P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks

AI 키워드	
날짜	@2025년 1월 6일
콘텐츠	논문
태그	유런

💡
- prompt tuning
- 2022.5.20

https://arxiv.org/pdf/2110.07602

Introduction

fine-tuning

- 전체 모델 파라미터 업데이트
- 메모리 사용량 많음, 각 task에 대한 전체 모델 파라미터 가지고 있어야 함

prompting

- 전체 파라미터 동결 후 자연어 query로 결과 이끌어냄

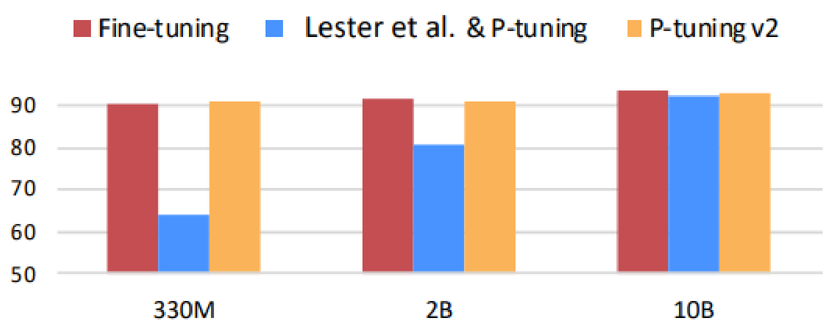
prompt tuning

- continuous prompt(=embedding)만 학습
- Lester et al. (2021)
- continuous 임베딩을 입력 임베딩에 더함 → continuous 임베딩만 업데이트 됨

- v1에서는 10B 이하의 모델/어려운 라벨링 task에 대해 fine-tuning보다 좋지 않은 성능
- 입력 레이어에만 continuous prompt

prompt tuning v2

- 모든 모델에 대해 최적화된 prompt tuning을 통해 fine-tuning을 능가하는 결과를 얻을 수 있음
- 모든 레이어에 continuous prompt



Preliminaries

NLU Task

1. simple classification task
2. hard sequence labeling task: named entity recognition, extractive question answering 등

prompt tuning

V: vocab

M: model

e: embedding layer

x: input text

→ input embedding sequence: [e(x), e("It"), e("is"), e("[MASK]")]

P-Tunign v2

Lack of Universalit

규모

- 10B 이상의 모델에서만 p-tuning이 fine-tuning 능가
- 100M에서 1B의 모델이 흔히 쓰임

tasks

- hard sequence tagging task에서 약함

Deep Prompt Tuning

- v1: input 임베딩 레이어에만 continuous prompt가 들어감

  → sequence 길이에 의해 tuning될 수 있는 파라미터가 한정됨

  → 모델 예측에 간접적인 영향만을 끼치게 됨

- v2: deep prompt tuning = 모든 레이어에 prefix token으로서 continuous prompt가 들어감

  → 모델 예측에 비교적 직접적인 영향

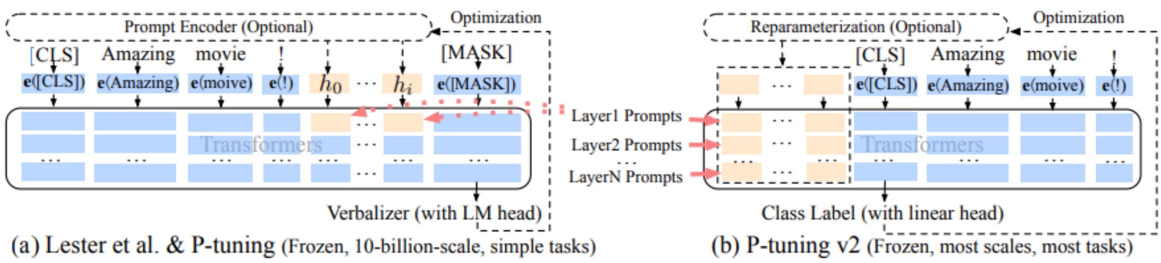


Figure 2: From Lester et al. (2021) & P-tuning to P-tuning v2. Orange blocks (i.e., $h_0, ..., h_i$) refer to trainable prompt embeddings; blue blocks are embeddings stored or computed by frozen pre-trained language models.

Optimization and Implementation

Method	Task	Re-param.	Deep PT	Multi-task	No verb.
P-tuning (Liu et al., 2021)	KP NLU	LSTM	-	-	-
PromptTuning (Lester et al., 2021)	NLU	-	-	✓	-
Prefix Tuning (Li and Liang, 2021)	NLG	MLP	✓	-	-
Soft Prompts (Qin and Eisner, 2021)	KP	-	✓	-	-
P-tuning v2 (Ours)	NLU SeqTag	*(depends)*	✓	✓	✓

Table 1: Conceptual comparison between P-tuning v2 and existing Prompt Tuning approaches (KP: Knowledge Probe; SeqTag: Sequence Tagging; Re-param.: Reparameterization; No verb.: No verbalizer).

- MLP와 같은 reparameterization 인코더 사용함 → task와 dataset에 따라 효과가 없을 수도 있음을 알아냄
- prompt 길이가 핵심적인 역할을 함. 간단한 task는 짧은 prompt를(<20), 어려운 task는 긴 prompt를(~=100)
- 각 task에 대해 fine-tuning 전 공통된 continuous prompt로 multi-task learning
- language modelnig head가 아니라 랜덤 초기화된 classification head 사용. BERT처럼

Experiments

NLU Tasks

- SuperGLUE, Named Entity Recognition, extractive Question Answering, d semantic role labeling

pre-trained models

- BERT-large, RoBERTa-large, DeBERTa-xlarge, GLM-xlarge/xxlarge

multitask learning

- 모든 task 유형의 데이터셋을 합침 → 각 데이터셋에 대해 분리된 선형 분류기 사용

P-tuning v2: Across Scales

	#Size	BoolQ			CB			COPA			MultiRC (F1a)		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
$\text{BERT}_{\text{large}}$	335M	**77.7**	67.2	75.8	**94.6**	80.4	**94.6**	69.0	55.0	**73.0**	70.5	59.6	**70.6**
$\text{RoBERTa}_{\text{large}}$	355M	**86.9**	62.3	84.8	98.2	71.4	**100**	**94.0**	63.0	93.0	**85.7**	59.9	82.5
$\text{GLM}_{\text{xlarge}}$	2B	**88.3**	79.7	87.0	**96.4**	76.4	**96.4**	**93.0**	92.0	91.0	84.1	77.5	**84.4**
$\text{GLM}_{\text{xxlarge}}$	10B	88.7	**88.8**	**88.8**	**98.7**	98.2	96.4	**98.0**	**98.0**	**98.0**	**88.1**	86.1	**88.1**

	#Size	ReCoRD (F1)			RTE			WiC			WSC		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
$\text{BERT}_{\text{large}}$	335M	70.6	44.2	**72.8**	70.4	53.5	**78.3**	74.9	63.0	**75.1**	**68.3**	64.4	**68.3**
$\text{RoBERTa}_{\text{large}}$	355M	89.0	46.3	**89.3**	86.6	58.8	**89.5**	**75.6**	56.9	73.4	63.5	**64.4**	63.5
$\text{GLM}_{\text{xlarge}}$	2B	91.8	82.7	**91.9**	**90.3**	85.6	**90.3**	**74.1**	71.0	72.0	**95.2**	87.5	92.3
$\text{GLM}_{\text{xxlarge}}$	10B	**94.4**	87.8	92.5	**93.1**	89.9	**93.1**	**75.7**	71.8	74.0	**95.2**	94.2	93.3

Table 2: Results on SuperGLUE development set. P-tuning v2 surpasses P-tuning & Lester et al. (2021) on models smaller than 10B, matching the performance of fine-tuning across different model scales. (FT: fine-tuning; PT: Lester et al. (2021) & P-tuning; PT-2: P-tuning v2; **bold**: the best; underline: the second best).

- v2 작은 모델에서도 fine-tuning과 견줄만한 성능 보여줌

⇒ p-tuning v2는 어느 크기의 모델에서도 fine-tuning과 견줄만 함. fine-tuning 대비 0.1%의 파라미터를 사용하면서도

P-tuning v2: Across Tasks

	#Size	CoNLL03				OntoNotes 5.0				CoNLL04			
		FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2
$\text{BERT}_{\text{large}}$	335M	**92.8**	81.9	90.2	91.0	**89.2**	74.6	86.4	86.3	85.6	73.6	84.5	**86.6**
$\text{RoBERTa}_{\text{large}}$	355M	92.6	86.1	**92.8**	**92.8**	**89.8**	80.8	**89.8**	**89.8**	88.8	76.2	88.4	**90.6**
$\text{DeBERTa}_{\text{xlarge}}$	750M	**93.1**	90.2	**93.1**	**93.1**	90.4	85.1	90.4	**90.5**	89.1	82.4	86.5	**90.1**

	#Size	SQuAD 1.1 dev (EM / F1)								SQuAD 2.0 dev (EM / F1)							
		FT		PT		PT-2		MPT-2		FT		PT		PT-2		MPT-2	
$\text{BERT}_{\text{large}}$	335M	**84.2**	**91.1**	1.0	8.5	77.8	86.0	82.3	89.6	**78.7**	**81.9**	50.2	50.2	69.7	73.5	72.7	75.9
$\text{RoBERTa}_{\text{large}}$	355M	**88.9**	**94.6**	1.2	12.0	88.5	94.4	88.0	94.1	**86.5**	**89.4**	50.2	50.2	82.1	85.5	83.4	86.7
$\text{DeBERTa}_{\text{xlarge}}$	750M	90.1	95.5	2.4	19.0	**90.4**	**95.7**	89.6	95.4	88.3	91.1	50.2	50.2	**88.4**	**91.1**	88.1	90.8

	#Size	CoNLL12				CoNLL05 WSJ				CoNLL05 Brown			
		FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2
$\text{BERT}_{\text{large}}$	335M	84.9	64.5.	83.2	**85.1**	**88.5**	76.0	86.3	**88.5**	82.7	70.0	80.7	**83.1**
$\text{RoBERTa}_{\text{large}}$	355M	**86.5**	67.2	84.6	86.2	**90.2**	76.8	89.2	90.0	85.6	70.7	84.3	**85.7**
$\text{DeBERTa}_{\text{xlarge}}$	750M	86.5	74.1	85.7	**87.1**	**91.2**	82.3	90.6	**91.2**	86.9	77.7	86.3	**87.0**

Table 3: Results on Named Entity Recognition (NER), Question Answering (Extractive QA), and Semantic Role Labeling (SRL). All metrics in NER and SRL are micro-f1 score. (FT: fine-tuning; PT: P-tuning & Lester et al. (2021); PT-2: P-tuning v2; MPT-2: Multi-task P-tuning v2; **bold**: the best; underline: the second best).

- v2 모든 task에 대하여 fine-tuning과 견줄만한 성능 보여줌

Ablation Study

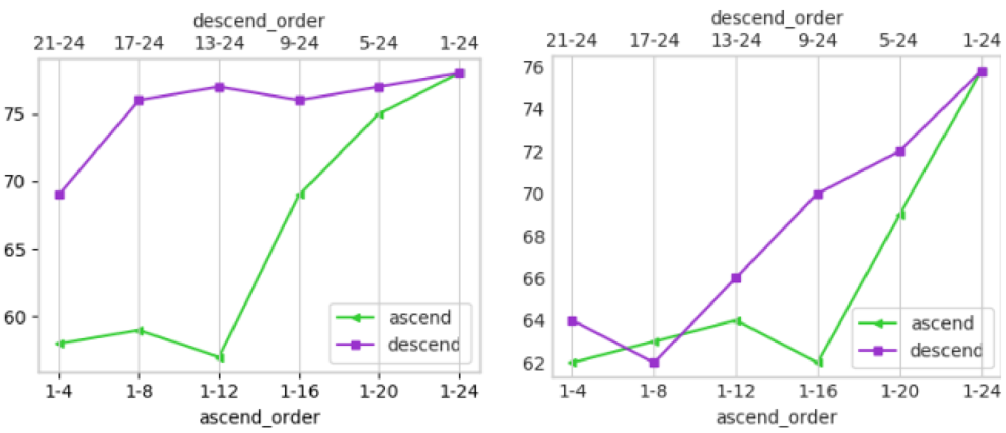
Verbalizer with LM head v.s. [CLS] label with linear head

	SST-2	RTE	BoolQ	CB
CLS & linear head	96.3	88.4	84.8	96.4
Verbalizer & LM head	95.8	86.6	84.6	94.6

Table 4: Comparison between [CLS] label with linear head and verbalizer with LM head on RoBERTa-large.

prompt depth

- multi-layer continuous prompt의 효과를 확인하기 위해 k개의 레이어에 프롬프트를 적용함
- 감소하는 순서로 적용하는 것이 언제나 더 좋은 결과



(a) RTE (b) BoolQ

Conclusion

contribution: 다양한 크기의 모델과 task에서 p-tuning이 fine-tuning과 견줄만한 결과를 보여줄 수 있음을 입증. parameter efficiency는 확연히 더 좋음