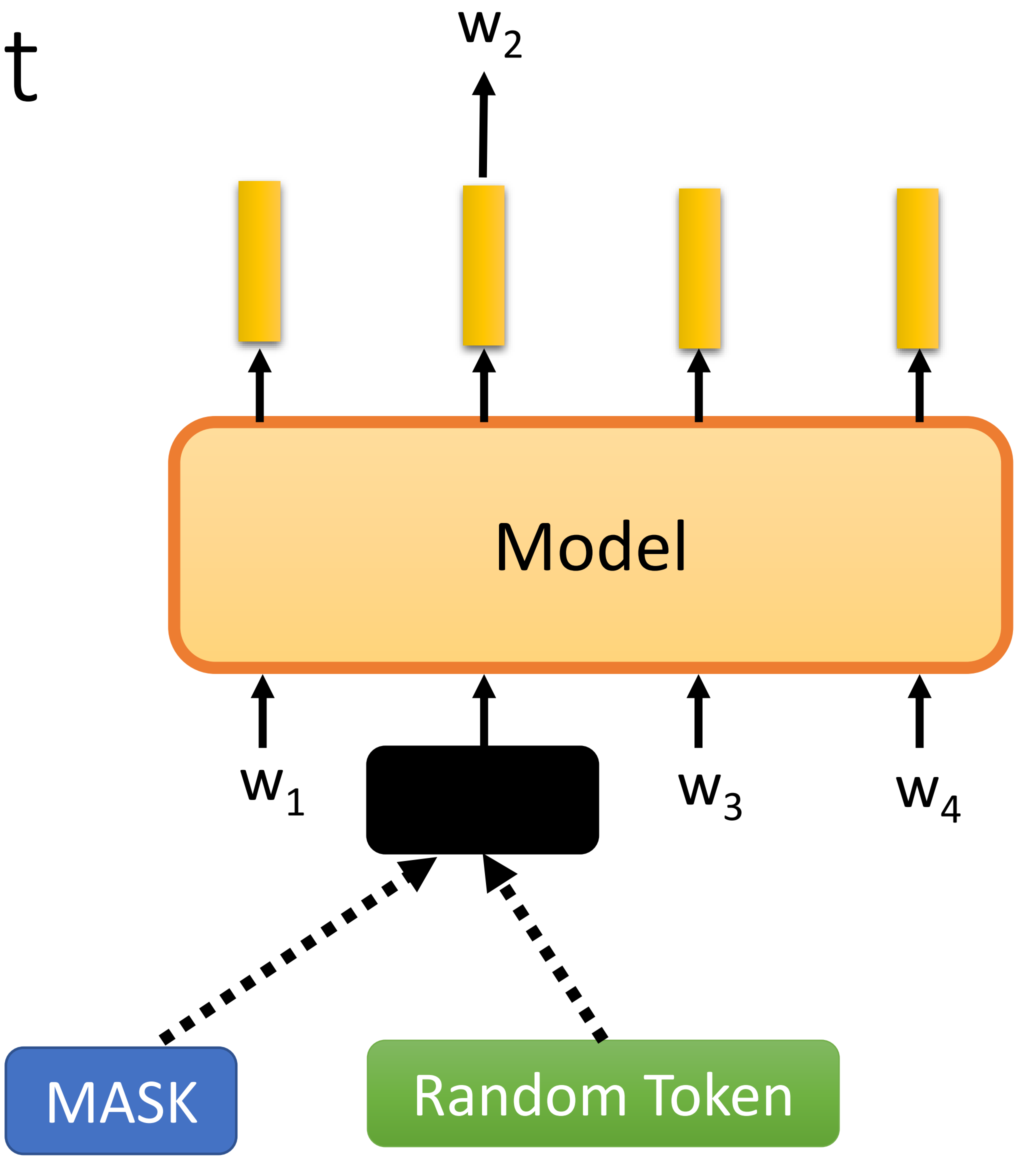
Masking Input

BERT

[Devlin, et al., NAACL'19]

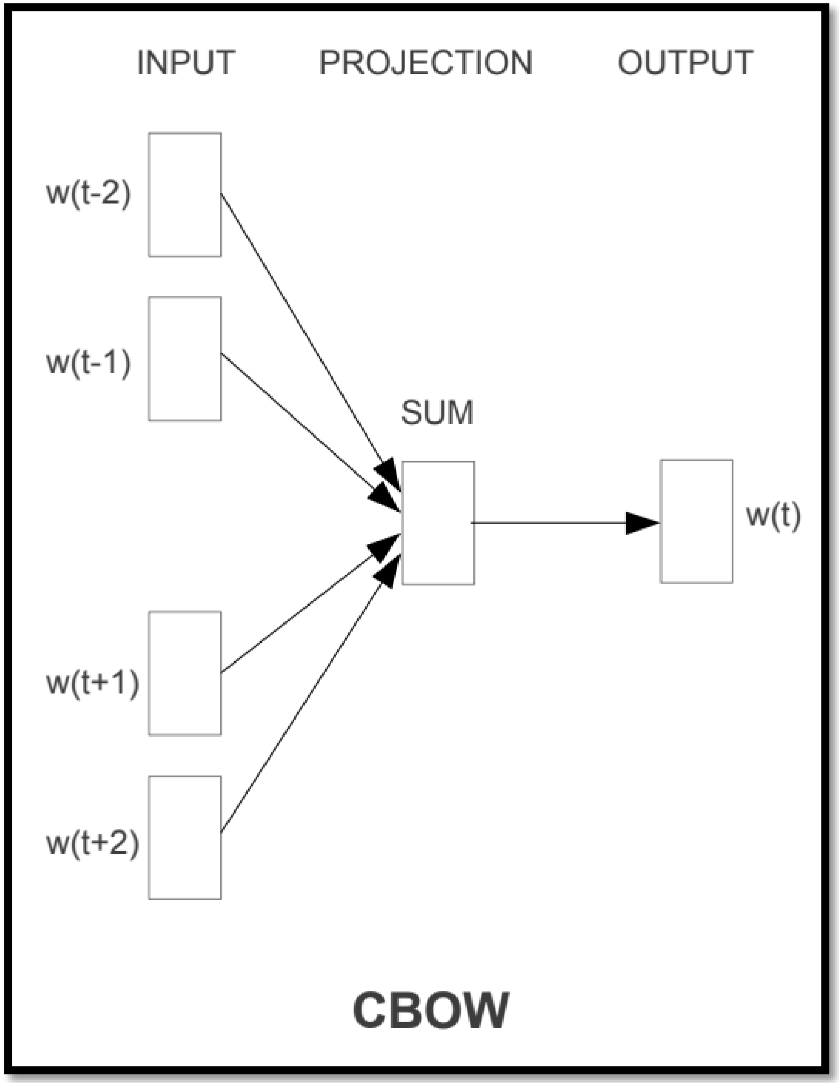
Transformer

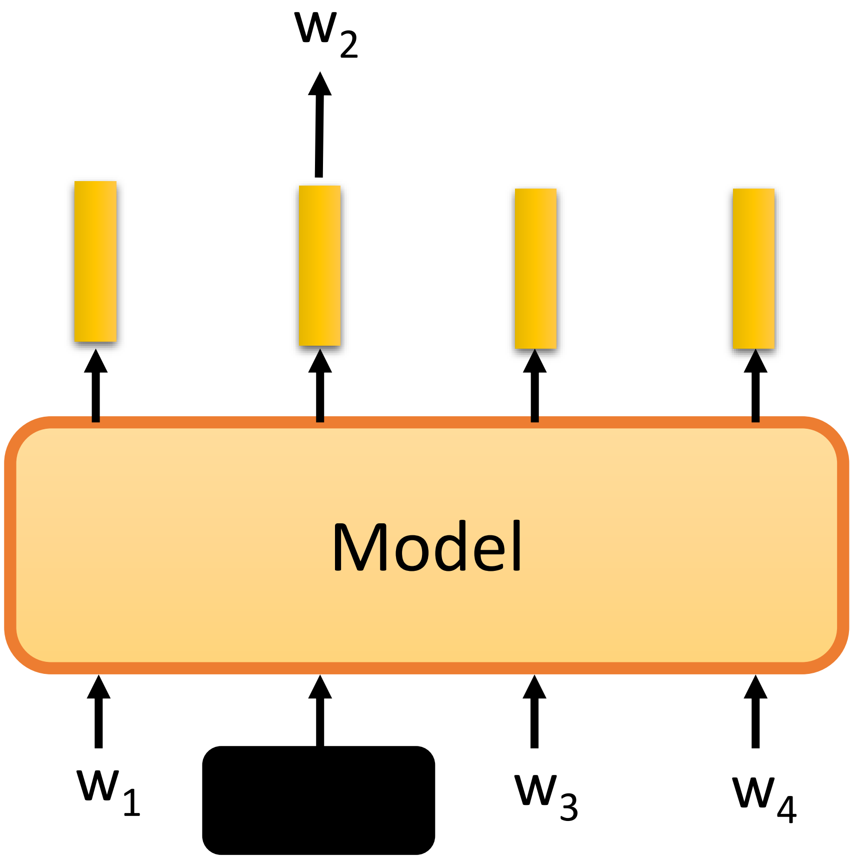
(no limitation on self-attention)



(special token)

Masking Input



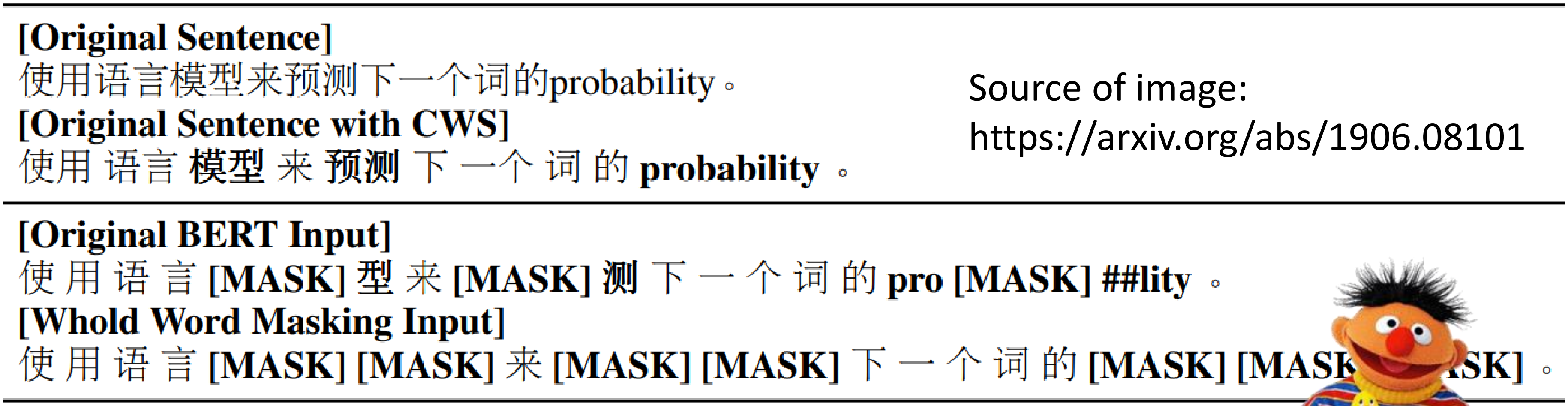


Using context to predict the missing token

Masking Input

Is random masking good enough?

- 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]
使用 语言 **模型** 来 **预测** 下 一个 词 的 **probability** 。

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

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

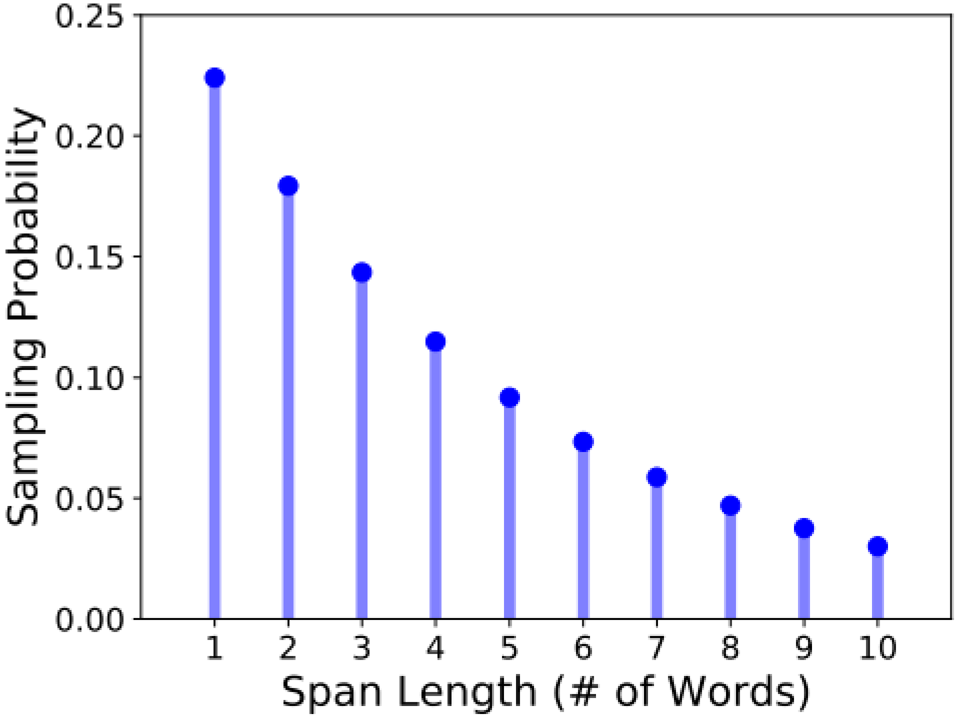
- Phrase-level & Entity-level
  [Sun, et al., ACL'19]

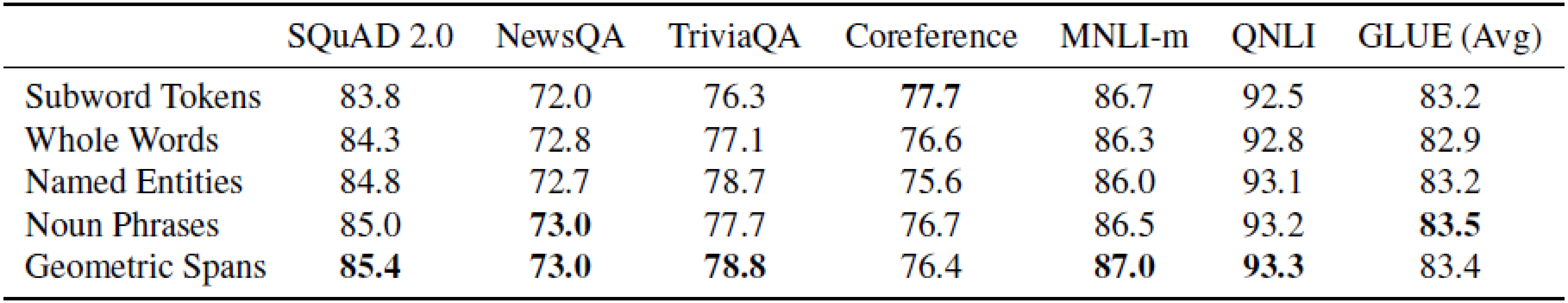
Enhanced Representation through Knowledge Integration (ERNIE)

SpanBert

[Joshi, et al., TACL’20]

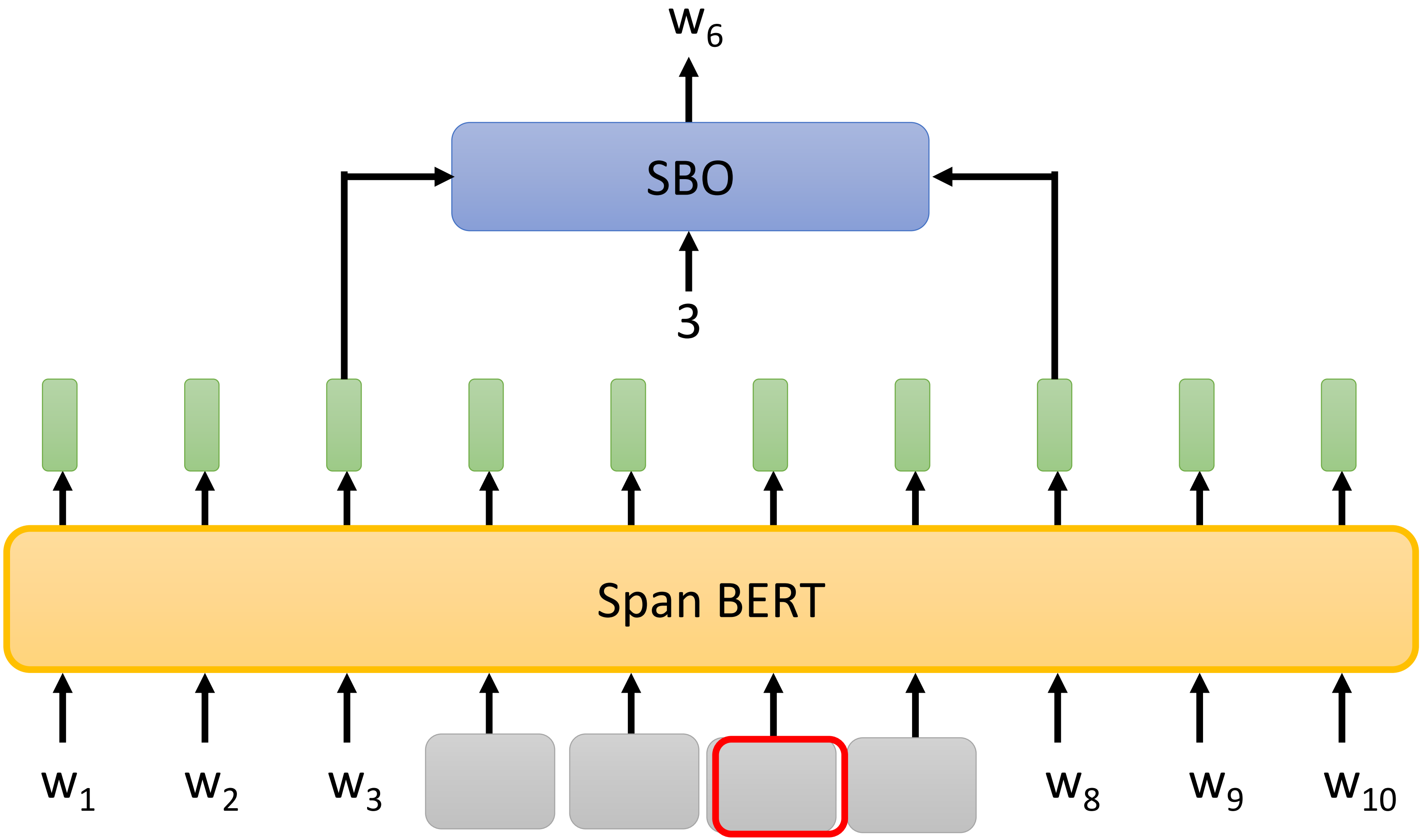
Source of image: https://arxiv.org/abs/1907.10529





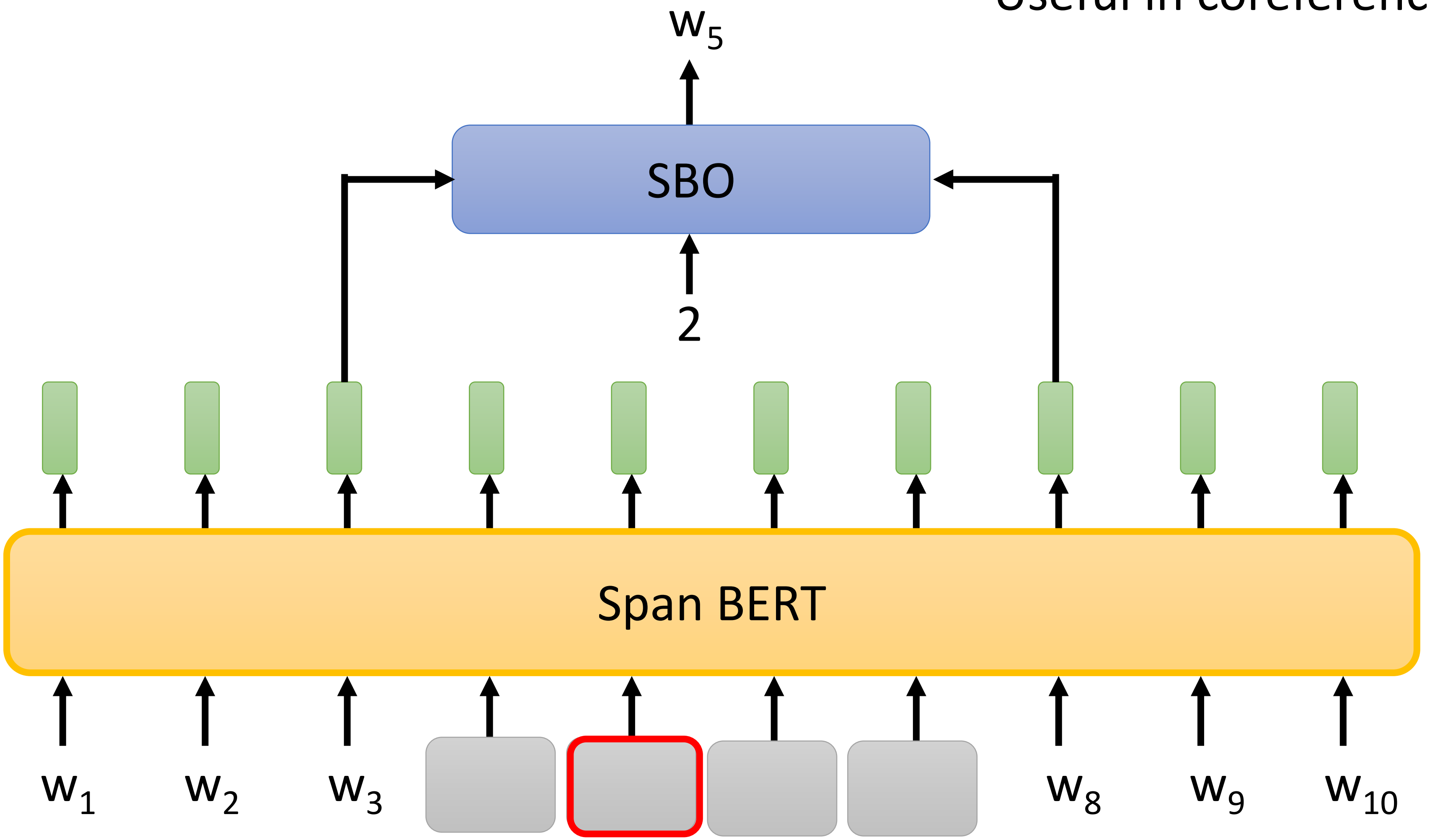
	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI	GLUE (Avg)
Subword Tokens	83.8	72.0	76.3	**77.7**	86.7	92.5	83.2
Whole Words	84.3	72.8	77.1	76.6	86.3	92.8	82.9
Named Entities	84.8	72.7	78.7	75.6	86.0	93.1	83.2
Noun Phrases	85.0	**73.0**	77.7	76.7	86.5	93.2	**83.5**
Geometric Spans	**85.4**	**73.0**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

SpanBert –
Span Boundary Objective (SBO)



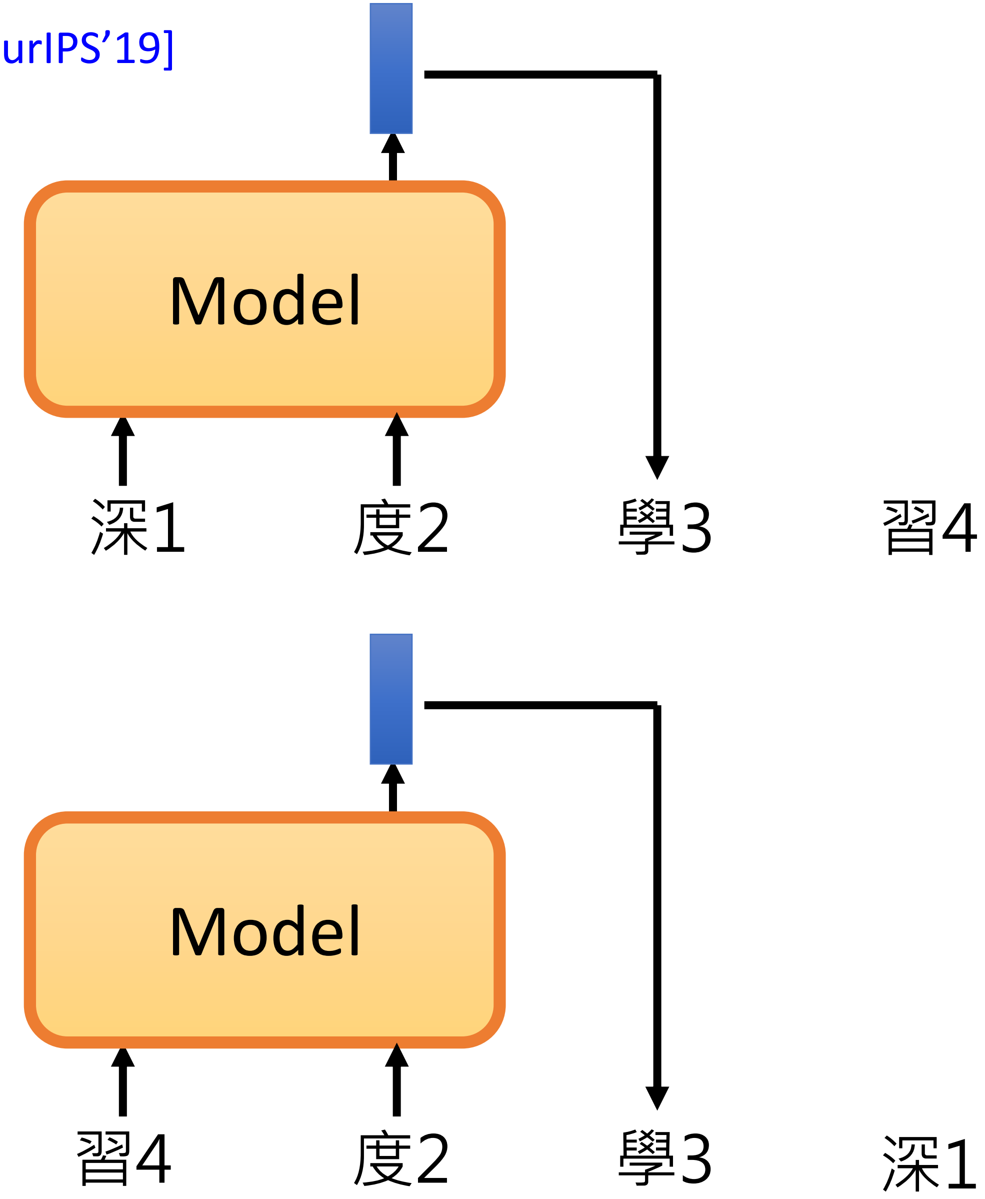
SpanBert –
Span Boundary Objective (SBO)

Useful in coreference?



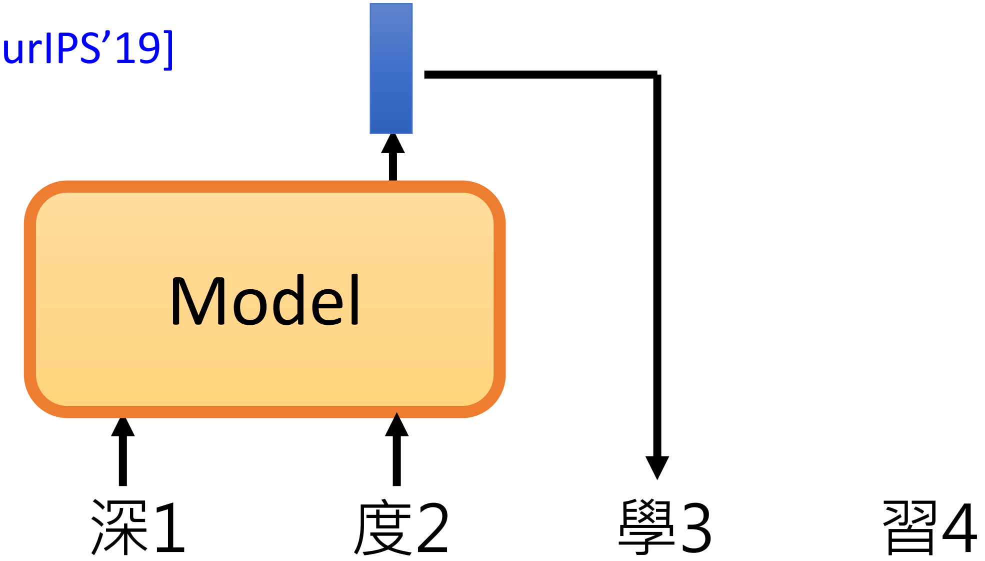
XLNet [Yang, et al., NeurIPS'19]

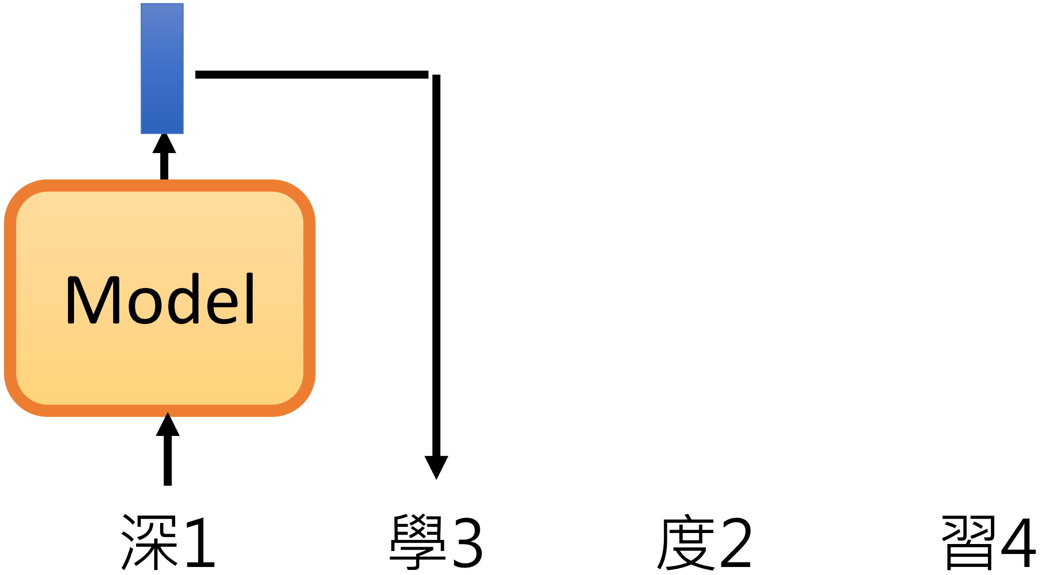
Transformer-XL



XLNet [Yang, et al., NeurIPS’19]

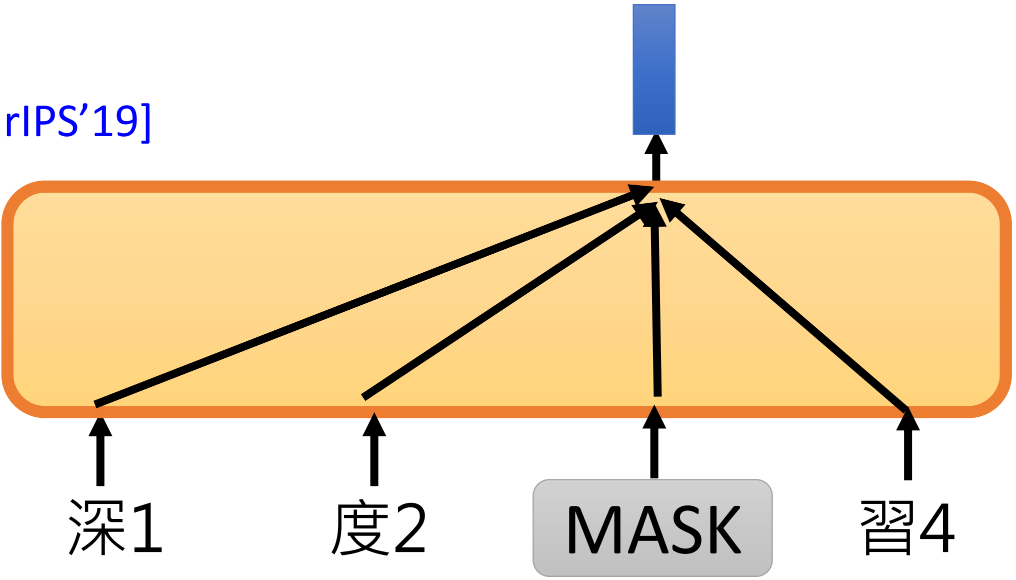
Transformer-XL

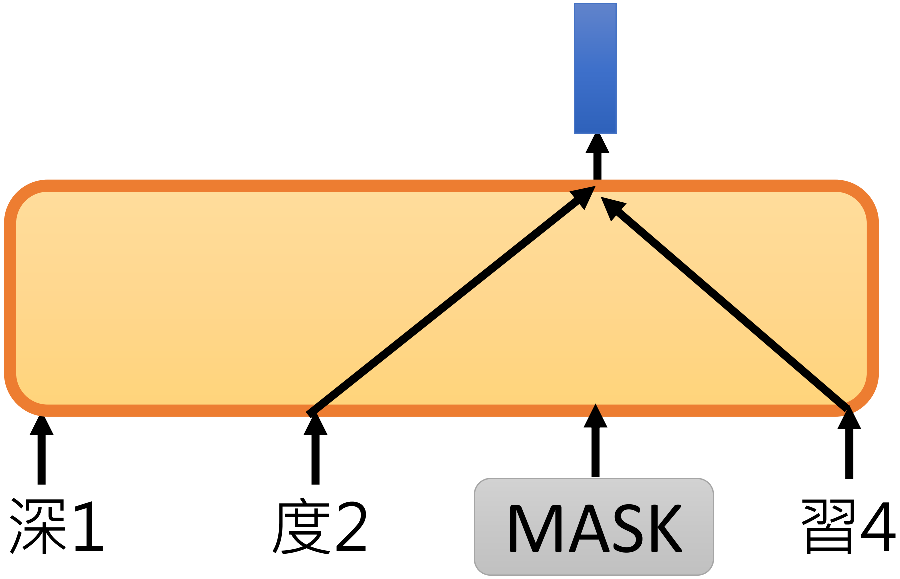


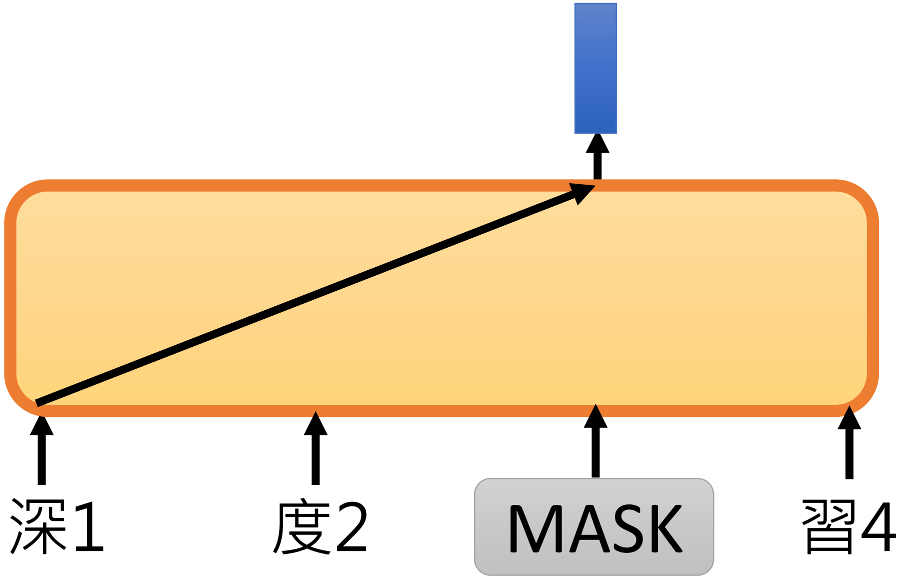


XLNet [Yang, et al., NeurIPS’19]

Transformer-XL



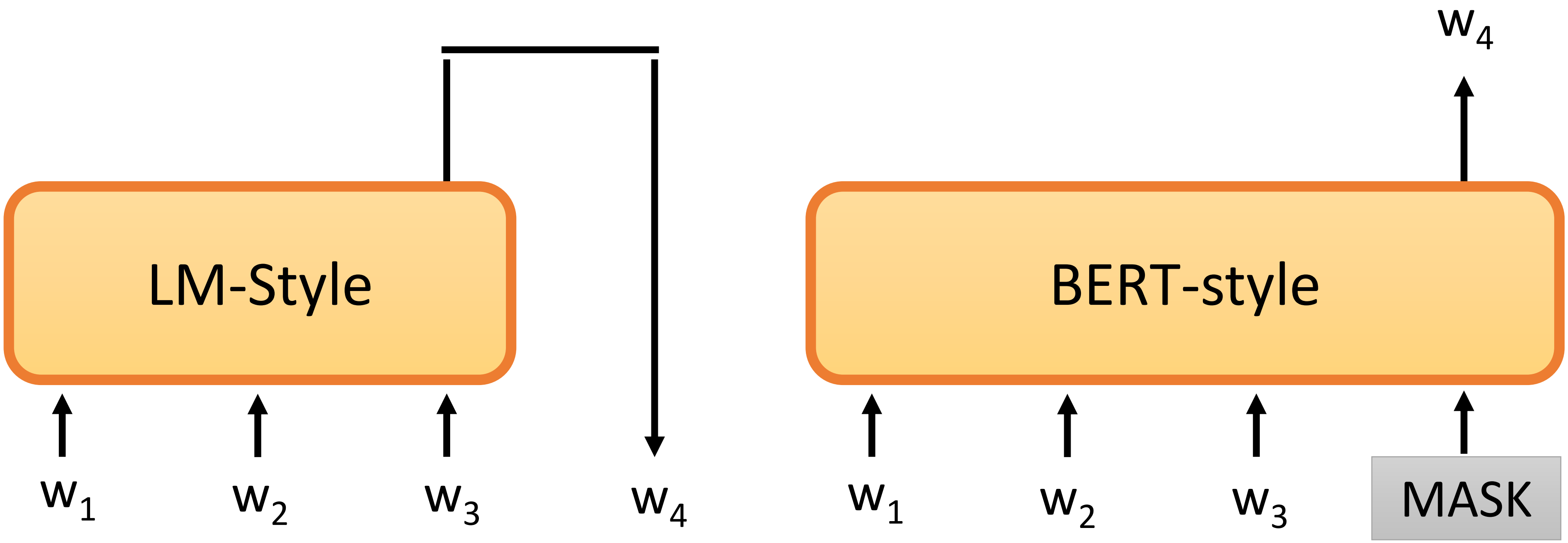




BERT cannot talk?

Limited to autoregressive model (non-autoregressive next time)

Given partial sequence, predict the next token

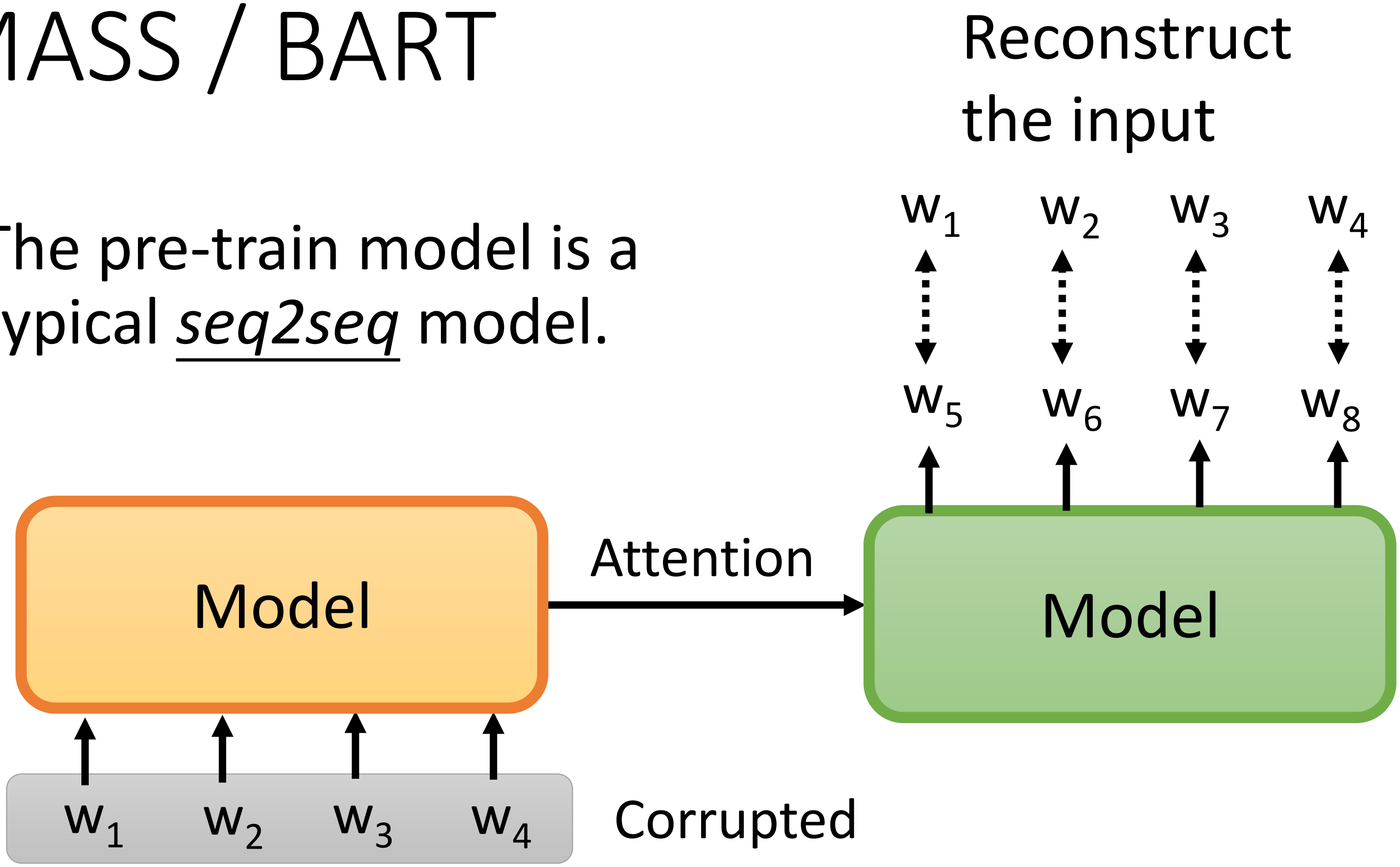


What LM born for

Never seen partial sequence

MASS / BART

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

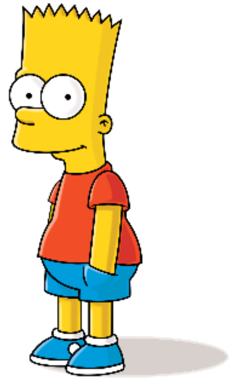


MAsked Sequence to Sequence pre-training (MASS) [Song, et al., ICML'19]

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

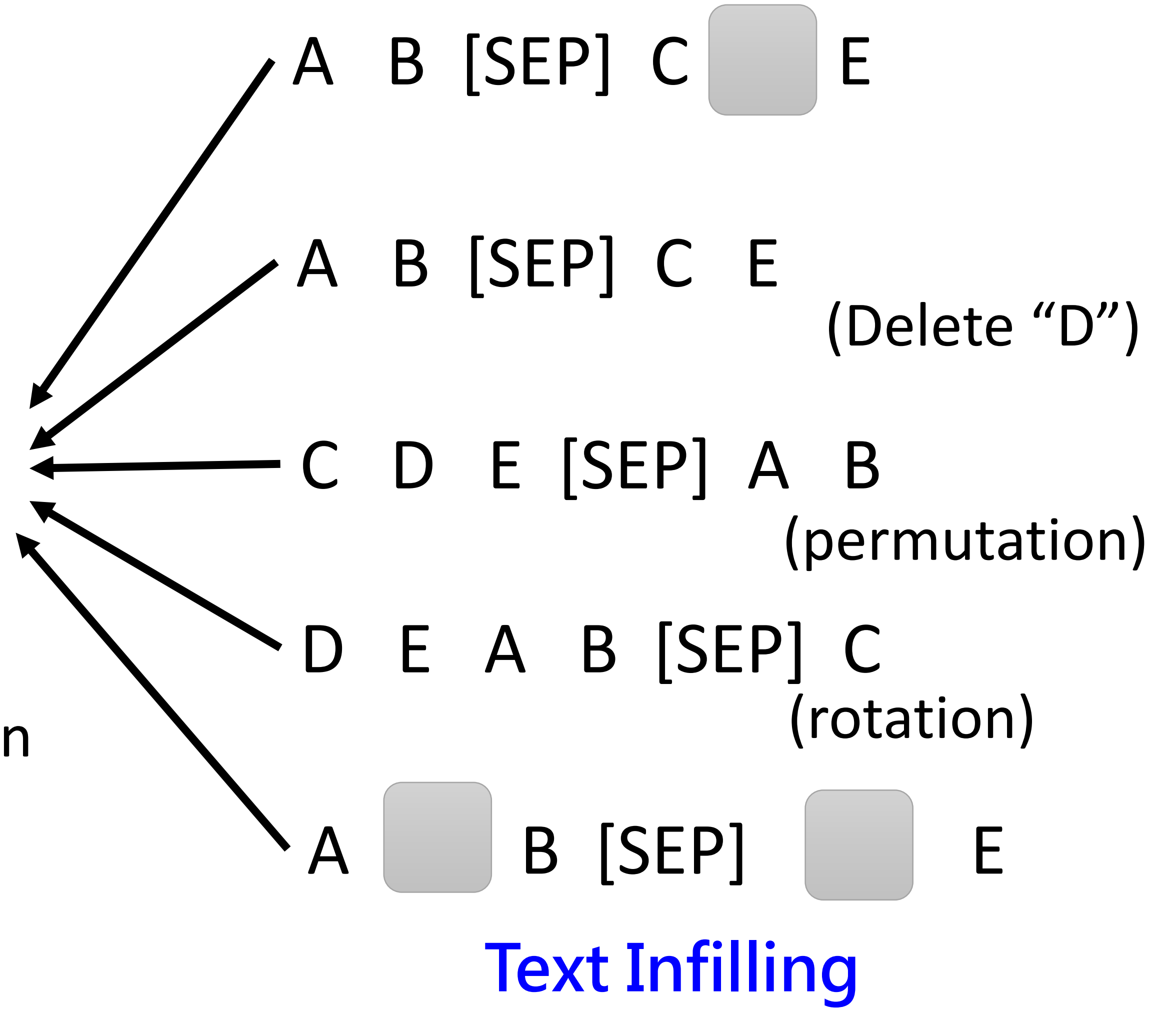
Input Corruption

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)

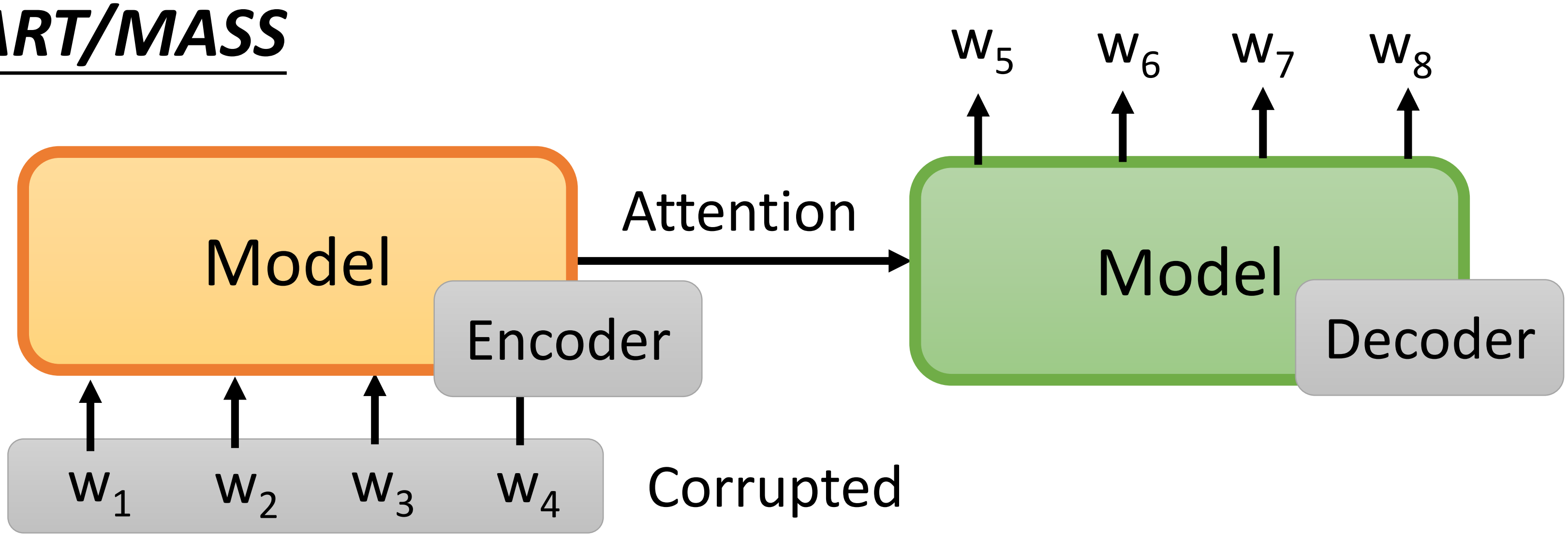
D E A B [SEP] C (rotation)

A □ B [SEP] □ E

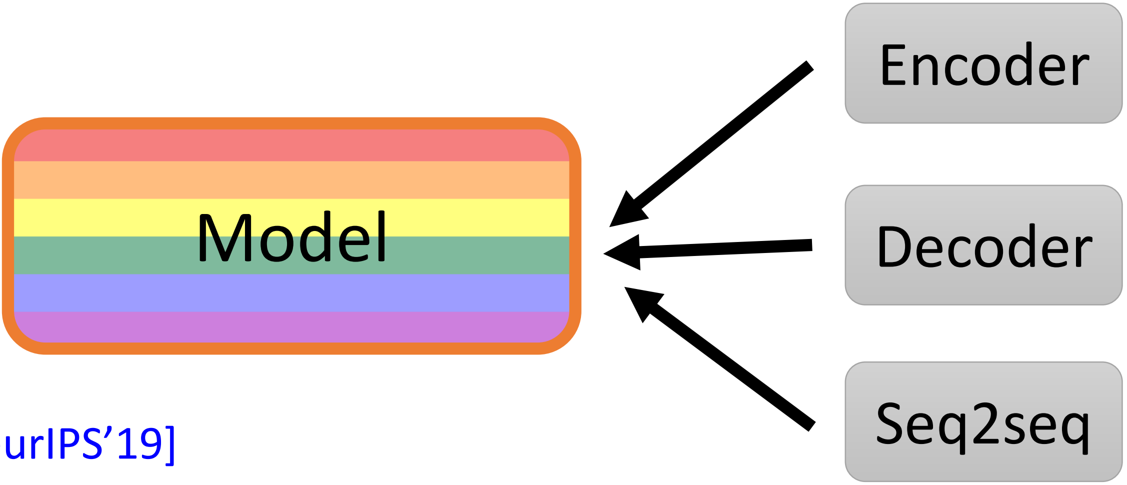
Text Infilling

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

BART/MASS

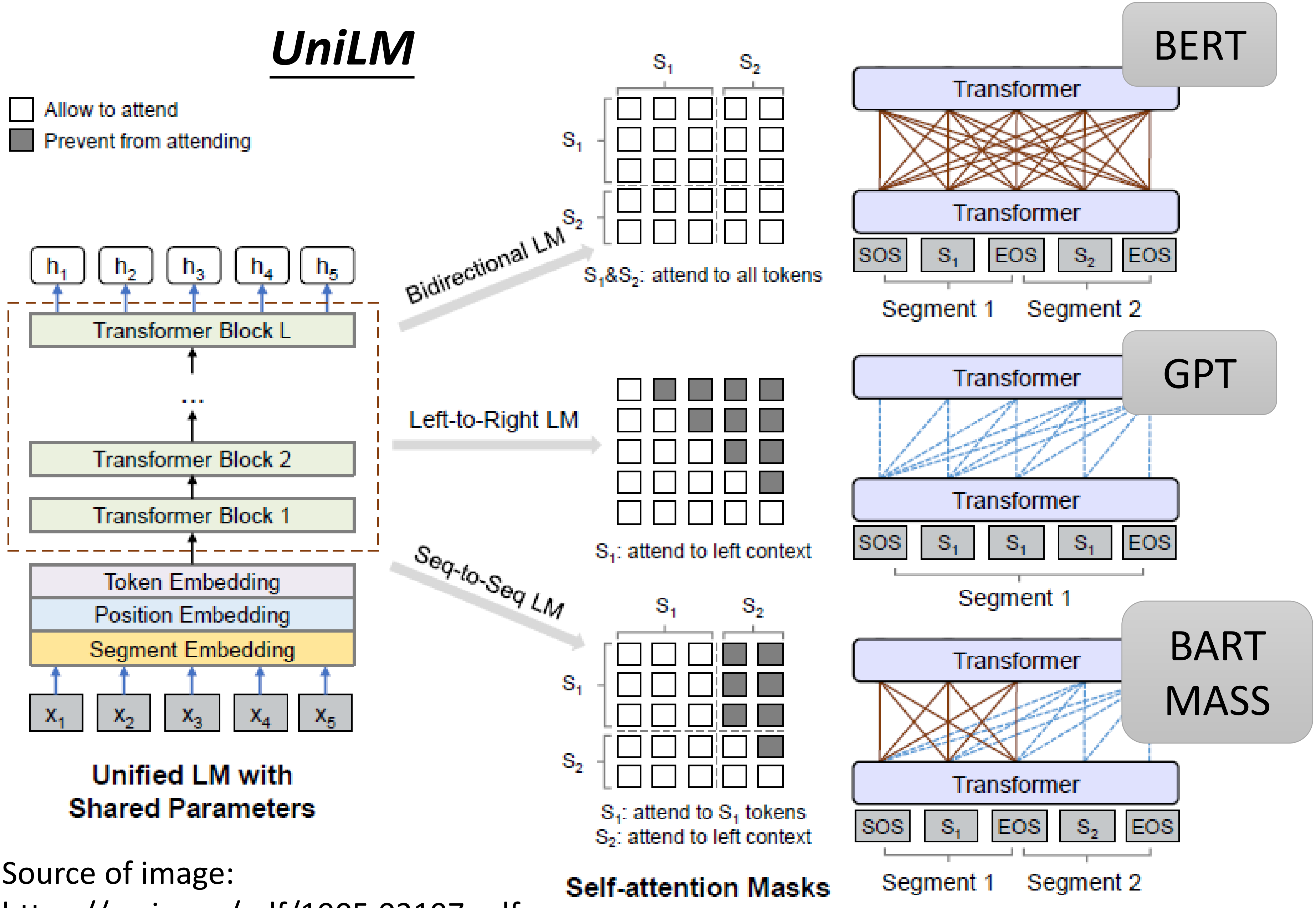


UniLM



[Dong, et al., NeurIPS’19]

UniLM

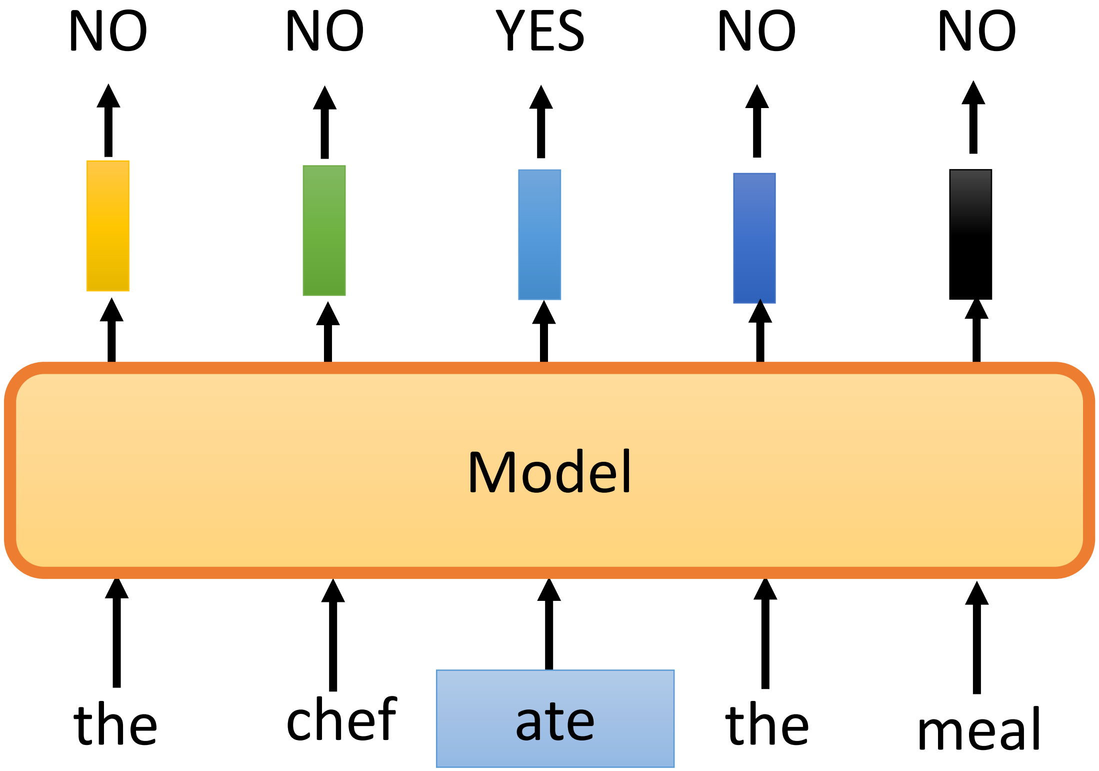


Source of image:
https://arxiv.org/pdf/1905.03197.pdf

Replace or Not?

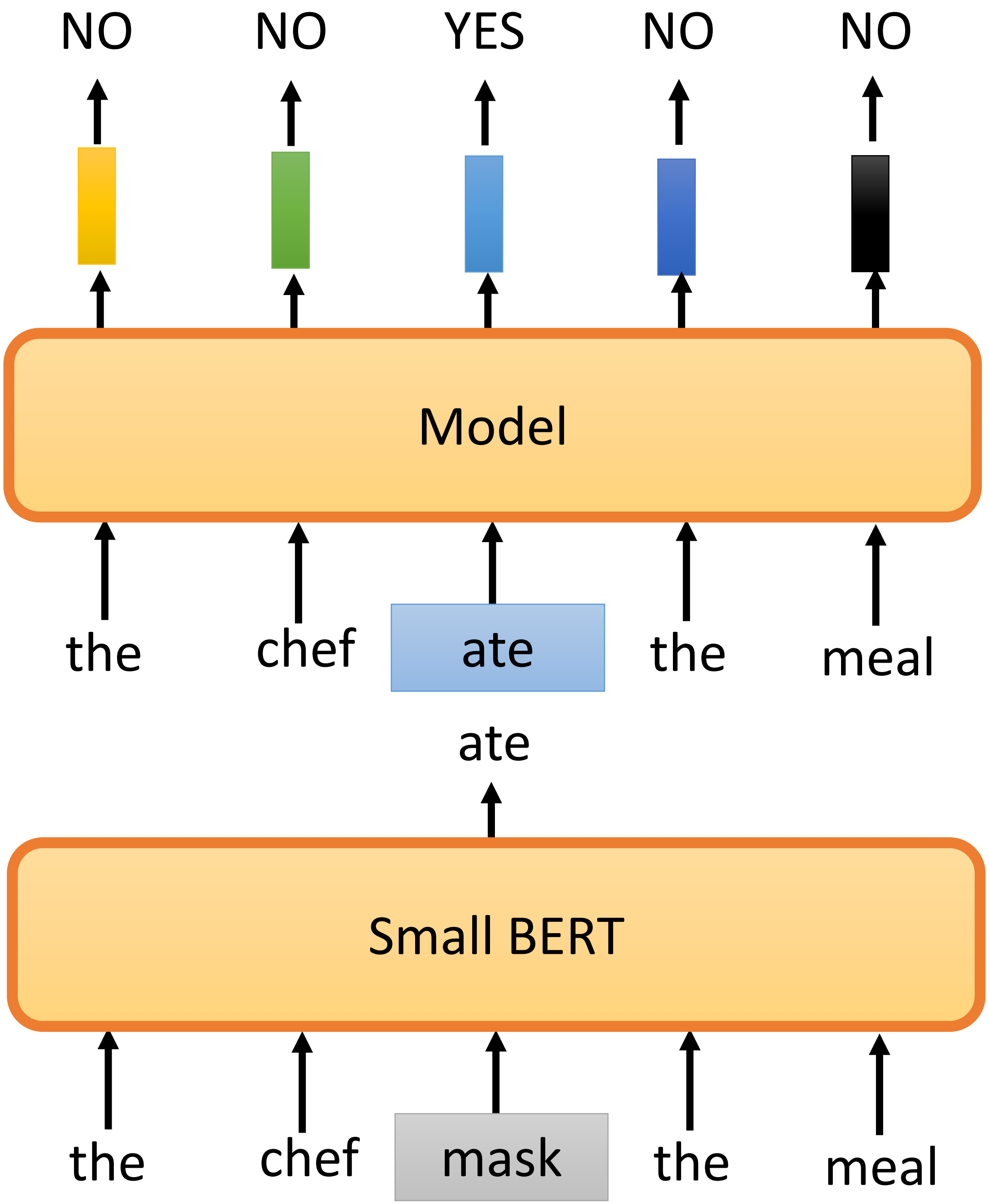
Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)

ELECTRA



Predicting yes/not is easier than reconstruction.

Every output position is used.



NO NO YES NO NO

Model

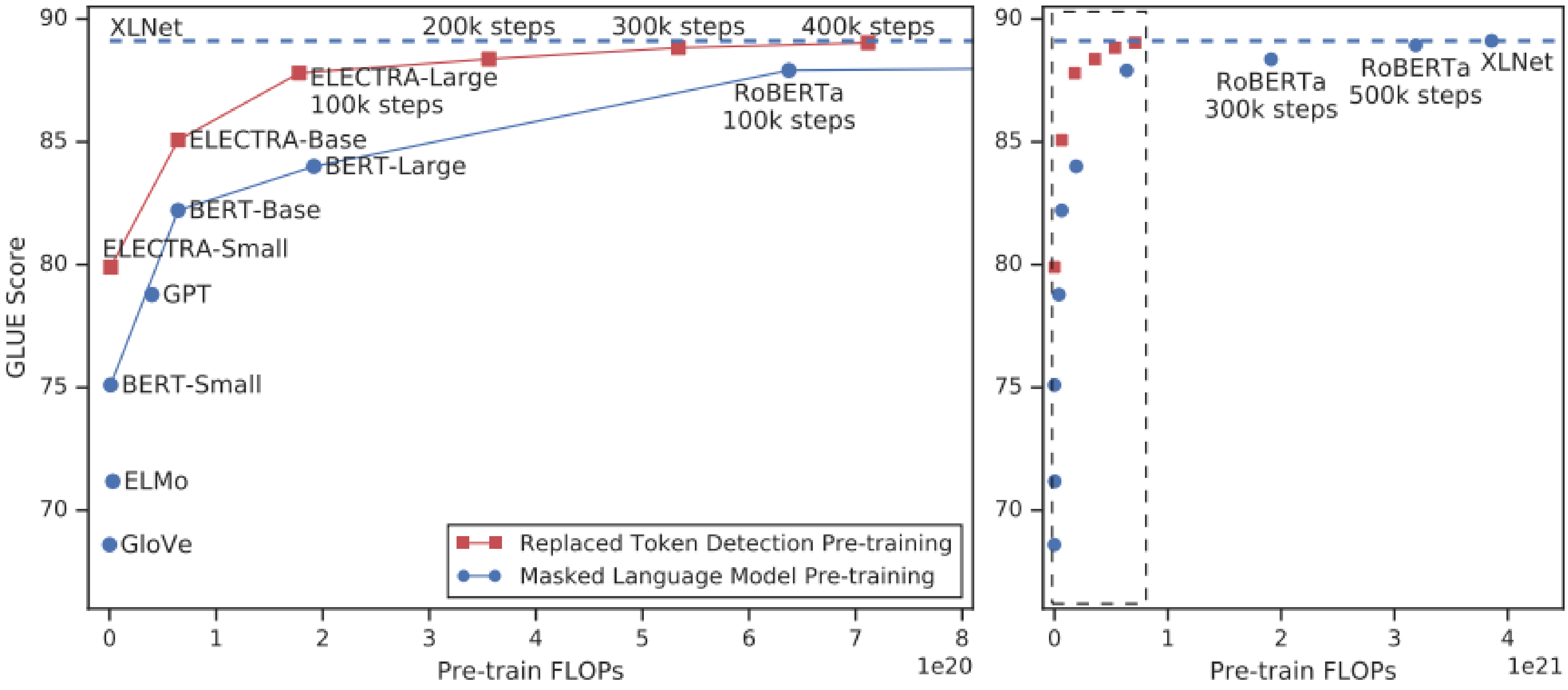
the chef ate the meal

ate

Small BERT

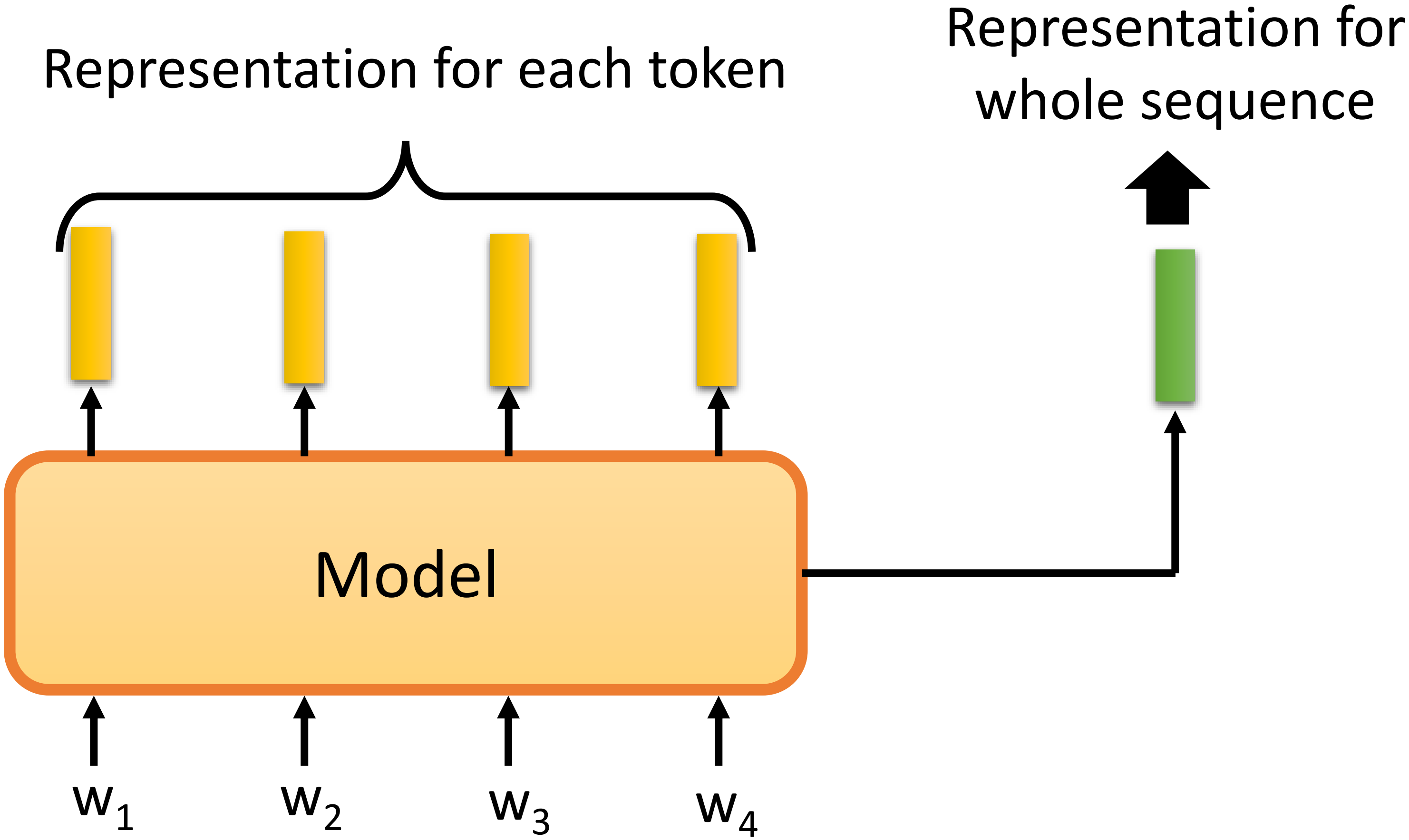
the chef mask the meal

Note: This is not GAN.



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

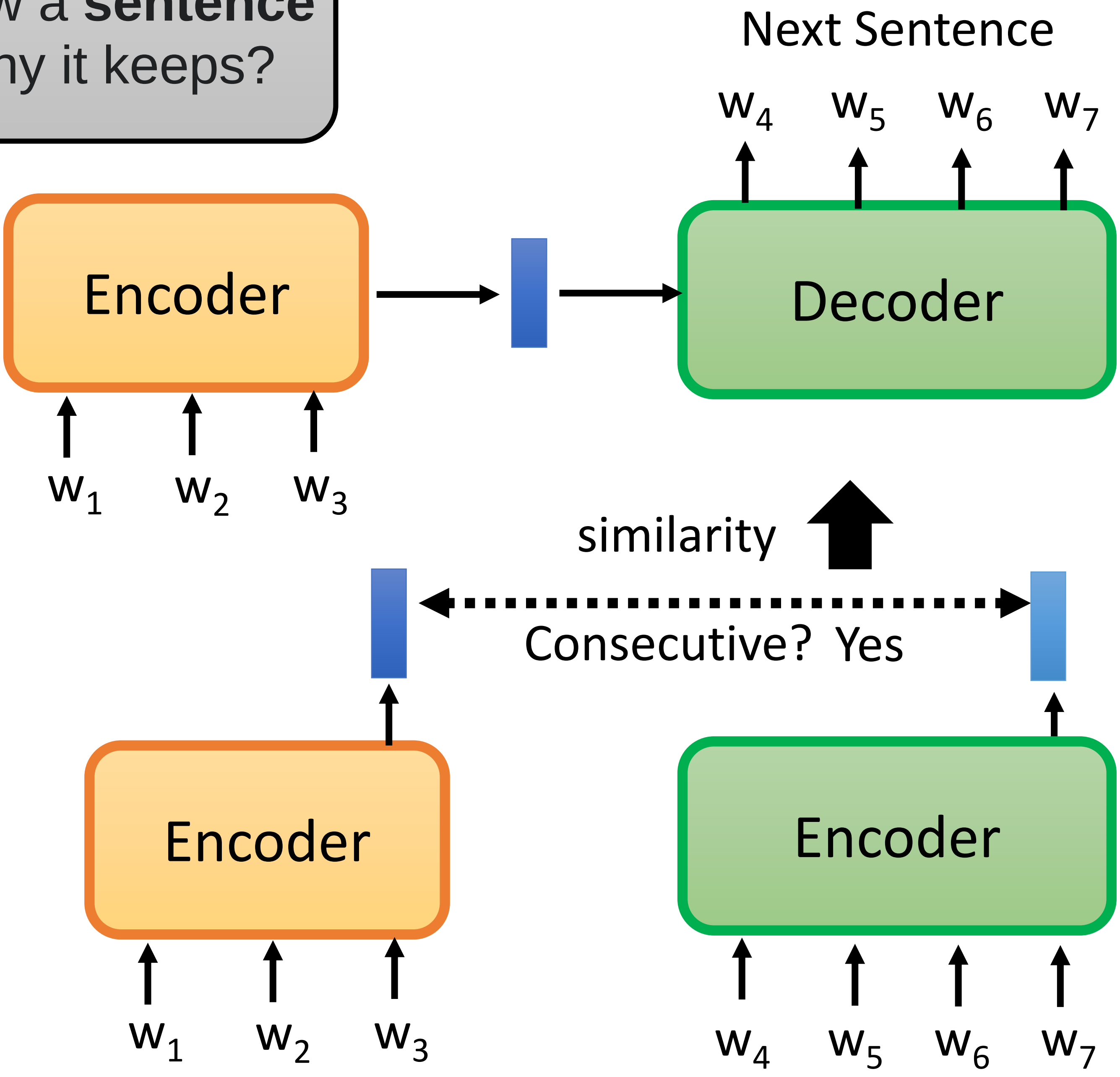
Sentence Level


Representation for each token
Representation for whole sequence
Model
w_1
w_2
w_3
w_4

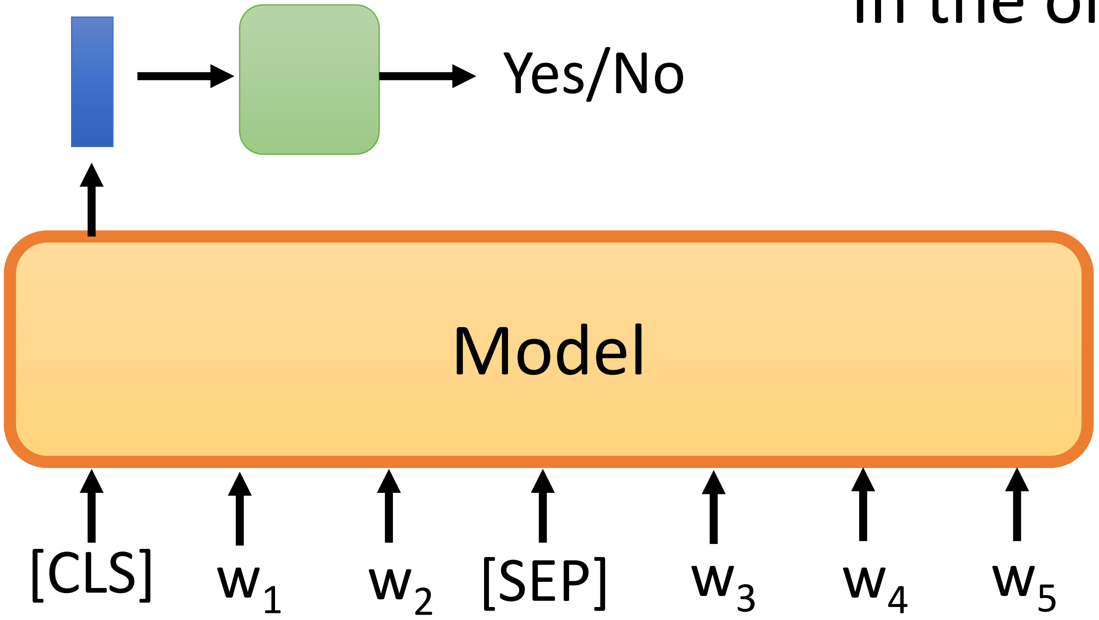
You shall know a **sentence** by the company it keeps?

Skip Thought

Quick Thought



In the original BERT, …..



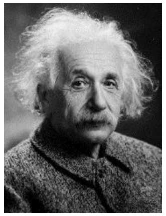
NSP: Next sentence prediction

Robustly optimized BERT approach (RoBERTa)

[Liu, et al., arXiv'19]

SOP: Sentence order prediction

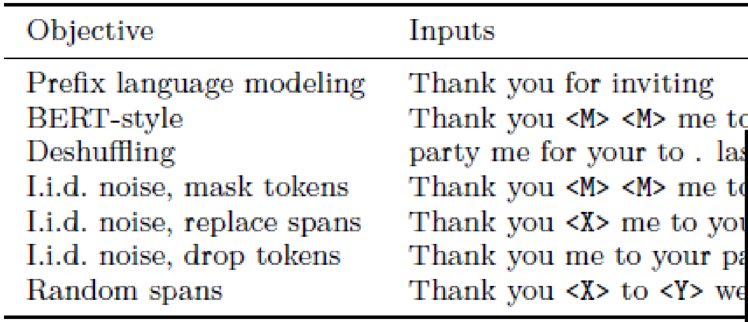
Used in ALBERT



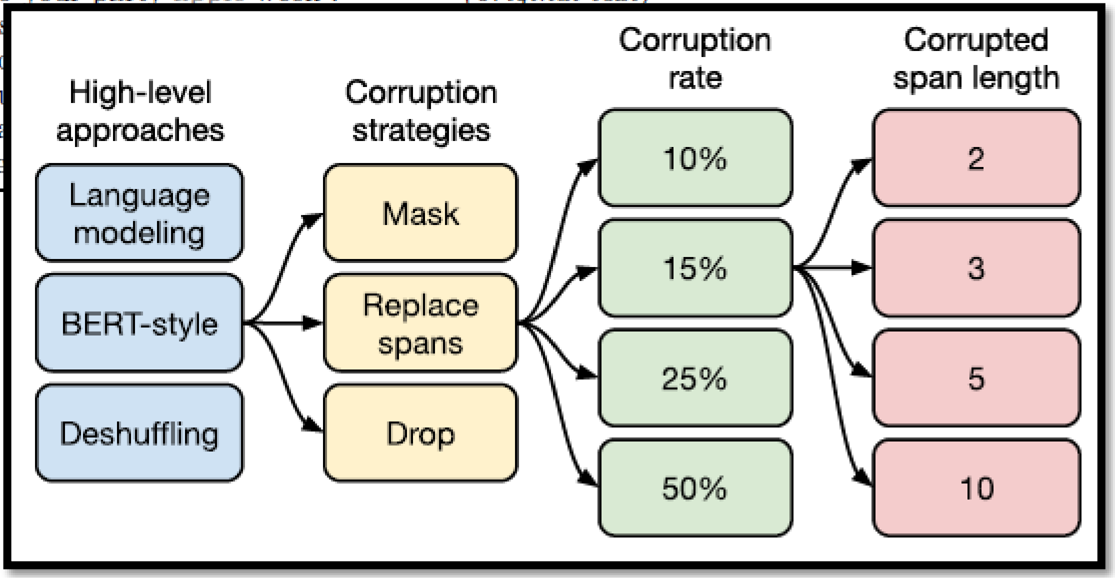
structBERT (Alice) [Want, et al., ICLR'20]

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

- 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 ……

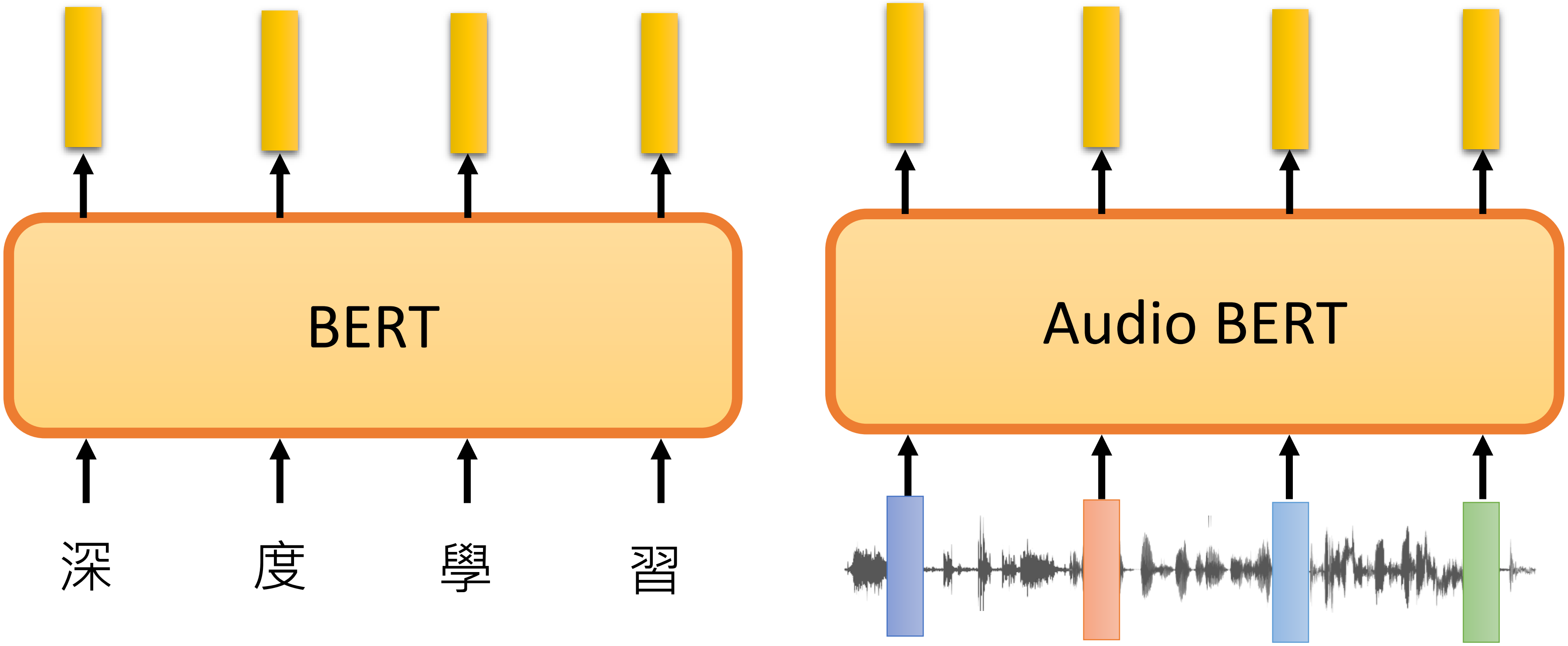
- **E**nhanced **L**anguage **R**epresentatio**N** with **I**nformative **E**ntities (ERNIE)

+

Concept
Concept
Concept
Concept
Concept
Concept

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