한국빅데이터학회지
제4권 제2호, 2019, pp. 1-12

데이터 증강을 통한 딥러닝 기반 주가 패턴 예측 정확도 향상 방안

Increasing Accuracy of Stock Price Pattern Prediction through Data Augmentation for Deep Learning

김영준[1]·김여정[2]·이인선[3]·이홍주[4]†

가톨릭대학교 법정경학부 법학과[1], 가톨릭대학교 컴퓨터정보공학부[2], 가톨릭대학교 수학과[3], 가톨릭대학교 경영학과[4]

요 약

인공지능 기술이 발전하면서 이미지, 음성, 텍스트 등 다양한 분야에 적용되고 있으며, 데이터가 충분한 경우 기존 기법들에 비해 좋은 결과를 보인다. 주식시장은 경제, 정치와 같은 많은 변수에 의해 영향을 받기 때문에, 주식 가격의 움직임 예측은 어려운 과제로 알려져 있다. 다양한 기계학습 기법과 인공지능 기법을 이용하여 주가 패턴을 연구하여 주가의 등락을 예측하려는 시도가 있어왔다. 본 연구는 딥러닝 기법 중 컨볼루셔널 뉴럴 네트워크(CNN)를 기반으로 주가 패턴 예측률 향상을 위한 데이터 증강 방안을 제안한다. CNN은 컨볼루셔널 계층을 통해 이미지에서 특징을 추출하여 뉴럴 네트워크를 이용하여 이미지를 분류한다. 따라서, 본 연구는 주식 데이터를 캔들스틱 차트 이미지로 만들어 CNN을 통해 패턴을 예측하고 분류하고자 한다. 딥러닝은 다량의 데이터가 필요하기에, 주식 차트 이미지에 다양한 데이터 증강(Data Augmentation) 방안을 적용하여 분류 정확도를 향상 시키는 방법을 제안한다. 데이터 증강 방안으로는 차트를 랜덤하게 변경하는 방안과 차트에 가우시안 노이즈를 적용하여 추가 데이터를 생성하였으며, 추가 생성된 데이터를 활용하여 학습하고 테스트 집합에 대한 분류 정확도를 비교하였다. 랜덤하게 차트를 변경하여 데이터를 증강시킨 경우의 분류 정확도는 79.92%였고, 가우시안 노이즈를 적용하여 생성된 데이터를 가지고 학습한 경우의 분류 정확도는 80.98%이었다. 주가의 다음날 상승/하락으로 분류하는 경우에는 60분 단위 캔들 차트가 82.60%의 정확도를 기록하였다.

■ 중심어 : 주가 패턴, 딥러닝, 컨볼루셔널 뉴럴 네트워크, 데이터 증강, 가우시안 노이즈

Abstract

As Artificial Intelligence (AI) technology develops, it is applied to various fields such as image, voice, and text. AI has shown fine results in certain areas. Researchers have tried to predict the stock market by utilizing artificial intelligence as well. Predicting the stock market is known as one of the difficult problems since the stock market is affected by various factors such as economy and politics. In the field of AI, there are attempts to predict the ups and downs of stock price by studying stock price patterns using various machine learning techniques. This study suggest a way of predicting stock price patterns based on the Convolutional Neural Network(CNN) among machine learning techniques. CNN uses neural networks to classify images by extracting features from images through convolutional layers. Therefore, this study tries to classify candlestick images made by stock data in order to predict patterns.

2019년 11월 21일 접수; 2019년 12월 10일 수정본 접수; 2019년 12월 31일 게재 확정.
† 교신저자 (hongjoo@catholic.ac.kr)

This study has two objectives. The first one referred as Case 1 is to predict the patterns with the images made by the same-day stock price data. The second one referred as Case 2 is to predict the next day stock price patterns with the images produced by the daily stock price data. In Case 1, data augmentation methods - random modification and Gaussian noise - are applied to generate more training data, and the generated images are put into the model to fit. Given that deep learning requires a large amount of data, this study suggests a method of data augmentation for candlestick images. Also, this study compares the accuracies of the images with Gaussian noise and different classification problems. All data in this study is collected through OpenAPI provided by DaiShin Securities. Case 1 has five different labels depending on patterns. The patterns are up with up closing, up with down closing, down with up closing, down with down closing, and staying. The images in Case 1 are created by removing the last candle(-1candle), the last two candles(-2candles), and the last three candles(-3candles) from 60 minutes, 30 minutes, 10 minutes, and 5 minutes candle charts. 60 minutes candle chart means one candle in the image has 60 minutes of information containing an open price, high price, low price, close price. Case 2 has two labels that are up and down. This study for Case 2 has generated for 60 minutes, 30 minutes, 10 minutes, and 5minutes candle charts without removing any candle.

Considering the stock data, moving the candles in the images is suggested, instead of existing data augmentation techniques. How much the candles are moved is defined as the modified value. The average difference of closing prices between candles was 0.0029. Therefore, in this study, 0.003, 0.002, 0.001, 0.00025 are used for the modified value. The number of images was doubled after data augmentation. When it comes to Gaussian Noise, the mean value was 0, and the value of variance was 0.01. For both Case 1 and Case 2, the model is based on VGG-Net16 that has 16 layers.

As a result, 10 minutes -1candle showed the best accuracy among 60 minutes, 30 minutes, 10 minutes, 5minutes candle charts. Thus, 10 minutes images were utilized for the rest of the experiment in Case 1. The three candles removed from the images were selected for data augmentation and application of Gaussian noise. 10 minutes -3candle resulted in 79.72% accuracy. The accuracy of the images with 0.00025 modified value and 100% changed candles was 79.92%. Applying Gaussian noise helped the accuracy to be 80.98%. According to the outcomes of Case 2, 60minutes candle charts could predict patterns of tomorrow by 82.60%.

To sum up, this study is expected to contribute to further studies on the prediction of stock price patterns using images. This research provides a possible method for data augmentation of stock data.

■ Keyword : Stock Price Pattern, Deep Learning, Convolutional Neural Network, Data Augmentation, Gaussian Noise

Ⅰ. 서 론

딥러닝을 활용하는 연구가 다양한 분야에서 진행되고 있으며, 주식 시장에 대한 예측과 분류에도 많이 적용되어 왔다(송현정, 이석준, 2018). 본 논문은 딥러닝을 활용하여 주가 패턴을 예측하는 문제에서 데이터 증강(Data Augmentation)을 통해 학습 데이터를 추가로 생성하여, 모형의 정확도를 높이는 방안을 제시한다. 주가 데이터는 실제 데이터만을 활용하는 경우에는 기간이 오래되더라도 학습에 많은 데이터를 필요로 하는 딥러닝 기법 적용에 충분한 데이터 확보가 어렵다. 따라서, 본 연구에서는 이미지 기반의 다양한 데이터 증강 방안을 고려하여 주가 패턴 정확도를 향상시키는 방안을 파악한다. 데이터는 대신증권 API를 이용하여 분 단위 데이터를 추출하였고, 전처리를 거쳐 노이즈 제거, 라벨링 및 캔들 차트 생성은 Plot.ly 라이브러리를 이용하였다. 하루의 주가 움직임을 캔들 차트로 그렸으며, 5분, 10분, 30분, 60분 단위로 다양하게 차트를 생성하였다. 분류 문제는 하루의 주가 패턴을 분류하는 문제와 내일의 주가가 상승할 지 하락할 지 분류하는 두 가지 문제로 구성하였다. 첫 번째 문제는 하루의 주가 패턴을 분류한 후에 과거 데이터를

가지고 학습하고, 캔들 차트에서 마지막 캔들(-1candle), 마지막에서 두번째 캔들(-2candle), 마지막에서 세번째(-3candle)을 제거한 테스트 데이터에 대해서 주가 패턴을 분류하는 방식으로 구성되어 있다. 두번째 분류 문제는 하루의 주가 캔들 차트를 가지고 내일의 주가 상승 여부를 분류하는 문제로 구성하였다.

CNN 기법을 통해서 모형을 학습하였으며, 데이터 증강방안으로는 주가 캔들 차트를 무작위로 변경하여 추가 데이터를 생성하는 방안과 가우시안 노이즈(Gausian Noise)(Hussain et al, 2018)을 적용하여 추가 데이터를 생성하는 방안을 활용하였다. 데이터 증강을 활용하지 않은 경우에 분류 범주에 따라 높은 정확도를 보이는 경우는 데이터의 양이 많은 경우임을 알 수 있었으며, 데이터 증강기법을 이용해서 데이터를 늘리고 학습하여 정확도를 비교하였다.

제 2장에서는 관련 연구에 대해 살펴보고, 제3장에서는 본 논문의 연구 방법에 대해 기술하였다. 제4장에서는 실험 및 결과를 비교하고 마지막으로 제5장에서 결론을 제시한다.

Ⅱ. 관련연구

기계학습을 활용한 주가 예측에는 Support Vector Machine, 유전자 알고리즘 등이 활용되어 왔다. 주가와 관련하여 다양한 데이터를 기반으로 주식과 관련된 정보 분석과 주가 예측 연구가 진행되었다. 다양한 원천의 텍스트 데이터 분석을 통해 주가 움직임을 예측하려는 연구들이 진행되었다(Jeong et al., 2015; Kim et al., 2012; Oh and Sheng, 2011; Schumaker, 2009). 뉴스 기사나 소셜 미디어의 반응에 따라 주식 거래를 하고 이에 따른 수익률을 분석하는 연구들이 진행되어 왔다(Ding et al. 2014). 뉴스 기사에 등장하는 단어를 고려하여 주가 변동을 예측 하는 방안들도 제시되었다(이민식, 이홍주, 2016).

시계열 데이터에 기반한 분석 방법으로 Jeantheau (2004)는 ARCH 모형, Amilon(2003)는 GARCH 모형을 통해 주가 예측을 진행했다. Park and Shin (2011)은 기존의 시계열 분석에 사용하는 정보와 함께 타기업이나 각종 경제지표들을 바탕으로 기계학습을 통한 주가 예측을 진행했다. 기계학습의 발전과 함께 인공신경망 기반의 주가 예측 연구도 진행되었다. Lee(2008)은 회귀모형, 인공신경망, SVM 모형을 결합하여 결합모형을 통한 주가 예측을 진행했다.

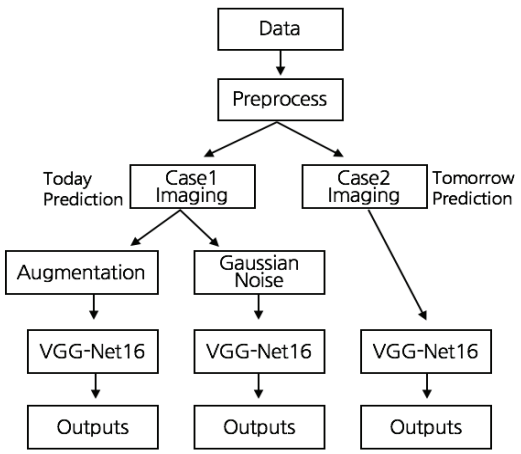
딥러닝 방안 또한 주가 예측에 적용되었는데 대표적으로 Long Shot-Term Memory(LSTM), Recurrent Neural Network(RNN)이 활용되었다(신동하, 최광호, 김창복, 2017). 딥러닝 방법 중 하나인 Convolutional Neural Network(CNN)은 이미지와 영상을 분류하는 문제에 사용되고 있다(LeCun et al., 1998). CNN 방안을 주가 예측에 적용하기 위한 방안으로 주가 움직임을 이미지로 표현하여야 하는데, 이에 캔들스틱 차트(Candlesticks Chart)로 표현하는 방안이 많이 이용된다. 캔들스틱 차트는 주가 움직임의 패턴을 시각화 하고, 패턴에 따라 매매 규칙을 정의하여 주식 매매 타이밍을 올바르게 결정할 수 있도록 돕는다(이강희, 양인실, 조근식, 1997). 캔들스틱 차트 이미지를 CNN 방안을 통해 학습하여 주가 등락을 예측하는 연구도 수행되어왔다(이모세, 2018). 그리고 상승, 하락 패턴을 이용하여 다양한 매매 빈도 패턴과 예측 시점 탐색 및 주가의 등락을 예측하는 연구도 진행되었다(송현정, 이석준, 2018).

Ⅲ. 연구방법

본 연구는 <Figure 1> 에 나온 것처럼 진행되었다. 주가 예측 실험은 두 가지로 나뉘어 진행하

였다. 첫 번째인 Case 1은 당일 주가 데이터로 생성한 캔들 이미지를 학습하여 당일 주가 패턴 예측에 관한 것이다. Case 2는 전일의 주가 데이터로 생성한 캔들 이미지로 다음날의 주가 상승 여부를 예측하는 것이다.

전처리를 통해서 주식 데이터의 노이즈를 제거한 후 이미지를 생성하였다. Case 1에서는 당일 주가 데이터를 통해 주가 패턴을 확인하여 라벨링한 후에 이미지를 생성하였다. 주가 패턴은 다음 절에서 자세히 설명하겠다. 데이터 증강 방안을 적용하여 학습데이터를 추가로 생성한 후에 딥러닝 모델에 학습시켰다. 그 후 테스트 데이터로 생성된 모델의 정확도를 확인하였다. Case 2는 전일 주가를 가지고 학습한 후에 다음날의 주가 상승 여부를 예측하는 것이기에 전일 이미지에 다음날 주가 상승여부에 대한 라벨링을 진행하였다. 그 후 라벨링된 이미지를 딥러닝 모델에 학습시키고 테스트 데이터로 정확도를 확인하였다.



〈Figure 1〉 Research Process

3.1 패턴 및 라벨링

수집한 주가 데이터를 이미지화할 때, 해당 이미지가 어떤 패턴인지를 함께 레이블링 해야 한다. Case 1의 경우에는 패턴을 <Table 1>과 같이 구분하였다. 패턴은 당일 주가가 상승하였는지 하락하였는지에 따라 상승(Up), 하락(Down)을 구분하고, 맨 마지막 캔들이 전 캔들보다 상승하면서 마감하였는지 하락하면서 마감하였는지에 따라 상승마감, 하락마감을 구분하였다. 패턴은 5개로 구분하였으며 당일 상승 하고 상승마감 혹은 하락마감한 패턴 두 개와 당일 하락 하고 상승마감 혹은 하락마감한 패턴 두 개 그리고 횡보이다. Case 2의 경우 패턴은 상승과 하락으로만 구분하였다.

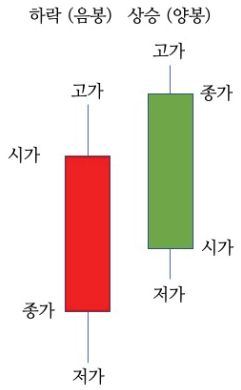
상승과 하락의 판단은 당일 마지막 캔들의 종가에서 첫 캔들의 시작가를 뺐을 때 양수이면 상승, 음수이면 하락이라 판단한다. 상승마감 혹은 하락마감 식도 마찬가지로 당일 마지막 캔들 종가에서 하나 전 캔들의 종가를 뺐을 때, 양수이면 상승 음수이면 하락이라 판단한다. 횡보는 당일의 시작가와 종가가 같았을 때이다. Case 2에서 상승하락은 전일 종가에서 다음날 종가를 뺐을 때 양수이면 상승, 음수이면 하락으로 구분하였다.

〈Table 1〉 Labels & Patterns

Case1		Case2	
Labels	**Patterns**	**Labels**	**Patterns**
Label_1	Up, Up Closing	Label_0	Up
Label_2	Up, Down Closing	Label_1	Down
Label_3	Down, Up Closing		
Label_4	Down, Down Closing		
Label_5	Staying		

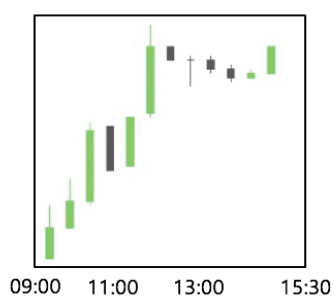
3.2 주가 이미지 생성

캔들스틱 차트(Candlestick Charts) 이미지를 만들기 위해서 'plotly' 패키지를 사용하여 jpg 형태로 저장하였다. 주가 데이터는 종목코드, 일자, 시간, 시가, 고가, 저가, 종가를 5분, 10분, 30분, 60분 단위로 수집하여 캔들스틱 차트 생성에 사용하였다. 하나의 캔들은 특정 시간 단위 동안의 시가, 고가, 저가, 종가 정보를 담고 있다. 상승, 하락에 따라서 캔들은 다음과 같은 정보를 포함하고 있다.



〈Figure 2〉 Candlestick chart

캔들스틱 이미지를 생성할 때 패턴 예측을 위해서 Case 1에서는 마지막 캔들(-1candle)만 제거한 이미지, 마지막과 마지막에서 두 번째 캔들(-2candles)을 제거한 이미지, 마지막에서부터 마지막 세 번째 캔들(-3candles)까지를 제거한 이미지를 각 분 단위마다 각각 생성하였다. 이렇게 제거된 이미지로 학습하여 해당 이미지가 어떤 주가 패턴을 보이는 지를 구분하는 데 활용되었다. <Figure 3>는 30분 단위로 주가 데이터를 캔들스틱 차트로 생성한 경우이며, 마지막 캔들(-1candle)이 제거된 상태이다. Case 2는 전일의 주가 데이터를 모두 사용하여 이미지를 생성하였다. <Figure 3>에서 색이 알려주는 의미는 초록색은 특정 분 단위의 종가가 시가보다 높은 상승을 뜻하며, 회색은 특정 분 단위의 종가가 시가보다 낮은 하락을 표현한다.



〈Figure 3〉 Image without the last candle(−1candle)

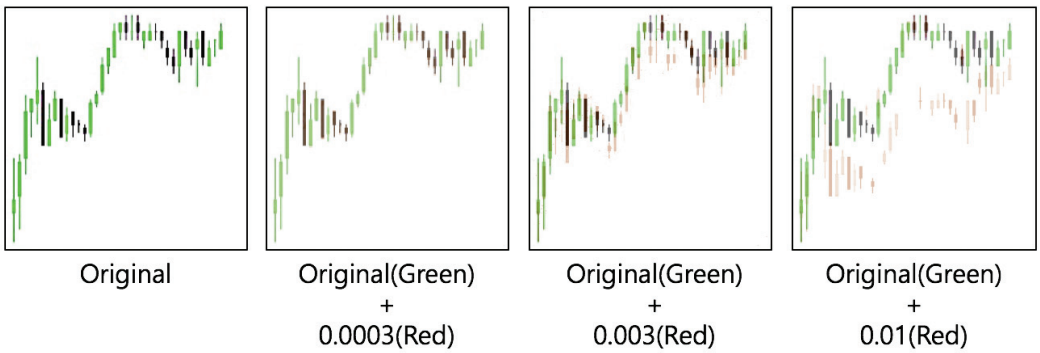
3.3 주가 이미지 데이터 증강

하루의 주가 움직임으로 하나의 이미지를 만들었기에 10년간의 주가 데이터를 수집한다고 하더라도 최대 30,000개 정도의 이미지 데이터를 만들 수 있다. 이미지 분류 문제의 성과를 개선하기 위해서는 학습 데이터를 더 많이 활용하는 것이 필요하다. 이를 위해 유사한 이미지의 데이터를 생성하여 모형의 학습에 활용하는 것이 데이터 증강이다.

이미지 데이터 증강은 보통 이미지를 반전하거나, 회전 혹은 특정 부분을 잘라서 추가 데이터를 생성하는 방식으로 이루어진다. 하지만, 주가 움직임을 나타내는 캔들스틱 차트에서는 이미지의 반전이나 회전, 크기 조절 등으로는 의미있는 추가 데이터 생성이 어렵다.

본 연구에서는 주가의 움직임을 나타내는 캔들스틱 차트에서 특정 캔들을 무작위로 움직여서 추가 이미지 데이터를 생성하는 방법을 활용하였으며, 절차는 다음과 같다.

1) 캔들스틱 차트에서 맨 왼쪽에 위치하는 첫 번째 캔들을 제외하고 나머지 캔들중에서 수정할 캔들을 100%, 50%, 10% 중에서 무

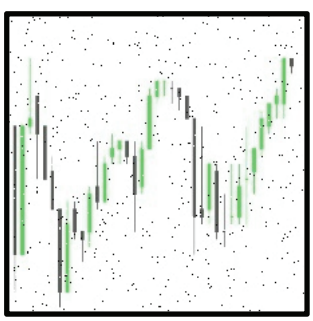


〈Figure 4〉 Data Augmentation

작위로 선택한다. 10%의 의미는 첫번째 캔들을 제외하고 나머지 캔들의 수의 10%에 해당하는 캔들을 선정한다는 것이다.

2) 무작위로 선정된 캔들을 상하로 어느 정도 움직일 것인지 정하는 값을 수정값이라 정의하였다. 수집한 데이터의 전체 종목에서 캔들 간의 상하 움직임의 평균 값이 0.0029였다. 이를 반올림하여 무작위로 선정된 캔들의 상하를 0.003만큼 수정해 보았을 때, 주식 차트의 모형을 어느 정도 유지하였다(<Figure 4>참조). 반면에 0.01만큼 수정해 본 경우에서는 캔들스틱 차트와 많이 차이가 있었다. 따라서, 본 연구에서 캔들의 상하 움직임 수정값을 0.003 0.002, 0.001, 0.00025로 정하였다.

3) 수정될 캔들의 수(100%, 50%, 10%)와 캔들의 움직임 정도(0.003 0.002, 0.001, 0.00025)에 따라 주가 이미지 데이터를 추가로 생성하였다.



〈Figure 5〉 Image with Gaussian Noise

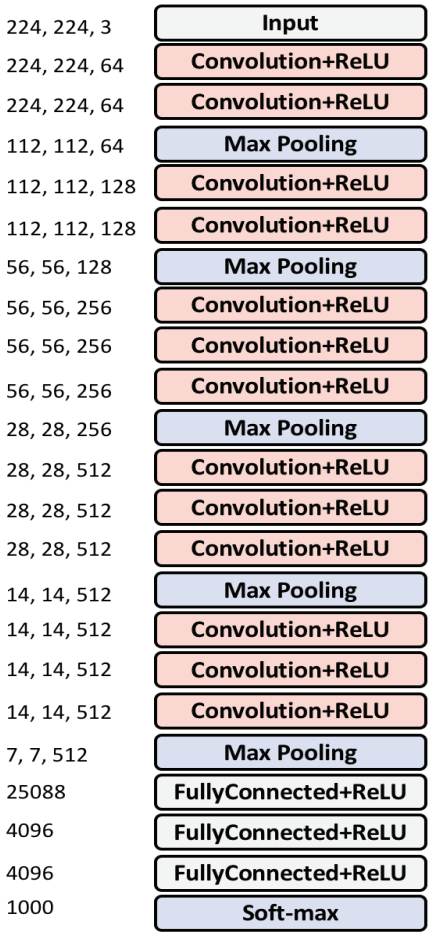
가우시안 노이즈(Gaussian Noise)는 <Figure 5>와 같이 주가 캔들스틱 차트에 노이즈를 추가하여 새로운 이미지를 만드는 방안이다. 기존 이미지에 평균 0, 분산 0.01만큼의 가우시안 노이즈 레이어를 추가하여, 추가 학습 이미지를 생성하였다.

3.4 딥러닝 모형

본 연구에서는 딥러닝 모형으로 Convolutional Neural Network (CNN)을 사용하였다. CNN은 이미지 형태의 데이터를 분류하는데 주로 사용되고 있는 딥러닝 모형이며, 컨볼루셔널 계층을 통해 이미지에 대한 특징을 추출하여 뉴럴 네트워크를 이용하여 분류가 이루어진다. 본 연구는 CNN 모형에서도 이미지 분류의 성과가 검증된 아키텍쳐인 VGGNet-16(<Figure 6 참조>)을 이용하였다. VGGNet-16은 단순한 구조이면서 동시에 단일 네트워크에서 더 좋은 성능을 보여, 다양한 분야에 폭넓게 활용되고 있다(고광은, 심귀보, 2017).

데이터는 85:15의 비율로 학습 데이터 셋(training data set)과 테스트 데이터(test data set)로 나누었다. 주식 데이터는 시간을 고려해야 하는 데이터이기

때문에, 과거 데이터로 학습하고 학습에 사용되지 않은 미래 데이터로 테스트가 이루어져야 한다.



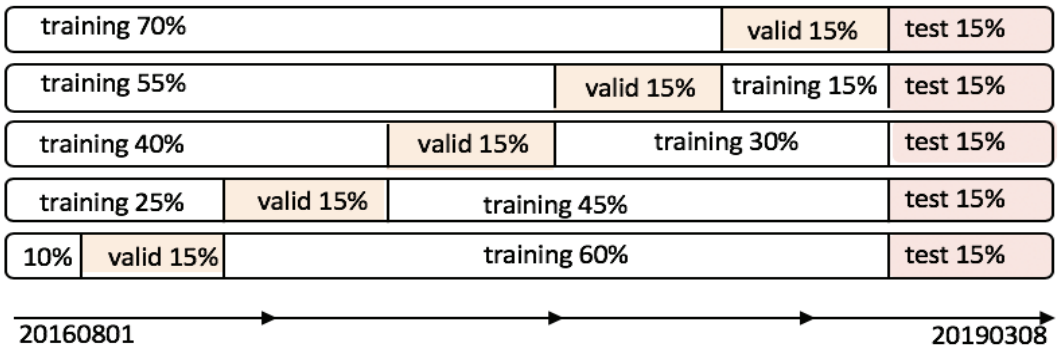
〈Figure 6〉 VGG-Net16

따라서, 전체 데이터에서 시간을 기준으로 최신의 15% 데이터를 테스트 데이터로 고정하였다. 성과의 교차검증을 위해 나머지 데이터에서 학습 데이터 셋과 검증 데이터(valid data)를 5가지 방식으로 나누었다. 테스트 데이터를 제외한 나머지 데이터에서 시간을 기준으로 최신으로부터 왼쪽으로 15%씩 움직이며 검증 데이터 집합으로 삼았다. 테스트 데이터를 제외한 데이터가 85%이기에 정확하게 5등분을 하지는 않았다. 이를 정리하면 <Figure 7>과 같다.

Ⅳ. 실험 및 결과

4.1 데이터 수집

본 연구에서는 대신증권 API를 이용하여 KOSDAQ 상위 150개 종목에 대해 종목코드, 일자, 시간, 시가, 고가, 저가, 종가를 5분, 10분, 30분, 60분 단위로 데이터를 수집하였다. 데이터 전처리를 통해서 노이즈 제거 후 104개 종목에 대한 데이터를 실험에 활용하였다. 데이터 수집 기간은 2016년 8월 1일부터 주식시장 종료 시간이 15시에서 15시 30분으로 변경되었기에, 2016년 8월 1일부터 2019년 3월 8일까지 수집했다. <Figure 8>는 수집된 데이터의 예이다.



〈Figure 7〉 Cross Validation

code	day	time	open_price	high_price	low_price	close_price	volume
A023410	20190308	15:30	7080	7120	7060	7060	49105
A023410	20190308	15:00	7140	7150	7040	7080	70080
A023410	20190308	14:00	7030	7180	7020	7130	176577
A023410	20190308	13:00	7010	7040	7000	7020	50814
A023410	20190308	12:00	6940	7030	6920	7000	76622
A023410	20190308	11:00	6980	6990	6930	6940	80864
A023410	20190308	10:00	6820	7060	6800	6980	179884
A023410	20190307	15:30	6940	6970	6860	6860	99245
A023410	20190307	15:00	7010	7020	6920	6930	118545
A023410	20190307	14:00	7040	7040	7000	7010	30575
A023410	20190307	13:00	7060	7080	7000	7040	34346
A023410	20190307	12:00	7050	7080	7020	7070	34931
A023410	20190307	11:00	7030	7130	7020	7050	40883
A023410	20190307	10:00	7130	7150	6960	7040	143130

〈Figure 8〉 Examples of Collected Data

4.2 원본 데이터 수

Case 1와 Case 2 실험 수행을 위해 생성한 캔들스틱 차트의 원본 이미지 수는 <Table 2>와 같다.

〈Table 2〉 Number of Images

Minutes	Case1	Case2
60 minutes	64511	64407
30 minutes	64951	64847
10 minutes	63155	63051
5 minutes	56859	56755

<Table 3>과 <Table 4>는 생성된 캔들스틱 이미지의 라벨별 분포를 보여준다. <Table 3>에서 캔들스틱차트의 단위시간이 60분에서 5분으로 줄어들면서 Label_1과 Label_3의 이미지가 급격하게 줄어들었다. 실제로 단위를 작게 할수록 하락마감이 더 많아진다는 의미이다. 반면, <Table 4>를 보면 Case 2는 Label_0과 Label_1이 각각 대략 46%, 53% 정도의 비율로 데이터가 균등하게 분포되어 있음을 볼 수 있다.

〈Table 3〉 Number of Images per Labels – Case 1

Labels	60 minutes	30 minutes	10 minutes	05 minutes
Label_1	16518	16683	123	104
Label_2	12400	12481	28318	25615
Label_3	11235	11330	84	65
Label_4	21905	21991	32286	29068
Label_5	2453	2466	2344	2007
Total	64511	64951	63155	56859

〈Table 4〉 Number of Images per Labels – Case 2

Labels	60 minutes	30 minutes	10 minutes	05 minutes
Label_0	29912	30104	29347	26538
Label_1	34495	34743	33704	30217
Total	64407	64847	63051	56755

4.3 학습 및 결과 비교

VGGNet-16을 이용해서 이미지를 학습하기 위해서 배치 크기(Batch Size)는 32, 학습률(Learning rate)는 0.01을 사용하였고, 가중치는 선행 학습된 가중치를 초기값으로 사용하였다. 반복 학습 횟수(Epoch)는 50에서 수렴하는 것을 확인할 수 있었기에 50으로 설정하였다.

Case 1에서 데이터 증강이나 가우시안 노이즈를 적용하지 않은 원본 이미지 집합을 먼저 학습하여 성과를 측정해 보았다. <Table 5>를 보면 10분 단위로 캔들스틱 차트 이미지를 생성하고 맨 마지막 하나의 캔들을 제거한 -1캔들 이미지를 가지고 학습한 모형이 0.808(80.8%)의 가장 높은 정확도를 보였다. 10분 단위 캔들스틱 차트에서 -1캔들, -2캔들, -3캔들의 분류정확도는 큰 차이가 없고, 마지막 3개의 캔들을 제거한 이미지를 가지고 주가 패턴을 예측하는 것이 더 큰 의미가 있다고 판단하였다. 따라서, 데이터 증강 및 가우시안 노이즈를 적용할 이미지는 10분 단위로 생성된 캔들스틱 차트와 마지막 세개의 캔들을 제거한 -3캔들을 사용하였다.

데이터 증강을 위해 캔들의 상하 움직임 값을 뜻하는 수정값을 다양하게 하여 증강 이미지 데이터를 생성하였다. 수정값으로 0.003, 0.002, 0.001, 0.00025를 적용하여 이미지를 만들었으며, 각각 원본데이터와 동일한 수의 추가 이미지를 생성하였다. 원본 데이터와 증강된 데이터를 활용하여 모형을 학습하였고, 성과를 측정하였다. <Table 6>은 테스트 집합의 분류 성과이며, 학습 데이터를 학습과 검증 집합으로 구분하는 것은 한번만 수행하였다.

수정값 0.003과 0.00025에서 동일하게 가장 높은 정확도를 보여주었다. 두 수정값을 활용하여 이미지 데이터 증강방안에 적용하였다. 무작위로 수정되는 캔들의 수는 100%, 50%, 10%로 나누어 적용하였다. <Figure 7>의 교차 검증방안을 적용하여 학습한 모형의 분류 정확도 결과가 <Tabel 7>이다. 수정값 0.00025로 캔들을 모두 (100%) 수정한 증강 이미지 데이터와 원본 이미지 데이터를 가지고 학습한 경우가 가장 높은 정확도를 보였다.

가우시안 노이즈는 10분 단위로 그린 캔들스

〈Table 5〉 Accuracy of Original Images – Case 1

Labels	60 minutes	30 minutes	10 minutes	05 minutes
-1candle	0.5266	0.4961	0.8080	0.7779
-2candles	0.4997	0.4763	0.7963	0.7781
-3candles	0.4930	0.4684	0.7972	0.7769

〈Table 6〉 Accuracy of Augmented Data with 10min –3candles

Modified Value	Accuracy
0.003	0.8030
0.002	0.7997
0.001	0.7898
0.00025	0.8030

〈Table 7〉 Accuracy of Modified Value

Modified Value	100%	50%	10%
0.003	0.7981	0.7960	0.7963
0.00025	0.7992	0.7974	0.7929

틱 차트에서 마지막 세개의 캔들을 제거한 (-3캔들) 원본 이미지와 수정값 0.003으로 50%의 캔들을 수정한 증강 이미지에 적용해보았다. 가우시안 노이즈를 적용했을 때는 CNN 모형이 50 Epoch까지 값이 수렴하지 않아서 200 Epoch까지 모형을 학습하였다. 그 결과 10분 -3캔들의 원본 이미지에 가우시안 노이즈를 적용한 경우의 분류 정확도는 0.8087이 나왔으며, 원본 이미지와 함께 데이터 증강(수정값 0.003, 50% 적용) 이미지에도 가우시안 노이즈를 적용한 경우의 분류 정확도는 0.8098이 나왔다(<Table 8>참조).

Case 2는 전일의 주가 데이터로 다음날 주가가 상승할지 하락할지 예측하는 문제이다. 60분, 30분, 10분, 5분 단위로 캔들스틱 차트 이미지를 생성하여 CNN 모형을 학습하였다. 마찬가지로 <Figure 7>의 교차 검증 방안을 활용하여 테스트하였으며, 분류 정확도는 <Table 9>와 같다. 60분 단위 캔들스틱 차트 이미지가 가장 높은 0.8260의 정확도를 보였으며, 30분 단위, 10분 단위, 5분 단위 이미지를 활용한 경우의 분류 정확도가 지속적으로 낮아졌다. 과도한 정보는 주가 움직임의 패턴을 파악하는데 방해가 된 것으로 볼 수 있다.

〈Table 8〉 Accuracy of Model applying Gaussian Noise

10minuste -3candles	Accuracy
Original	0.7972
Original (Gaussian Noise)	0.8087
Augmentation (Gaussian Noise)	0.8098

〈Table 9〉 Accuracy of Binary Images(Case 2)

Minutes	Accuracy
60 minutes	0.8260
30 minutes	0.8099
10 minutes	0.7909
5 minutes	0.7358

Ⅴ. 결 론

본 연구는 두 가지 목적을 이루기 위해 진행되었다. 첫 번째 목적(Case 1)은 당일 주가 데이터로 생성한 캔들스틱 이미지로 딥러닝을 적용하여 당일 주가 패턴 예측에 관한 것이고, 두 번째 목적(Case 2)은 전일 주가 데이터로 생성한 캔들스틱 이미지로 딥러닝을 적용하여 다음날 주가 상승여부를 예측하는 것이다.

Case 1에서는 데이터 증강 및 가우시안 노이즈를 적용하여 결과를 비교하였다. 60분, 30분, 10분, 5분 단위로 캔들스틱 차트 이미지를 생성하였으며, 마지막 캔들(-1candle), 마지막 두 개의 캔들(-2candles), 마지막 세 개의 캔들(-3candles)을 제거한 이미지를 각각 생성 하여 분류정확도를 측정하였다. 그 결과 10분 단위의 캔들스틱 차트 이미지를 활용한 경우가 가장 좋은 성과를 보였다. 실제로 -1candle이 높은 정확도를 보여주나, -3candle이 예측에 있어 더 큰 의미가 있다고 판단하여, Case 1의 연구에는 10분 -3candle을 이용하였다. 데이터 증강은 여러 수정값 중 0.00025를 모든 캔들(100%)에 적용하였을 때, 0.7992로 가장 높은 정확도를 보였다. 가우시안 노이즈를 적용하였을 때는 데이터 증강을 한 이미지가 0.8098로 가장 높게 나왔다. Case 2는 60분 단위 이미지가 0.8260로 가장 높게 나왔다.

본 연구를 통해서, 주가 데이터를 이용한 캔들스틱 이미지를 다양한 방법으로 증강하여 적용하는 방안을 제시하였다. 본 연구에서 제안한 방안 이외에 다양한 방안을 적용하여 증강이미지를 생성할 수 있으며, 시계열 데이터를 이미지로 생성함에 있어서 데이터 수가 부족한 문제를 해결하는 좋은 방안이 될 수 있을 것이다.

본 연구의 한계점은 코스닥 데이터만 활용하였고, 2016년 8월 1일부터 2019년 3월 8일까지의 데이터만을 사용하였다는 점이다. 주식시장에는 크고 작은 흐름이 존재하기 때문에, 특정 시점에

서만 좋은 결과가 나올 가능성이 있다. 그리고 VGGNet-16이 우수한 성과를 보이는 모형이지만, 본 연구에서는 하나의 모형만 사용하였다. 추후 다양한 모형을 적용하여 주가 패턴 예측을 위한 캔들스틱 이미지에 적합한 딥러닝 모형을 찾아 검증해 볼 필요가 있다.

참 고 문 헌

[1] Amilon, H., "GARCH estimation and discrete stock prices: an application to low-priced Australian stocks", *Economics Letters*, Vol.81, No.2, pp.215-222, 2003.

[2] Ding, X., Zhang, Y., Liu, T., and Duan, J., "Using Structured Events to Predict Stock Price Movement: An Empirical Investigation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1415-1425, 2014.

[3] Guo, S. J., Hung, C. C., and Hsu, F. C., "Deep Candlestick Predictor: A Framework toward Forecasting the Price Movement from Candlestick Charts", *2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, pp.219-226, 2018.

[4] Hoseinzade, E. and Haratizadeh, S., "CNNpred: CNN-based stock market prediction using a diverse set of variables", *Expert Systems with Applications*, Vol.129, No.-, pp.273-285, 2019.

[5] Hussain, Z., Gimenez, F., Yi, D., Rubin, D., "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks", *Proceedings of 2018 AMIA Annual Symposium*, pp.979-984, 2018.

[6] Jeantheau, T., "A link between complete models with stochastic volatility and ARCH models", *Finance and Stochastics*, Vol. 8, No.1, pp.111-131, 2004.

[7] Ko, D. G., Song, S. H., Kang, K.M., and Han, S. W., "Convolutional Neural Networks for Character-level Classification", *IEIE Transactions on Smart Processing & Computing*, Vol.6, No.1, pp.53-59, 2017.

[8] LeCun, Y., Bottou, L., and Haffner, P., "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, Vol.86, pp.2278-2324, 1998.

[9] Oh, C., and Shen, O. R. L., "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement", Proceedings of ICIS 2011, Shanghai, China.

[10] Schumaker, R. P., and Chen, H., "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System", *ACM Transactions on Information Systems*, Vol.27, No.2, Article No. 12, 2009.

[11] 고광은, 심귀보, "딥러닝을 이용한 객체 인식 및 검출 기술 동향", 제어로봇시스템학회지, 제23권 제3호, pp.17-24, 2017.

[12] 김유신, 김남규, 정승렬, "뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사결정모형", 지능정보연구, 제18권 제2호, pp.143-156, 2012.

[13] 박강희, 신현정, "시계열 네트워크에 기반한 주가예측", 經營 科學(Korean management science review), 제28권 제1호, pp.53-60, 2011.

[14] 손현정, 이석준, "딥러닝을 활용한 실시간 주식 거래에서의 매매 빈도 패턴과 예측 시점에 관한 연구: KOSDAQ 시장을 중심으로", 情報시스템 硏究(Journal of information systems), 제27권 제3호, pp.123-140, 2018.

[15] 신동하, 최광호, 김창복, "RNN과 LSTM을 이용한 주가 예측율 향상을 위한 딥러닝 모델", 한국

정보기술학회논문지, 제15권 제10호, pp.9-16, 2017.

[16] 이강희, 양인실, 조근식, “캔들스틱 차트 분석을 이용한 주식 매매 타이밍 예측을 위한 전문가 시스템”, 지능정보연구, 제3권 제2호, pp.57-70, 1997.

[17] 이모세, 안현철, “효과적인 입력변수 패턴 학습을 위한 시계열 그래프 기반 합성곱 신경망 모형”, 지능정보연구, 제24권 제1호, pp.167-181, 2018.

[18] 이민식, 이홍주, “중립도 기반 선택적 단어 제거를 통한 유용 리뷰 분류 정확도 향상 방안”, 지능정보연구, 제22권 제3호, pp.129-142, 2016.

[19] 이윤선, “시간흐름을 반영하는 캔들스틱과 거래량차트”, 금융공학연구, 제5권 제1호, pp.113-127, 2006.

[20] 이형용, “한국 주가지수 등락 예측을 위한 유전자 알고리즘 기반 인공지능 예측기법 결합모형”, Entrue Journal of Information Technology, 제7권 제2호, pp.33-43, 2008.

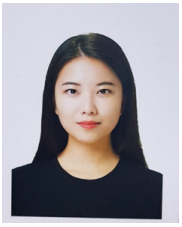
[21] 정지선, 김동성, 김종우, “온라인 언급이 기업 성과에 미치는 영향 분석”, 지능정보연구, 제21권 제4호, pp.37-51. 2015.

사 사

이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A3A2066740).

저 자 소 개

김 영 준(Youngjun Kim)
현재 가톨릭대학교 법학과 재학중이며 빅데이터인문경영융복합전공을 복수전공하고 있다. 주요 관심분야는 데이터 분석, 기계학습, 딥러닝 등이다.



김 여 정(Yeojeong Kim)
현재 가톨릭대학교 컴퓨터정보공학부 재학중이며 관심분야는 기계학습과 딥러닝, HCI 등이다.

이 인 선(Insun Lee)
현재 가톨릭대학교 수학과 재학중이며 컴퓨터정보공학을 복수전공하고 있다. 관심분야는 머신러닝, 음성인식, NLP 등이다.

이 홍 주(Hong Joo Lee)
현재 가톨릭대학교 경영학과에 재직중이며 KAIST 테크노경영대학원에서 석사학위와 박사학위를 취득하였다. 관심분야는 데이터 분석, 추천시스템, 텍스트 마이닝 등이다.