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

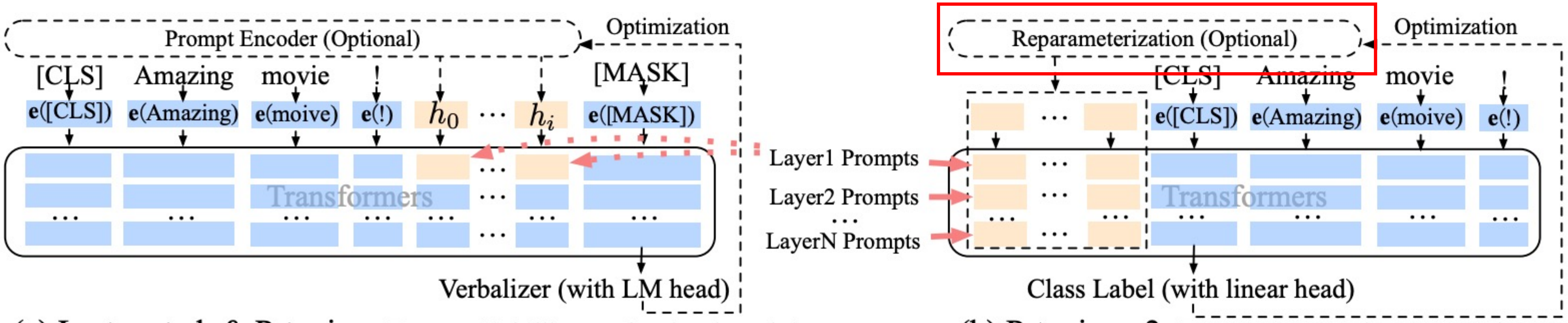
백인진 22.09.14

Introduction

- To present a structure to overcome the disadvantages of P-tuning v1
  - If the size of the model is less than 10B, the performance is lower than finetuning.
  - Low performance for relatively challenging tasks such as Sequence Labeling.

- It presents an optimized prompt engineering structure regardless of the size of the model or the work to be performed.
  - Deep Prompt Tuning
  - Continuous Prompt on all layers
    - In P-tuning V1, Continuous Prompt on a layer
  - Save cost with learning parameters ranging from 0.1% to 3% compared to Finetuning method

- P-tuning v2



(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks) (b) P-tuning v2 (Frozen, most scales, most tasks)

- P-tuning v1
  - Not enough parameters to train.
  - Sequence length constraint exists.
  - The effect of input embedding is relatively indirect.

- P-tuning v2
  - Create a prompt for each layer
    - To create more parameters to train
  - Increase from 0.01% to 0.1% ~ 3% of all model parameters

Architect

- **Reparameterization**
  - Use optionally according to the task.
- **Prompt Length**
  - Based on existing experimental results, the prompt length that produces good performance depends on the task
  - Use shorter prompt length for relatively easy tasks and longer prompt length for difficult tasks
- **Multi-task Learning**
  - To improve performance by learning various tasks at once.
- **Classification Head**
  - Remove the verbalizer that was used primarily in P-tuning v1
  - Attach the randomly-initialized classification head

Experiments

	#Size	BoolQ			CB			COPA			MultiRC (F1a)		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
$BERT_{large}$	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**
$RoBERTa_{large}$	355M	**86.9**	62.3	84.8	98.2	71.4	**100**	**94.0**	63.0	93.0	**85.7**	59.9	82.5
GLM_{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**
$GLM_{xxlarge}$	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
$BERT_{large}$	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**
$RoBERTa_{large}$	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
GLM_{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
$GLM_{xxlarge}$	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).

- Unlike V1, V2 shows similar performance to finetuning in the smaller scale model
    - At this time, the number of parameters is about 0.1% compared to the Finetuning method

Experiments

NER >>

	#Size	CoNLL03				OntoNotes 5.0				CoNLL04			
		FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2
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**
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**
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**

QA >>

	#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	
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
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
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

SRL >>

	#Size	CoNLL12				CoNLL05 WSJ				CoNLL05 Brown			
		FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2
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**
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**
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).

- Good performance on QA Task, which was low performance on traditional V1.
- Even when multi-task learning is used, it shows good performance except for QA tasks.

References

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

https://arxiv.org/abs/2110.07602