Efficient Fine-Tuning

Shih Heng Wang, Min Han Shih 2023.10.16

Outline

- Efficient Fine-Tuning
- Types of Efficient Fine-Tuning
- HW-Prompt
- HW-Adapter
- Requirements
- Source

Efficient Fine-Tuning

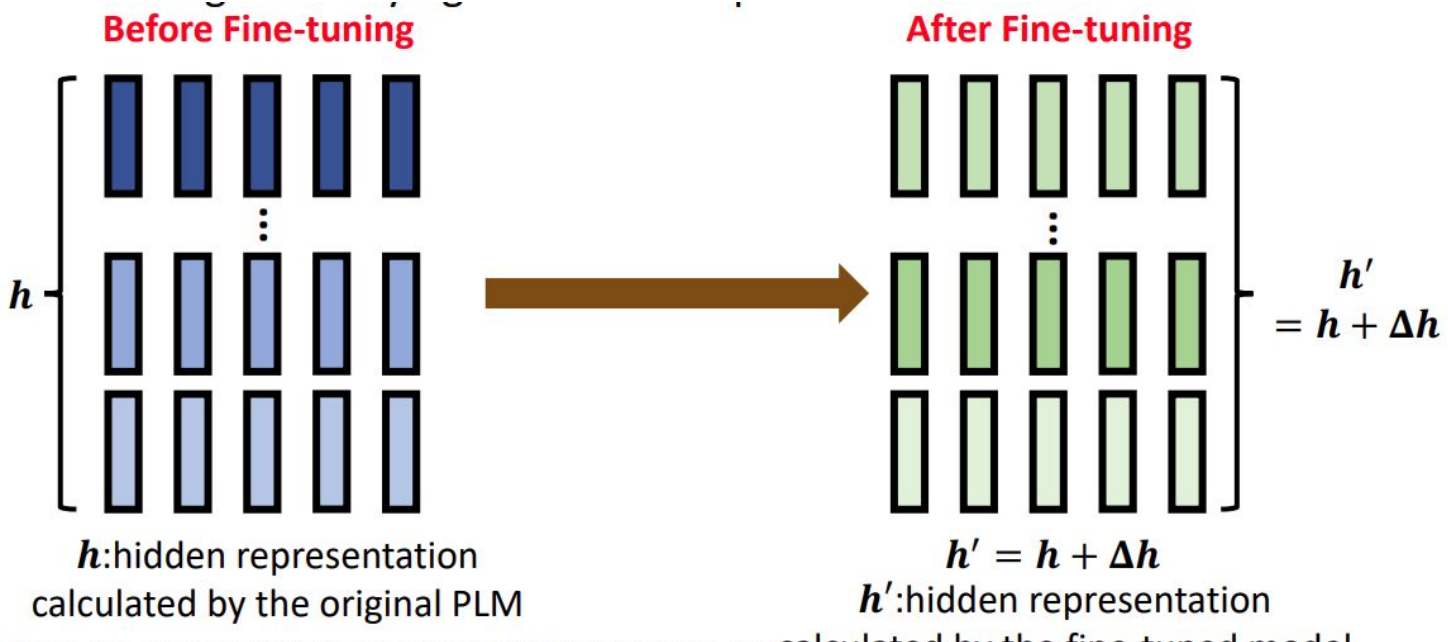
- Fine-tuning the whole PLM takes huge computation resource
- Ex GPT-3 has **175B parameters** **(float32 / float 16 / int 8 :4 / 2 /1 GB)**
- We want to fine-tune the model with less resources

Efficient Fine-Tuning

What Efficient Fine-Tuning is doing ?

Modified the hidden representation

reference: AACL Tutorial



He, Junxian, et al. "Towards a Unified View of Parameter-Efficient Transfer Learning." *International Conference on*

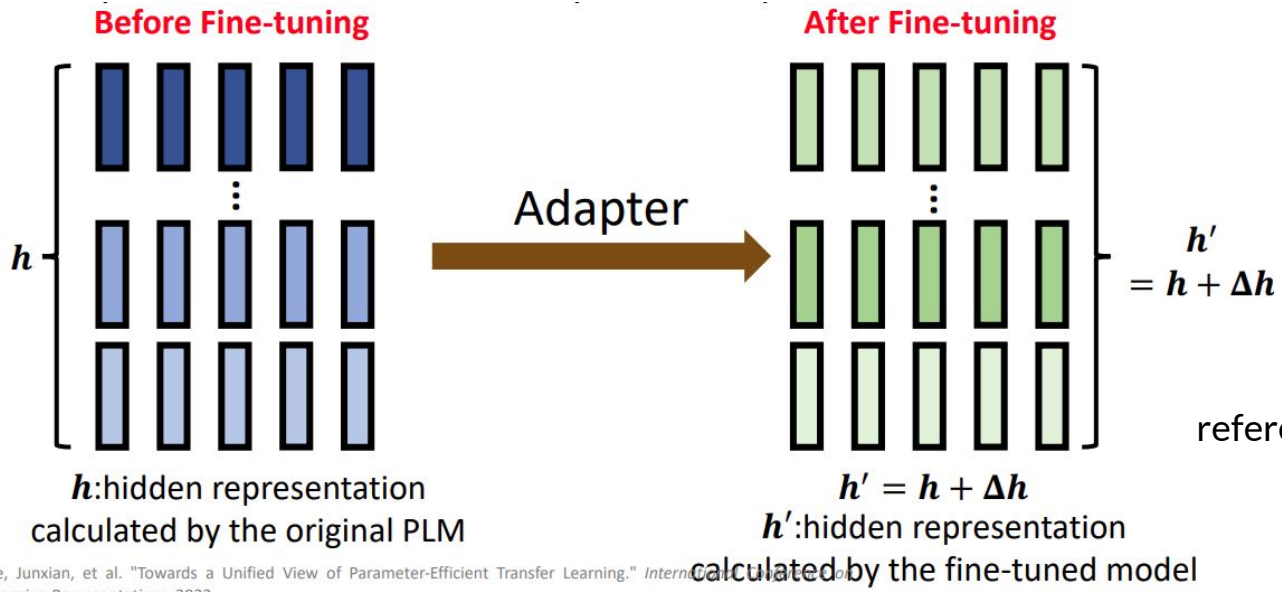
Types of Efficient Fine-Tuning

Types of Efficient Fine-Tuning

- Adapter (Houlsby, Adapter Bias, BitFit)
- Prompt (Prefix-tuning, Prompt-Tuning)

Adapter

- Use special submodules to modify hidden representations!



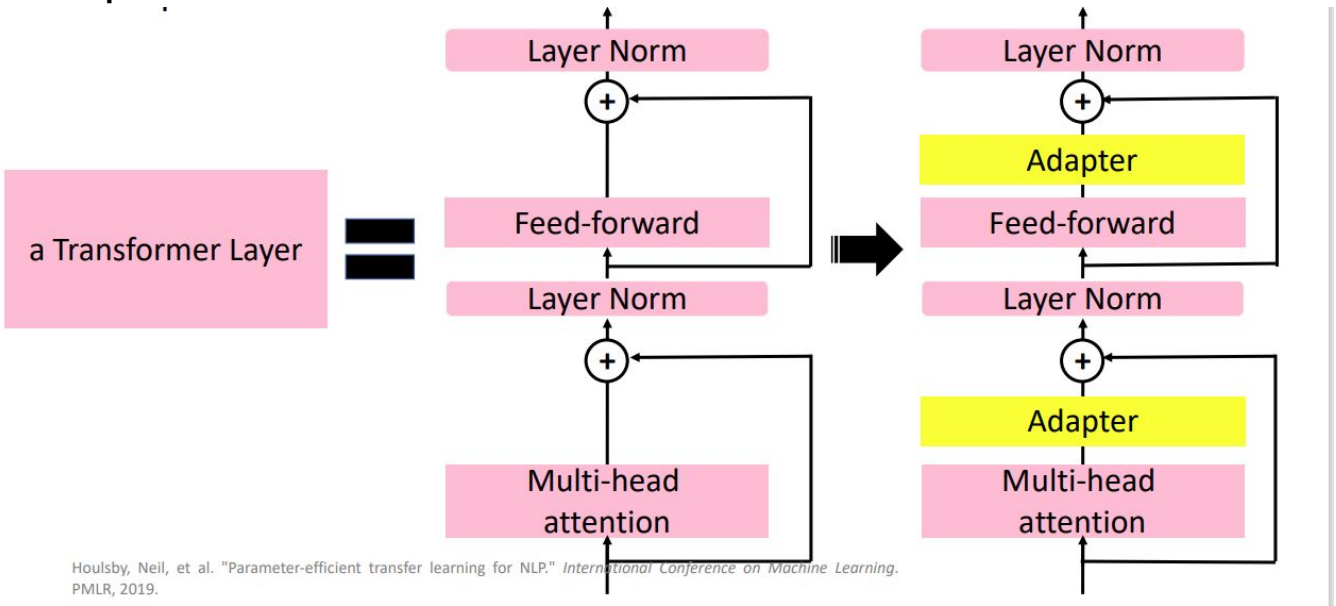
reference: AACL Tutorial

He, Junxian, et al. "Towards a Unified View of Parameter-Efficient Transfer Learning." *International Conference on Learning Representations*. 2022.

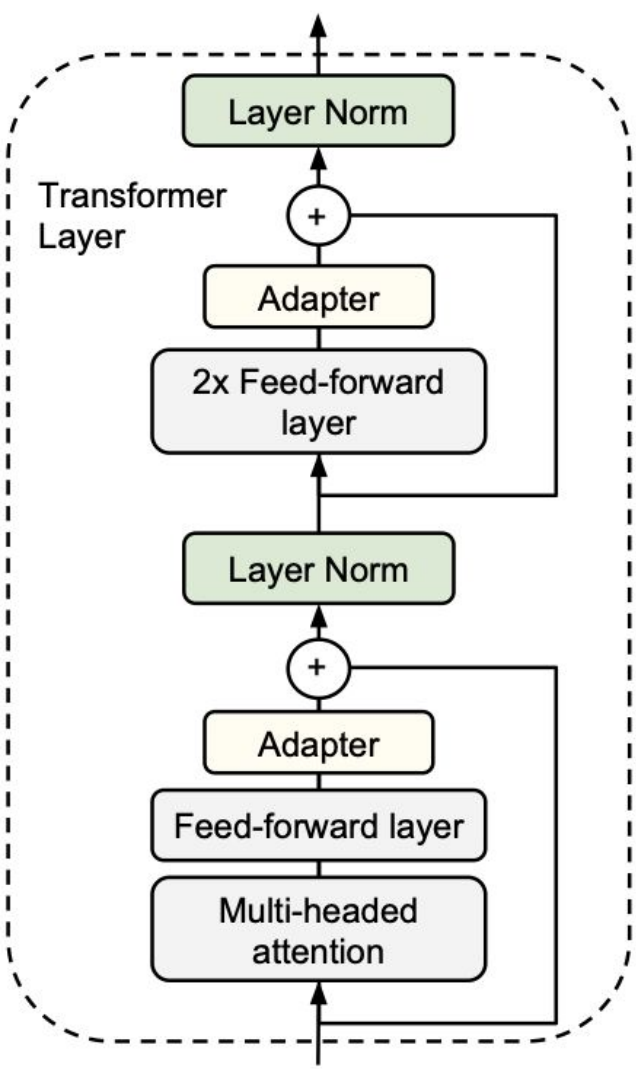
Adapter Module

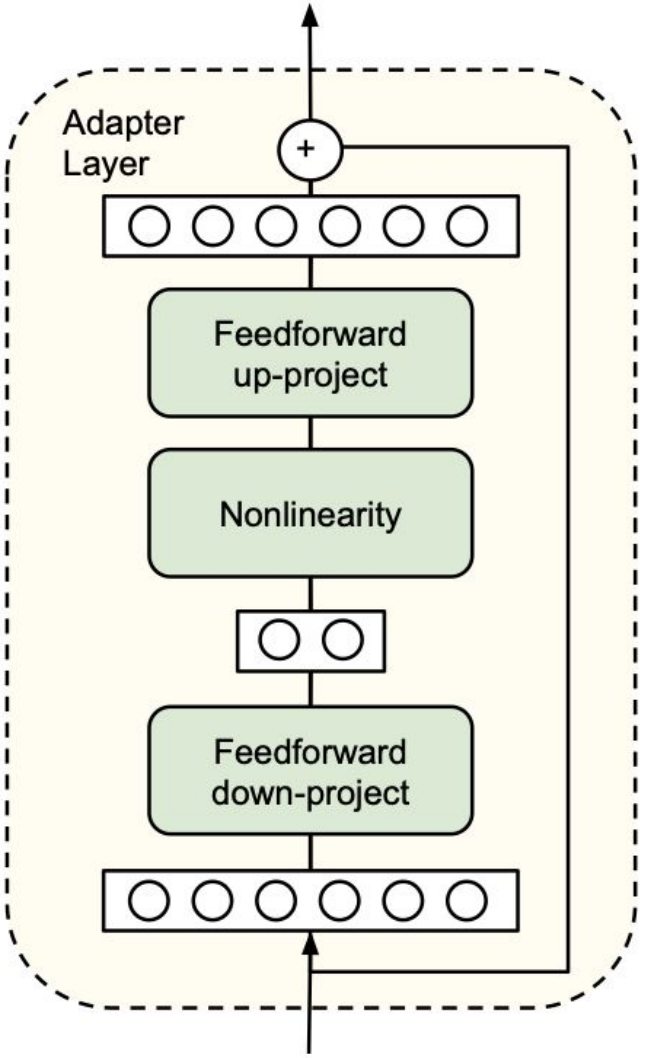
reference: AACL Tutorial

- Adapters: small trainable submodules inserted in transformers



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

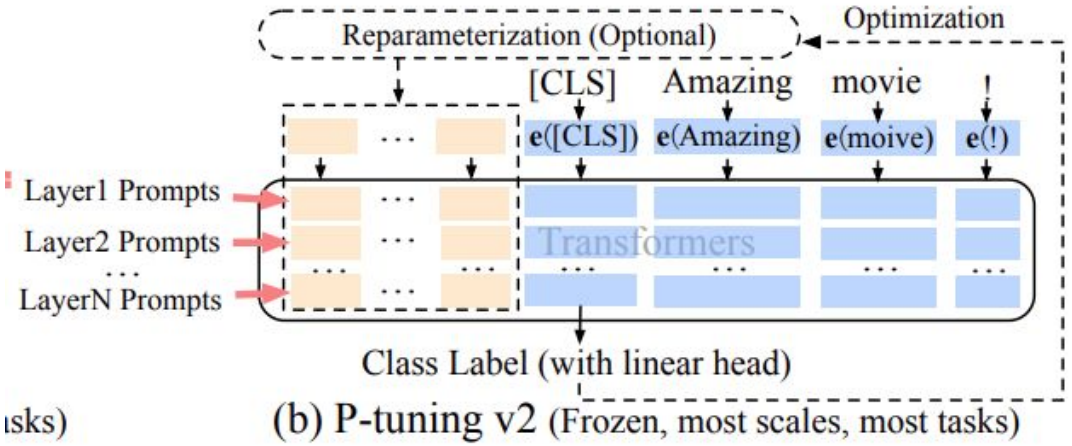

Transformer Layer
Layer Norm
+
Adapter
2x Feed-forward layer
Layer Norm
+
Adapter
Feed-forward layer
Multi-headed attention


Adapter Layer
+
Feedforward up-project
Nonlinearity
Feedforward down-project

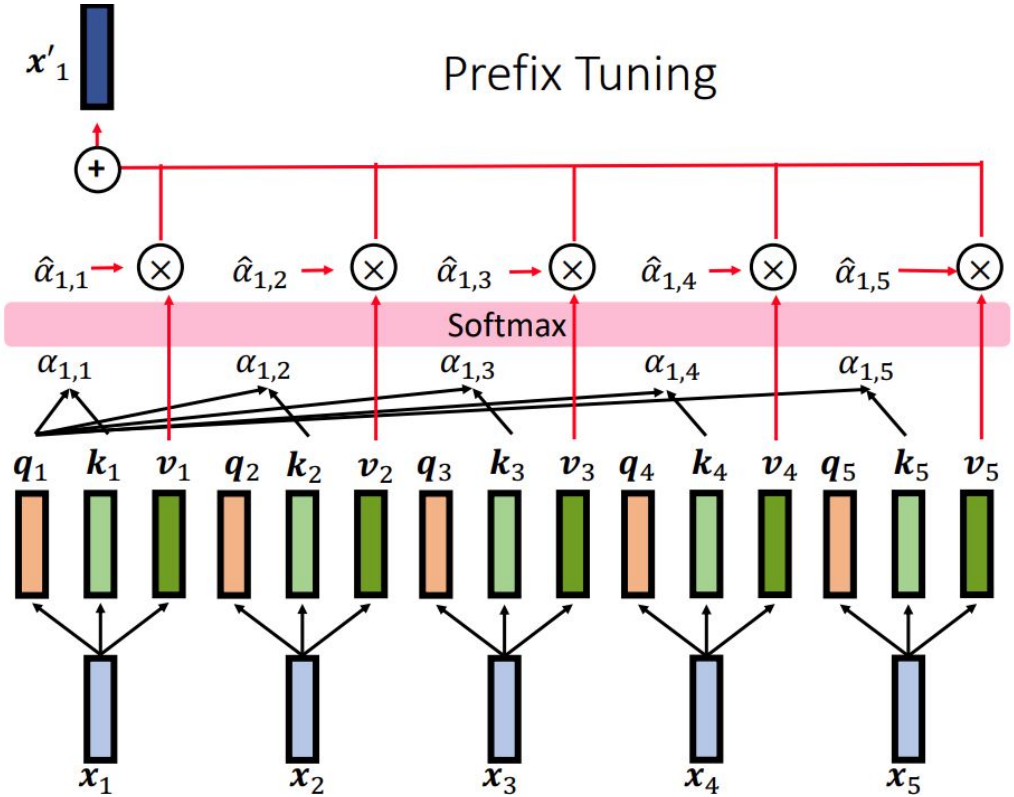
P-tuning V2

- Detail can be found at [p-tuning implementation]

reference: [AACL Tutorial]

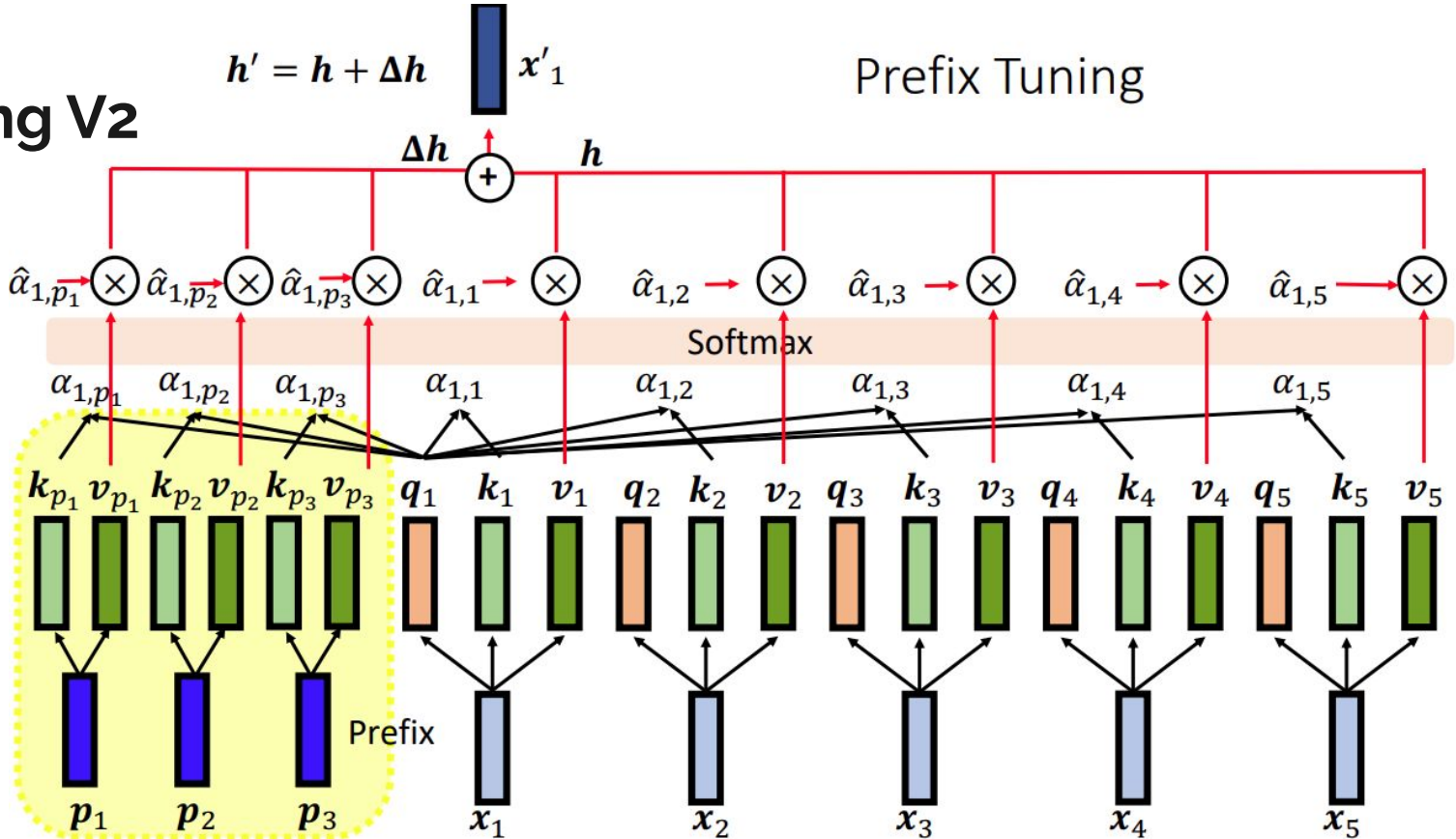


P-tuning V2



reference: [AACL Tutorial]

P-tuning V2



reference: [AACL Tutorial]

Prompt-Tuning

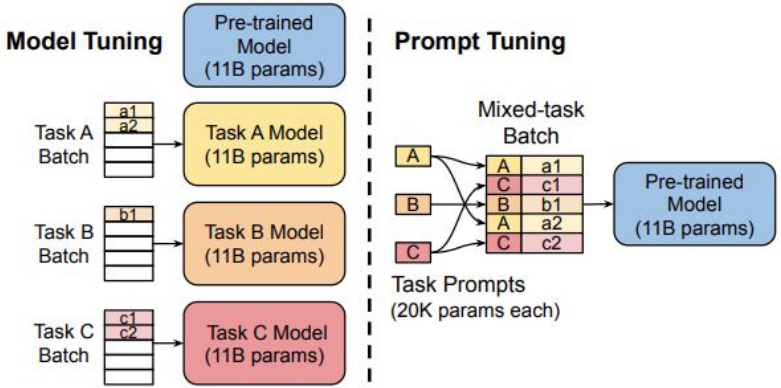
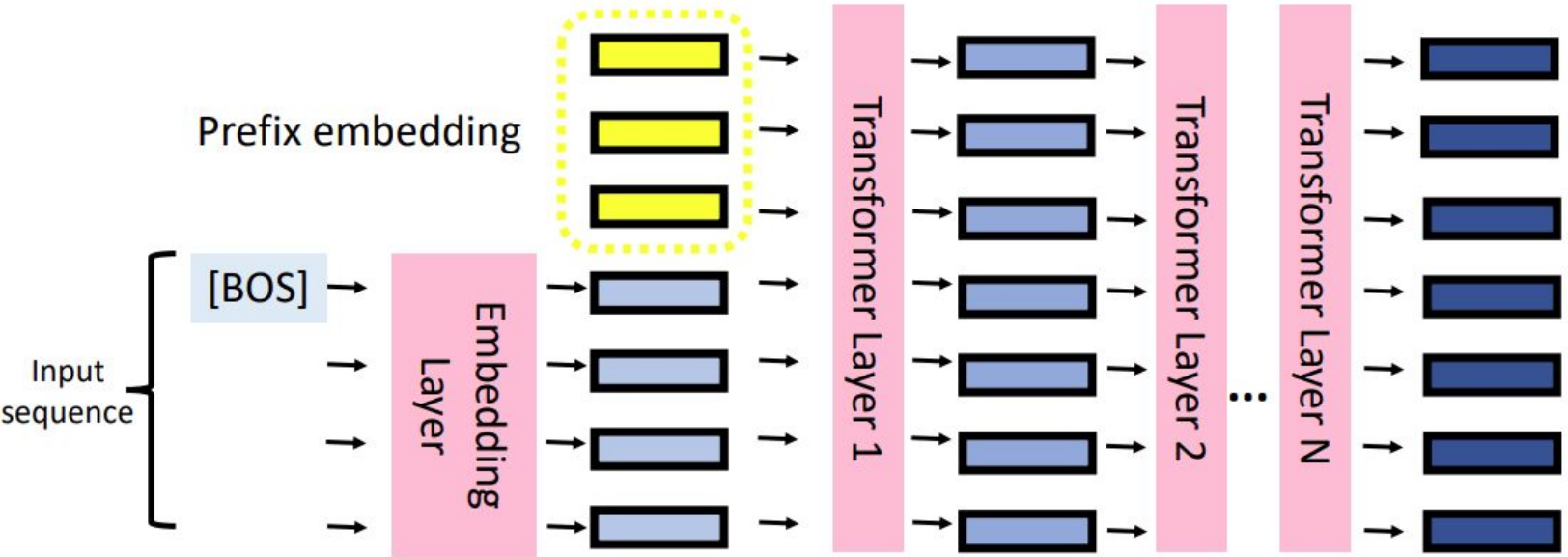


Figure 2: **Model tuning** requires making a task-specific copy of the entire pre-trained model for each downstream task and inference must be performed in separate batches. **Prompt tuning** only requires storing a small task-specific prompt for each task, and enables mixed-task inference using the original pre-trained model. With a T5 "XXL" model, each copy of the tuned model requires 11 billion parameters. By contrast, our tuned prompts would only require 20,480 parameters per task—a reduction of *over five orders of magnitude*—assuming a prompt length of 5 tokens.

Prompt-tuning

reference: AACL Tutorial

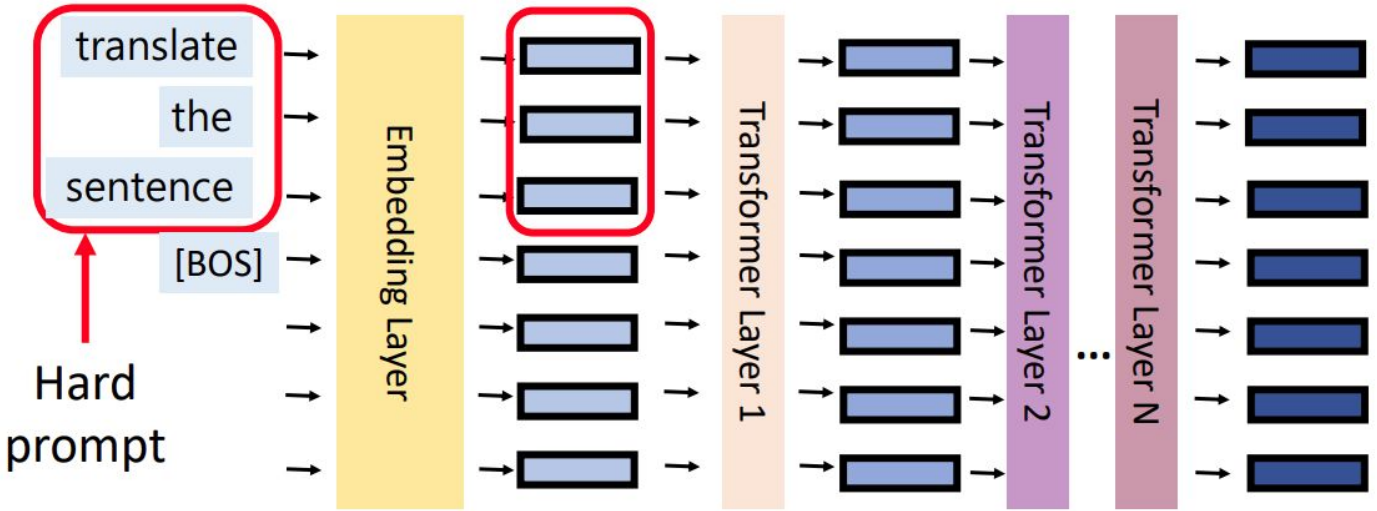


Lester, Brian, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

Prompt-tuning

reference: AACL Tutorial

- Soft Prompting can be considered as the soften version of prompting



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.

HW - Prompt

- Repo : P-tuning-v2
- Backbone PLM: BERT-large or RoBERTa-large
- Prompt: Prefix-tuning or Prompt-Tuning
- Run at least **3** tasks and report the result
- **At least one task should outperform the provided result in the Repo**
- **Get to know the implementation of these two well-known prompt**

Released results on BERT-large

	BoolQ	COPA	RTE	WiC	WSC	CoNLL04	OntoNotes 5.0	CoNLL12
Result	74.3	77.0	80.1	75.1	68.3	84.5	86.4	85.3
Total Epochs	100	80	60	80	80	40	30	45
Best Epoch	58	12	30	56	17	33	24	43

Released results on RoBERTa-large

	BoolQ	COPA	RTE	WiC	WSC	CoNLL03	CoNLL04	OntoNotes 5.0	CoNLI
Results	84.0	92.0	86.6	73.7	64.4	91.8	88.4	90.1	84.7
Total Epochs	100	120	100	50	10	30	80	60	45
Best Epoch	86	78	65	31	3	28	45	59	37

Provided result of the P-tuning-v2

Note

- When building up the environment, add huggingface_hub==0.7.0 if encounter error of huggingface_hub

HW - Adapter

- Repo : Easy Adapter
- Run at least **2** types of adapter
- **Report the evaluation accuracy**
- **Get to know the implementation of these well-known adapter**

Deadline

- **10/30**: 5, 6, 7, 8
- **11/06**: 1, 2, 3, 4
- You only have to do **one** of adapter or prompt HW

Reference

Repo:

P-tuning-v2, Easy Adapter

Paper:

Adapter (Adapter Bias, BitFit)

Prompt (Prefix-tuning, Prompt-Tuning)

Other:

P-tuning implementation, AACL 2022 Tutorial