036 분산학습

Distributed Training

여러 장비가 협력해 AI 모델을 병렬 학습하는 방식

- 대규모 데이터나 복잡한 AI 모델을 여러 서버·GPU에 나누어 병렬 처리함으로써 학습 속도와 효율을 높이는 기술

분산학습의 배경

AI 모델의 규모가 커지면서 단일 장비로는 연산량과 메모리를 감당하기 어려워졌습니다. 수십억 개 이상의 매개변수를 가진 LLM이나 멀티모달 모델은 학습에 막대한 시간이 필요합니다. 이를 해결하기 위해 여러 장비가 동시에 연산을 수행하는 분산학습이 등장했습니다. 하나의 모델을 여러 서버·GPU가 나누어 학습함으로써 시간을 단축하고, 메모리 한계를 넘어 대형 모델을 효율적으로 훈련할 수 있습니다. 분산학습은 초대규모 AI 개발의 핵심 인프라로 기능합니다.

분산학습의 유형

분산학습은 데이터 병렬화와 모델 병렬화로 구분됩니다. 데이터 병렬화는 동일 모델을 여러 장비에 배치해 데이터의 일부를 학습하고 결과를 통합하는 방식으로, 구현이 단순하고 효율적입니다. 반면 모델 병렬화는 초거대 모델같이 모델이 너무 커서 단일 GPU 메모리에 들어가지 않을 때 주로 사용되고 각 장비는 일부 계산만 수행합니다. 최근 두 방식을 결합한 하이브리드 병렬화가 주로 사용되며, 통신 지연을 줄이기 위한 최적화 기술도 발전하고 있습니다.

분산학습의 의의

분산학습은 학습 속도 단축과 확장성 확보에 탁월합니다. 대규모 학습이 가능하고, 여러 장비를 병렬로 활용해 자원 효율도 높지만 노드 간 통신 오버헤드와 동기화 지연, 네트워크 병목 등의 문제는 여전히 과제입니다. 장애 복구와 자원 관리 비용 부담도 크며, 일관된 결과 통합을 위한 기술적 보완이 필요합니다. 향후에는 통신 효율 향상과 자동화된 운영 관리 기술이 병행 발전해야 분산학습의 효율성이 극대화될 것입니다.

관련 용어

연합학습(federated learning)

여러 장치나 기관이 데이터를 공유하지 않은 채로 협력 학습을 수행하는 방식입니다. 분산학습이 하나의 모델을 여러 장비로 나누어 계산 효율을 높이는 기술이라면, 연합학습은 데이터 프라이버시 보호와 분산 데이터 활용을 목표로 합니다. 각 참여 노드는 자신이 가진 데이터를 로컬에서 학습한 뒤, 모델의 매개변수만 중앙 서버로 전송해 통합합니다. 이 과정에서 개인 데이터는 외부로 이동하지 않아 보안성과 개인정보 보호가 강화되기 때문에 의료, 금융, 모바일 기기 등 데이터 이동이 제한된 환경에서 특히 유용합니다.