Neuromorphic model

Solve curse of Length

Fast inference performance

Low power Low cost

memonode

Neuromorphic network model

- **특징 ***

Neuromorphic Model 및 Hybrid Attention모델은 GPT Attention 모델과 비교하여 테스트 환경에서

GPU메모리 사용량, 학습 오차 수렴 속도, 수행 속도, 추론 정확도 모든면에서 성능 우위를 보이고

길이의 증가에 따라 이차적으로 비용이 증가하는 어텐션 모델과 달리 선형적으로 증가함으로서

대용량 LLM 구성시 월등한 성능우위를 나타낼 수 있다.

- **특징 ****

오토 리그레션 추론 과정에서 문장의 모든 토큰을 처음부터 매번 연산하는 것이 아닌

이전 문맥으로부터 다음 토큰을 바로 추론함으로써 최고 속도의 추론을 최저의 비용으로 수행한다.

- **특징 *****

길이 제약 없는 추론을 하드웨어 비용 증가 없이 저비용으로 수행할 수 있다.

Pre-training 성능 비교

	New Model	Hybrid Attention	GPT Attention
100 에포크 학습 오차	0.674164	1.076424	1.118022
200 에포크 학습 오차	0.323494	0.403430	0.422084
100 에포크 학습 정확도	0.650000	0.141406	0.029687
200 에포크 학습 정확도	0.999219	0.807813	0.833594
메모리 사용률	64%	67%	90%
1 에포크 학습 수행 시간	1분 14초	1분 20초	1분 42초
5 batch 추론 수행 시간	2초	3초	6초
문맥(context) 추론	지원	지원	미지원

GPU: RTX3090 Memory: 24G

Pre-trainning 학습 구성: 512 길이단위 chunk, 32 batch

추론 정확도 측정: 512 길이의 전반부 입력으로주고 후반부 일치 스코어 계산