딥러닝을 이용한 암호화폐 도지코인 가격 예측

Deep Learning Models for Dogecoin Price Prediction

김진우[1], 안병천[1], 신택수[1]

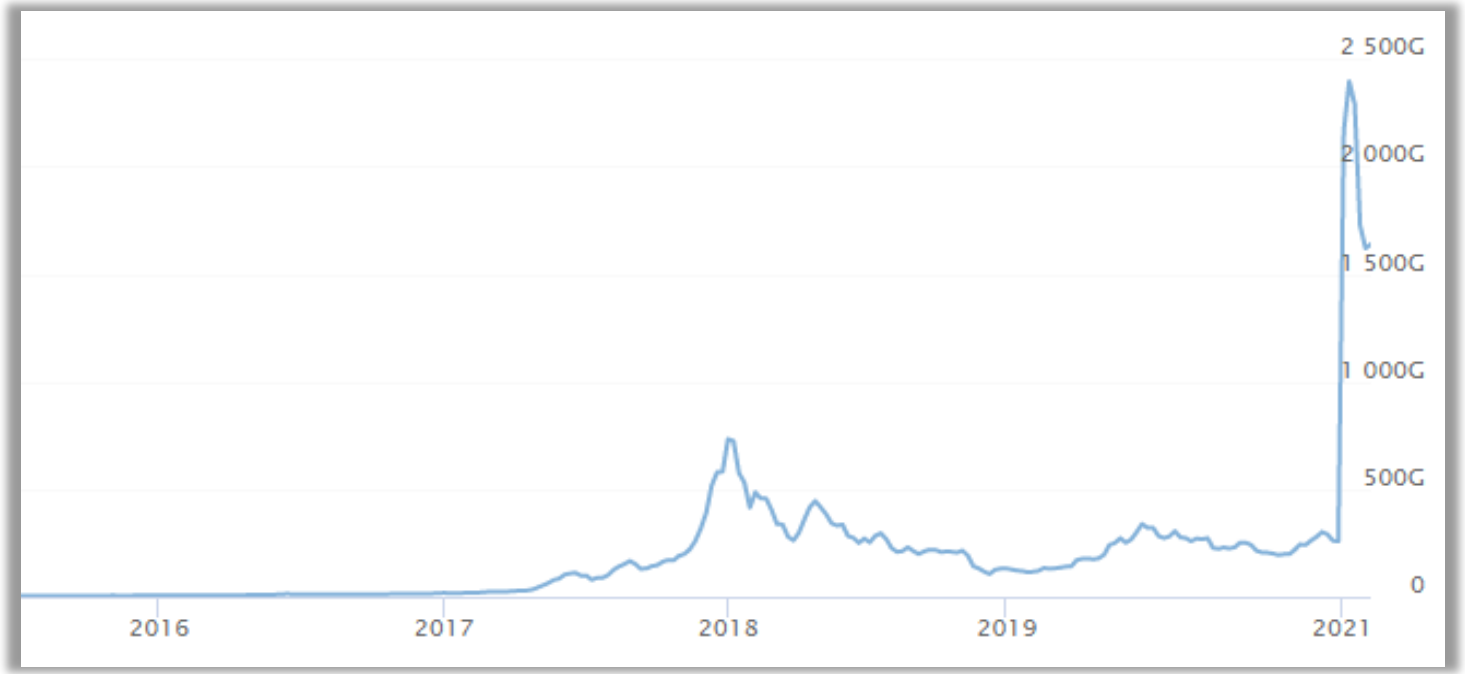
[1]연세대학교 글로벌창의융합대학 경영학부

2021. 08. 18

Outline

I. Introduction - Background

Fig. 1. 암호화폐 시가총액



2009년 사토시 나카모토에 의해 공개된 비트코인을 시작으로, 다양한 암호화폐가 등장하고 사람들의 관심과 투자가 증가하여 암호화폐 시장은 2021년을 기점으로 크게 성장하였다.

Table. 1. 사용 데이터에 따른 금융시장 가격 예측 종류

금융시장 (주식, 암호화폐) 가격 예측	정형 데이터	OHLVC데이터	• 시가,고가,저가,종가, 거래량 자체 • 종가만 사용
		기술적 지표	• **불린저 밴드** • **단순 이동 지표** • **지수 이동 평균** • **이동평균 수렴확산** • **MACD 히스토그램**
	비정형 데이터	텍스트데이터	• **뉴스 데이터** • **SNS 데이터** • **공시 자료**
		이미지 데이터	• **캔들스틱 차트** • **그래프**

I. Introduction - Financial market pricing prediction

가격 예측

기본적 분석

상품의 내재적 가치를 분석하여 미래의 상품 가격을 예측하는 방법. 상품의 펀더멘털 정보를 주로 활용한다.

기술적분석

과거 상품의 가격이나 거래량 같은 자료를 이용하여 상품 가격 변화의 추세를 발견해내어 미래의 상품 가격을 예측하는 방법

I. Introduction - Deep learning study for Crptocurrency pricing prediction

- 암호화폐 가격 흐름은 기존 주식 가격이 가진 흐름과 비슷한 특징을 가지고 때문에 암호화폐 예측 연구는 기존 주가 예측에 사용되었던 방법들이 사용되고 있다.

- 주식 가격 예측에는 기술적 분석과 기본적 분석이 동시에 존재하지만, 암호 화폐 예측에는 펀더멘털에 대한 정보가 주어지지 않기 때문에 기본적 분석은 사용되지 않으며 과거 가격의 흐름을 통해 미래의 가격을 예측하고자 하는 연구가 주를 이루고 있다.

I. Introduction - Deep learning study for Crptocurrency pricing prediction

- 대부분의 암호화폐 선행 연구는 데이터셋으로 비트코인을 사용

  (Seo, 2018, Heo, 2019, Chowhury, 2020, Kim, 2020, Gang, 2020, Kim,2021, Won, 2021)

- Kim(2019)는 대시, 라이트 코인, 모네를, Heo(2019)는 ETH, XRP, BCH, LTC, DASH, ETC를, Chowhury(2020)은 DOGE 코인, Ethereum, IOTA, Litecoin, NEM, NEO를, Choi(2020)은 이더리움을 사용하기도 함

I. Introduction - Deep learning study for Crptocurrency pricing prediction

- 대부분의 암호화폐 선행 연구는 일별 데이터를 사용

(Seo, 2018, Heo, 2019, Chowhury, 2020, Kim, 2020, Gang, 2020, Kim,2021, Won, 2021)

- Heo(2019)는 다양한 코인의 부족한 데이터양을 채우기 위해 10분 데이터를 사용

I. Introduction - Dogecoin

- 발행일자 : 2013년 12월
- Joke currency : 장난에서 시작된 코인
- 시가총액(2021-06-01) :46.2조원, 전체4위
- 거래량(업비트,2021-06): 2795억 7500만개
- 부족한 기술 바탕에 비해 많은 시장 참여자

Dogecoin의 하루동안 거래는 기존 주식 시장에서 발생하던 거래량을 훨씬 초과하여 발생하고 있다. 따라서 하루동안의 변동성이 매우 크며, 장기적인 추세 없이 빠르게 상승과 하락을 반복하는 특성을 보인다.

도지코인 1분 거래량과 삼성전자 1일 거래량 비교

도지코인 1분 평균	삼성전자 하루 평균
7,519,857	약 12,000,000

*분석기간 기준(업비트, 삼성 공시자료)

- OHLCV 정형 데이터를 입력값으로 사용하는 딥러닝 모델과 Candlestick chart 이미지 데이터를 입력값을 사용하는 딥러닝 모델을 가지고 DOGE COIN의 가격 등락 예측하고 비교, 분석 연구

- DOGE Coin의 과거 30분 동안의 캔들스틱 차트 이미지를 가지고 1분 뒤의 상승과 하락을 예측

II. Literature review - Study for financial market prediction using unstructed data

저자	제목	데이터셋			분석모델	예측형태	분석기간
Won (2021.06)	텍스트 마이닝과 딥러닝을 활용한 암호화폐 가격 예측 : 한국과 미국시장 비교	Bitcoin	가격 데이터, 뉴스 데이터	일별	ARIMA, RNN, Separated RNN	가격자체	2020.01.20 ~ 2020.09.16
Kim (2021.06)	감성분석을 이용한 뉴스정보와 딥러닝 기반의 암호화폐 수익률 변동 예측을 위한 통합모형	비트코인	과거 가격 정보, 뉴스감성분석 사례기반추론	1일	로짓, 인공신경망(ANN), SVM, LSTM	가격의 상승과 하락	2018~2020(학습) 2021(검증)
Choi (2020.12)	암호화폐 가격 예측을 위한 딥러닝 앙상블 모델링 : Deep 4-LSTM Ensemble Model	이더리움	거래날짜, OHLCV	일별 /30일	Single RNN, Single LSTM, 2-LSTM, 3-LSTM, D4LE	가격등락	
Gang (2020.11)	LSTM 기반 감성분석을 이용한 비트코인 가격 등락 예측	Bitcoin	전날의 비트코인 관련 Google 기사 (감정 수치)	일별 / 1일 평균	LSTM	가격등락	2013.01.01 ~ 2020.07.30

II. Literature review - Study for financial market prediction using unstructed data

저자	제목	데이터셋			분석모델	예측형태	분석기간
Kim (2020.10)	뉴스의 감성정보와 딥러닝을 이용한 암호화폐 가격 예측모형	비트코인	가격 정보 뉴스정보	일별	ARIMA, BPN, SVR RNN	가격자체	2018.01.01 ~2020.08.31
Chowhury (2020.03)	An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning	Bitcoin Cash; Bitcoin; Dash; Dogecoin (DOGE); Ethereum; IOTA (MIOTA); Litecoin; NEM; NEO	Date, OHLCV, Market capital	daily	Gradient boosted trees, Neural net, Ensemble learning, K-NN	가격자체	~2018.12.31 2019.01.01 ~ 2019.01.31
Kim (2019.10)	머신러닝 기법을 활용한 암호화폐 유통 가격 예측 연구	대시, 라이트 코인, 모네	가격데이터	일별	ANN, SVM	가격자체	2018.04.28 ~2018.12.25
Heo (2019.02)	기계학습을 활용한 암호화폐 가격동향 예측	BTC, ETH, XRP, BCH, LTC, DASH, ETC	API 데이터	10분	그레디언트 부스팅	상승하락	2017.8 ~ 2018.5

II. Literature review - Deep learning study for Crptocurrency pricing prediction

- 선행연구들은 주로 과거 주가 예측에 사용되었던 연구방법들을 암호화폐 시장에 적용시켜 암호화폐의 가격을 예측하자 시도하고있다.

- 정형 데이터인 가격 데이터를 입력값으로 머신러닝, 딥러닝 모델에 입력하여 예측하는 연구에서 최근에는 텍스트 마이닝을 통한 SNS, 뉴스 데이터의 감성 수치를 이용한 예측이 진행되고 있다.

II. Literature review - Study for financial market prediction using unstructed data

저자	제목	데이터셋		분석모델	예측형태	분석기간
Guo et al.(2018)	Deep candlestick predictor: A framework toward forecasting the price movement from candlestick charts	TAIFEX	Candlestick chart	AE+CNN	가격상승 또는 하락	1998 ~ 2016
Kim and Kim(2019)	데이터 증강을 통한 딥러닝 기반 주가 패턴 예측 정확도 향상 방안	S&P 500	Candlestick chart	CNN, LSTM	가격자체	2016.10 ~ 2017.10
Kusuma et al.(2019)	Using Deep Learning Neural Networks and Candlestick Chart Representation to Predict Stock Market	S&P BSE SENSEX; NIFTY 50	Candlestick	CNN	가격상승 또는 하락	NA
Bae (2021)	변이형 오토인코더와 어텐션 메커니즘을 결합한 차트기반 주가 예측	S&P 500	Candlestick chart, VIX chart	VAE+ Attention model and Dense layer	가격상승 또는 하락	1993.7 ~ 2019.7

Literature review - Study for financial market prediction using unstructed data

- 최근 들어 주식 가격 예측을 위해 비정형 데이터를 활용하는 연구들이 많이 등장하고 있다. 텍스트 데이터 기반의 연구 뿐만 아니라 다양한 그래프를 이미지로 학습하여 주식 가격을 예측하는 연구들도 등장하고 있다.

- 이미지 데이터를 사용하는 대부분의 연구들은 이미지 학습에 특화된 CNN(합성곱 신경망)을 사용하거나 CNN에 LSTM, BLSTM, AE를 결합한 딥러닝 모델을 사용하여 주식 가격을 예측하고자한다.

Data source

- Upbit API : DOGE Coin OHLVC Data

Time period :

- 2021.04.12 – 2021.05.12 (31 Day, 44400 Min)

Data labeling

- 다음날 종가가 시가보다 클 시 UP(상승) 레이블을, 작을 시 Down(하락) 레이블을 부여
- Up = 1, Down = 0

III. Data - Descriptive statistics table

Table. Descriptive statistics.

	시가(Open)	고가(High)	저가(Low)	종가(Close)	거래량 (Volume)
Mean	544.097	545.618	542.624	544.099	7519857
std	133.017	133.494	132.550	133.029	117201100
25%	426	427	425	426	1429776
50%	522	524	521	522	3547487
75%	641	642	639	6340	8550610
min	317	321	316	317	5278
max	886	889	880	885	180514200

III. Data - Close graph & Label rate

Fig. Close graph

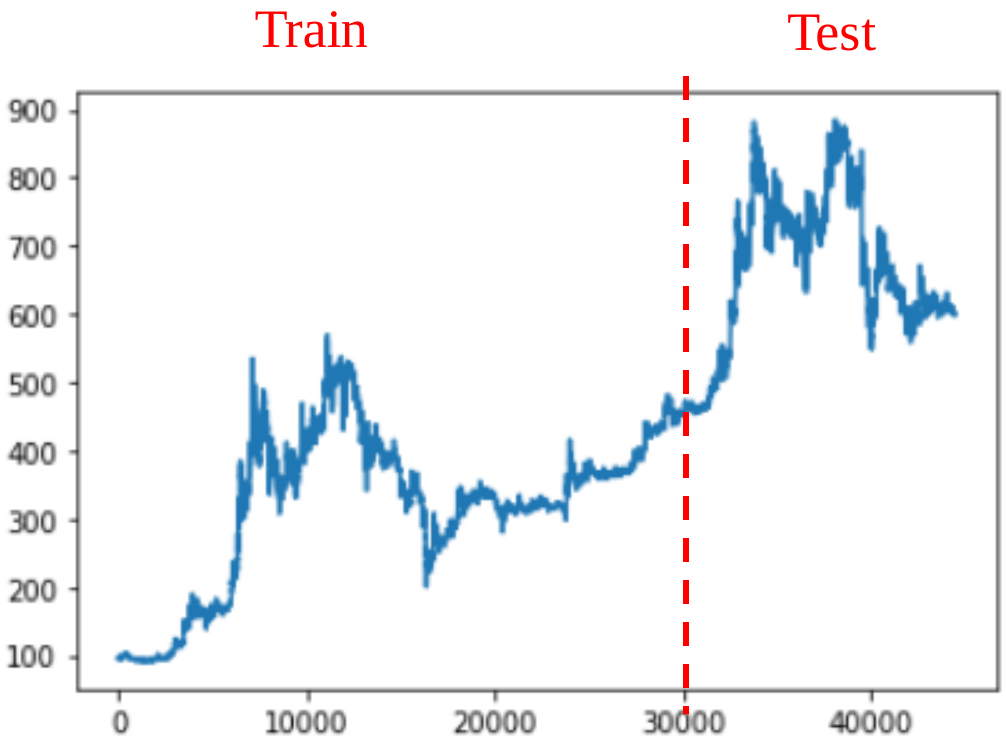
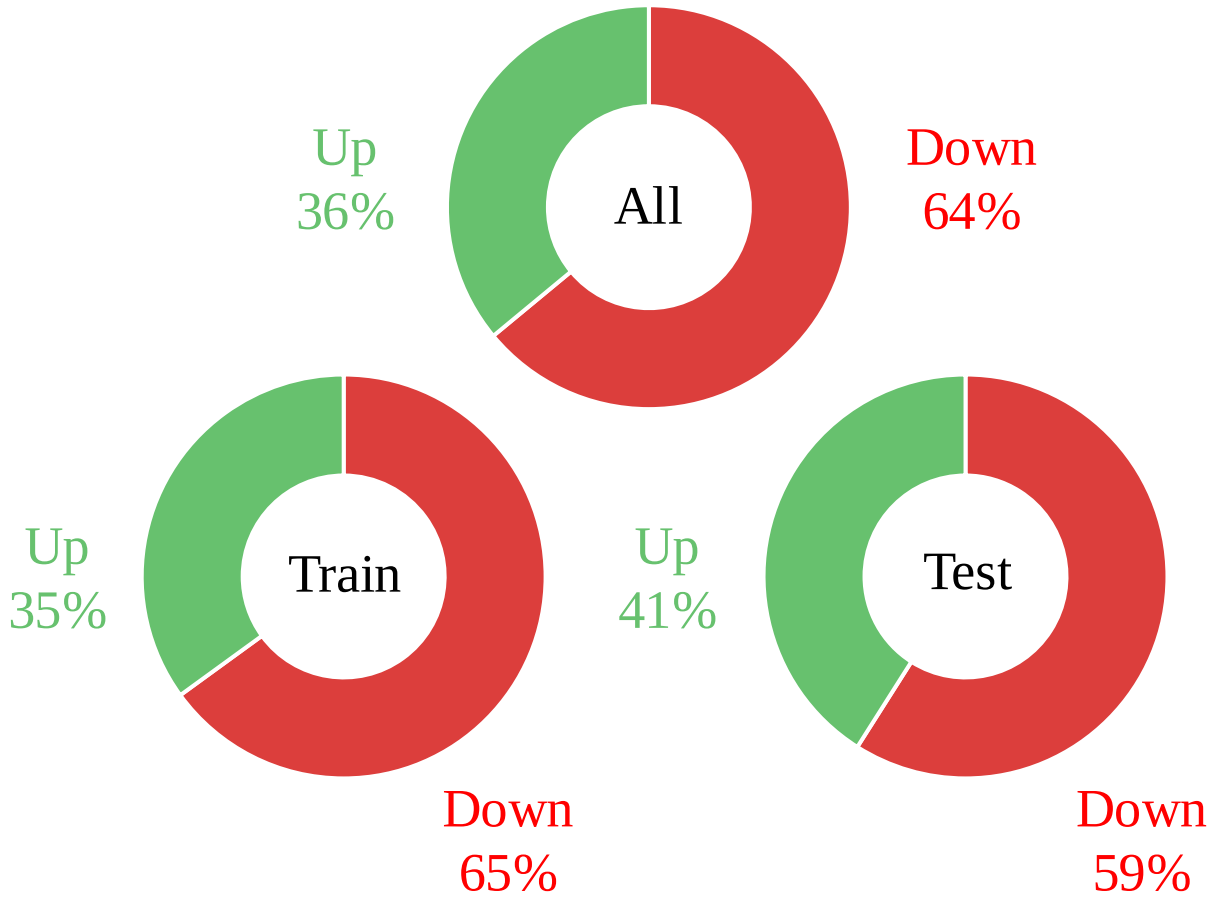


Fig. 2.Up/Down rate



III. Data – OHLCV Table

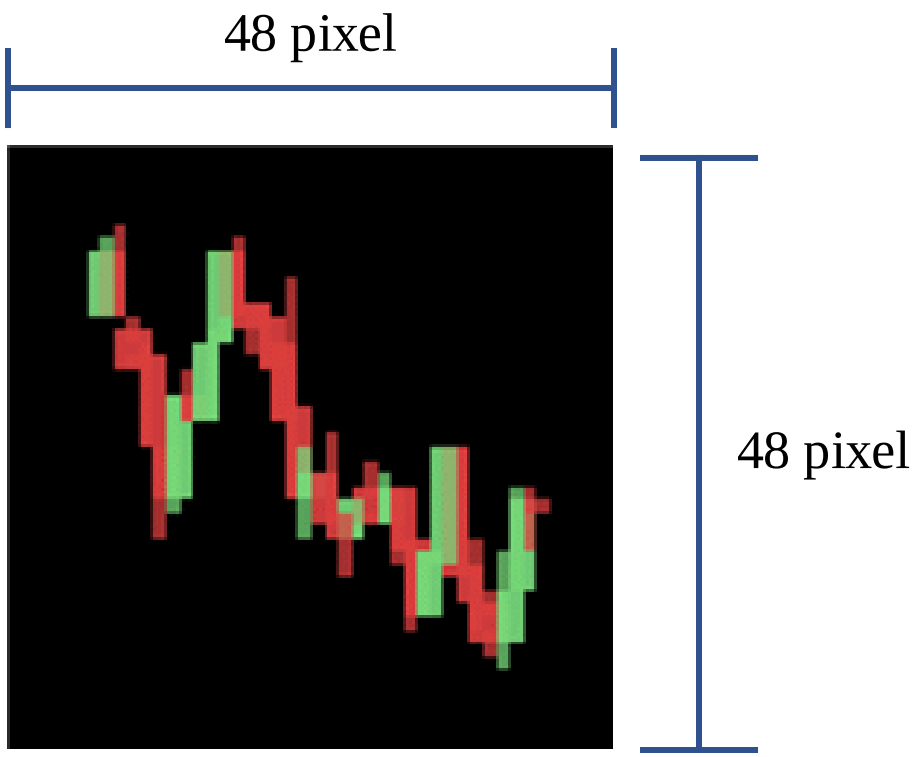
Img. OHLCV data example

	candleDateTimeKst	openingPrice	highPrice	lowPrice	tradePrice	candleAccTradeVolume
1	2021-04-12T00:01:00+09:00	94.5	94.5	94.2	94.3	4.739030e+06
2	2021-04-12T00:02:00+09:00	94.4	94.7	94.3	94.6	4.767889e+06
3	2021-04-12T00:03:00+09:00	94.7	94.8	94.5	94.7	5.741046e+06
4	2021-04-12T00:04:00+09:00	94.6	94.7	94.1	94.2	4.927677e+06
5	2021-04-12T00:05:00+09:00	94.3	94.5	94.1	94.5	2.178284e+06
6	2021-04-12T00:06:00+09:00	94.4	94.4	94.2	94.4	1.924161e+06
7	2021-04-12T00:07:00+09:00	94.4	94.8	94.2	94.7	5.311308e+06
8	2021-04-12T00:08:00+09:00	94.8	94.9	94.5	94.6	3.122199e+06
9	2021-04-12T00:09:00+09:00	94.6	95.2	94.6	95.0	7.563511e+06
10	2021-04-12T00:10:00+09:00	95.0	95.6	94.8	95.6	9.744812e+06

- Length : 44400
- Window_size, Sequence_length : 30
- Open(시가), High(고가), Low(저가), Close(종가), Volume(거래량)
- Shape : (44400,30,5)

III. Data - Candlestic chart image data for CNN

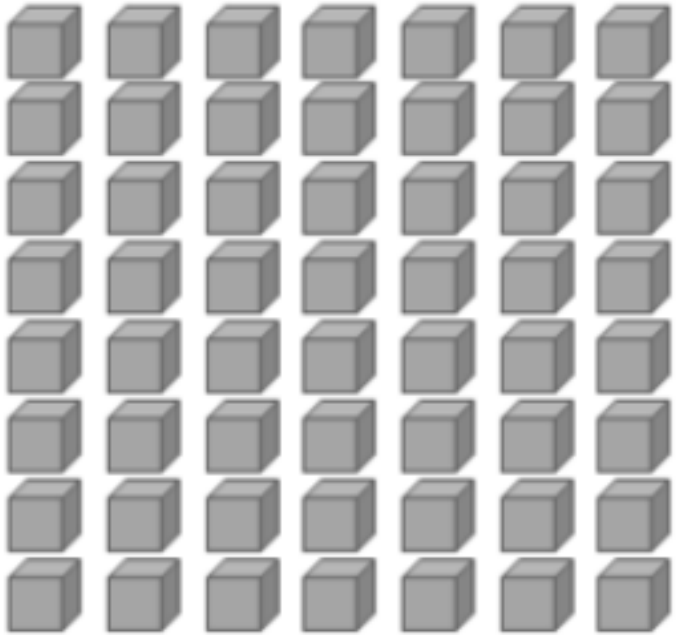
Img. Candlestic chart image



- Length : 44439
- Height x Width : 48x48
- Depth : 3
- Batch_size : 100
- 한 이미지에 30개의 캔들스틱 저장
- Shape : (44439, 48, 48, 3)
- 초록색은 종가가 시가보다 높은 상승
- 빨간색은 종가가 시가보다 낮은 하락

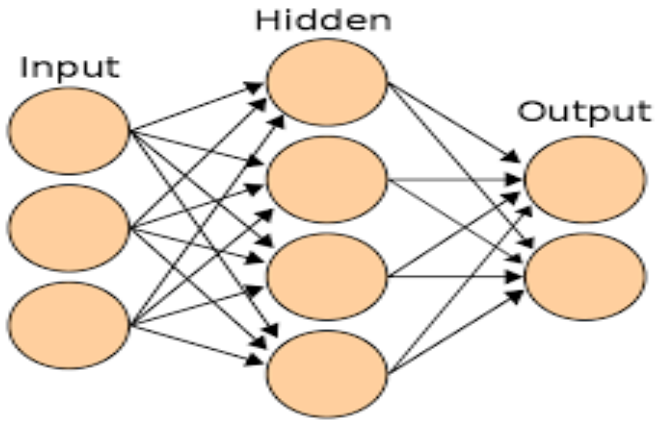
III. Data – 5D tensor data for CNN conbined model

Img. 5D tensor data



- Length : 44439
- Height x Width : 48x48
- Depth : 3
- Batch_size : 100
- Frame: 30
- Shape : (44439, 30, 48, 48,3)

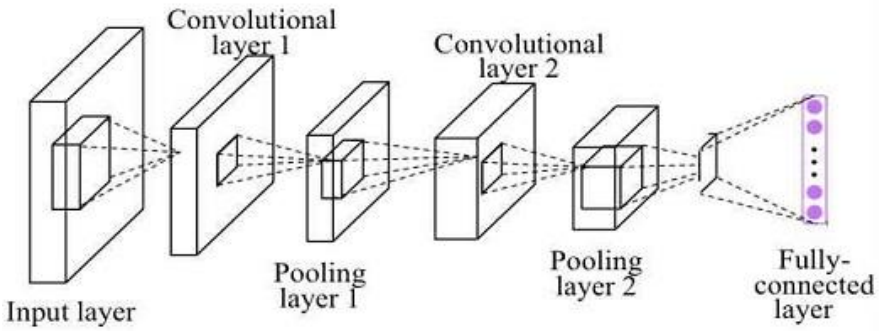
ANN (Artificial Neural Network)



인공신경망(ANN)은 시냅스의 결합으로 네트워크를 형성한 인공뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 가리킨다.

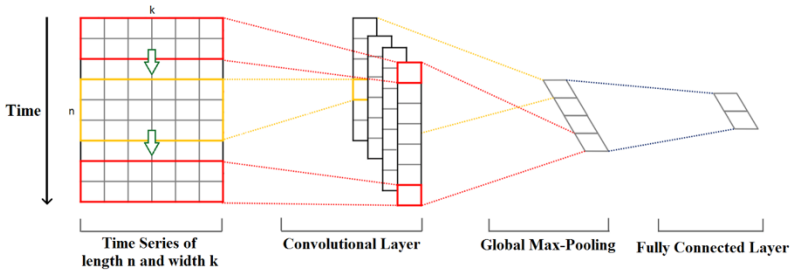
IV. Methodology - Single model

CNN (Convolutional Neural Network)



CNN은 이미지 처리에 탁월한 성능을
보이는 신경망이다. CNN은 크게 합성곱층
(Convolution layer)과 풀링층(Pooling layer)으로 구성된다

1D CNN

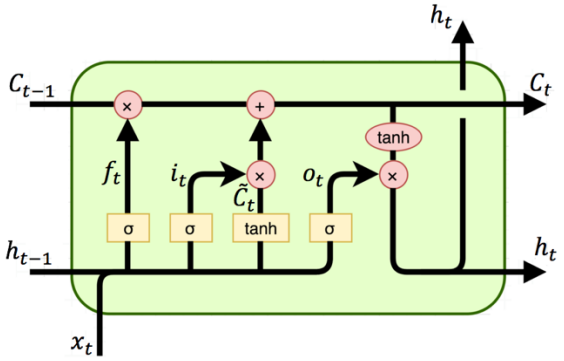


CNN의 일종인 1D CNN은
자연어 처리에 특화되어 있다.
일방향적인 데이터처리와 배열로 인해 일반적인 CNN
과 달리
시계열적 특성이 남아있다.

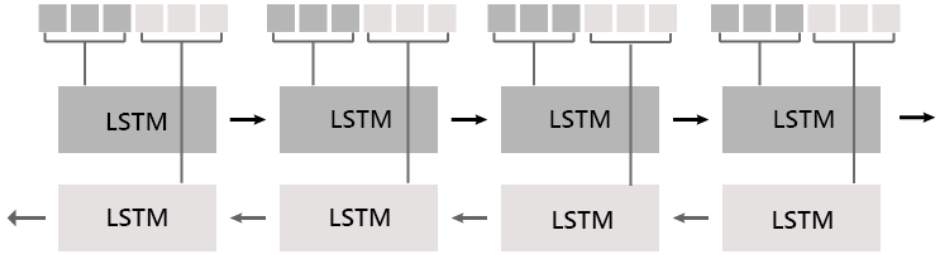
IV. Methodology - Single model

LSTM

(long-short term memory)

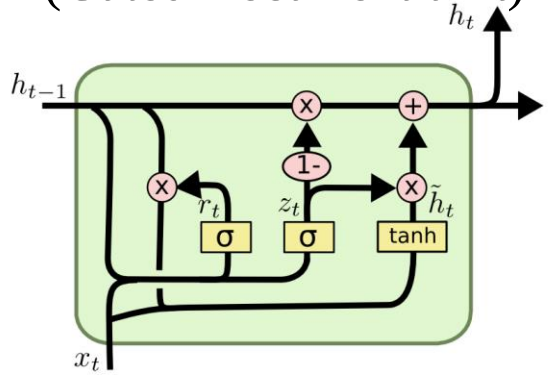


BLSTM (Bi-directional LSTM)



GRU

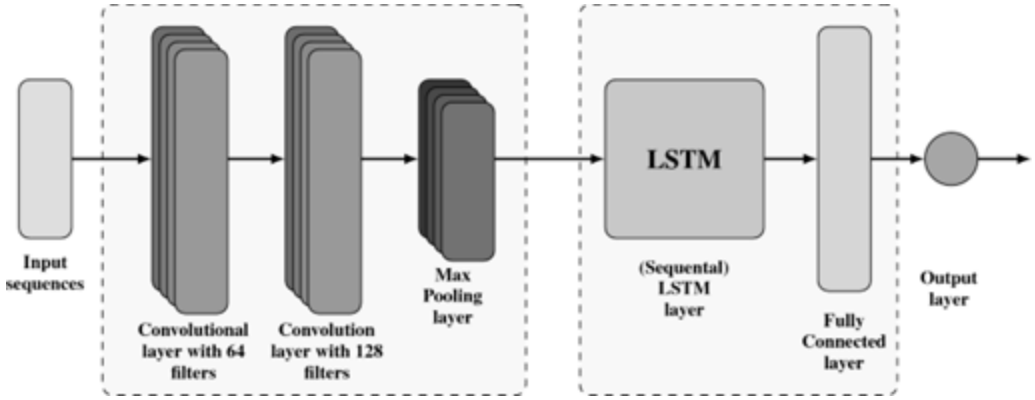
(Gated Recurrent unit)



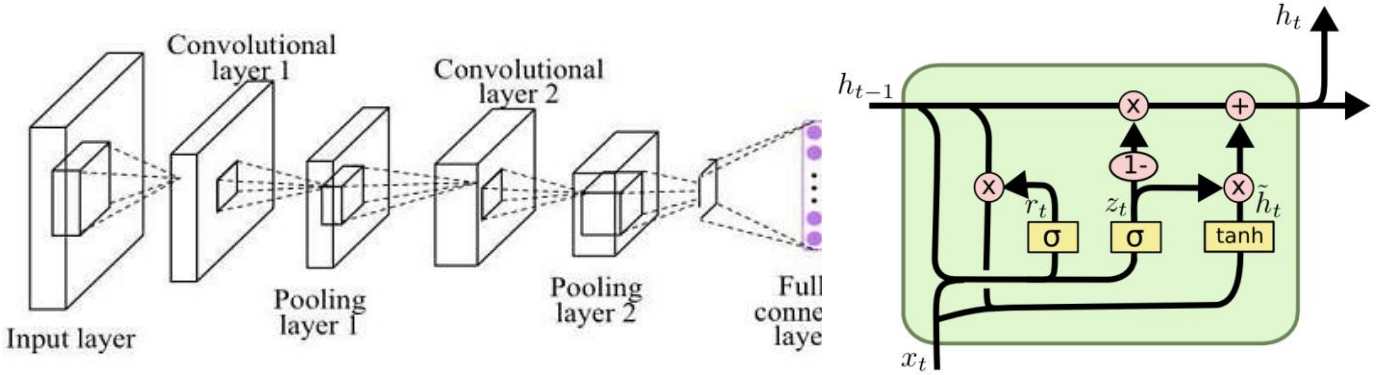
LSTM은 RNN(Recurrent Neural Network)에서 파생된 알고리즘이다. RNN은 입력과 출력을 시퀀스 단위로 처리하는 모델의 일종으로 은닉층의 노드에서 활성화 함수를 통해 나온 결과값을 출력층 방향으로도 보내면서, 다시 은닉층 노드의 다음 계산의 입력으로 보내는 특징을 갖고 있다, 하지만 RNN의 시점(time step)이 길어질수록 앞의 정보가 뒤로 충분히 전달되지 못하는 현상이 발생한다. 이러한 현상을 해결하기 위해서 은닉층의 메모리 셀에 입력 게이트, 망각 게이트, 출력 게이트를 추가하여 불필요한 기억을 지우고, 기억해야 할 것들을 정한 것이 LSTM이다.

IV. Methodology – CNN Conbined model

CNN – LSTM, BLSTM

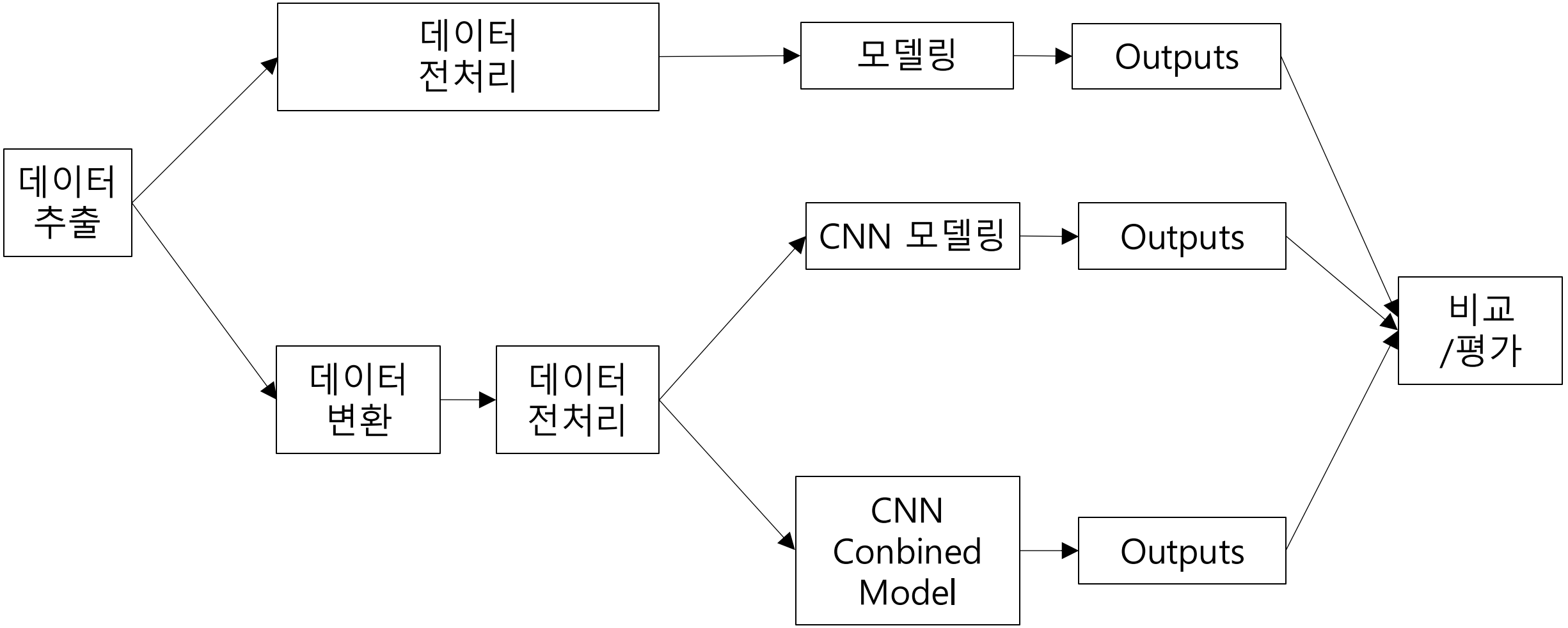


CNN – GRU, BGRU



캔틀스틱 차트 이미지가 입력된 CNN 신경망은 이미지의 패턴을 도출해내고, 도출된 데이터는 LSTM 층을 지나 시계열적 특성을 지닌 채 결과를 나타낸다.

IV. Methodology – Research model design



V. Empirical results - Parameter

Table . hyper parameter

Hyper-parmeter	ANN	CNN1D	GRU	LSTM	BLSTM
Hidden Layer	1	4	1	2	1
Neurons	32	224	50	64	20
Dropout Ratio	-	0.8	-	0.5	0.5
Gate Activation	relu	relu	relu	relu	relu
Recurrent Activation	Sigmoid	Sigmoid	Sigmoid	Sigmoid	sigmoid
Batch Size	30	30	30	30	30
Epochs	10	10	10	10	10
Window Size	5	5	5	5	5
Sequence Length	30	30	30	30	30
Output Dimension	1	1	1	1	1
Optimizer	Adam	Adam	Adam	Adam	Adam
Loss Function	binary_ crossentropy	binary_ crossentropy	binary_ crossentropy	binary_ crossentropy	binary_ crossentropy

V. Empirical results - Parameter

Table . Model hyper parameter

Hyper-parmeter	CNN	CNN-LSTM	CNN-BLSTM	CNN-GRU	CNN-BGRU
Hidden Layer	4	6	6	6	6
Neurons	496	560	560	596	560
Dropout Ratio	0.5	0.5	0.5	0.5	0.5
Gate Activation	relu	relu	relu	relu	Relu
Recurrent Activation	Sigmoid	Sigmoid	Sigmoid	sigmoid	Sigmoid
Batch Size	100	10	10	10	10
Epochs	30	30	30	30	30
Window Size	3x3	30	30	30	30
Sequence Length	30	30	30	30	30
Output Dimension	1	1	1	1	1
Optimizer	Adam	Adam	Adam	Adam	Adam
Loss Function	binary_ crossentropy	binary_ crossentropy	binary_ crossentropy	binary_ crossentropy	binary_ crossentropy

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

- 미래 가격의 증감을 예측하는 선행 연구들이 대부분 정확도(Accuracy)를 성과평가의 척도로 사용
- 해당 연구에서도 성과 평가의 척도로 다음날 종가의 상승이냐 아니냐를 맞추는 BinaryAccuracy를 사용

Table . Train&Test acurracy

모델명/정확도		학습용	평가용
정형 데이터	ANN	55.44%	45.80%
	CNN1D	54.64%	56.65%
	GRU	53.85%	59.08%
	LSTM	65.36%	59.12%
	BLSTM	62.93%	59.12%
	CNN1D-LSTM	65.36%	59.12%
	CNN1D-BLSTM	65.36%	59.12%

Table . Train&Test acurracy

모델명/정확도		Train	Test
이미지 데이터	CNN	70.7%	71.0%
	CNN-LSTM	80.65%	80.8%
	CNN-BLSTM	82.36%	82.5%
	CNN-GRU	75.87%	76.1%
	CNN-BGRU	80.45%	80.6%

- 정형 데이터를 사용한 연구는 50~60%의 정확도를, 이미지 데이터를 사용한 연구는 70~80%의 정확도를 보여 유의미한 차이를 보인다.

- 선행 연구들과 같이 정형 데이터 기반 연구와 이미지 데이터 기반 연구 모두에서 시계열 특성을 고려할 수 있는 LSTM, GRU 모델이 성과가 높게 나타났다.

VI. Conclusion - Empirical results review

- 이미지 데이터를 사용한 연구가 정형 데이터를 사용한 연구보다 좋은 성과를 보였다. 정형 데이터를 사용한 연구는 70%~80%의 정확도를 가졌는데, 이는 데이터 레이블의 불균형을 고려하더라도 높은 성과라고 할 수 있다.

VI. Conclusion - Contribution

- 본 연구에서는 기존 주식 가격 예측에 사용되었던 차트 이미지 기반 딥러닝 모형들을 코인 가격에 적용시켜보았다. 정형데이터를 사용한 연구보다 높은 성과를 보였고, 향후 진행될 이미지 데이터 기반 후속 연구들의 근거가 될 수 있을 것으로 기대

- 최근에 발표되었던 코인 관련 연구를 살펴보게 되면 거의 텍스트 데이터 감성 수치 기반의 연구가 진행되었다. 따라서, 캔들스틱 차트 이미지 데이터를 입력값으로 암호화폐 예측 연구를 진행한 점은 기존 연구들과 차별

- 일봉 데이터를 통한 하루 뒤의 가격 예측 연구가 아닌 분봉 데이터를 사용한 1분 뒤의 가격 예측 연구를 진행한 점도 기존 연구들과 차별

VI. Conclusion - Limitations & Future research

- 우선 본 연구는 딥러닝 모델들을 코인 가격 데이터에 최적화시키지 못했다는 한계점이 있다. 후속 연구를 계속해서 진행하여 하이퍼 파라미터를 충분히 설명할 수 있다면 연구의 정확도는 더욱 높아질 것이다.

- 데이터 레이블이 불균형하고 장기적 추세없이 단기 변동성이 심한 암호화폐의 가격 데이터 특성상 다른 기간에 본 연구에서 사용한 모델을 적용하여 예측/평가를 진행할 시 그 정확도가 본 연구와 다르게 나올 수 있다.

Q&A