Cross-lingual/ Fine-tuning

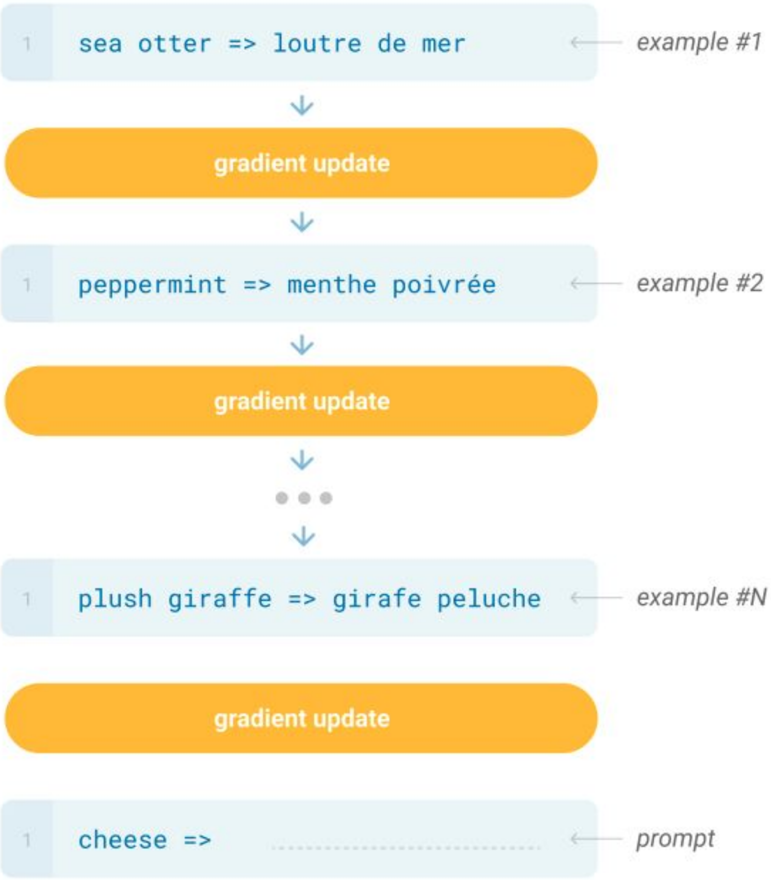
第六組 余奇恩 林熙哲 金家逸

Outline

- Cross Lingual
  - Task
  - Observation
  - Experiment
  - Inference Time and Performance
  - Conclusion
- Efficient Tuning : P-tuning v2
  - Introduction
  - Task
  - Performance
  - Conclusion

Cross-Lingual

Traditional fine-tuning



"In-context" Learning

Few-shot Learning

(no gradient descent)

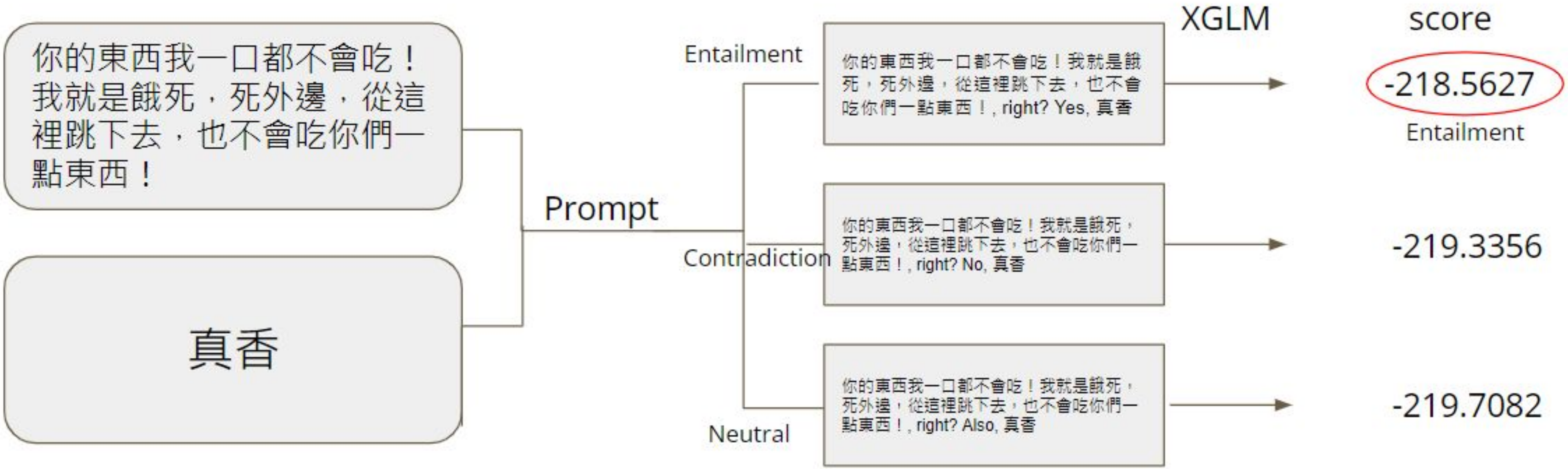
One-shot Learning

Zero-shot Learning

From training or dev set

http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/GPT3%20(v6).pdf

Cross lingual - Task



Cross lingual - Task

From XNLI dev set sample k example

ex:(k=3)

hello stackoverflow. right? yes, she is a programmer.

hello world. right? Also, she is a programmer.

hi. right? No, she is a programmer.

Use XNLI test set to evaluate accuracy

hola~ right? yes/no/also , she is a programmer

->比較三者的可能性, 來決定句子的前後關係

Cross lingual - Task

From XNLI dev set sample k example

ex:(k=3)

hello stackoverflow. right? yes, she is a programmer.

hello world. right? Also, she is a programmer.

hi. right? No, she is a programmer.

Use XNLI test set to evaluate accuracy

hola~ right? yes/no/also , she is a programmer

->比較三者的可能性，來決定句子的前後關係

K shot translation - Some Observation

模型在最後重複翻譯 "這個問題已經解決了"

input:

```
translate from Chinese to english:
印第安纳波利斯还有很多艺术和文化的机会, 但没有比思域剧院更好的。 => Many more arts and cultural opportunities exist in
Indianapolis, yet none finer than Civic Theatre.
但你为什么不开始呢, 因为你已经有更多的时间去思考它, 如果你不介意的话。 => well why don't you start because you've had m
ore time to think about it if you don't mind
我想这就是为什么我记得。 => I think that's why I remember that.
嗯, 我根本没想过, 但是我很沮丧, 最后我又和他说话了。 =>
translate from Chinese to english:
市场代表着这座城市的战后重生。 => The marketplace represents the city's post-war rebirth.
该香水工厂位于南非区域后面。 => The Perfume Factory is behind the South African Area.
可能是因为出错, 而导致会计师最近被解雇了。 => There was the potential for error as the accountant was recently fired.
我还没有和他再次谈论。 =>
```

output:

```
translate from Chinese to english: 印第安纳波利斯还有很多艺术和文化的机会,但没有比思域剧院更好的。 => Many more arts an
d cultural opportunities exist in Indianapolis, yet none finer than Civic Theatre. 但你为什么不开始呢,因为你已经有更多
  时间去思考它,如果你不介意的话。 => well why don't you start because you've had more time to think about it if you don
't mind 我想这就是为什么我记得。 => I think that's why I remember that. 嗯,我根本没想过,但是我很<unk>丧,最后我又和他说
  了。 => Well, I didn't think about it, but I was really disappointed, and then I spoke to him again.
translate from Chinese to english: 市场代表着这座城市的战后重生。 => The marketplace represents the city's post-war reb
irth. 该香水工厂位于南非区域后面。 => The Perfume Factory is behind the South African Area. 可能是因为出错,而导致会计师
  近被解雇了。 => There was the potential for error as the accountant was recently fired. 我还没有和他再次谈论。 => I h
aven't discussed it with him again. 这个问题已经解决了。 => The problem has been resolved. 这个问题已经解决了。 => The
problem has been resolved. 这个问题已经解决了。 => The problem has been resolved. 这个问题已经解决了。 => The problem h
as been resolved. 这个问题已经解决了。 => The problem has been resolved. 这个问题已经解决了。 => The problem has been r
esolved. 这个问题已经解决了。 => The problem has been resolved. 这个问题已经解决了。 => The problem has been resolved.
  个问题已经解决
```

k shot translation - 0 shot vs. 3 shot

```
zh  contradiction          嗯，我根本没想过，但是我很沮丧，最后我又和他说话了。    我还没有和他再次谈论。  2
zh  entailment          嗯，我根本没想过，但是我很沮丧，最后我又和他说话了。    我非常沮丧，我刚刚开始跟他说话
zh  neutral          嗯，我根本没想过，但是我很沮丧，最后我又和他说话了。    我们谈得很好。  2   6   facetofac
```

0-shot

```
zh  contradiction    嗯,我根本没想过,但是我很<unk>丧,最后我又和他说话了。    I haven't talked to him again.
zh  entailment   嗯,我根本没想过,但是我很<unk>丧,最后我又和他说话了。    我非常<unk>丧,我刚刚开始跟他说话。
zh  neutral  嗯,我根本没想过,但是我很<unk>丧,最后我又和他说话了。    我们谈得很好。
```

3 shot

```
zh  contradiction   Well, I didn't think about it, but I was really disappointed, and then I spoke to him again.    I haven't discussed it with him again.
zh  entailment  Well, I didn't think about it, but I was really disappointed, and then I spoke to him again.    I was very upset, I was just starting to talk to him.
zh  neutral Well, I didn't think about it, but I was really disappointed, and then I spoke to him again.    We're on good terms.
```

->在translation 上，in-context learning大幅提升了模型的翻譯能力

Cross lingual - Experiment

1. 0 shot vs. 12 shot (prompt is en, example and inference are the same leng)
   a. choose examples randomly from 3 labels
   b. choose examples equally from 3 labels and arrange it in specific order
2. Translate into high resouce language(en) and evaluate 0 shot vs. 12 shot
3. Change to our prompt to evaluate
4. Cross lingual test - example and inference are in different lengauge

Cross lingual - inference time

evaluation:

0 shot zh : 15min

2 shot zh : 18 min

12shot zh : 40 min

Translation:

0 shot :

3 shot :

Cross lingual - Experiment

1. 0 shot vs. 12 shot (prompt is en, example and inference are the same leng)
   a. choose examples randomly from 3 labels
   b. choose examples equally from 3 labels and arrange it in specific order
2. Translate into high resouce language(en) and evaluate 0 shot vs. 12 shot
3. Change to our prompt to evaluate
4. Cross lingual test - example and inference are in different lengauge

0 shot vs. 12 shot - example order design

At first, we just choose examples randomly -> bad performance

(In English case, there are 2 yes, 6 no and 4 also)

=>give example uniformly(4 yes, 4 no and 4 also cases) in following order

Pattern1 : yes x4, no x4, also x4 + inference data

Cross lingual - Performance

example and inference case are in the same lang

	en	fr	ru	zh	hi	ur	bg	vi
0 shot	52.30	47.62	44.79	44.93	42.10	40.78	47.56	46.85
12 shot	43.95	38.74	34.45	39.34	37.05	35.11		
12 shot pattern1	55.53	52.38	47.35	44.35	45.39	45.39	48.82	51.30

Cross lingual - Performance

example and inference cases translating into English with 0 shot

	en	fr	ru	zh	hi	ur	bg	vi
0 shot (no translation)	52.3	47.62	44.79	44.93	42.10	40.78	47.56	46.85
0 shot	x	45.39	42.28	44.85	41.50	39.44	45.93	44.61
12 shot	x	37.62	35.93	39.38	39.52	33.62	41.04	39.26

Cross lingual - Performance (3 shot translation)

	en	fr	ru	zh	hi	ur	bg	vi
0 shot (no translation)	52.3	47.62	44.79	44.93	42.10	40.78	47.56	46.85
0 shot (3 shot translation)	x	50.12	48.51	48.25	46.02	43.99	50.34	48.93
12 shot	x	41.96	40.12	41.54	40.92	35.84	43.21	42.63
12 shot pattern1	x		52.18	50.72		46.37		

Cross lingual - Performance

prompt 為該語言

	en	zh
0 shot	52.3	
12 shot	55.53	33.49

right? yes -> 對嗎？ 是的

right? no -> 對嗎？不是的

right? Also-> 對嗎？ 並且

Cross lingual - Performance

from English transfer to fr and zh

	en - fr	en - zh	zh - vi
12 shot	35.19	33.75	46.45

Cross lingual - Performance

改成自己的prompt

	en	fr	ru	zh	hi	ur	bg	vi
0 shot								
12 shot								

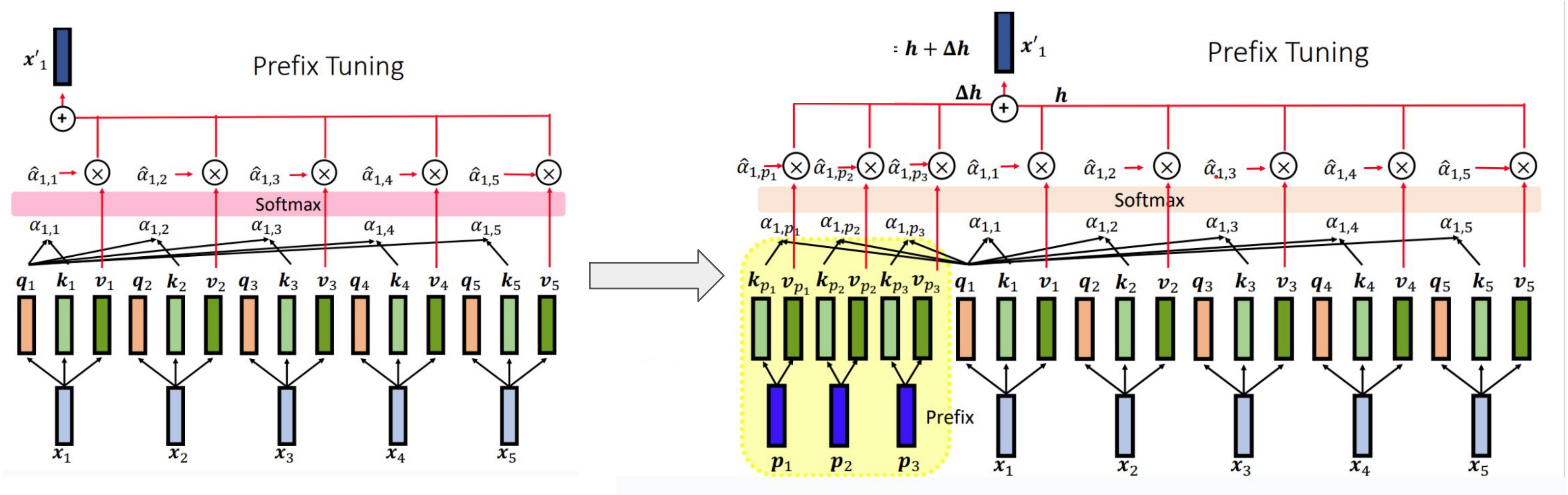
```
sentence 1: excuse me we pay for any you know the child care but we don't pay as much as they do off base
sentence 2: Childcare costs $2000 more off base.
The relation between sentence 1 and sentence 2 is neutral

sentence 1: They took Joe with them, and my Granny said, she said it was such a sad time in the house because, you know, everybody was missi
sentence 2: Everyone in the house was moping because they missed Joe so much.
The relation between sentence 1 and sentence 2 is entailment

sentence 1: Flanking it, a modern octagonal church to the east and a chapel and hexagonal tower to the west represent the city's post-war re
sentence 2: The marketplace represents the city's post-war rebirth.
The relation between sentence 1 and sentence 2 is contradiction

sentence 1: Well, I wasn't even thinking about that, but I was so frustrated, and, I ended up talking to him again.
sentence 2: I havent spoken to him again.
The relation between sentence 1 and sentence 2 is entailment
```

Cross Lingual - Conclusion

1. In-context learning do raise the performance of LM about 5%, but only if example data is given in specific way.
2. Note that if example data is given in an improper way(unbalenced labels), in-context learning method may drop accuracy.
3. When a model performs poorly in a certain language, translating it into a high resource lengauge(such as English) can improve performance.

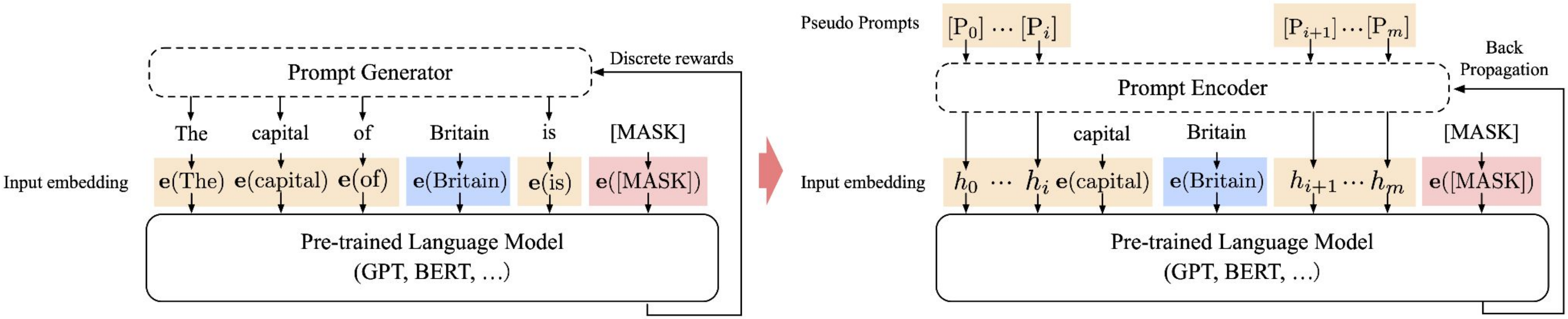
Efficient Tuning : P-tuning v2

Prefix tunning



1. Add a prefix in front of each layer in the Transformer
2. During training, only update the parameters of the prefix while keeping the pretrained parameters in the Transformer fixed.

P-tuning v1



(a) Discrete Prompt Search (b) P-tuning

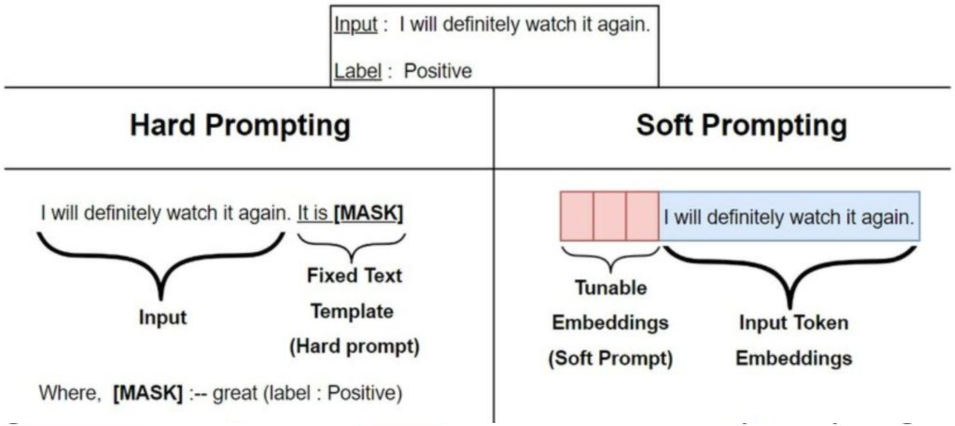
- Principle: Automatic Template Construction :

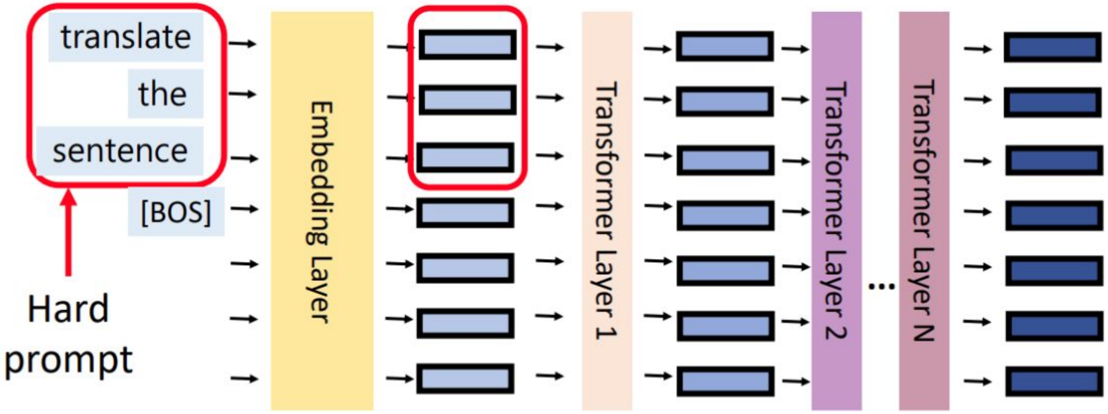
  Add Pseudo Prompts at the input layer and optimize the Pseudo Prompts and the Prompt Encoder using gradient-decent methods.

- Fix problem : natural language template are sensitive to variation:

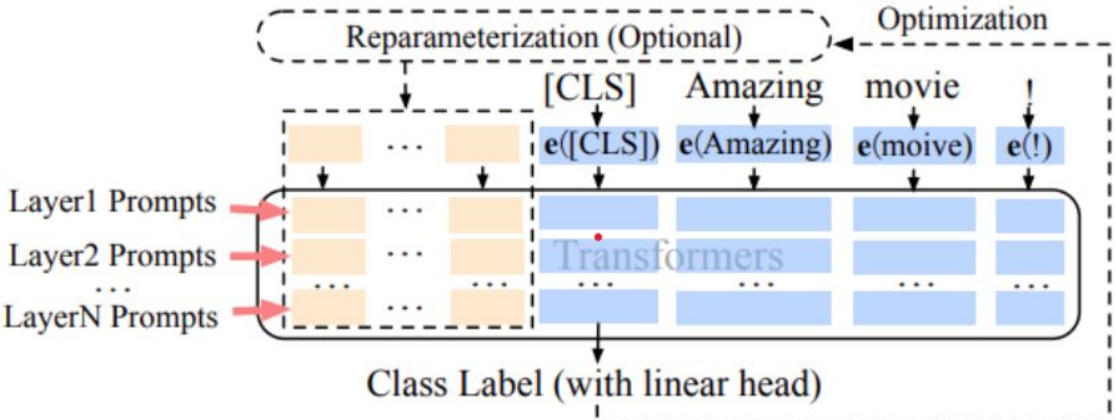
  P-tuning transforms previous templates, which were constructed from natural language (discrete) prompts, into parameterized (continuous) and learnable embedding layers.

Comparision : hard (discrete) prompt / soft (continuous) prompt





P-tuning v2



(b) P-tuning v2 (Frozen, most scales, most tasks)

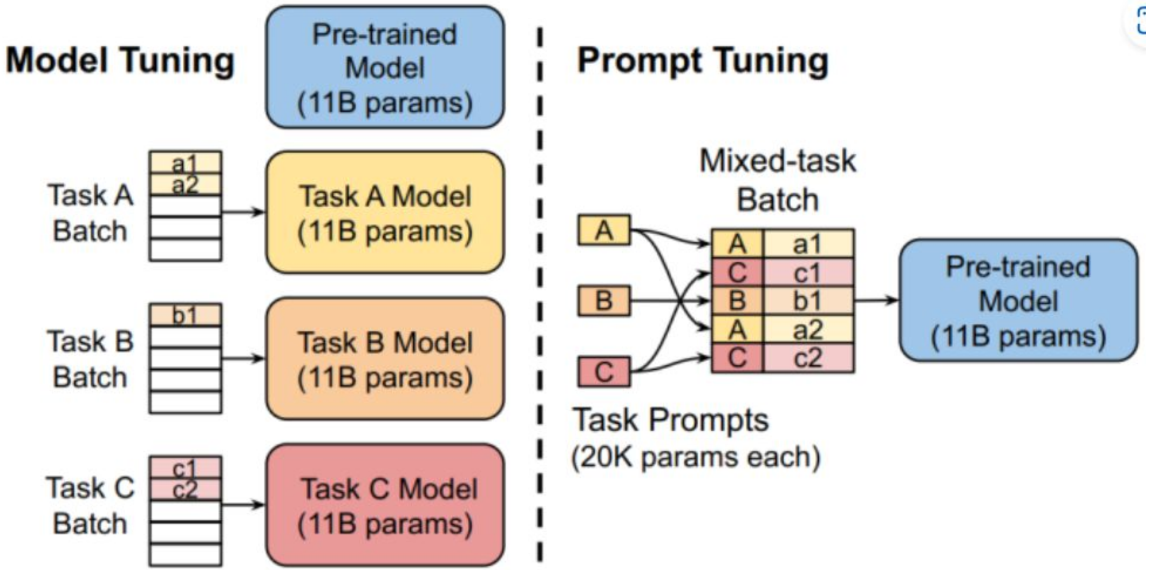
- Principle :

  Add custom-sized layer prompts in front of the original input, and in subsequent training for downstream tasks, freeze all parameters of the pretrained model while training only these prompts.

implementation in code

```
class PrefixEncoder(torch.nn.Module):
    r'''
    The torch.nn model to encode the prefix

    Input shape: (batch-size, prefix-length)

    Output shape: (batch-size, prefix-length, 2*layers*hidden)
    '''
    def __init__(self, config):
        super().__init__()
        self.prefix_projection = config.prefix_projection
        if self.prefix_projection:
            # Use a two-layer MLP to encode the prefix
            self.embedding = torch.nn.Embedding(config.pre_seq_len, config.hidden_size)
            self.trans = torch.nn.Sequential(
                torch.nn.Linear(config.hidden_size, config.prefix_hidden_size),
                torch.nn.Tanh(),
                torch.nn.Linear(config.prefix_hidden_size, config.num_hidden_layers * 2 * config.hidden_size)
            )
        else:
            self.embedding = torch.nn.Embedding(config.pre_seq_len, config.num_hidden_layers * 2 * config.hidden_size)

    def forward(self, prefix: torch.Tensor):
        if self.prefix_projection:
            prefix_tokens = self.embedding(prefix)
            past_key_values = self.trans(prefix_tokens)
        else:
            past_key_values = self.embedding(prefix)
        return past_key_values
```

Prompt Tuning



- Principle :

  For each task, define its own prompt, concatenate it to the data as input (only at the input layer) ,and simultaneously freeze the pretrained model for training.

P-tuning - Task

1. Bert to wic
2. Bert to wsc
3. Robert to wic
4. Robert to wsc

1. Bert to wic

	learning rate	batch size	dropout	epoch	seed	performance
default	1e-4	16	0.1	80	44	75.1
ours	1e-5	8	0.1	100	1000	72.41

2. Bert to wsc (最後一次加了 prefix projection)

	learning rate	batch size	dropout	epoch	psl	seed	performance
default	5e-3	16	0.1	80	20	44	68.3
ours	5e-3	16	0.1	80	20	44	65.38
	3e-4	16	0.1	80	20	44	64.42
	3e-5	16	0.1	80	20	44	67.31
	1e-5	16	0.1	80	20	44	66.35
	3e-6	16	0.1	80	20	44	65.38
	3e-5	16	0.2	80	20	44	63.46
	3e-4	16	0.1	80	8	44	67.31
	3e-5	16	0.1	80	20	44	64.42

3. roberta-wic

	learning rate	batch size	dropout	epoch	seed	performance
default	1e-2	32	0.1	50	11	73.7
ours	5e-3	32	0.1	50	225	68.97
	7e-3	64	0.15	60	11	68.18
	8e-3	32	0.1	80	172	72.57
	2e-2	32	0.1	40	172	71.16
	9e-3	29	0.08	40	172	71.16

4. roberta-wsc

	learning rate	batch size	dropout	epoch	seed	performance
default	1e-2	16	0.1	10	44	64.4
ours	1e-2	16	0.1	10	1	63.46
	9e-3	16	0.1	20	1	63.46

Comparison : best performence of four case

	default	ours
Bert to wic	75.1	72.41
roberta-wic	73.7	72.57
Bert to wsc	68.3	67.31
roberta-wsc	64.4	63.46

1. for model performance : Bert > roberta
2. for task performance : wic > wsc

Efficient tuning - conclusion

1. Efficient finetuning addresses the issue of the high cost of retraining the entire pretrained model.
2. Prefix tuning, P-tuning v1, and P-tuning v2 have all improved upon traditional hard prompt methods, transforming them into trainable soft prompt methods, which can enhancing the stability of the model's performance.
3. These methods all rely on freezing the parameters of the pretrained model while adjusting the embedding parameters, resulting in reduced costs and faster training times.


Q&A