참조 용어 | 지식 증류 p.74

# 030 모델 압축

Model Compression

**AI 모델의 크기와 연산량을 줄여 효율을 높이는 기법**

- 불필요한 매개변수를 제거하거나 구조를 단순화해 모델을 가볍게 만드는 기술
- 모바일·에지 환경에서 빠르고 경제적으로 AI를 실행하기 위해 사용

## 모델 압축의 개념

모델 압축은 AI 모델의 성능을 크게 훼손하지 않으면서 크기·메모리·연산량을 줄이는 기술을 의미합니다. 한국에서는 경량화라는 표현도 사용하며, 학계에서는 최적화라는 표현을 사용하기도 합니다. 최신 LLM과 비전 모델은 학습과 추론에 막대한 자원과 비용이 필요하기 때문에, 이를 실제 서비스나 모바일 기기에서 실행하기 위해서는 모델을 더 작고 가볍게 만드는 과정이 필수적입니다. 모델 압축은 단순히 매개변수 수를 줄이는 것이 아니라, 모델이 의사결정을 위해 필요로 하지 않는 부분을 찾아 제거하고, 구조를 재배열하거나 표현 방식을 바꾸는 방식으로 효율성을 높입니다. 이를 통해 대규모 모델의 성능을 유지하면서도 더 적은 비용으로 추론을 수행할 수 있는 실용적 형태로 전환할 수 있습니다.

## 모델 압축의 주요 방식

모델 압축의 방식에는 가지치기(pruning), 양자화(quantization), 지식 증류(distillation) 등이 있습니다.

- 가지치기(Pruning): 모델의 출력에 거의 기여하지 않는 매개변수나 뉴런을 제거해 구조를 간소화하는 방식으로, 불필요한 연결을 제거하면 모델 크기와 연산량이 감소하면서도 핵심 정보는 유지
- 양자화(Quantization): 매개변수를 고정밀 숫자 대신 더 작은 비트 수로 표현하는 방식으로, 예를 들어 16비트를 8비트나 4비트로 줄여 계산량과 메모리 사용량을 크게 절감
- 지식 증류(Distillation): 크고 복잡한 모델이 가진 지식을 작은 모델(학생 모델)에 이전해, 더 작은 구조로 유사한 성능을 내도록 만드는 기술

이러한 방식들은 단독으로도 쓰이지만, 실제로는 서로 결합해 더 높은 효율을 얻는 경우가 많습니다.

## 모델 압축의 중요성

모델 압축은 AI를 실제 환경에서 활용하기 위한 핵심 기술로 평가됩니다. 대규모 모델은 강력한 성능을 제공하지만, 서버 비용이 높고 응답 속도가 느리며 모바일·에지 기기에서는 실행 자체가 어려운 경우가 많습니다. 모델 압축을 통해 연산 비용을 크게 낮추면 AI 서비스를 더 저렴하게 제공할 수 있고, 사용자 단말에서도 빠르고 지속적인 추론이 가능해집니다. 특히 생성형 AI가 산업 전반에 확산되면서, 경량 모델을 기반으로 한 온디바이스 AI, 개인화 모델, 보안이 중요한 로컬 처리 환경에서 모델 압축의 활용도가 더욱 커지고 있습니다. 또한 에너지 효율 개선과 탄소 배출 감소 측면에서도 중요한 기술로 주목받고 있습니다. 결국 모델 압축은 AI 성능을 유지하면서도 접근성을 높이고 비용 구조를 개선하는 데 필수적인 역할을 수행합니다.