062 지식 증류

Knowledge Distillation

큰 모델의 지식을 작은 모델로 전달해 성능을 유지하는 기법

- 대규모 모델의 예측 과정과 정보 구조를 간추려 작은 모델이 학습하도록 하는 경량화 방식
- 작은 모델이 복잡한 모델의 판단 패턴을 모방해 효율적 성능을 내도록 만드는 기술

지식 증류의 개념

지식 증류는 대규모 모델(교사 모델)의 지식·표현·추론 패턴을 작은 모델(학생 모델)에 압축해 전달하는 모델 압축 기법입니다. 대규모 모델은 방대한 데이터와 연산을 통해 복잡한 판단 구조를 형성하지만, 이를 그대로 서비스 환경에 적용하기에는 비용과 자원 요구량이 매우 큽니다. 지식 증류는 이 문제를 해결하기 위해 교사 모델이 가진 정보를 간소화해 학생 모델이 학습할 수 있도록 만들며, 성능과 효율 간 균형을 맞춥니다. 일반적인 학습이 정답 레이블만을 활용하는 것과 달리, 증류는 교사 모델이 정답에 이르는 과정에서 생성하는 확률 분포나 중간 표현까지 활용해 더 풍부한 신호를 제공합니다. 이를 통해 학생 모델은 훨씬 작은 규모임에도 교사 모델의 패턴을 모방해 안정적인 성능을 유지할 수 있습니다.

지식 증류의 작동 방식

지식 증류는 교사 모델이 출력한 정보를 학생 모델의 학습 신호로 사용하는 방식으로 작동합니다. 교사 모델은 입력에 대해 단순 정답뿐 아니라, 여러 선택지에 부여한 확률값, 토큰 간 관계, 특징 표현 등 다양한 형태의 정보를 제공합니다. 학생 모델은 이러한 신호를 통해 "교사가 어떻게 판단하는지"를 학습하며 데이터의 구조와 의미를 자연스럽게 이해하게 됩니다. 또한 일부 설정에서는 소프트 타깃(soft target)을 활용하는데, 이는 정답과 오답을 1과 0으로 구분하는 방식이 아니라 각 선택지의 가능성을 확률로 표현한 값으로, 교사 모델의 판단 경향·유사도·강약 관계까지 반영된 정보입니다. 이를 통해 학생 모델은 적은 매개변수로도 높은 일반화 성능을 확보할 수 있으며, 제한된 자원 환경에서도 대규모 모델에 가까운 성능을 제공합니다. 지식 증류는 단일 모델 축소뿐 아니라 특정 도메인에 맞는 경량 모델을 빠르게 만들 때에도 널리 활용됩니다.

관련 용어

양자화 (Quantization)

양자화 또한 모델 압축 기법 중 하나로, 모델의 가중치와 연산을 더 낮은 비트폭으로 표현해 연산량과 메모리 사용을 줄입니다. 일반적으로 모델은 32비트 부동소수점 값으로 매개변수를 저장하지만, 대신 16·8·4비트와 같이 더 작은 표현을 사용해 모델 크기와 계산 부하를 줄입니다. 비트폭이 낮아지면 계산이 단순해져 추론 속도가 빨라지고, 에너지 소비도 줄어들어 모바일·에지 기기처럼 자원이 제한된 환경에서 특히 효과적입니다. 다만 정밀도가 저하될 수 있어, 이를 보완하기 위해 보정 기법이나 혼합정밀도 방식이 함께 사용됩니다