

Seminar

On Transferability of Prompt Tuning for Natural Language Processing

Indicator: Anh-Thu Le, Nhat-Truyen Nguyen, Minh-Dat Nguyen



Table of contents



1 Background

Fine Tuning, Prompt Tuning

2 Prompt Transfer

2.1 Settings for the experiment

2.2 Cross-Task Transfer

2.3 Cross-Model Transfer

3 Conclusion





+01

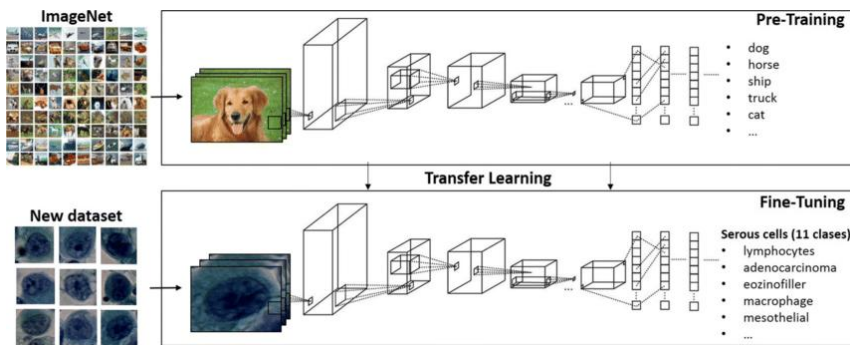


Background

Fine Tuning, Prompt Tuning

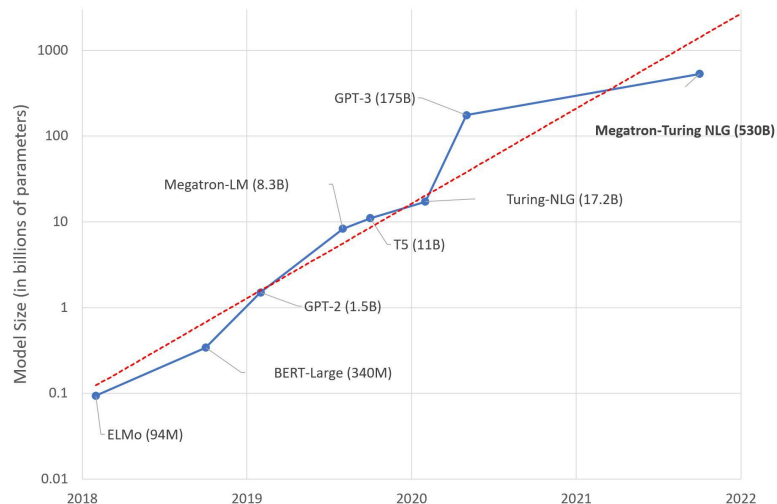
Definition

- Transfer learning method
- Use the pre-trained weights to train new dataset
- Better performance compare with directly train on our small dataset



Challenge

- Pre-trained Language Models (PLMs) development
- Require a large number of parameters
- Take a lot of computational resources

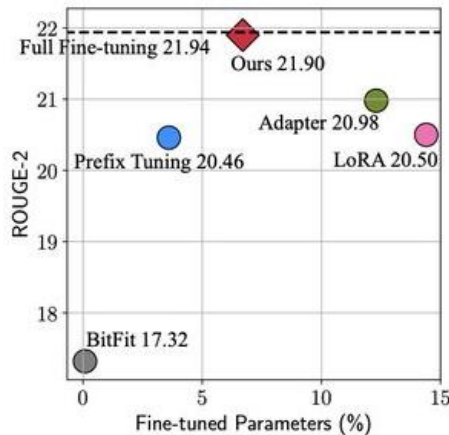
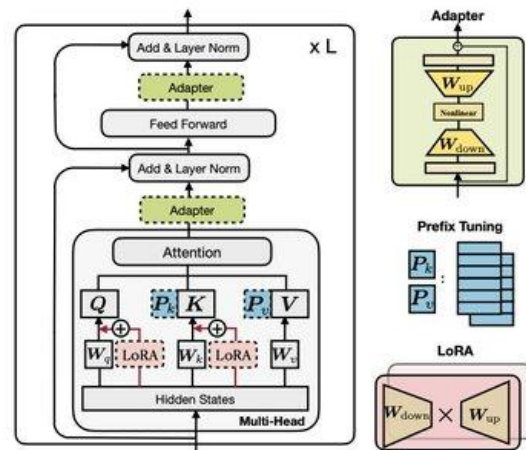


Overcome this challenge

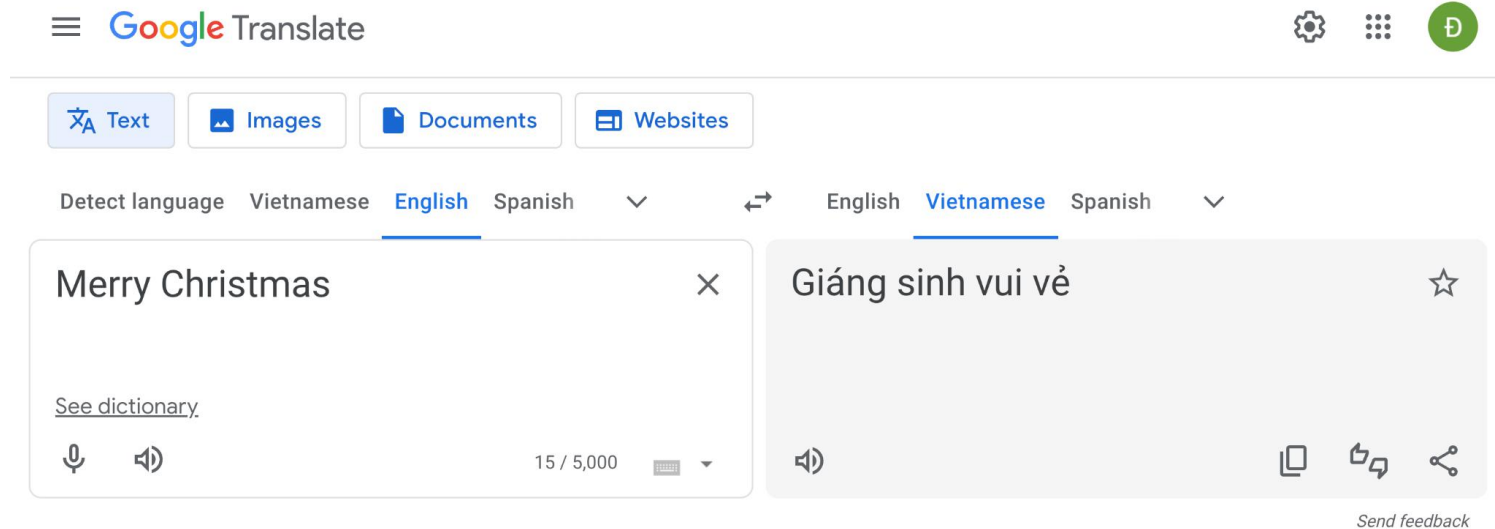
Parameter-Efficient Fine Tuning (PEFT) Methods

- Optimize a small part of parameters for downstream tasks while freezing the rest of the parameters of the PLM

Adapter	Neil, et al., Parameter-Efficient Transfer Learning Learning for NLP, ICML, 2019.
Prefix	Li, et al., Prefix-Tuning: Optimizing Continuous Prompts for Generation, ACL, 2021.
LoRA	Hu, et al., LoRA: Low-Rank Adaptation of Large Language Models, ICLR, 2022.
BitFit	Zaken, et al., BitFit: Simple Parameter-efficient Fine-Tuning for Transformer-based Masked Language-models, ACL, 2022.
Prompt	Lester et al., The Power of Scale for Parameter-Efficient Prompt Tuning, EMNLP, 2021.
Combine	He at al, Towards a Unified View of Parameter-Efficient Transfer Learning, ICLR, 2022.



Fine Tuning



How to improve?

Update the model itself!

Prompt Tuning



ChatGPT 3.5 ▾



You

Translate the sentence from English to Vietnamese: "Merry Christmas"



ChatGPT

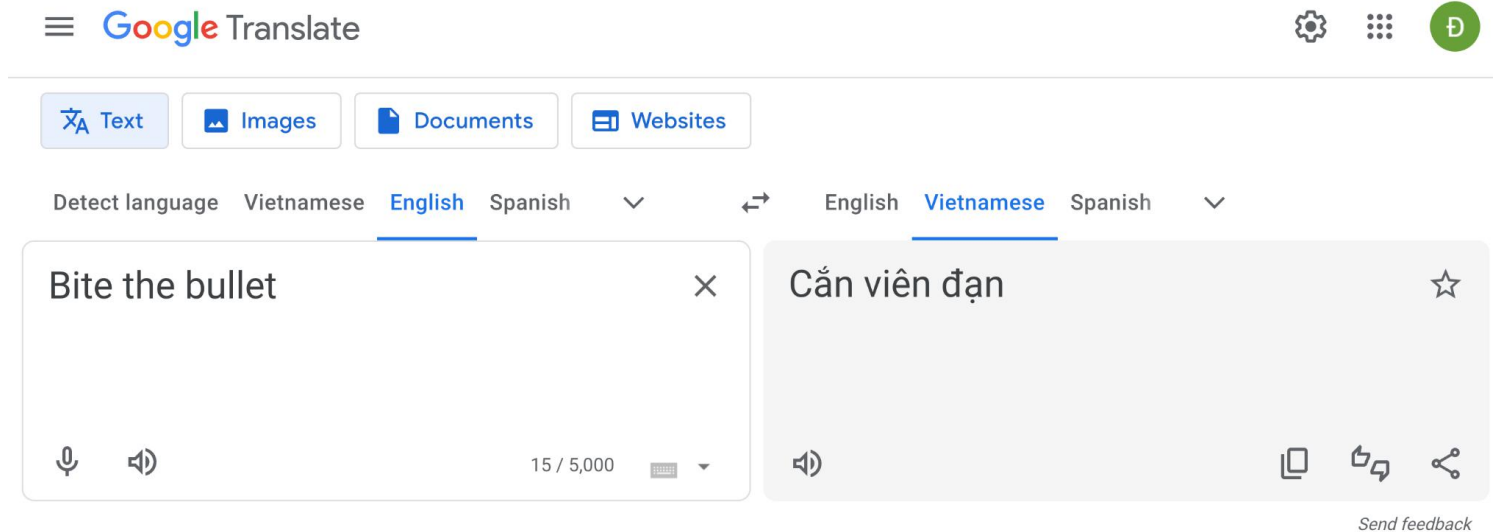
The translation of "Merry Christmas" from English to Vietnamese is "Chúc Mừng Giáng Sinh."



How to improve?

Just tune the prompt!

When Google Translate fails



The screenshot shows the Google Translate web interface. At the top, there's a menu icon, the Google Translate logo, and settings, grid, and profile icons. Below the header, there are tabs for Text, Images, Documents, and Websites. The language selection bar shows 'Detect language' and 'Vietnamese' on the left, and 'English' and 'Vietnamese' on the right, with a double-headed arrow between them. The input text 'Bite the bullet' is in the left box, and the output 'Cẩn viên đạn' is in the right box. The right box also contains a star icon, a speaker icon, and icons for copy, share, and feedback. The feedback link is labeled 'Send feedback'.

Google Translate

Text Images Documents Websites

Detect language Vietnamese English Spanish ↕ English Vietnamese Spanish

Bite the bullet ×

Cẩn viên đạn ☆

15 / 5,000

Send feedback

When ChatGPT fails

N

You

Translate the sentence from English to Vietnamese: "Bite the bullet"



ChatGPT

The phrase "Bite the bullet" translates to "Nhai viên đạn" in Vietnamese.

N

You

The sentence "Bite the bullet" is actually a collocation in English

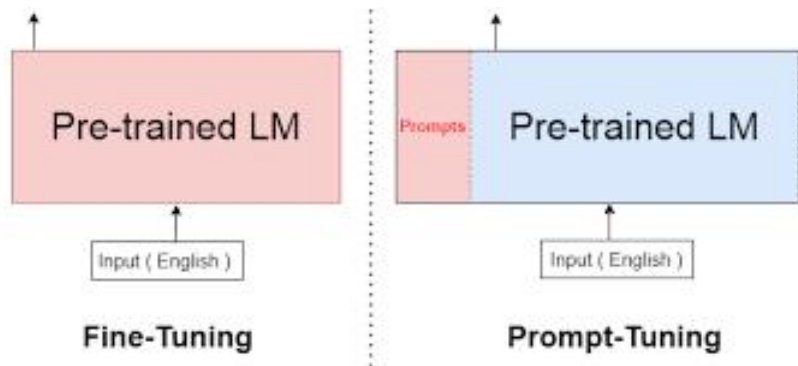


ChatGPT

You're absolutely right! "Bite the bullet" is an English idiom meaning to endure a painful or difficult situation with courage and determination. In Vietnamese, this expression could be translated as "Nhấn nhịn vượt qua khó khăn" or "Đối mặt với khó khăn một cách dũng cảm."

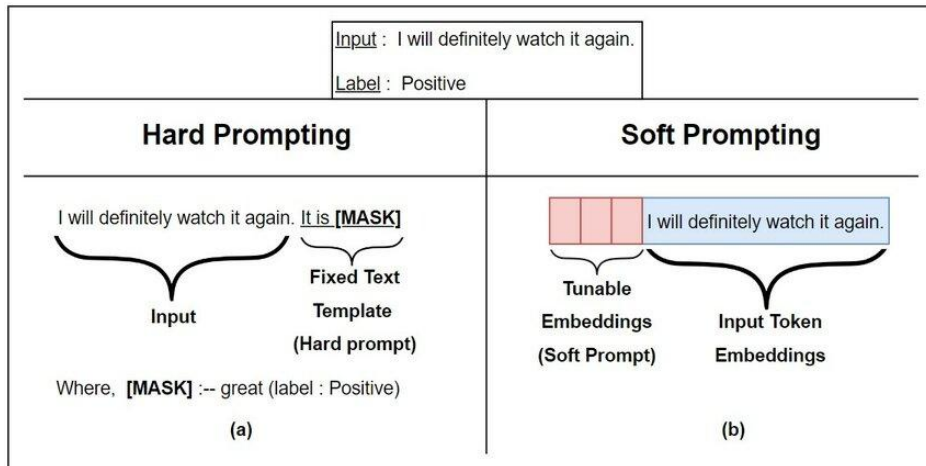
Definition

- Effectively fine-tuning pre-trained models on small datasets
- Use soft prompt to tune the output of a pre-trained model



Hard / Soft Prompt

<Prompt><Instance>



Pros/Cons

- **Advantage:** Lowest computation costs
- **Challenge:** Slow Convergence

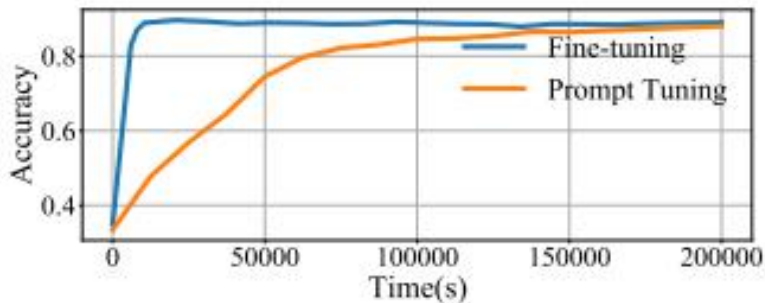


Figure 2: Validation accuracies against training time of fine-tuning and PT for RoBERTa_{LARGE} on MNLI. PT takes much more training time.

Proposed Method

Key idea of this paper

- **Should we look at knowledge transfer?**
 - Because only soft prompts have been learned anyway, wouldn't it be possible to transfer knowledge by simply detaching and pasting the tokens? **Prompt Transfer**
 - Accordingly, this study investigates whether **Prompt Transfer** can contribute to improving **Prompt Tuning** effectively



+ 02



Prompt Transfer

Cross-Task Transfer, Cross-Model Transfer



+



Settings for the experiment

Cross-Task Transfer:

- **17 NLP Tasks in 6 Categories:**
 - **Sentiment Analysis:** IMDB, SST-2, Laptop, Restaurant, Movie, Regionales, TweetEvel
 - **NLI:** MNLI, QNLI, SNLI
 - **Ethical Judgment:** Deontology, Justice
 - **Paraphrase Identification:** QQP, MRPC
 - **QA:** SQuAD, NQ-Open
 - **Summarization:** Multi-News, SAMSum

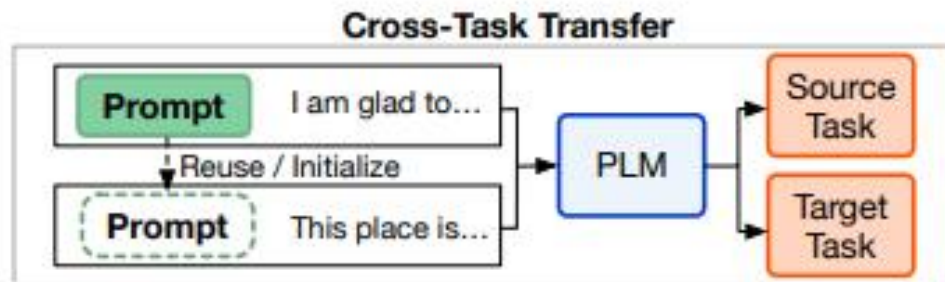
Cross-Model Transfer:

- **RoBERTa**
 - RoBERTa-large: MLM representative
- **T5**
 - T5-XXL: Representative seq2seq training

Target

- Check prompt transferability in zero-shot setting
- Check effectiveness and efficiency of prompt tuning with transfer

Implementation



- Since only the task is different, we experimented with only prompts being removed from the source task and attached to the target task

Prompt Input

- Input structure:
<soft><mask><text_a><text_b>
- Example:
soft: Does the first sentence entails the second?
mask:
text_a: A soccer game with multiple males playing.
text_b: Some men are playing a sport.

Implementation



✓
0s

```
[8] # cấu trúc input
trainer.template.text
```

```
[{'add_prefix_space': ' ', 'soft': None, 'duplicate': 100, 'same': True},
 {'add_prefix_space': ' ', 'mask': None},
 {'add_prefix_space': ' ', 'placeholder': 'text_a'},
 {'add_prefix_space': ' ', 'placeholder': 'text_b'}]
```

✓
0s

```
[9] # độ dài soft prompt
trainer.template.num_tokens
```

```
100
```

✓
0s

```
[10] # kích thước prompt embedding (num_tokens, embedding_dim)
trainer.template.soft_embeds.shape
```

```
torch.Size([100, 768])
```



Train prompt

- Backbone: RoBERTa-base
- Dataset: **SST2**
- Task type: Sentiment analysis
- Evaluate accuracy: **77,17%**

Cross-task eval

- Zero-shot transfer
- Target dataset: **rotten_tomatoes**
- Task type: Sentiment analysis
- Evaluate accuracy: **72,13%**

PT on target dataset

- Evaluate accuracy: **74,10%**

SST2: analyzing sentiment within sentences, at the sentence level.

rotten_tomatoes: expressing the overall sentiment towards movies.



Results

Similar or better performance than soft prompts starting from random initialization, and less training time

→ Soft prompts can be used for **similar tasks** without additional learning

Source Task		IMDB	SST-2	laptop	restaurant	Movie	Tweet	MNLI	QNLI	SNLI	deontology	justice	QQP	MRPC	random prompt
	IMDB	100	95	65	84	101	52	37	55	37	58	63	43	82	
	SST-2	91	100	88	92	93	66	50	59	38	61	62	57	66	
	laptop	76	91	100	93	84	74	38	55	37	59	63	43	84	
	restaurant	80	92	95	100	81	70	38	55	37	59	62	44	81	
	Movie	98	80	70	40	100	54	37	55	37	59	62	62	69	
	Tweet	88	94	66	90	94	100	41	55	37	59	62	43	80	
	MNLI	55	61	70	62	61	54	100	79	62	60	62	72	81	
	QNLI	75	53	3	69	80	54	60	100	65	59	61	65	39	
	SNLI	55	53	64	68	58	54	87	82	100	59	62	51	84	
	deontology	63	54	5	5	59	58	38	55	38	100	80	48	75	
	justice	55	79	64	58	82	46	38	55	37	83	100	49	51	
	QQP	55	53	68	8	59	54	43	58	37	59	62	100	78	
	MRPC	59	53	3	1	59	54	38	54	36	59	62	78	100	
	random prompt	54	52	3	2	59	54	38	55	36	58	62	46	75	
Target Task		IMDB	SST-2	laptop	restaurant	Movie	Tweet	MNLI	QNLI	SNLI	deontology	justice	QQP	MRPC	

(a) RoBERTa_{LARGE}

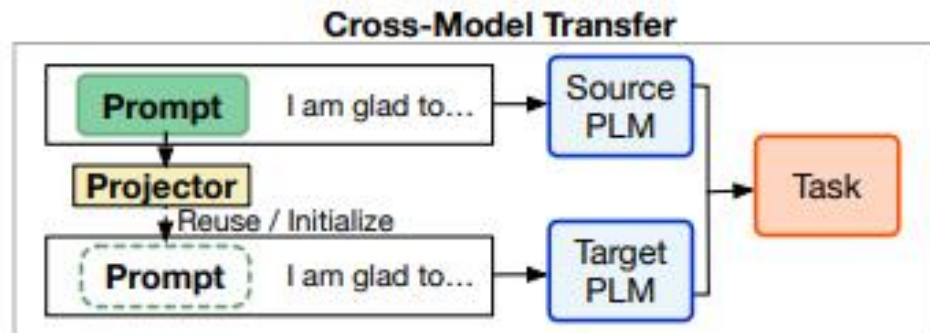
Source Task		IMDB	SST-2	laptop	restaurant	Movie	Tweet	MNLI	QNLI	SNLI	deontology	justice	QQP	MRPC	SQuAD	NQ-Open	Multi-News	SAMSum	random prompt
	IMDB	100	96	79	87	98	65	36	52	34	58	54	67	39	0	1	0	0	
	SST-2	84	100	89	88	67	69	35	55	35	58	56	45	67	0	0	0	0	
	laptop	90	86	100	90	83	76	36	53	36	57	54	41	63	0	0	0	0	
	restaurant	90	92	101	100	81	77	36	53	33	57	57	42	68	0	0	0	0	
	Movie	100	91	81	87	100	68	38	53	37	62	59	55	46	0	1	0	0	
	Tweet	96	92	100	91	84	100	33	53	36	57	56	45	67	0	0	0	0	
	MNLI	65	81	60	45	53	43	100	81	98	57	54	41	69	1	2	4	0	
	QNLI	62	52	70	73	52	56	59	100	64	57	54	41	69	1	1	1	0	
	SNLI	64	66	18	20	53	22	96	76	100	57	54	70	33	0	1	1	0	
	deontology	53	60	41	42	53	30	37	56	36	100	74	63	59	0	0	0	0	
	justice	51	50	26	19	53	55	44	52	41	58	100	41	69	0	0	0	0	
	QQP	51	51	26	20	53	22	36	53	36	58	54	100	78	1	0	0	0	
	MRPC	51	50	28	20	53	21	49	56	48	58	54	84	100	0	0	0	0	
	SQuAD	73	82	69	73	60	63	40	53	38	58	58	48	62	100	20	33	33	
	NQ-Open	73	75	62	47	53	55	42	58	36	56	62	51	50	16	100	23	13	
	Multi-News	62	76	26	19	53	21	39	52	36	57	54	70	33	6	25	100	28	
	SAMSum	76	77	68	75	51	57	36	53	36	57	54	43	62	14	15	67	100	
	random prompt	52	50	26	19	53	22	35	51	35	57	54	41	69	0	0	0	0	
	IMDB																		
	SST-2																		
	laptop																		
	restaurant																		
	Movie																		
	Tweet																		
	MNLI																		
	QNLI																		
	SNLI																		
	deontology																		
	justice																		
	QQP																		
	MRPC																		
	SQuAD																		
	NQ-Open																		
	Multi-News																		
	SAMSum																		
Target Task		IMDB	SST-2	laptop	restaurant	Movie	Tweet	MNLI	QNLI	SNLI	deontology	justice	QQP	MRPC	SQuAD	NQ-Open	Multi-News	SAMSum	

(b) T5_{XXL}

Target

- Find an answer to whether soft prompts can be transferred from *small PLM* to *large PLM*
 - Source model: **RoBERTa** (small PLM)
 - Target model: **T5** (large PLM)

Implementation



- Because the models are different, soft prompts cannot be used directly as cross-tasks
- **Prompt Projectors** were used to ensure that soft prompts can be easily moved from the source model to the target model



+

Min distance

- Prompt trained on source model \mathbf{P}^s
- Prompt trained on target model \mathbf{P}^t
- Minimize the distance between $\text{Project}(\mathbf{P}^s)$ and \mathbf{P}^t

Example

\mathbf{P}^s shape: (100, 768)

\mathbf{P}^t shape: (100, 1024)



Task tuning

- Directly tune the projected prompts $\text{Project}(\mathbf{P}^s)$
- Backpropagate the supervision signals to train the projector weights

+

Results

Task tuning is generalized even for **unseen tasks** in **similar type tasks**

Even if the model is **different**, if there is a **prompt projector**, learned soft prompts can be **used**

→ cross model prompt transfer seems **possible**

Method		SA						NLI		
		IMDB	SST-2	laptop	restaurant	Movie	Tweet	MNLI	QNLI	SNLI
PT on T5 _{XXL}		96.5	97.4	76.6	88.1	97.9	72.5	90.5	95.2	93.4
Random Prompt		49.7	49.0	19.8	17.0	51.6	15.5	31.8	49.3	31.9
(a) Zero-shot Transfer Performance (%)										
laptop	Distance Minimizing	49.6	49.0	76.6	17.5	51.5	14.4	31.8	48.1	32.8
	Task Tuning	82.9	89.3	80.3	85.7	78.6	58.4	32.4	50.7	33.6
MNLI	Distance Minimizing	49.6	50.1	19.8	18.3	51.2	15.0	90.5	49.0	32.9
	Task Tuning	49.7	48.8	19.8	17.0	51.6	16.0	89.8	82.7	88.2

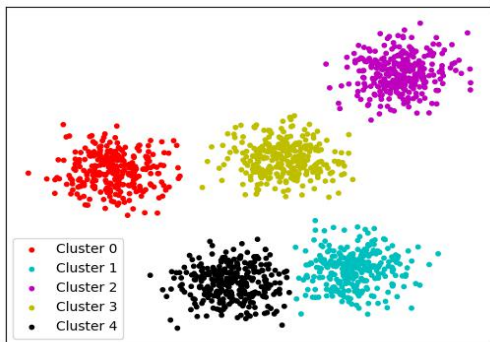


Transferability indicators

Why can soft prompts transfer across tasks?

Embedding Similarity

- Compare similarities between trained prompts (Euclidean, cosine)
- Prompts of different tasks form distinguishable clusters



Model Stimulation Similarity

- Compare similarities between responses of PLM
- Compute the overlapping rate of activated neurons (ON)

→ Model stimulation is more important than embedding distances



+ 03

Conclusion



Contribution

- This study shows the possibility that prompt transition can contribute to improving the efficiency of Prompt Transfer
 - + Cross-task
 - + Cross-model
- Explore possibilities for soft prompts initialization: accelerating training, ensuring efficiency & effectiveness
- Create metrics to determine which elements (and ultimately model stimulation) are effective for soft prompts in PLM





Limitation



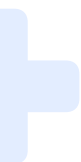
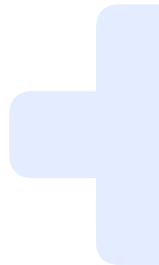
Interpretability

Interpretability issues of soft prompts



Implementation


Incomplete implementation of prompt projector





The end

Thank you for listening
and feel free to ask any questions.



References:

- [1]: [On Transferability of Prompt Tuning for Natural Language Processing](#)
- [2]: [The Power of Scale for Parameter-Efficient Prompt Tuning](#)
- [3]: [Towards a Unified View of Parameter-Efficient Transfer Learning](#)
- [4]: [Prompt as Parameter-Efficient Fine-Tuning](#)

*Implementation
Notebooks*

