비트코인 트랜잭션 수 예측을 위한 LSTM 모델 설계

지세현, 백의준, 신무곤, 구영훈, 윤성호*, 김명섭

고려대학교, *LG전자

{sxzer, pb1069, tm0309, gyh0808, tmskim}@korea.ac.kr, *sungho.sky.yoon@lge.com

Design of LSTM Model for Prediction of Number of Bitcoin Transactions

Se-Hyun Ji, Ui-Jun Baek, Mu-Gon Shin, Young-Hoon Goo, Sung-Ho Yoon*, Myung-Sup Kim

Korea Univ. *LG Electronics

요 약

블록체인 기술을 기반으로 만들어진 비트코인은 peer-to-peer 기술을 사용하는 온라인 암호화폐이다. 지난 몇 년간 비트코인 트랜잭션 수의 증가와 함께 비트코인 네트워크는 급속도로 발전하고 있다. 트랜잭션 수를 예측하는 것은 비트코인 생태계를 유지 및 성장시키는 것에 있어 중요하다. 본 논문은 비트코인 통계데이터를 이용해 기계학습 알고리즘 중 하나인 Long Short Term Memory(LSTM) 알고리즘을 적용한 비트코인 트랜잭션 수 예측 모델 설계방법을 제안한다. 제안하는 방법은 완성된 모델의 Mean Square Error(MSE) 수치를 통해 적합성을 검증한다.

Ⅰ. 서론 및 관련 연구

2009년 Satoshi Nakamoto에 의해 개발된 비트코인은 peer-to-peer 기술을 사용하는 세계 최초의 온라인 암호화폐이다 [1]. 지난 몇 년간 비트코인의 시장 규모는 트랜잭션 수의 증가와 함께 급속도로 커지고 있다. CoinMarketCap에 의하면, 비트코인의 시가총액은 약 170조원에 달한다 [2]. 비트코인 네트워크는 꾸준히 발전하고 있지만, 그에 따른 부작용이 발생한다. 트랜잭션 처리비용은 증가했지만, 트랜잭션 확인시간은 지연되는 문제가 발생한다 [3]. 이러한 이유로 비트코인 트랜잭션 수를 예측하는 것은 비트코인 생태계를 유지 및 성장시키는 것에 있어 중요하다.

비트코인 데이터를 분석 및 예측하기 위해 다양한 기계학습 알고리즘을 적용한 연구가 진행 중이다. 기계학습 알고리즘 중 시계열 데이터예측에 특화된 알고리즘은 Recurrent Neural Network(RNN) 와 LSTM이다. [4]는 비트코인 블록데이터를 학습데이터로 사용하여 비트코인의 가격을 예측하는 RNN 모델을 설계한 뒤, 블록데이터를 RNN 모델의 학습데이터로 사용하는 것에 대한 적합성을 검증한다. [5]는 비트코인 가격을 예측하기 위해 RNN 모델의 학습데이터로 Twitter 데이터를 사용하여 77.62%의 예측 정확도를 나타내는 모델을 설계한다. [6]은 RNN 모델의 학습데이터로 비트코인 가격과 관련된 14종류의 데이터를 사용하여 비트코인 시장의 추세를 예측한다. [7]은 비트코인의 가격을 예측하기 위한 LSTM 모델의 학습데이터로 비트코인의 가격을 사용하여 Auto Regressive Integrated Moving Average(ARIMA) 모델의 성능과 비교한다. 성능 비교 결과 LSTM 모델의 성능이 ARIMA 모델보다 정확한 예측을 한다. [8]은 Artificial Neural Network(ANN) 모델을 설계하여 비트코인의 환율을 예측한다. LSTM 알고리즘을 적용하면 ANN 모델보다 정확한 예측을 할 것이라고 제안한다. [9]는 비트코인의 가격예측을 하는 LSTM 모델을 설계했다. 학습데이터의 종류에 따른 두 가지 LSTM 모델을 설계했고, 설계된 LSTM 모델의 성능을 비교했다. 학습데이터의 종류에 따라 LSTM 모델의 성능에 변화가 있음을 확인한다.

본 논문의 목적은 비트코인 트랜잭션 수를 예측하기 위한 LSTM 모델 설계방안을 제안하는 것이다. 적합한 LSTM 모델을 설계하기 위해서는 적절한 학습데이터를 사용하는 것과 LSTM 모델을 구성하는 Hyper-parameter Optimization이 필요하다. 학습데이터의 종류에 따라 LSTM 모델의 성능은 달라진다. 본 연구팀은 비트코인 블록으로부터 84종류의 비트코인 블록 및 트랜잭션의 통계데이터를 수집했다. 수집된 84종류의 데이터를 전부 사용하는 것은 모델의 성능을 보장할 수 없을 뿐만 아니라 많은 시간이 소요된다. 따라서 비트코인 트랜잭션 수와 관계가 있는 데이터를 선별하기 위해 상관분석을 적용한다. 상관분석을 적용해 학습데이터를 선별한 뒤, LSTM 모델의 성능을 높이기 위해 Hyper-parameter Optimization을 하여 비트코인 트랜잭션 수를 예측하는 모델을 완성한다.

Ⅱ. 비트코인 트랜잭션 수 예측 LSTM 모델 설계

2.1 비트코인 블록 통계데이터

표 1. 비트코인 블록의 통계데이터

Data Unit	Raw Feature	1st Statistical Process	2nd Statistical Process	Number of Features
Block	nTx			1
	Weight			1
	Size			1
	Vsize			1
Transaction	nVin		Sum Max Min Avg Stdv	5
	nVout			5
	Value			5
	Fee			5
	Tx_Size			5
	Tx_Vsize			5
	Vout_value	Sum Max Min Avg Stdv		25
	Vin_value			25

※ 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2018R1D1A1B07045742)과 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2018-0-00539-001,블록체인의 트랜잭션 모니터링 및 분석 기술개발)

초기 비트코인 블록은 비트코인 네트워크가 본격적으로 활성화되기 이전의 데이터가 포함되어 있으므로, 100,001 높이의 블록부터 200,000 높이의 비트코인 블록데이터를 수집한다. 수집한 데이터 종류는 표 1과 같다. 비트코인 블록에 포함된 원시 데이터로부터 합, 평균, 최댓값, 최솟값, 표준편차를 구하는 통계처리를 적용해 84종류의 통계데이터를 수집한다.

2.2 피어슨 상관분석을 적용한 학습데이터 선택

피어슨 상관분석은 두 변수 사이의 선형 상관관계를 피어슨 상관계수로 나타내는 방법이다. 피어슨 상관계수는 두 변수 사이의 선형 상관관계를 계량화한 값이다. 피어슨 상관계수는 -1부터 +1 사이의 값을 갖고 수식 1에 의해 값을 구한다. p는 피어슨 상관계수, B는 블록 통계데이터, T는 트랜잭션 수, n은 데이터의 수, S는 표준편차, bar는 표본 평균을 의미한다. 피어슨 상관계수의 값이 -1의 경우 완벽한 음의 선형적 관계, +1의 경우 완벽한 양의 선형적 관계를 의미한다. 학습데이터의 피어슨 상관계수의 값이 +1에 가까울수록 LSTM 모델의 성능은 좋아진다 [10]. 수집한 84종류의 통계데이터 항목과 비트코인 트랜잭션 수 사이에 피어슨 상관분석을 적용해 상관계수가 가장 높은 데이터를 선택한다.

$$p = \frac{\sum_{i=1}^{n}(B_i - \overline{B})(T_i - \overline{T})}{(n-1)S_B S_T} \qquad (1)$$

2.3 Hyper-parameter Optimization

LSTM 모델을 구성하는 주요 Hyper-parameter는 1회의 학습에 사용할 데이터의 길이인 Sequence Length, LSTM Cell에 존재하는 Hidden Unit의 수, 모델의 성능을 평가하는 Loss Function, Loss Function을 줄이기 위한 함수인 Optimizer이다. Sequence Length와 Hidden Unit의 수를 제외한 나머지 Hyper-parameter는 보편적으로 우수한 성능을 나타내는 구성이 존재한다. Loss Function은 MSE를 사용하여 모델의 성능을 평가하고, Optimizer로는 효율적인 학습을 위해 Adam Optimizer를 사용한다. 그러나 적절한 길이의 Sequence Length와 Hidden Unit의 수는 실험을 통해 찾아야 한다.

Ⅲ. 실험 및 결과

피어슨 상관분석을 통해 선택한 학습데이터를 사용하여 실험을 진행한다. 총 100,000개의 데이터를 학습 80%, 검증 10%, 실험 10%의 비율로 구분한다. 모든 데이터를 LSTM 모델에 최적화하기 위해 0과 1 사이의 값으로 정규화를 한다. 적합한 Hidden Unit의 수를 찾기 위해 Sequence Length를 1로 설정하여, Hidden Unit의 개수를 늘려가며 모델을 학습시킨 뒤 학습에 사용하지 않은 검증, 실험데이터를 사용하여 성능을 확인한다. 모델 성능 평가 지표인 MSE는 0에 가까운 값일수록 모델의 성능이 좋다. 실험 결과는 그림 1과 같다.

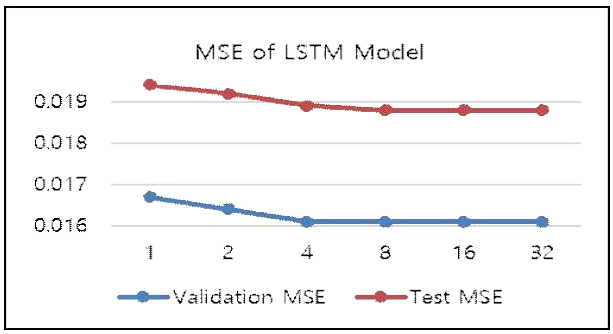

그림 1. Hidden Unit의 수 최적화 실험 결과

Hidden Unit의 수가 8개까지는 MSE 값이 줄어들지만, 8개 이상의 개수를 설정해도 값이 줄지 않는다. 따라서 트랜잭션 수를 예측하는 LSTM 모델의 Hidden Unit의 수는 8개로 설정하는 것이 적합하다. 마지막으로 적절한 길이의 Sequence Length를 찾기 위한 실험을 진행한다. Hidden Unit의 수를 8개로 설정한 뒤, Sequence Length를 점차 늘려가며 모델의 성능을 확인한다. 실험 결과는 그림 2와 같다. Sequence Length의 길이가 32일 때까지는 MSE 값이 줄어들지만, 32개 이상의 길이를 설정해도 값이 줄지 않는다. 따라서 적절한 길이의 Sequence Length는 32이다. MSE의 변화를 보았을 때, Sequence Length가 Hidden Unit의 수보다 LSTM 모델의 성능에 더 영향을 주는 요소로 파악할 수 있다.

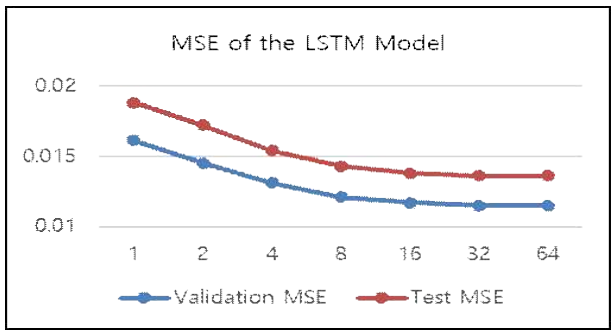

그림 2. Sequence Length 최적화 실험 결과

Ⅳ. 결론 및 향후 연구

본 논문은 비트코인 네트워크를 유지 및 성장시키는데 중요한 비트코인 트랜잭션 수를 예측하는 LSTM 모델 설계방법을 제안하였다. 제안하는 방법은 실험을 통해서 적합성을 검증하였다. 향후 연구로는 기본 LSTM 모델 이외의 Stacked-LSTM, Stateful-LSTM 모델을 설계하는 방안을 연구할 계획이다.

참 고 문 헌

[1] NAKAMOTO, Satoshi, et al. Bitcoin: A peer-to-peer electronic cash system. 2008.

[2] CoinMarketCap, last Modified May 23, 2019, accessed May 23, 2019, https://coinmarketcap.com

[3] Gabriel Bianconi, Mahesh Agrawal, Predicting Bitcoin Transactions with Network Analysis, snap.stanford.edu, last modified Sep 10, 2018, accessed May 20, 2019, https://snap.stanford.edu/class/cs224w2017/projects/cs224w-65-final.pdf.

[4] KODAMA, Osamu; PICHL, Lukáš; KAIZOJI, Taisei. Regime change and trend prediction for Bitcoin time series data. In: CBU International Conference Proceedings. 2017. p.384-388.

[5] PANT, Dibakar Raj, et al. Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis. In: 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS). IEEE, 2018. p. 128-132.

[6] SEO, Yunbeom; HWANG, Changha. Predicting Bitcoin Market Trend with Deep Learning Models. Quantitative Bio-Science, 2018, 37.1: 65-71.

[7] KARAKOYUN, E. S.; CIBIKDIKEN, A. O. Comparison of ARIMA Time Series Model and LSTM Deep Learning Algorithm for Bitcoin Price Forecasting. In: The 13th Multidisciplinary Academic Conference in Prague 2018 (The 13th MAC 2018). 2018. p. 171-180.

[8] PICHL, Lukáš; KAIZOJI, Taisei. Volatility analysis of bitcoin. Quantitative Finance and Economics, 2017, 1: 474-485.

[9] WU, Chih-Hung, et al. A New Forecasting Framework for Bitcoin Price with LSTM. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018. p. 168-175.

[10] 지세현, 구영훈, 백의준, 신무곤, 윤성호, 김명섭, "비트코인 트랜잭션 수 예측을 위한 LSTM 학습데이터 선택기법", KNOM Conference 2019, accepted May 14, 2019.