오픈소스 기초설계 개인 프로젝트 제안서.

20211771 김진영

소형주 모멘텀 예측을 위한 트랜스포머 모델 개발 및 오픈소스 구현

---

0. 프로젝트 목표 (Project Objectives)

본 프로젝트의 최종 목표는 **트랜스포머(Transformer) 모델을 활용하여 국내 소형주 시장에 특화된 고성능 주가 모멘텀 예측 시스템을 구축하고, 전체 프로세스를 재사용 가능한 오픈소스 코드로 공개**하는 것입니다.
이를 달성하기 위한 세부 목표는 다음과 같습니다.

- **소형주 모멘텀 특성 정의 및 데이터셋 구축:** 국내 주식 시장 데이터를 기반으로 '시장 마찰'이 존재하며 '유의미한 거래가 가능한' 소형주를 식별하는 기준을 정의하고, 모멘텀 분석에 필요한 학습 데이터셋을 구축합니다.
- **Transformer 기반 모멘텀 예측 모델 개발:** 시계열 데이터의 장기 의존성 파악에 뛰어난 Transformer 아키텍처를 기반으로 소형주 모멘텀 패턴을 효과적으로 학습하는 예측 모델을 설계하고 구현합니다.
- **모델 성능 및 투자 전략 검증:** 개발된 모델의 예측 정확도를 평가하고, 가상 투자 시뮬레이션(백테스팅)을 통해 실제 투자 전략으로서의 유효성과 수익성을 검증합니다.
- **프로세스 오픈소스화:** 데이터 처리, 모델 학습, 백테스팅에 이르는 전 과정을 모듈화된 Python 코드로 작성하고 GitHub 에 공개하여 금융 AI 분야의 다른 연구자 및 개발자들에게 기여합니다.

---

1. 프로젝트 배경 (Background & State-of-the-Art)

- **모델의 진화:** 전통적인 시계열 분석 이후, 주가 예측 분야에서는 LSTM 과 같은 순환 신경망(RNN) 계열의 딥러닝 모델이 널리 사용되어 왔습니다. 하지만 LSTM 은 장기 의존성(long-term dependency) 문제, 즉 아주 먼 과거의 정보가 현재 예측에 미치는 영향을 포착하는 데 한계가 있습니다. 최근 자연어 처리(NLP) 분야에서 압도적인 성능을 보인 **트랜스포머 모델**이 이러한 장기 기억 문제를 효과적으로 해결하며 금융 시계열 데이터 분석의 새로운 대안으로 부상하고 있습니다.
- **기존 연구의 초점:** 현재까지의 모멘텀 예측 연구는 대부분 시장 전체(대형주, 중형주 포함)를 대상으로 하는 **'일반(general) 모델'** 개발에 집중되어 있습니다. 하지만 소형주 시장은 대형주 시장과는 다른 독특한 특성을 가집니다.

---

2. 문제 정의 (Problem Definition)

1. **기술적 한계 (LSTM 의 한계):** 주가 모멘텀은 단기적 추세뿐만 아니라 장기적인 시장 심리와 과거 데이터 패턴에 영향을 받습니다. 기존 LSTM 모델은 이러한 장기적인 시계열 정보를 효과적으로 인식하고 학습하는 데 구조적인 한계가 있습니다.
2. **데이터의 함정 (지나친 소형주의 노이즈):** 필터링 되지 않은 극소형주는 다음과 같은 심각한 문제를 야기하여 모델의 학습을 방해하고 전략의 실효성을 떨어뜨립니다.
    - **가격 왜곡 (Price Distortion):** 낮은 유동성으로 인해 몇 건의 거래만으로 주가가 급등락하여 '가짜 모멘텀' 신호를 만듭니다.
    - **높은 거래 비용 (Transaction Cost):** 큰 매수-매도 호가 차이(Bid-Ask Spread)와 슬리피지(Slippage)로 인해 백테스트 상의 수익이 실제 거래에서는 손실로 전환될 수 있습니다.
3. **기존 모델의 한계 (General 모델의 한계):**
    - **소형주 시장의 특수성 간과:** 소형주 시장은 높은 **Market Friction** 때문에 기관 투자자의 대규모 자본 진입이 어렵고, 유동성이 낮아 **차익거래(Arbitrage) 기회가 적습니다.**
    - **비효율성 존재:** 이러한 특성으로 인해 소형주 시장은 대형주 시장보다 비효율적일 가능성이 높으며, 이는 일반화된 모델로는 포착하기 어려운 독특한 모멘텀 패턴이 존재할 수 있음을 시사합니다.

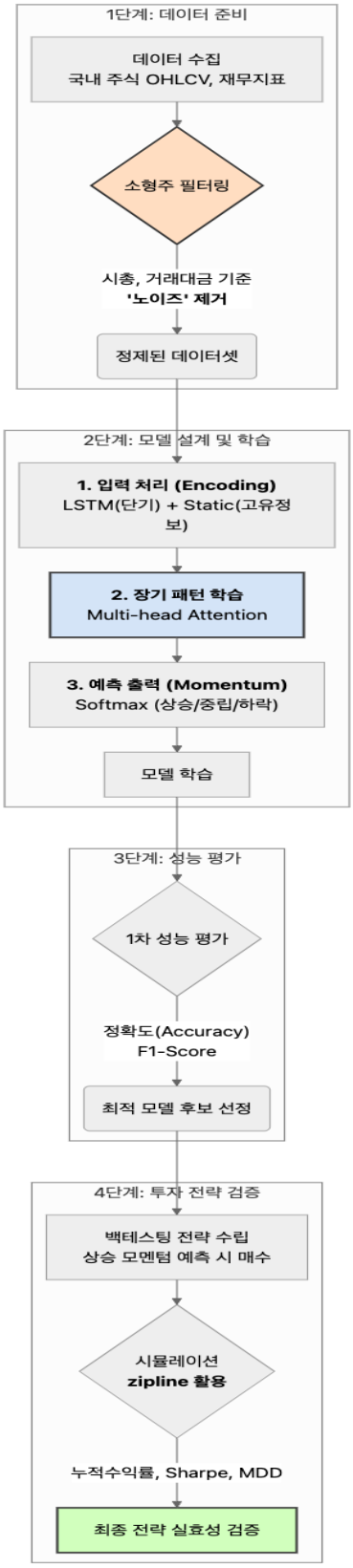
---

3. 해결 방안 (Proposed Solution)

위에서 정의한 두 가지 문제를 해결하기 위해 다음과 같은 접근 방식을 제안합니다.

1. **트랜스포머 모델 도입 (기술적 문제 해결):**
    - LSTM 의 장기 의존성 문제를 해결하기 위해 **트랜스포머 아키텍처**를 주가 예측 모델의 핵심 엔진으로 도입합니다.
    - 트랜스포머의 '어텐션(Attention) 메커니즘'을 통해 과거의 어떤 시점 데이터가 현재 모멘텀 예측에 더 중요한지 가중치를 부여하여, 더 정교하고 장기적인 패턴을 학습합니다.
2. **소형주 필터링 (데이터의 함정 문제 해결):**
    - '노이즈'를 제거하고 '유의미한 모멘텀'을 탐색하기 위해 단순히 시가총액 하위 그룹이 아닌, **최소한의 유동성 기준(예: 일평균 거래대금, 최소 시가총액)을 적용**하여 거래 자체가 불가능하거나 왜곡이 심한 종목을 **사전 제거**합니다.
    - **도구 활용:** mlfinlab 라이브러리를 활용하여 금융 데이터 정제 및 레이블링 작업을 효율화합니다.
3. **소형주 특화 데이터셋 구축 및 모델 학습 (General 모델의 문제 해결):**
    - **실용성 확보 가능:** 이 필터링 과정은 실제 매매 전략 구사 시 발생할 수 있는 과도한 슬리피지(Slippage)나 **가격 왜곡(Price Distortion)** 문제를 사전에 방지하는 역할을 합니다.
    - **특화 모델 학습 가능:** 선별된 소형주 데이터를 기반으로 '소형주 특화 모멘텀 예측 모델'을 학습시켜, 해당 세그먼트에서만 나타나는 고유한 패턴을 포착합니다.

4. 주요 연구 내용 및 방법 (Methodology)



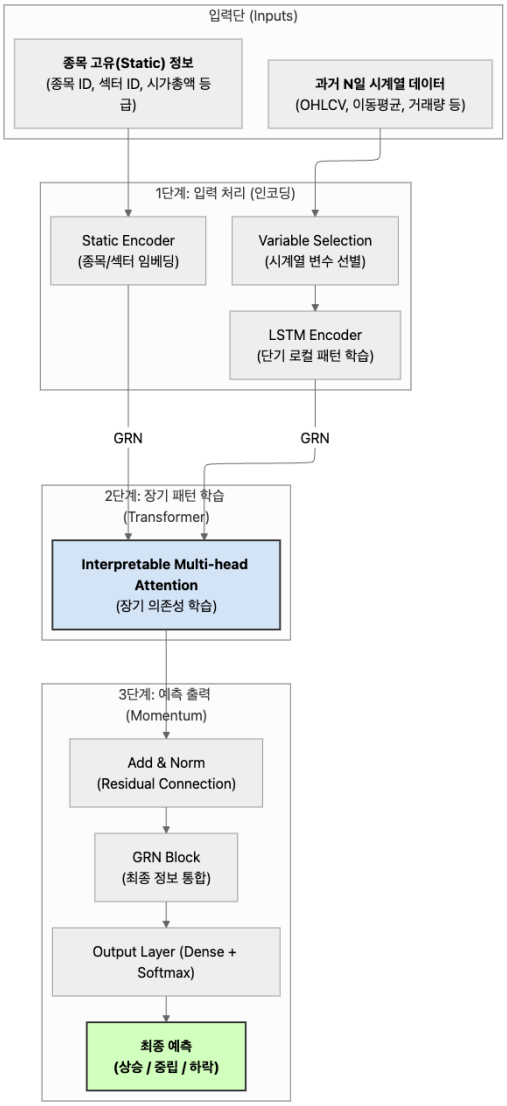
<1 단계: 데이터 수집 및 전처리 (소형주 데이터셋 구축)>

- **데이터 확보:** 국내 상장 주식의 시계열 데이터(시가, 고가, 저가, 종가, 거래량)와 시가총액 등 재무 지표를 수집합니다.
- **핵심 - 소형주 필터링:** '노이즈'를 제거하고 '유의미한 모멘텀'을 탐색하기 위해 단순히 시가총액 하위 그룹이 아닌, **최소한의 유동성 기준(예: 일평균 거래대금, 최소 시가총액)을 적용**하여 거래 자체가 불가능하거나 왜곡이 심한 종목을 **사전 제거**합니다.
- **도구 활용:** mlfinlab 라이브러리를 활용하여 금융 데이터 정제 및 레이블링 작업을 효율화합니다.

<2 단계: Transformer 기반 예측 모델 설계>

- **모델 아키텍처: TFT-Lite (Temporal Fusion Transformer-Lite) 하이브리드 모델**
    - 본 프로젝트는 단순 Transformer 인코더 구조를 넘어, Google AI 에서 발표한 **Temporal Fusion Transformer (TFT) 모델**의 핵심 아이디어를 차용하여 '소형주 모멘텀 예측'에 최적화된 **TFT-Lite** 모델을 설계합니다.
    - 이 모델은 **LSTM 과 트랜스포머의 장점을 결합한 하이브리드 구조**를 가집니다.
        - **LSTM Encoder:** 시계열 데이터(과거 N 일의 주가, 거래량 등)를 입력받아 단기적인 지역 패턴(local pattern)을 먼저 학습하고 요약합니다.
        - **Transformer (Attention):** LSTM 이 요약한 단기 패턴들과 종목 고유 정보(섹터, 시총 등급)를 함께 입력받아, 예측에 중요한 장기적인 의존성(long-term dependency)을 찾아냅니다.
        - **Static Encoder:** 종목의 고유 정보를 학습하여, 모든 주식에 동일한 패턴을 적용하는 것이 아닌 '소형주'라는 특성을 모델이 이해하도록 돕습니다.
- **구현:** PyTorch 또는 TensorFlow 프레임워크를 사용하여 모델을 구현합니다.
- **참고 모델 및 논문:**
    - 아래 논문에서 제안된 TFT 아키텍처를 기반으로, 본 프로젝트의 목표에 맞게 구조를 경량화하고 출력부를 수정하여 적용합니다. GRN(Gated Residual Network)과 같은 핵심 부품은 해당 논문의 설계를 따릅니다.
    - **Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting". *International Journal of Forecasting*, 37(4).**

- **모델 설계 예상 구조도**



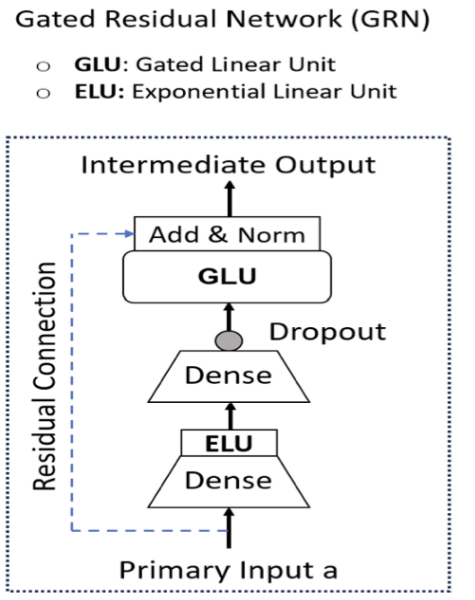


Image 2: Gated Residual Network Architecture [10]

1 단계: 입력 처리 (단기 패턴 학습)

- **목적:** 데이터를 받아서 '단기적인 특징'과 '종목의 특성'을 파악합니다.
- **작동:**
  - Static (고유 정보) : "이 종목은 IT 섹터의 소형주다" 같은 '변하지 않는 정보'를 Static Encoder 가 학습합니다.
  - Time-Varying (시계열): "과거 60 일간의 주가, 거래량" 같은 '변하는 정보'를 LSTM Encoder 가 학습합니다. LSTM 은 여기서 '단기적인 추세'나 '로컬 패턴'을 요약하는 역할을 합니다.

2 단계: 장기 패턴 학습 (트랜스포머)

- **목적:** 1 단계에서 요약된 정보들을 모아 '장기적인 핵심 패턴'을 찾아냅니다.

- **작동:**
    - Interpretable Multi-head Attention (트랜스포머의 핵심)이 1 단계의 결과물들을 입력받습니다.
    - "이 IT 소형주의 60 일치 단기 패턴들 중, **미래 모멘텀 예측에 가장 중요한 날은 언제였지?**"를 스스로 찾아내 가중치를 부여합니다. (예: 30 일 전의 거래량 폭발 패턴)
    - 이것이 바로 LSTM 의 한계였던 **장기 의존성(Long-term dependency)을 해결**하는 부분입니다.

** 3 단계: 예측 출력 (모멘텀 판단)**

- **목적:** 모든 정보를 종합하여 최종 결론을 내립니다.
- **작동:**
    - 트랜스포머가 찾아낸 '핵심 정보'가 GRN Block 과 Output Layer 를 통과합니다.
    - Softmax 함수가 "앞으로 상승할 확률 70%, 횡보 20%, 하락 10%"처럼 최종 예측 결과를 확률로 출력합니다.

<3 단계: 모델 학습 및 성능 평가>

- **학습:** 정제된 소형주 그룹의 과거 데이터를 사용하여 모델을 학습시킵니다.
- **평가:** 예측된 주가 방향과 실제 방향의 일치도를 측정하는 **정확도(Accuracy), F1-Score** 등의 지표로 모델의 예측 성능을 평가합니다.

<4 단계: 백테스팅을 통한 투자 전략 검증>

- **전략 수립:** 모델이 '상승 모멘텀'을 예측한 종목을 매수하는 가상 투자 전략을 수립합니다.
- **성과 검증:** 과거 데이터를 이용한 시뮬레이션을 통해 **누적수익률, 샤프 지수(Sharpe Ratio), 최대 낙폭(MDD)** 등 투자 성과 지표를 산출하여 전략의 실효성을 검증합니다.
- **도구 활용:** 퀀트 전략 백테스팅 라이브러리인 zipline 을 활용하여 시뮬레이션의 정확성과 효율성을 높입니다.

---

5. 기대 효과 및 기여 방안 (Expected Outcomes)

- **기술적 기여:** Transformer 모델이 금융 시계열 데이터, 특히 비효율성이 존재하는 소형주 시장 예측에 효과적으로 적용될 수 있음을 실증적으로 보입니다.
- **오픈소스 기여:** 데이터 수집 및 정제, 모델 학습, 백테스팅에 이르는 **전체 파이프라인 코드를 GitHub 에 공개**함으로써 금융 AI 분야 입문자들에게 실용적인 가이드를 제공하고, 특히 **국내 주식 데이터 분석 커뮤니티 활성화에 기여**하고자 합니다.

---

6. 참고 자료 (References)

가. 주요 학술 논문 (Academic Papers)

1. **Vaswani, A., et al. (2017). "Attention Is All You Need". *Advances in neural information processing systems*, 30.**
    - **내용:** 트랜스포머 아키텍처를 최초로 제안한 **필수적인 기초 논문**입니다. 모델의 핵심인 '어텐션 메커니즘'의 이론적 기반을 제공하여, 본 연구에서 트랜스포머를 채택한 기술적 타당성을 뒷받침합니다.
2. **Lim, B., et al. (2021). "Time-series forecasting with deep learning: a survey". *Philosophical Transactions of the Royal Society A*, 379(2194).**
    - **내용:** 금융을 포함한 시계열 예측 분야에서 딥러 NING 모델(LSTM, Transformer 등)이 어떻게 활용되는지 전반적으로 정리한 논문입니다. LSTM 의 한계와 트랜스포머의 가능성을 비교하며 **연구의 큰 그림**을 그릴 때 유용합니다.
3. **Zhou, H., et al. (2021). "Informer: Beyond efficient transformer for long sequence time-series forecasting". *In Proceedings of the AAAI conference on artificial intelligence*, 35(12).**
    - **내용:** 기존 트랜스포머를 장기 시계열 예측에 더 효율적으로 사용하기 위해 개선한 'Informer' 모델을 제안합니다. 본 프로젝트의 **모델 성능을 고도화하거나 확장**할 때 중요한 아이디어를 제공할 수 있습니다.
4. **Lim, B., et al. (2019). "Deep Momentum Networks: A Deep Learning Approach for Price Momentum-Based Trading".**
    - **내용:** LSTM 을 사용하여 모멘텀 패턴을 직접 학습하는 방법을 제안한 논문입니다. 본 연구가 **비교 기준으로 삼는 기존 딥러닝 접근 방식**을 이해하는 데 중요한 참고 자료가 됩니다.
5. **Freyberger, B., et al. (2020). "Dissecting Characteristics-Based Anomalies with Machine Learning".**
    - **내용:** 어떤 주식 특성(시가총액, 가치 지표 등)이 주가 이상 현상(수익률)에 영향을 미치는지 머신러닝으로 분석한 논문입니다. '소외된 소형주'의 특성을 **정량적으로 정의하고 필터링하는 근거**를 마련하는 데 아이디어를 제공합니다.
6. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting". *International Journal of Forecasting*, 37(4).
    - 내용 : TFT 구조 제안

나. 오픈소스 및 기술 자료 (Open-Source & Technical Resources)

1. **mlfinlab (GitHub Repository)**
    - **URL:** https://github.com/hudson-and-thames/mlfinlab
    - **역할:** 금융 데이터의 구조적 특징을 고려한 데이터 전처리, 피처 생성, 레이블링(Labeling) 등 복잡한 작업을 효율적으로 처리하기 위해 활용합니다.
2. **zipline (GitHub Repository)**
    - **URL:** https://github.com/quantopian/zipline

- **역할:** 개발된 예측 모델을 기반으로 투자 전략을 수립하고, 과거 데이터를 이용해 성과(수익률, MDD 등)를 검증하는 백테스팅 시뮬레이터로 활용합니다.

3. **Moon, T. (2023). "Momentum Transformers". The Moonlight Blog.**
   - **URL:** https://www.themoonlight.io/ko/review/enhanced-momentum-with-momentum-transformers
   - **역할:** 트랜스포머를 모멘텀 전략에 적용하는 아이디어와 구현 사례를 쉽게 설명한 기술 블로그로, **초기 아이디어 구체화 및 구현**에 참고가 되었습니다.
   - **논문 Enhanced Momentum with Momentum Transformers**
     Max Mason[1], Waasi A Jagirdar[2], David Huang[1], Rahul Murugan[2]**:**
     https://arxiv.org/pdf/2412.12516