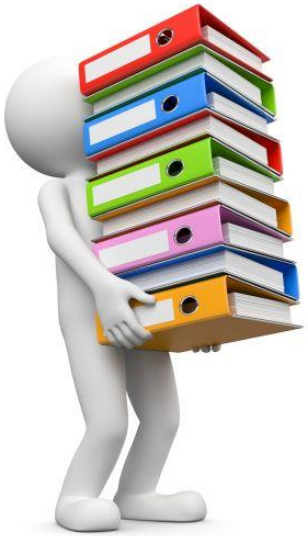
BERT and its family

Hung-yi Lee 李宏毅



Text without annotation

A model that can read text

Pre-train

Model

Task-specific data with annotation

Fine-tune

Task Specific

Model

Task Specific

Model

Task Specific

Model

Outline

What is pre-train model

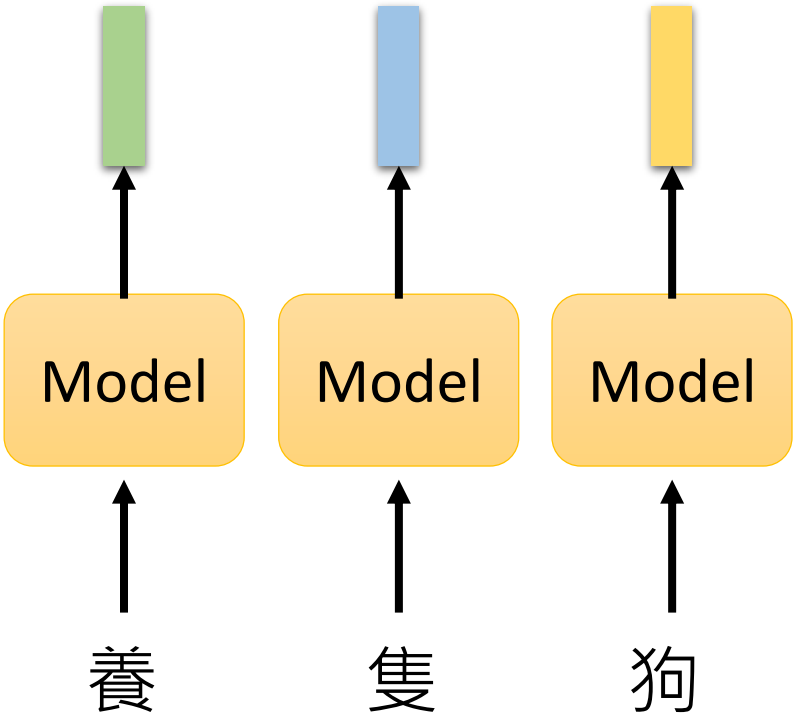
How to fine-tune

How to pre-train

Pre-train Model

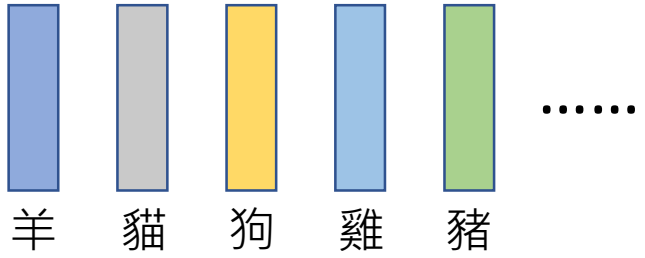
Pre-train Model

Represent each token by a embedding vector



The token with the same type has the same embedding.

Simply a table look-up

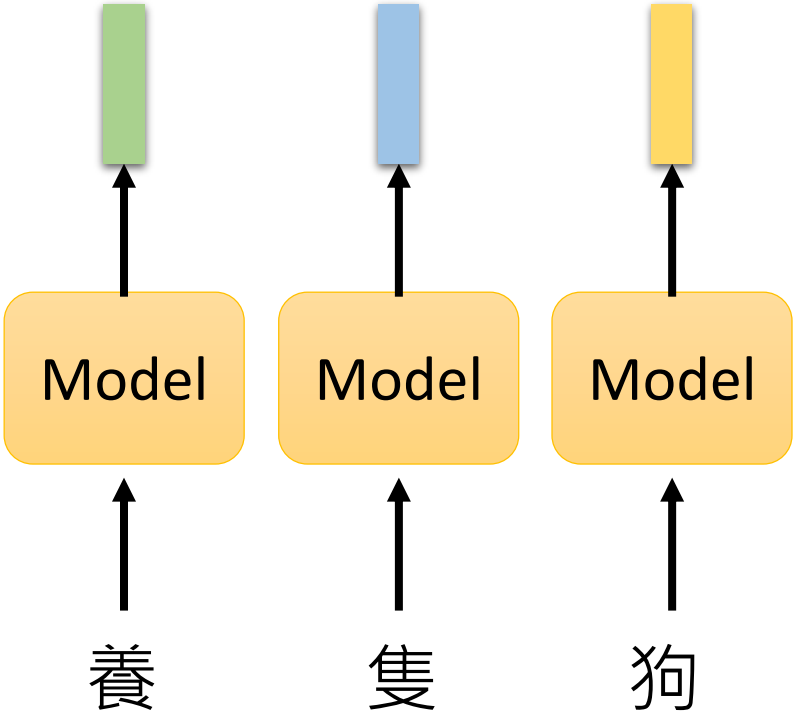


Word2vec [Mikolov, et al., NIPS'13]

Glove [Pennington, et al., EMNLP'14]

Pre-train Model

Represent each token by a embedding vector

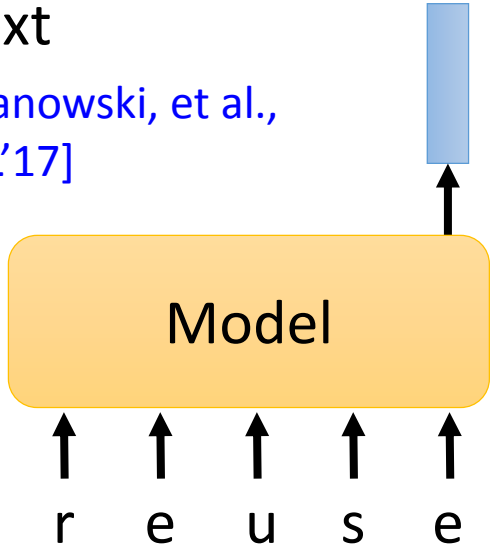


The token with the same type has the same embedding.

English word as token …

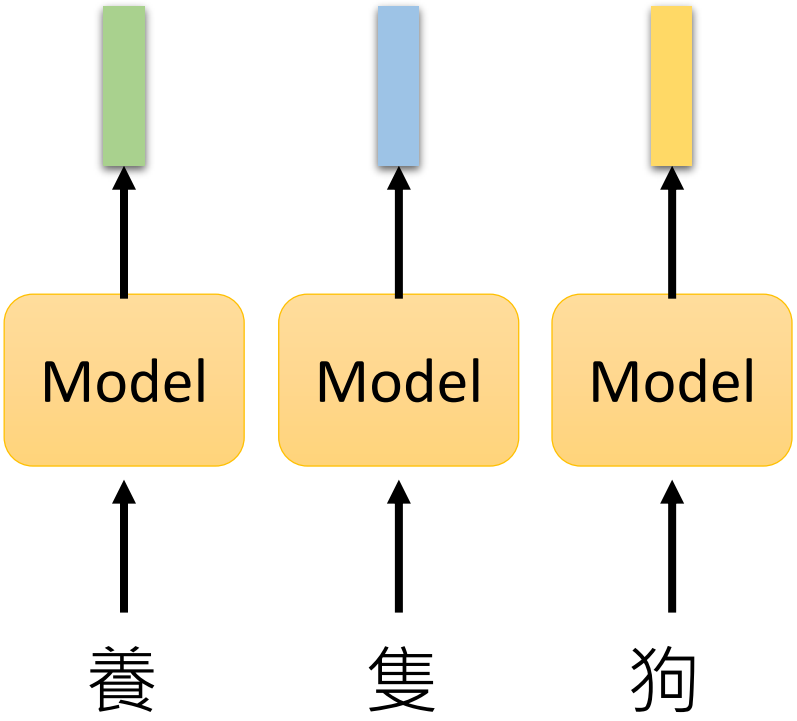
FastText

[Bojanowski, et al., TACL'17]



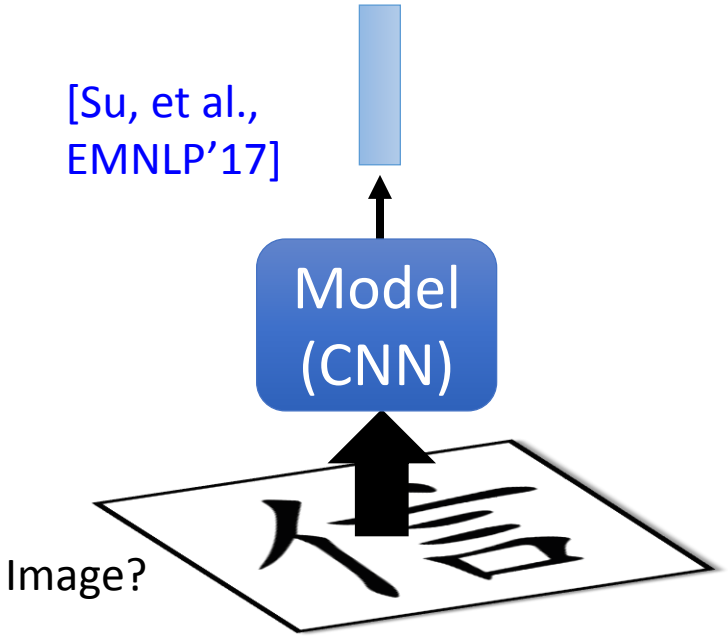
Pre-train Model

Represent each token by a embedding vector



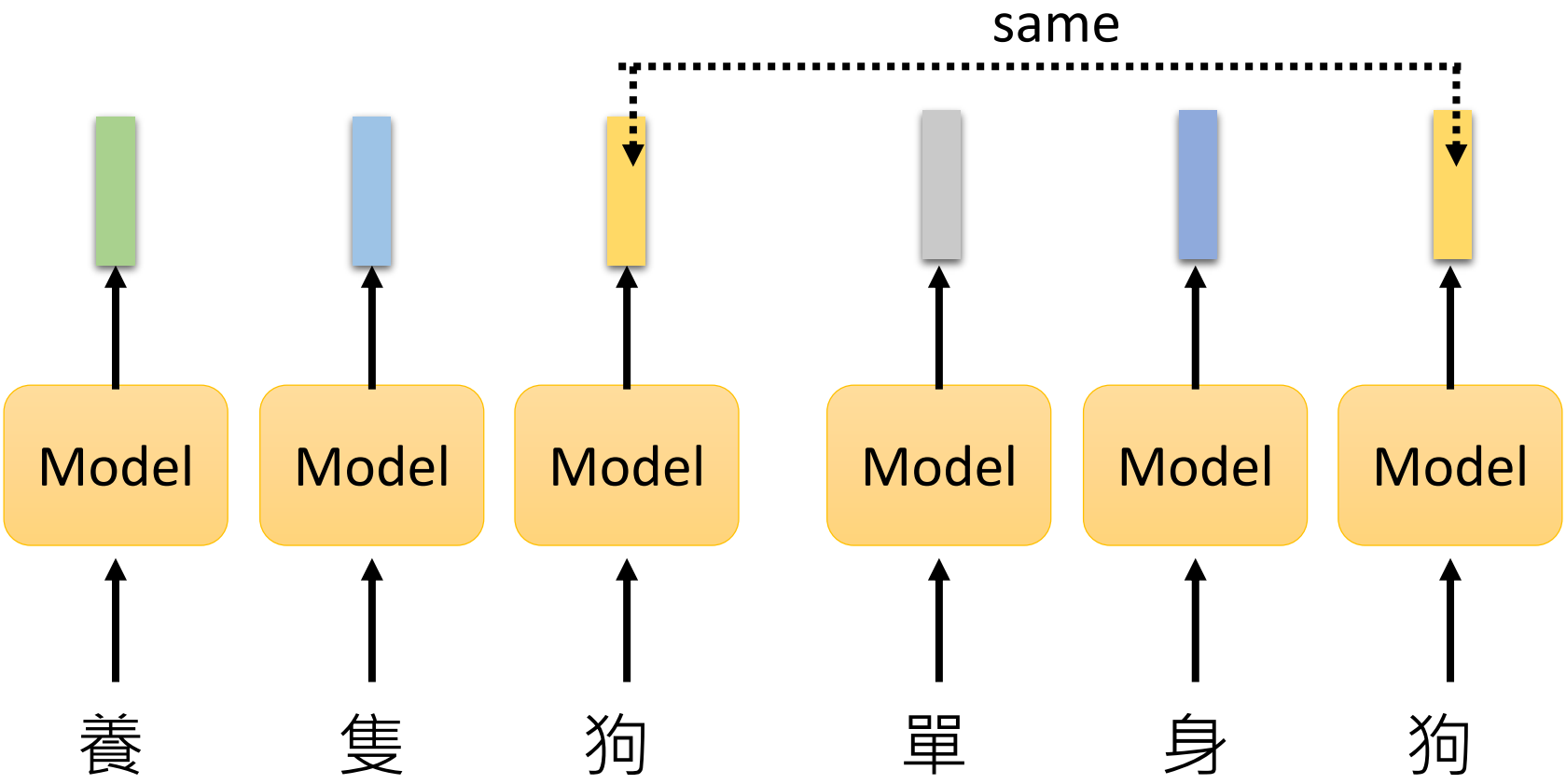
The token with the same type has the same embedding.

Chinese character as token …



Pre-train Model

Represent each token by a embedding vector



Pre-tra

Represent e

Model

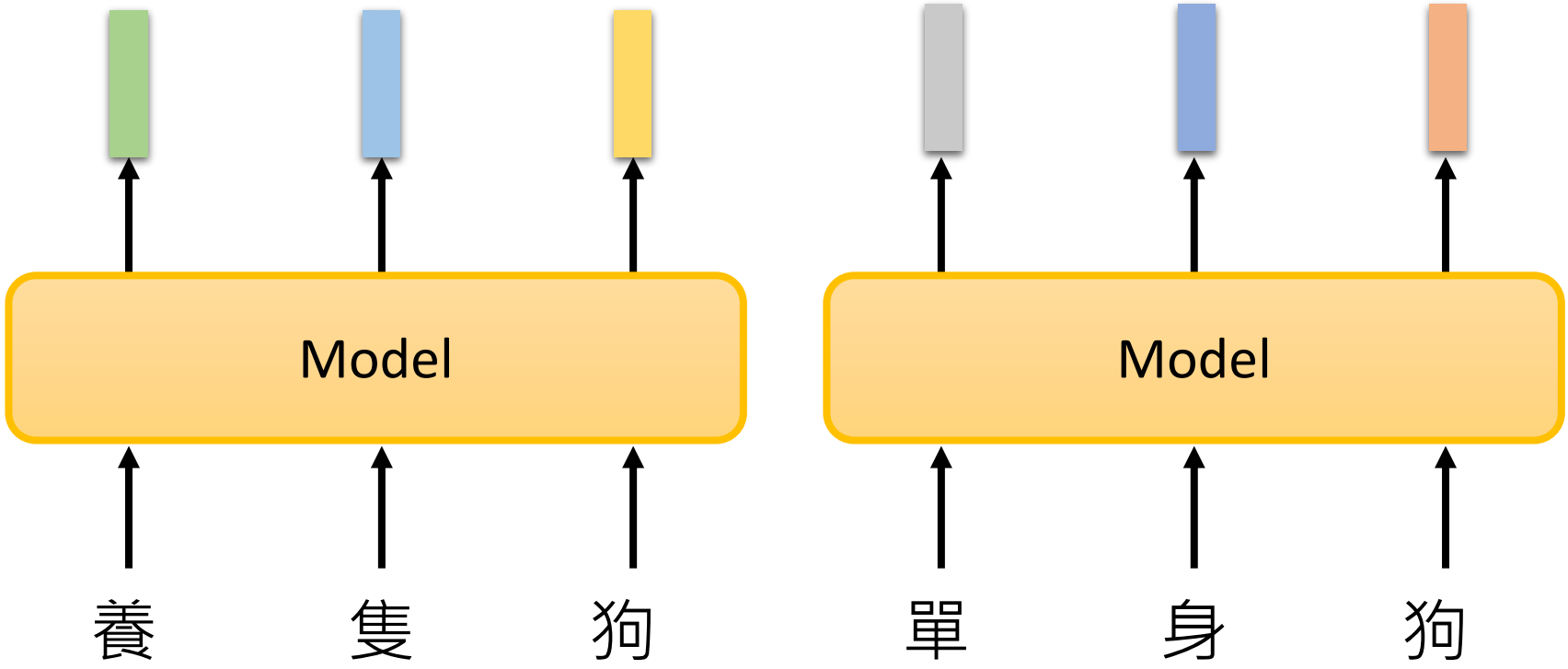
Model

養 隻 狗 單 身 狗

different

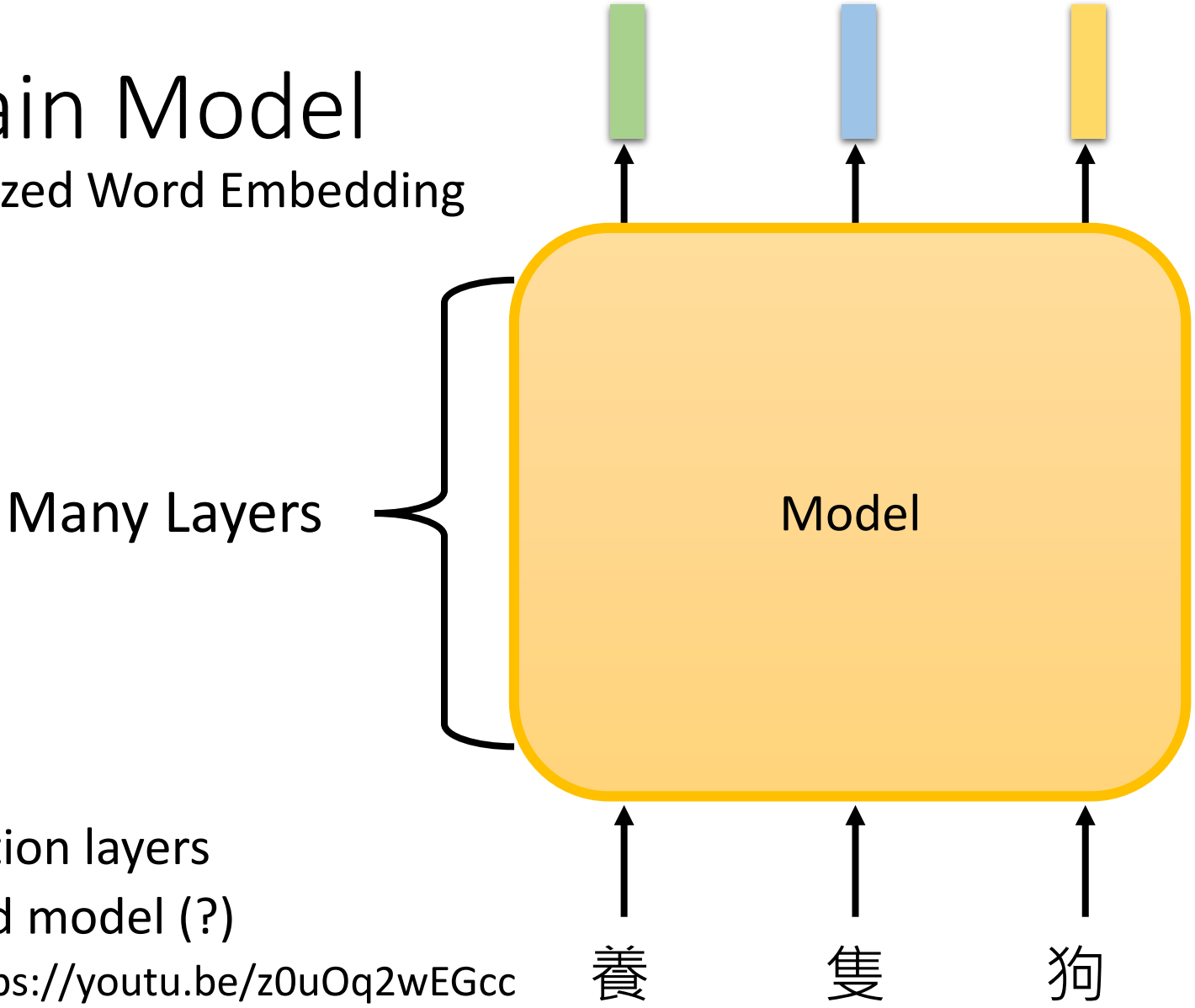
Pre-train Model

Contextualized Word Embedding

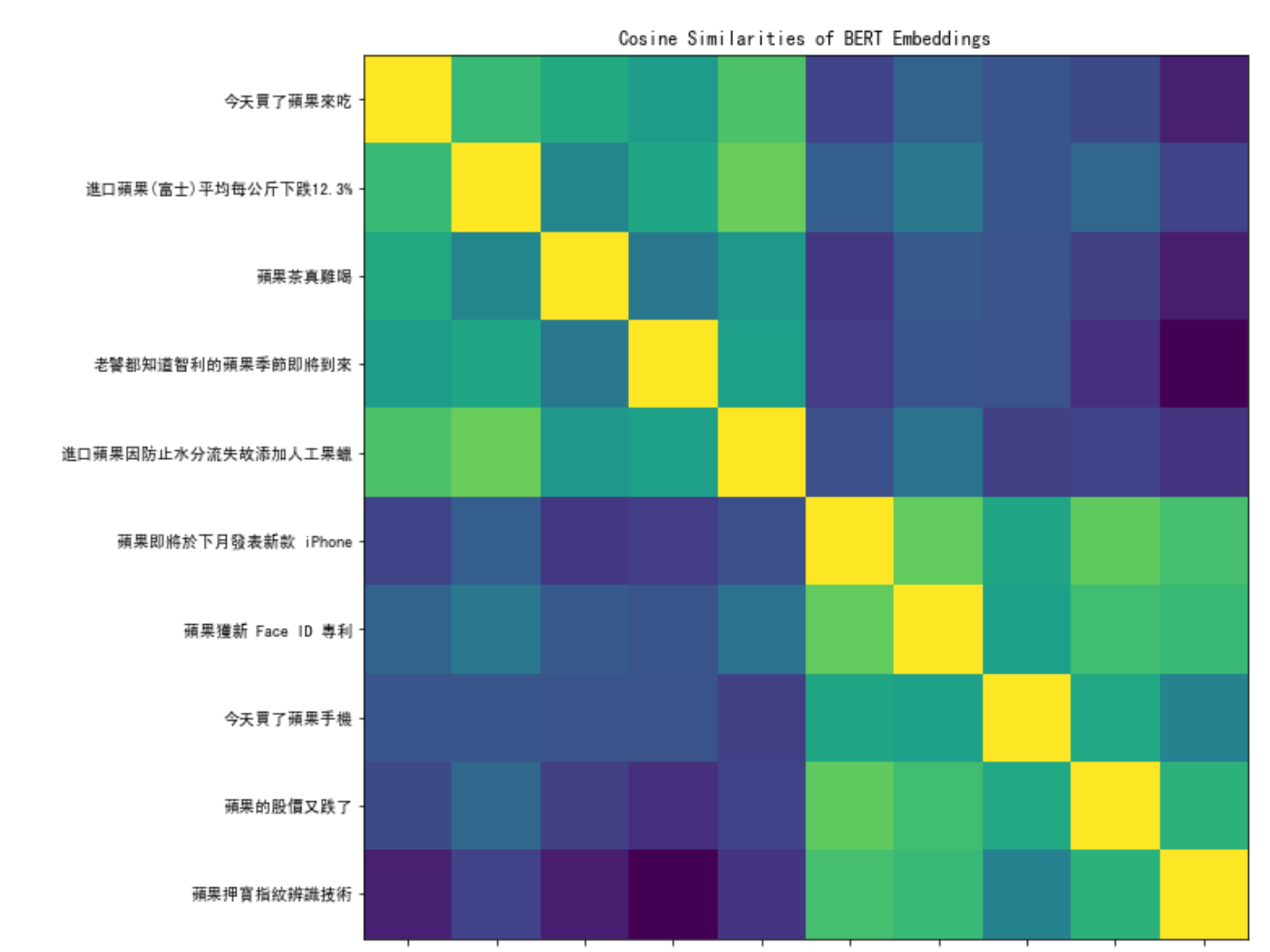


Pre-train Model

Contextualized Word Embedding

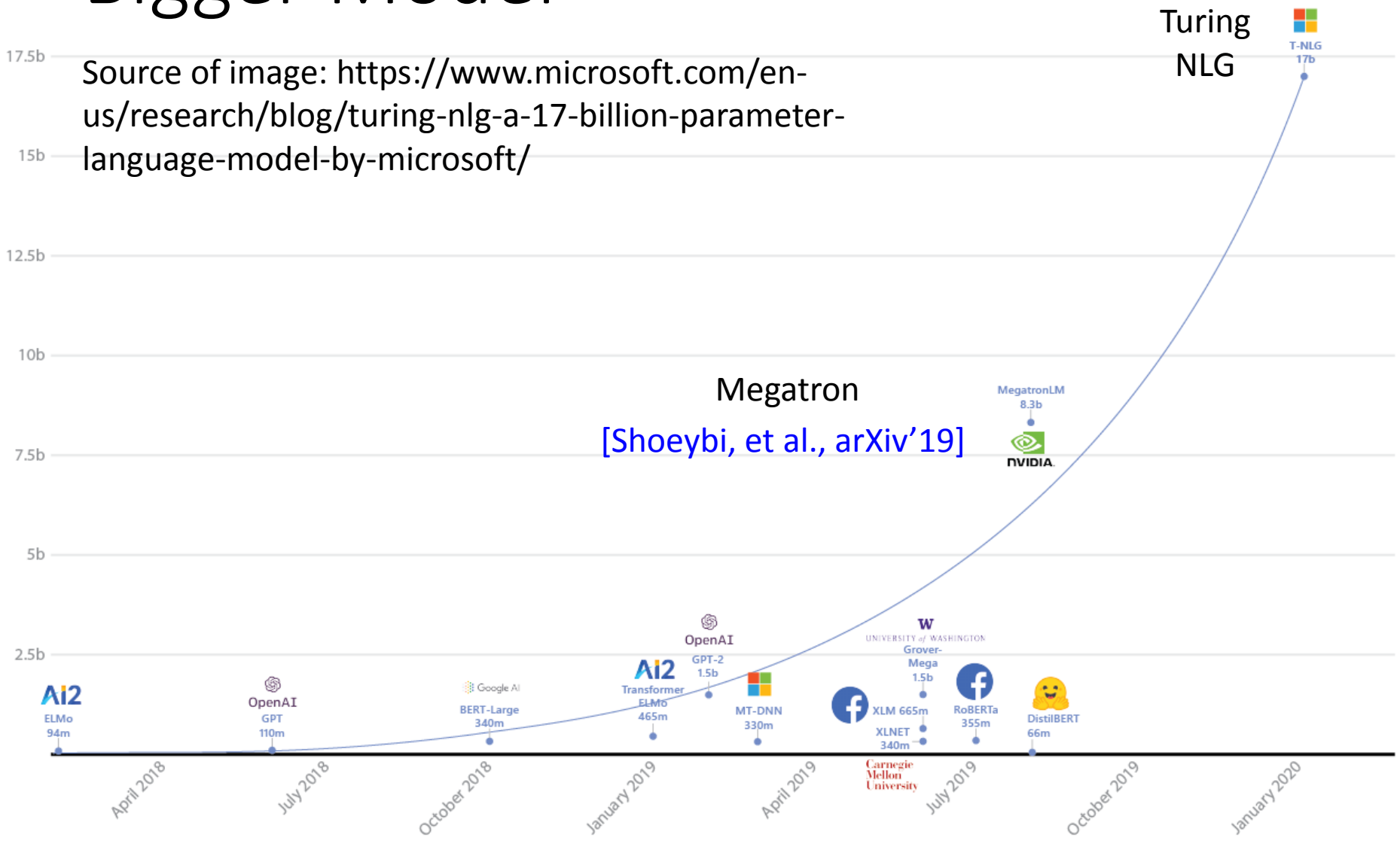


- LSTM
- Self-attention layers
- Tree-based model (?)
  - Ref: https://youtu.be/z0uOq2wEGcc


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/



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

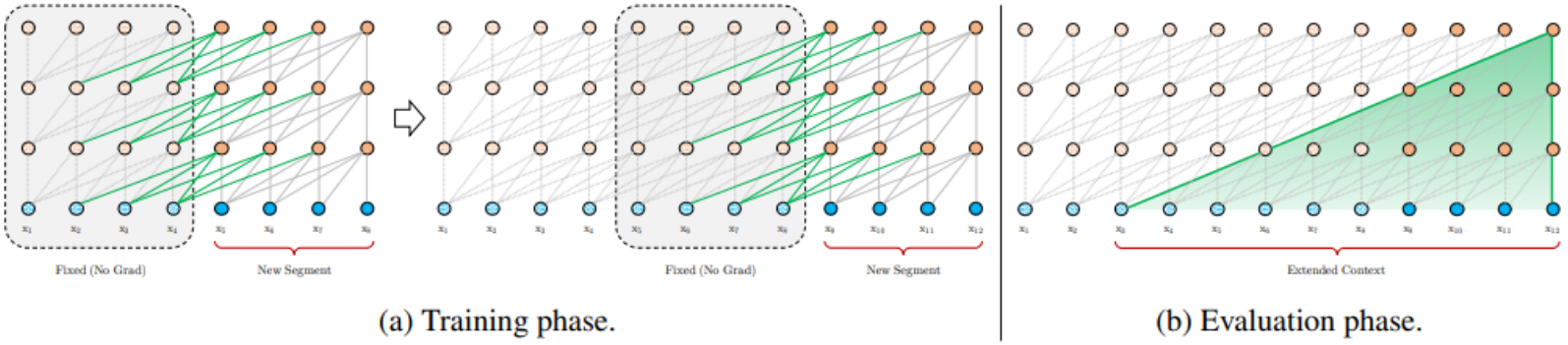
Ref: https://youtu.be/dPp8rCAnU_A

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]

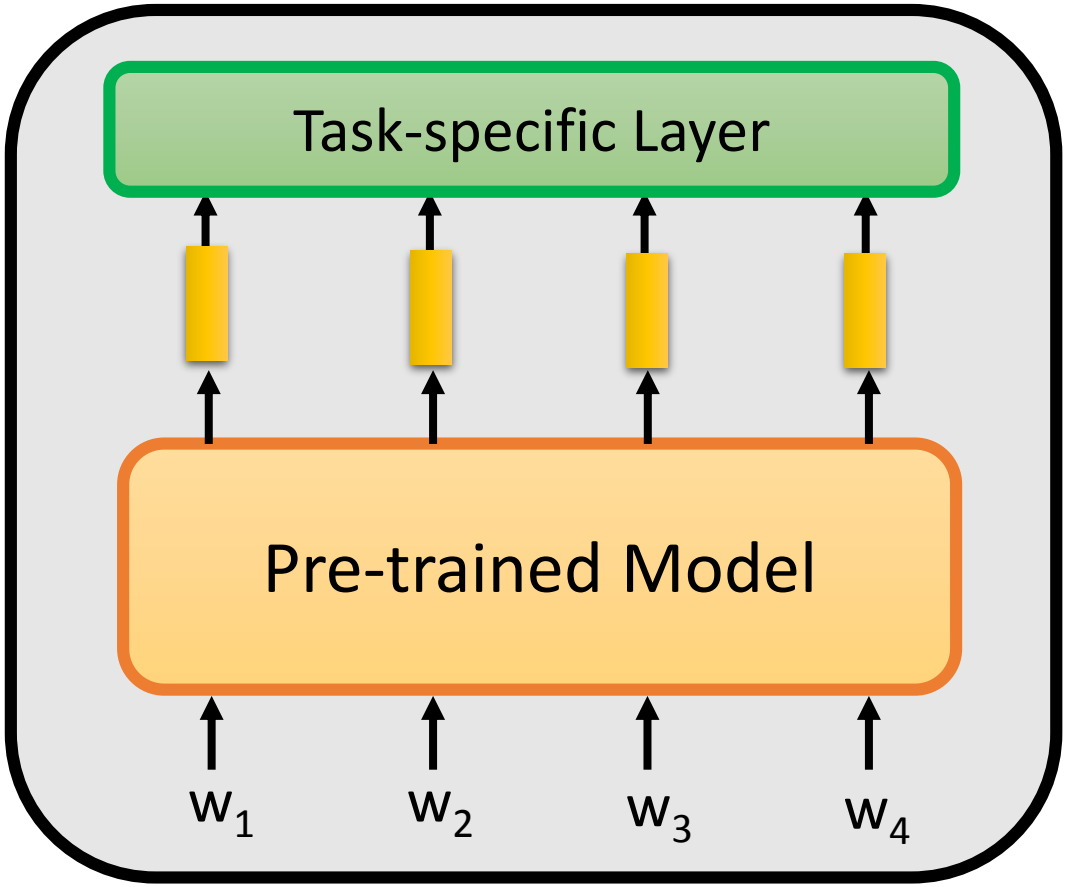


(a) Training phase. (b) Evaluation phase.

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

} Reduce the complexity of self-attention

How to fine-tune



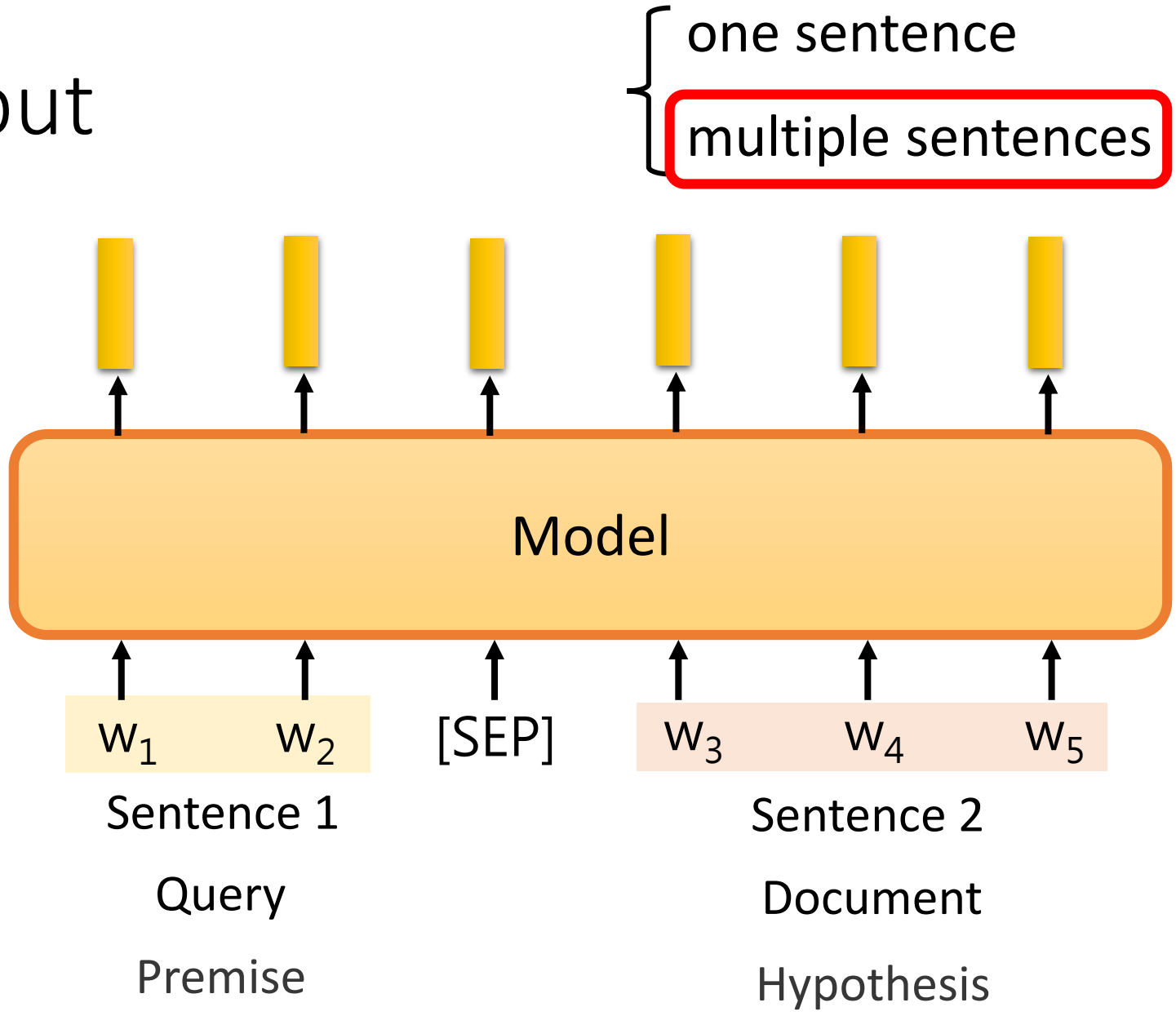
For a specific NLP task

NLP tasks

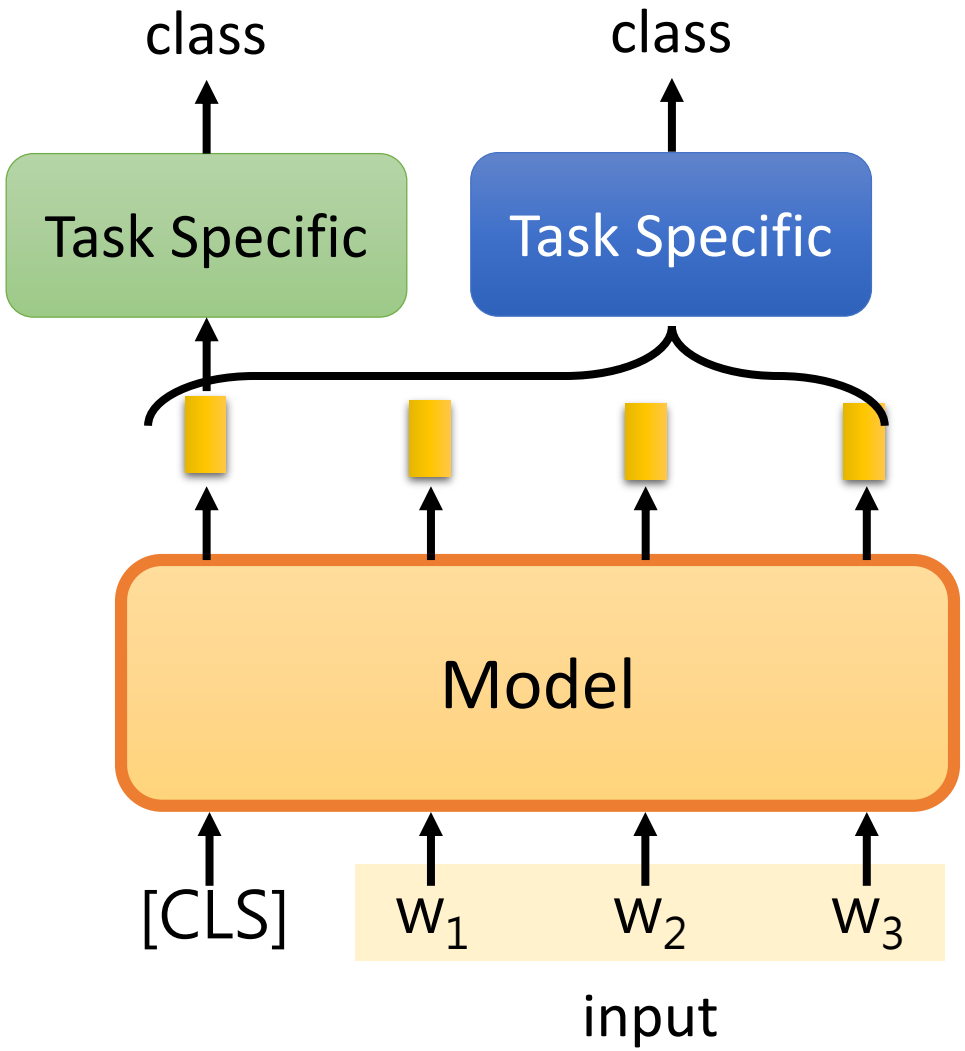
Input
- one sentence
- multiple sentences

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

Input

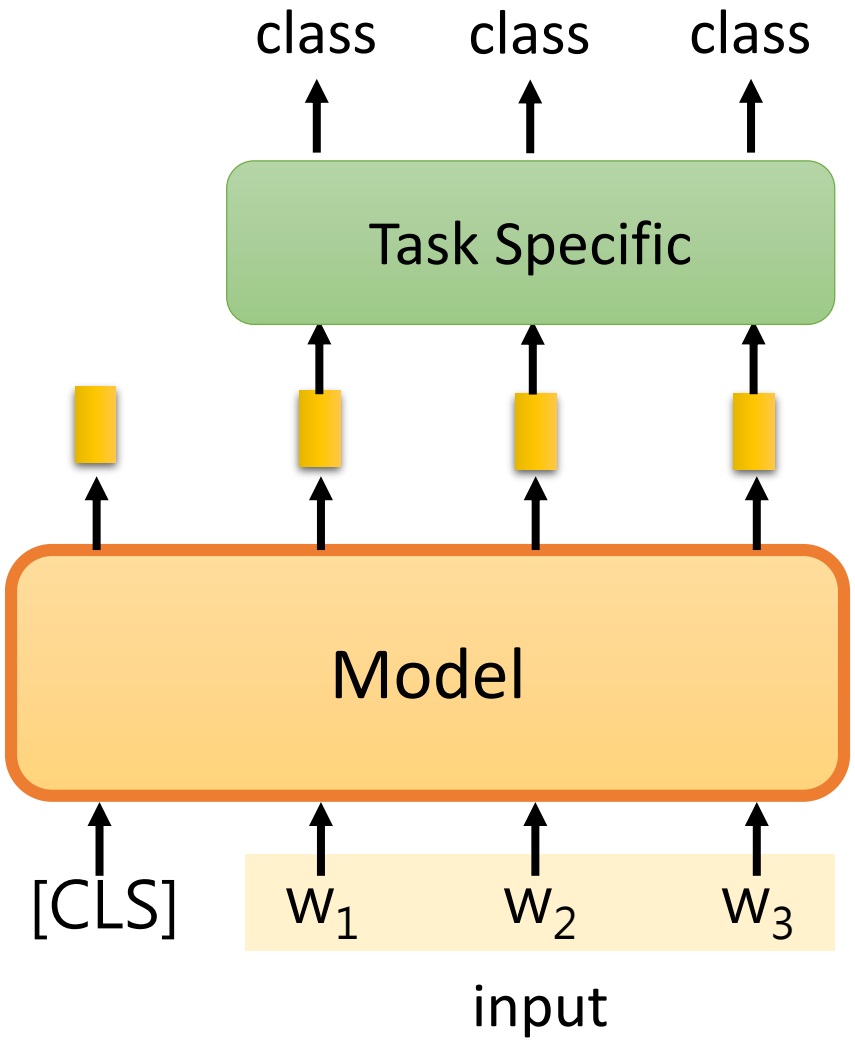

one sentence
multiple sentences
Model
w_1
w_2
[SEP]
w_3
w_4
w_5
Sentence 1
Query
Premise
Sentence 2
Document
Hypothesis

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

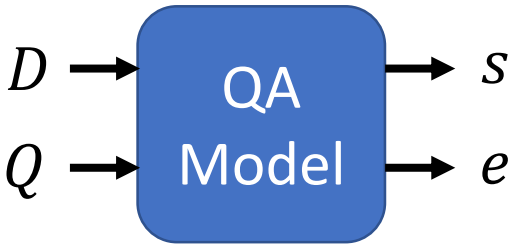
class for each token

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

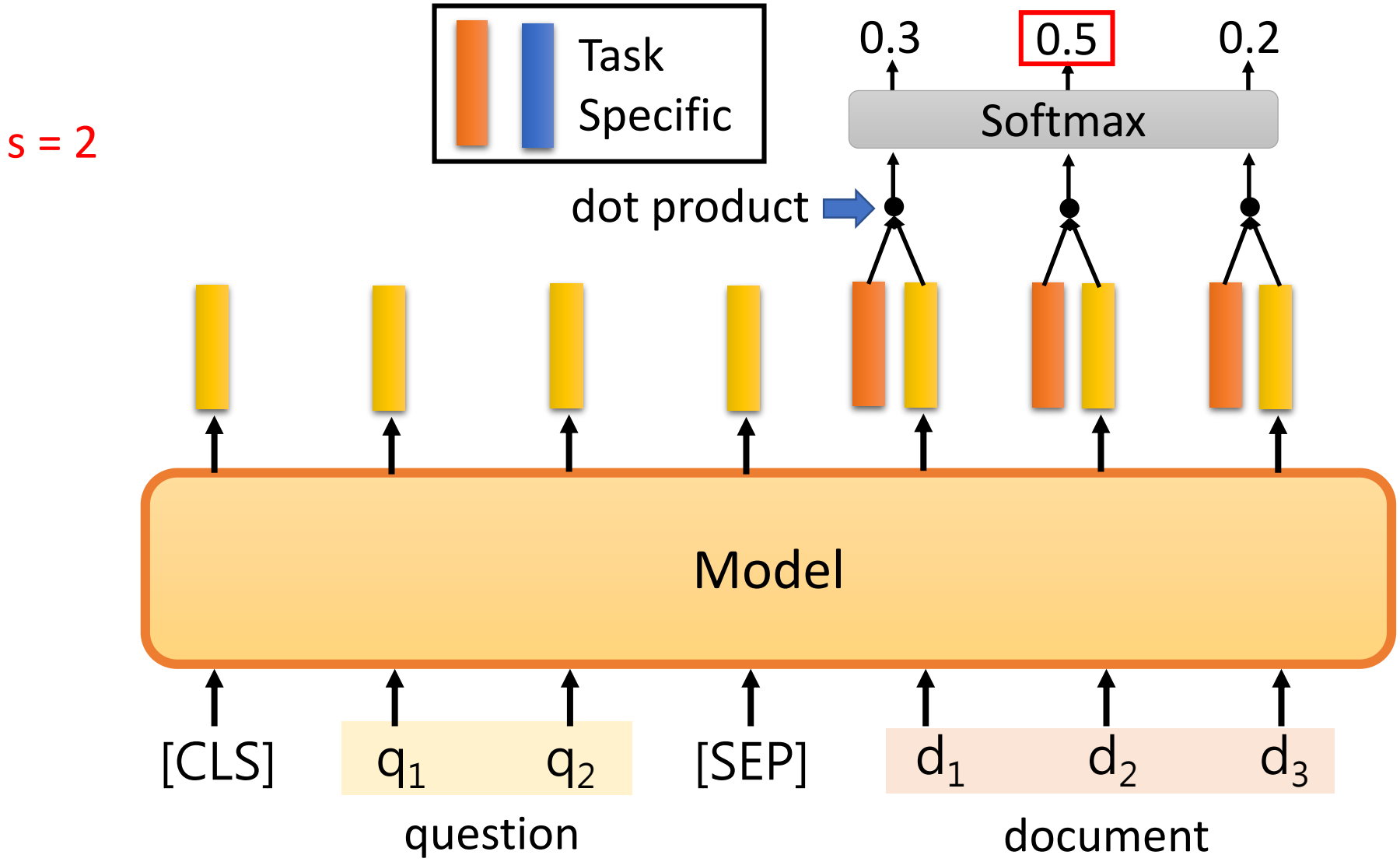
In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Preci 77 on 79 as smaller droplets coalesce via ion other rain drops or ice crystals within a cloud.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

What causes precipitation to fall?
gravity

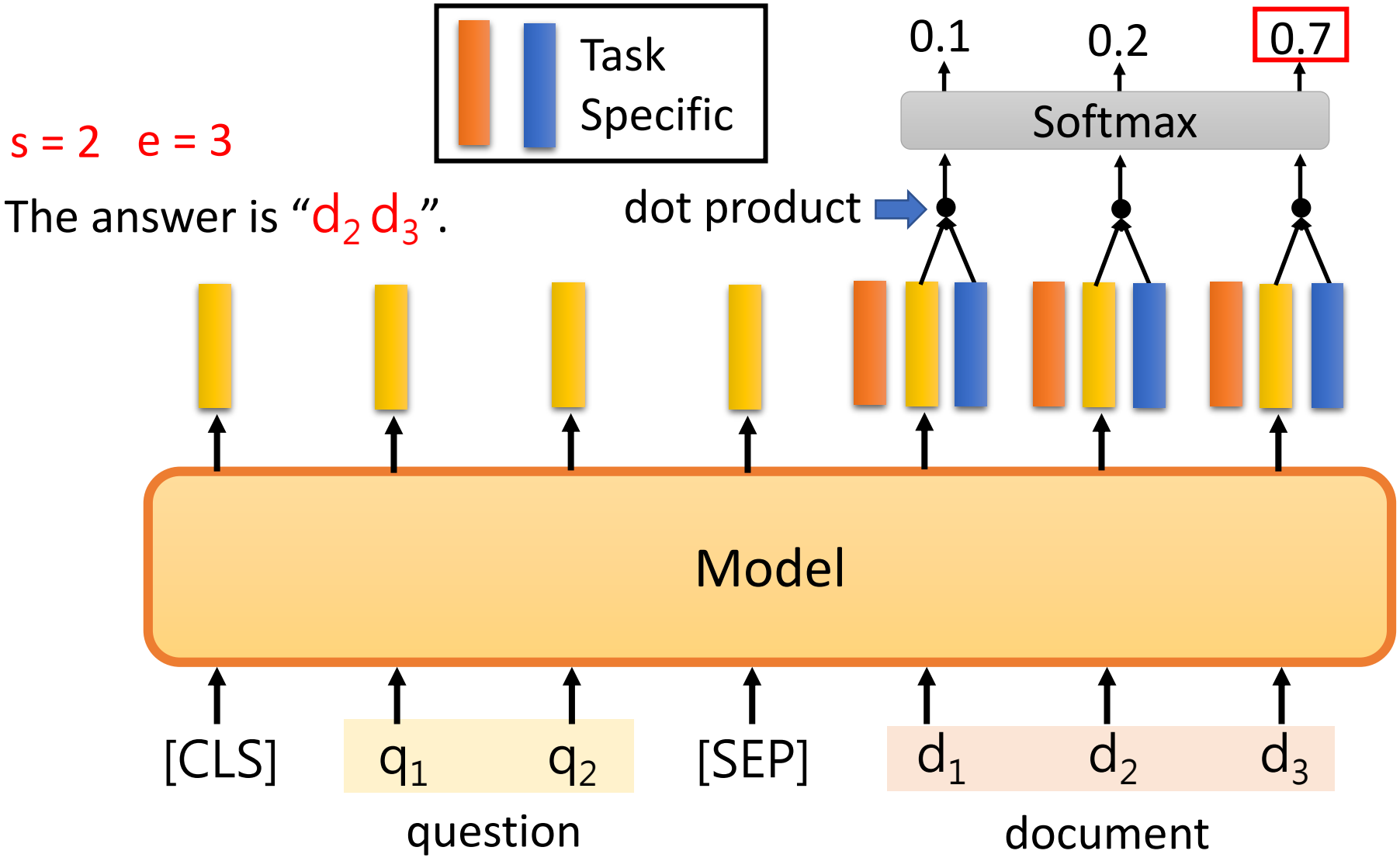
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

$s = 77, e = 79$

Copy from Input (BERT)

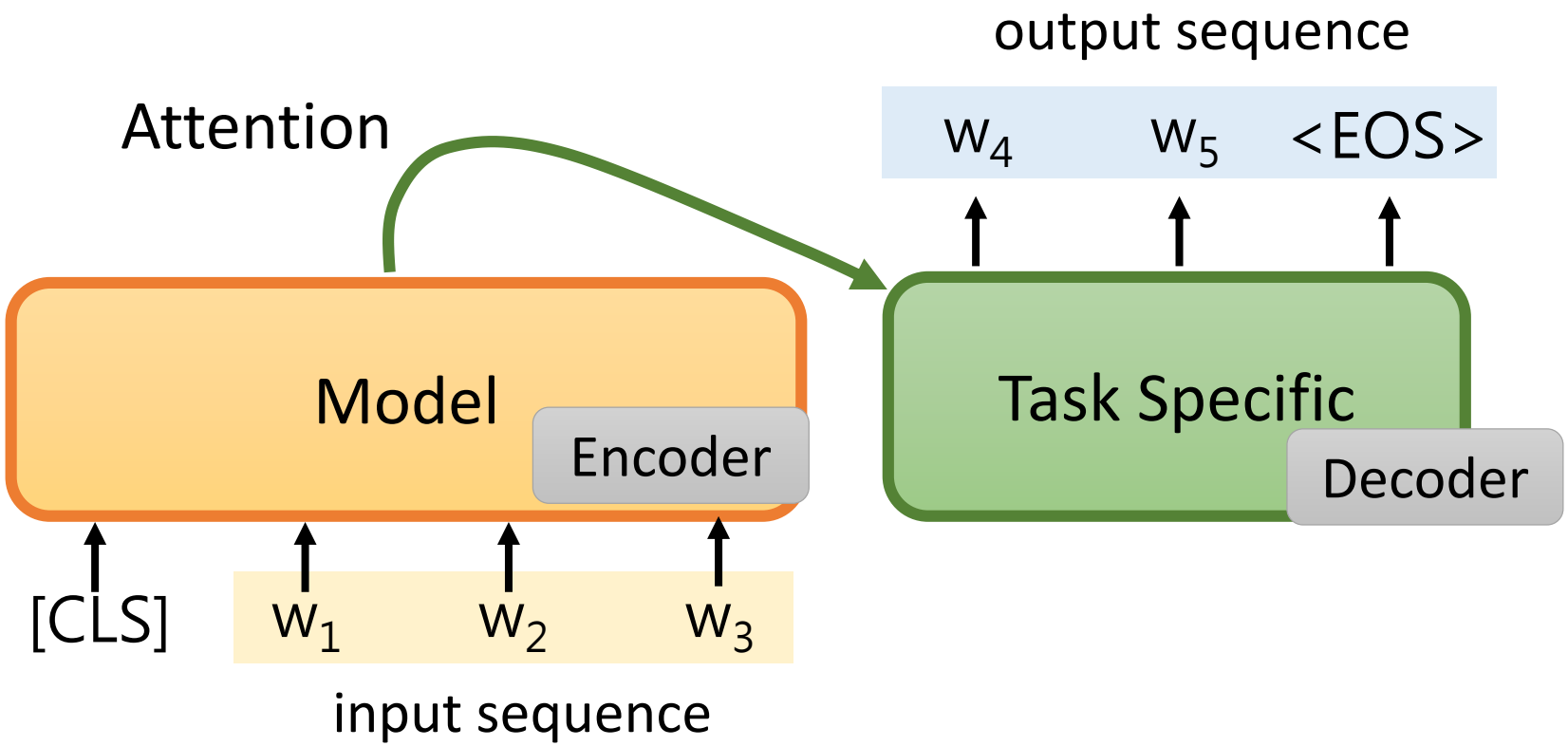


Copy from Input (BERT)

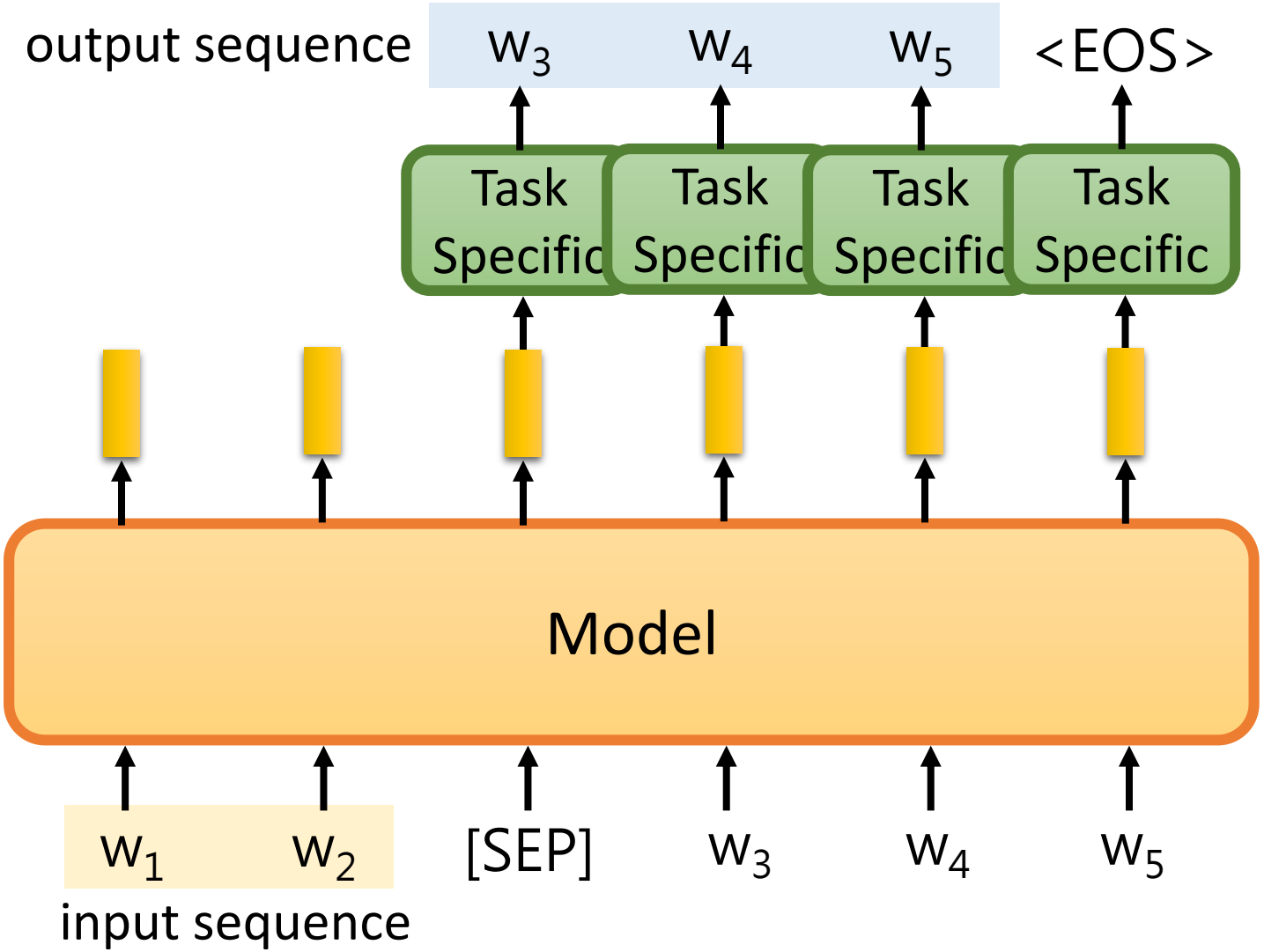


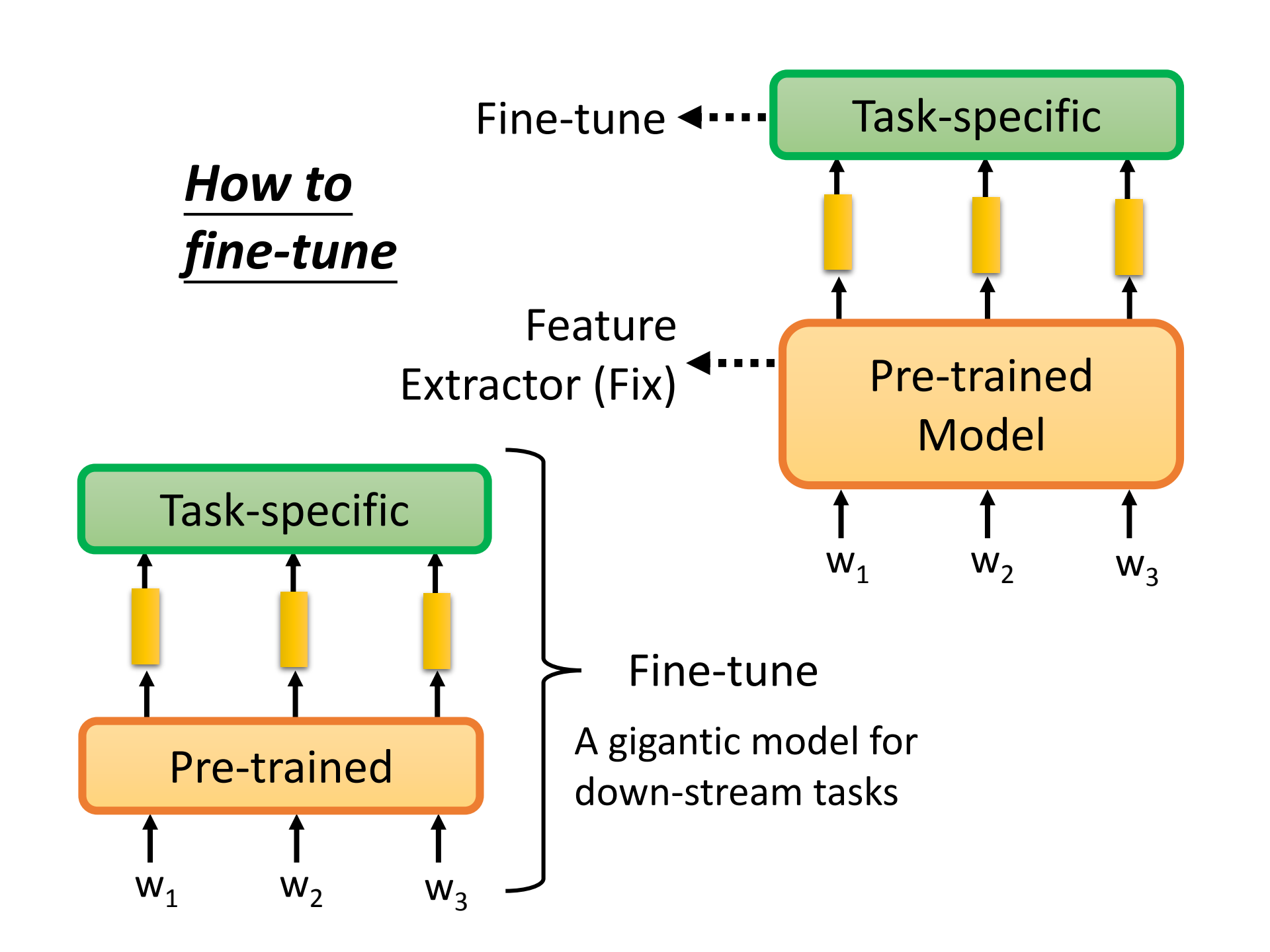
Output – General Sequence (v1)

- Seq2seq model

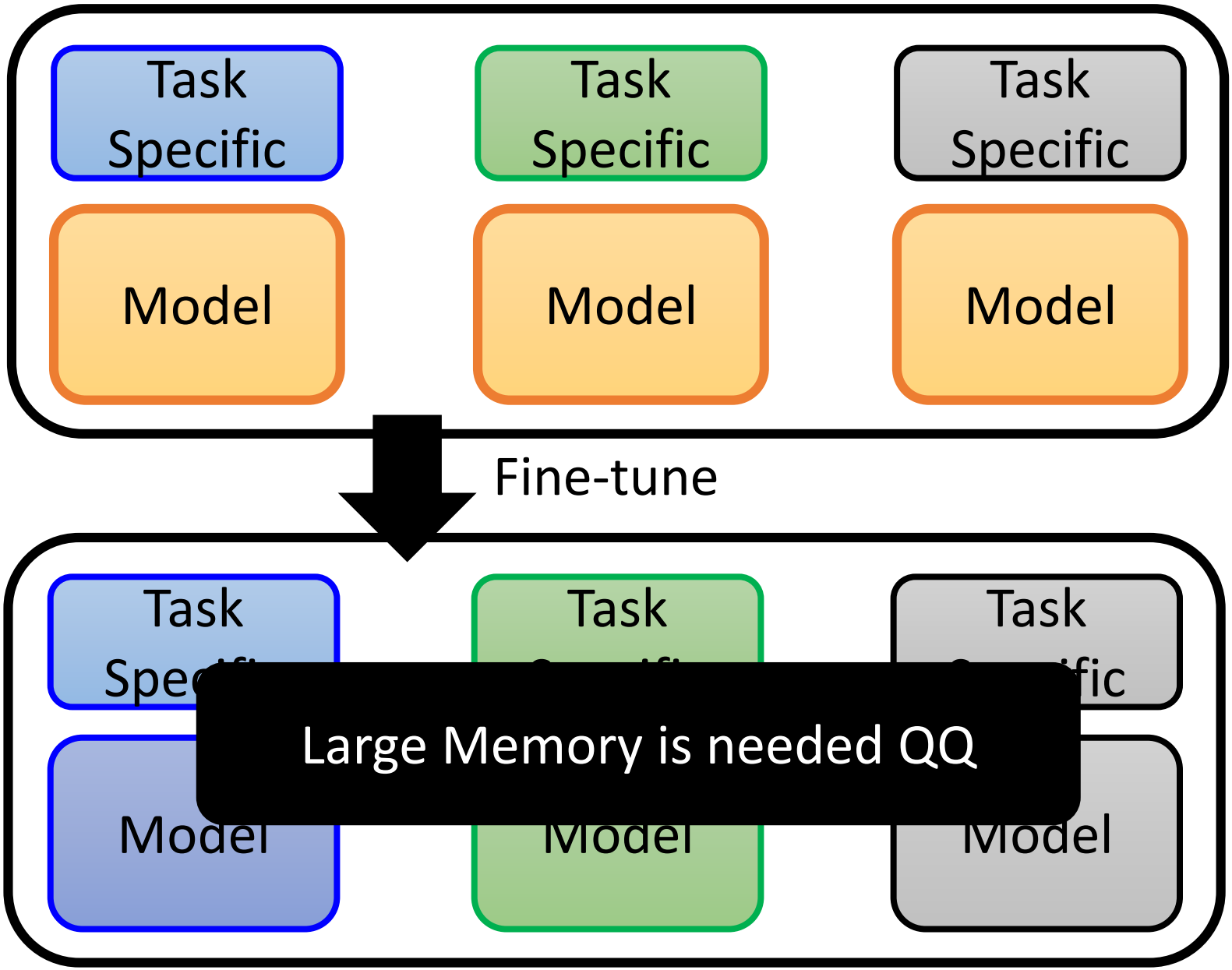


Output – General Sequence (v2)

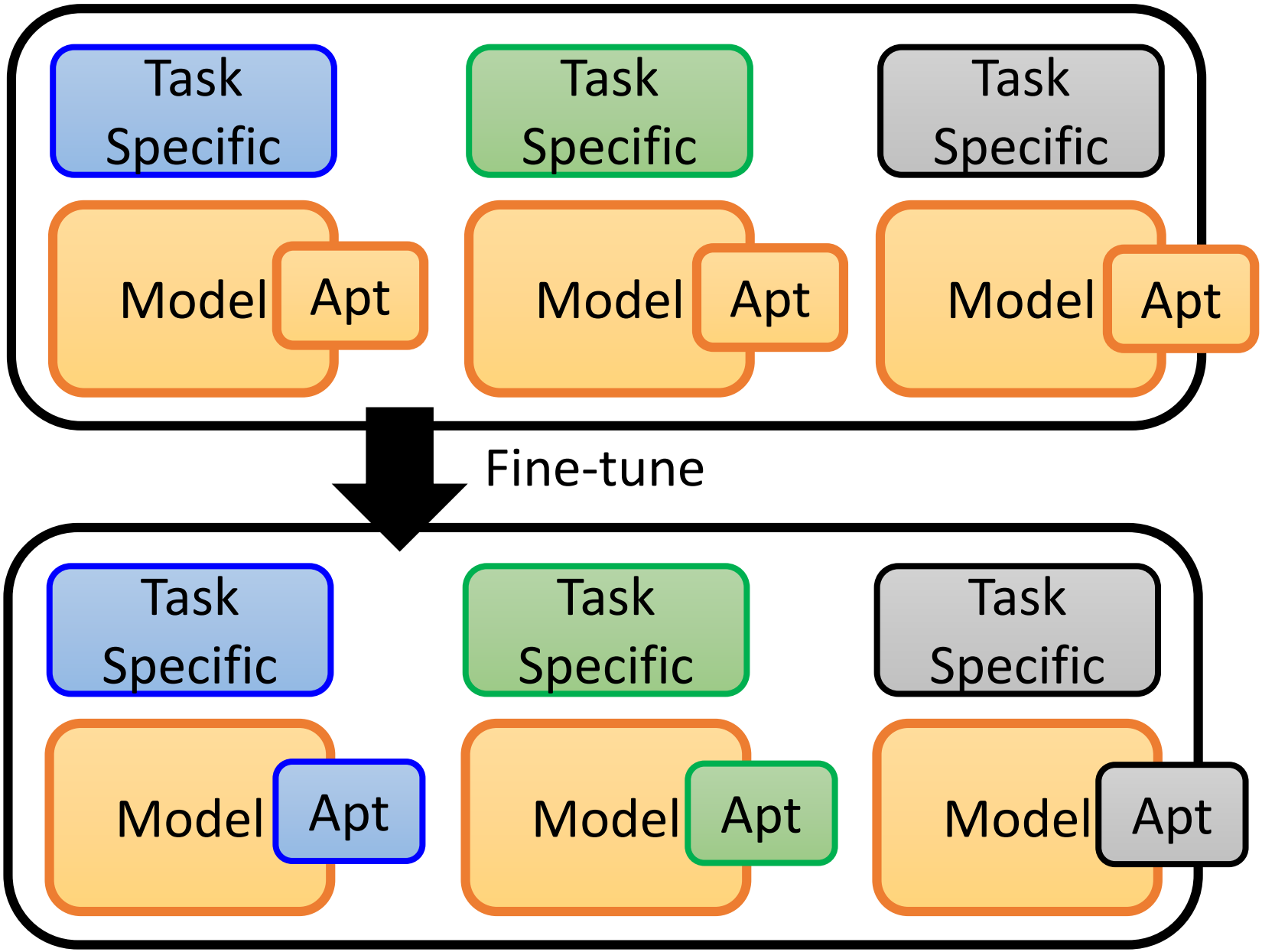


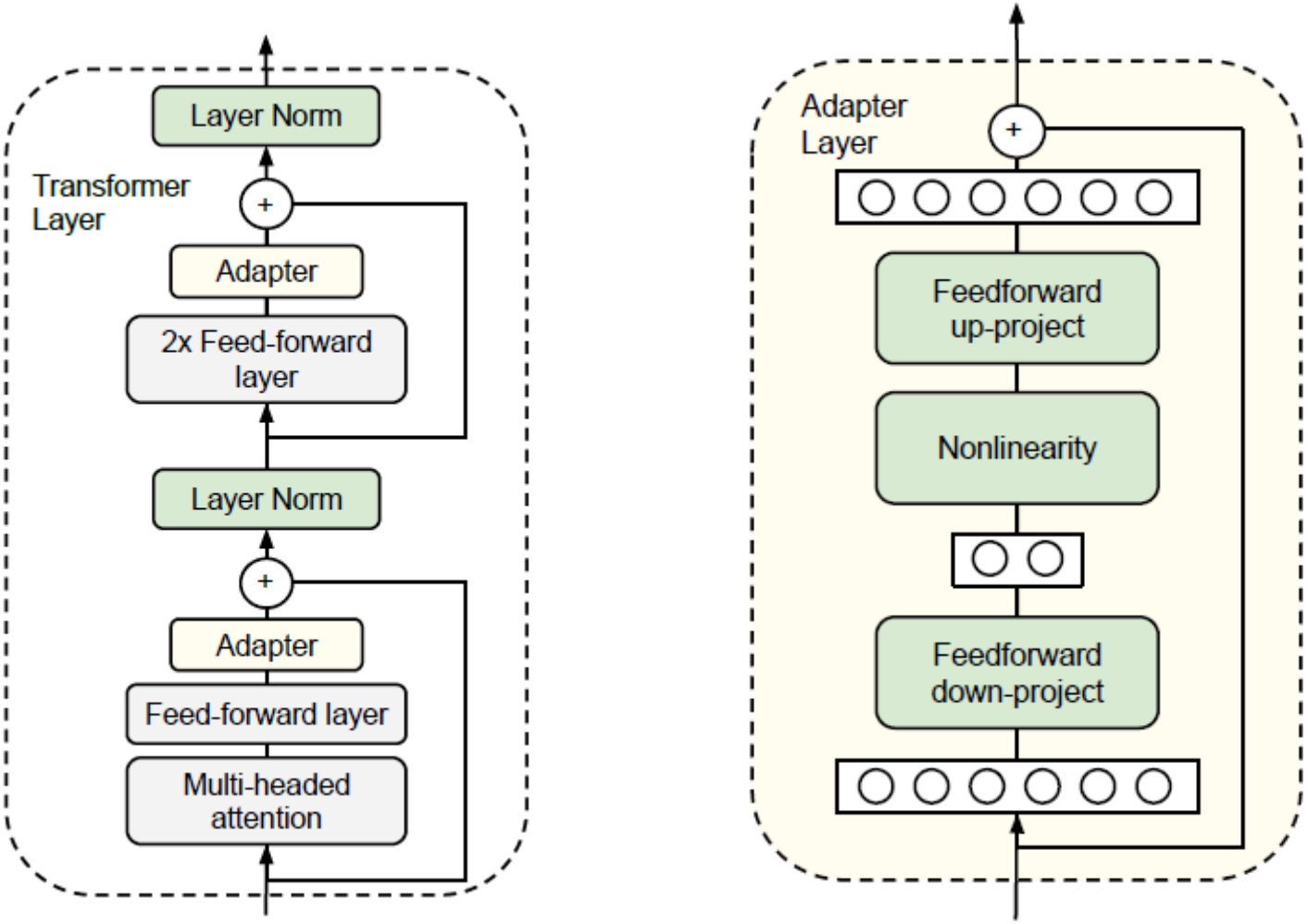

How to fine-tune
Fine-tune
Task-specific
Feature Extractor (Fix)
Pre-trained Model
w1
w2
w3
Task-specific
Pre-trained
w1
w2
w3
Fine-tune
A gigantic model for down-stream tasks

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



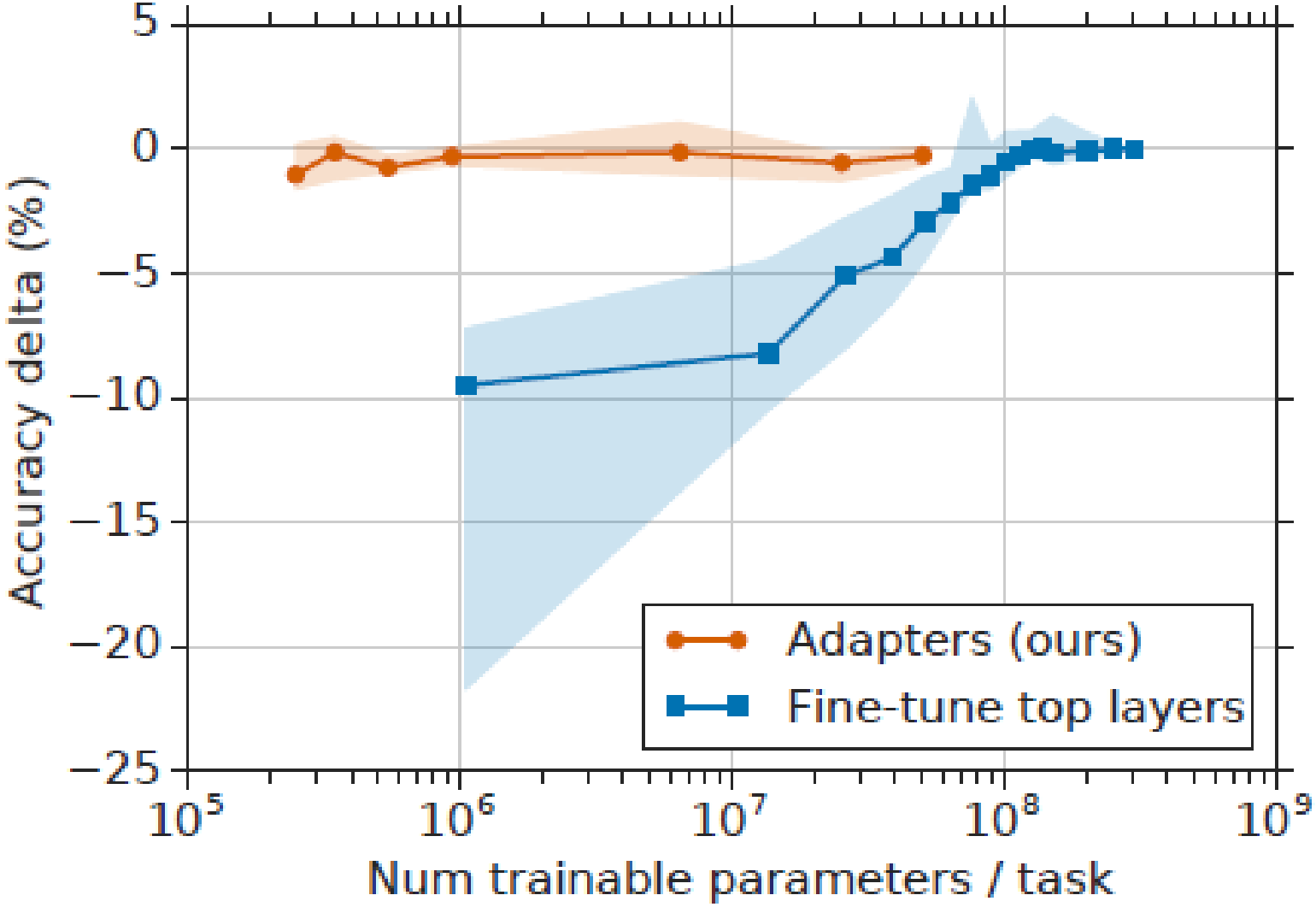
Adaptor [Stickland, et al., ICML'19] [Houlsby, et al., ICML'19]





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

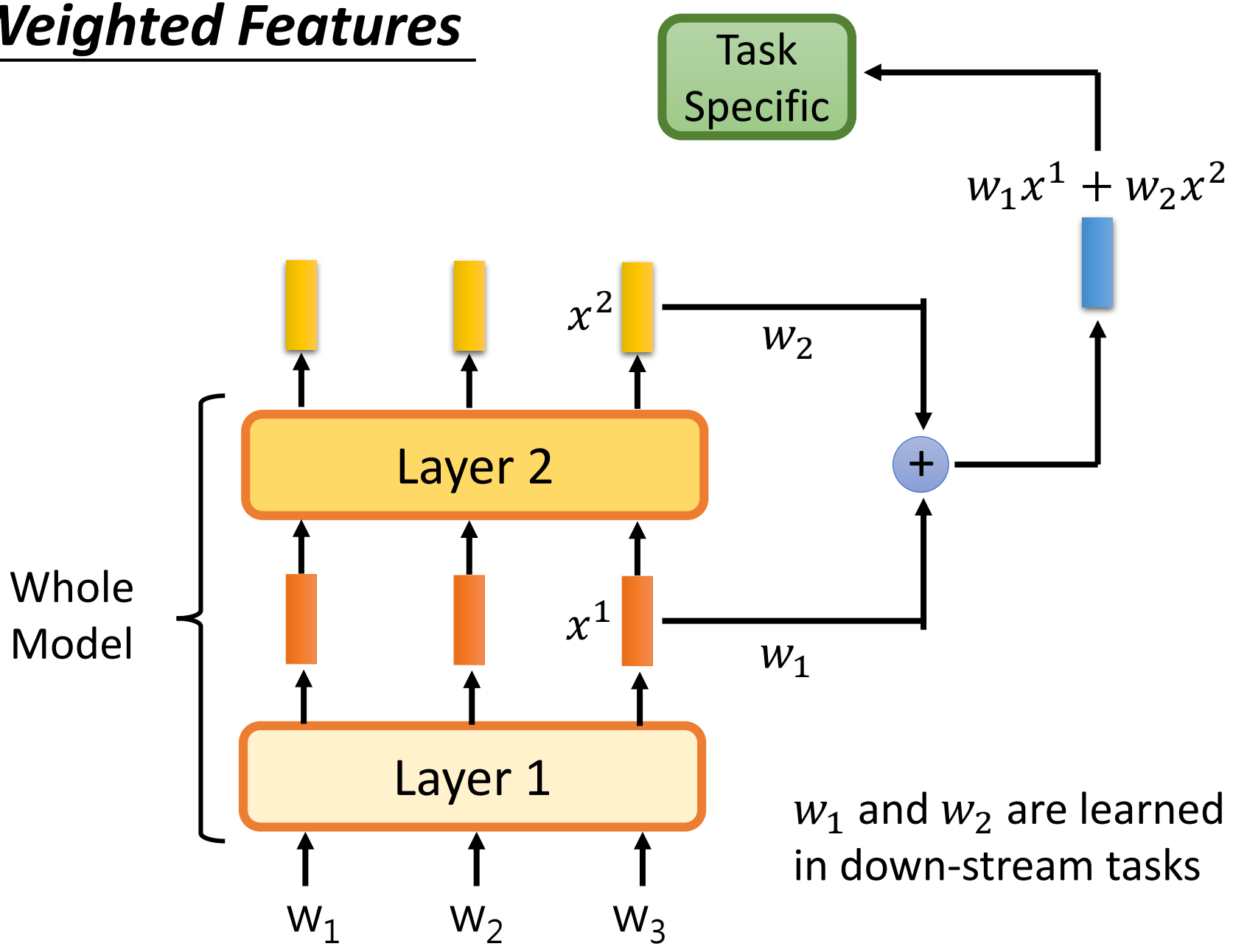
[Houlsby, et al., ICML'19]



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

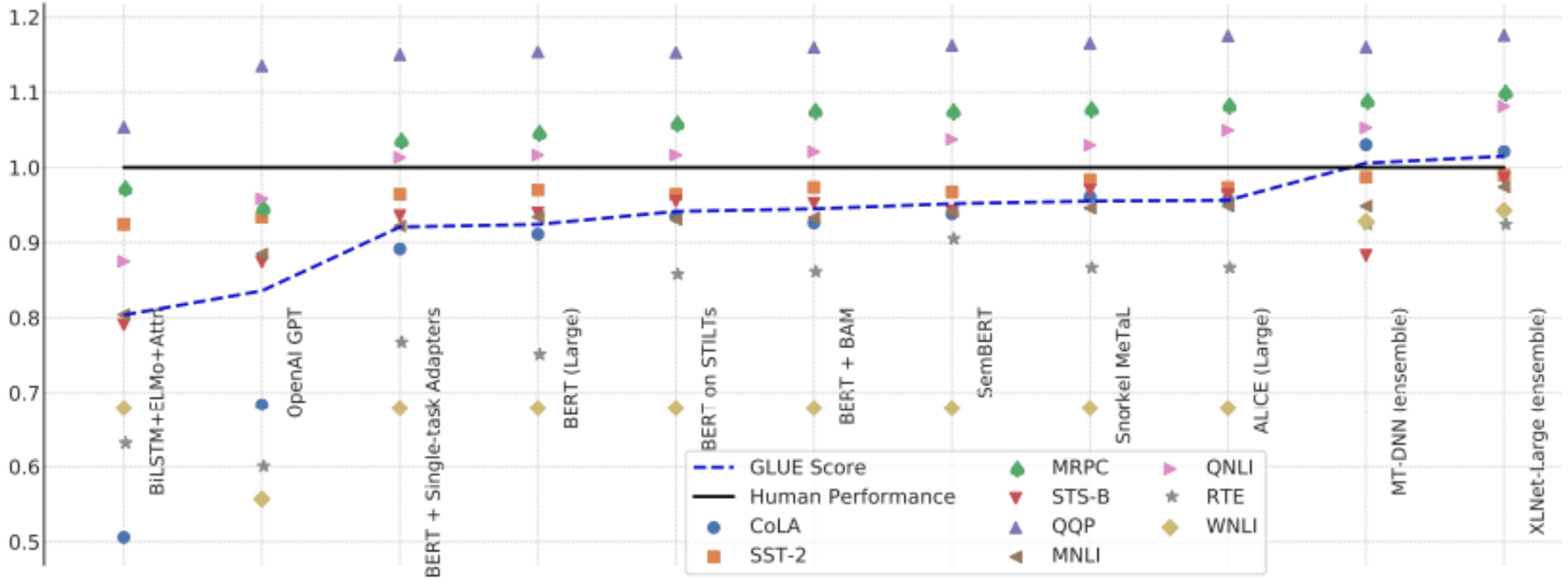
[Houlsby, et al., ICML'19]

Weighted Features



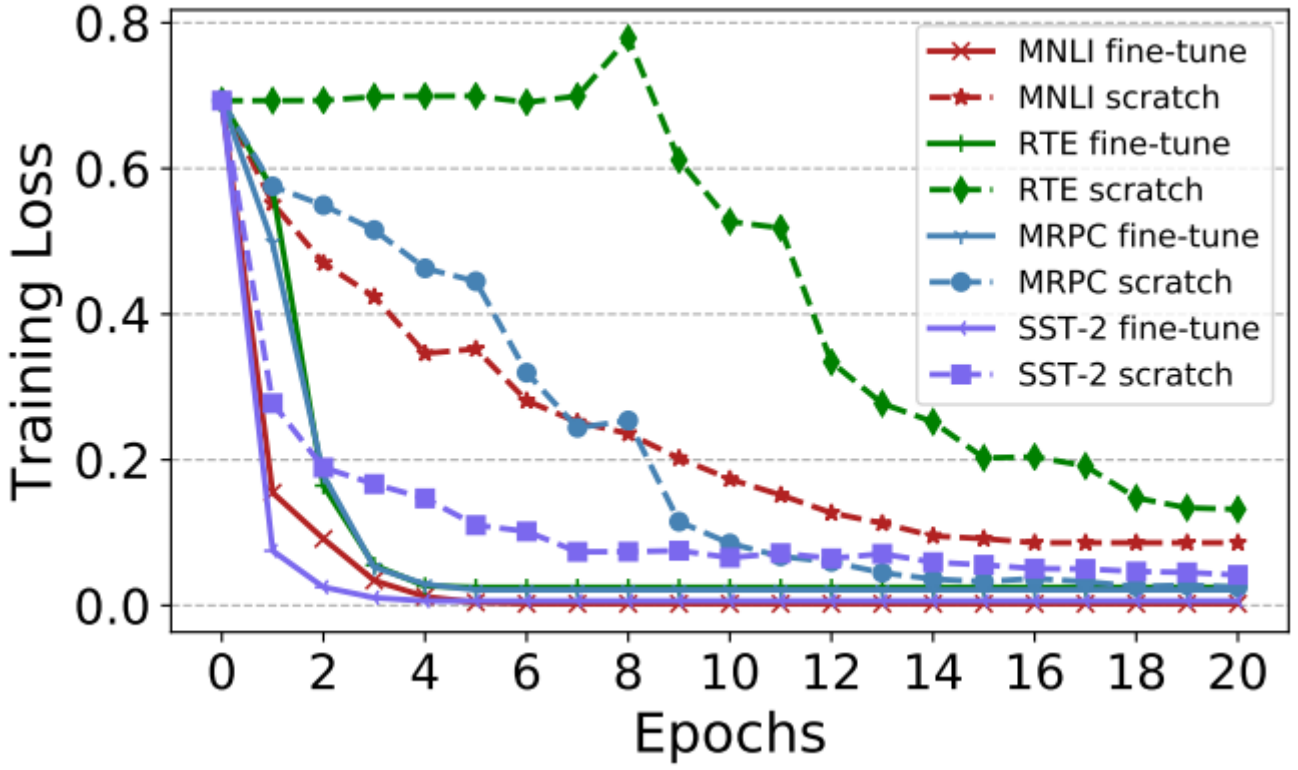
Why Pre-train Models?

- GLUE scores



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

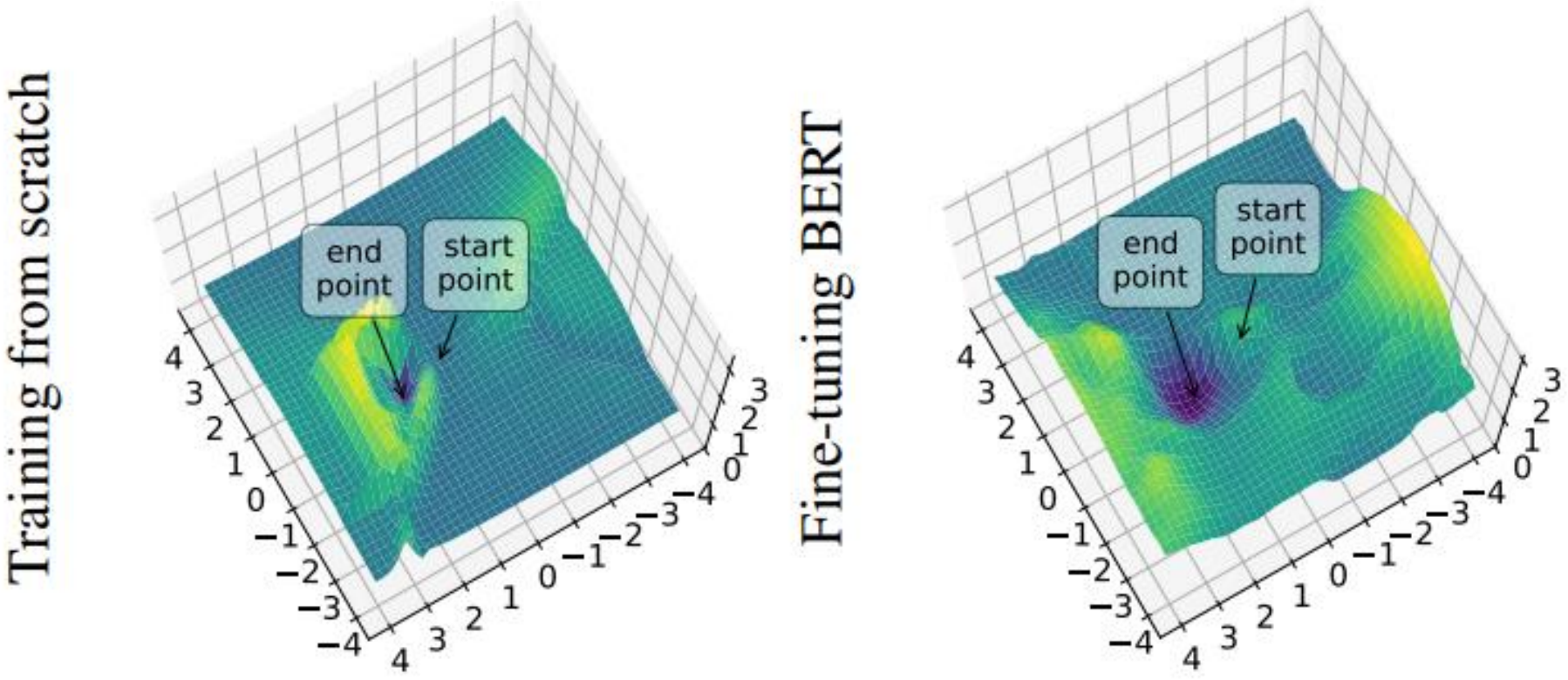
Why Fine-tune?



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

Why Fine-tune?

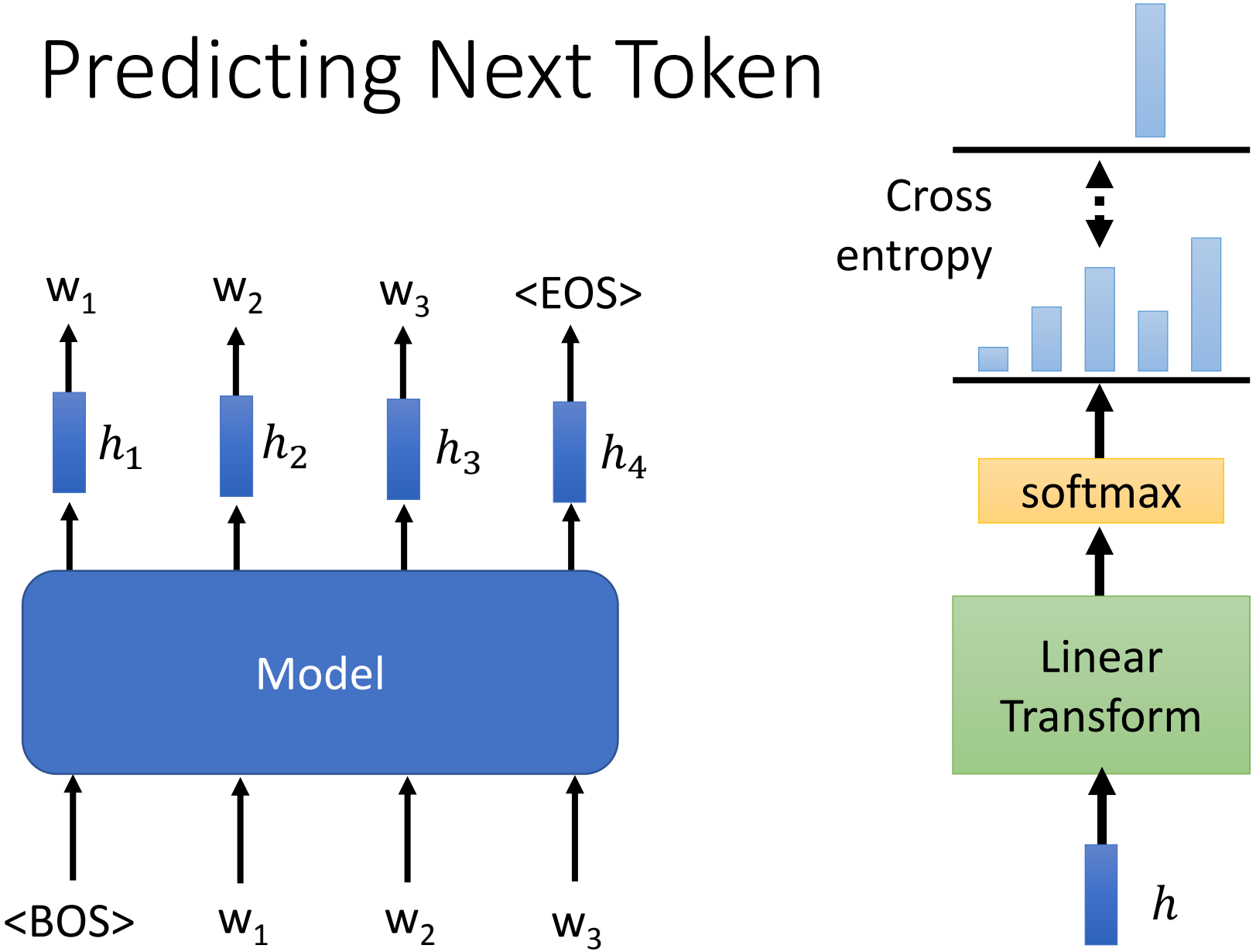
How to generate the figures below?
https://youtu.be/XysGHdNOTbg



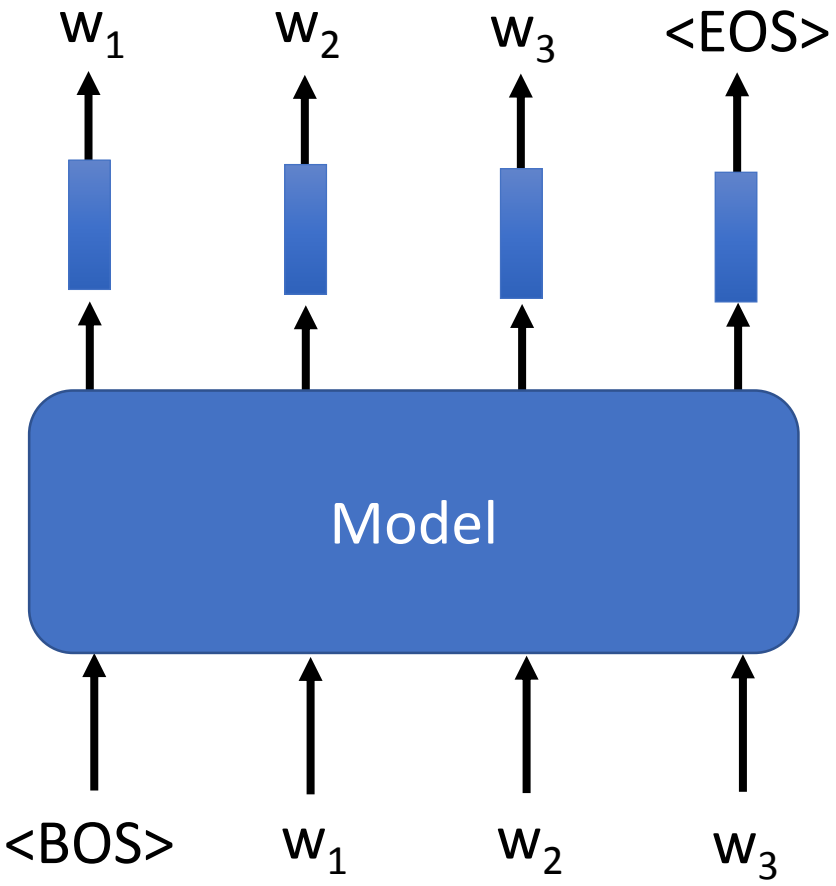
[Hao, et al., EMNLP'19] Source of image: https://arxiv.org/abs/1908.05620

Training Methods

Predicting Next Token



Predicting Next Token



[Peters, et al., NAACL'18]

ELMO

LSTM

GPT

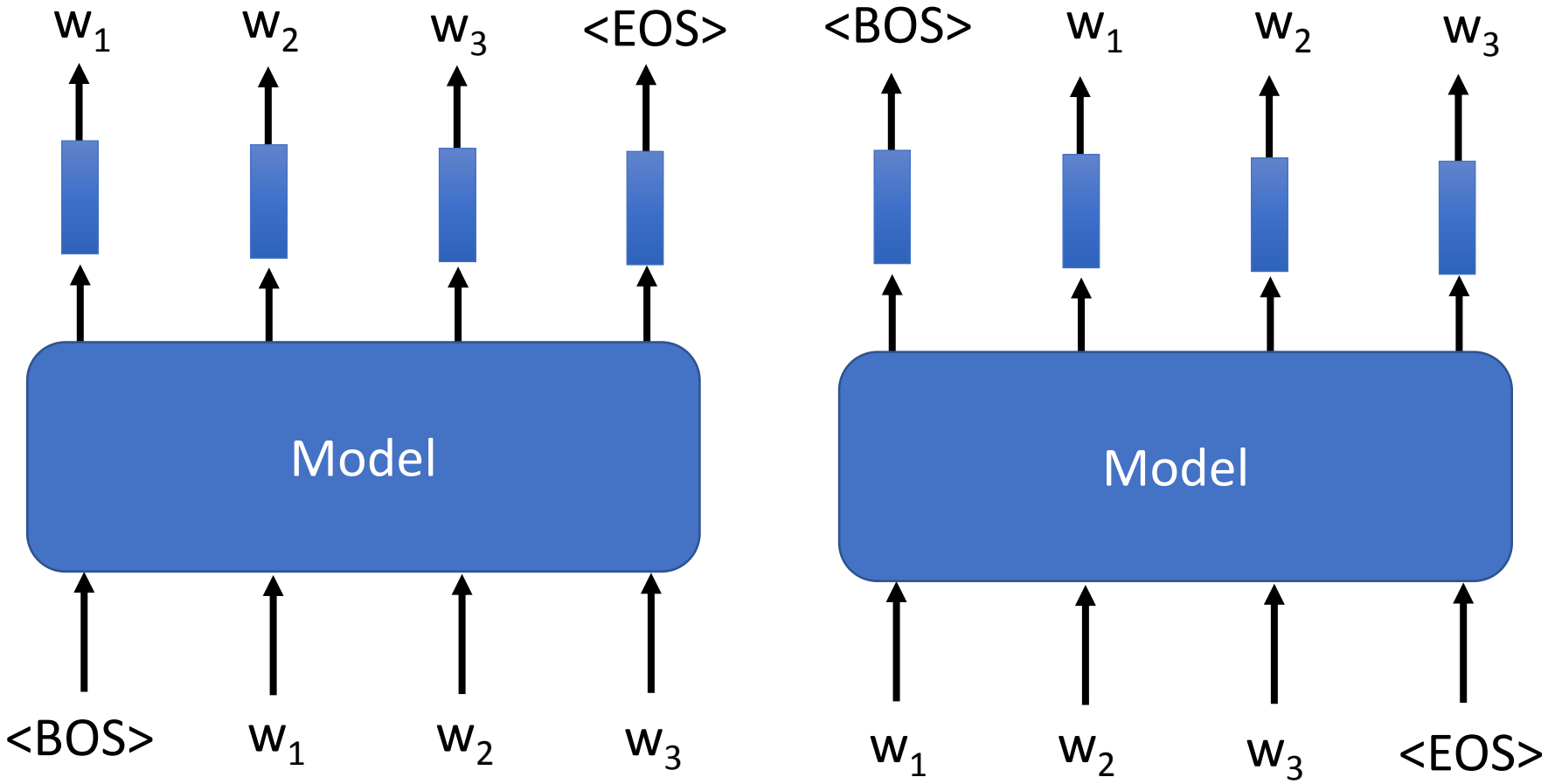
Transformer-Decoder

Don't use Transformer-Encoder

Predicting Next Token

ELMO

LSTM



Predicting Next Token

ELMO

LSTM

Model

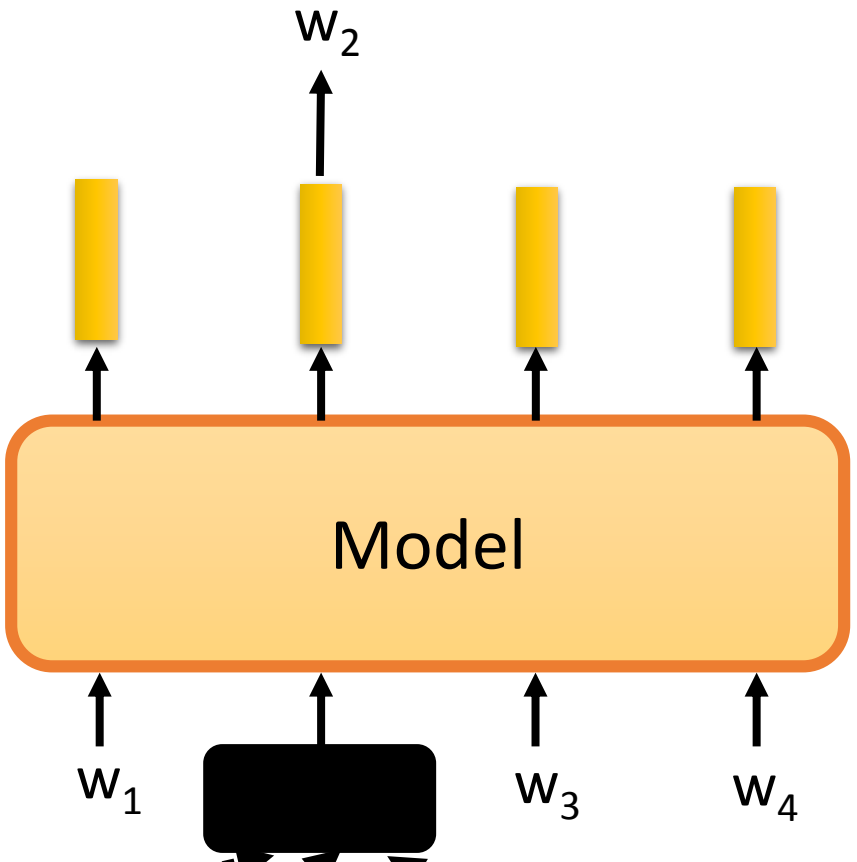
Model

<BOS> w_1 w_2 w_3 w_1 w_2 w_3 <EOS>

Masking Input

BERT

Transformer-Encoder



MASK

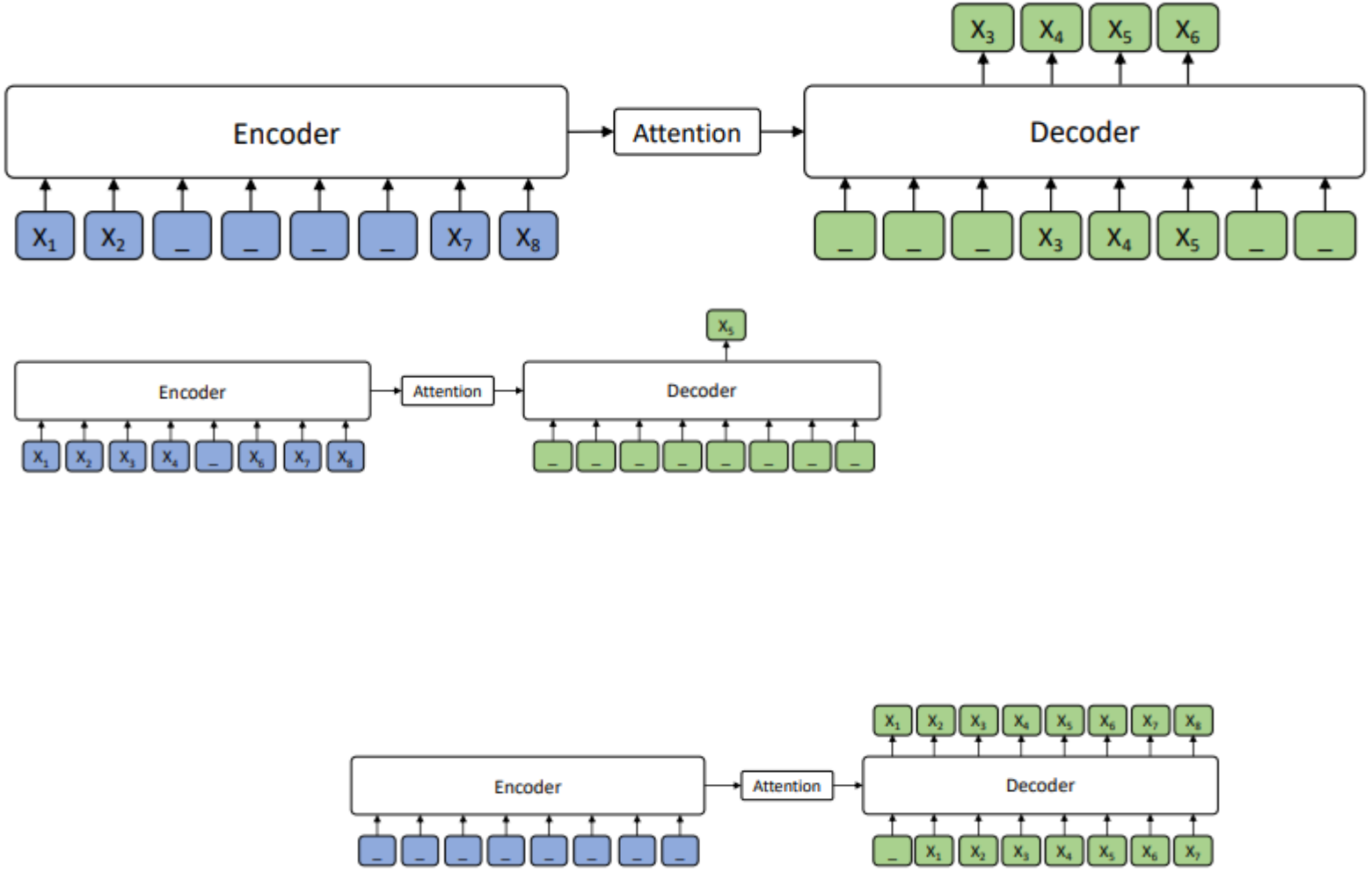
Random Token

Delete (not used in BERT)

MASS

MAsked Sequence to Sequence pre-training (MASS)

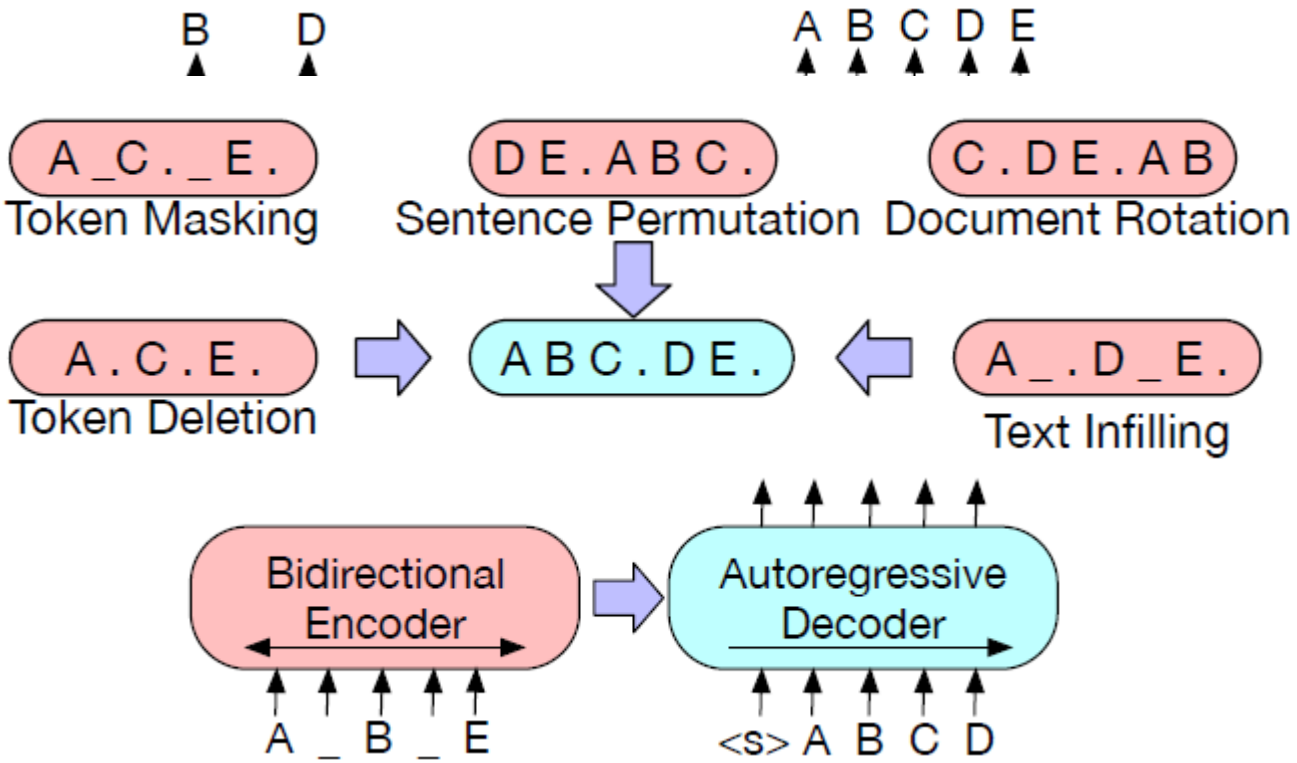
[Song, et al., ICML'19]



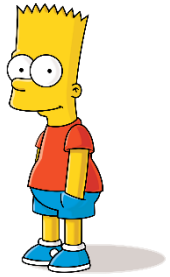
BART

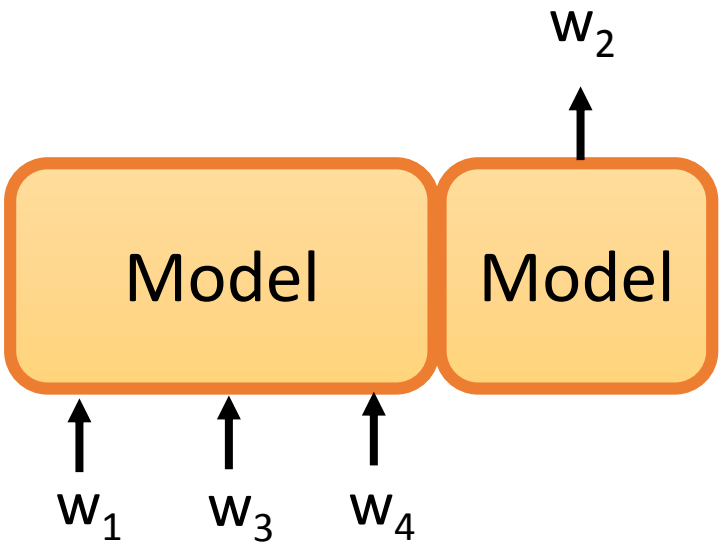
Bidirectional and Auto-Regressive Transformers

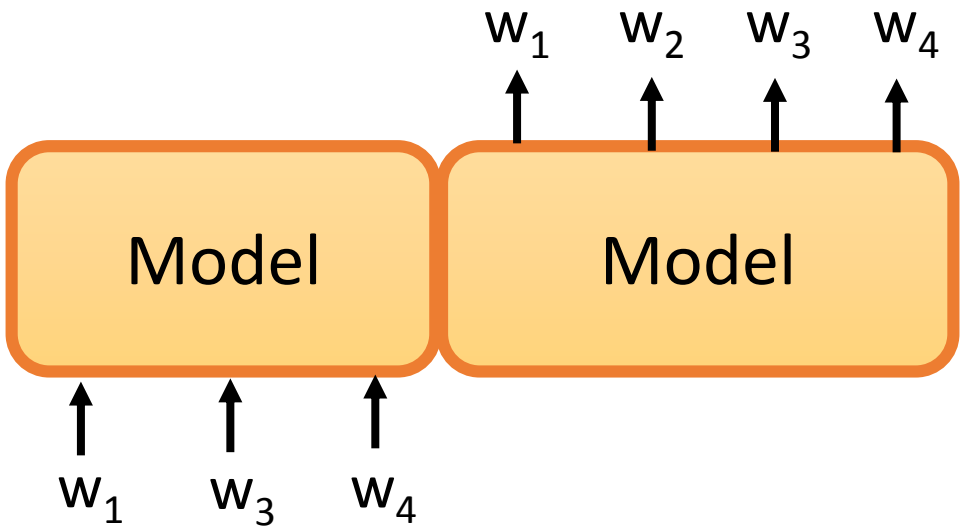
[Lewis, et al., arXiv'19]



Masking Input







w_2 is deleted

Masking Input

Is random masking good enough?

- Whole Word Masking (WWM) https://arxiv.org/abs/1906.08101

[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]** 。

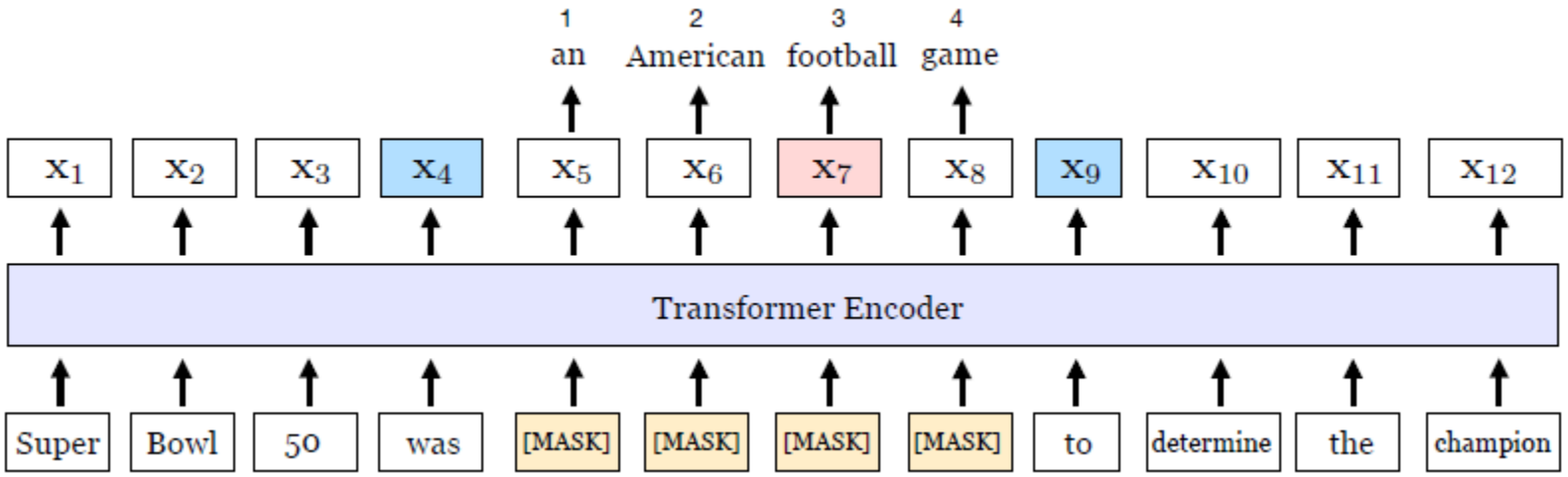
- ERNIE

Enhanced Representation through Knowledge Integration (ERNIE)

SpanBert

NSP is not good

[Joshi, et al., TACL'20]



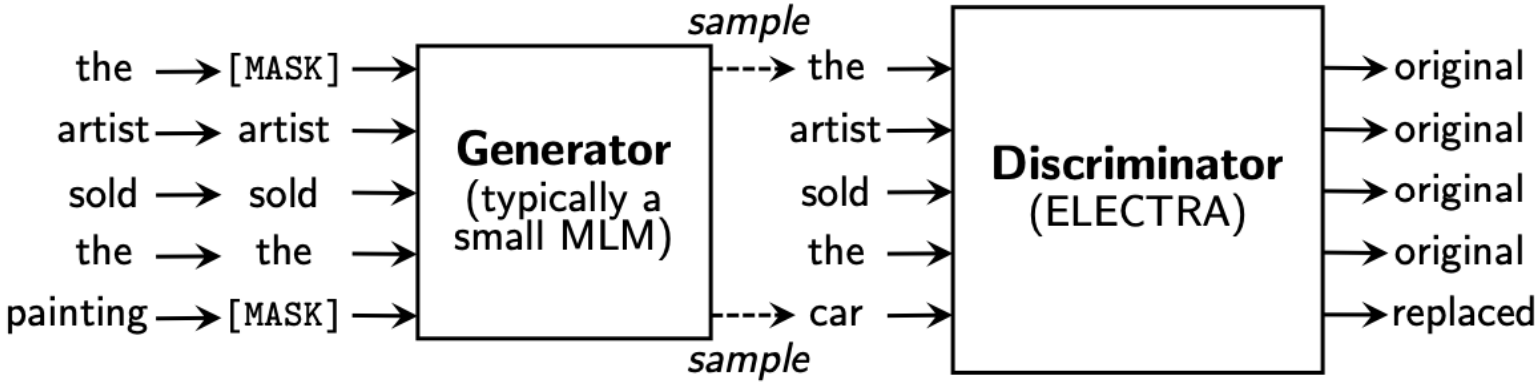
	SQuAD 2.0	NewsQA	TriviaQA	Coreference	MNLI-m	QNLI	GLUE (Avg)
Subword Tokens	83.8	72.0	76.3	**77.7**	86.7	92.5	83.2
Whole Words	84.3	72.8	77.1	76.6	86.3	92.8	82.9
Named Entities	84.8	72.7	78.7	75.6	86.0	93.1	83.2
Noun Phrases	85.0	**73.0**	77.7	76.7	86.5	93.2	**83.5**
Geometric Spans	**85.4**	**73.0**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

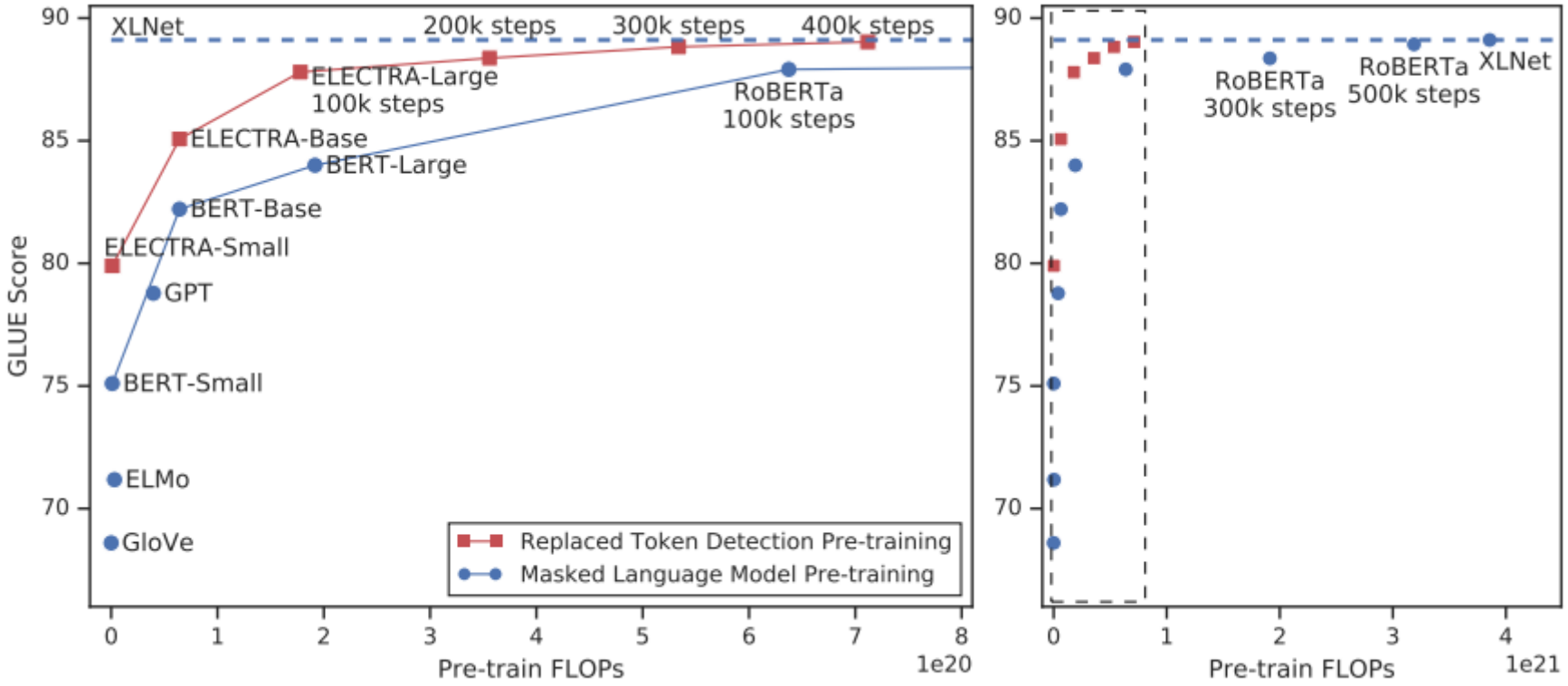
Discriminator

Efficiently Learning an Encoder that Classifies Token Replacements Accurately.”

https://arxiv.org/abs/2003.10555

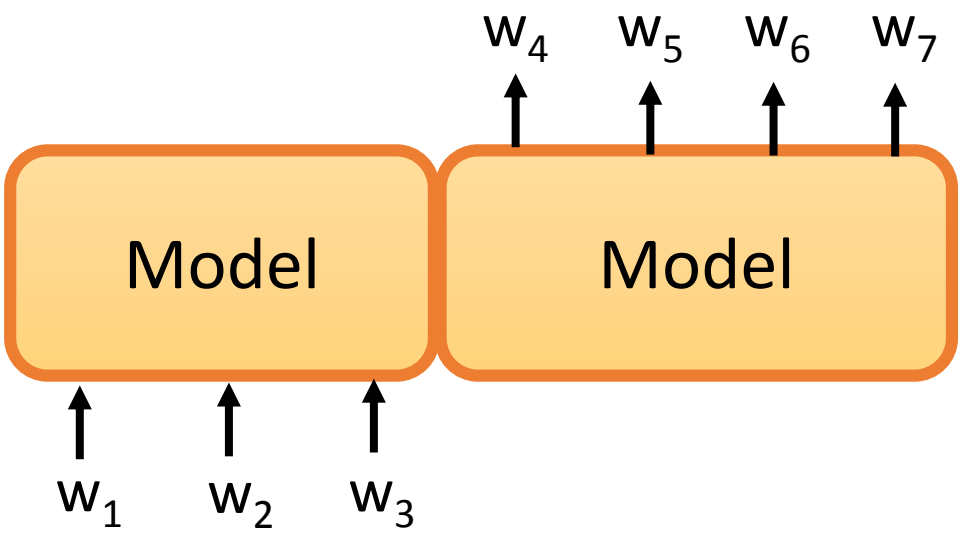
ELECTRA




90
85
80
75
70
GLUE Score
XLNet
200k steps
300k steps
400k steps
ELECTRA-Large
100k steps
ELECTRA-Base
BERT-Large
BERT-Base
ELECTRA-Small
GPT
BERT-Small
ELMo
GloVe
RoBERTa
100k steps
Replaced Token Detection Pre-training
Masked Language Model Pre-training
0
1
2
3
4
5
6
7
8
1e20
Pre-train FLOPs
RoBERTa
300k steps
RoBERTa
500k steps
XLNet
1e21

Sentence Level

- Input one sentence, predict another



COVE:

Input: A language

output: B language

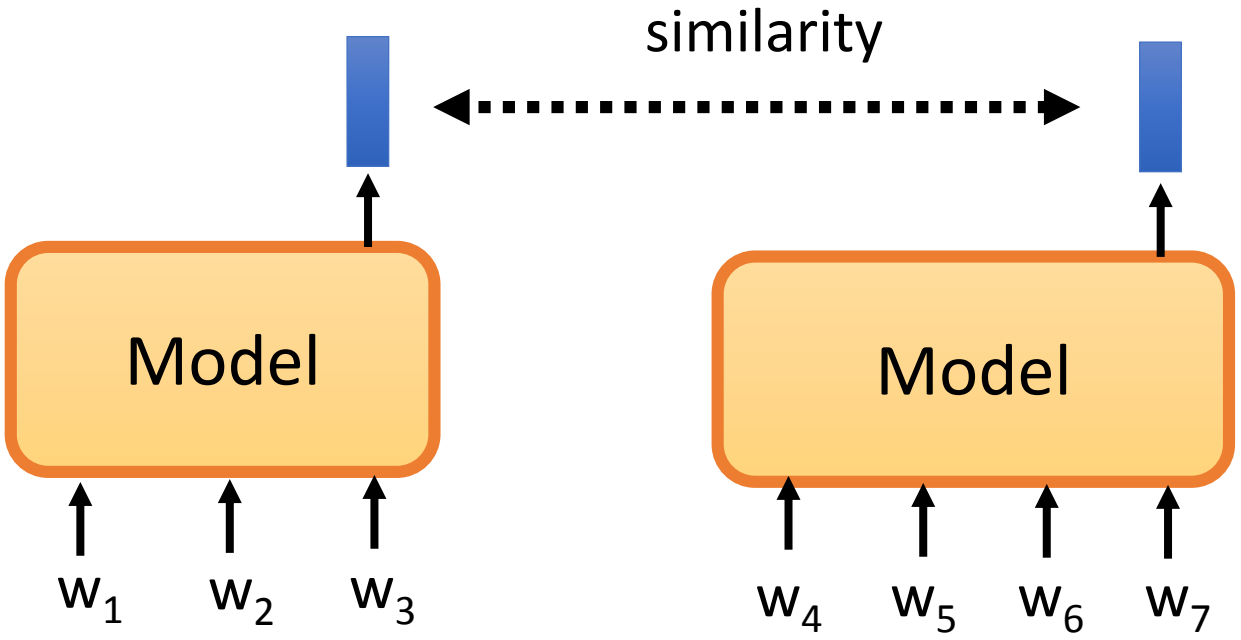
Skip-Thought

Input: previous sentence

output: next sentence

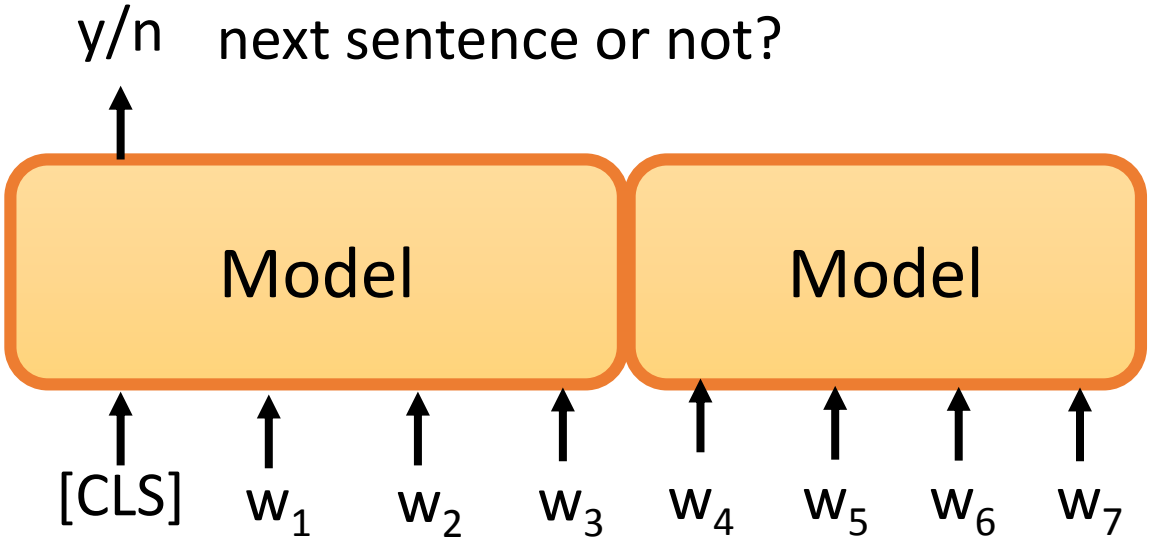
Sentence Level

- Quick Though



SOP

Sentence Level



Robustly optimized BERT approach (ROBERTA)

[Liu, et al., arXiv'19]

Concluding Remarks

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