Artficial Intelligence Laboratory

The Power of Scale for Parameter-Efficient Prompt Tuning

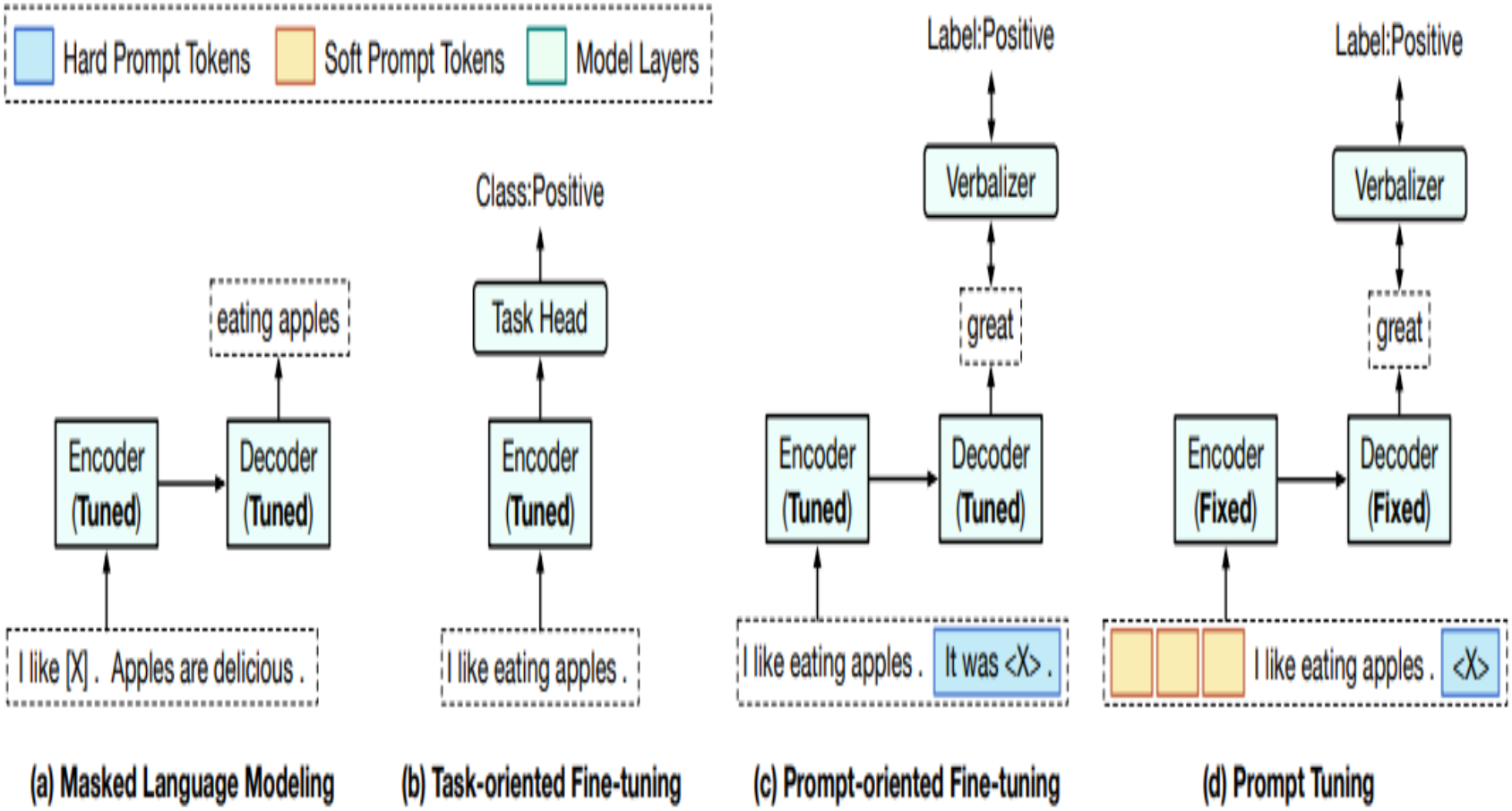
Google Research

정보융합공학과 AI전공
정주경

부산대학교
PUSAN NATIONAL UNIVERSITY

목 차

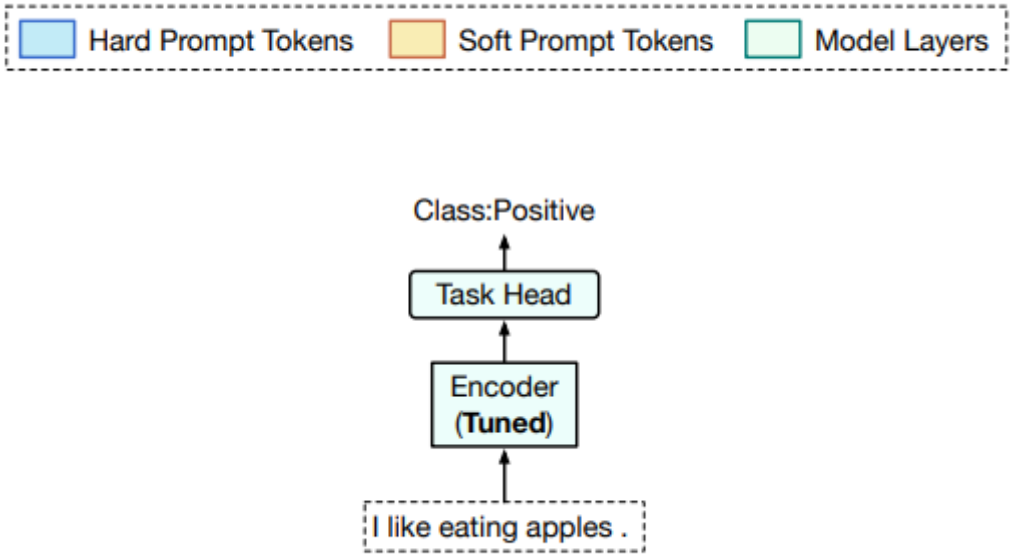
Introduction: Pre-training의 paradigms



(a) Masked Language Modeling (b) Task-oriented Fine-tuning (c) Prompt-oriented Fine-tuning (d) Prompt Tuning

Introduction: Pre-training의 paradigms

FT 방법의 두 가지 mainstream (1. Task-oriented Fine-tuning)

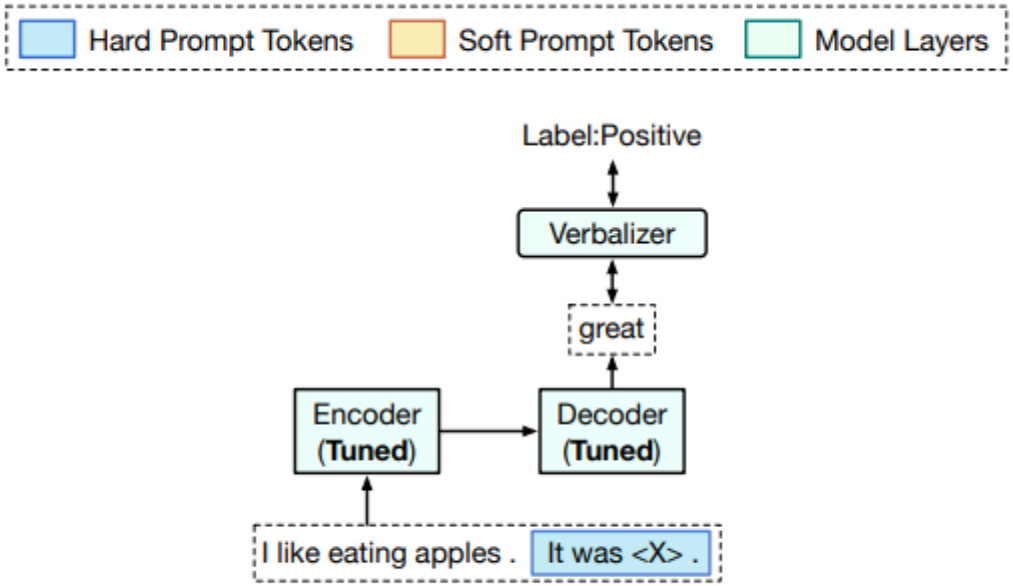


(b) Task-oriented Fine-tuning

- **Task-oriented Fine-Tuning(Full-model tuning) : PLM위에 task-specific head가 추가된 다음 training data에서 task-specific objective를 최적화하여 전체 모델 fine-tuned**
- **Ex) BERT**

Introduction: Pre-training의 paradigms

- FT 방법의 두 가지 mainstream (2. Prompt-oriented Fine-tuning



(c) Prompt-oriented Fine-tuning

- Prompt-oriented Fine-tuning: PLM의 지식을 조사하기 위해 prompt를 사용하는 GPT-3에서 영감을 얻음 ex) PET
- 데이터 샘플은 prompt token을 포함하는 sequence로 변환
- "It was <X>" prompt를 문장에 추가, mask 위치에 "great" or "terrible"를 예측
- Task-oriented fine-tuning과 비교하면 pre-training objective(MLM)와 비슷해서 PLM에서 지식을 더 잘 사용하고 더 나은 성능

Introduction

Model Tuning

- 대규모 pre-trained language model은 많은 NLP benchmark에서 SOTA 결과 달성
- GPT와 BERT 이후, 표준 관행은 네트워크의 모든 가중치 조정 (model tuning)과 함께 downstream task에서 모델을 fine-tuning 하는 것

⇒ 모델이 계속 커짐에 따라 downstream task마다 튜닝 된 모델의 복사본 저장과 제공이 어려워짐

대안 : Frozen Pre-trained Model

- 모든 downstream task에서 가중치가 고정되는 single frozen pre-trained language model을 공유
- GPT-3는 frozen model이 “in-context” 학습을 통해 다른 작업을 수행하도록 조건화 되는 것을 보여줌

Prompt

Frozen model방식을 통해 특정 task의 모델에 대해 prompt design

⇒ 즉, task에 대한 설명 또는 예를 포함한 text prompt를 hand-crafting으로 만듬

Ex) “이 영화는 놀라워“ input sequence 앞에 “다음 영화 리뷰는 긍정적인가 부정적인가?”라는 prompt를 붙일 수 있음

Introduction

Prompt

- 동일한 frozen model을 task간 공유 => 높은 비용 문제 해결과 효율적인 task 추론이 가능하지만 성능이 떨어짐
- Text prompt는 수동 설계가 필요하며, 잘 설계된 prompt도 model tuning에 비해 성능이 떨어짐
- Ex) frozen GPT-3 175B parameter 모델의 SuperGLUE benchmark 성능은 16배 적은 parameter를 사용하는 fine-tuned T5-XXL(11B)모델이 비해 17.5점 낮음

8	T5 Team - Google	T5		89.3
22	Ben Mann	GPT-3 few-shot - OpenAI		71.8

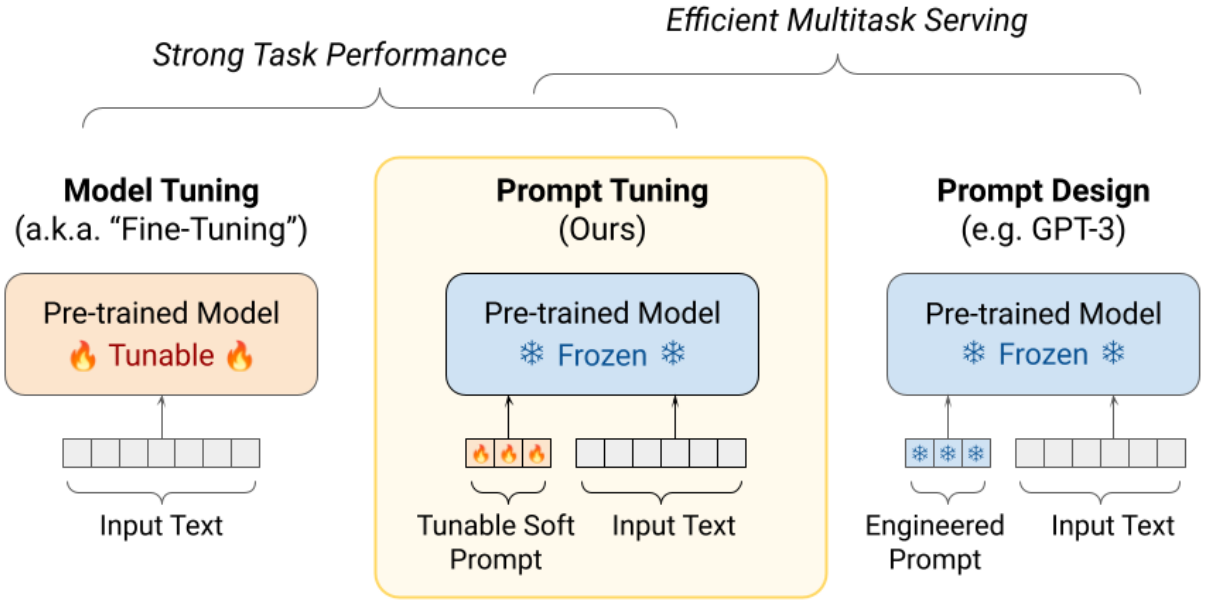
Prompt tuning

Soft prompt

- 조정 가능한 soft prompt를 사용하여 frozen model을 조절하기 위한 효과적인 방법
- engineered text prompt와 마찬가지로, soft prompt는 입력 텍스트에 연결
- 기존 vocabulary 항목에서 선택X, soft prompt의 "token"은 학습 가능한 벡터

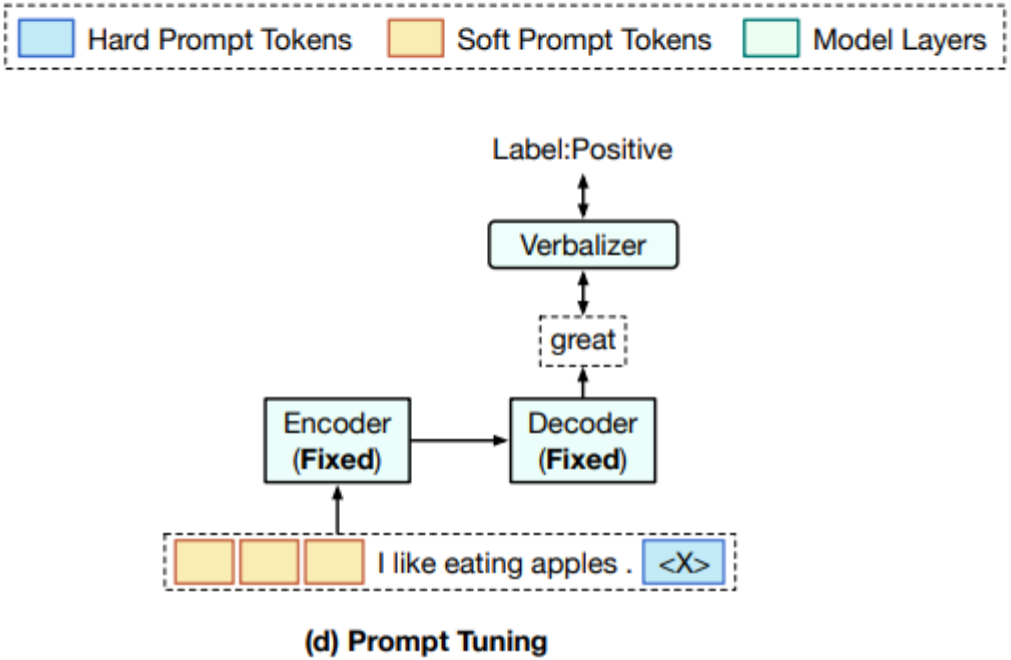
=> training dataset에 대해 soft prompt를 end-to-end로 최적화 할 수 있음을 의미

- 수동 설계 제거를 통해 prompt는 수백만 개의 예제를 포함하는 데이터셋 정보 압축 가능
- 이에 비해 engineered text prompt는 모델 입력 길이의 제약으로 인해 일반적으로 50개 미만의 예제로 제한



Prompt tuning

Soft prompt



(d) Prompt Tuning

- soft prompt를 만들려면 prompt를 고정된 길이 벡터 시퀀스(예: 20 토큰 길이)로 초기화
- 이 벡터를 각 embedded input 시작 부분에 넣고 이러한 시퀀스를 모델에 준다
- 모델의 예측을 target과 비교하여 loss계산, error는 back-propagation하여 gradients를 계산하지만 gradient update는 새로운 학습가능한 벡터에만 적용하며 core model은 frozen을 유지

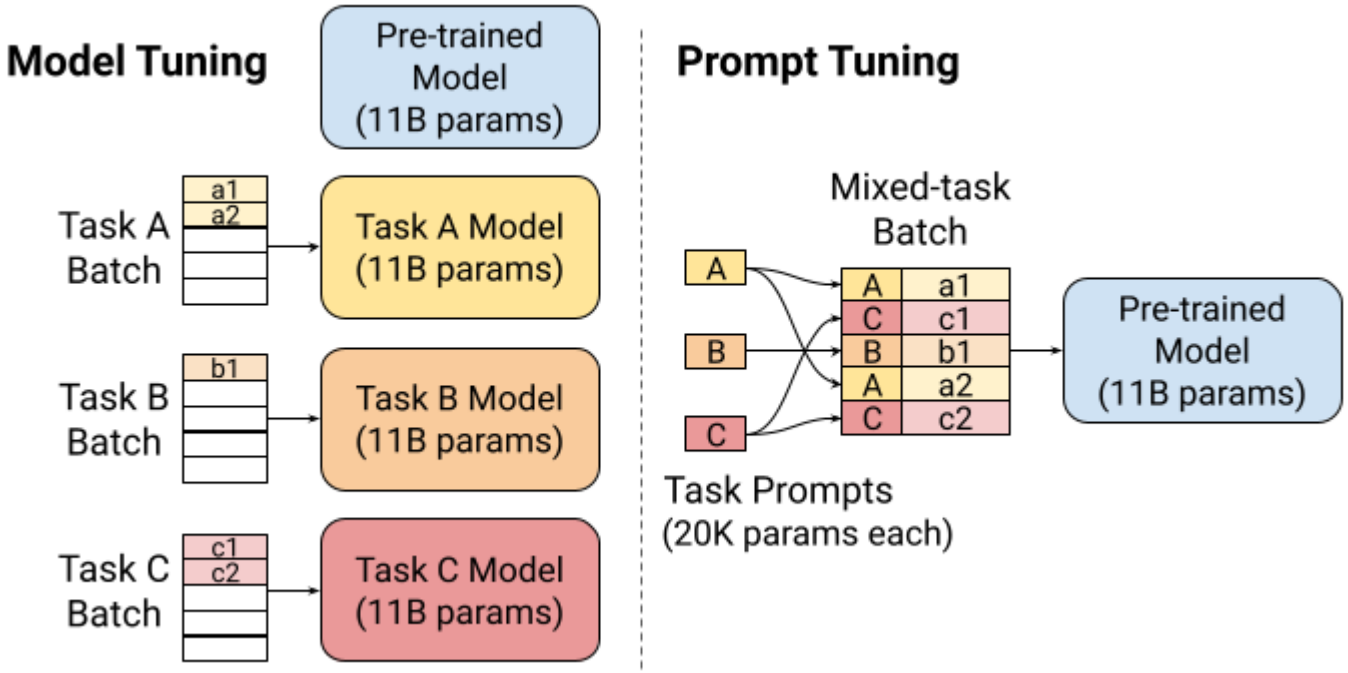
Prompt tuning

Soft prompt

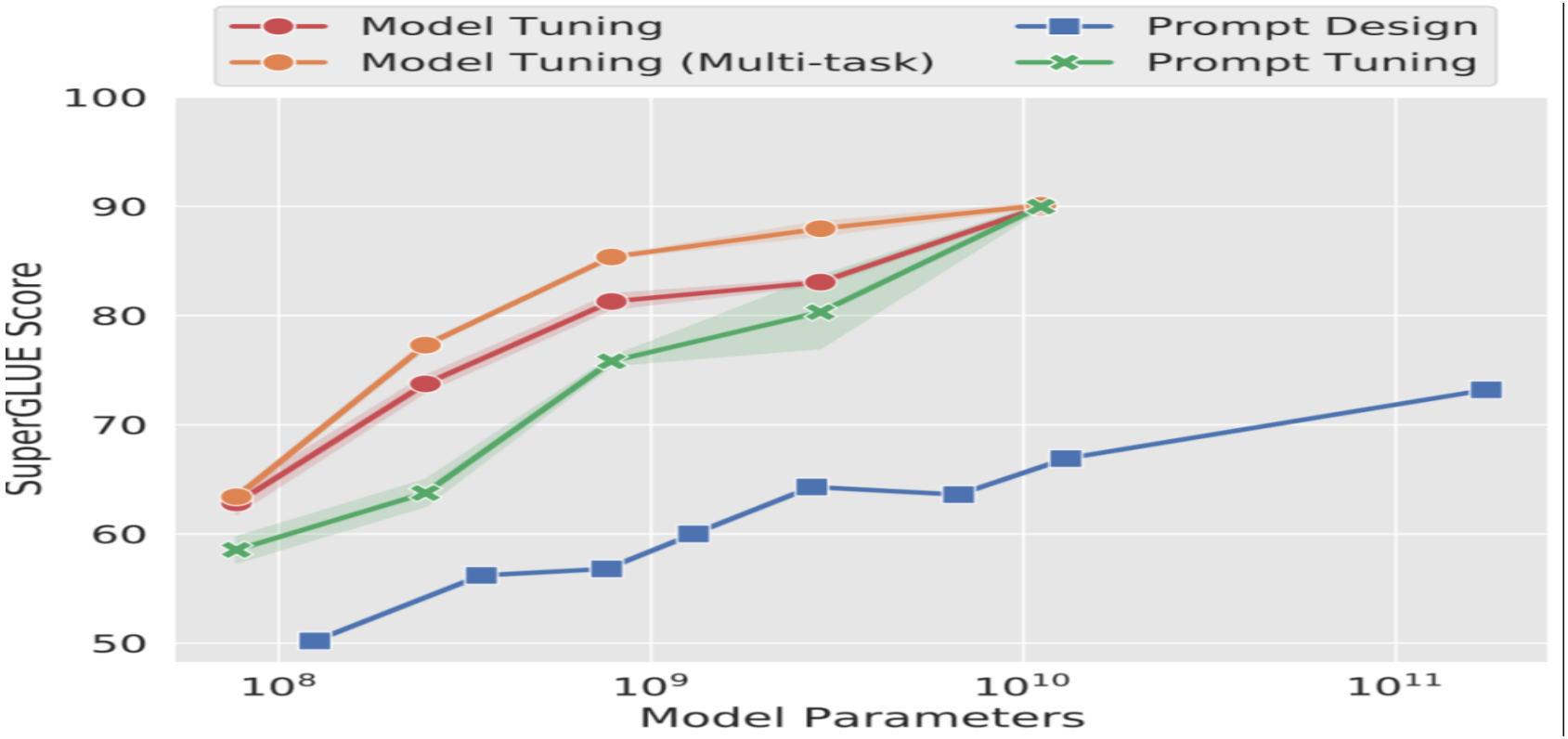
Parameter 공간이 작기 때문에 각 입력 예제와 함께 다른 prompt를 모델에 쉽게 전달

이를 통해 mixed-task 추론 배치가 가능하며, 이는 많은 작업에서 하나의 핵심 모델을 공유하여 서비스를 간소화

T5 XXL(11B) 기준



Improvement with Scale



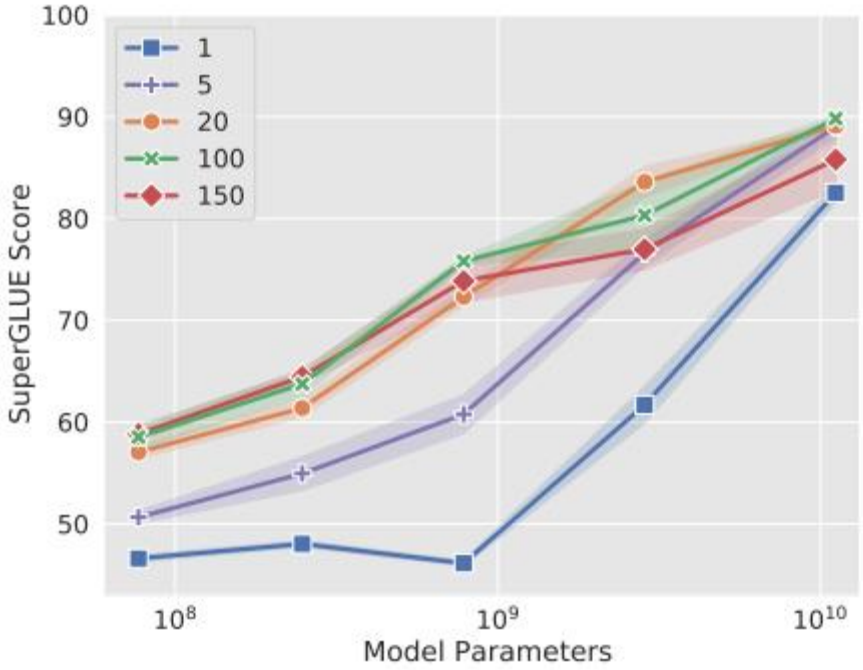
- Prompt tuning은 모든 작업에 대해 single frozen model을 재사용
- Frozen T5 model 사용해 SuperGLUE에서 prompt tuning은 prompt design을 크게 능가
- 규모가 커짐에 따라 parameter가 적음에도 prompt tuning이 model tuning과 일치

=> 모델 크기가 계속 증가함에 따라 pre-trained model을 “freezing”하는 것은 좋은 대안

=> 대형 모델에서 개별 복사본을 제공하면 큰 계산 overhead가 발생할 수 있어 대규모 모델에서의 prompt tuning 효과는 특히 중요

Ablation Study

Prompt length

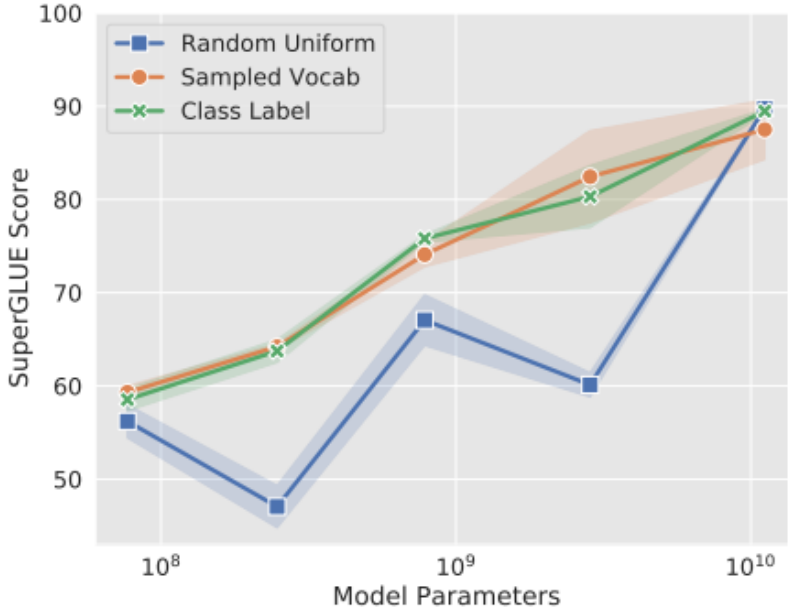


(a) Prompt length

- {1, 5, 20, 100, 150}의 prompt 길이를 변경해 각 모델 크기에 대한 prompt 학습
- 우수한 성능을 얻으려면 single token을 넘어 prompt길이를 늘리는 것이 매우 중요
- 20개 이상의 토큰으로 증가하면 성능 향상, XXL은 single token prompt에서도 우수한 성능
- 100개 토큰 초과는 성능이 저하되는 패턴 관찰

Ablation Study

Prompt initialization



(b) Prompt initialization

- **Random : 랜덤하게 표본 추출**
- **Sampled Vocab: pre-training corpus의 likelihood로 정렬된 T5의 Sentence Piece vocabulary에서 가장 "일반적인" 토큰 5000개로 제한**
- **Class Label: downstream task에서 각 클래스의 sentence representation에 대한 임베딩을 가져와 prompt에서 token중 하나를 초기화**
- **Random uniform initialization는 "advanced" initialization(sampled vocab, class label)에 뒤쳐지지만 XXL크기에서는 차이가 사라짐**

Ablation Study

T5

1. encoder-decoder 구조 사용
2. Span-corruption objective로 학습됨 (masked span reconstruction)

Ex) Thank you for inviting me to your party last week

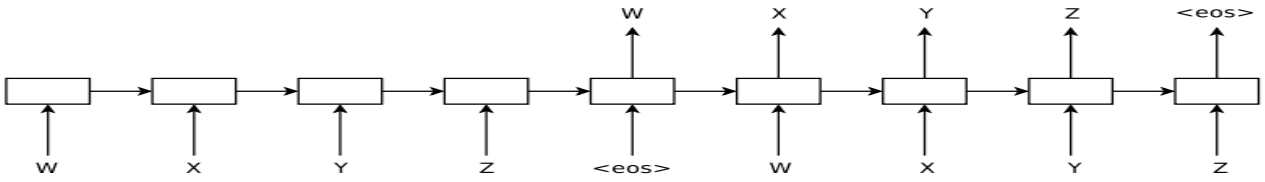
Input : Thank you <X> me to your party <Y> week

Output: <X> for inviting <Y> last <Z>

=> Pre-trained T5는 natural input text(sentinel token이 없는)를 본 적이 없어 single frozen model을 만들기에는 적합하지 않음

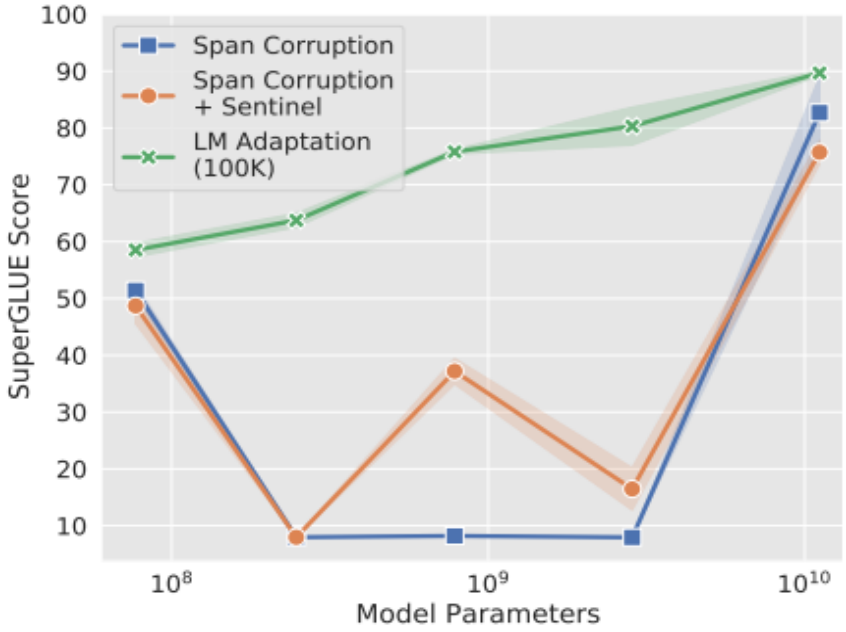
Pre-training method

1. Span Corruption : pre-trained T5를 frozen model로 사용
2. Span Corruption + Sentinel : pre-trained T5 + 모든 downstream task에 sentinel을 추가
3. LM Adaptation : T5의 self-supervised training은 계속하지만, 입력으로 natural text prefix 가 주어지면 모델은 출력으로 natural text를 연속으로 생성하는 “LM objective”를 사용



Ablation Study

Pre-training method

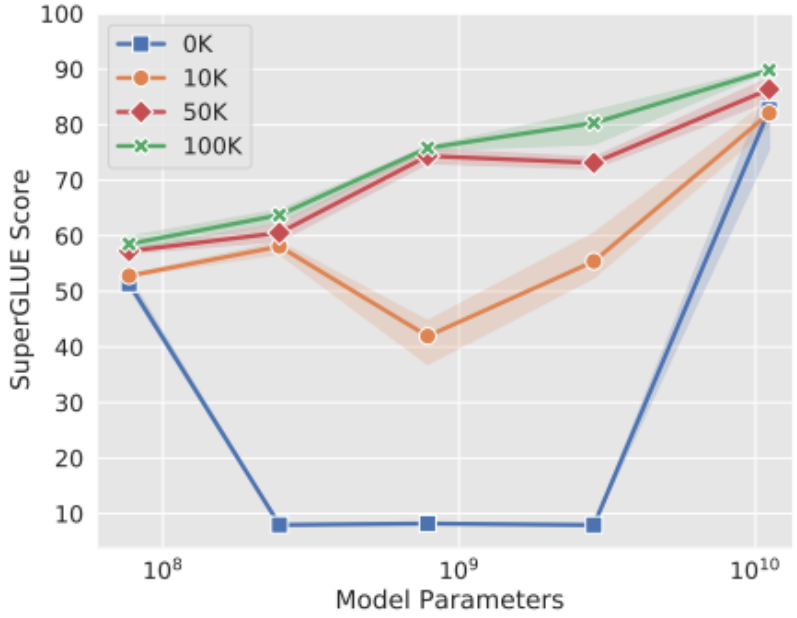


(c) Pre-training method

- sentinel이 downstream task에 추가되더라도 LM adaptation 성능이 더 우수
- XXL에서는 모든 방법이 잘 작동

Ablation Study

LM adaptation steps



(d) LM adaptation steps

- Adaptation이 길수록 일반적으로 성능이 좋지만 XXL은 짧은 adaption에도 robust
- 이는 span corruption -> LM objective로의 “transition” 이러한 변화가 중요

Resilience to Domain Shift

Model tuning: 네트워크의 모든 가중치에 영향을 미쳐 fine-tuning 데이터에 쉽게 overfit, 추론 시 다양한 작업에서 잘 일반화되지 못할 수 있음

Learned Soft prompt: parameter의 수가 적기 때문에 잘 일반화될 수 있음

zero-shot performance 1

F1 average stddev

Dataset	Domain	Model	Prompt	Δ
SQuAD	Wiki	94.9 ±0.2	94.8 ±0.1	−0.1
TextbookQA	Book	54.3 ±3.7	**66.8** ±2.9	+12.5
BioASQ	Bio	77.9 ±0.4	**79.1** ±0.3	+1.2
RACE	Exam	59.8 ±0.6	**60.7** ±0.5	+0.9
RE	Wiki	88.4 ±0.1	**88.8** ±0.2	+0.4
DuoRC	Movie	**68.9** ±0.7	67.7 ±1.1	−1.2
DROP	Wiki	**68.9** ±1.7	67.1 ±1.9	−1.8

- **SQuAD에서 학습되고 MRQA 2019 shared task(Evaluating Generalization in Reading Comprehension)의 도메인 외 데이터셋에서 평가된 모델의 F1 평균 및 stddev**

- Prompt tuning은 TextbookQA 같은 도메인 이동이 큰 데이터셋에서 model tuning보다 zero-shot 성능이 좋음

Resilience to Domain Shift

zero-shot performance 2

Train	Eval	Tuning	Accuracy	F1
QQP	MRPC	Model	73.1 ±0.9	81.2 ±2.1
		Prompt	**76.3** ±0.1	**84.3** ±0.3
MRPC	QQP	Model	74.9 ±1.3	**70.9** ±1.2
		Prompt	**75.4** ±0.8	69.7 ±0.3

- GLUE의 두 가지 paraphrase detection task간의 transfer
- QQP(커뮤니티 Q&A 사이트 Quora에서 두 가지 질문이 중복인지), MRPC(뉴스 기사에서 도출된 두 문장이 paraphrases인지)로 테스트
- 이전과 마찬가지로 “in-domain”작업에 대해 학습, “out-of-domain”작업에 대해 zero-shot 평가
- QQP 데이터에서 학습과 MRPC에서 평가가 prompt tuning이 model tuning보다 성능이 좋음 (정확도+3.2, F1 +3.1)
- 반대의 경우 prompt tuning의 정확도는 model tuning보다 약간 좋지만 F1 약간 성능저하

=> Model tuning이 over-parameterized되고 training task에 overfit

Prompt Ensembling

Neural model ensemble

- 신경 모델의 앙상블은 작업 성능 향상과 모델 불확실성을 추정하는 데 유용
- N개의 모델 저장하는 데 필요한 공간과 N개의 개별 모델을 실행하는 데 상당한 비용 발생

Prompt ensemble

- Prompt tuning은 pre-trained 언어 모델의 여러 adaptation을 합치는 것보다 효율적인 방법 제공
- 동일한 task에 대해 N개의 prompt 학습, parameter를 공유하면서 한 task에 대해 N개의 개별 "model" 만듬
- Prompt ensemble은 storage 비용 절감과 추론을 효율적으로 만듬

Prompt Ensembling

Prompt ensemble

Dataset	Metric	Average	Best	Ensemble
BoolQ	acc.	91.1	91.3	**91.7**
CB	acc./F1	99.3 / 99.0	100.00 / 100.00	**100.0 / 100.0**
COPA	acc.	98.8	100.0	**100.0**
MultiRC	$EM/F1_a$	65.7 / 88.7	66.3 / 89.0	**67.1 / 89.4**
ReCoRD	EM/F1	92.7 / 93.4	92.9 / 93.5	**93.2 / 93.9**
RTE	acc.	92.6	**93.5**	**93.5**
WiC	acc.	76.2	76.6	**77.4**
WSC	acc.	95.8	**96.2**	**96.2**
SuperGLUE (dev)		90.5	91.0	**91.3**

- Single frozen T5-XXL 모델 사용하여 SuperGLUE task에 대해 5개의 prompt를 학습
- Ensemble에서 예측을 계산하기 위해 다수결로 정함

- 모든 작업에서 앙상블이 단일 prompt 평균을 능가, 최상의 개별 prompt를 능가하거나 일치함을 보여줌

Conclusion

- Prompt tuning이 GPT-3의 few-shot learning 성능을 크게 능가

- T5를 사용한 모델 크기의 축소를 통해 Prompt tuning이 규모에 있어서 경쟁력 있음을 보여줌

- 대형 모델은 공유 및 서비스 비용이 많이 들고 하나의 frozen model을 여러 downstream task에 재사용할 수 있기 때문에 이러한 결과는 특히 중요

- SuperGLUE benchmark에서의 성능은 model tuning에 필적하며 모델 크기가 증가에 따라 격차가 사라짐

- Zero-shot domain transfer에서 prompt tuning은 일반화가 잘 되는 것을 알 수 있음

- 일반화가 잘 되기 때문에 prompt ensembling이 가능

앞으로 task-defining parameter를 일반 언어 모델링 parameter와 구별하는 것이 새로운 연구를 위한 많은 길을 열어주는 흥미로운 단계라고 믿습니다.