

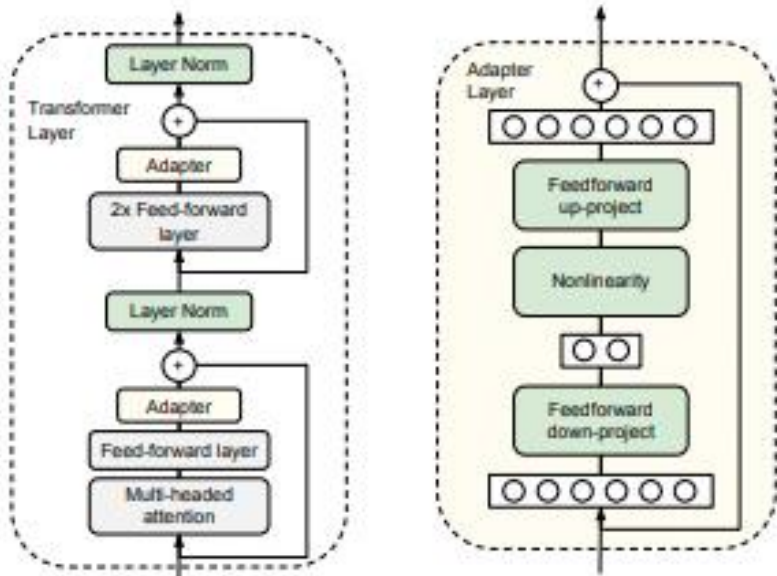
LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* ([2021](#)).

Background

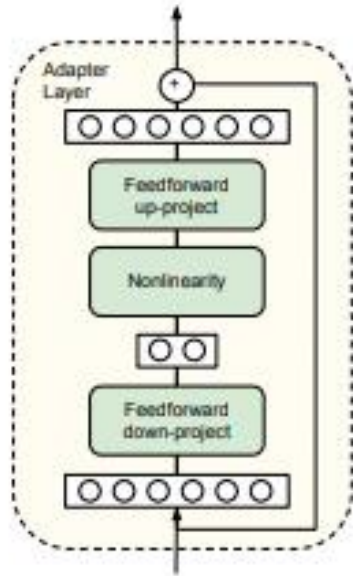
다운스트림 태스크는 일반적으로 사전 학습된 모델의 모든 파라미터를 업데이트하는 fine-tuning을 통해 수행된다.

추가되는 어댑터에 대해서만 학습
→ 어댑터 레이어들은 latency가 증가한다



prompt 최적화

adaptation을 위해 시퀀스 길이의 일부를 사용하면 다운스트림 task를 처리하는 데 사용할 수 있는 시퀀스 길이가 필연적으로 줄어들어 다른 방법에 비해 프롬프트 튜닝 성능이 떨어진다.



LLM Tuning Methods

Prompt Engineering

PEFT
(Parameter Efficient Fine-Tuning)

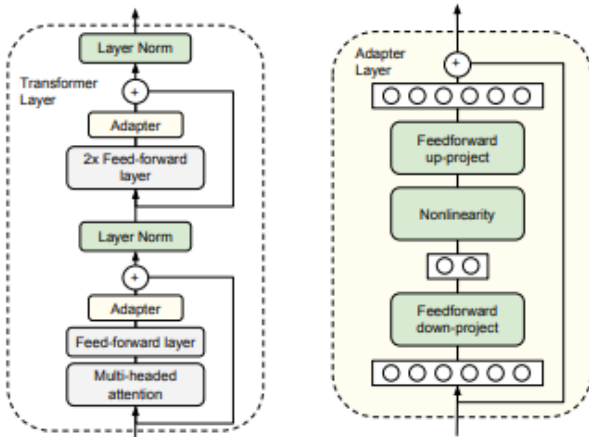
RAG
(Retrieval Augmented Generation)

Full Fine-Tuning

Cost ~ Complexity ~ Quality

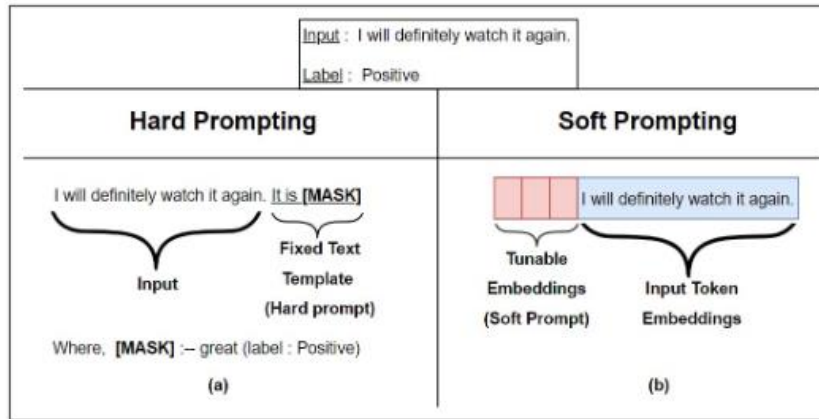
1. Adapter Layers

pre-trained model 사이사이에
feed-forward networks(학습 가능한
작은 신경망 층)를 삽입



2. Prompt Tuning

프롬프트의 일부를 학습 가능한 매개변수로 변환하여
훈련할 수 있도록 한 것 (입력 프롬프트에 대하여)



Hard Prompt vs. Soft Prompt (Senadeera & Iye, 2022)

3. LoRA / QLoRA (Low-Rank Adaptation)

pre-trained model의 원래 가중치는 유지하면서 학습 가능한 저차원
행렬인 lank decomposition 행렬을 삽입하여 소수의 파라미터만 조
정하는 기법

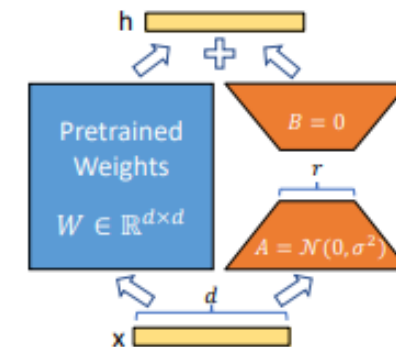


Figure 1: Our reparametrization. We only train A and B .

LoRA

행렬에서의 Rank

: 행렬에서 선형적으로 독립적인 벡터의 개수

랭크가 낮을 수록 행렬의 정보가 중복되거나 불필요하다는 것을 의미
즉, 랭크가 낮다면 행렬의 차원을 줄일 수 있다

LoRA(Low-Rank Adaptation)

모델의 파라미터를 낮은 랭크로 근사하여, 원래 고차원 공간에서의 변화를 낮은 랭크의 행렬로 표현해서 비슷한 결과를 낼 수 있도록 하는 것

Fine tuning과정에서 원래 모델의 파라미터는 freeze,
low-rank decomposition 행렬 쌍을 추가로 주입하여 훈련

$$h = W_0x + \Delta Wx = W_0x + BAx$$

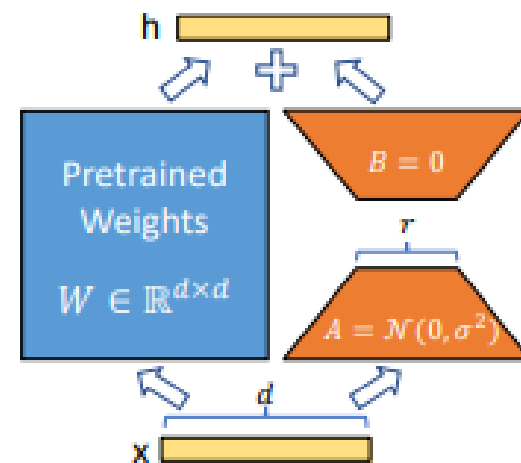


Figure 1: Our reparametrization. We only train A and B .

Applying LoRA to Transformer

학습 가능한 파라미터의 수를 줄이기 위해
신경망에서 **가중치 행렬의 모든 부분집합에 LoRA를 적용할 수 있다.**

	# of Trainable Parameters = 18M						
Weight Type Rank r	W_q 8	W_k 8	W_v 8	W_o 8	W_q, W_k 4	W_q, W_v 4	W_q, W_k, W_v, W_o 2
WikiSQL ($\pm 0.5\%$)	70.4	70.0	73.0	73.2	71.4	73.7	73.7
MultiNLI ($\pm 0.1\%$)	91.0	90.8	91.0	91.3	91.3	91.3	91.7

LoRA 기법을 사용해 적은 수의 파라미터로도 높은 성능을 유지할 수 있음을 시사

Result

Model & Method	# Trainable Parameters	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	Avg.
RoB _{base} (FT)*	125.0M	87.6	94.8	90.2	63.6	92.8	91.9	78.7	91.2	86.4
RoB _{base} (BitFit)*	0.1M	84.7	93.7	92.7	62.0	91.8	84.0	81.5	90.8	85.2
RoB _{base} (Adpt ^D)*	0.3M	87.1 \pm .0	94.2 \pm .1	88.5 \pm 1.1	60.8 \pm .4	93.1 \pm .1	90.2 \pm .0	71.5 \pm 2.7	89.7 \pm .3	84.4
RoB _{base} (Adpt ^D)*	0.9M	87.3 \pm .1	94.7 \pm .3	88.4 \pm .1	62.6 \pm .9	93.0 \pm .2	90.6 \pm .0	75.9 \pm 2.2	90.3 \pm .1	85.4
RoB _{base} (LoRA)	0.3M	87.5 \pm .3	95.1\pm.2	89.7 \pm .7	63.4 \pm 1.2	93.3\pm.3	90.8 \pm .1	86.6\pm.7	91.5\pm.2	87.2
RoB _{large} (FT)*	355.0M	90.2	96.4	90.9	68.0	94.7	92.2	86.6	92.4	88.9
RoB _{large} (LoRA)	0.8M	90.6\pm.2	96.2 \pm .5	90.9\pm1.2	68.2\pm1.9	94.9\pm.3	91.6 \pm .1	87.4\pm2.5	92.6\pm.2	89.0
RoB _{large} (Adpt ^P)†	3.0M	90.2 \pm .3	96.1 \pm .3	90.2 \pm .7	68.3\pm1.0	94.8\pm.2	91.9\pm.1	83.8 \pm 2.9	92.1 \pm .7	88.4
RoB _{large} (Adpt ^P)†	0.8M	90.5\pm.3	96.6\pm.2	89.7 \pm 1.2	67.8 \pm 2.5	94.8\pm.3	91.7 \pm .2	80.1 \pm 2.9	91.9 \pm .4	87.9
RoB _{large} (Adpt ^H)†	6.0M	89.9 \pm .5	96.2 \pm .3	88.7 \pm 2.9	66.5 \pm 4.4	94.7 \pm .2	92.1 \pm .1	83.4 \pm 1.1	91.0 \pm 1.7	87.8
RoB _{large} (Adpt ^H)†	0.8M	90.3 \pm .3	96.3 \pm .5	87.7 \pm 1.7	66.3 \pm 2.0	94.7 \pm .2	91.5 \pm .1	72.9 \pm 2.9	91.5 \pm .5	86.4
RoB _{large} (LoRA)†	0.8M	90.6\pm.2	96.2 \pm .5	90.2\pm1.0	68.2 \pm 1.9	94.8\pm.3	91.6 \pm .2	85.2\pm1.1	92.3\pm.5	88.6
DeB _{XXL} (FT)*	1500.0M	91.8	97.2	92.0	72.0	96.0	92.7	93.9	92.9	91.1
DeB _{XXL} (LoRA)	4.7M	91.9\pm.2	96.9 \pm .2	92.6\pm.6	72.4\pm1.1	96.0\pm.1	92.9\pm.1	94.9\pm.4	93.0\pm.2	91.3

Model & Method	# Trainable Parameters	BLEU	E2E NLG Challenge				CIDEr
			NIST	MET	ROUGE-L		
GPT-2 M (FT)*	354.92M	68.2	8.62	46.2	71.0		2.47
GPT-2 M (Adapter ^L)*	0.37M	66.3	8.41	45.0	69.8		2.40
GPT-2 M (Adapter ^L)*	11.09M	68.9	8.71	46.1	71.3		2.47
GPT-2 M (Adapter ^H)	11.09M	67.3 \pm .6	8.50 \pm .07	46.0 \pm .2	70.7 \pm .2		2.44 \pm .01
GPT-2 M (FT ^{Top2})*	25.19M	68.1	8.59	46.0	70.8		2.41
GPT-2 M (PreLayer)*	0.35M	69.7	8.81	46.1	71.4		2.49
GPT-2 M (LoRA)	0.35M	70.4\pm.1	8.85\pm.02	46.8\pm.2	71.8\pm.1		2.53\pm.02
GPT-2 L (FT)*	774.03M	68.5	8.78	46.0	69.9		2.45
GPT-2 L (Adapter ^L)	0.88M	69.1 \pm .1	8.68 \pm .03	46.3 \pm .0	71.4 \pm .2		2.49\pm.0
GPT-2 L (Adapter ^L)	23.00M	68.9 \pm .3	8.70 \pm .04	46.1 \pm .1	71.3 \pm .2		2.45 \pm .02
GPT-2 L (PreLayer)*	0.77M	70.3	8.85	46.2	71.7		2.47
GPT-2 L (LoRA)	0.77M	70.4\pm.1	8.89\pm.02	46.8\pm.2	72.0\pm.2		2.47 \pm .02

이점

- 메모리와 스토리지 사용량 감소
- 모든 파라미터가 아닌 LoRA 가중치만 교환함으로써 훨씬 저렴한 비용으로 배포 중에 task 사이를 전환