GPT understands, too.

백인진 22.08.31

Introduction

- Language Models
  - Uni-directional : GPT style
  - Bi-directional : BERT style
  - Hybrid : XLNet, UniLM, etc.

- Drawbacks of GPT-style
  - Low performance in NLU task
  - Difficult prompt engineering process s

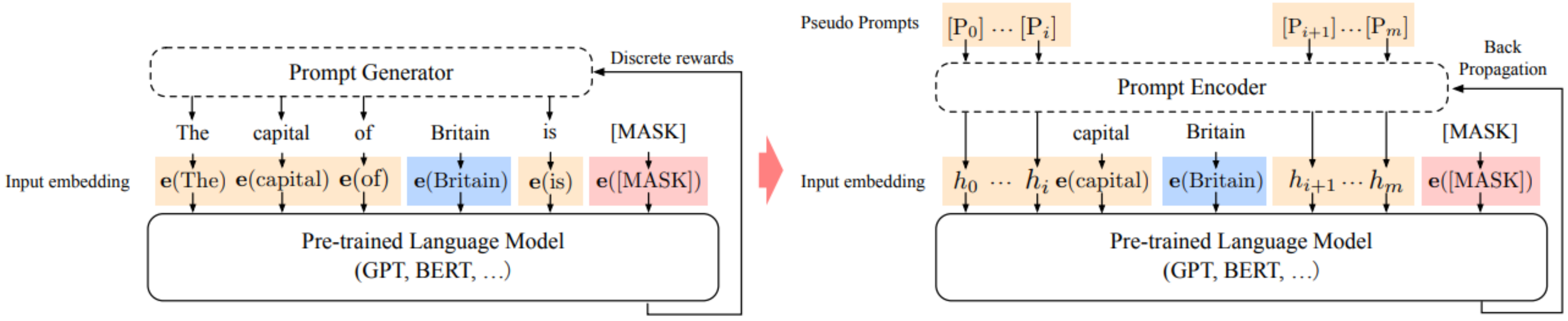
Motivation

- Giant models : suffer from poor transfer-ability
  - Too large to finetune
- Handcraft prompt searching : performance is volatile
  - Overfitting for test dataset
  - easy to create advertising prompts that result in significant degradation
- Recent works : automating the search of discrete prompts
  - Since neural networks are inherently continuous, discrete prompts can be sub-optimal.

>> finding continuous prompts that can be differentially optimized

Methods

- **P-tuning**
  - A new way to automatically search for prompts in continuous space
  - Template : $\{[P_{0:i}], x, [P_{i+1:m}], y\}$
    - x : *input sequence*, y : *target token*



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

- Traditional discrete prompts
  - $\{e([P_{0:i}]), e(x), e([P_{i+1:m}]), e(y)\}$
    - $[P_i] \in V$
- P-tuning
  - $\{h_0, \ldots, h_i, e(x), h_{i+1}, \ldots, h_m, e(y)\}$
    - h_i: trainable embedding tensors

Optimization

- Discreteness
  - h:initialized with random distribution & optimized with SGD
- Association
  - The values of prompt embeddings h_i : should be dependent on each other

$$h_i = MLP([\overrightarrow{h_i} : \overleftarrow{h_i}]) = MLP([LSTM(h_{0:i}) : LSTM(h_{i:m})])$$

  - The use of LSTM add some parameters, but added size is several times smaller than PLM.
  - Only output embedding h is required in inference, and LSTM may be discarded.
  - Can use anchor token like '?' for specific tasks.

Result

- Knowledge probing

Prompt type	Model	P@1
Original (MP)	BERT-base	31.1
	BERT-large	32.3
	E-BERT	36.2
Discrete	LPAQA (BERT-base)	34.1
	LPAQA (BERT-large)	39.4
	AutoPrompt (BERT-base)	43.3
P-tuning	BERT-base	48.3
	BERT-large	**50.6**

Model	MP	FT	MP+FT	P-tuning
BERT-base (109M)	31.7	51.6	52.1	52.3 (+20.6)
-AutoPrompt (Shin et al., 2020)	-	-	-	45.2
BERT-large (335M)	33.5	54.0	55.0	54.6 (+21.1)
RoBERTa-base (125M)	18.4	49.2	50.0	49.3 (+30.9)
-AutoPrompt (Shin et al., 2020)	-	-	-	40.0
RoBERTa-large (355M)	22.1	52.3	52.4	53.5 (+31.4)
GPT2-medium (345M)	20.3	41.9	38.2	46.5 (+26.2)
GPT2-xl (1.5B)	22.8	44.9	46.5	54.4 (+31.6)
MegatronLM (11B)	23.1	OOM*	OOM*	**64.2** (+41.1)

* MegatronLM (11B) is too large for effective fine-tuning.

  - Manual prompt < Discrete prompt < P-tuning

- SuperGLUE

Method	BoolQ (Acc.)	CB (Acc.)	CB (F1)	WiC (Acc.)	RTE (Acc.)	MultiRC (EM)	MultiRC (F1a)	WSC (Acc.)	COPA (Acc.)	Avg.
BERT-base-cased (109M)										
Fine-tuning	72.9	85.1	73.9	71.1	68.4	16.2	66.3	63.5	67.0	66.2
MP zero-shot	59.1	41.1	19.4	49.8	54.5	0.4	0.9	62.5	65.0	46.0
MP fine-tuning	73.7	87.5	90.8	67.9	70.4	13.7	62.5	60.6	70.0	67.1
P-tuning	73.9	89.2	92.1	68.8	71.1	14.8	63.3	63.5	72.0	68.4
GPT2-base (117M)										
Fine-tune	71.2	78.6	55.8	65.5	67.8	17.4	65.8	63.0	64.4	63.0
MP zero-shot	61.3	44.6	33.3	54.1	49.5	2.2	23.8	62.5	58.0	48.2
MP fine-tuning	74.8	87.5	88.1	68.0	70.0	23.5	69.7	66.3	78.0	70.2
P-tuning	75.0 (+1.1)	91.1 (+1.9)	93.2 (+1.1)	68.3 (-2.8)	70.8 (-0.3)	23.5 (+7.3)	69.8 (+3.5)	63.5 (+0.0)	76.0 (+4.0)	70.4 (+2.0)

  - Finetuned < P-tuning
  - BERT < GPT

Conclusion

- Contributions
  - New methods : P-tuning
    - Augmenting pretrained model's ability in NLU by automatically searching better prompts in the continuous space
    - Relying less on a large validation dataset
    - Suffering less from adversarial prompts
    - Alleviating over-fitting
    - Also, helping bi-directional models

References

- GPT Understands, Too.

https://arxiv.org/abs/2103.10385