

056 저랭크 적응 / LoRA

Low-Rank Adaptation

가중치 전체가 아닌 저차원 행렬만 조정하는 효율적 미세조정 기법

- 기존 모델의 가중치는 고정한 채, 추가된 저랭크 행렬만 학습해 계산량과 메모리 사용을 줄이는 경량 학습 방식
- 대규모 모델의 성능은 유지하면서 빠르고 저비용으로 특정 작업에 맞게 적응시키는 미세조정 기법

LoRA란?

LoRA는 LLM이나 이미지 생성 모델을 특정 목적에 맞게 조정할 때, 전체 매개변수를 학습하지 않고 일부만 효율적으로 조정하는 경량 미세조정 기술입니다. 기존의 미세조정(Fine-tuning)은 모델의 모든 가중치를 업데이트해야 하므로, 막대한 GPU 메모리와 연산 자원이 필요했습니다. 반면 LoRA는 학습 효율을 극대화하기 위해 모델의 가중치 행렬을 저차원(Low-Rank) 형태로 분해하고, 이 중 추가된 보조 행렬만 학습합니다. 그러면 모델의 주요 구조를 유지하면서도 학습해야 할 가중치 수를 크게 줄일 수 있습니다.

LoRA의 작동 원리

LoRA의 핵심은 모델 전체를 바꾸지 않고, 필요한 부분만 조정하는 것입니다. 대형 AI 모델은 수십억 개의 가중치를 가지고 있지만, 실제로 특정 작업을 새로 학습할 때는 그중 일부만 변화가 필요합니다. LoRA는 이 점에 착안해, 기존 모델의 가중치는 그대로 두고, 아주 작은 보조 구조만 추가해 그 부분만 학습합니다. 예를 들어, LoRA는 모델의 큰 가중치 행렬을 그대로 두고, 가중치 변화량만 계산하는 별도의 작은 행렬 경로를 추가합니다. 이 경로는 두 개의 작은 행렬로 구성되며, 학습 과정에서는 이 부분만 업데이트 됩니다.

이렇게 하면 전체 모델을 다시 훈련하지 않아도 되기 때문에, 메모리 사용량과 학습 속도를 크게 줄일 수 있습니다. 학습이 끝나면 보조 행렬이 만들어낸 조정 결과만 저장하고, 원래 모델에 덧붙여 사용할 수 있습니다. 즉, 하나의 기본 모델을 유지한 채로 여러 LoRA 모듈을 만들어, 필요할 때마다 주제-언어-스타일에 맞게 교체할 수 있습니다.

LoRA의 활용

LoRA는 초거대 AI 모델의 맞춤형 활용을 가능하게 한 핵심 기술로 평가됩니다. 기존 미세조정 대비 학습 매개변수 수를 수백분의 1로 줄이면서도 성능 저하가 거의 없어, 저비용·고효율 모델 커스터마이징이 가능합니다. 이미지 생성, LLM, 음성합성 등에서 폭넓게 활용되며, 하나의 모델에 여러 LoRA를 조합하는 모듈형 학습 방식으로 확장되고 있습니다. 특히 공공·산업 부문에서는 LoRA를 통해 AI 모델을 특정 업무 환경이나 언어, 정책 도메인에 맞게 빠르게 적응시킬 수 있어, AI 확산의 현실적 대안으로 주목받고 있습니다.