Recent Advances in Pre-trained Language Models

姜成翰

Cheng-Han Chiang

04.01.2022

防疫訊息

更新日期：111年3月30日

本校防疫訊息–依據CDC 3月28日公告提醒本校教職員工生落實各項防疫措施

全校教職員工生大家好：

依據中央流行疫情指揮中心 3月28日發布的公告，請全校教職員工生在校內活動時，除用餐、室內外運動、以及指揮中心規定的例外情形之外，其他活動均應全程配戴口罩。講課、以及室內外拍攝個人/團體照，雖屬於指揮中心規定得免戴口罩的活動，但本校建議如無法與他人保持社交距離時，仍應該配戴口罩。其他防疫措施提醒事項如下：

大家好：

疫情升溫，保健中心也接獲同學反應：有許多學生在上課時並未配戴口罩，可能造成防疫漏洞。
尤其曾發生配戴口罩的同學被迫與不戴口罩的同學分在同一組討論，令配戴口罩的同學感到很大的壓力。

保健中心在此籲請本校各單位再度對內加強宣導：
學生上課時應一律配戴口罩，也請各單位提醒教師上課時應提醒學生全程配戴口罩，
尊重自己也尊重他人，不應該讓遵守規定配戴口罩的同學承受他人不配戴口罩的防疫壓力，謝謝大家～

https://www.ntu.edu.tw/spotlight/2022/2045_20220330.html

Prerequisites

- 【機器學習2021】自注意力機制 (Self-attention) (上)
- 【機器學習2021】自注意力機制 (Self-attention) (下)

Highly Related Topics

- 【機器學習2021】BERT簡介
- 【機器學習2021】GPT的野望

Highly Related Topics

- 【深度學習於人類語言處理 2020】BERT and its family

Outline

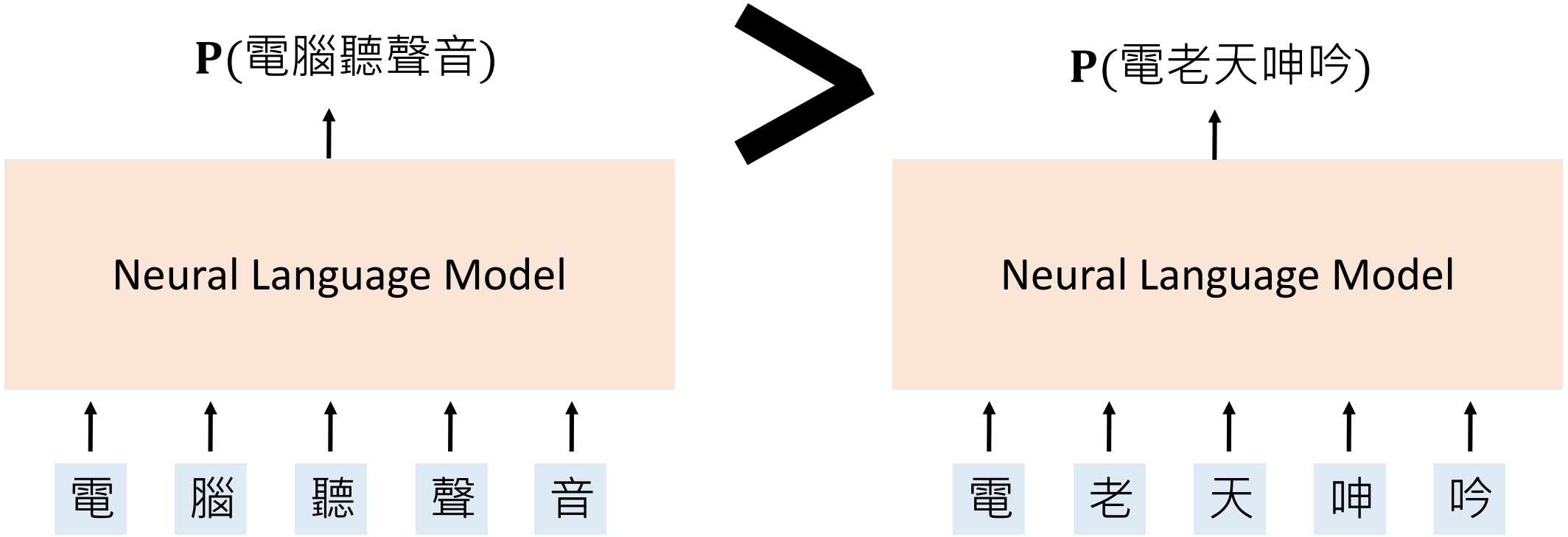
- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
- Closing Remarks

Outline

- Background knowledge
  - Pre-trained Language Models
- The Problems of PLMs
- The Solutions of Those Problems
- Closing Remarks

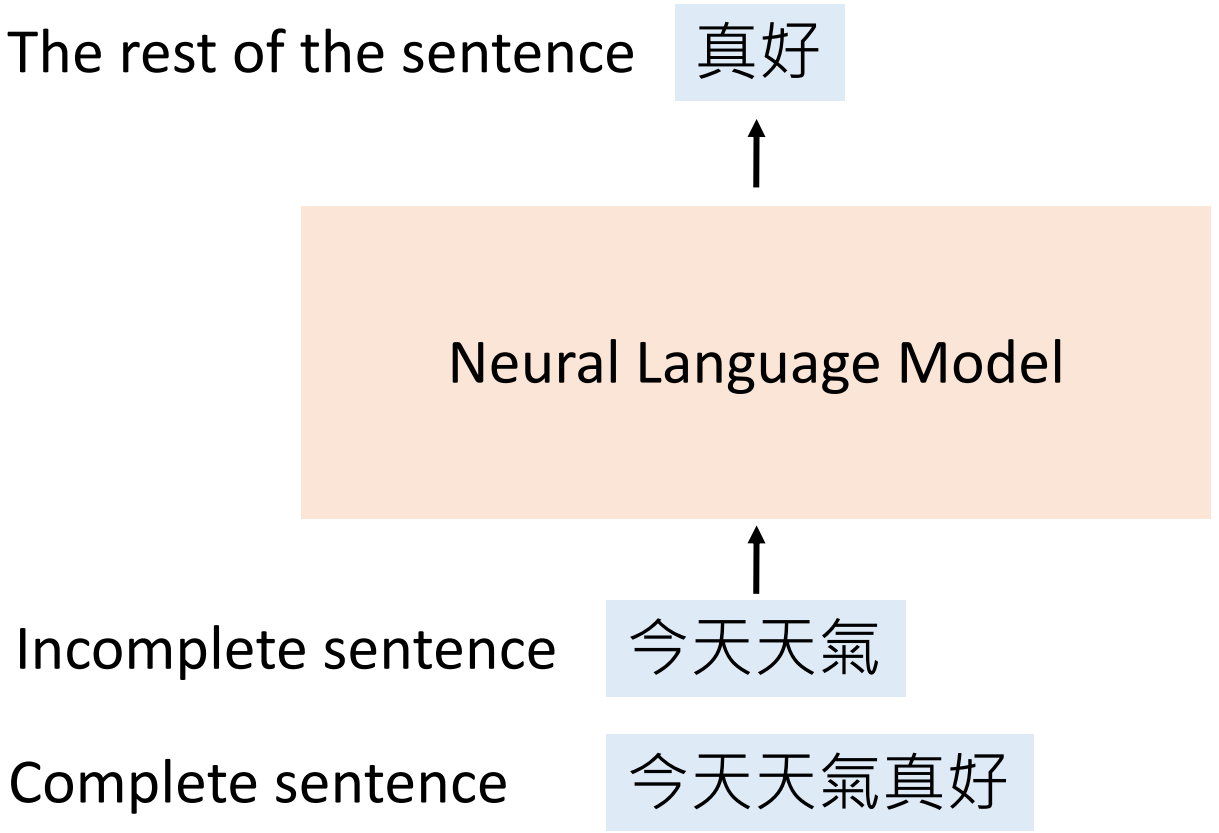
Background: Pre-trained Language Models

- Neural Language Models: A neural network that defines the probability over sequences of words



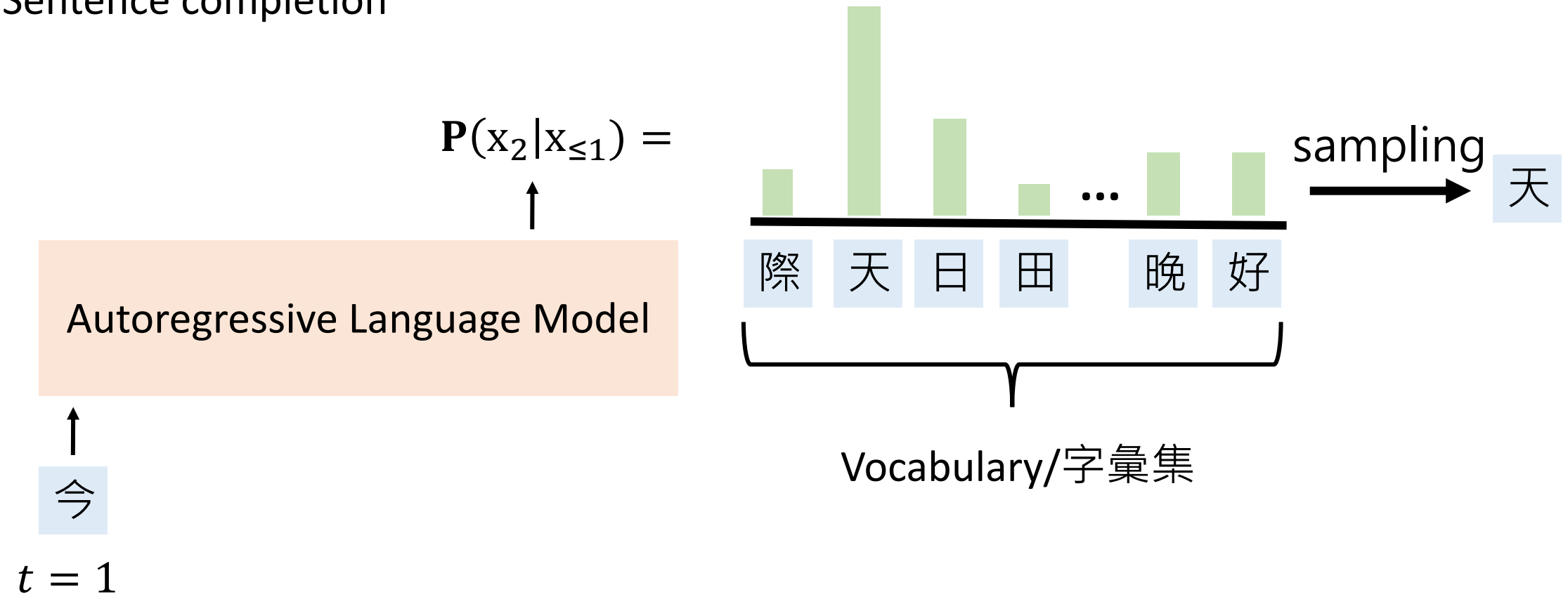
Background: Pre-trained Language Models

- How are these language models trained?
  - Given an incomplete sentence, predict the rest of the sentence



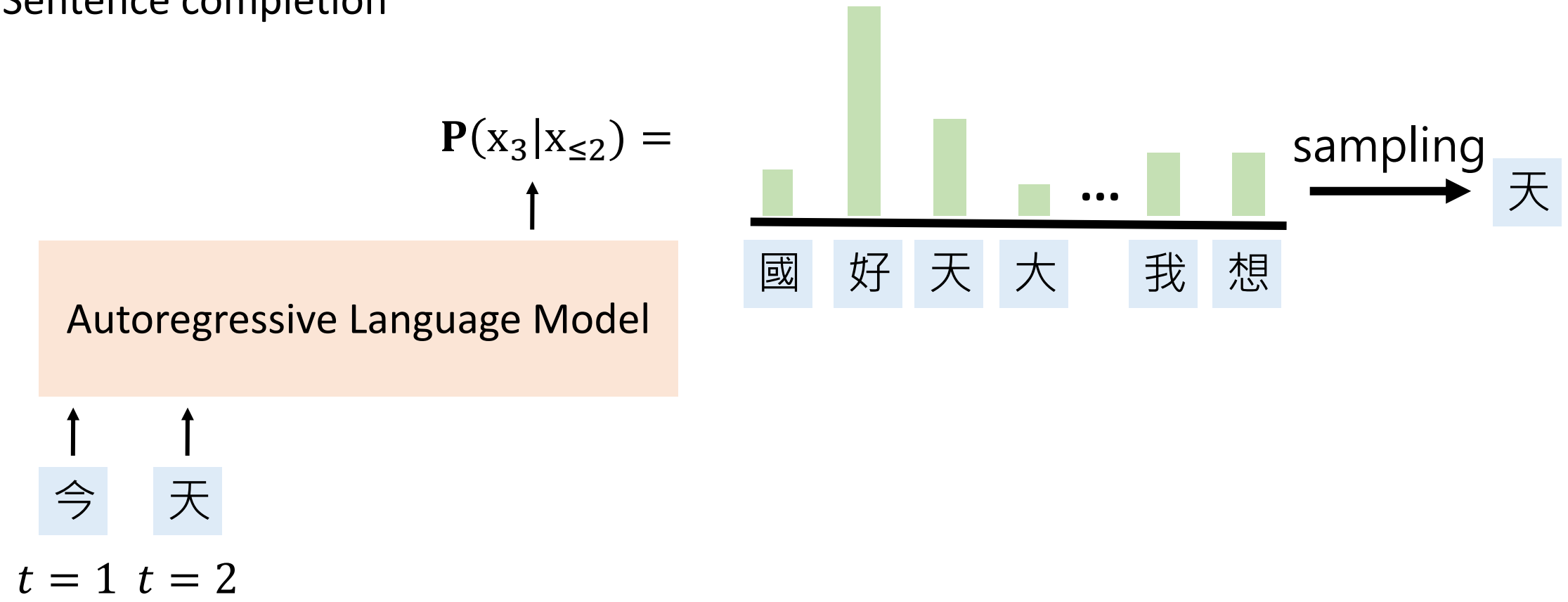
Background: Pre-trained Language Models

- Autoregressive Language Models (ALMs): Complete the sentence given its prefix
  - Sentence completion



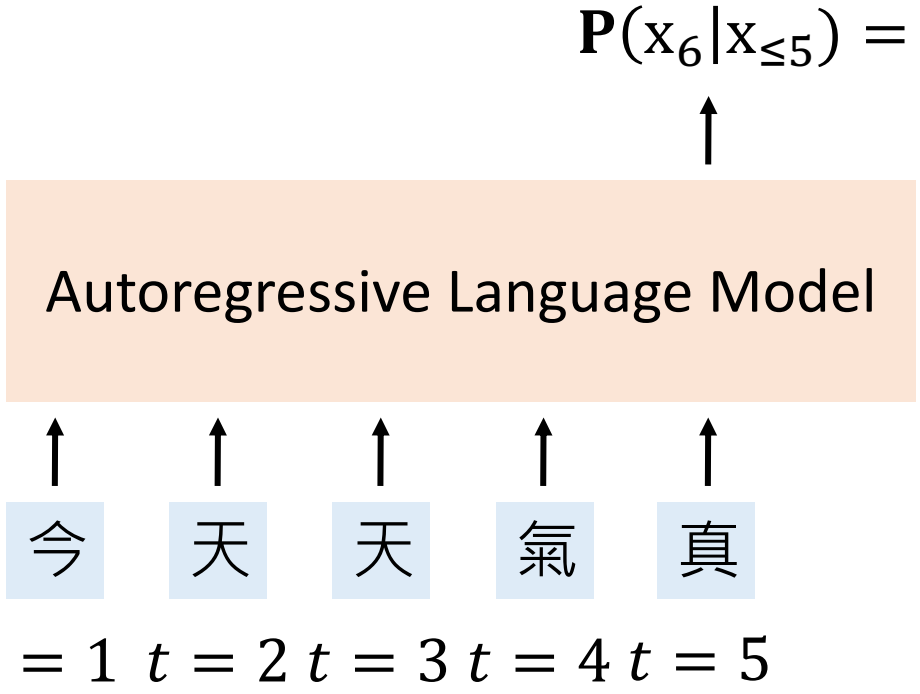
Background: Pre-trained Language Models

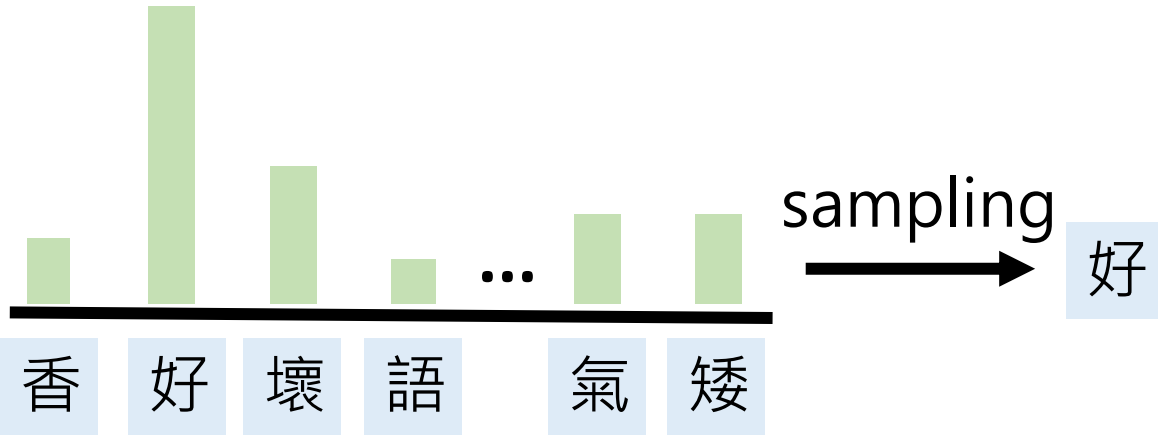
- Autoregressive Language Models (ALMs): Complete the sentence given its prefix
  - Sentence completion



Background: Pre-trained Language Models

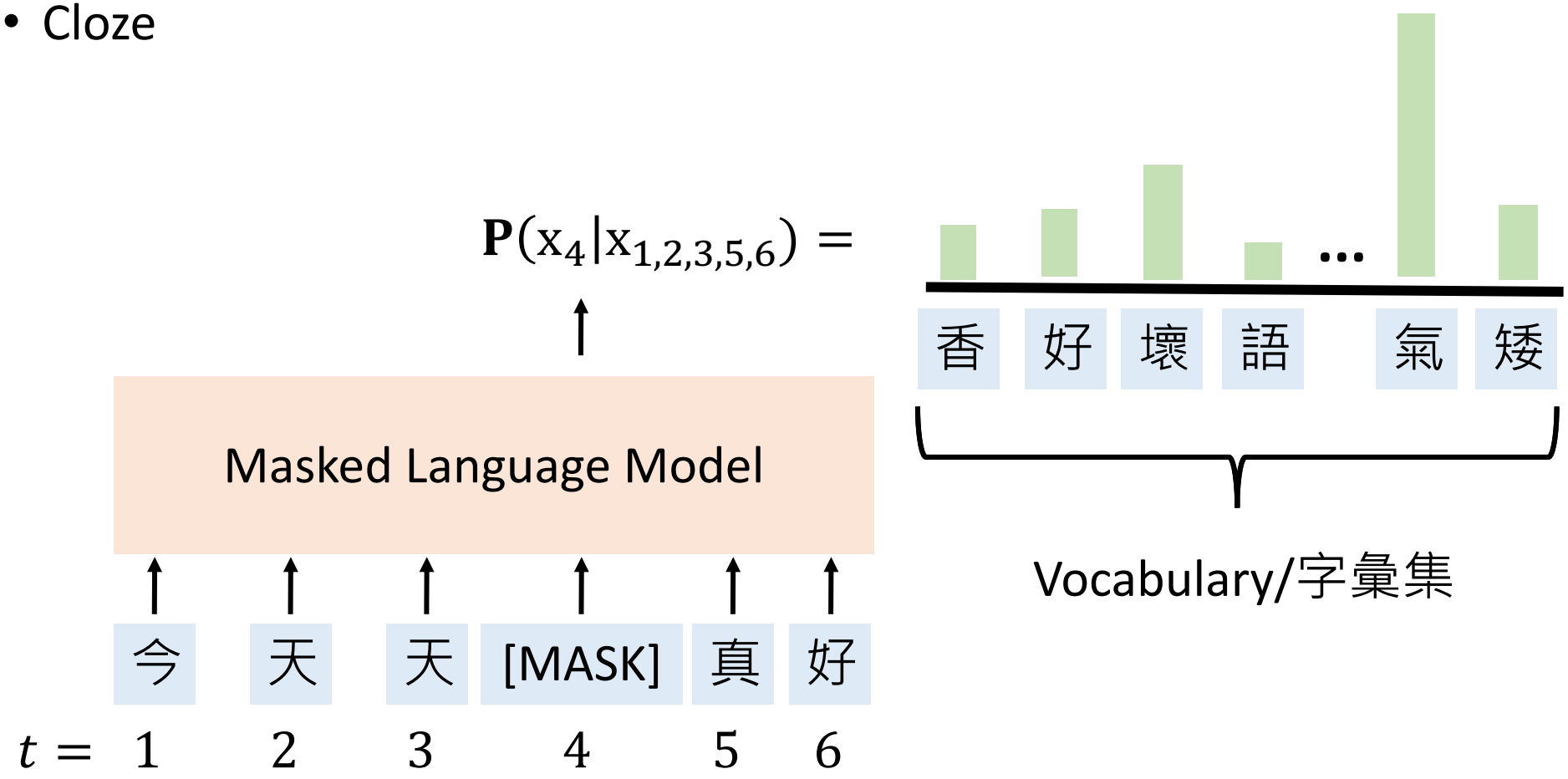
- Autoregressive Language Models (ALMs): Complete the sentence given its prefix
  - Sentence completion





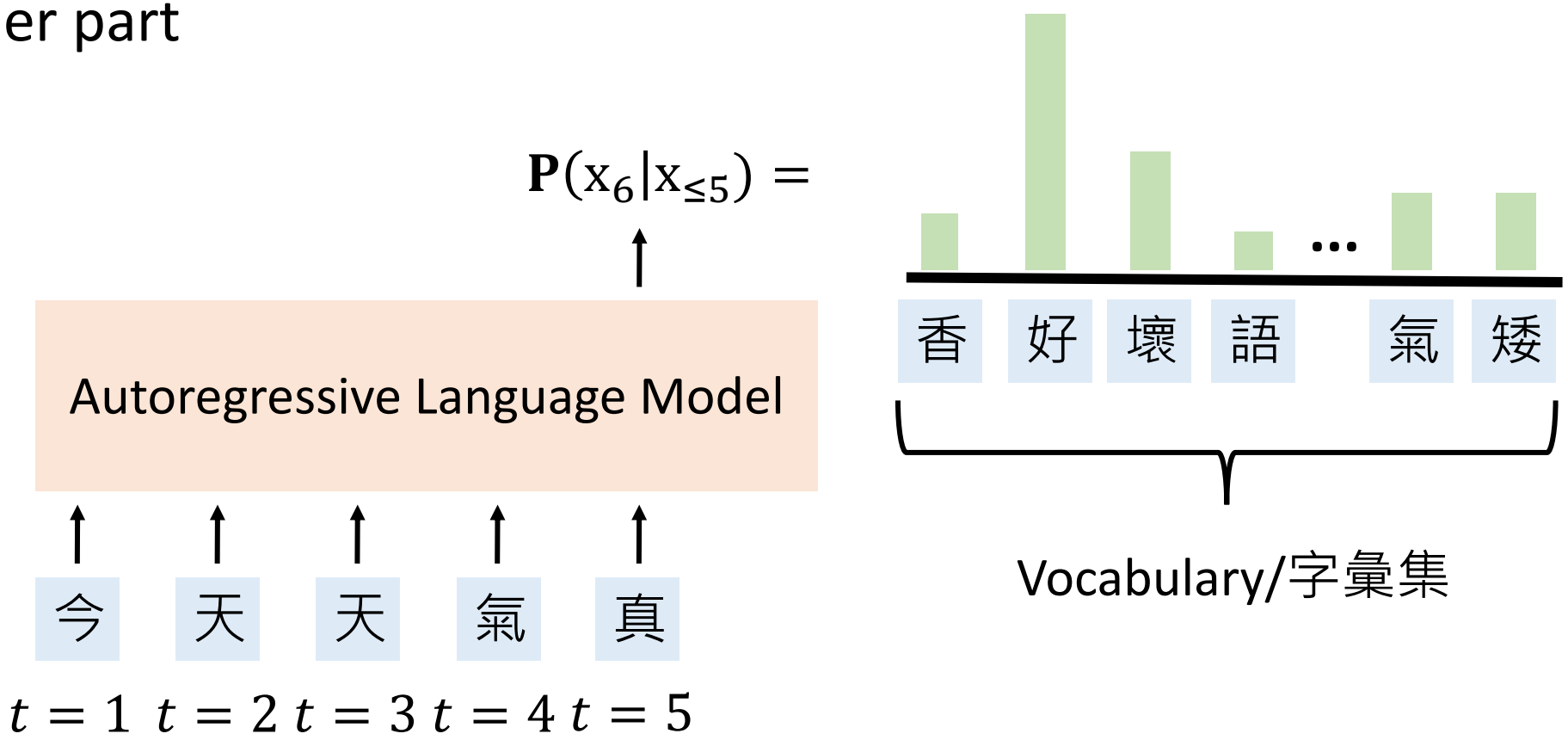
Background: Pre-trained Language Models

- Masked Language Models (MLMs): Use the unmasked words to predict the masked word
  - Cloze



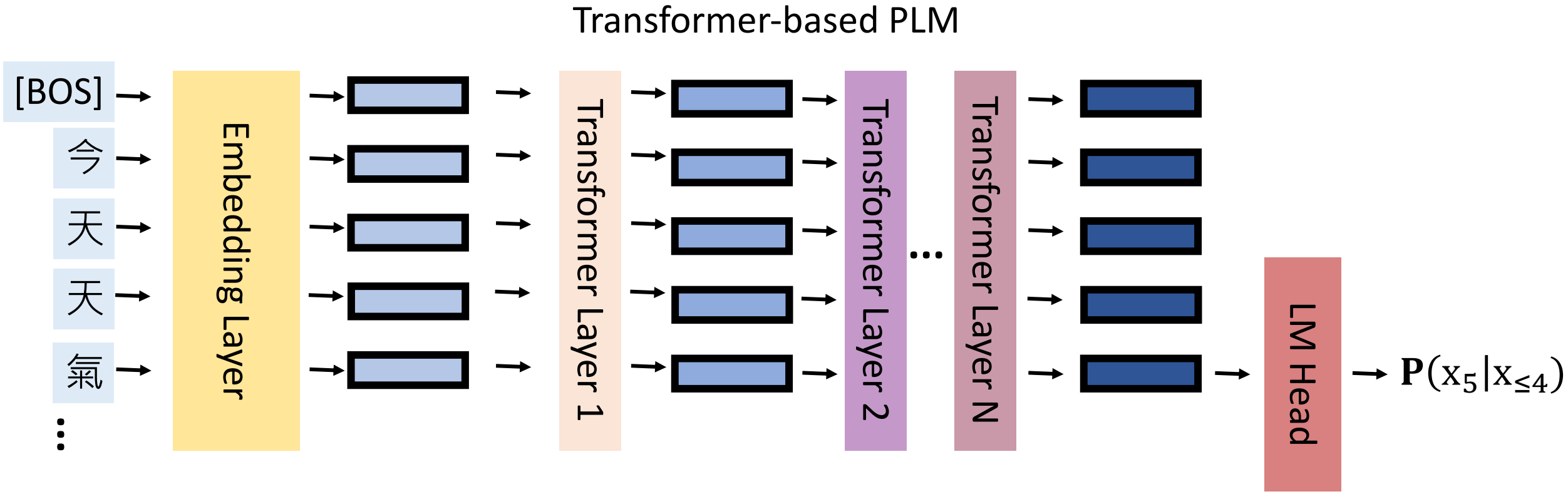
Background: Pre-trained Language Models

- Training a langauge model is self-supervised learning
- Self-supervised learning :Predicting any part of the input from any other part



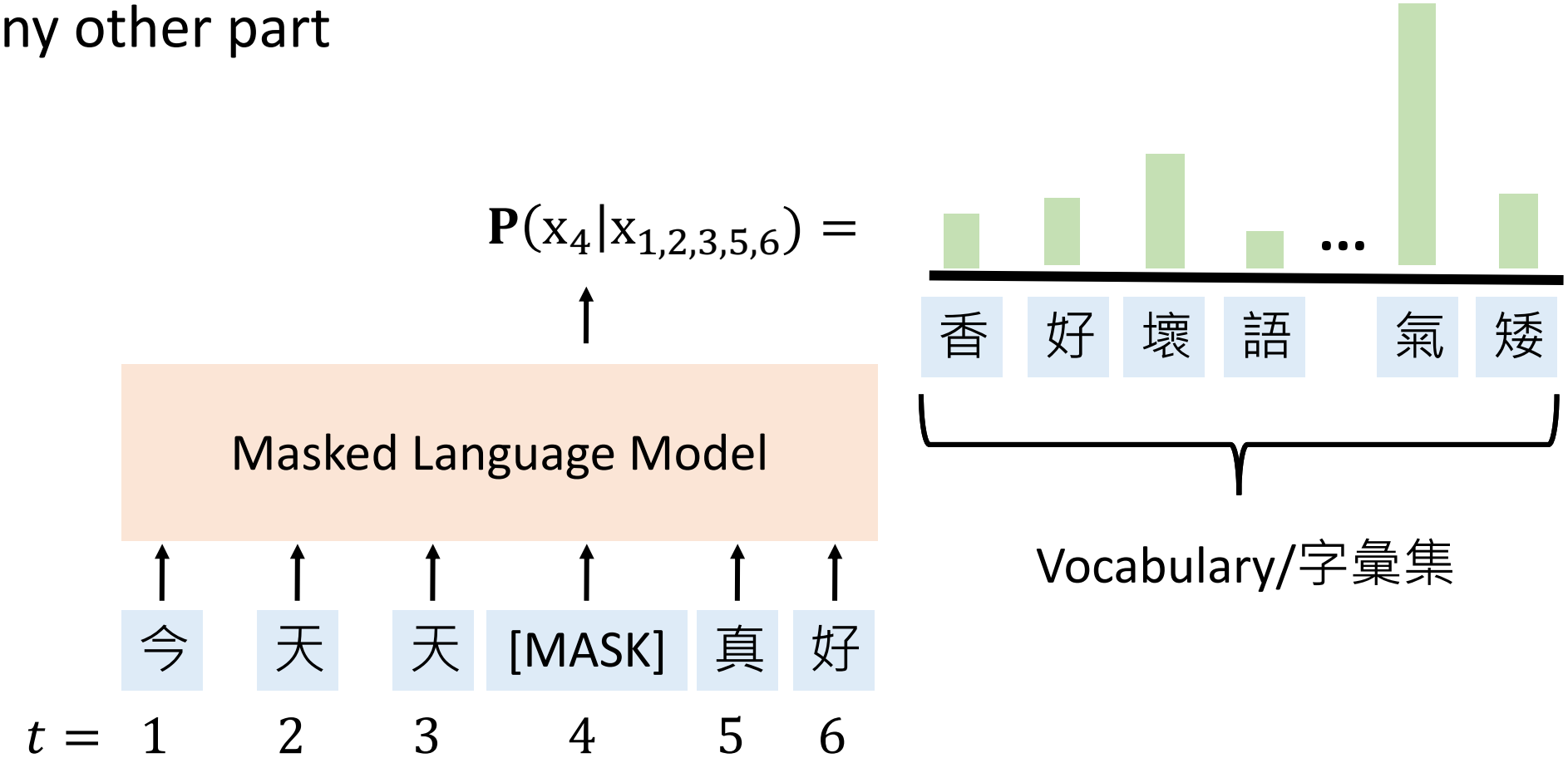
Background: Pre-trained Language Models

- Transformer-based ALMs: Composed of stacked layers of transformer layers



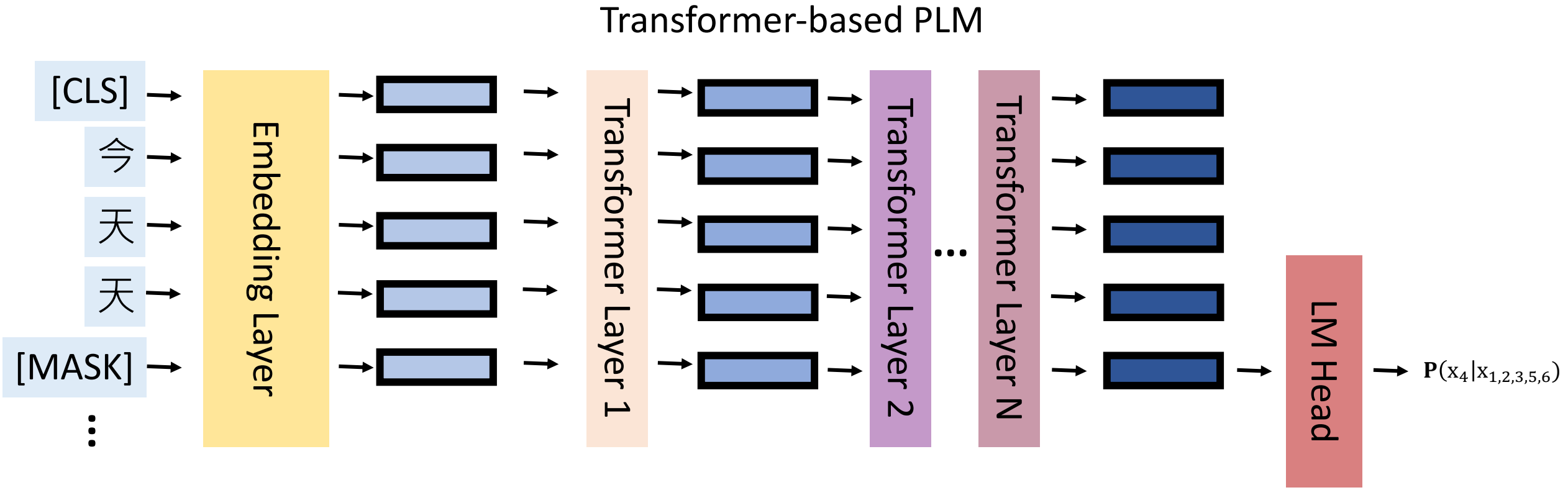
Background: Pre-trained Language Models

- Training a langauge model is self-supervised learning
- Self-supervised learning :Predicting any part of the input from any other part



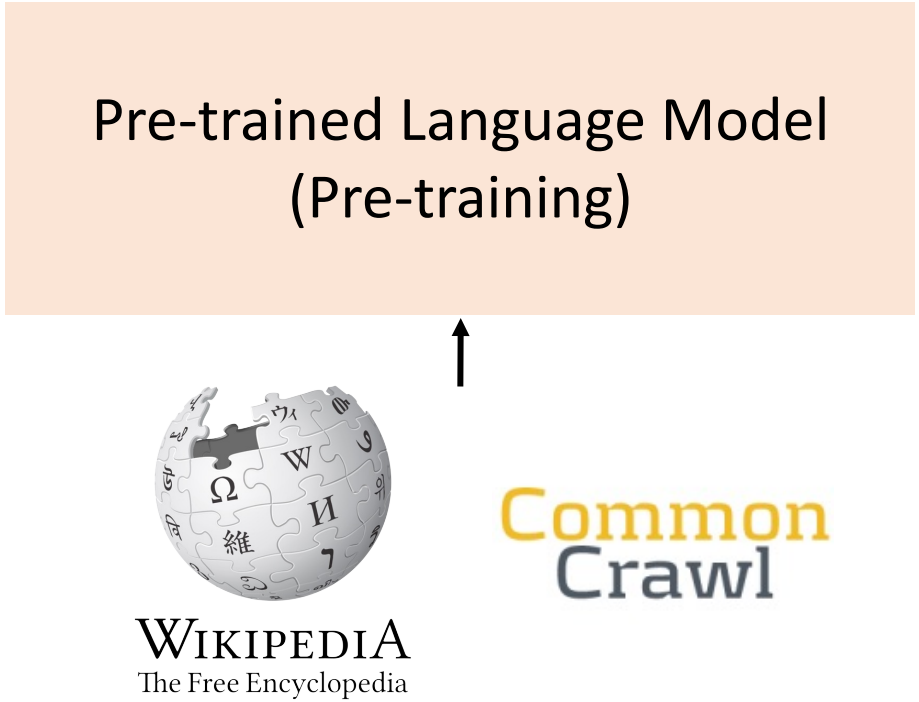
Background: Pre-trained Language Models

- Transformer-based PLMs: Composed of stacked layers of transformer layers



Background: Pre-trained Language Models

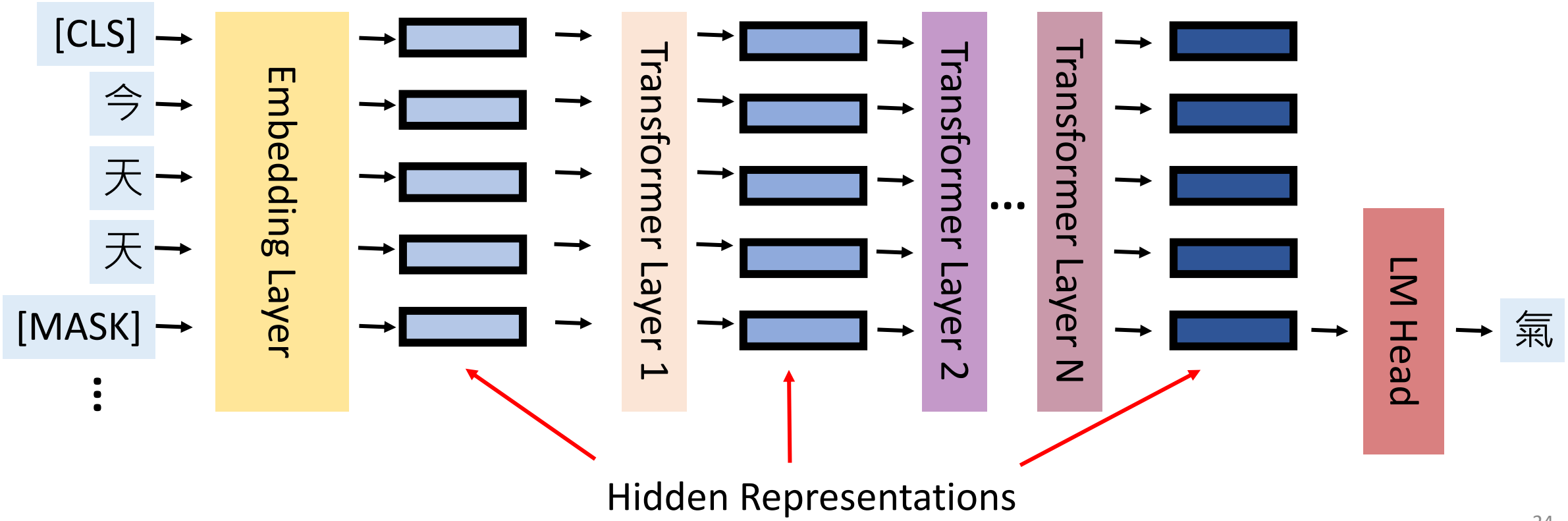
- Pre-trained Language Models (PLMs)/ 預訓練語言模型
  - **Pre**-training/**預**訓練: Using a large corpora to train a neural language model
    - Autoregressive pre-trained: GPT 系列 (GPT, GPT-2, GPT-3)
    - MLM-based pre-trained: BERT 系列 (BERT, RoBERTa, ALBERT)



https://zh.wikipedia.org/wiki/%E8%8B%B1%E8%AF%AD%E7%BB%B4%E5%9F%BA%E7%99%BE%E7%A7%91

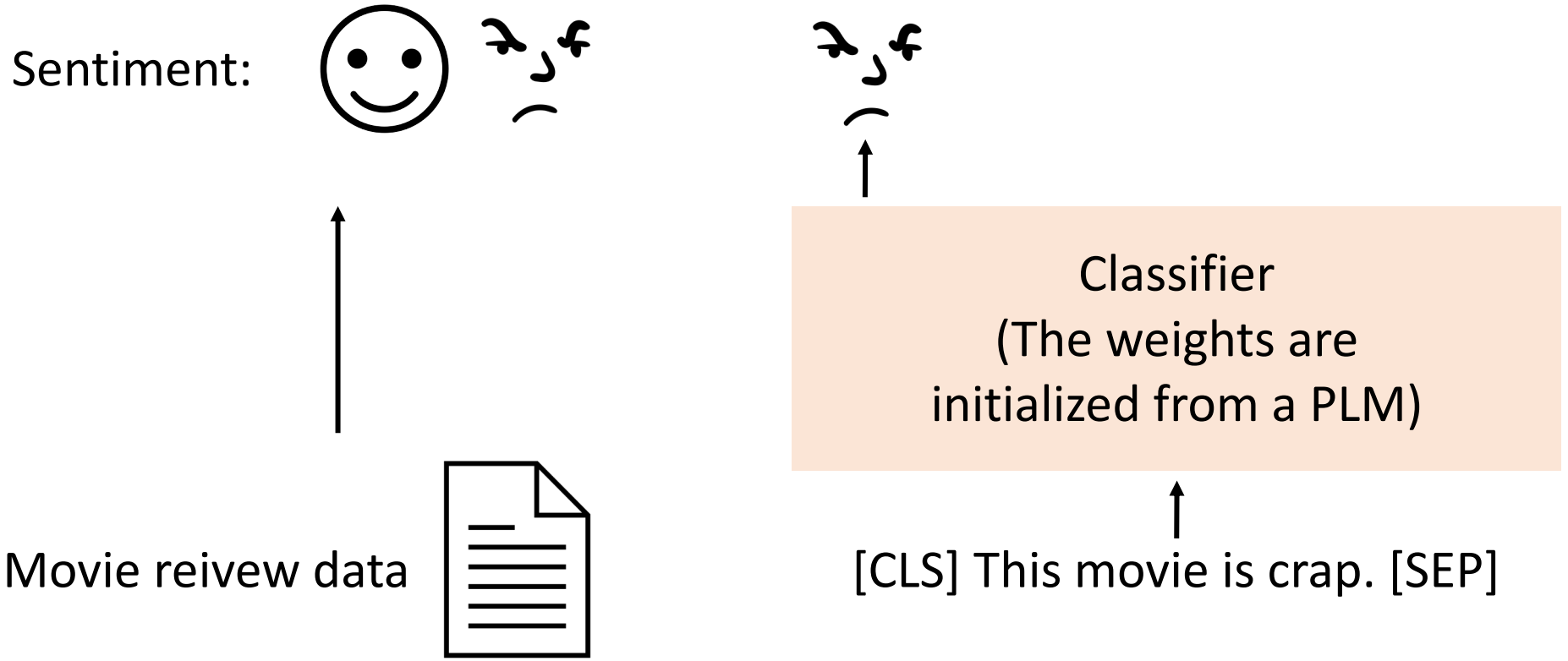
Background: Pre-trained Language Models

- We believe that after pre-training, the PLM learns some knowledge, encoded in its hidden representations, that can transfer to downstream tasks



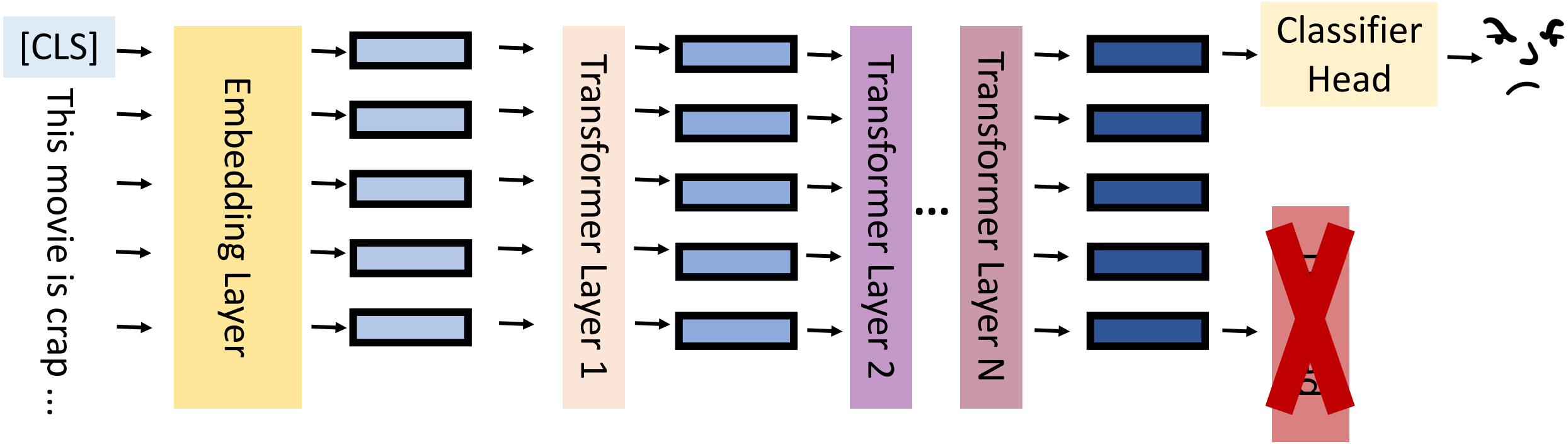
Background: Pre-trained Language Models

- Pre-trained Language Models (PLMs)/ 預訓練語言模型
  - (Standard) fine-tuning/微調: Using the pre-trained weights of the PLM to initialize a model for a **downstream task**



Background: Pre-trained Language Models

- Pre-trained Language Models (PLMs)/ 預訓練語言模型
  - (Standard) fine-tuning/微調: Using the pre-trained weights of the PLM to initialize a model for a **downstream task**



Background: Pre-trained Language Models

- Pre-trained Language Models (PLMs)/ 預訓練語言模型
  - Fine-tuning PLMs on downstream tasks achieves exceptional performance on many kinds of downstream tasks

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Background:Pre-trained Language Models

- Pre-trained Language Models (PLMs)/ 預訓練語言模型
  - PLMs are widely applied to many different scenarios in different realms

KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs

A Simple

with

Cross-Lingual Lemmatization and Morphology Tagging with Two-Stage Multilingual BERT Fine-Tuning

Zulfat Miftahu

Kazar

Transfe

Yasuh

TMU Transformer System Using BERT for Re-ranking at BEA 2019 Grammatical Error Correction on Restricted Track

Incorporating medical knowledge in BERT for clinical relation extraction

machi

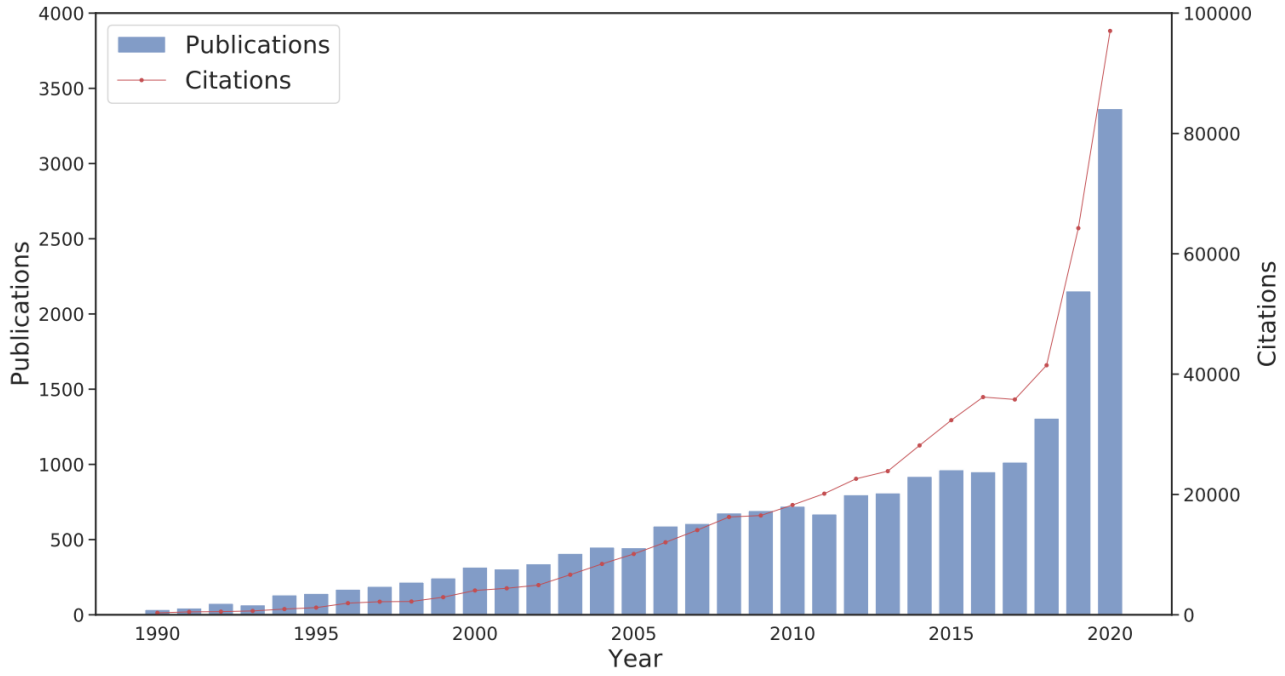
S

Arpita Roy and **Shimei Pan**
Department of Information Systems
University of Maryland, Baltimore County
Maryland, USA
{arpita2, shimei}@umbc.edu

https://aclanthology.org/events/acl-2019/#p19-1

Background:Pre-trained Language Models

- Pre-trained Language Models (PLMs)/ 預訓練語言模型
  - PLMs are every where



(a) The number of publications on "language models" and thei citations in recent years.

Han, Xu, et al. "Pre-trained models: Past, present and future." *AI Open* 2 (2021): 225-250.

Background:Pre-trained Language Models

- PLMs has shown great success on a variety of benchmark datasets in NLP
- The next goal is to make PLMs fit in real-life use case
  - How unrealistic is PLMs nowadays?

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
- Closing Remarks

The Problems of PLMs

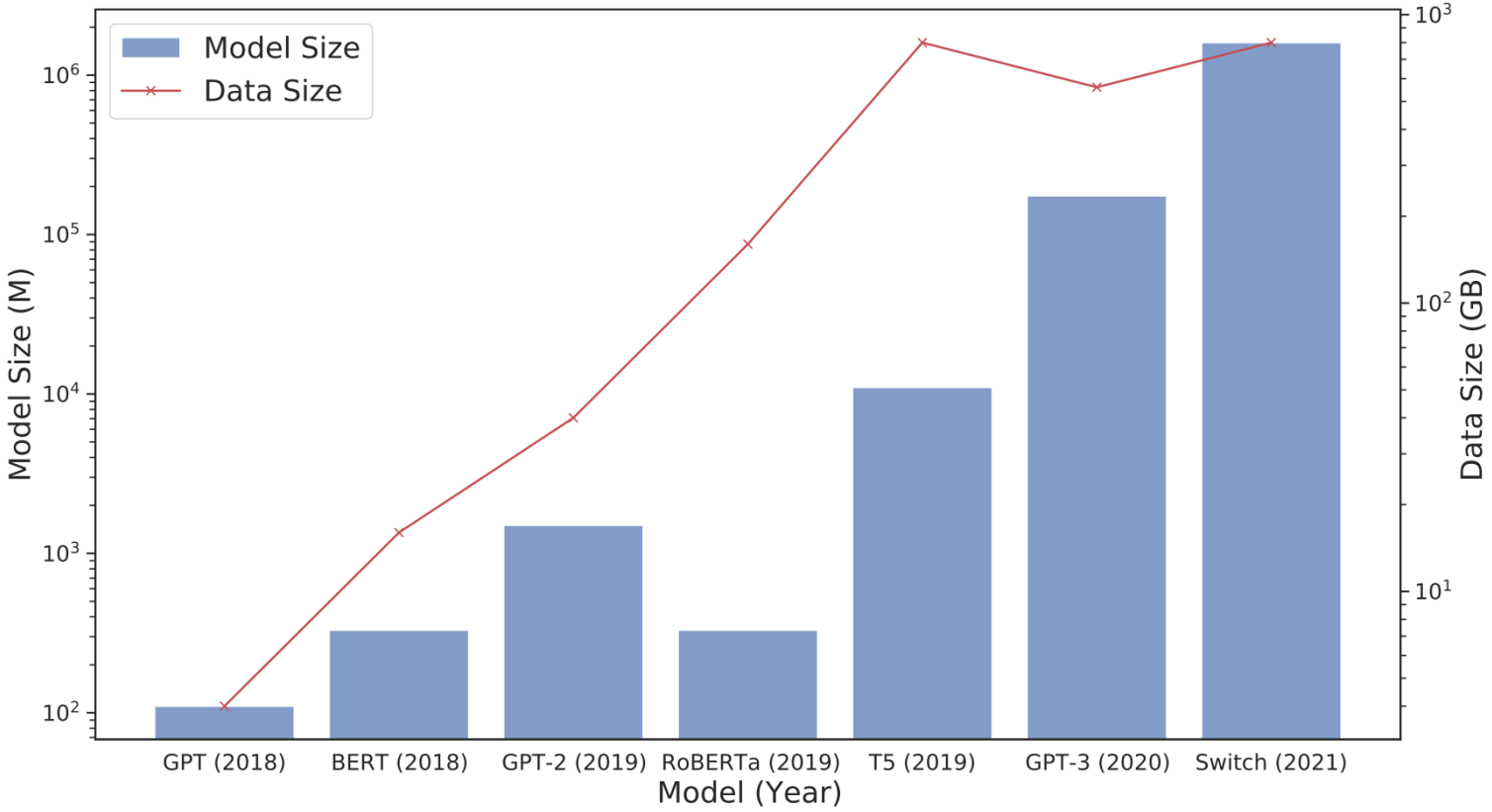
- Problem 1: **Data scarcity** in downstream tasks
- A large amount of labeled data is not easy to obtain for each downstream task

Pre-trained Language Model
(Fine-tuning)

MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k

The Problems of PLMs

- Problem 2: The PLM is too big, and they are still getting bigger

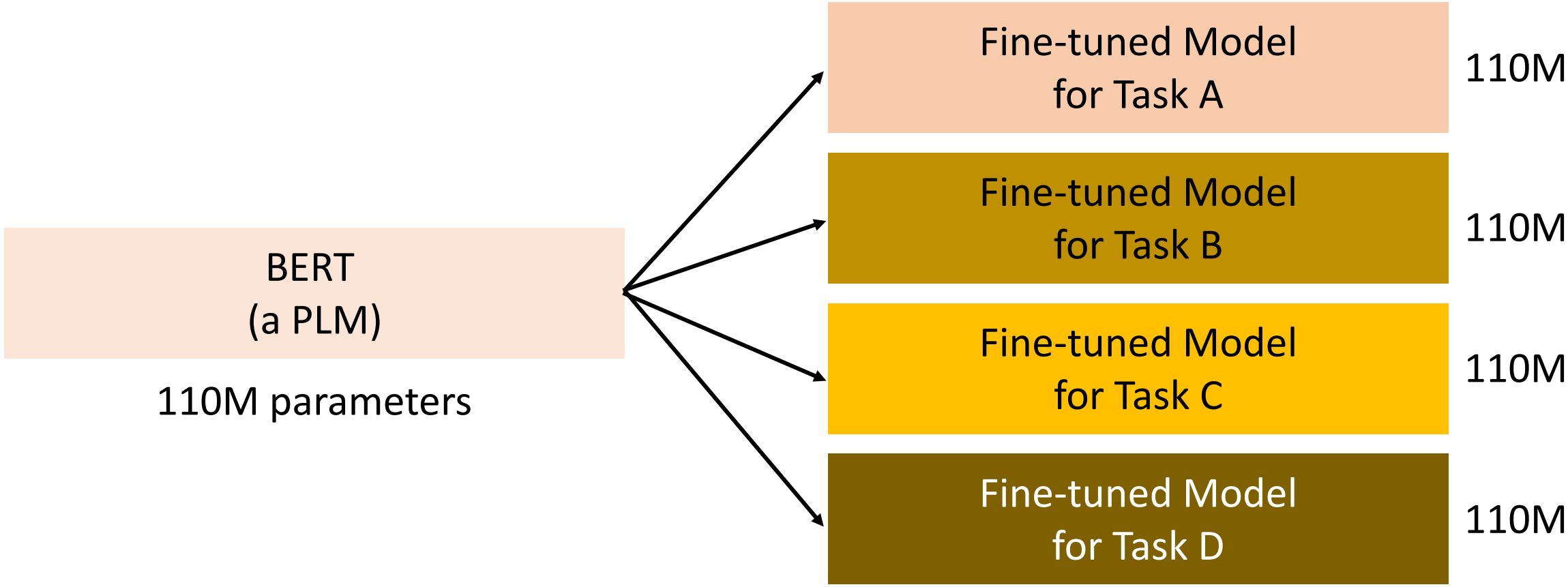


(b) The model size and data size applied by recent NLP PTMs.

Han, Xu, et al. "Pre-trained models: Past, present and future." *AI Open* 2 (2021): 225-250.

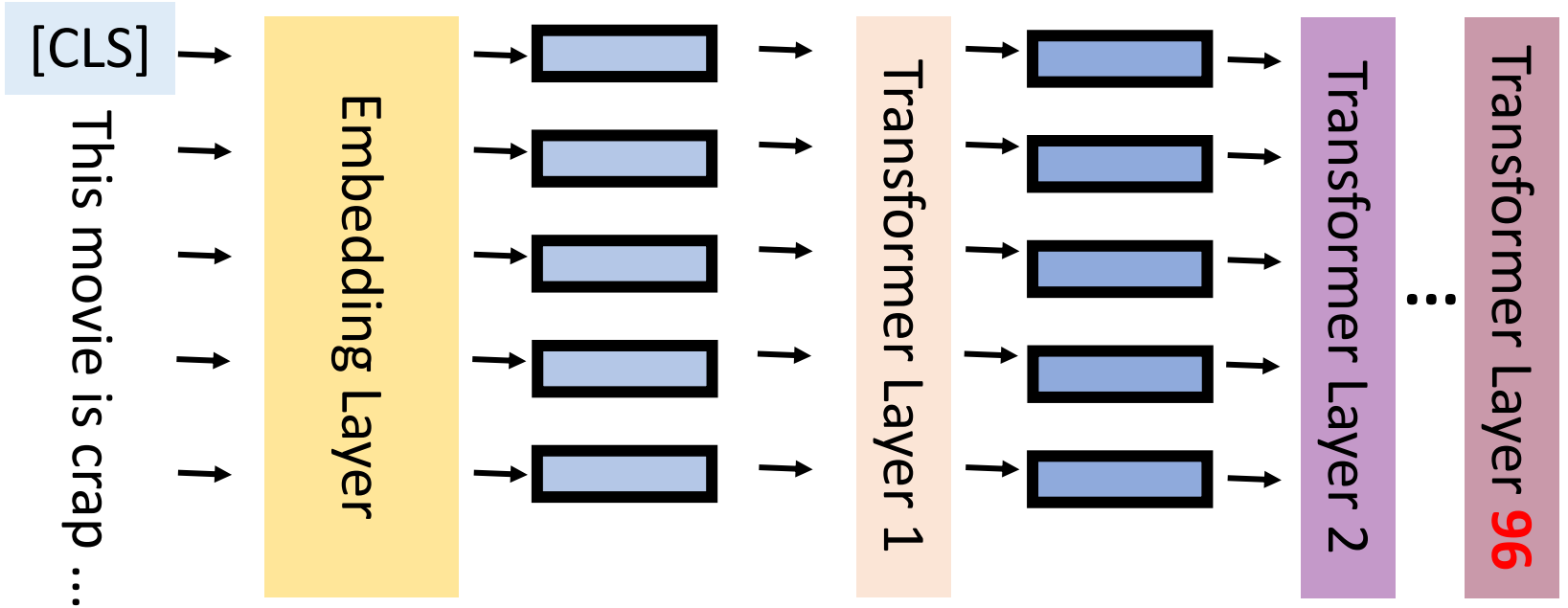
The Problems of PLMs

- Problem 2: The PLM is too big
  - Need a copy for each downstream task



The Problems of PLMs

- Problem 2: The PLM is too big
  - Inference takes too long
  - Consume too much space

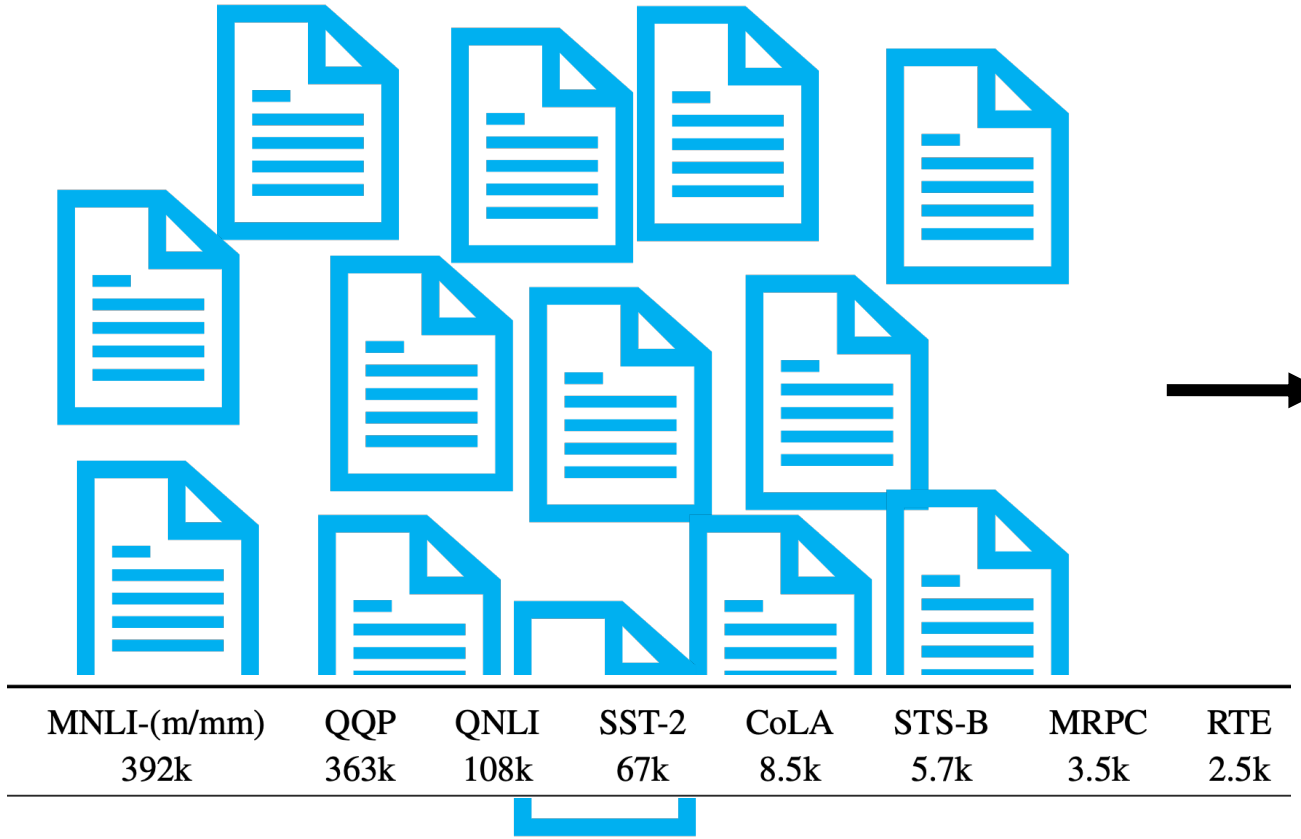


Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

The Problems of PLMs

- Problem 1: **Data scarcity** in downstream tasks
- A large amount of labeled data is not easy to obtain for each downstream task



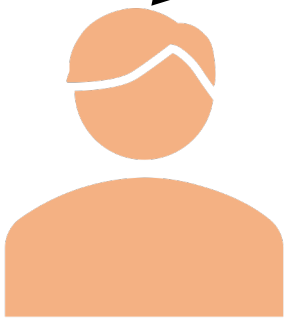
MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k

Pre-trained Language Model
(Fine-tuning)

Outline

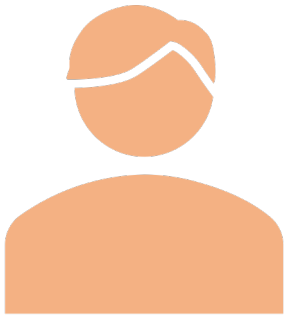
- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
    - Prompt Tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

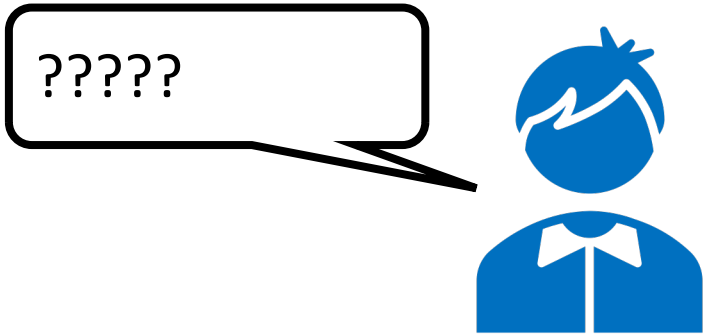
- [CLS] Jack likes dog. [SEP] Jack loves ice cream. [SEP] >>> **neutral**
- [CLS] The spring break is coming soon. [SEP] The spring break was over. [SEP] >>> **contradiction**
- [CLS] I am going to have dinner. [SEP] I am going to eat something. [SEP] >>> **entailment**
- ……
- [CLS] Mary likes pie. [SEP] Mary hates pie. [SEP] >>> ?



Natural language inference

- [CLS] The spring break is coming soon. [SEP] The spring break was over. [SEP] >>> **contradiction**
- [CLS] I am going to have dinner. [SEP] I am going to eat something. [SEP] >>> **entailment**
- [CLS] Mary likes pie. [SEP] Mary hates pie. [SEP] >>> ?

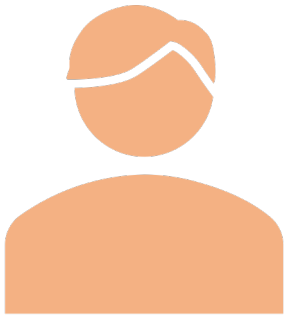




Natural language inference

Data-Efficient Fine-tuning: Prompt Tuning

- [CLS] The spring break is coming soon. Is it true that the spring break was over? >>> **no**
- [CLS] I am going to have dinner. Is it true that I am going to eat something? >>> **yes**
- [CLS] Mary likes pie. Is it true that Mary hates pie. [SEP] >>> ?



Natural language inference

Data-Efficient Fine-tuning: Prompt Tuning

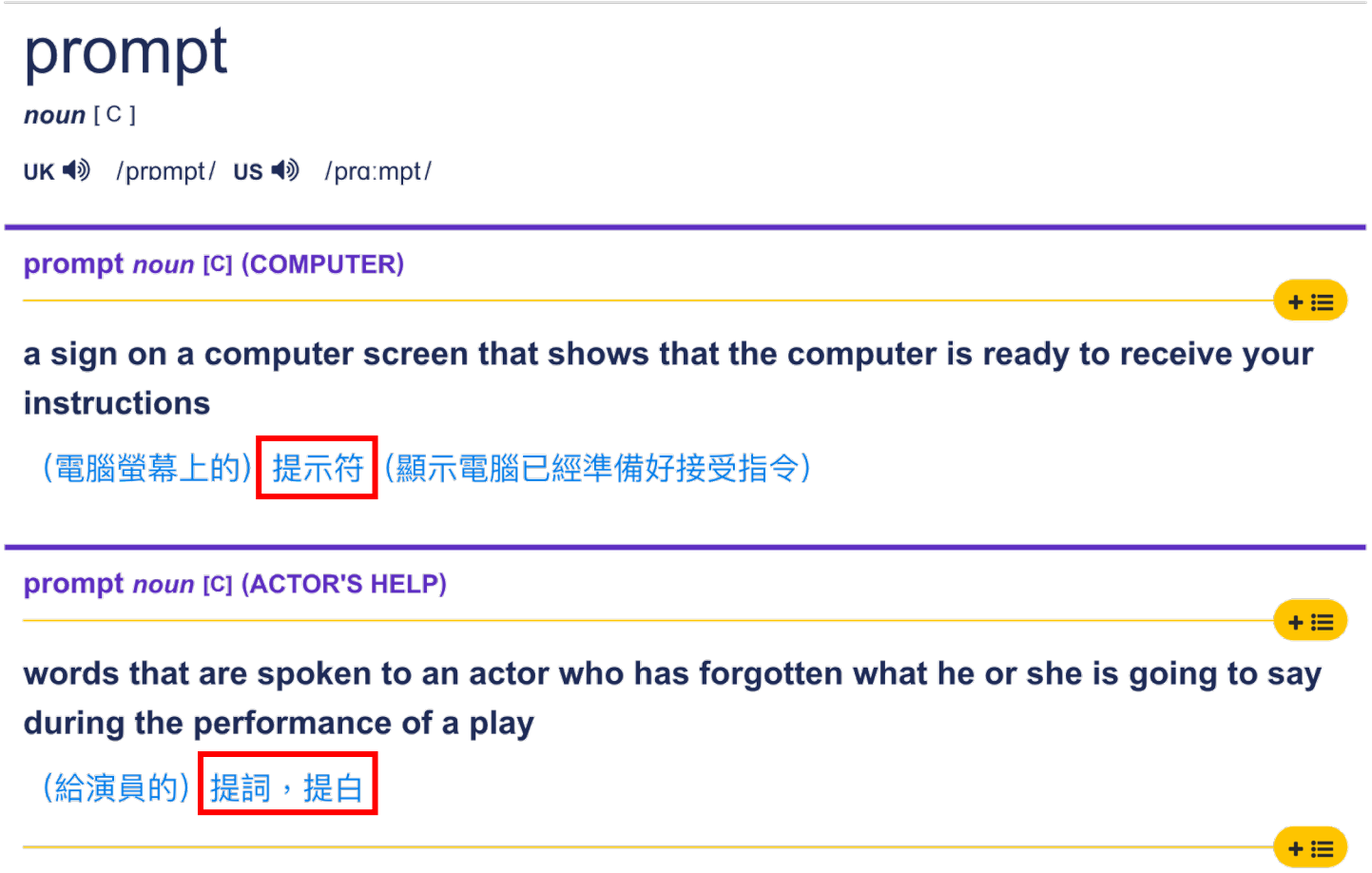
- By converting the data points in the dataset into natural language prompts, the model may be easier to know what it should do

- [CLS] The spring break is coming soon. [SEP] The spring break was over. [SEP] >>> **contradiction**
- [CLS] I am going to have dinner. [SEP] I am going to eat something. [SEP] >>> **entailment**
- [CLS] Mary likes pie. [SEP] Mary hates pie. [SEP] >>> ?

- [CLS] The spring break is coming soon. Is it true that the spring break was over? >>> **no**
- [CLS] I am going to have dinner. Is it true that I am going to eat something? >>> **yes**
- [CLS] Mary likes pie. Is it true that Mary hates pie. [SEP] >>> ?

Data-Efficient Fine-tuning: Prompt Tuning

- Format the downstream task as a language modelling task with pre-defined templates into natural language **prompts**



prompt

noun [C]

UK /prɒmpt/ US /prɑːmpt/

prompt *noun* [C] (COMPUTER)

a sign on a computer screen that shows that the computer is ready to receive your instructions

（電腦螢幕上的）提示符（顯示電腦已經準備好接受指令）

prompt *noun* [C] (ACTOR'S HELP)

words that are spoken to an actor who has forgotten what he or she is going to say during the performance of a play

（給演員的）提詞，提白

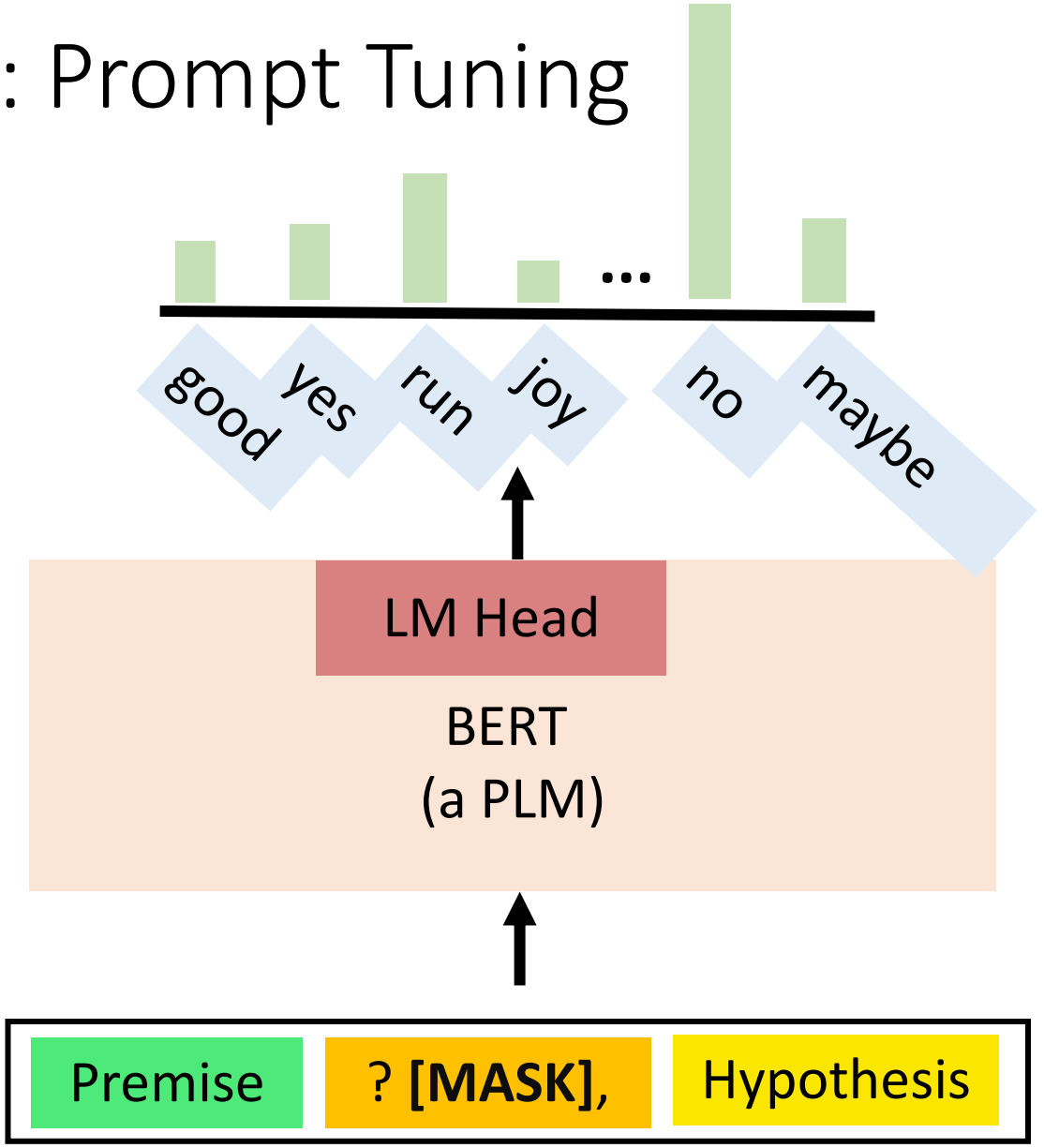
https://dictionary.cambridge.org/zht/%E8%A9%9E%E5%85%B8/%E8%8B%B1%E8%AA%9E-%E6%BC%A2%E8%AA%9E-%E7%B9%81%E9%AB%94/prompt

Data-Efficient Fine-tuning: Prompt Tuning

- What you need in prompt tuning
  1. A prompt template
  2. A PLM
  3. A verbalizer

Premise	Mary likes pie.
Hypothesis	Mary hates pie.
Label	2


```
"label" : [
  0 : "entailment"
  1 : "neutral"
  2 : "contradiction"
]
```

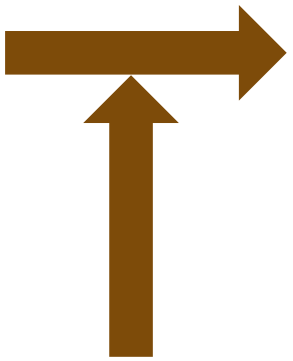
Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

- What you need in prompt tuning
  1. A prompt template: convert data points into a natural language prompt

Premise	Mary likes pie
Hypothesis	Mary hates pie
Label	2


```
"label" : [
  0 : "entailment"
  1 : "neutral"
  2 : "contradiction"
]
```

Prompt template

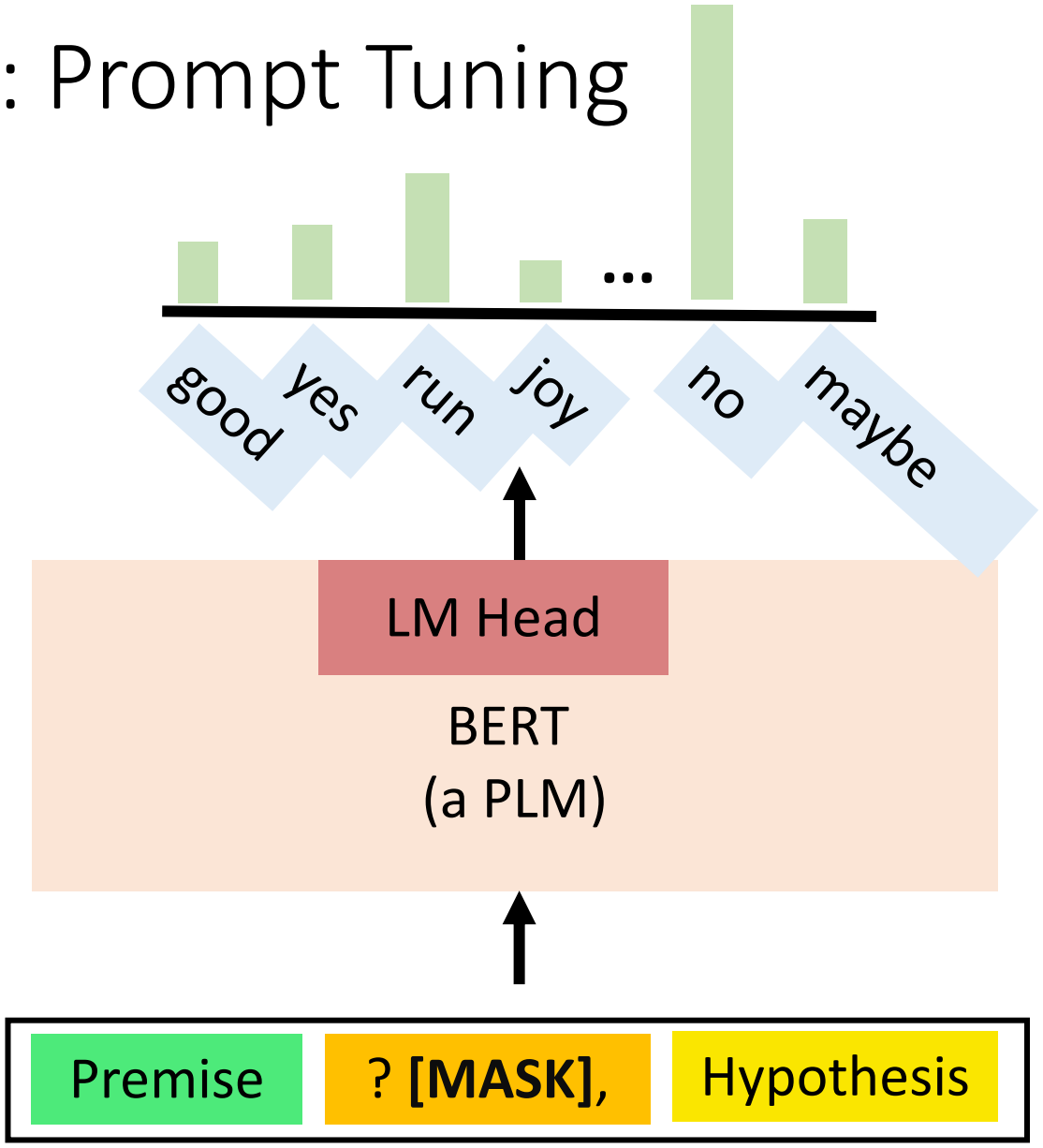
Mary likes pie | ? **[MASK]**, | Mary hates pie

Premise | ? **[MASK]**, | Hypothesis

Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

- What you need in prompt tuning
  2. A PLM: perform language modeling

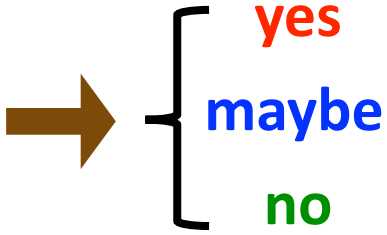


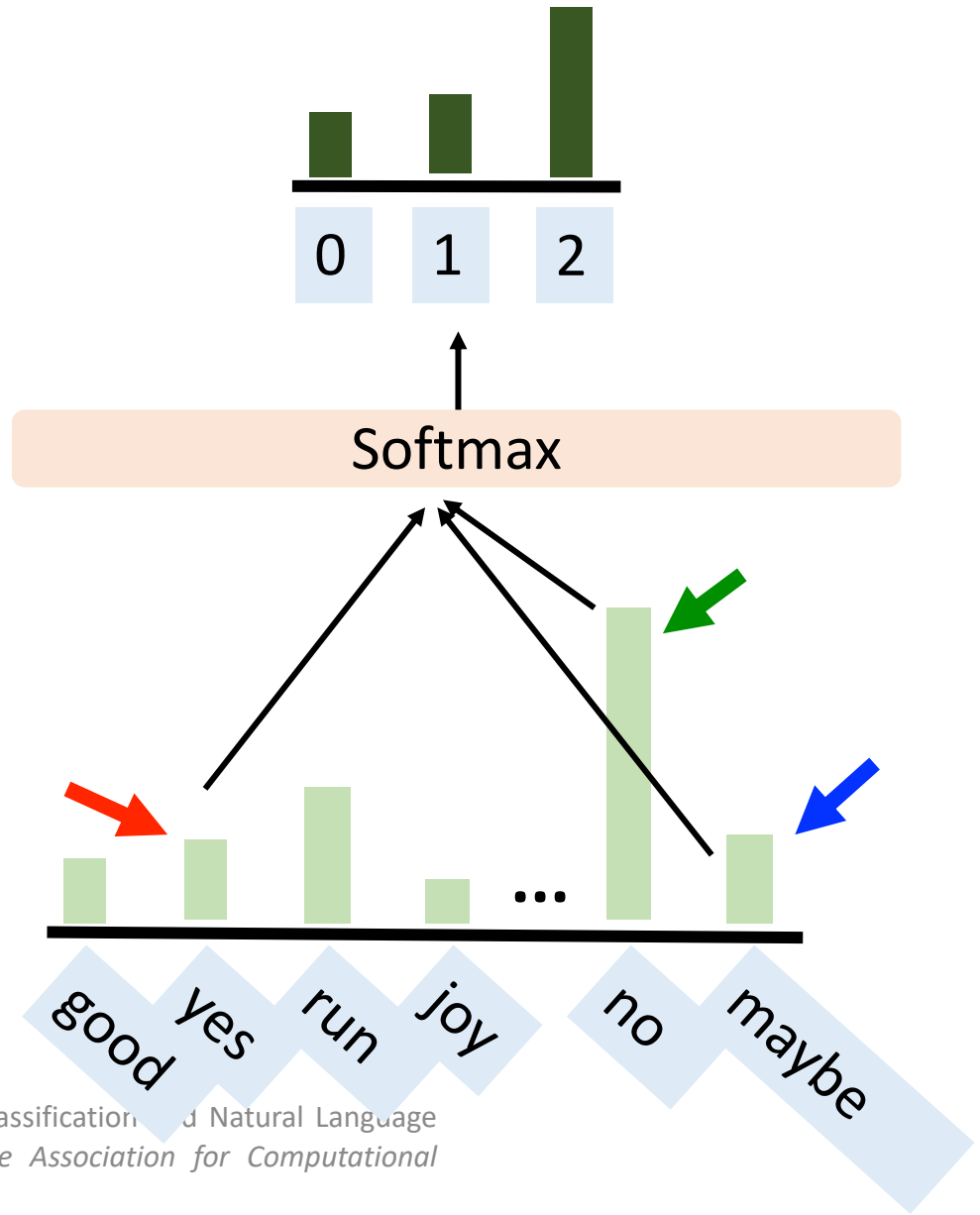
Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

- What you need in prompt tuning
  3. A verbalizer: A mapping between the label and the vocabulary
     - Which vocabulary should represents the class “entailment”

```
"label" : [
    0 : "entailment"
    1 : "neutral"
    2 : "contradiction"
]
```

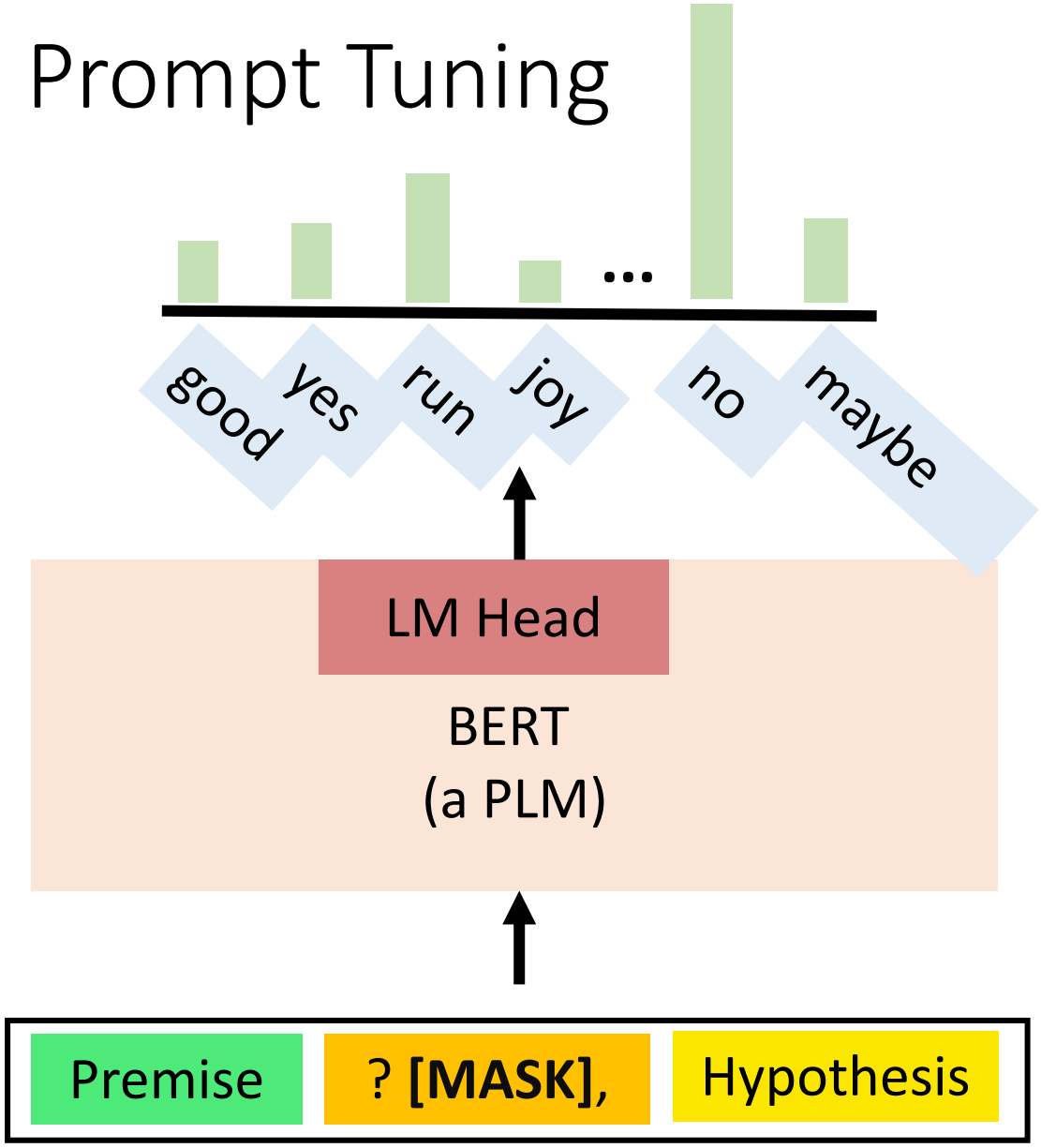




Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

- Prompt tuning
  - The whole PLM will be fine-tuned



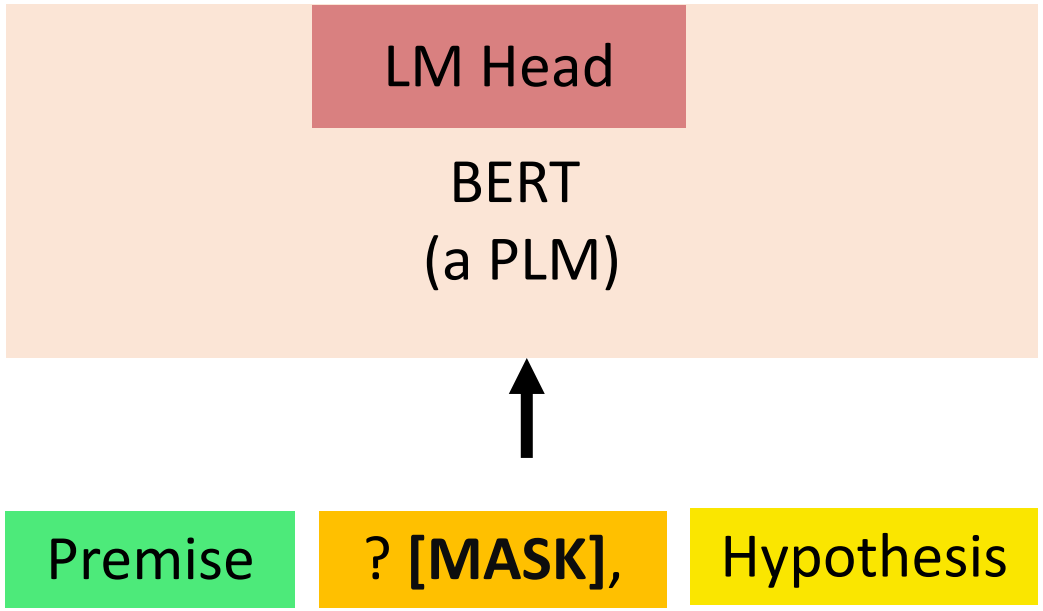
Premise	Mary likes pie.
Hypothesis	Mary hates pie.
Label	2

```
"label" : [
    0 : "entailment"
    1 : "neutral"
    2 : "contradiction"
]
```

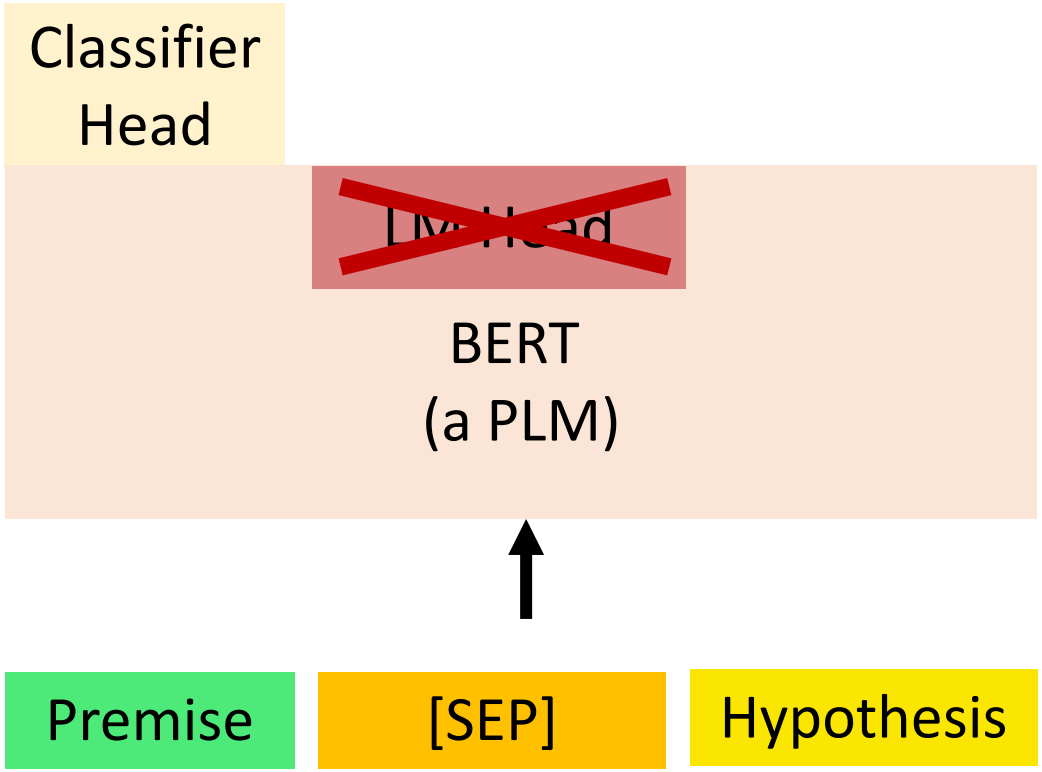
Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

- Prompt tuning



- Standard fine-tuning

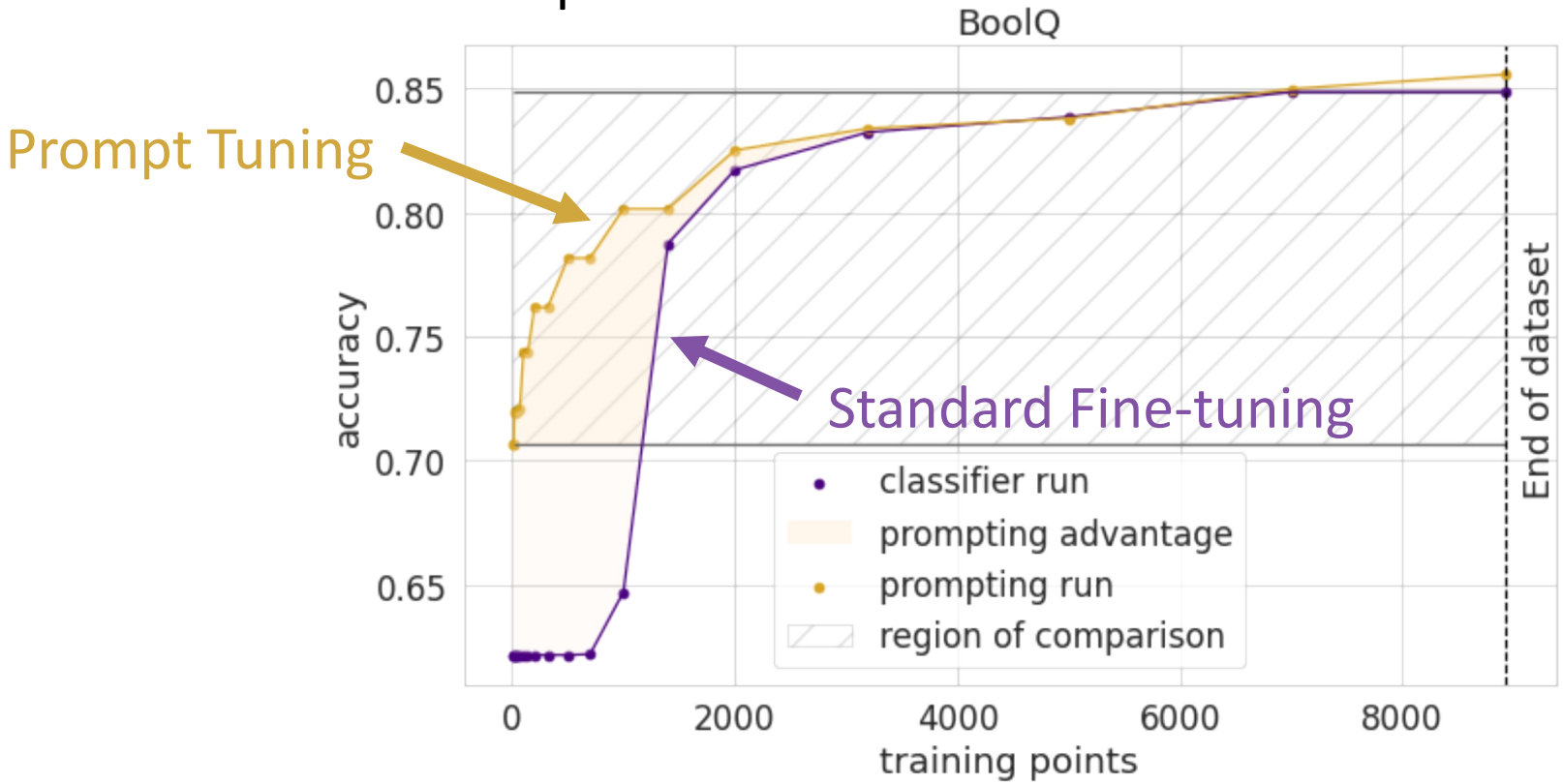


* I omit the [CLS] at the beginning and the [SEP] at the end

Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

- Prompt tuning has better performance under data scarcity because
  - It incorporates human knowledge
  - It introduces no new parameters

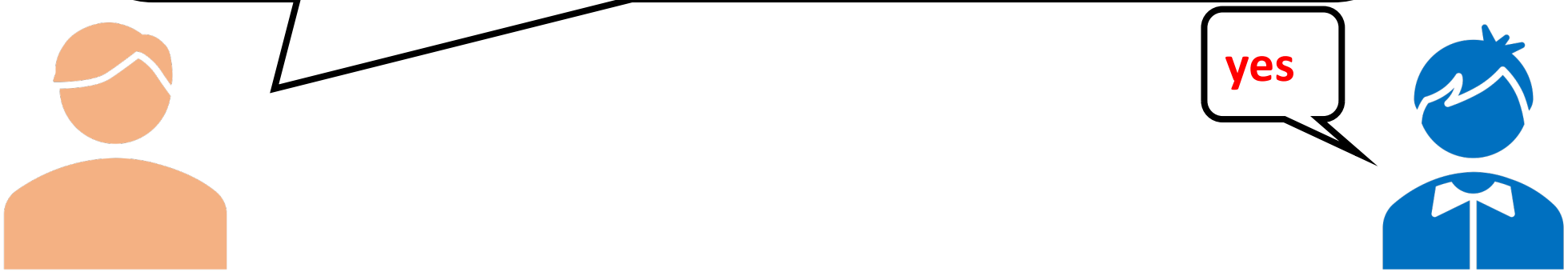


Le Scao, Teven, and Alexander M. Rush. "How many data points is a prompt worth?." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Data-Efficient Fine-tuning: Prompt Tuning

Lets see how prompts can help us under different level of data scarcity

- [CLS] The spring break is coming soon. Is it true that the spring break was over? >>> **no**
- [CLS] I am going to have dinner. Is it true that I am going to eat something? >>> **yes**
- [CLS] Mary likes pie. Is it true that Mary hates pie. [SEP] >>> ?



Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
    - Few-shot Learning
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

Data-Efficient Fine-tuning: Few-shot Learning

- Few-shot learning: We have some labeled training data
  - ”Some” ≈ 10 幾筆

Labeled Training data

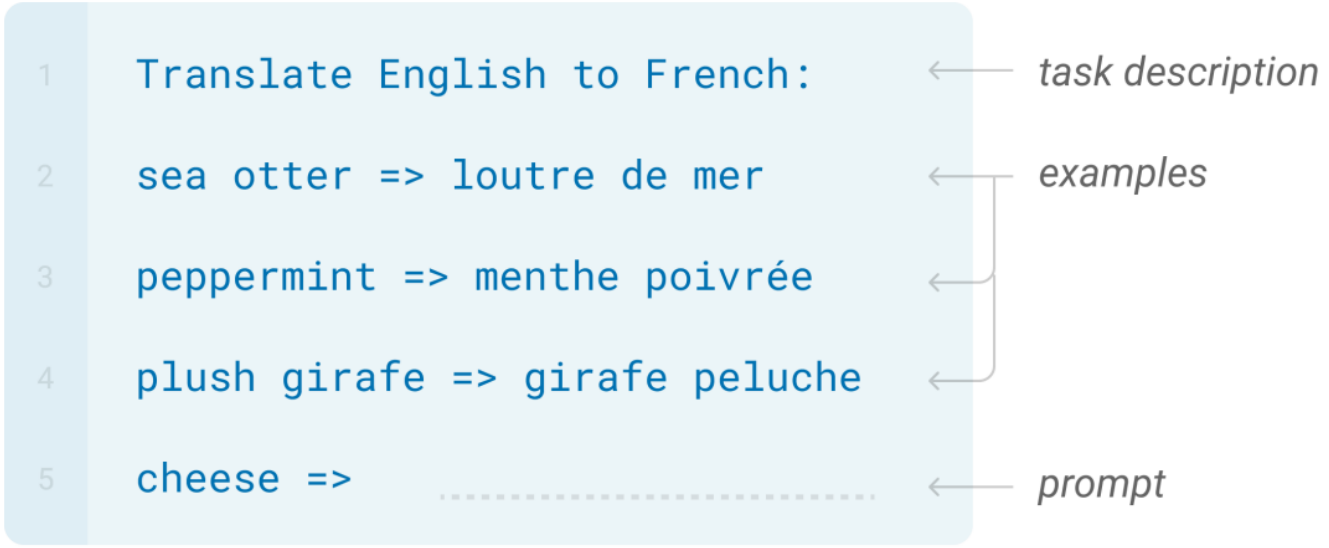
Pre-trained Language Model

Data-Efficient Fine-tuning: Few-shot Learning

- Good news: GPT-3 can be used for few-shot setting
- Bad news: GPT-3 is not freely available and contains 175B parameters

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



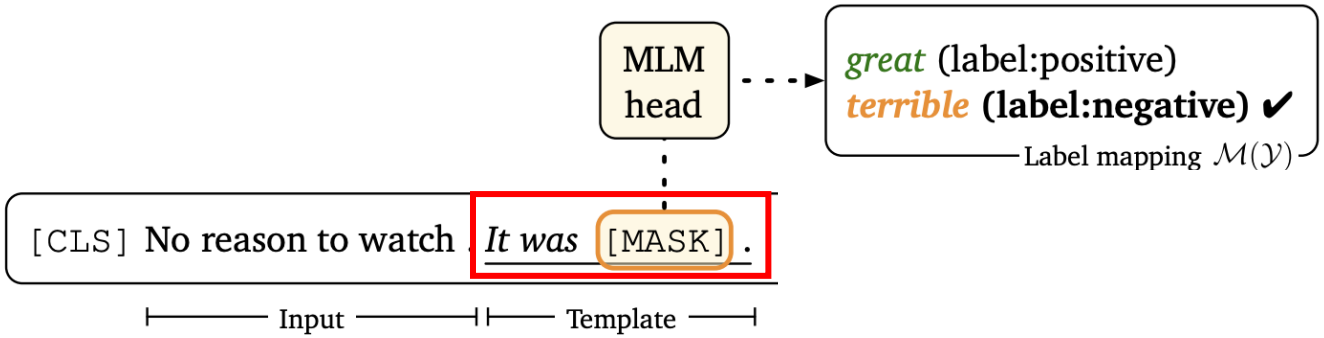
Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Data-Efficient Fine-tuning: Few-shot Learning

- Can we use smaller(?) PLMs and make them to perform well in few-shot learning?
- LM-BFF: **b**etter **f**ew-shot **f**ine-tuning of **l**anguage **m**odels

[1] Alternatively, language models' best friends forever.

  - Core concept: **prompt** + demonstration



Gao, Tianyu, Adam Fisch, and Danqi Chen. "Making Pre-trained Language Models Better Few-shot Learners." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

Data-Efficient Fine-tuning: Few-shot Learning

- LM-BFF
  - Prompt tuning: No new parameters are introduced during fine-tuning
  - Automatic template searching

	MNLI (acc)	MNLI-mm (acc)	SNLI (acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (Pear.)
Majority†	*32.7*	*33.0*	*33.8*	*49.5*	*52.7*	*81.2*	*0.0*	-
Prompt-based zero-shot‡	50.8	51.7	49.5	50.8	51.3	61.9	49.7	-3.2
Fine-tuning	45.8 (6.4)	47.8 (6.8)	48.4 (4.8)	60.2 (6.5)	54.4 (3.9)	76.6 (2.5)	60.7 (4.3)	53.5 (8.5)
$K=16$ Prompt-based FT (man)	68.3 (2.3)	70.5 (1.9)	77.2 (3.7)	64.5 (4.2)	69.1 (3.6)	74.5 (5.3)	65.5 (5.3)	71.0 (7.0)
+ demonstrations	**70.7** (1.3)	**72.0** (1.2)	**79.7** (1.5)	**69.2** (1.9)	68.7 (2.3)	77.8 (2.0)	**69.8** (1.8)	73.5 (5.1)
Prompt-based FT (auto)	68.3 (2.5)	70.1 (2.6)	77.1 (2.1)	68.3 (7.4)	**73.9** (2.2)	76.2 (2.3)	67.0 (3.0)	75.0 (3.3)
+ demonstrations	70.0 (3.6)	**72.0** (3.1)	77.5 (3.5)	68.5 (5.4)	71.1 (5.3)	**78.1** (3.4)	67.7 (5.8)	**76.4** (6.2)
Fine-tuning (full)†	*89.8*	*89.5*	*92.6*	*93.3*	*80.9*	*91.4*	*81.7*	*91.9*

Gao, Tianyu, Adam Fisch, and Danqi Chen. "Making Pre-trained Language Models Better Few-shot Learners." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
    - Semi-supervised Learning
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

Data-Efficient Fine-tuning: Semi-supervised Learning

- Semi-Supervised learning: We have some labeled training data and a large amount of unlabeled data

Data-Efficient Fine-tuning: Semi-supervised Learning

- Semi-Supervised learning: We have some labeled training data and a large amount of unlabeled data

It's Not Just Size That Matters:
Small Language Models Are Also Few-Shot Learners

Timo Schick[1,2] and **Hinrich Schütze**[1]

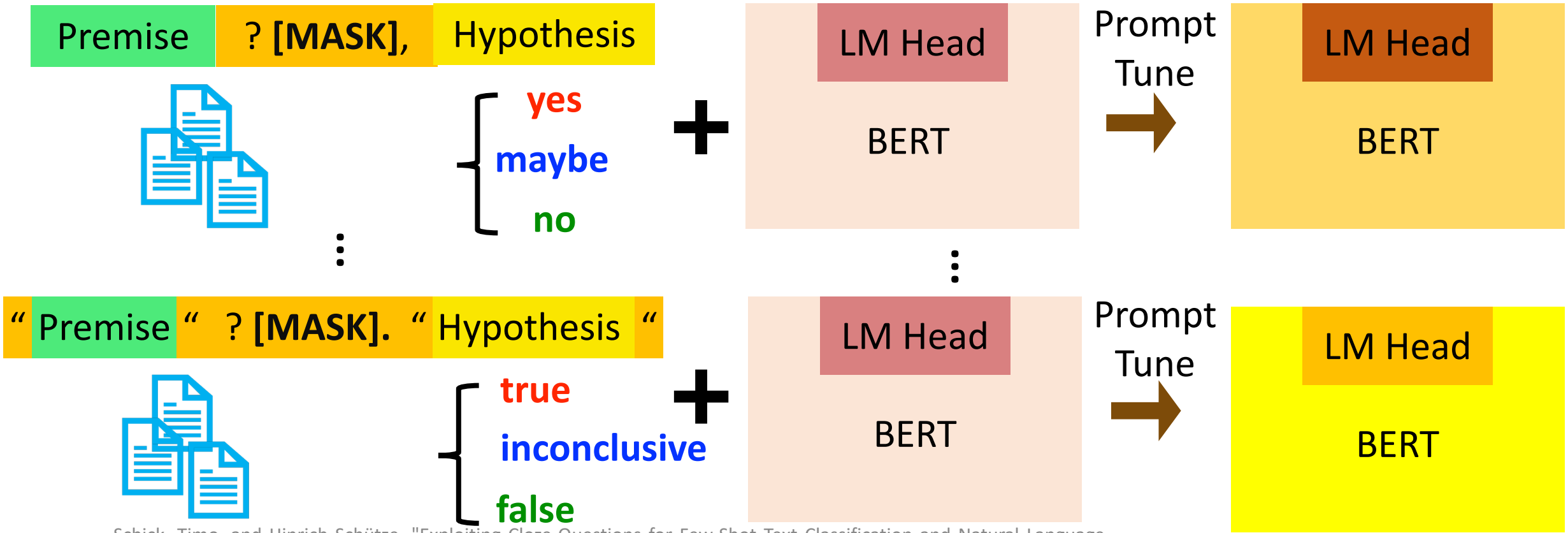
[1] Center for Information and Language Processing, LMU Munich, Germany
[2] Sulzer GmbH, Munich, Germany

timo.schick@sulzer.de

Schick, Timo, and Hinrich Schütze. "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

Data-Efficient Fine-tuning: Semi-supervised Learning

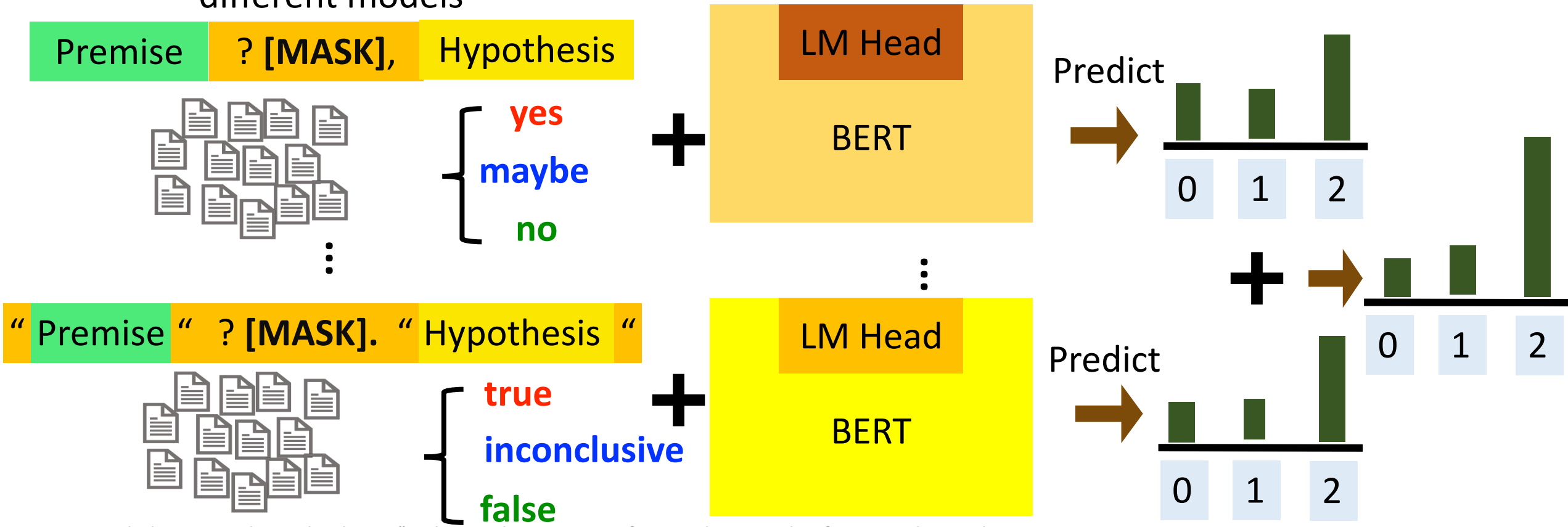
- Pattern-Exploiting Training (PET)
  - Step 1: Use different prompts and verbalizer to prompt-tune different PLMs on the labeled dataset



Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Semi-supervised Learning

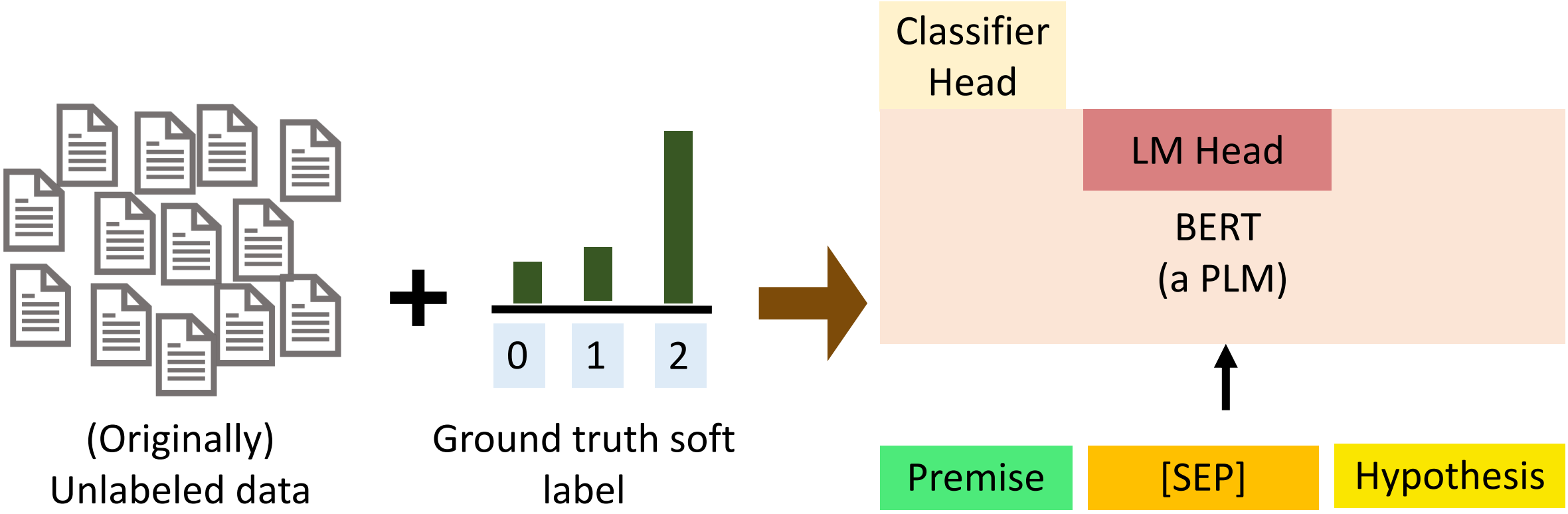
- Pattern-Exploiting Training (PET)
  - Step 2: Predict the unlabeled dataset and combine the predictions from different models



Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Data-Efficient Fine-tuning: Semi-supervised Learning

- Pattern-Exploiting Training (PET)
  - Step 3: Use a PLM with classifier head to train on the soft-labeled data set



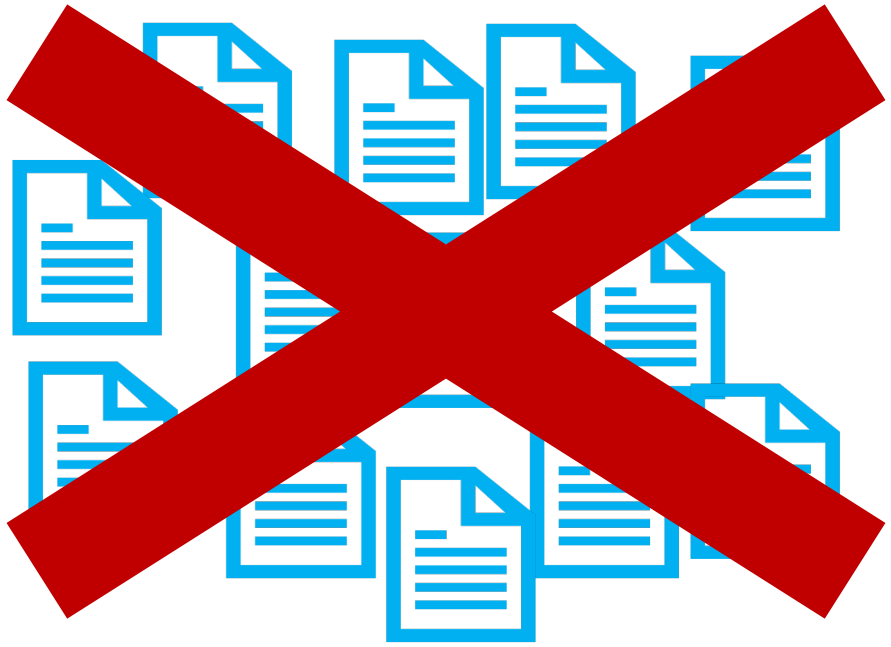
Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
    - Zero-shot Learning
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

Data-Efficient Fine-tuning(?): Zero-shot

- Zero-shot inference: inference on the downstream task without any training data
- If you don't have training data, then we need a model that can zero-shot inference on downstream tasks



Training data

Pre-trained Language Model

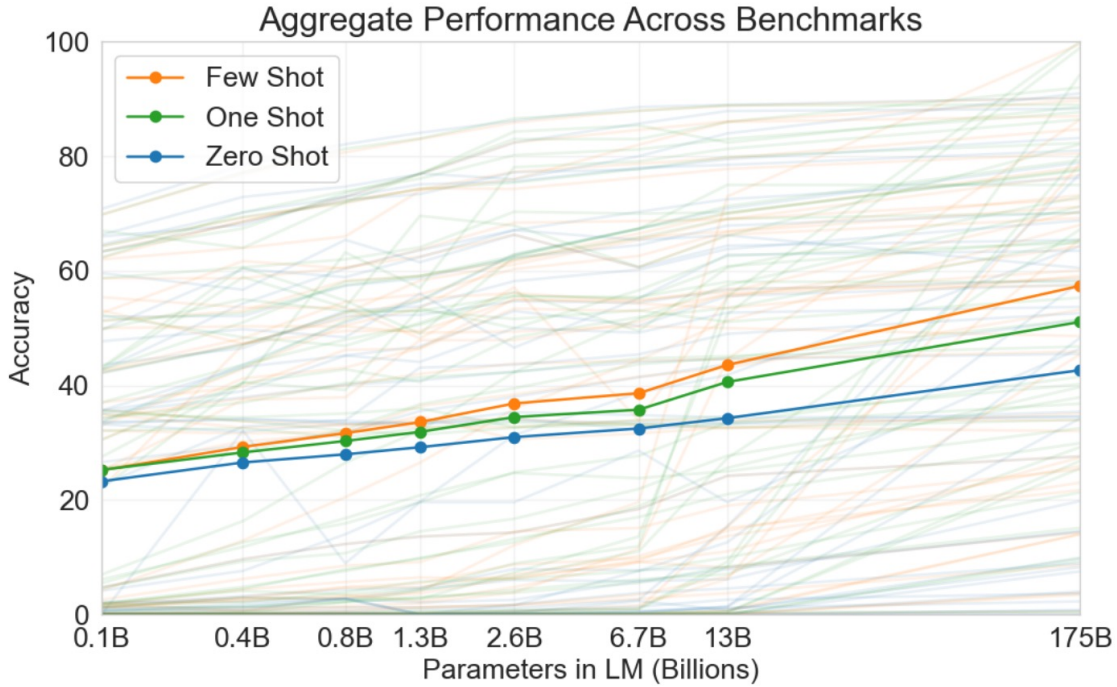
Data-Efficient Fine-tuning(?): Zero-shot

- GPT-3 shows that zero-shot (with task description) is possible
  - Only if your model is large enough

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```



Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Data-Efficient Fine-tuning(?): Zero-shot

- Where does this zero-shot ability spring from?
  - Hypothesis: during pre-training, the training datasets implicitly contains a mixture of different tasks
    - QA

      Q: I got 4 papers. Should I expect this load in the future?

      A: The average monthly load for reviewers should be much closer to 2, but in certain periods (close to large conferences), it's possible that the load is higher.
    - Summarization

Finetuned Language Models are Zero-Shot Learners

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, Quoc V Le

29 Sept 2021 (modified: 10 Feb 2022) ICLR 2022 Oral Readers: Everyone Show Bibtex Show Revisions

Keywords: natural language processing, zero-shot learning, language models

Abstract: This paper explores a simple method for improving the zero-shot learning abilities of language models. We show that instruction tuning—finetuning language models on a collection of datasets described via instructions—substantially improves zero-shot performance on unseen tasks. We take a 137B parameter pretrained language model and instruction tune it on over 60 NLP datasets verbalized via natural language instruction templates. We evaluate this instruction-tuned model, which we call FLAN, on unseen task types. FLAN substantially improves the performance of its unmodified counterpart and surpasses zero-shot 175B GPT-3 on 20 of 25 datasets that we evaluate. FLAN even outperforms few-shot GPT-3 by a large margin on ANLI, RTE, BoolQ, AI2-ARC, OpenbookQA, and StoryCloze. Ablation studies reveal that number of finetuning datasets, model scale, and natural language instructions are key to the success of instruction tuning.

One-sentence Summary: "Instruction tuning", which finetunes language models on a collection of tasks described via instructions, substantially boosts zero-shot performance on unseen tasks.

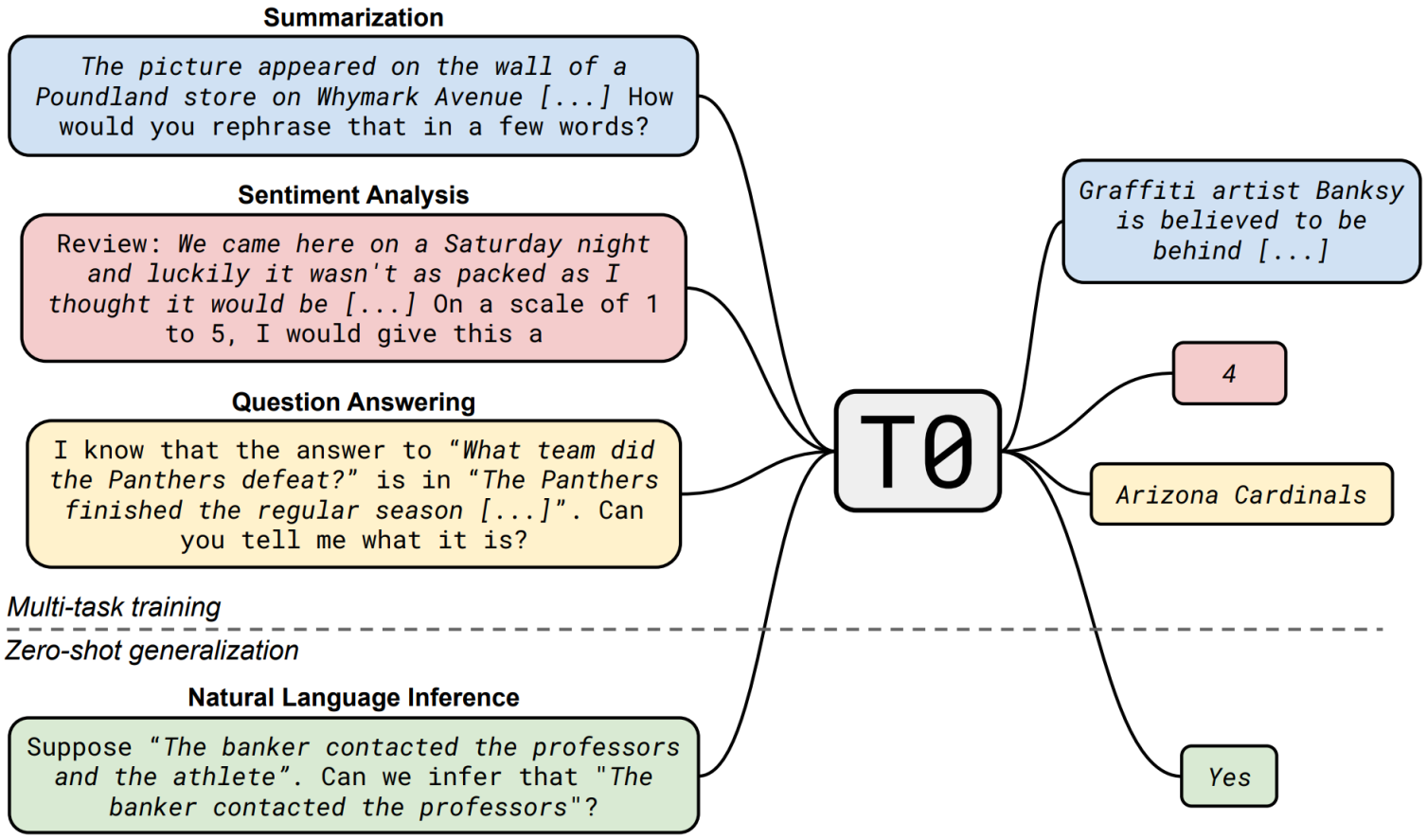
Wei, Jason, et al. "Finetuned Language Models Are Zero-Shot Learners." (2021).

Data-Efficient Fine-tuning(?): Zero-shot

- Hypothesis: multi-task training enables zero-shot generalization
  - Why not train a model with multi-task learning on a bunch of dataset?

Multi-task fine-tuning

Zero-shot Generalization



Sanh, Victor, et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization." *The Tenth International Conference on Learning Representations*. 2022.

Data-Efficient Fine-tuning(?): Zero-shot

- Multi-task fine-tuning using a PLM
  - Convert the task into a natural language **prompts**
  - Example: Natural Language Inference

```
"label" : [
  0 : "entailment"
  1 : "neutral"
  2 : "contradiction"
]
```

NLI dataset

premise	hypothesis	label
Conceptually cream skimming has two basic dimensions - product and geography.	Product and geography are what make cream skimming work.	1

Natural language prompt

```
{{premise}} Based on the previous passage, is it true that "{{hypothesis}}"? Yes, no, or maybe?
```

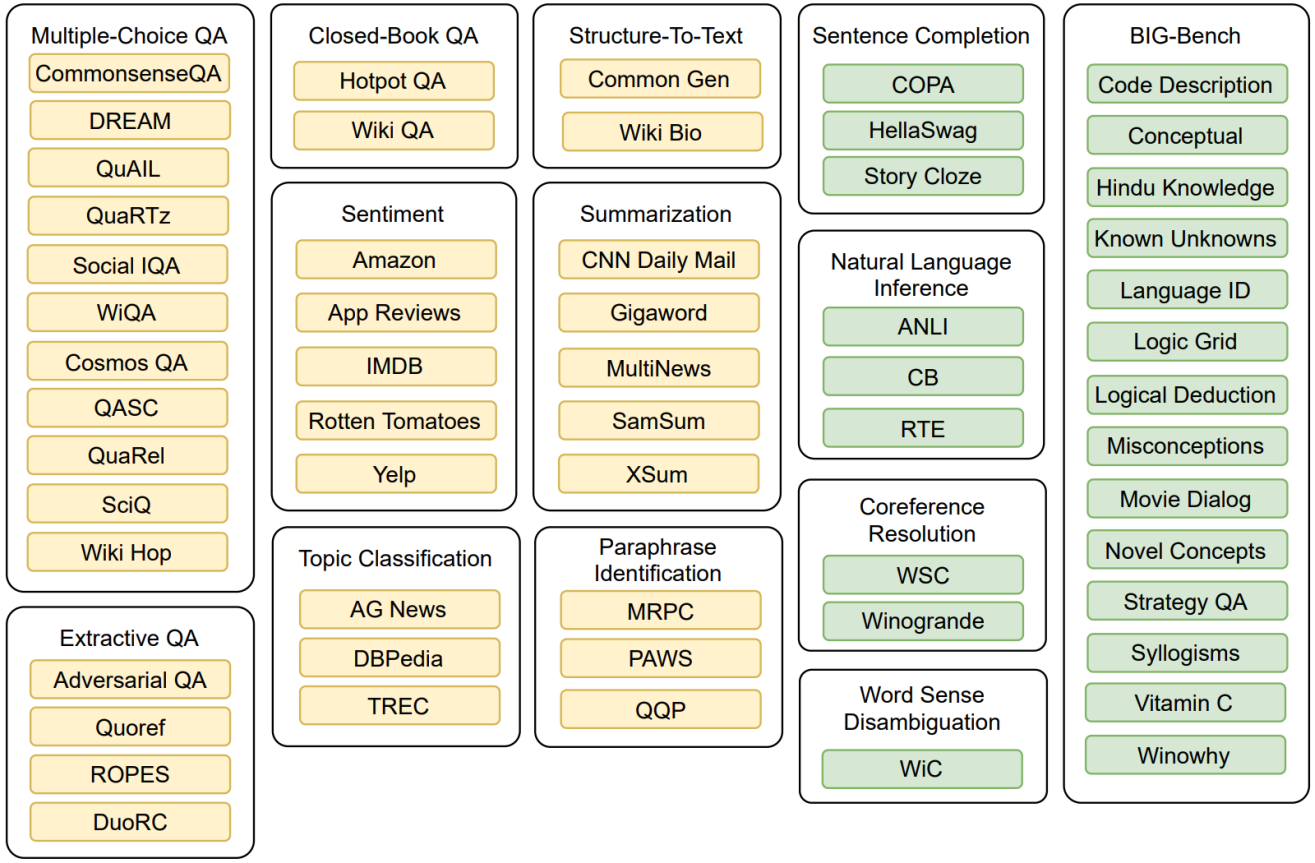
```
{{premise}} Based on that information, is the claim: "{{hypothesis}}" {{"true"}}, {{"false"}}, or {{"inconclusive"}}?
```

```
Given that {{premise}} Does it follow that {{hypothesis}} Yes, no, or maybe?
```

Sanh, Victor, et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization." *The Tenth International Conference on Learning Representations*. 2022.

Data-Efficient Fine-tuning(?): Zero-shot

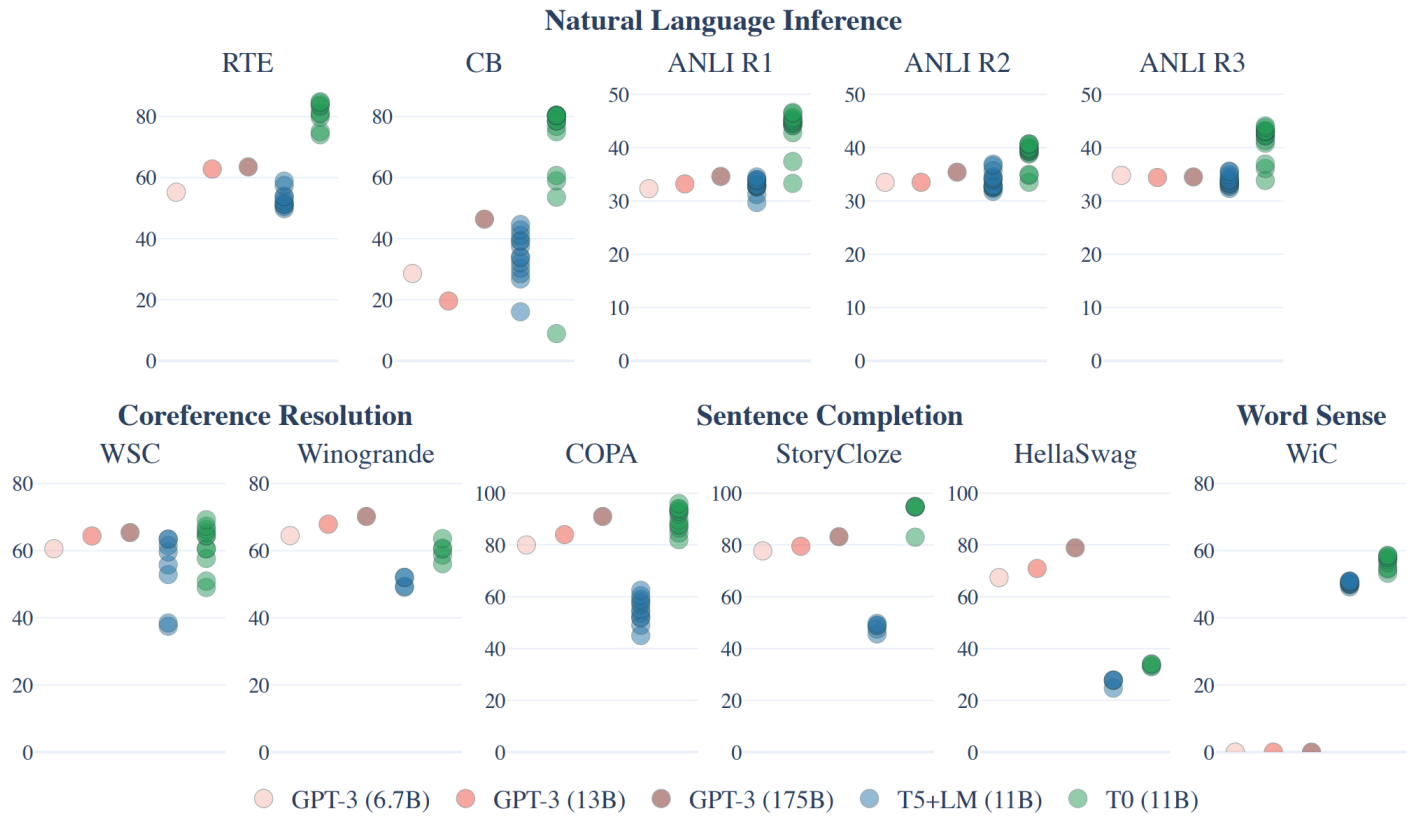
- Fine-tuning with some types of tasks and zero-shot inference on other types of tasks



Sanh, Victor, et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization." *The Tenth International Conference on Learning Representations*. 2022.

Data-Efficient Fine-tuning(?): Zero-shot

- Sometimes achieves performance better than GPT-3 (175B parameters) with ***only 11B*** parameters



Sanh, Victor, et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization." *The Tenth International Conference on Learning Representations*. 2022.

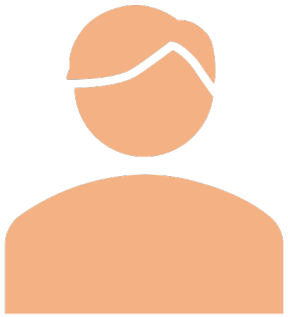
Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
    - Summary
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

Data-Efficient Fine-tuning: Summary

- Use natural language prompts and add scenario-specific designs

- [CLS] The spring break is coming soon. Is it true that the spring break was over? >>> **no**
- [CLS] I am going to have dinner. Is it true that I am going to eat something? >>> **yes**
- [CLS] Mary likes pie. Is it true that Mary hates pie. [SEP] >>> ?

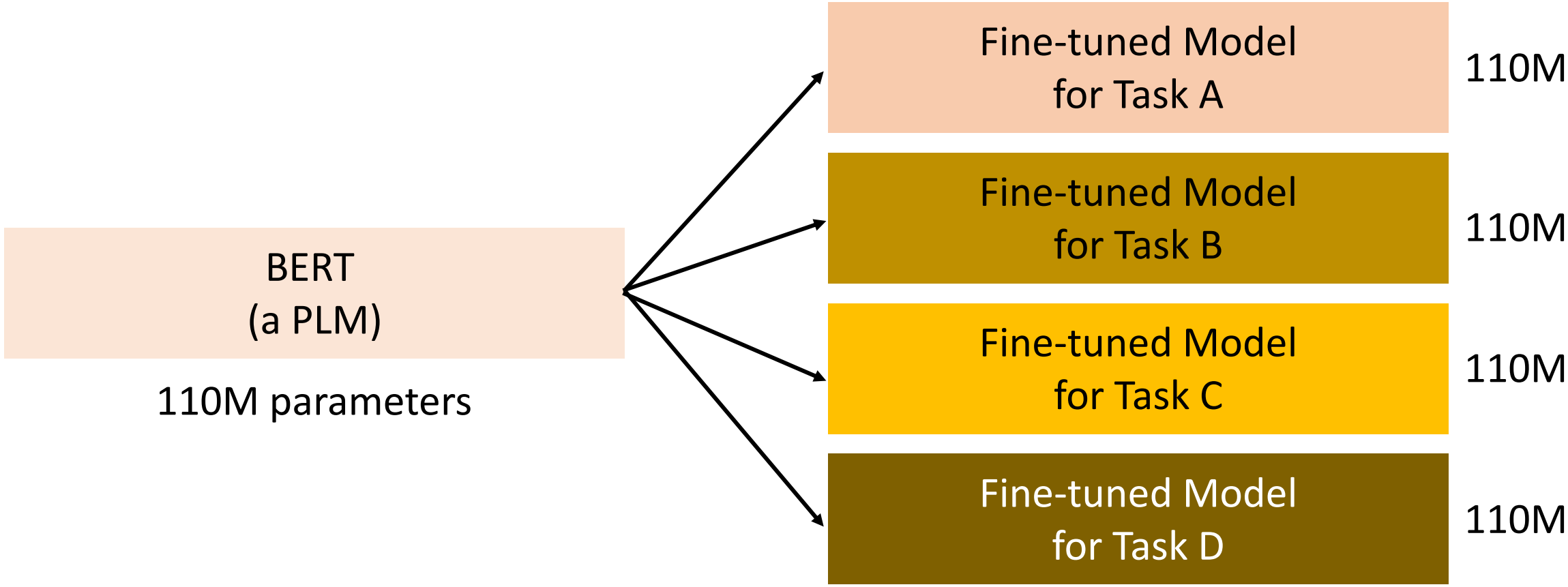


Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
- Closing Remarks

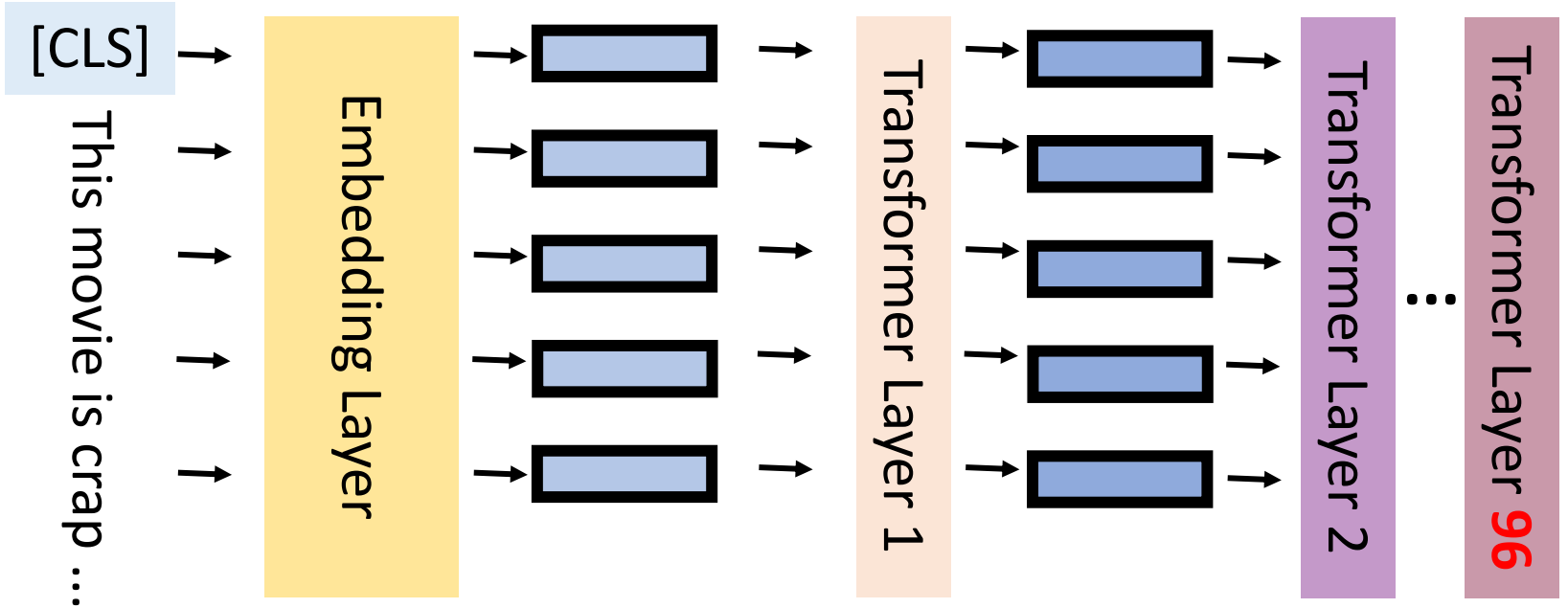
The Problems of PLMs

- Problem 2: The PLM is too big
  - Need a copy for each downstream task



The Problems of PLMs

- Problem 2: The PLM is too big
  - Inference takes too long
  - Consume too much space

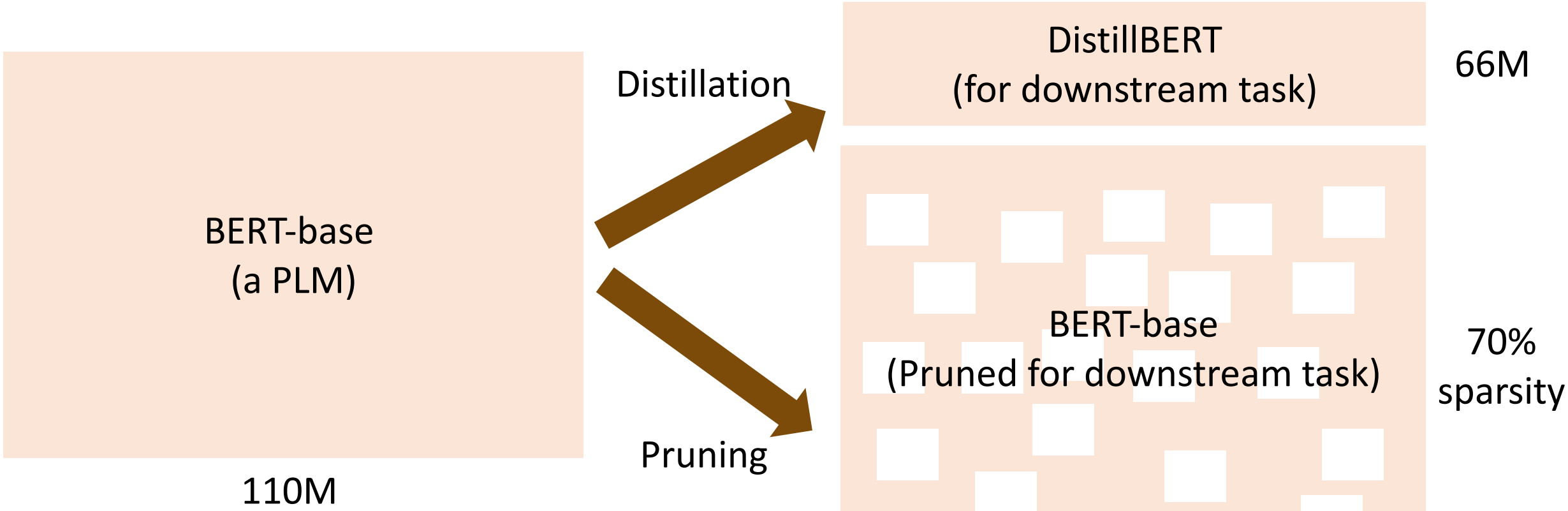


Reducing the Number of Parameters

- Problem: PLM is too large (in terms of numbers of parameters, model size, and the storage needed to store the model)
- Solution: Reduce the number of parameters
  - Smaller pre-trained model?

Reducing the Number of Parameters

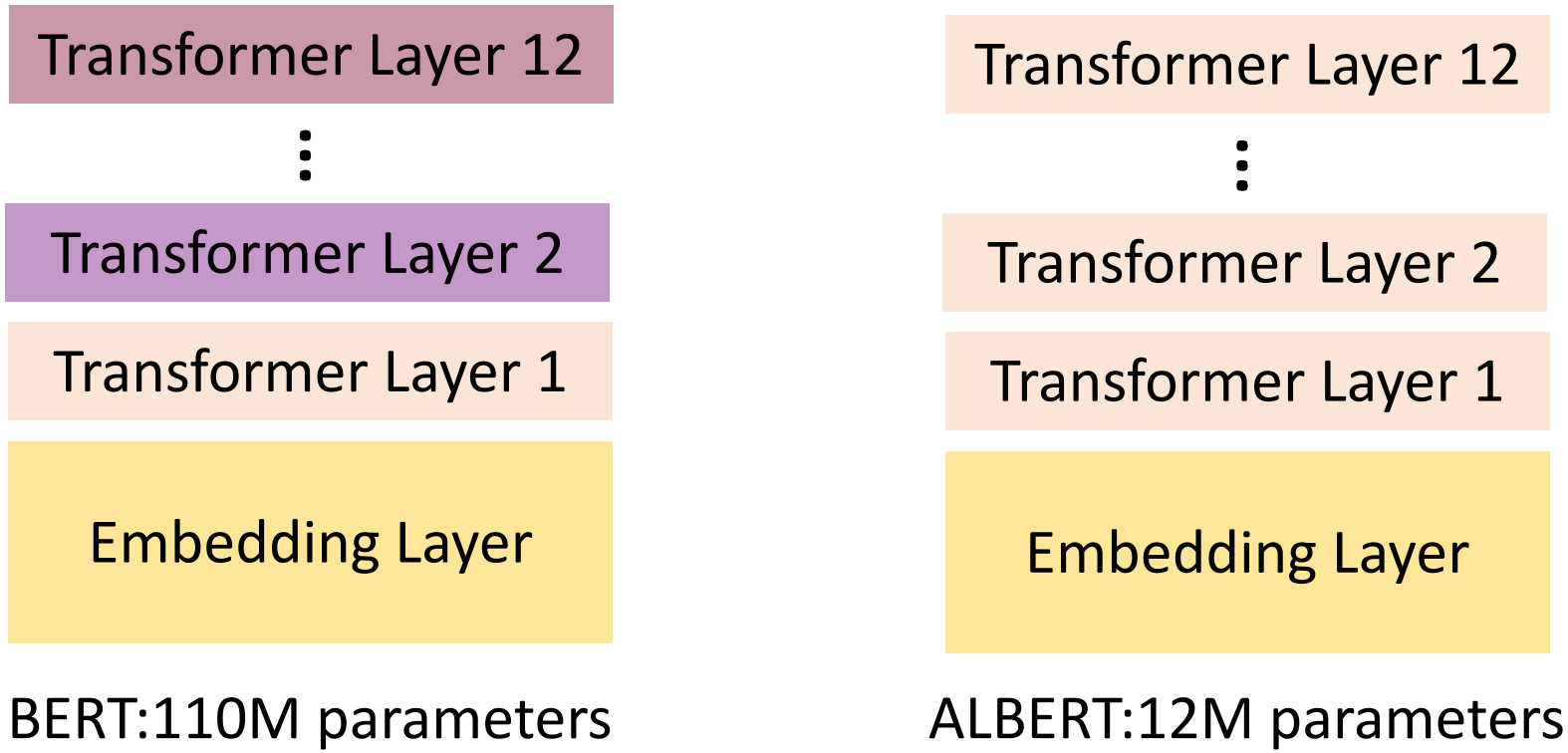
- Pre-train a large model, but use a smaller model for the downstream tasks



Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper an
arXiv:1910.01108
(2019).
Lai, Cheng-I. Jeff, et al. "Parp: Prune, adjust and re-prune for self-supervised speech recognition." *Advances in Neural Information Processing Systems* 34 (2021)

Reducing the Number of Parameters

- Share the parameters among the transformer layers



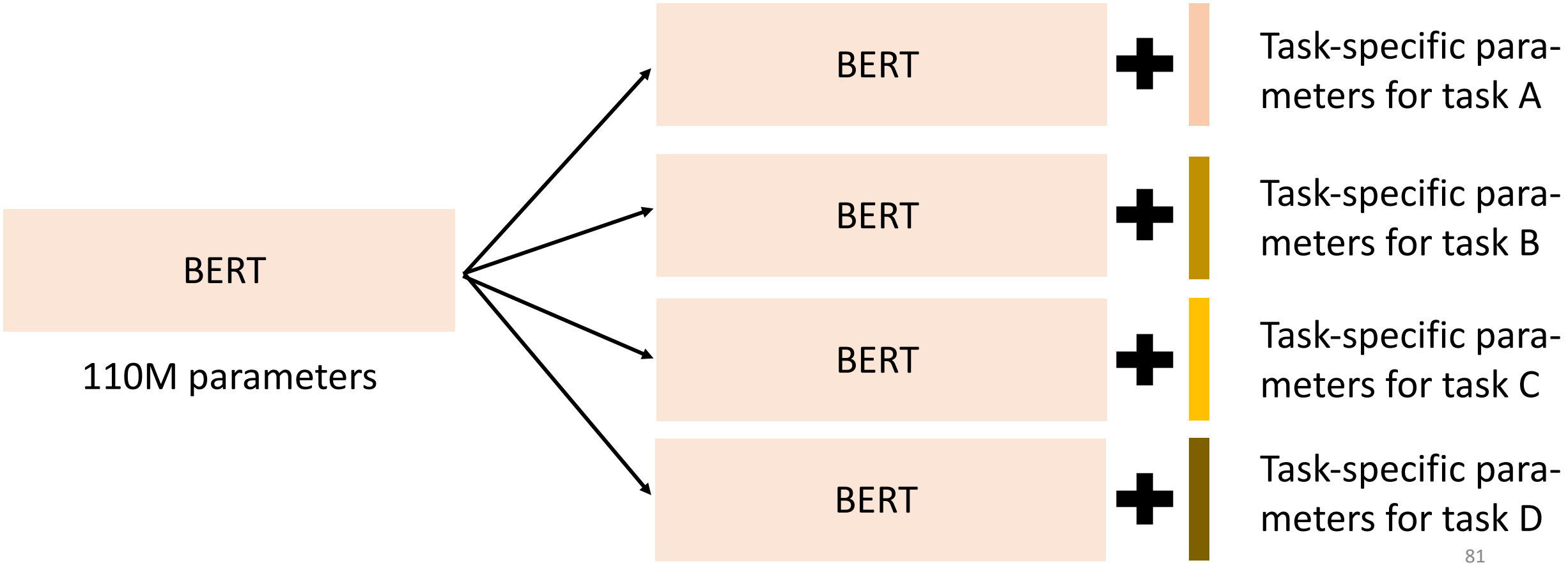
Lan, Zhenzhong, et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." *International Conference on Learning Representations*. 2019.

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Parameter-Efficient Fine-tuning
- Closing Remarks

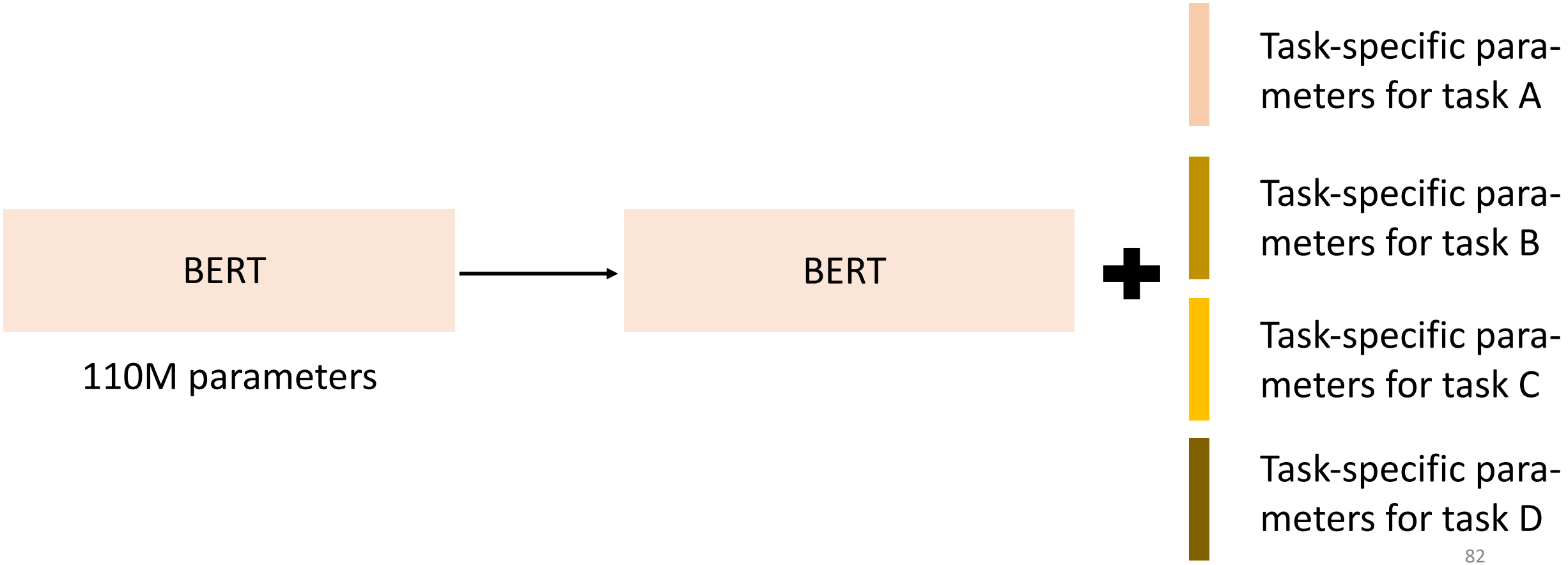
Parameter-Efficient Fine-tuning

- Use a small amount of parameters for each downstream task



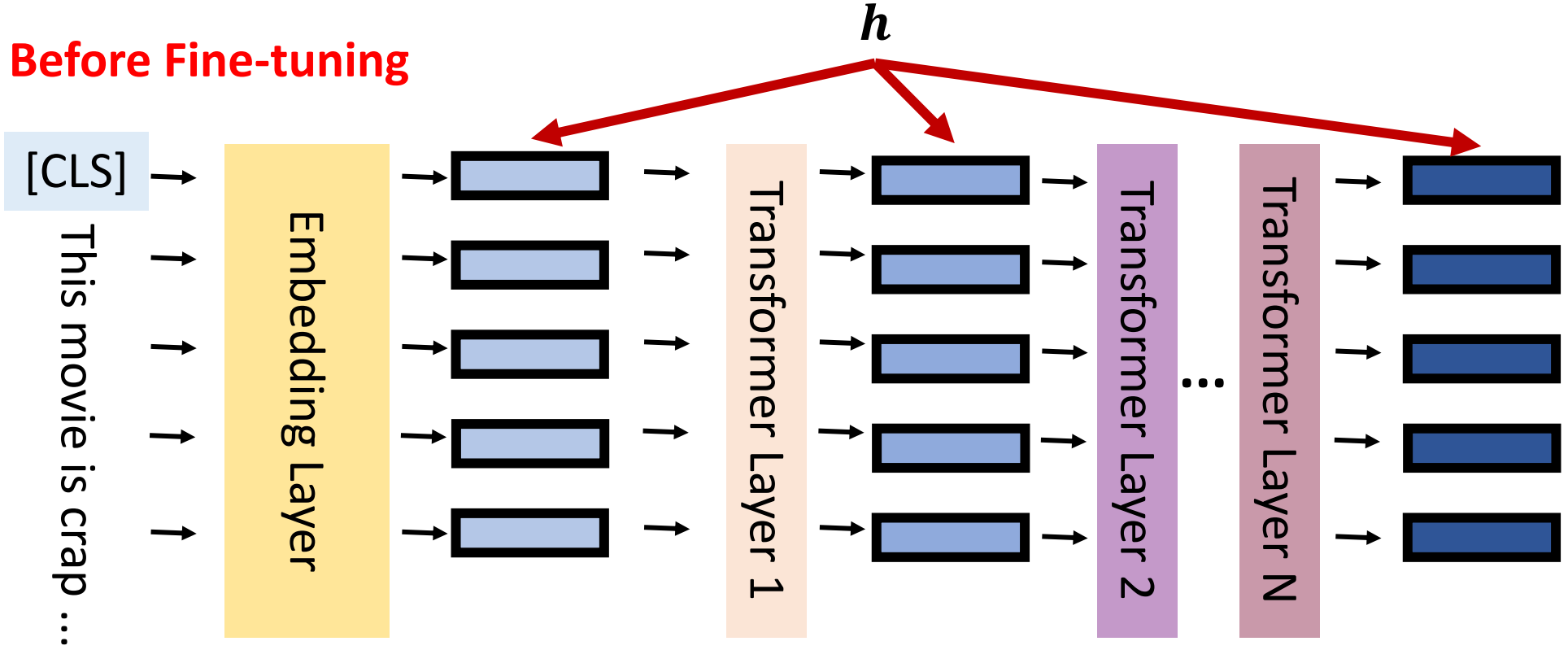
Parameter-Efficient Fine-tuning

- Use a small amount of parameters for each downstream task



Parameter-Efficient Fine-tuning

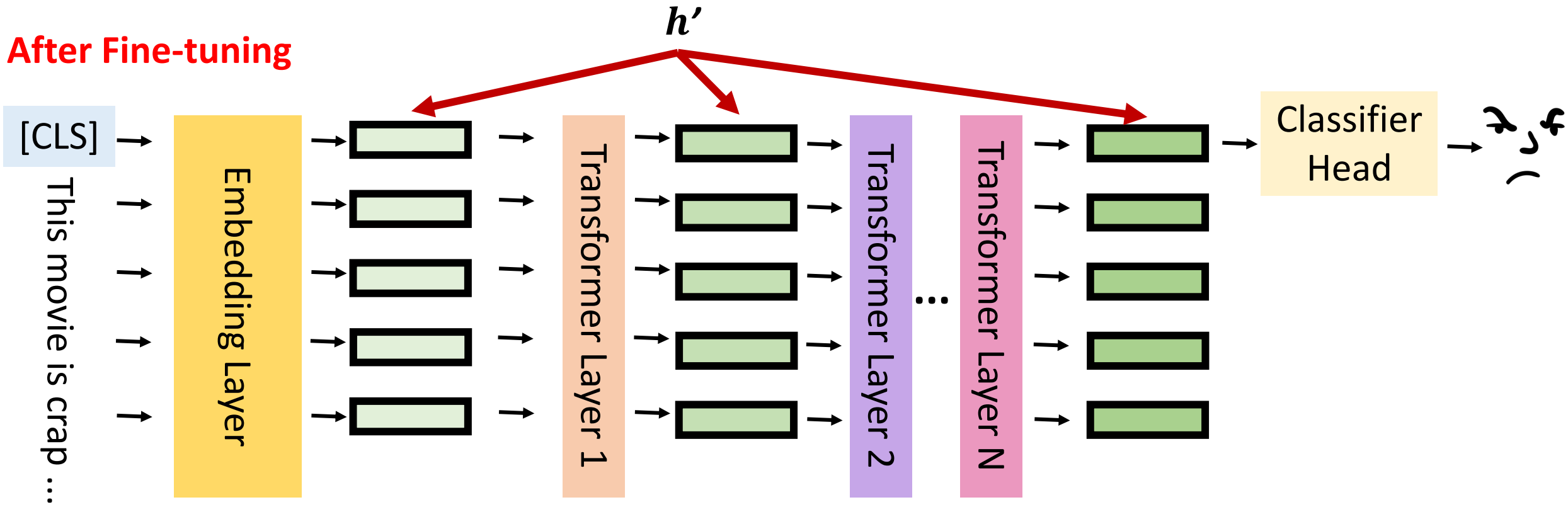
- What is standard fine-tuning really doing?
  - Modify the *hidden representations* ($\boldsymbol{h}$) of the PLM such that it can perform well on downstream task



He, Junxian, et al. "Towards a unified view of parameter-efficient transfer learning." *arXiv preprint arXiv:2110.04366* (2021).

Parameter-Efficient Fine-tuning

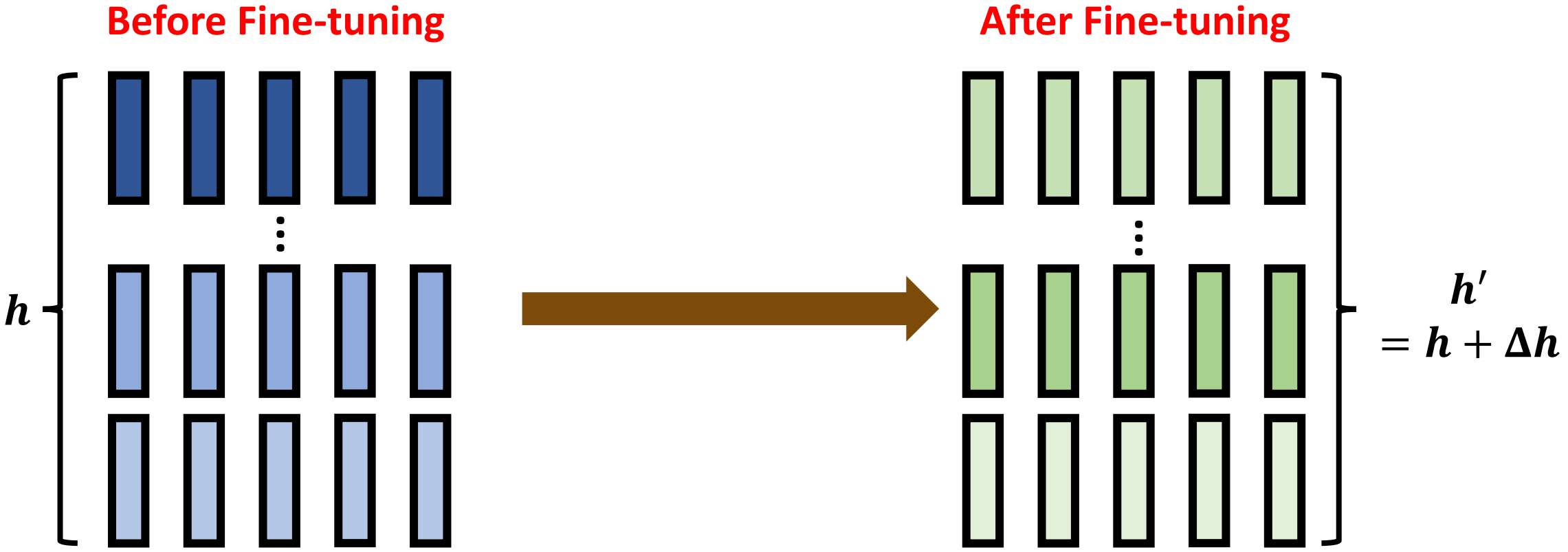
- What is standard fine-tuning really doing?
  - Modify the *hidden representations* ($\boldsymbol{h}$) of the PLM such that it can perform well on downstream task



He, Junxian, et al. "Towards a unified view of parameter-efficient transfer learning." *arXiv preprint arXiv:2110.04366* (2021).

Parameter-Efficient Fine-tuning

- Fine-tuning = modifying the hidden representation based on a PLM



$\boldsymbol{h}$:hidden representation
calculated by the original PLM

$\boldsymbol{h}' = \boldsymbol{h} + \boldsymbol{\Delta h}$
$\boldsymbol{h}'$:hidden representation
calculated by the fine-tuned model

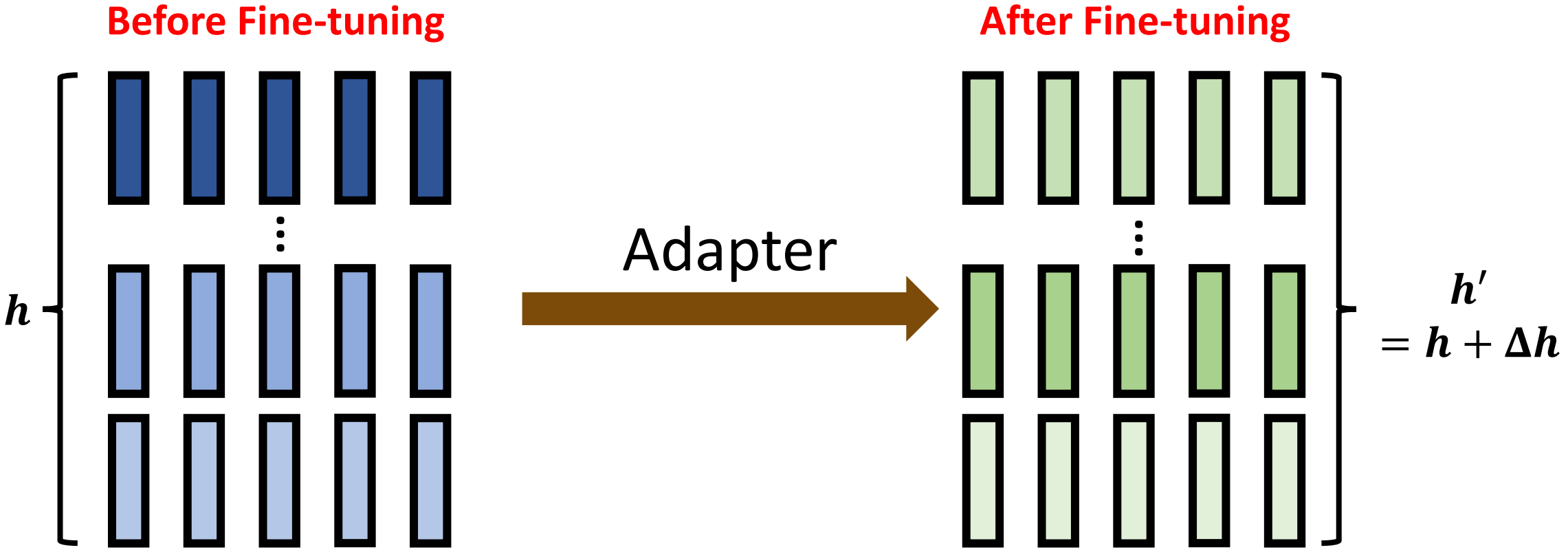
He, Junxian, et al. "Towards a unified view of parameter-efficient transfer learning." *arXiv preprint arXiv:2110.04366* (2021).

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Parameter-Efficient Fine-tuning
      - Adapter
- Closing Remarks

Parameter-Efficient Fine-tuning: Adapter

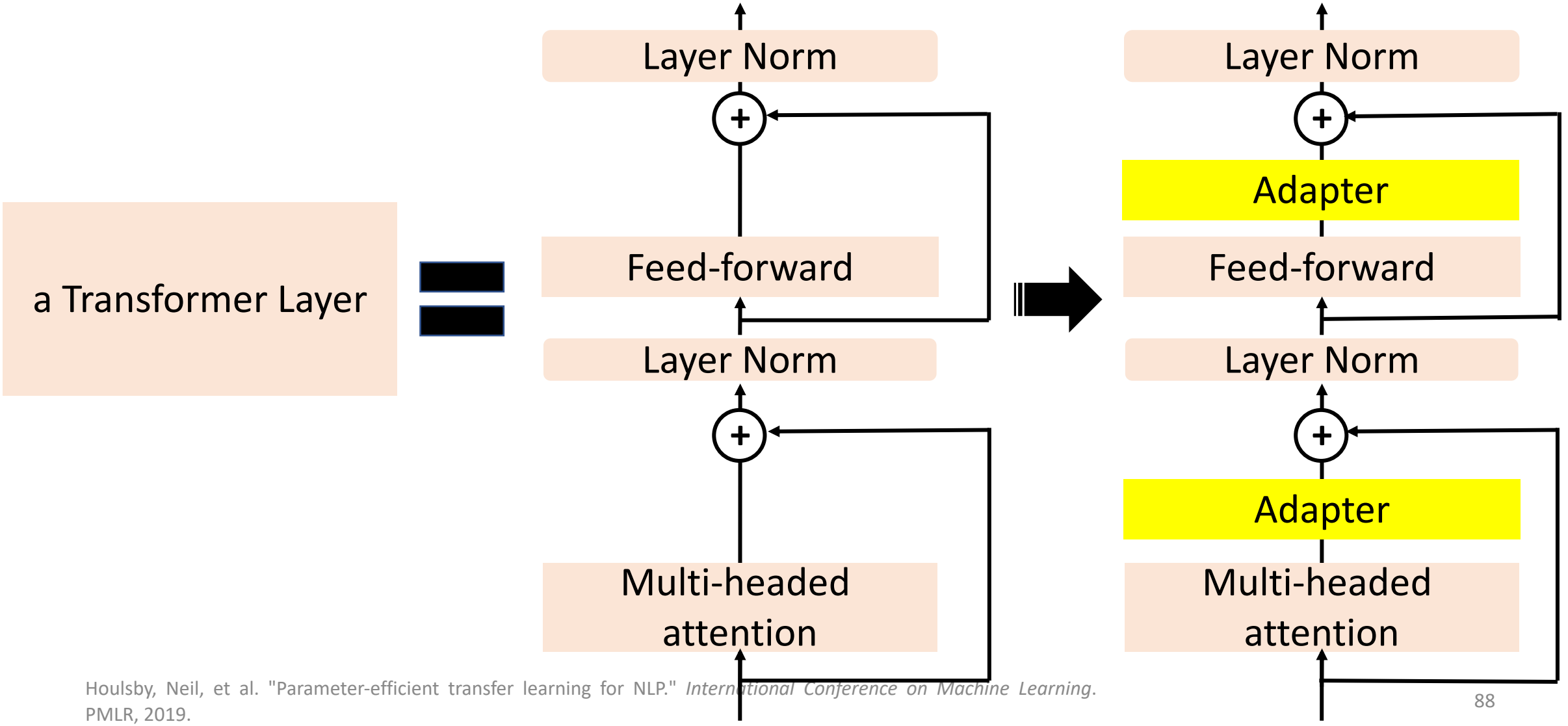
- Use special submodules to modify hidden representations!



He, Junxian, et al. "Towards a unified view of parameter-efficient transfer learning." *arXiv preprint arXiv:2110.04366* (2021).

Parameter-Efficient Fine-tuning: Adapter

- Adapters: small trainable submodules inserted in transformers

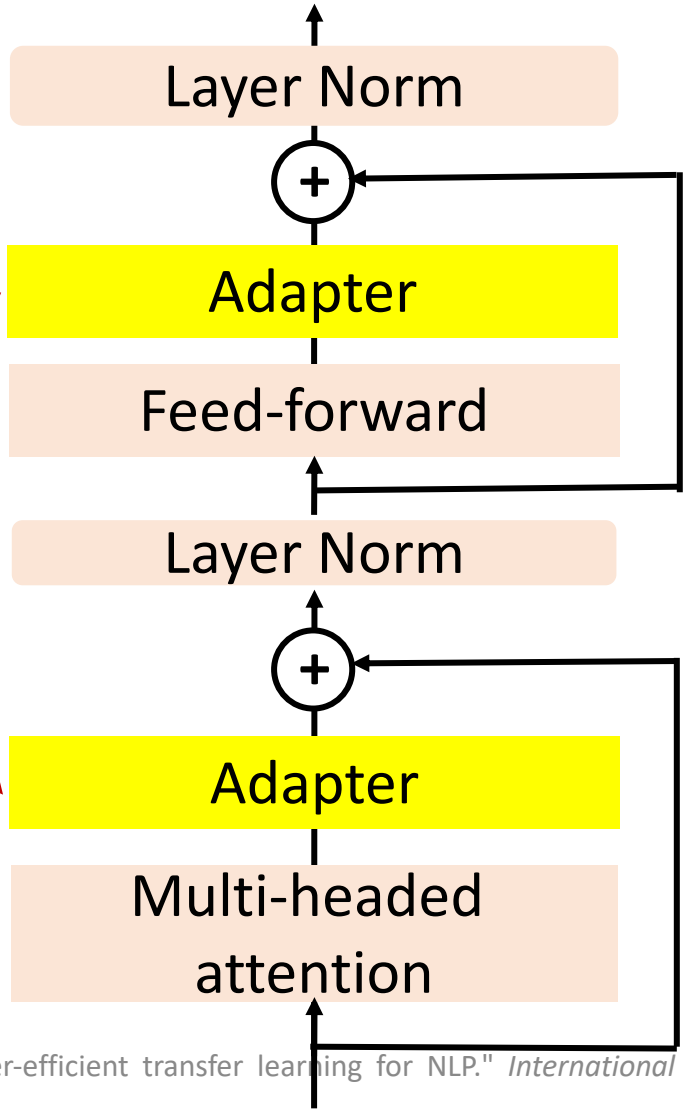


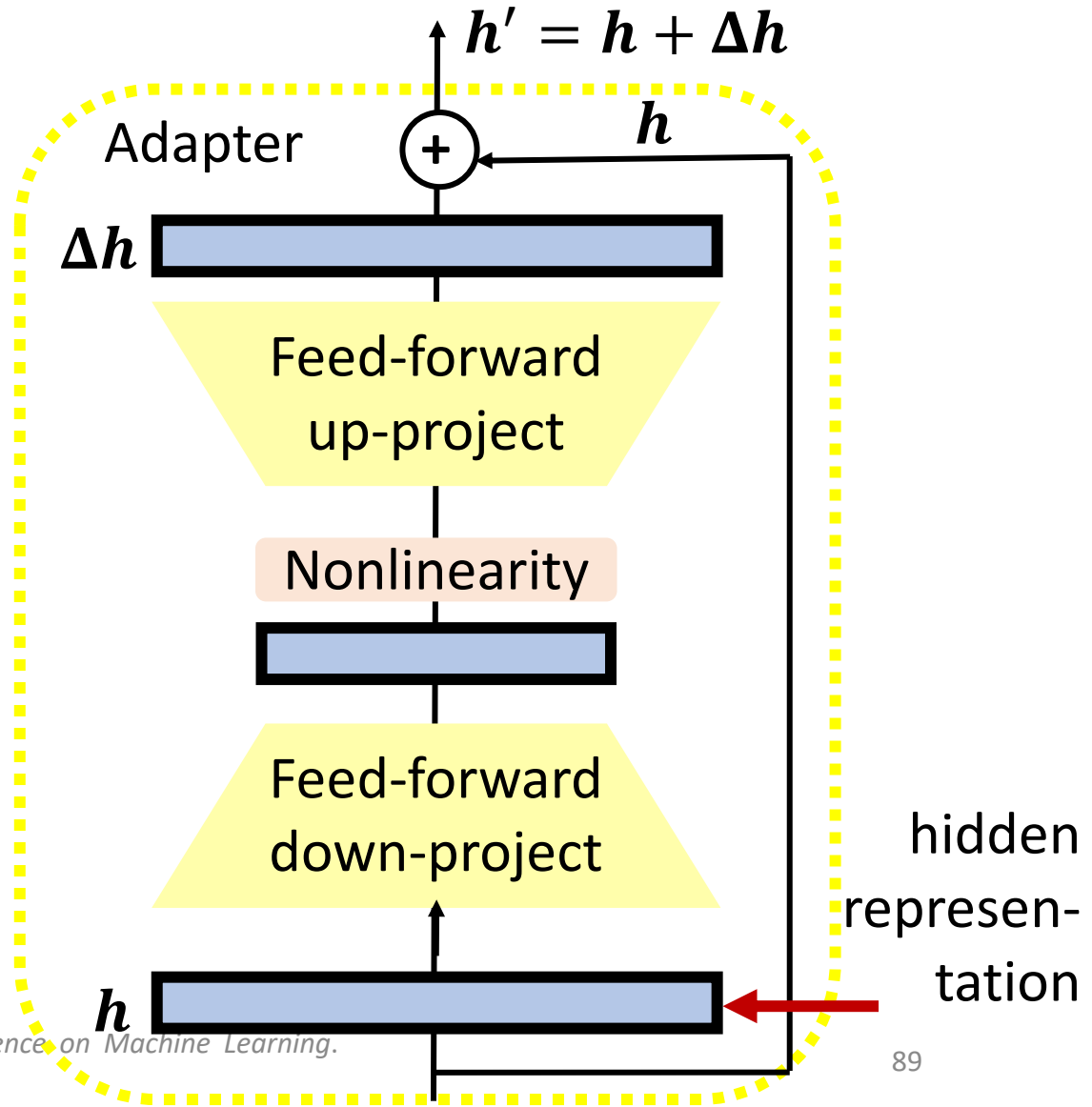
Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International Conference on Machine Learning*. PMLR, 2019.

Parameter-Efficient Fine-tuning: Adapter

- Adapters

Inside of the transformer layer, only adapters are updated

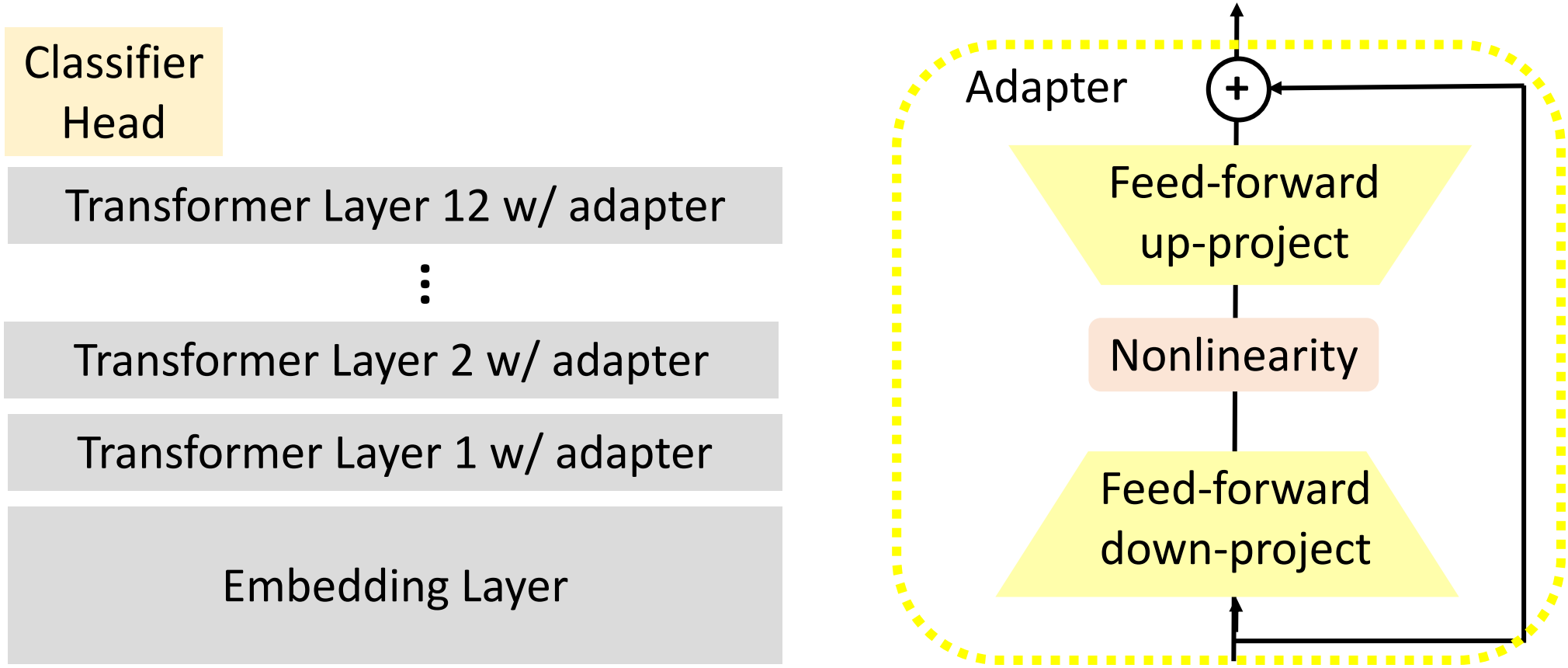




Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International Conference on Machine Learning*. PMLR, 2019.

Parameter-Efficient Fine-tuning: Adapter

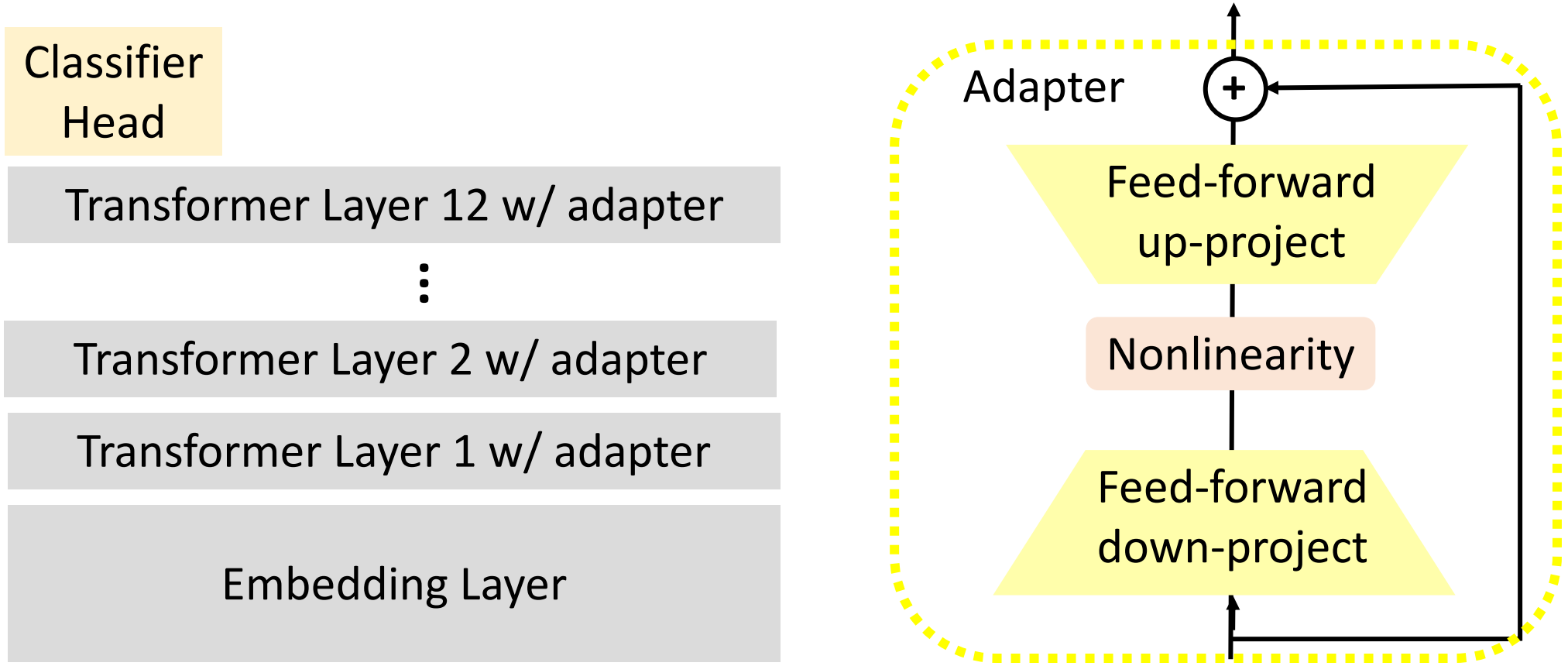
- Adapters: During fine-tuning, only update the adpaters and the classifier head



Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International Conference on Machine Learning*. PMLR, 2019.

Parameter-Efficient Fine-tuning: Adapter

- Adapters: All downstream tasks share the PLM; the adapters in each layer and the classifier heads are the task-specific modules



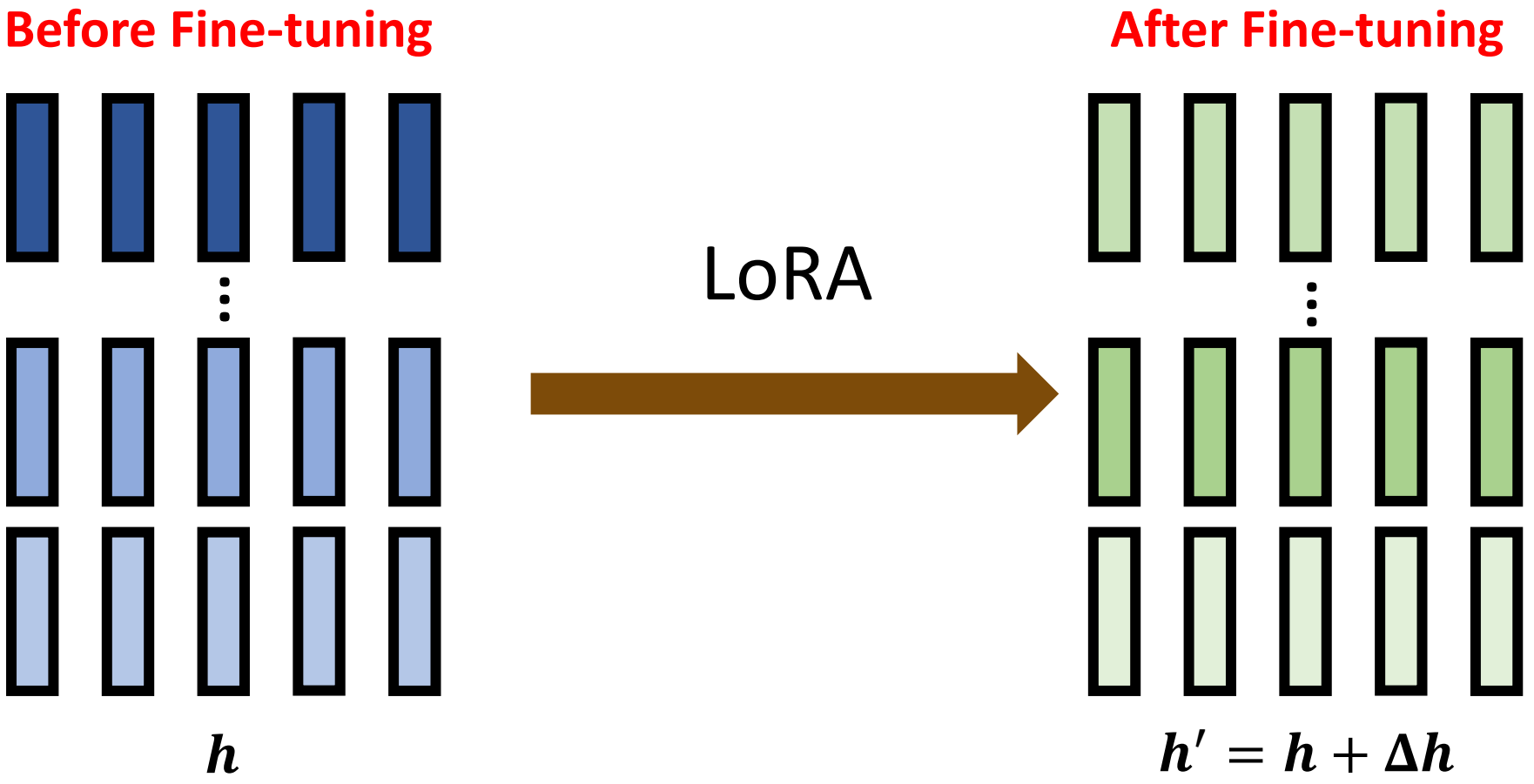
Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International Conference on Machine Learning*. PMLR, 2019.

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Parameter-Efficient Fine-tuning
      - LoRA
- Closing Remarks

Parameter-Efficient Fine-tuning: LoRA

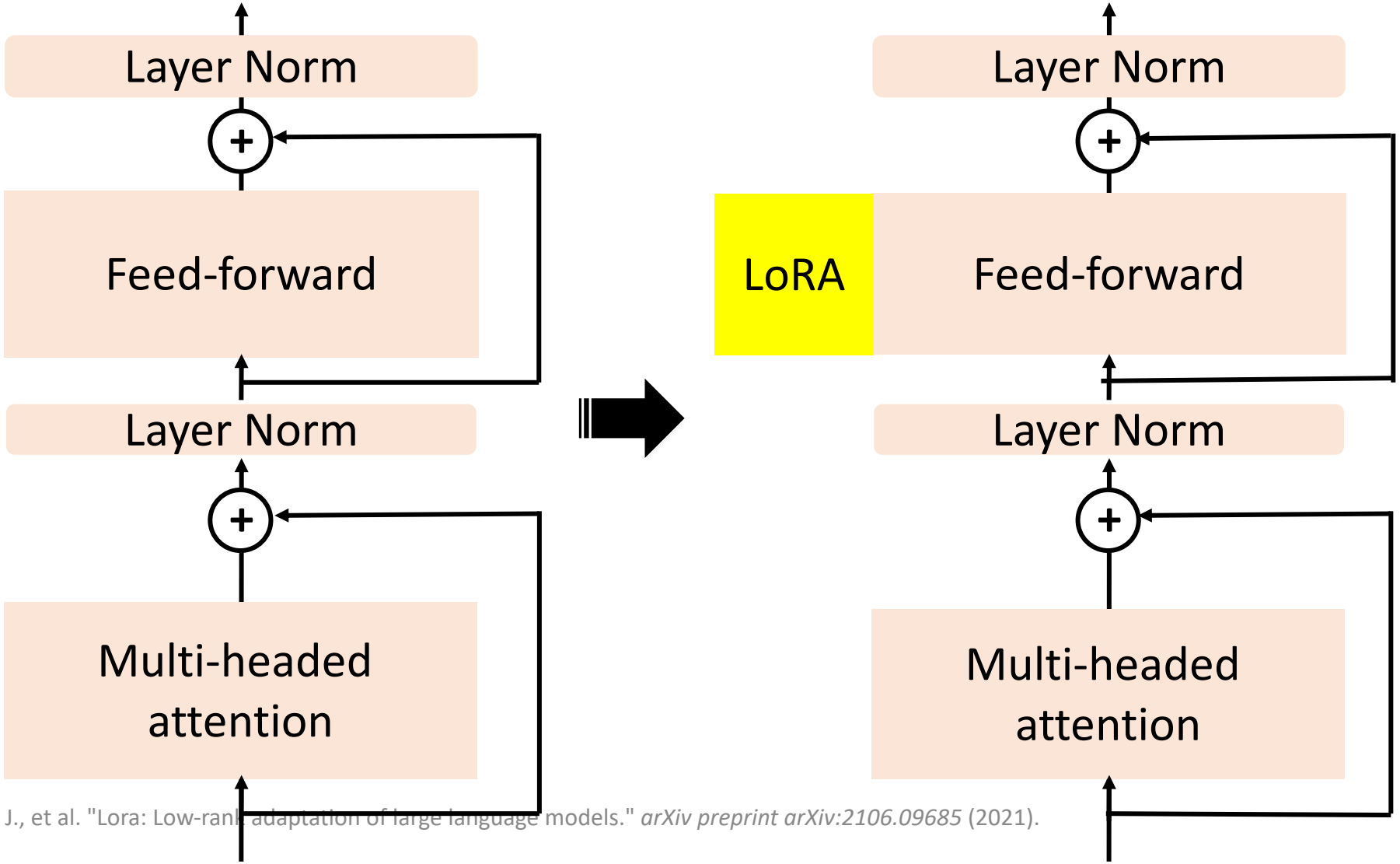
- Use special submodules to modify hidden representations!



He, Junxian, et al. "Towards a unified view of parameter-efficient transfer learning." *arXiv preprint arXiv:2110.04366* (2021).

Parameter-Efficient Fine-tuning: LoRA

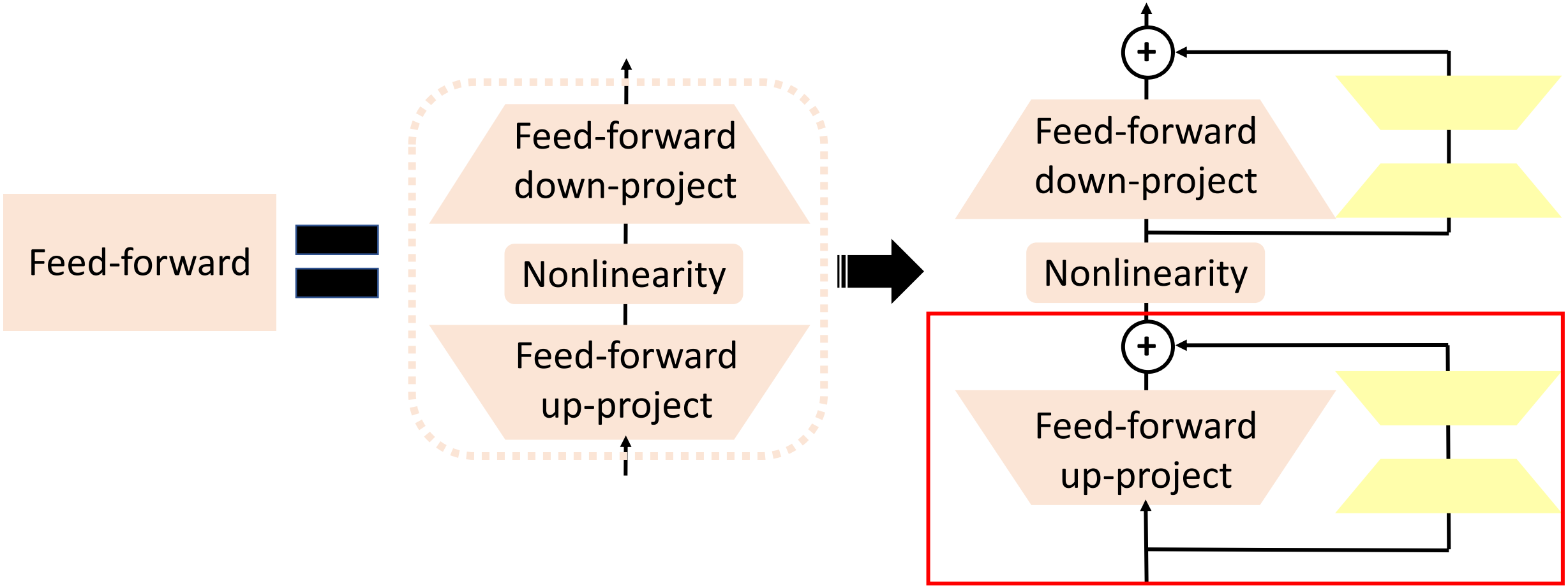
- LoRA: Low-Rank Adaptation of Large Language Models



Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Parameter-Efficient Fine-tuning: LoRA

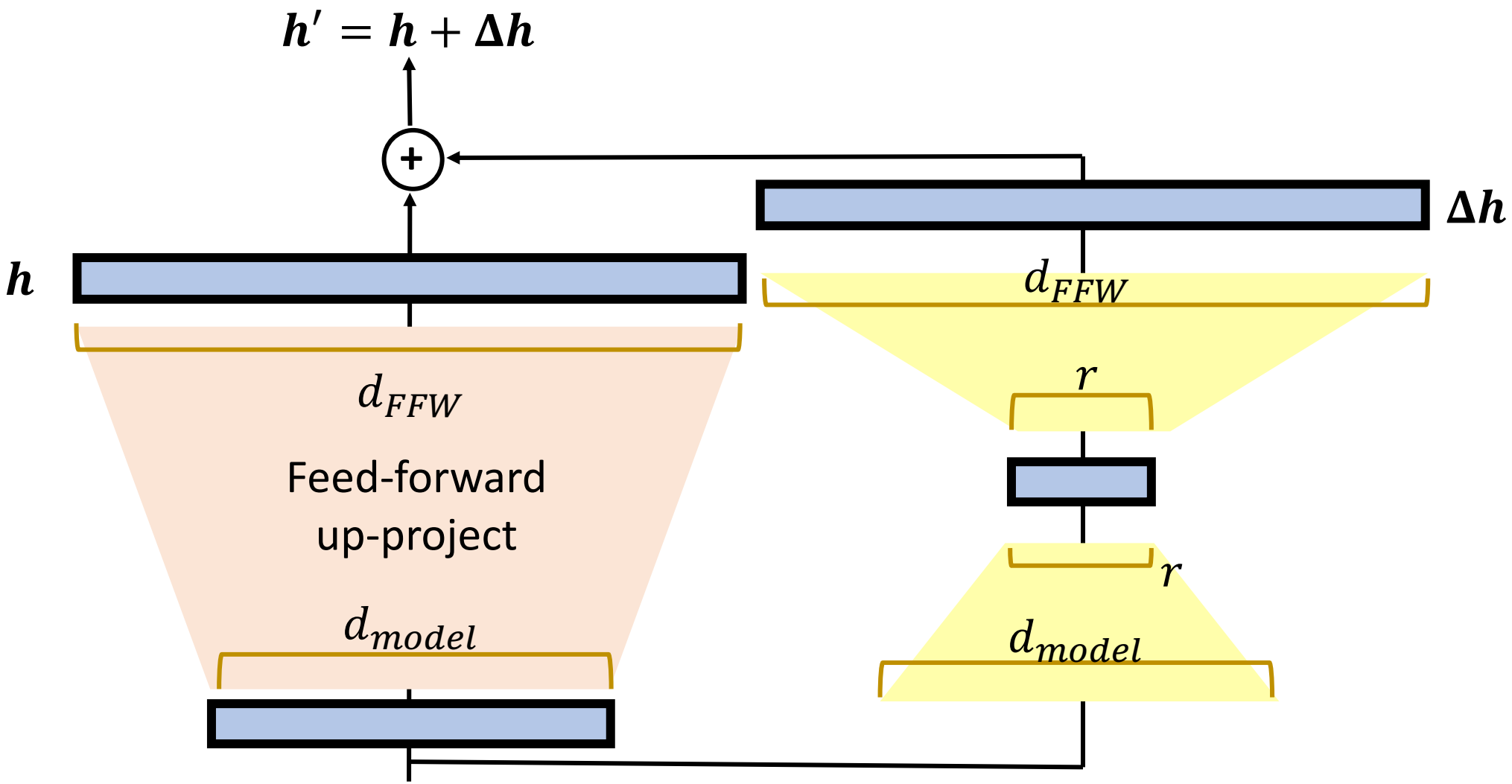
- LoRA



Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Parameter-Efficient Fine-tuning: LoRA

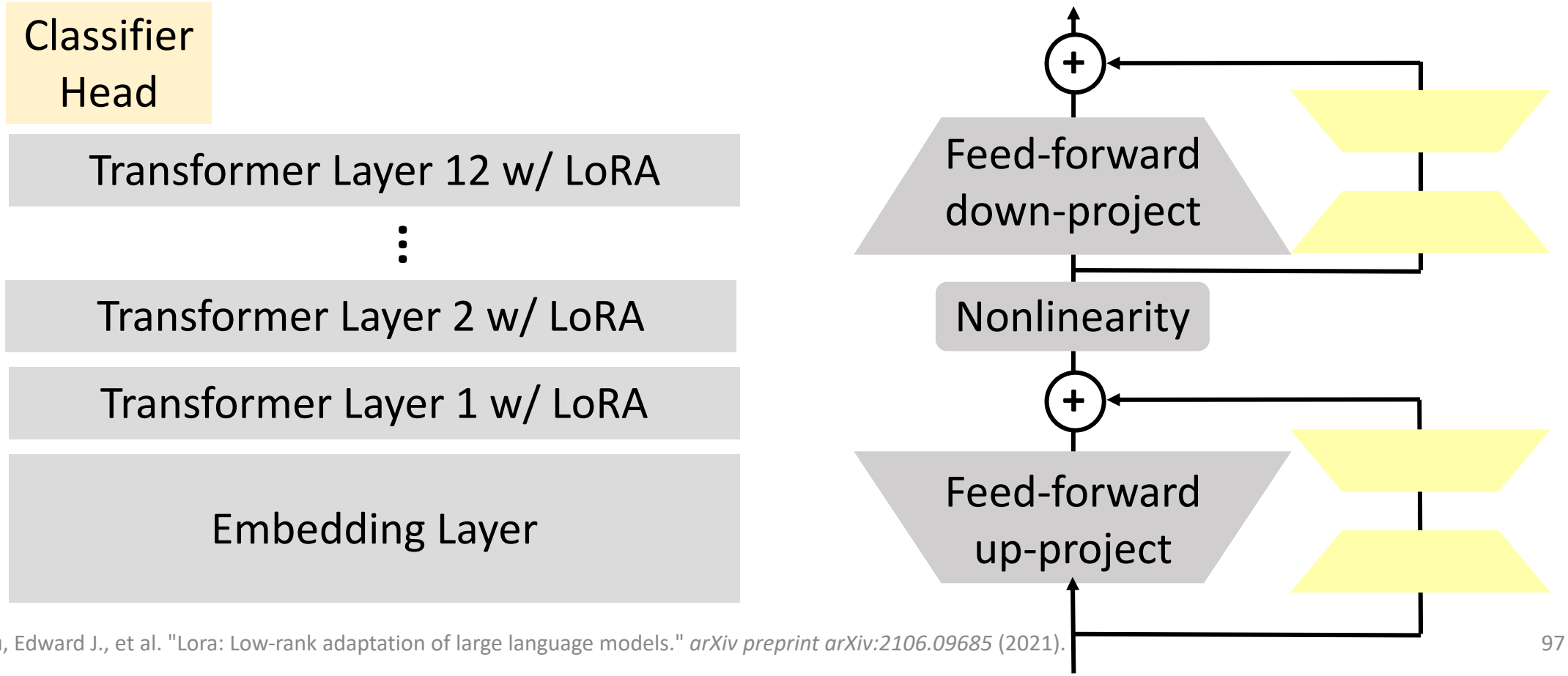
- LoRA



Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Parameter-Efficient Fine-tuning: LoRA

- LoRA: All downstream tasks share the PLM; the LoRA in each layer and the classifier heads are the task-specific modules



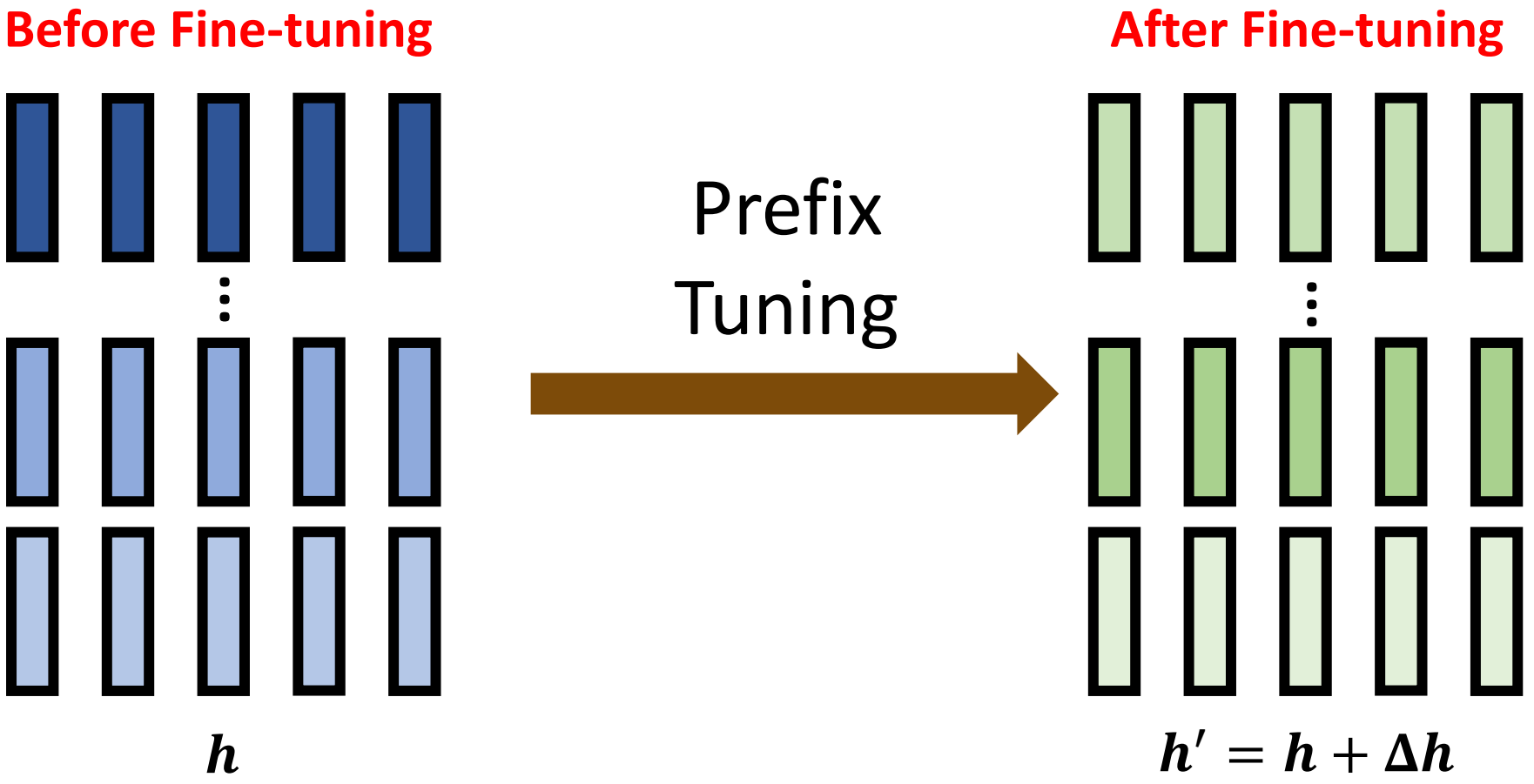
Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Parameter-Efficient Fine-tuning
      - Prefix Tuning
- Closing Remarks

Parameter-Efficient Fine-tuning: Prefix Tuning

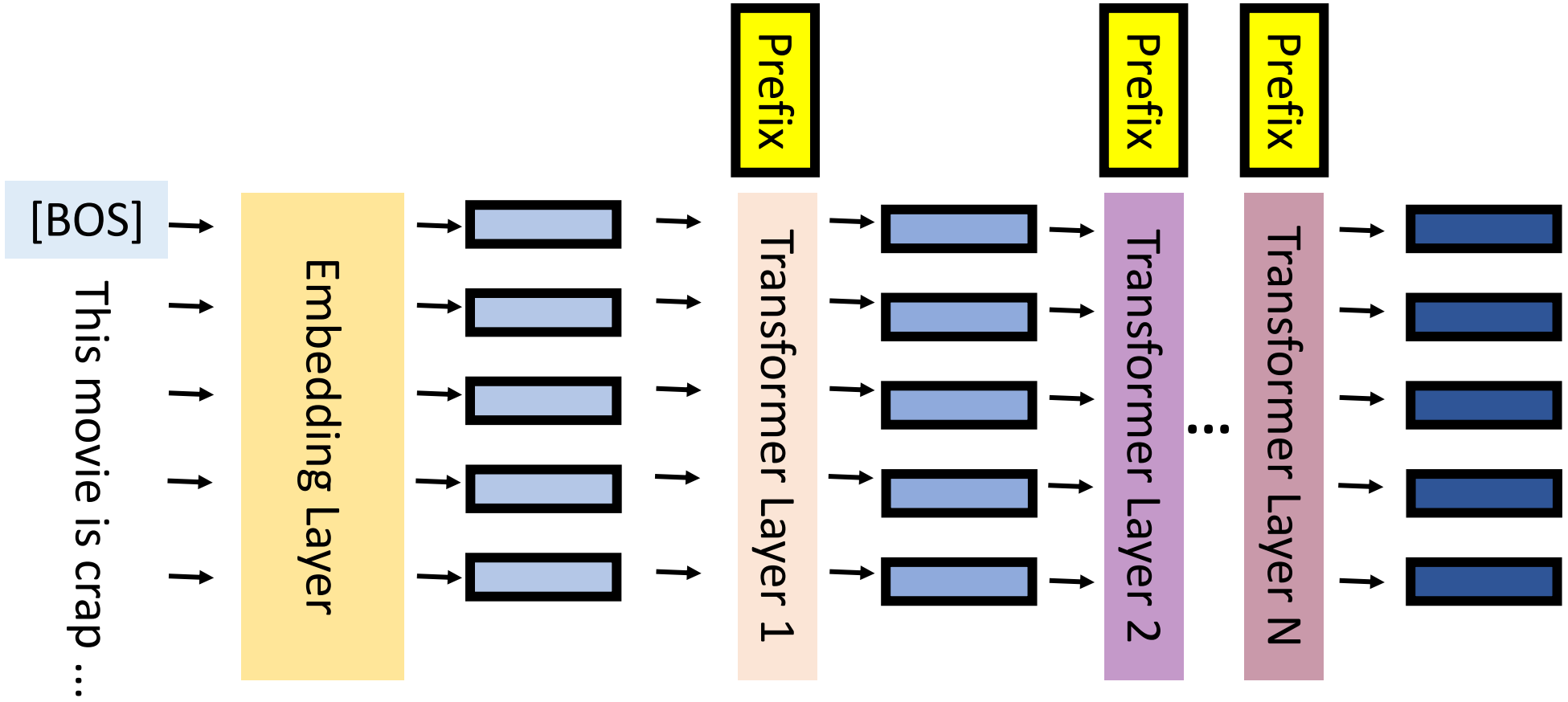
- Use special submodules to modify hidden representations!



He, Junxian, et al. "Towards a unified view of parameter-efficient transfer learning." *arXiv preprint arXiv:2110.04366* (2021).

Parameter-Efficient Fine-tuning: Prefix Tuning

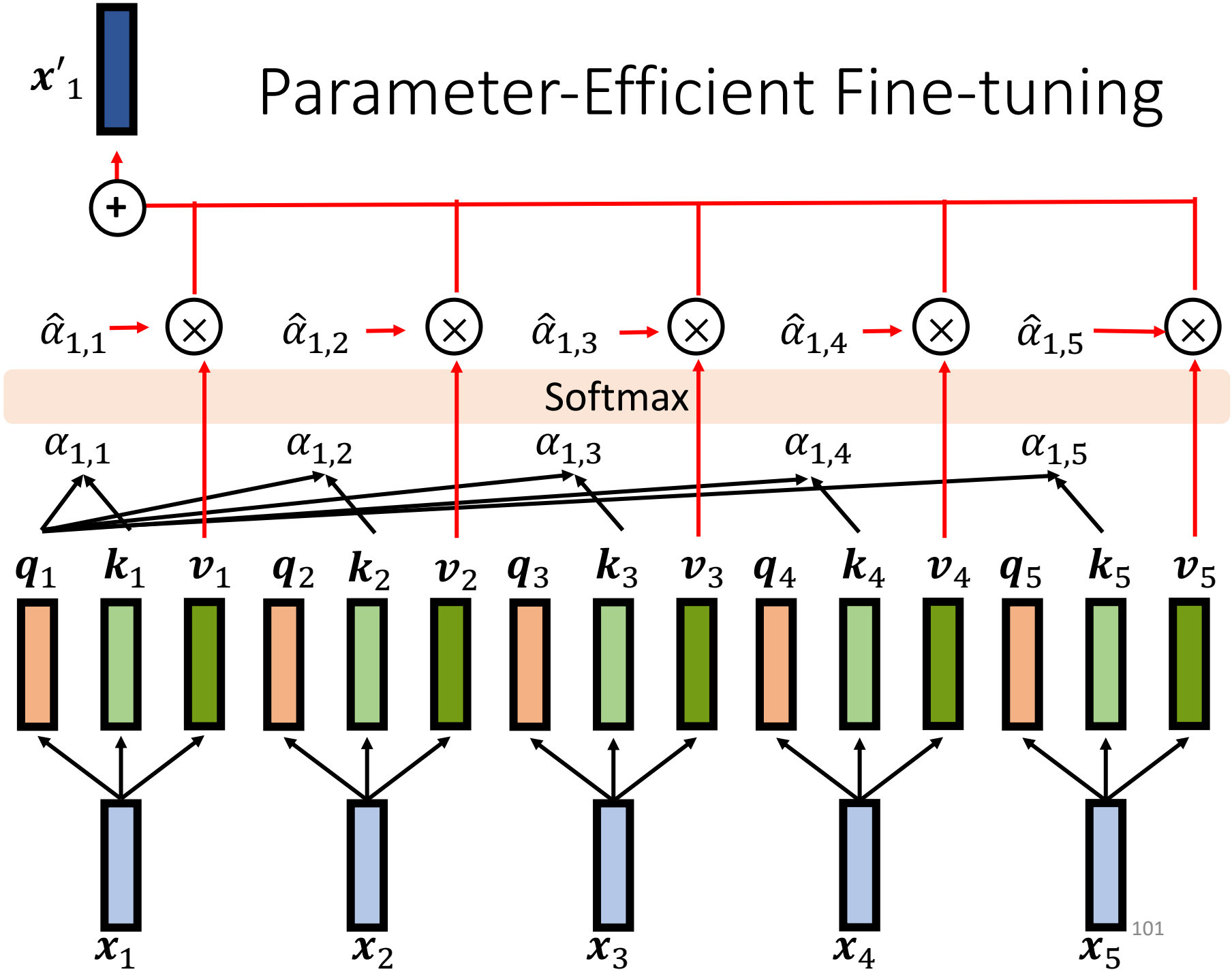
- Prefix Tuning: Insert trainable prefix in each layer



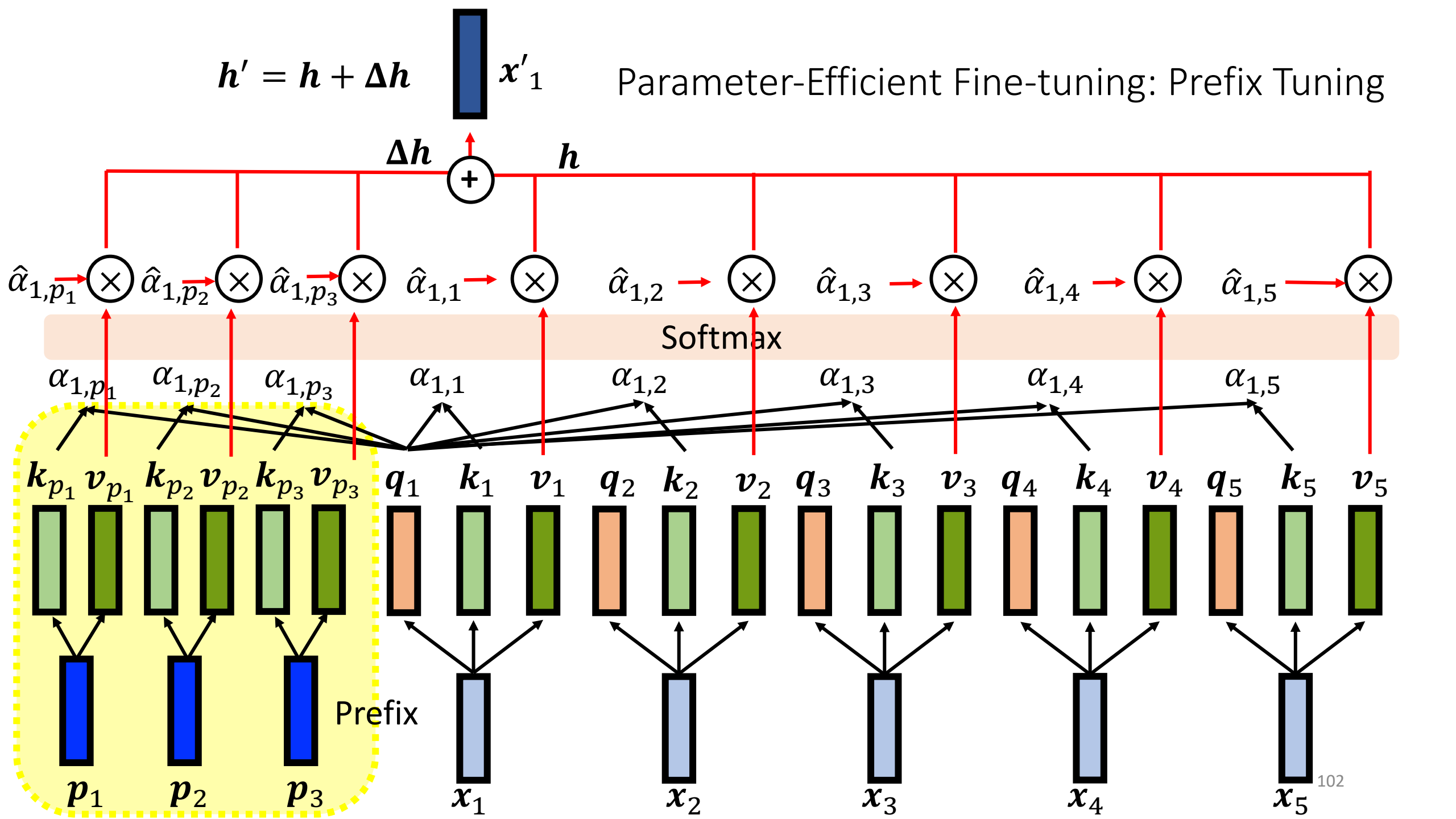
Li, Xiang Lisa, and Percy Liang. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

Parameter-Efficient Fine-tuning

Standard Self-Attention

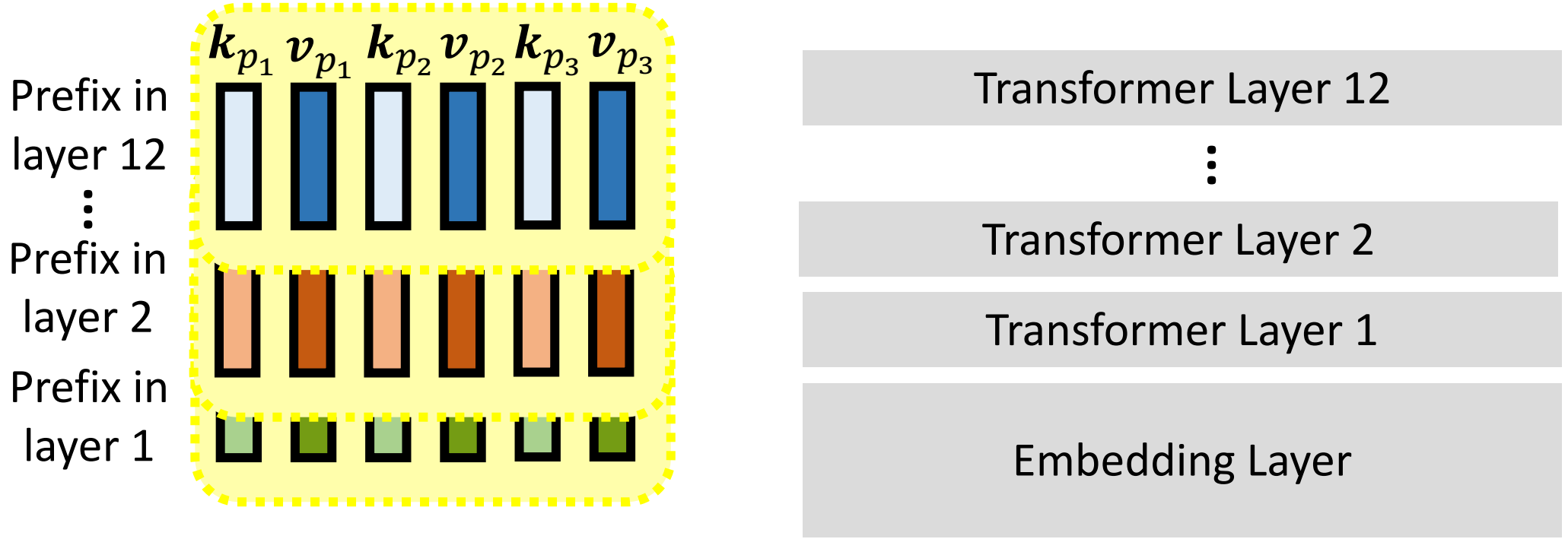


Parameter-Efficient Fine-tuning: Prefix Tuning



Parameter-Efficient Fine-tuning: Prefix Tuning

- Prefix Tuning: Only the prefix (key and value) are updated during fine-tuning



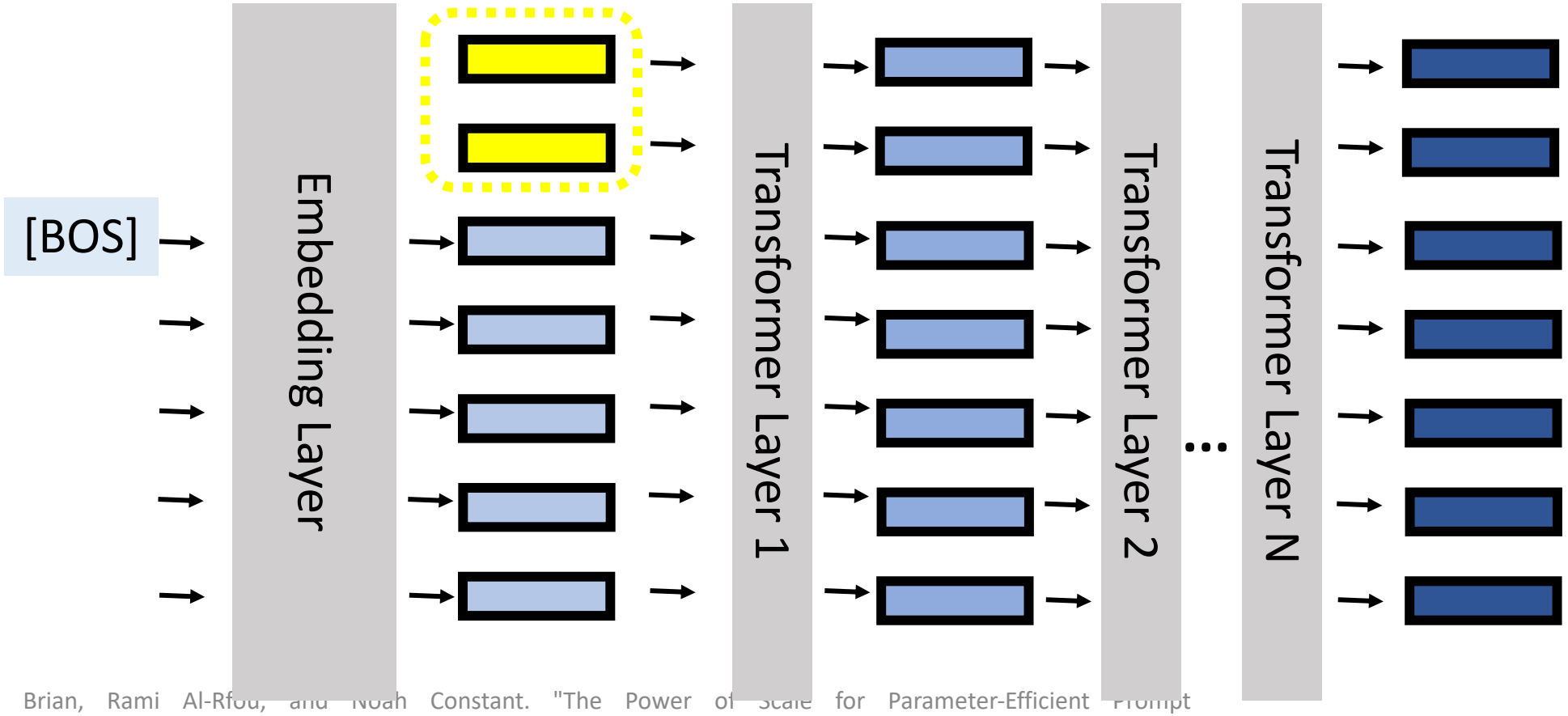
Li, Xiang Lisa, and Percy Liang. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Parameter-Efficient Fine-tuning
      - Soft Prompting
- Closing Remarks

Parameter-Efficient Fine-tuning: Soft Prompting

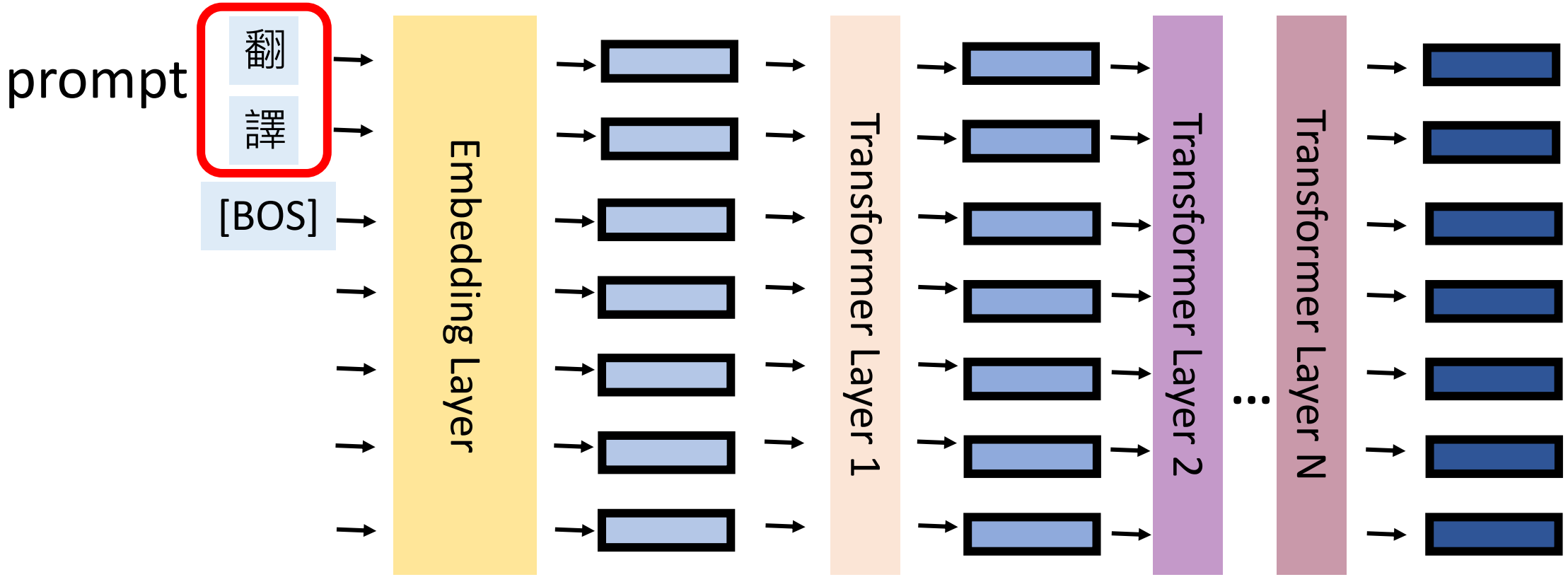
- Soft Prompting
  - Prepend the prefix embedding at the input layer



Lester, Brian, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Parameter-Efficient Fine-tuning: Soft Prompting

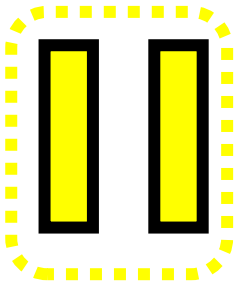
- Soft Prompting can be considered as the ***soften*** version of prompting
  - (Hard) prompting: add words in the input sentence (fine-tune the model while fixing the prompts)



Lester, Brian, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Parameter-Efficient Fine-tuning

- Soft Prompts: vectors (can be initialized from some word embeddings)



- Hard Prompts: words (that are originally in the vocabulary)

prompt

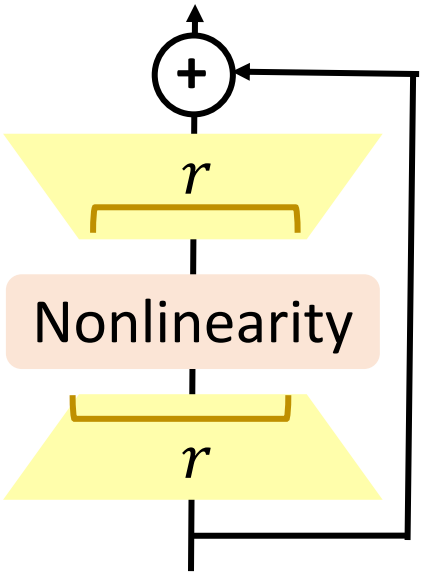
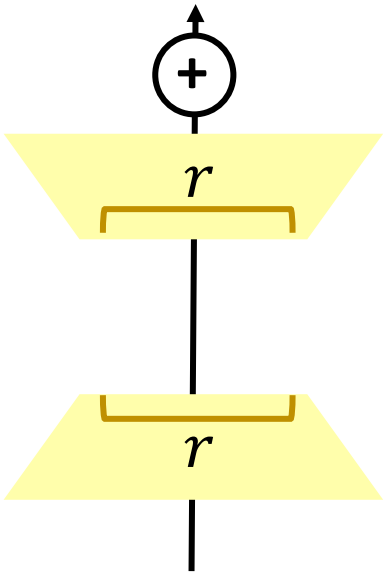
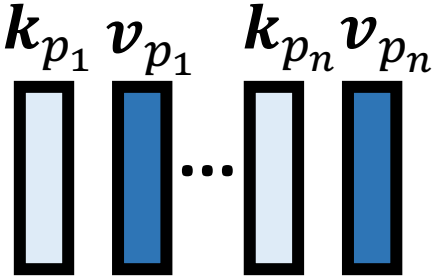
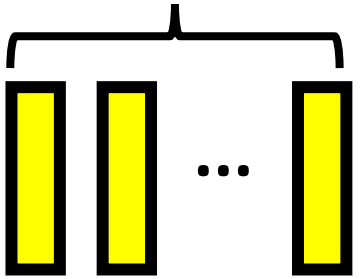
翻

譯

Lester, Brian, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Parameter-Efficient Fine-tuning

- Benefit 1: Drastically decreases the task-specific parameters

	Adapter	LoRA	Prefix Tuning	Soft Prompt
Task-specific parameters*	$\Theta(d_{model}rL)$	$\Theta(d_{model}rL)$	$\Theta(d_{model}nL)$	$\Theta(d_{model}n)$
Percent Trainable	<5%	<0.1%	<0.1%	<0.05%
Illustration			n:Prefix length 	n:Prefix length 

*not including the classifier head

Parameter-Efficient Fine-tuning

- Benefit 2: Less easier to overfit on training data; better out-of-domain performance

In-domain dataset

OOD test dataset

Dataset	Domain	Model	Soft Prompt	Δ
SQuAD	Wiki	94.9 ±0.2	94.8 ±0.1	−0.1
TextbookQA	Book	54.3 ±3.7	**66.8** ±2.9	+12.5
BioASQ	Bio	77.9 ±0.4	**79.1** ±0.3	+1.2
RACE	Exam	59.8 ±0.6	**60.7** ±0.5	+0.9
RE	Wiki	88.4 ±0.1	**88.8** ±0.2	+0.4
DuoRC	Movie	**68.9** ±0.7	67.7 ±1.1	−1.2
DROP	Wiki	**68.9** ±1.7	67.1 ±1.9	−1.8

Parameter-Efficient Fine-tuning

- Benefit 3: Fewer parameters to fine-tune; a good candidate when training with small dataset

Model	low-resource				high-resource			
	CHEMPROT (4169)	ACL-ARC (1688)	SCIERC (3219)	HYP. (515)	RCT (180k)	AGNEWS (115k)	HELPFUL. (115k)	IMDB (20k)
RoBa.-ft†	$81.9_{1.0}$	$63.0_{5.8}$	$77.3_{1.9}$	$86.6_{0.9}$	$\mathbf{87.2}_{0.1}$	$\mathbf{93.9}_{0.2}$	$65.1_{3.4}$	$95.0_{0.2}$
RoBa.-ft*	$81.7_{0.8}$	$65.0_{3.6}$	$78.5_{1.8}$	$88.9_{3.3}$	$87.0_{0.1}$	$93.7_{0.2}$	$\mathbf{69.1}_{0.6}$	$95.2_{0.1}$
RoBa.-adapter$_{256}$	$\mathbf{82.9}_{0.6}$	$\mathbf{67.5}_{4.3}$	$\mathbf{80.8}_{0.7}$	$\mathbf{90.4}_{4.2}$	$87.1_{0.1}$	$93.8_{0.1}$	$69.0_{0.4}$	$\mathbf{95.7}_{0.1}$

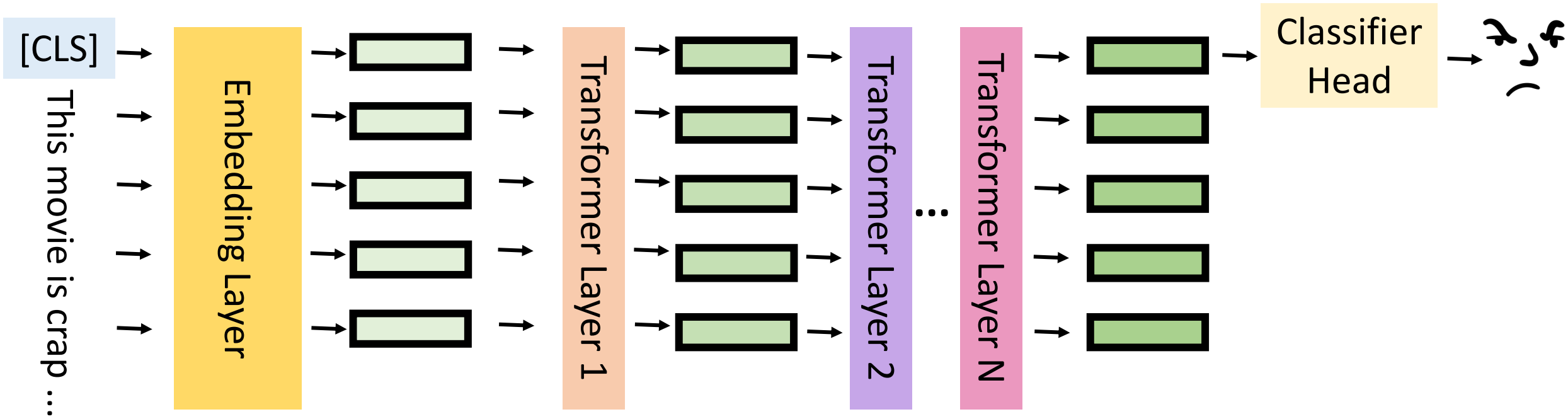
He, Ruidan, et al. "On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Early Exit
- Closing Remarks

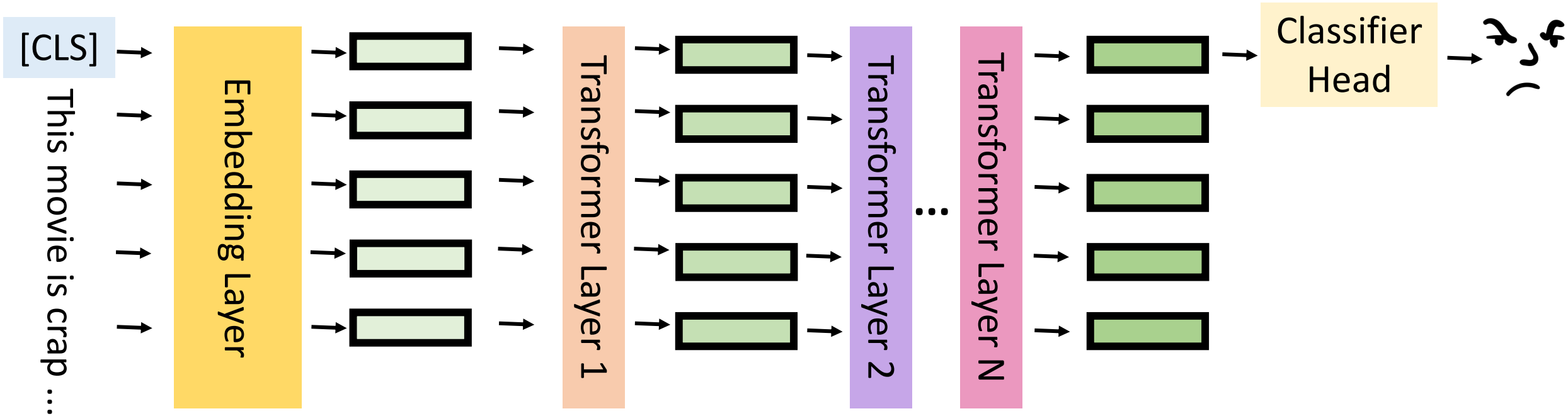
Early Exit

- Problem 1: The PLM is too big
  - Inference takes too long



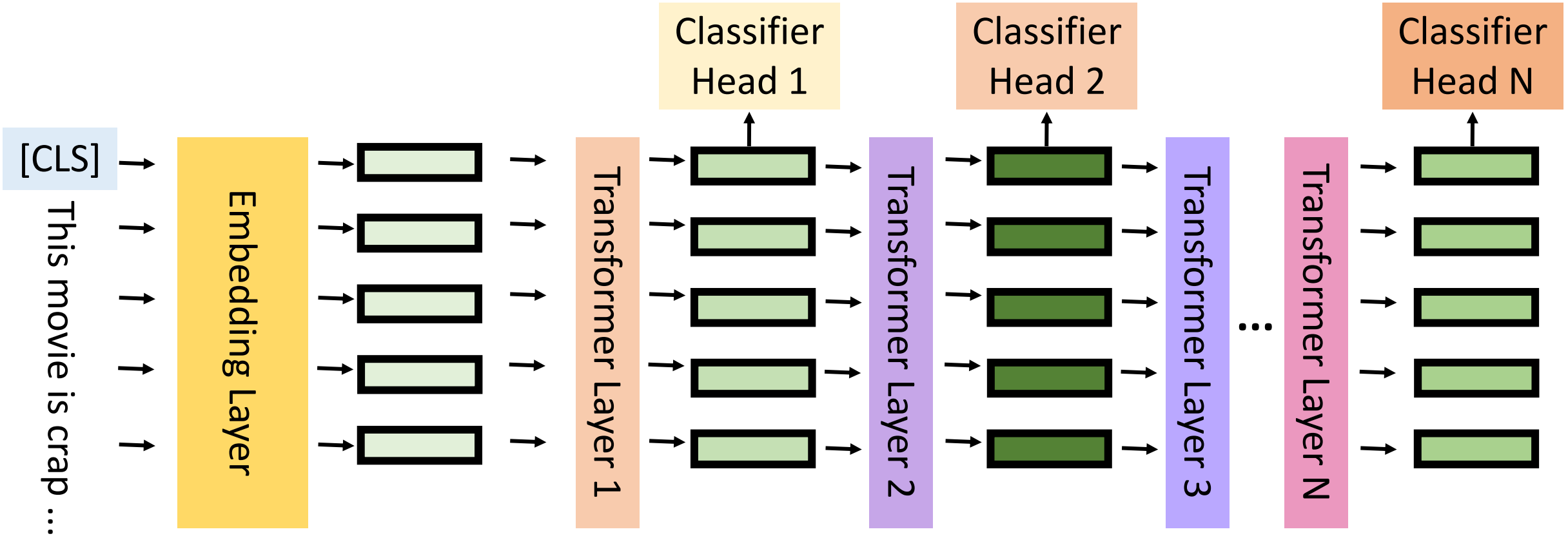
Early Exit

- Inference using the whole model takes too long
- Simpler data may require lesser effort to obtain the answer
- Reduce the number of layers used during inference



Early Exit

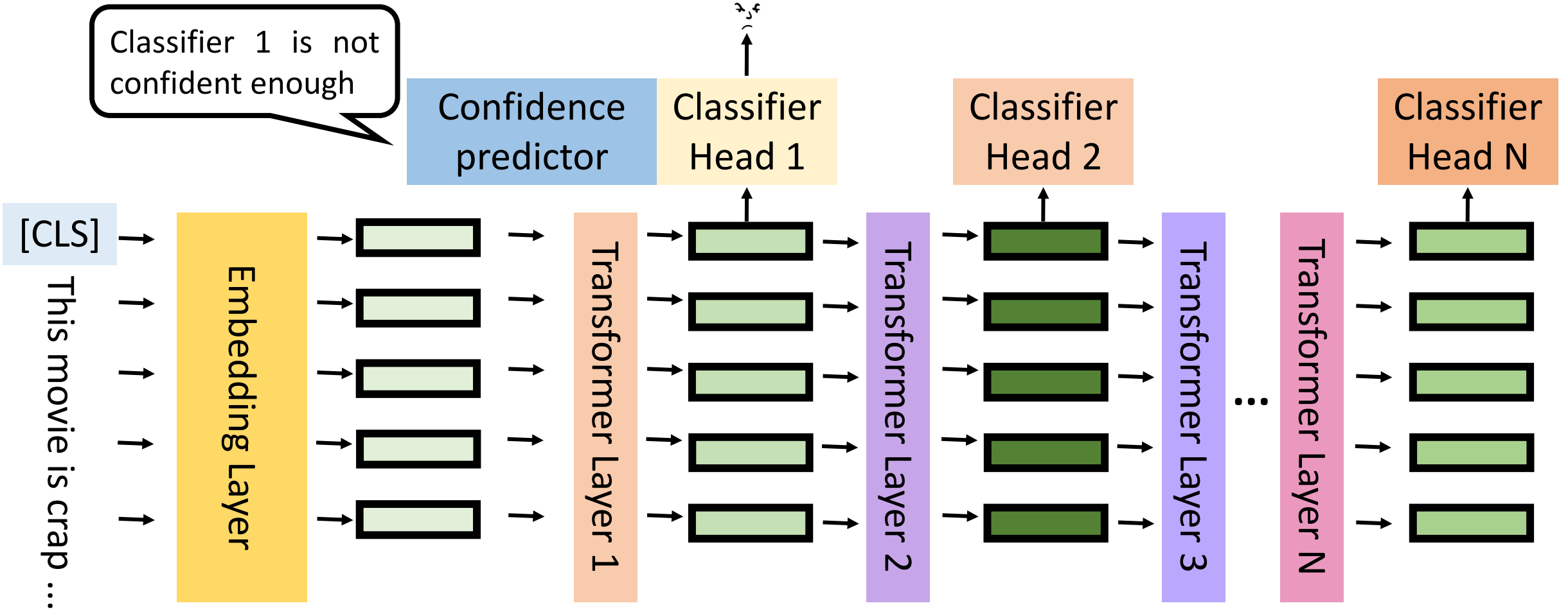
- Add a classifier at each layer



Xin, Ji, et al. "BERxiT: Early Exiting for BERT with Better fine-tuning and extension to regression." *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume*. 2021

Early Exit

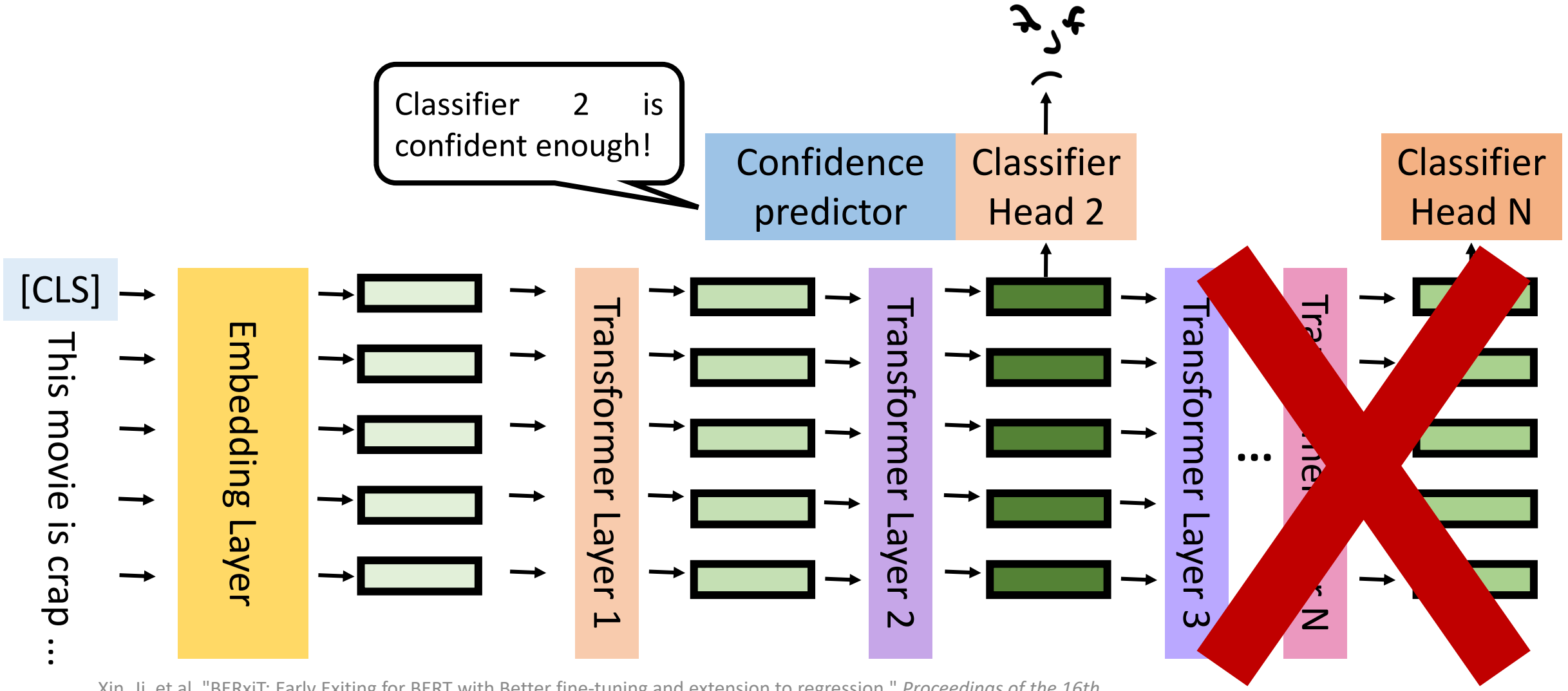
- How do we know which classifier to use?



Xin, Ji, et al. "BERxiT: Early Exiting for BERT with Better fine-tuning and extension to regression." *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume*. 2021

Early Exit

- How do we know which classifier to use?



Xin, Ji, et al. "BERxiT: Early Exiting for BERT with Better fine-tuning and extension to regression." *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume*. 2021

Early Exit

- Early exit reduces the inference time while keeping the performance

	RTE		MRPC		SST-2		QNLI		QQP		MNLI-(m/mm)		STS-B	
	Score	Layer	Score	Layer	Score	Layer	Score	Layer	Score	Layer	Score	Layer	Score	Layer
							BERT$_{\text{BASE}}$							
RAW	66.4	12	88.9	12	93.5	12	90.5	12	71.2	12	84.6/83.4	12	85.8	12
ALT	101%	−44%	99%	−30%	98%	−65%	99%	−42%	99%	−56%	99%/99%	−37%	95%	−50%
	99%	−54%	97%	−56%	96%	−79%	98%	−63%	97%	−75%	97%/97%	−57%	91%	−67%
	96%	−64%	94%	−74%	94%	−87%	95%	−71%	93%	−84%	93%/92%	−72%	85%	−75%
							BERT$_{\text{LARGE}}$							
RAW	70.1	24	89.3	24	94.9	24	92.7	24	72.1	24	86.7/85.9	24	86.5	24
ALT	95%	−33%	99%	−32%	100%	−32%	97%	−62%	98%	−74%	99%/99%	−36%	97%	−39%
	94%	−46%	98%	−46%	99%	−61%	95%	−73%	96%	−82%	96%/97%	−57%	90%	−62%
	88%	−62%	94%	−71%	96%	−78%	91%	−83%	91%	−89%	90%/90%	−75%	76%	−80%

Xin, Ji, et al. "BERxiT: Early Exiting for BERT with Better fine-tuning and extension to regression." *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume*. 2021

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
  - Labeled Data Scarcity → Data-Efficient Fine-tuning
  - PLMs Are Gigantic → Reducing the Number of Parameters
    - Summary
- Closing Remarks

Reducing the Number of Parameters: Summary

- Parameter-efficient fine-tuning: Reduce the task-specific parameters in downstream task
- Early exit: Reduce the models that are involved during inference

Outline

- Background knowledge
- The Problems of PLMs
- The Solutions of Those Problems
- Closing Remarks

Closing Remarks

- What we address in this lecture
  - Making PLM smaller, faster, and more parameter-efficient
  - Deploying PLMs when the labeled data in the downstream task is scarce
- The problems are not completely solved yet
- The problems we discuss are just a small part of problems of PLMs
  - Why does self-supervised pre-training work
  - Interpretability of the model's prediction
  - Domain adaptation
  - Continual learning/lifelong learning
  - Security and privacy

To Learn More

- AACL-IJCNLP 2022 Tutorial (11.24.2022)
  - **Recent Advances in Pre-trained Language Models: Why Do They Work and How Do They Work.**

Cheng-Han Chiang
National Taiwan University
dcml0714@gmail.com

Yung-Sung Chuang
CSAIL, MIT
yungsung@mit.edu

Hung-yi Lee
National Taiwan University
hungyilee@ntu.edu.tw

https://www.aacl2022.org/home