KOSA
Korea Software Industry Association

업무 자동화 및 워크플로 최적화를 위한

프라이빗 GPT 모델 만들기

2024. 01.

정 준 수 PhD

과정 목표

“언어모델(LLM) 작동 원리 이해를 이해하고

전문적 업무 효율성을 높이는 맞춤형 GPT 모델 구현”

생성형(Generative)의 의미

생성(Generative)이란?

문장의 시작에서 다음에 올 수 있는 가장 가능성이 높은 단어를 예측

generative purposes by predicting the next most likely word(s) given the start of a text.

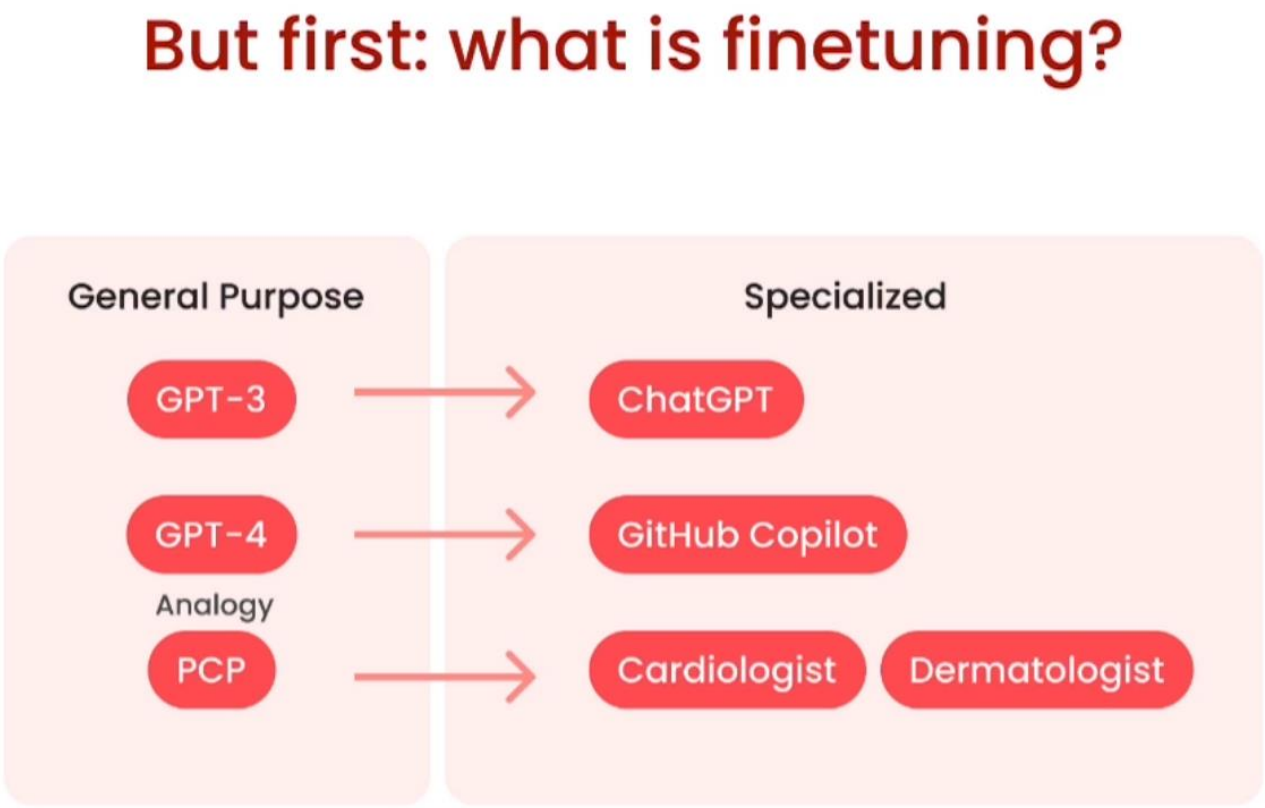
Prompt 란?

- 프롬프트란? 모델로부터 원하는 답을 얻기 위해 입력하는 내용
  모델이 학습한 내용의 특정 영역을 탐색하도록 안내하여 결과가 원하는 내용에 관련된 내용을 제공 하도록 함

- 비유하자면 언어 모델을 코드 인터프리터로 생각하면 프롬프트는 해석할 입력 코드임

- 주의할 점은 모델에는 작업할 앵커(기준점)를 지정하지 않으면 무작위로 상태를 지정하여 답을 생성함

LLM과 Pre-trained, Fine-Tuning 모델의 범용성과 특화 방법



LLM 활용에서 "미세 조정"은 특정 작업의 성능을 높이거나 특정 애플리케이션의 세부사항을 반영하기 위해 사전 훈련된 모델을 특정 데이터셋에 맞게 조정하는 과정

Prompt Engineering vs. Finetuning 의 장단점 비교

기준	프롬프트 엔지니어링	파인튜닝
데이터 요구 사항	시작할 데이터 필요 없음	더 많은 고품질 데이터 필요
초기 비용	적은 초기 비용	초기 컴퓨팅 비용 더 많음
기술 지식	기술 지식 필요 없음	데이터 관련 일부 기술 지식 필요
데이터 처리	데이터 검색을 통해 연결 (RAG) 가능	거의 무한한 데이터 수용, 새로운 정보 학습 가능
한계점	적은 데이터만 가능, 데이터를 잊어버림, 환각 현상, RAG 미스나 잘못된 데이터를 얻을 수 있음	명시되지 않음; 그러나 더 정교한 데이터 처리 능력을 시사함
모델 크기 이점	명시되지 않음	더 작은 모델 사용 시 초기 투자 후 비용 절감 가능
오류 수정	명시되지 않음	잘못된 정보 수정 가능
RAG 사용 가능 여부	가능	가능

LLM 파인튜의 장점

개인화된 LLM을 파인튜닝할 때 얻을 수 있는 성능, 개인 정보 보호, 비용, 그리고 신뢰성과 관련된 주요 장점들을 요약하면 아래 표와 같습니다.

이점 분류	자체 LLM 파인튜닝의 이점
성능	환각 중단, 일관성 증가, 원치 않는 정보 감소
개인 정보 보호	온프레미스 또는 VPC, 유출 방지, 침해 없음
비용	요청당 비용 감소, 투명성 증가, 더 큰 통제력
신뢰성	가동 시간 통제, 더 낮은 지연 시간, 조정

프리트레이닝(Pretraining) 모델

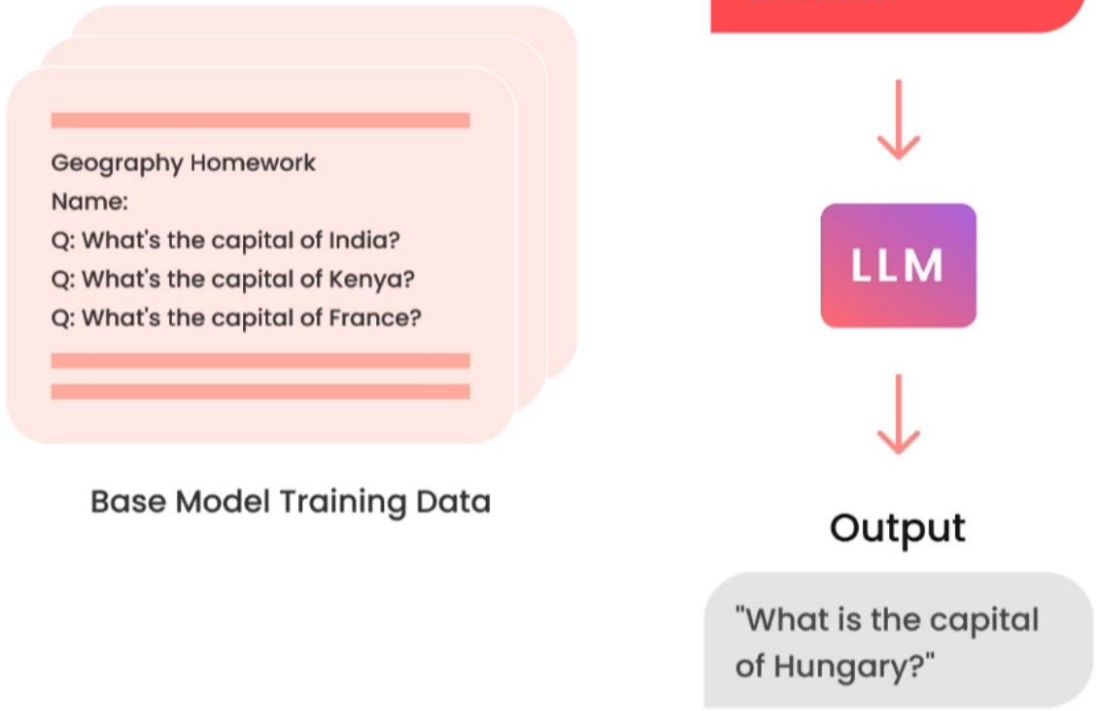
프리트레이닝
시작할 때 모델:
- 사전 지식이 전혀 없음
- 영어 단어를 형성할 수 없음
다음 토큰 예측:
- 방대한 텍스트 데이터 코퍼스
- 인터넷에서 자주 스크랩됨: “unlabeled"
- Self-supervised learning
트레이닝 후:
- 언어 학습
- 지식 학습

데이터 수집과 모델 학습, 튜닝의 특정 업무 적용

인터넷에서 스크랩된 데이터란 무엇인가?

- 프리트레인 방법이 종종 공개되지 않음
- 오픈 소스 프리트레이닝 데이터: “The Pile"
- 트레이닝에 비용이 많이 들고 시간이 오래 걸림

프리트레이닝(Pretraining) 모델 사용의 제약점



사전 훈련된 기본 모델의 제약점:

- 입력: "멕시코의 수도는 무엇입니까?"
- 출력: "헝가리의 수도는 무엇입니까?"

기본 모델 훈련 데이터에서 다양한 지리에 관한 질문이 포함되어 있었음에도 불구하고, 모델이 입력받은 질문 "멕시코의 수도는 무엇입니까?"에 대해 관련 없는 "헝가리의 수도는 무엇입니까?"라고 출력하는 것을 볼 수 있습니다. 이는 모델이 때때로 관련성이 낮은 정보를 출력할 수 있다는 것을 보여주는 예시로, 사전 훈련된 모델의 제한을 나타냅니다.

프리트레이닝(Pretraining) 모델의 Finetuning 결과

사전 훈련 모델의 파인튜닝 이후:

- 파인튜닝은 일반적으로 추가 훈련을 의미합니다.
  - 자기 감독되는 라벨 없는 데이터일 수도 있음
  - 당신이 큐레이션한 "라벨된" 데이터일 수도 있음
  - 훨씬 적은 데이터가 필요
  - 당신의 도구 상자 안의 도구

- 생성 작업을 위한 파인튜닝은 잘 정의되어 있지 않음:
  - 모델의 일부가 아닌 전체 모델을 업데이트
  - 동일한 훈련 목표: 다음 토큰 예측
  - 얼마나 많이 업데이트할지를 줄이는 더 진보된 방법들이 있음

파인튜닝의 목적은 무엇인가요?

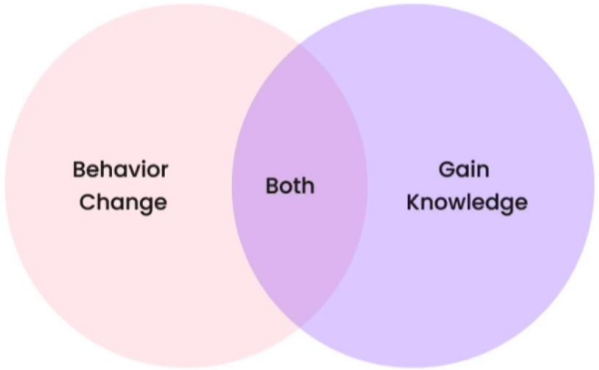
행동 변화

- 더 일관되게 응답하는 법을 배움
- 집중하는 법을 배움, 예를 들어 조정
- 능력을 이끌어내는 것, 예를 들어 대화에서 더 나아짐

지식 획득

- 새로운 특정 개념의 지식 증가
- 잘못된 오래된 정보 수정

위 3개 내용을 포함하면

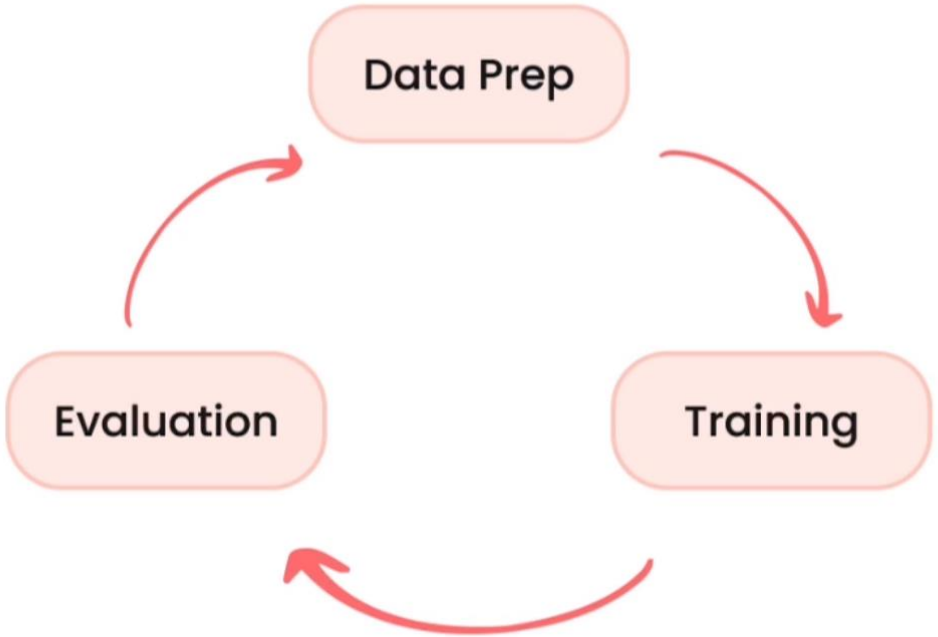


파인튜닝 작업의 종류

파인튜닝 작업 대상:

- **단순히 텍스트 입력, 텍스트 출력:**
  - **추출: 텍스트 입력, 적은 텍스트 출력**
    - **"읽기"**
    - **키워드, 주제, 라우팅, 에이전트(계획, 추론, 자기 비판, 도구 사용 등)**
  - **확장: 텍스트 입력, 더 많은 텍스트 출력**
    - **"쓰기"**
    - **채팅, 이메일 작성, 코드 작성**
- **작업의 명확성은 성공의 주요 지표입니다.**
- **명확성은 나쁜 것과 좋은 것, 더 나은 것을 아는 것을 의미합니다.**

파인튜닝 작업 과정

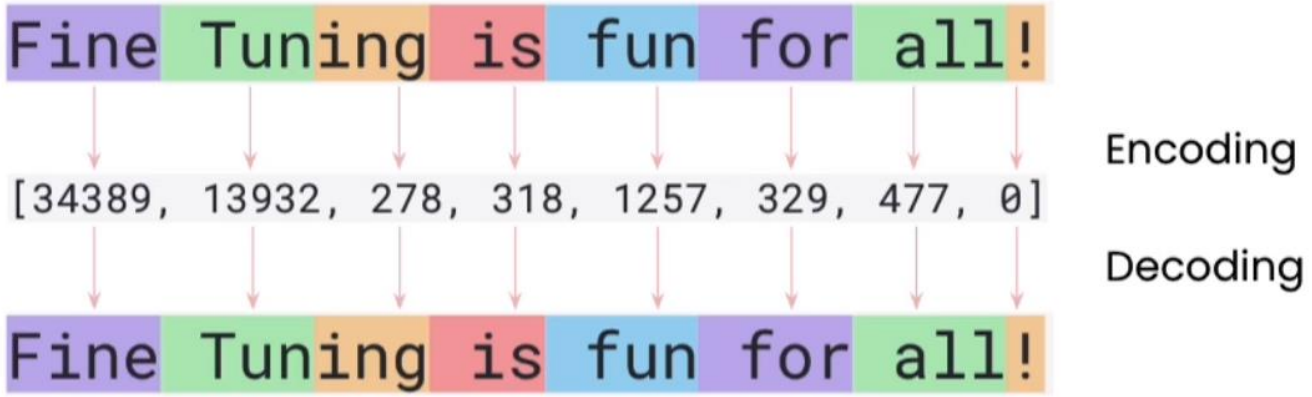


좋은 학습 데이터와 나쁜 학습 데이터

데이터 품질	좋은 데이터	나쁜 데이터
품질	더 높은 품질	낮은 품질
다양성	다양성	동질성
실제성	실제	생성된
양	더 많음	더 적음

데이터 토크나이징

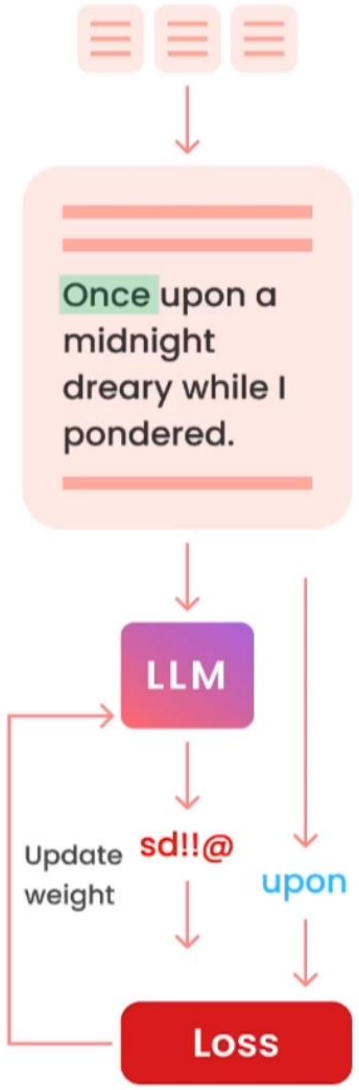
- Tokenize the data



There are multiple popular tokenizers:

- Use the tokenizer associated with your model!

데이터 토크나이징

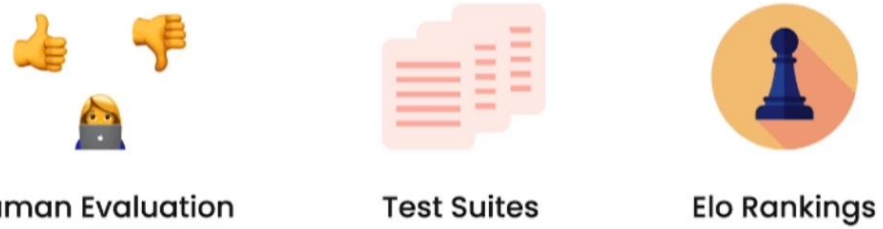


훈련 과정: 다른 신경망과 동일

- 무슨 일이 일어나고 있는가?
  - 훈련 데이터 추가
  - 손실 계산
  - 모델을 통한 역전파
  - 가중치 업데이트

- 하이퍼파라미터
  - 학습률
  - 학습률 스케줄러
  - 최적화기 하이퍼파라미터

매우 어려운 생성 모델 평가



- Human expert evaluation is most reliable
- Good test data is crucial
  - High-quality
  - Accurate
  - Generalized
  - Not seen in training data
- Elo comparisons also popular

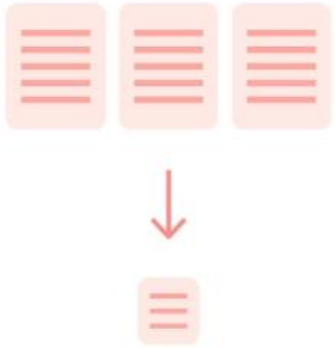
인간 전문가 평가가 가장 신뢰할 수 있습니다

좋은 테스트 데이터가 매우 중요합니다

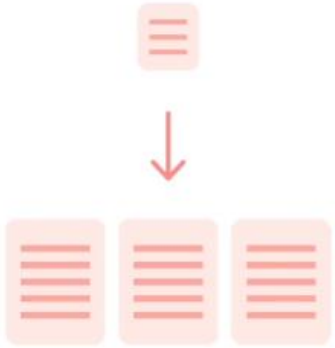
1. **고품질**
2. **정확함**
3. **일반화됨**
4. **훈련 데이터에서 본 적 없음**

일반적인 '엘로' 비교 평가

파인튜닝 작업 종류에 따른 모델 크기 비교



Extraction
(Smaller Model)



Expansion
(Larger Model)

복잡성: 토큰이 많아질수록 더 어려워짐

- 추출("읽기")은 더 쉬움
- 키워드, 주제, 라우팅, 에이전트
- 확장("쓰기")은 더 어려움
- 채팅, 이메일 작성, 코드 작성

작업의 조합은 한 가지 작업보다 어려움

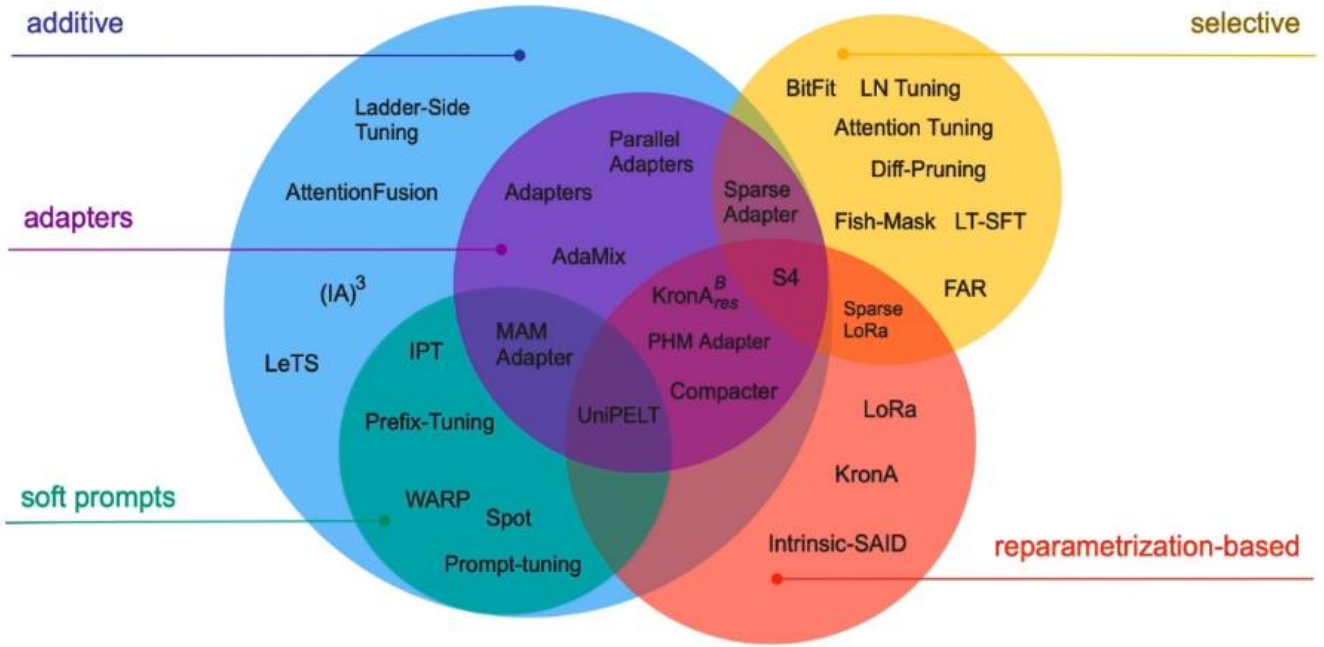
더 어렵거나 더 일반적인 것 = 더 큰 모델

- 예: 여러분은 에이전트가 유연하게 여러 가지 일을 한 번에 또는 한 단계로 수행하기를 원할 수 있음

파인튜닝 작업 규모에 따른 컴퓨팅 환경

AWS Instance	GPUs	GPU Memory	Max inference size (# of params)	Max training size (# of tokens)
p3.2xlarge	1 V100	16GB	7B	1B
p3.8xlarge	4 V100	64GB	7B	1B
p3.16xlarge	8 V100	128GB	7B	1B
p3dn.24xlarge	8 V100	256GB	14B	2B
p4d.24xlarge	8 A100	320GB HBM2	18B	2.5B
p4de.24xlarge	8 A100	640GB HBM2e	32B	5B

PEFT: Parameter-Efficient Finetuning



https://arxiv.org/abs/2303.15647

Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning

Vladislav Lialin, Vijeta Deshpande, Anna Rumshisky

PEFT(매개변수 효율적인 파인튜닝)은 기존의 파인튜닝 방법에 비해 모델의 매개변수를 보다 효율적으로 조정하여, 적은 자원을 사용하면서도 모델의 성능을 개선하는 방법입니다. 이는 특히 매우 큰 모델을 다룰 때 중요한데, 모든 매개변수를 전체적으로 조정하는 대신, 모델의 일부분만을 조정하거나, 조정할 매개변수의 수를 제한함으로써 계산 비용을 줄이는 데 도움을 줍니다.

모든 매개변수를 파인튜닝하는 이유는 무엇인가요?

모든 매개변수를 파인튜닝하는 이유는 무엇인가요?

- **LoRA: LLM의 저랭크 적응**
  - 더 적은 훈련 가능 매개변수: GPT3의 경우 10,000배 적음
  - 적은 GPU 메모리 사용: GPT3의 경우 3배 적음
  - 파인튜닝에 비해 약간 낮은 정확도
  - 동일한 추론 지연 시간

- **일부 레이어에서 새 가중치를 훈련하고, 주요 가중치는 고정**
  - 새 가중치: 원본 가중치의 랭크 분해 행렬의 변화
  - 추론 시, 주요 가중치와 병합

- **새롭고 다른 작업에 적응하기 위해 LoRA 사용**

실습 파일

https://github.com/JSJeong-me/GPT-Finetuning/

강사 소개

정 준 수 PhD

jsjeong@hansung.ac.kr

【학력】

- 고려대학교 전기공학사
- 뉴욕 공대 전산학석사(AI 전공)
- 한성대학교 컨설팅학박사

【경력】

- (前) 삼성전자, 삼성의료원, 삼성SDS 연구원
- (前) 한성대학교 겸임교수 – 머신러닝과 인지과학 강의
- (現) (주)퍼즐시스템즈 연구소장

Thank you!