Not Transformer Model(NT Model)

Beyond Transformer based GPT Model

How to build the best small LLM

Best low cost LLM build method

memonode

Not Transformer, Not Attention Model

NT-model는 어텐션(attention) 연산을 사용하지 않는 시계열 신경망 모델이다.

어텐션 연산은 시퀀스와 디멘젼의 증가에 따라 지수적으로 계산량이 증가하여 고비용 학습의 주 원인

현재까지 장기 의존성에 있어 자기주의(self-attention) 집중을 넘어서는 모델은 없었으나 본 모델은 어텐션 연산을 사용하지 않고 트랜스포머 GPT 대비 우수한 학습 수렴 속도 및 실행 속도, 메모리 저 사용량에서 대략 200% 정도의 우위를 보이고 이는 토큰량 및 시퀀스 길이의 증가에 따라 지수적인 차이를 발생시킨다.

시퀀스 길이와 차원 확대에 따른 계산비용이 선형적으로 증가하여 트랜스포머 대비 확장이 용이하고 비용이 대폭 감소되어 초 대용량 학습 및 추론에 제약이 없다.

비 언어 수치 시계열 데이터에서 어텐션 대비 우수한 추론 능력을 나타낸다.

LLM (언어 사전학습 GPT 비교)

- KorQuAD(1.0) 말뭉치를 512개 토큰 시퀀스 단위로 청크하여 사전학습 시킨 결과 GPT 모델의 경우 에 포크 800번에 손실오차 0.6596을 나타내고 하습 데이터의 재현률은 85%를 나타낸다

- NT-model의 경우 에포크 200번에 손실오차 0.3819에 도달하여 빠른 학습 수렴 속도를 나타내고 학습데이터 재현률은 99%를 나타낸다.

- 2023년 신문 기사 말뭉치(2.16G)를 4K 토큰 시퀀스로 사전학습 시킨 결과 GPT 는 1개 layer 구성으로도 H100 80G 메모리를 초과하여 1개 H100으로 학습이 불가능 하다. H100 한대의 제약 조건을 fully 활용하여 512 길이로 학습시킬 경우 학습오차 3.80 대에 수렴한다.

- NT-model은 신문 기사 데이터를 4K 토큰 시퀀스, 20개 레이어 구성으로 로 학 습 시킬 경우 activation 메모리가 66G 소요되어 23억 파라미터를 H100한대로 사전 학습이 가능하다. H100 한대의 제약 조건에서 512 길이로 학습시킬 경우 학습오차 3.74 대에 수렴한다

- 위 실헝에서 어휘 량애 관계없이 NT-model이 트랜스포머 모델 보다 학습 수렴이 더 잘됨을 알수있고 가속기 1대 제한 조건에서 학습 수행 속도는 2배 가량 빠르고 엑티베이션 메모리는 ½로 적게 사용하며 이 차이는 대량 토큰일 수록 지수적으로 벌어진다.

- NT-model 모델은 토큰 시퀀스 길이 증가에 따라 메모리가 2차 지수적이 아닌 선형 적으로 증가하므로 멀티 gpu 서버 한대로서 LLM 학습이 가능하다.

- 38억 매개변수와 4k토큰 시퀀스를 지원하는 파이3미니와 같은 SLM도 사전학습에 H100 100대가 필요한 것에 반해 본 파운데이션 모델은 가속기 몇 대로 학습이 가능하고 학습 파라미터와 토큰 시퀀스 길이의 증가에 비례하여 그 차이는 지수적으로 커져 현재의 대용량 LLM 사전학습에 따른 비용문제를 근본적으로 해결한다.

Sign Curve (비 언어 수치 데이터 예측 정확도 비교)

- 테스트 데이터의 전체 길이의 반을 입력으로 주고 나머지 반을 auto regression으로 예측

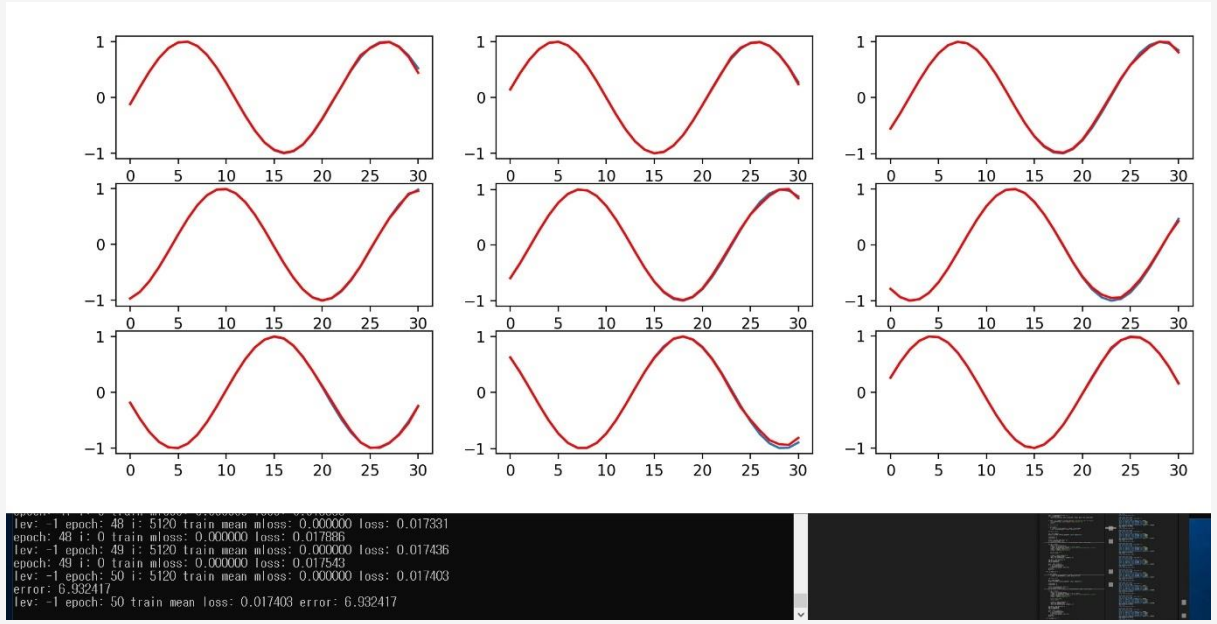
- GPT 와 비교하여 테스트 데이터 셋 추론에서 그래프 육안 식별 과 평균제곱 오차 모두 NT-model이 정확

- 이와 같이 규칙성이 명확한 케이스는 정확한 학습 및 추론이 되야 하나 학습 오차를 보 면 트랜스포머 계열 모델은 오차가 계속 수렴하지 않고 진동하며 정확도가 떨어지고 테스트 셋 실험 결과 그래프 역시 육안으로도 미세한 오차가 확인된다. 이에 반해 NT-model 모델은 오차가 계속 수렴하며 정밀한 추론 결과를 보여준다.

- Red는 예측값, blue는 정답, GPT의 경우 blue가 약간 나타나나 NT-model 의 경우 완전히 오버랩되어 red만 보여짐.

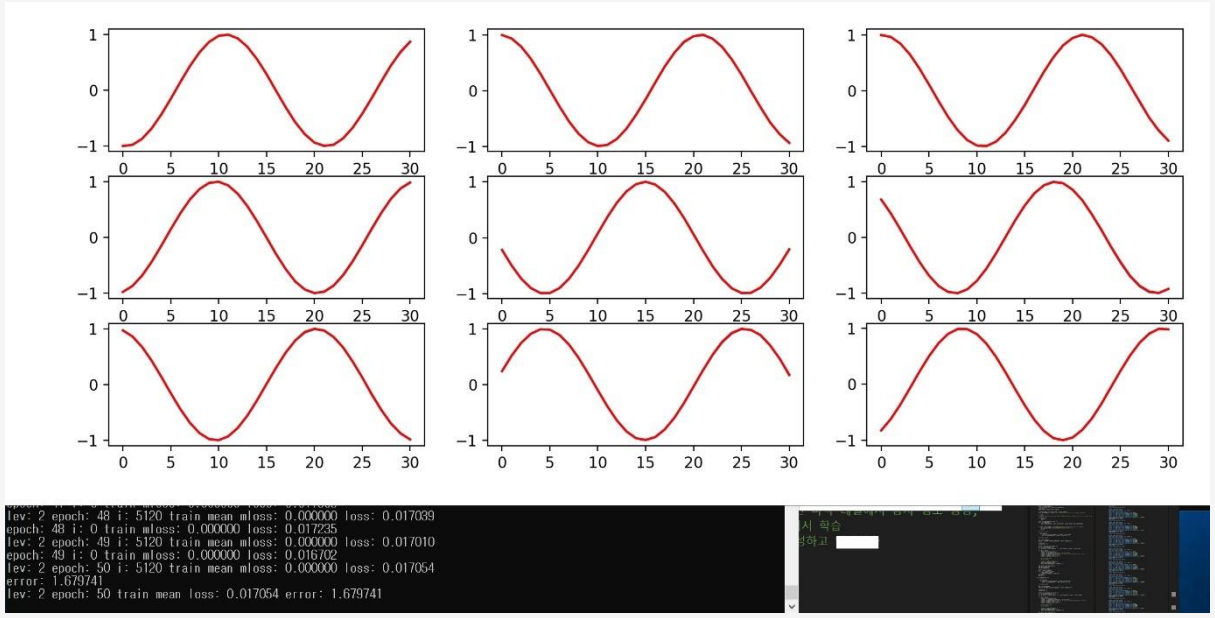
Sign Curve (비 언어 수치 데이터 예측 정확도 비교 그래프)

GPT



Error: 6.94471

NT-model



Error: 1.679741

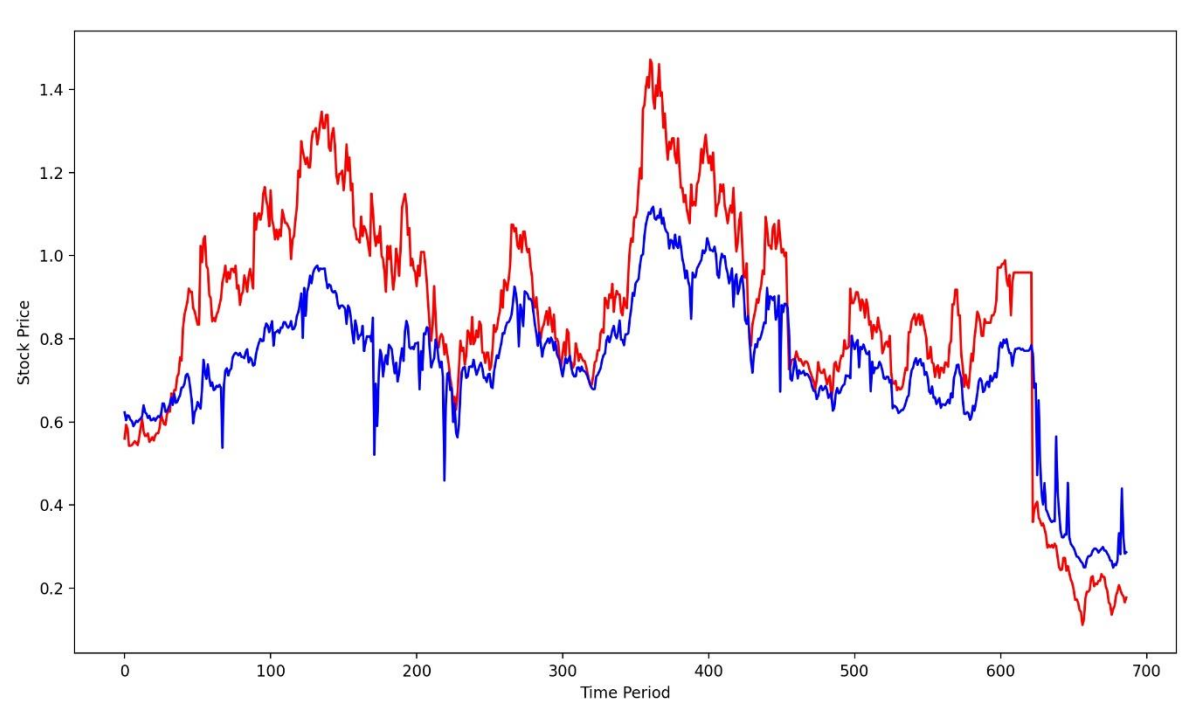
주가 예측 (비 언어 수치 데이터 추론 능력 비교)

- 과거 추세 데이터로 학습 및 예측하는 케이스에 있어 일반적인 경우와 달리 입력 데이터에 목표 주가에 관한 정보와 미래 값을 포함하지 않아 추종할(lagging) 입력값이 없어 예측이 어려운 케이스로 학습하여 다음 날 종가를 예측한다.

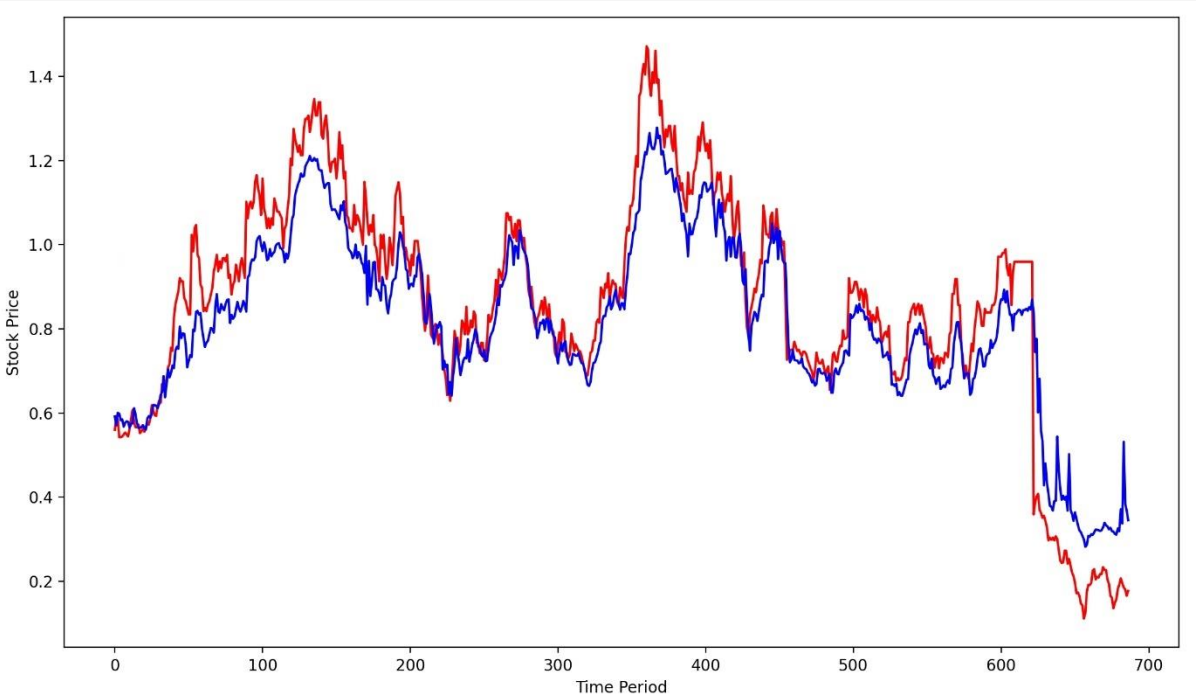
- GPT 의 경우보다 목표치에 더 근접한 추론 결과를 나타낸다.

주가 예측 (비 언어 수치 데이터 추론 능력 비교 그래프)

GPT

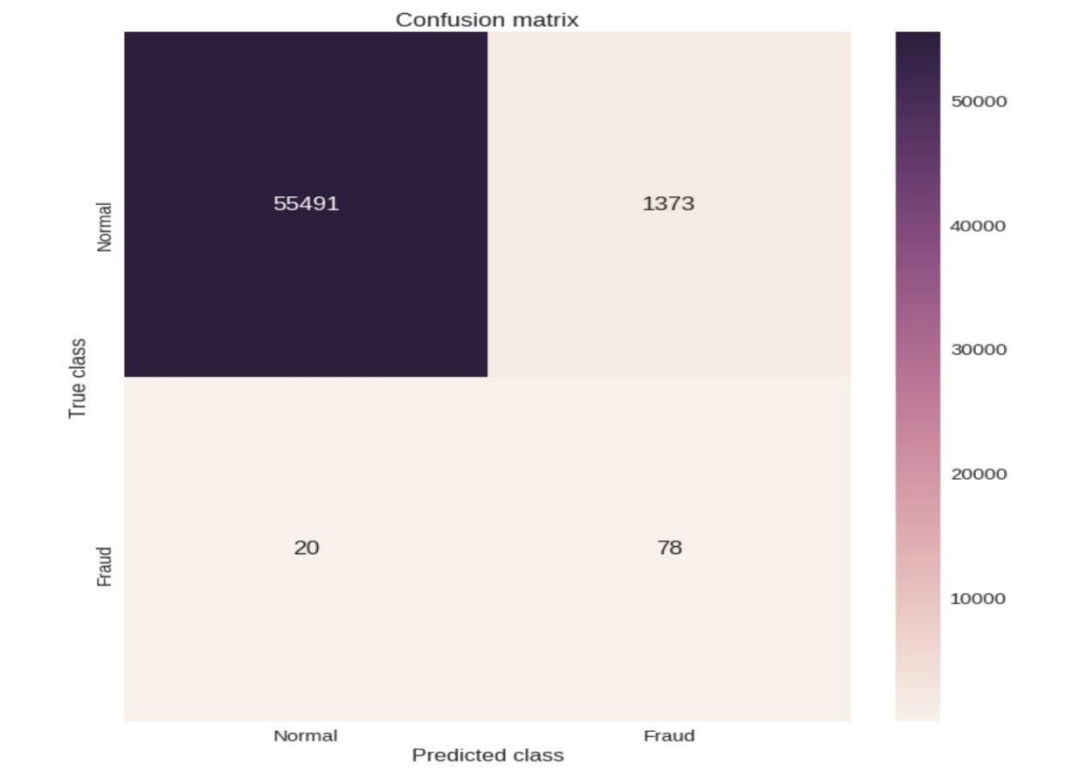


NT-model



Anormal Detection (이상탐지, 희소 데이터 학습 추론 능력)

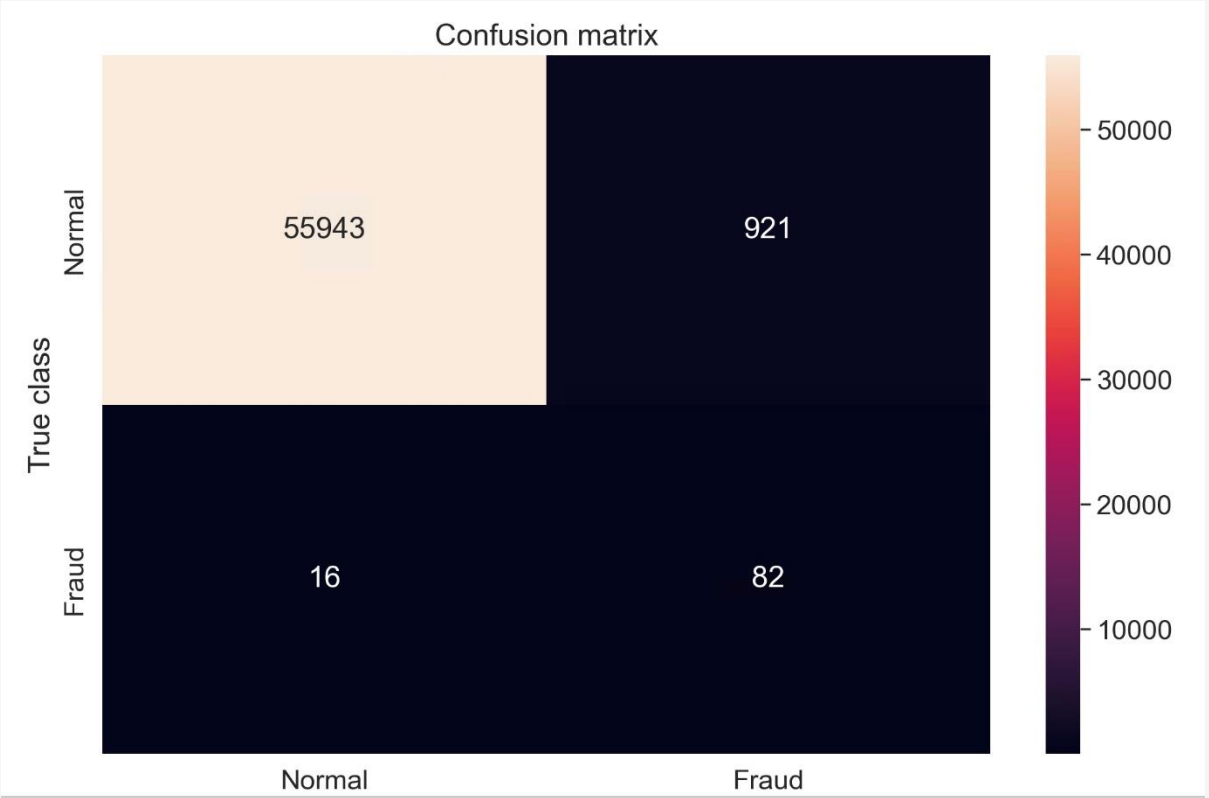
Auto Encoder



좌상단에서 우하단 대각선 방향의 값들은 크고

우상단에서 좌하단 대각선 방향의 값들은 작을 수록 탐지 능력이 좋음

NT-model



세로방향: 정답 class

가로방향: 예측 class

감사합니다.

모든 예제는 깃허브 : mempute (github.com)

Ai framework: mempute/flux: AI framework, c++ & python 1: 1 matching, API supporting host and gpu(cuda) libraries, powerfull time series neural net (github.com)

LLM: mempute/v4: Time series models built with PyTorch C++ and supporting Python interfaces, including: Spiking model using som(self-organizing map), quantum attention model, Bayesian inference network using rdbms (github.com)

LLM 사전 학습 결과: mempute/learn_results (github.com)

트랜스포머 기반 LLM의 대안을 원하거나 비 언어 시계열 데이터에서 기존 어텐션 과 차별화된 추론 성능을 원하시는 분은 연락 바랍니다.

기타 문의 사항은 mempute@gmail.com 또는 010-3254-8223으로 연락 바랍니다.