064 추론-시점 연산량 / TTC

Test-Time Compute

추론 단계에서 입력 하나의 처리에 사용되는 연산 자원의 총량

- 학습이 끝난 모델이 실제 데이터를 입력받아 결과를 생성할 때 수행하는 계산량으로, 모델의 효율성과 성능을 평가하는 핵심 지표
- 모델 구조와 토큰 길이, 하드웨어 자원 등에 따라 결정되며, AI 서비스의 비용·속도·에너지 효율에 직접적인 영향을 미치는 요소

TTC란?

TTC는 AI 모델이 학습을 마친 후, 실제 추론 단계에서 데이터를 처리하는 데 필요한 연산량을 의미합니다. 쉽게 말해, 모델이 "정답을 내놓는 순간"에 얼마나 많은 계산을 수행하느냐를 측정하는 지표이며, AI 모델의 효율성·확장성·경제성을 판단하는 핵심 기준이 됩니다. 입력 토큰 수, 모델의 매개변수 크기, 추론 시 반복 횟수 등에 따라 연산량이 크게 달라지기 때문입니다. 예를 들어, 동일한 모델이라도 더 긴 문장을 처리하거나 더 많은 응답 후보를 생성할 경우 TTC가 기하급수적으로 증가합니다.

TTC의 중요성

TTC는 AI 서비스의 속도·비용·에너지 효율을 결정짓는 핵심 요인입니다. 연산량이 많을수록 응답 시간이 길어지고, GPU·전력 소비가 늘어나며, 운영비용도 급등합니다. 반대로 TTC를 최적화하면 같은 성능을 유지하면서도 더 빠르고 경제적인 서비스 제공이 가능합니다. 이를 위해 모델 구조 단순화, 양자화 등을 적용해 불필요한 연산을 줄입니다. 또한 추론 과정에서 계산 경로를 선택적으로 활성화하는 방식도 사용됩니다.

관련 용어

Test-Time Augmentation (TTA)와 Test-Time Scaling (TTS)

TTA와 TTS는 모두 모델이 추론 단계에서 사용하는 연산량(TTC)과 밀접하게 연관된 기법입니다. TTA는 하나의 입력을 여러 형태로 변형해 모델이 반복적으로 예측한 뒤 결과를 평균하거나 통합하는 방식으로, TTC를 늘려 정확도와 안정성을 높이는 전략입니다. 예를 들어 이미지 반전·회전 등 다양한 변형을 통해 모델이 입력의 노이즈나 왜곡에 덜 민감하게 학습된 지식을 활용하도록 합니다.

반면 TTS는 모델 구조를 바꾸지 않고 추론 시 투입되는 연산 자원을 늘려 성능을 향상시키는 방법입니다. 예를 들어 언어모델이 여러 응답을 생성하고 자기 검증(Self-Consistency)을 수행하거나, 더 긴 문맥을 분석하는 방식이 이에 해당합니다.

TTA가 입력 데이터를 다양화해 예측 품질을 높인다면, TTS는 연산 규모를 조정해 결과의 정밀도를 높이는 것입니다. 즉 두 방법 모두 TTC를 전략적으로 확장해 모델의 출력 품질을 개선하는 추론 고도화 기법입니다.