101 전문가 조합 / MoE

Mixture of Experts

여러 전문가 모델을 선택적으로 활용해 효율성을 높이는 AI 구조

- 여러 개의 전문가(서브네트워크) 중 입력에 맞는 일부만 활성화해 연산 효율을 높이고, 다양한 기능을 수행할 수 있도록 설계된 모델 구조
- 대규모 모델을 효율적으로 확장하고 역할을 분담해 성능과 자원 활용의 균형을 확보하는 분산 학습 방식

전문가 조합의 개념

전문가 조합(MoE)은 하나의 모델을 여러 '전문가'로 나누고, 입력에 따라 일부만 작동시키는 구조의 AI 모델입니다. 기존 대형 모델이 전체 매개변수를 매번 사용하는 데 비해, MoE는 상황에 맞는 전문가만 선택적으로 활성화해 연산 효율을 높이면서도 성능 저하가 적습니다. 특히 언어·이미지·코드 등 다양한 데이터를 처리하는 멀티모달 AI에서는 입력 특성에 따라 적합한 전문가가 선택되어 작동하므로 효율성을 더욱 높입니다

전문가 조합의 구조

MoE는 크게 전문가 집단, 선택 모듈(게이팅 네트워크), 결합부로 구성됩니다. 전문가들은 각각 다른 패턴을 학습한 작은 신경망들로 이루어져 있으며, 입력이 들어오면 선택 모듈은 이를 분석해 가장 적합한 전문가를 선택합니다. 선택된 전문가만 활성화 되므로 전체 매개변수 중 일부만 사용하게 되어, 적은 연산자원으로도 고성능을 유지할 수 있습니다. 각 전문가의 출력은 결합부에서 통합되어 최종 결과를 생성합니다.

전문가 조합의 중요성

최근 AI 모델의 규모와 연산 요구가 크게 증가하면서, 고성능을 유지하면서도 비용을 줄일 수 있는 구조가 중요해지고 있습니다. MoE는 이러한 효율성 중심 접근의 대표적 해결책으로 평가되며, DeepSeek-R1은 MoE 구조를 적극 활용해 적은 연산 자원으로도 고성능을 달성한 대표 사례입니다. 다만 전문가 간 조합 불균형이나 편중이 발생하면 효율이 떨어지고, 구조가 복잡해질수록 결과 해석이 어려워지는 한계가 있어 안정적인 MoE 설계와 로드 밸런싱 기술에 대한 관심도 함께 커지고 있습니다.

관련 용어

로드 밸런싱(Load Balancing)

여러 전문가 중 일부에게 연산이 과하게 집중되지 않도록 작업을 고르게 분배하는 과정입니다 선택 모듈이 특정 전문가만 반복적으로 활성화하지 않도록 학습 단계에서 균형 조정 규칙을 적용해, 모든 전문가가 일정 비율로 활용되도록 합니다. 이를 통해 연산 효율을 유지하고 편향이나 과적합을 방지합니다.