# LLM 학습

## 1. Full Fine-Tuning이 어려운 이유

1. **모델 파라미터의 규모**
    - 일반적인 LLM은 수십억(1.5~3B) 이상의 파라미터를 가짐.
    - 모델을 *단순히 로드*하는 것만으로도 고사양 GPU가 필요하고, 파인 튜닝할 때는 더 많은 VRAM이 요구됨.
2. **학습 과정에서의 자원 소모**
    - Forward & Backward 연산, Gradient 계산, Optimizer 상태 유지 등을 위해 *수 배 이상의 메모리*가 추가로 필요함.
    - 예를 들어, 1B 파라미터를 가진 모델을 학습하려면 2~3B 이상의 GPU 메모리가 필요한 경우가 많음.
3. **결과적으로**
    - 기업이나 개인이 LLM을 통째로 미세 조정(Full Fine-Tuning)하기 어려운 가장 큰 이유는 **비용과 자원 한계**

---

## 3. PEFT(파라미터 효율화 미세 조정)

**PEFT**는 모델 전체 파라미터가 아닌 **일부만 조정**해서 모델을 튜닝하는 기법

대표적으로 **LoRA**(Low-Rank Adaptation)가 있고, 최근에는 양자화(quantization)까지 접목한 **QLoRA**로 확장됨.

### 3.1 LoRA(Low-Rank Adaptation)란?

### (1) 핵심 아이디어

- **모델의 대규모 파라미터(Weight)는 고정(Freeze)**
- 대신에, 각 Transformer 레이어에 **저차원(낮은 랭크, r) 행렬 A, B**를 추가로 삽입해서 ***학습 가능한 파라미터***만 업데이트함.
- 이렇게 하면 모델의 전체 파라미터를 학습할 필요가 없어서 **메모리 사용량과 계산량**이 대폭 줄어듦.

### (2) 저차원 구조(Low-Rank)의 이해

- 행렬 분해(Matrix Factorization) 기법을 이용해 큰 차원의 행렬을 작은 행렬 A, B의 곱으로 표현함.
- 이미지 추가 해야함..

### (3) 작동 방식

- 사전훈련된 모델(Pretrained Weight)을 **변경 없이 고정**
- 실제 Forward/Backward 시, **ΔW = B × A**를 해당 레이어의 Pretrained Weight에 **덧셈**해서 미세 조정 효과를 줌.
- 역전파(Backpropagation) 때는 관련 파라미터에 대해서만 기울기를 계산하니까, **학습해야 할 파라미터**가 현저히 줄어듦.

### (4) 장점

- **VRAM, 메모리 사용량 절감**: 거대한 모델 가중치를 업데이트하지 않아도 됨.
- **학습 속도 향상**: 적은 파라미터만 학습하니까 연산량도 줄어듦.
- **추론(Inference) 시에도 동일한 복잡도**: 추가된 ΔW를 단순히 원래 가중치에 더하기만 하면 됨. (수식 이미지 캡처로 추가 할까 아니면 캡처할까)
- **원상복구가 쉬움**: 모델의 원본 가중치는 변화가 없고, LoRA 파라미터만 제거하면 바로 초기 상태로 복구 가능함.

---

### 3.2 QLoRA(Quantized LoRA)란?

**QLoRA**는 LoRA의 파라미터 효율성에 **양자화(Quantization)** 기법을 결합해서, **더욱 적은 메모리**로 **유사한 성능**을 얻는 방법

1. **양자화(Quantization)**
    - 파라미터(Weights)를 16비트, 8비트 혹은 4비트 등 **낮은 정밀도로 표현**해서 메모리 절약하고, 계산 효율성 향상시킴.

- 예: `bnb_4bit_quant_type="nf4"` 옵션 등을 사용해 4비트 정밀도에 맞춰 모델 파라미터를 양자화.

2. **LoRA와의 결합**
   - LoRA는 중요한 레이어에 저랭크(A, B) 행렬을 추가해서 학습.
   - 여기에 양자화된 모델을 활용해서 전체 모델 로드를 훨씬 가볍게 수행할 수 있음.
   - 결과적으로 **Low-Rank 행렬**과 **Quantization**이라는 두 가지 효율화 기법이 합쳐져 **최소 자원으로도 준수한 성능**을 낼 수 있음.

---

# 4. 마무리

- **Full Fine-Tuning**: 모델 성능을 최적화하기 좋지만, **방대한 메모리와 계산량**이 필요함.
- **LoRA**:
  - 사전훈련된 거대한 파라미터는 그대로 두고, 일부 저차원 파라미터(A, B)만 학습.
  - 메모리 사용과 계산량을 크게 줄이면서도 **준수하거나 더 나은 성능**을 제공.
- **QLoRA**:
  - LoRA에 **양자화**까지 더해서, 훨씬 적은 정밀도로 모델 파라미터를 표현.
  - **추가 메모리 절감**과 **학습 속도 향상**, 그럼에도 **성능은 큰 손실 없이** 유지.

결론 **PEFT 기법(LoRA, QLoRA)**은 **한정된 자원**으로도 대형 언어 모델을 효과적으로 활용할 수 있게 해준다.