Advanced Transformer

Solve curse of Length

Fast inference performance

memonode

Advanced Transformer

- **Solve curse of length**

문장 길이에 지수적으로 폭발하는 기존 트랜스포머 모델의 높은 메모리 및 연산 비용을 근본적이고
완벽하게 해결하여 완만한 선형적 하드웨어 증설로 제한 없는 토큰 컨택스트 윈도우 지원

길이의 저주를 해결함과 동시에 단위 환경에서 기존 트랜스포머 대비 학습 수렴 속도 및 추론 능력에서
200%의 우위 달성

- **Fast inference performance**

오토 리그레션 추론 과정에서 컨택스트의 모든 토큰을 처음부터 매번 연산하는 것이 아닌
이전 문맥으로부터 바로 추론함으로써 COT(Chain of Thought) 추론 과정에서 비 효율성을 제거하고
매우 적은 비용과 빠른 속도를 지원

비교

--ICLR 2025에 accept(spotlight)된 **"Hymba: A Hybrid-head Architecture for Small Language Models"** 논문--
Transformer와 Mamba block을 Parallel 하게 결합한 하이브리드 아키텍처로 Transformer의 높은 메모리 및 연산 비용과 Mamba의 성능 한계를 서로 보완하여, 효율성과 성능을 동시에 향상시킨 모델

Mamba의 성능 한계를 서로 보완 한다는 것은 지금까지 모든 트랜스포머의 고비용 문제를 개선한 모델은 학습 및 추론 성능 관점에서 트랜스포머를 능가하지 못했다는 것을 의미하고 하이브리드 모델의 성능 상한은 트랜스포머 architecture 이고 비용개선 반대 급부로서 필연적으로 트랜스포머 보다 성능 저하를 나타낼 수 밖에 없다는 사실을 보여준다.

본 Advance Transformer 모델은 기존 트랜스포머의 높은 계산 비용을 근원적으로 개선함과 동시에 모델 구조 자체로서 단위 환경에서 200%의 성능 우위를 나타내고 긴 컨택스트 윈도우 및 대량 코퍼스 학습에 비례하여 지수적 성능 향상을 제공한다.