석사학위논문

변이형 오토 인코더와 어텐션 메커니즘을 결합한 차트 기반 주가 예측

Chart-based Stock Price Forecasting Combining Variational Auto Encoders and Attention Mechanisms

국민대학교 일반대학원

데이터사이언스학과 데이터사이언스전공

배 상 현

2019

변이형 오토 인코더와 어텐션 메커니즘을 결합한 차트 기반 주가 예측

Chart-based Stock Price Forecasting Combining Variational Auto Encoders and Attention Mechanisms.

지도교수 최 병 구

이 논문을 데이터사이언스학석사학위 청구논문으로 제출함

2019년 12월 13일

국민대학교 일반대학원

데이터사이언스학과 데이터사이언스전공

배 상 현

2019

배상현의

데이터사이언스학석사학위 청구논문을

인준함

2020년 01월 10일

심사위원장 안 성 만 ㊞

심사위원 조 윤 호 ㊞

심사위원 최 병 구 ㊞

국민대학교 일반대학원

차 례

그림 차례

표 차례

국문 요약

변이형 오토 인코더와 어텐션 메커니즘을 결합한 차트 기반 주가 예측

배 상 현

국민대학교 데이터사이언스학과

주식시장의 전망에 대한 예측은 오래전부터 많은 사람들의 관심을 받아오던 주제였다. 이에 많은 연구들이 진행되었지만 주식시장은 Random walk의 특성을 가지고 있어 미래의 주식시장을 예측하는 것은 쉬운 일이 아니었다. 이를 극복하기 위해 여러가지 분석 기법들이 등장하였지만 여러가지 한계점 때문에 만족할 만한 성과를 내기 힘들었다.

최근 딥 러닝 기법의 발달로 이러한 방법론들을 금융시장에 적용하려고 하는 시도가 점점 더 많아지고 있다. 이에 본 연구에서는 변이형 오토 인코더와 어텐션 모델을 이용하여 익일 주가의 등락을 예측하는 방법을 제안하고자 한다. 이를 위해 S&P 500지수 구성 종목 중 50개의 기업을 랜덤으로 추출하여 각 기업의 시가, 상한가, 하한가, 종가데이터와 거래량으로 캔들 스틱 차트를 그린다. 그 후 시장의 감성점수라고도 볼 수 있는 변동성 지수를 가져와 만들어 놓은 캔들 스틱 차트에 그래프를 덧붙여 그리게 된다. 이 데이터 셋을 변이형 오토 인코더를 이용하여 특징을 추출하고 이 특징을 어텐션 모델에 입력 값으로 주게 되고 다음날 주가의 등락을 예측하게 된다.

모델의 성능은 손실와 AUC로 평가하였으며 제안된 모델이 단순 합성곱 신경망이나 LSTM모델보다 더 좋은 성능을 보였으며 기존 연구들과 비교하였을 때도 우수한 성능을 보였다. 또한 학습된 모델로 시뮬레이션 투자를 진행하였으며, 본 연구의 모델이 벤치마크보다 우수한 수익률을 거두었다.

제 1장 서론

여러가지 경제적, 정치적 상황에 크게 영향을 받는 주식시장은 경제의 대표적인 지표라고 볼 수 있다. 주식시장은 개인들이 쉽게 접근할 수 있는 금융상품 중 하나이고, 기업들은 주식시장으로 자금을 조달할 수 있다. 미래의 주가를 예측할 수 있다면 상당한 이익을 얻을 수 있지만 주식시장을 예측하기 위한 정보는 크거나 적은 상관관계를 가지고 있기 때문에 예측이 쉽지가 않은 것이 사실이다. 그럼에도 불구하고 주식시장의 전망을 예측하여 이익을 얻으려는 노력은 끊임없이 시도되어 왔고 항상 사람들의 관심을 끌어왔다. 이에 여러가지 분석방법들이 등장했는데 대표적으로 기본적 분석과 기술적 분석가 있다.

기본적 분석은 현재 경제 상태, 전체 산업 상황 및 회사의 재무상태 등과 같은 주가에 영향을 미치는 기본 조건 및 결정 요인을 분석하여 미래 주가 추세를 예측한다. 반대로 기술적 분석은 주식 시장의 행동 측면에서 미래의 주가 추세를 분석하는 방법이다(Murphy, 2011). 캔들 스틱 차트는 기술적 분석 방법 중 가장 널리 채택된 방법 중 하나이다. Fama(1965)가 제안한 효율적 시장 가설(efficient market theory, EMH)은 주가에 영향을 미치는 알려진 모든 요소가 현재 주가에 반영되어 있기 때문에 주식에 대한 기본적 분석 및 기술적 분석은 유효하지 않다고 주장하였다.

하지만 점점 더 많은 전문가들이 효율적 시장 가설을 보수적 의견으로 받아들이고 있고 기술적 분석은 여전히 주식 예측에 널리 사용되고 있다. Marshall et al.(2006) 은 과거의 데이터를 기반으로 한 예측이 높은 수익

률을 보일 수 있다는 것을 실증하였고 이는 많은 주식시장에서 통용되는 사실이다.

컴퓨터 공학 기술의 발달로 점점 더 많은 기업들이 빅 데이터와 인공지능 기술을 금융투자 분야에 적용하고 있다. 골드만 삭스는 2000년, 뉴욕에 600명 이상의 트레이더를 고용하고 있었지만 2017년에는 2명의 트레이더만이 남았으며 컴퓨터가 나머지를 대체하게 되었다(Maney, 2017). 최근 많은 연구들도 머신 러닝과 딥 러닝 기법을 이용하여 주가 예측을 시도하고 있다. Guo et al.(2018) 은 캔들 스틱 차트와 Lecun et al.(1989)가 제안한 합성곱 신경망(convolusional neural network, CNN)을 사용하여 주가의 등락을 예측하였고, Ghoshal & Roberts(2018)는 캔들 스틱 차트에 다양한 딥 러닝 모델을 적용하여 각각의 성능을 비교하였다. 서포트 벡터 머신(support vector machine, SVM)과 캔들 스틱 차트를 이용하여 주가의 등락 예측(Zhipeng & Chao, 2019)을 시도하는 연구도 진행되었다.

하지만 많은 캔들 스틱 차트를 기반으로 한 주가 등락 예측 연구들이 시장 참여자들의 감정 상태가 주가에 미치는 영향이 크다는 연구 결과가 있음에도 불구하고(Ding et al., 2015.; Huynh et al., 2017; Nguyen et al., 2015)기술적 시장 지표에만 집중하는 경향을 보였다. 따라서 본 연구에서는 기존 연구들의 이러한 한계점을 보완하기 위해 시장 참여자들의 시장 전망에 대한 기대라 할 수 있는 변동성(volatility)지수를 추가로 사용하여 예측 성과를 높이려 시도 하였다. 또한 기존 연구에서 사용되던 딥 러닝 모델의 구조적 결함을 해결하기 위해 자연어 처리에 쓰이던 모델을 주가 등락 예측에 적용하는 방법을 제안한다.

본 논문은 다음과 같이 구성되었다. 2장에서는 기존 주가 예측, 본 연

구에서 사용되는 캔들 스틱 차트와 변동성 지수, 그리고 딥 러닝 기반 주가 예측에 관해 살펴보며 선행 연구들의 진행 방식과 한계점에 대해 서술한다. 3장에서는 본 연구의 연구 모형과 관련 모델들에 관해 서술하였다. 4장에서는 사용된 데이터에 관한 내용과 평가 방법, 5장에서는 결과와 추가적 연구에 관해 기술하였다. 마지막 6장에서는 본 연구의 의의와 한계점에 대해 알아본다.

제 2장 관련 연구

2.1 주가 예측

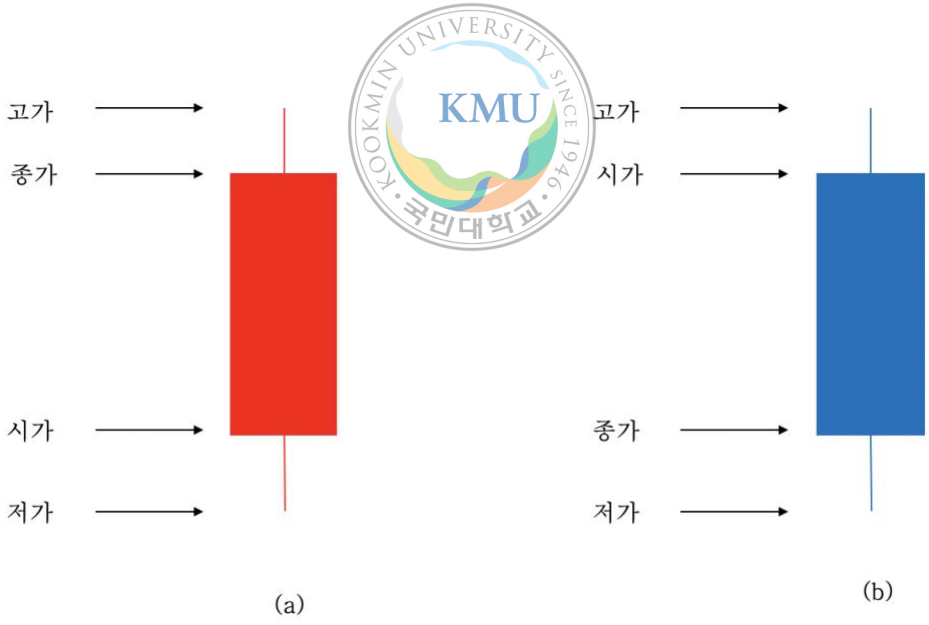
기존 주가 예측은 기본적 분석이나 기술적 분석으로 이루어졌다. 기본적 분석은 주가는 기업의 내재가치에 달려 있다고 보는 입장이다. 기업은 이익을 내기 위해 존재하는 집단이고 주식은 해당 기업의 가치를 반영하므로 주식은 기업의 이익의 성과에 맞추어 가치가 변동하고 그렇기에 기본적 분석은 재무제표를 기반으로 분석을 하게 된다. 회사의 주당순이익(earning per share, EPS), 주가 수익 비율(price earning ratio, PER)등 다양한 기본적 분석 지표 등이 등장하였고 최근에도 주가 예측에 꾸준히 사용되고 있다. 하지만 재무제표는 1년에 4번 밖에 발표 되지 않아 예측 적시성의 문제점을 가지고 있고 기업의 분식회계 등 재무제표의 신뢰성의 문제로 기본적 분석은 여러가지 한계점을 가지고 있다.

기술적 분석은 비교적 전통적인 주가 예측 방식이다. 주가의 추세를 발견하여 주가를 예측하는 방식인데 주로 차트를 가지고 분석을 하게 된다. 기술적 분석은 두 가지 가정을 전제로 하고 있다. 주가는 반복되는 움직임을 가지고 있고 추세를 파악한다면 기업의 가치와는 무관하게 주가의 변동을 예측할 수 있다는 것이 그것이다. 대표적인 기술적 분석방법으로는 이동평균선의 움직임이나 캔들 스틱 차트의 특정 패턴을 보고 주식을 변동을 예측하는 방법이 있는데, 이러한 기술적 분석은 표본이 적을수록 예측의 불확실성이 늘어나고 작전 세력, 기사 등에 의해 예상치 못한 변동이 발생할 수 있는 단점이 존재한다. 이러한 기존 주가 예측의 문제점

들을 해결하기 위해 본 연구에서는 딥 러닝 모델을 사용한 주가 예측 방법을 제안한다.

2.2 캔들 스틱 차트

캔들 스틱 차트는 가격 분석에 유용한 몇 가지 상황을 보여 줄 수 있다. 캔들 스틱 차트에서 특정 그래프 조합은 시장의 변동성 감소를 나타낼 수 있고 다른 조합은 변동성 증가를 시사한다. <그림 1>에 표시된 캔들 스틱 차트는 각 시간 단위의 고가, 저가, 시가 및 종가를 나타낸다.



<그림 1> 캔들 스틱 차트

촛대의 두꺼운 부분을 몸통이라고 하며 시가(open price)와 종가(close price) 사이의 간격을 표시한다. 파란색의 캔들 스틱 차트 (b)는 종가가

시가보다 낮았음을 의미하고 빨간색의 캔들 스틱 차트 (a)는 종가가 시가보다 높아졌음을 의미한다. 몸통에서 위아래로 움직이는 선을 그림자라고 하며 하루 중 고가와 저가를 나타내게 된다.

캔들 스틱 차트는 주식, 상품 및 옵션 거래의 의사 결정에 시각적인 도움을 줄 수 있다. 만약 막대가 빨간색이고 몸통의 길이가 길면 시장의 참여자들이 낙관적인 판단을 했다는 것을 의미한다. 파란 몸통의 경우에는 그 반대일 수 있다. 숙련된 투자자는 캔들 스틱 차트를 직관적으로 이해하고 투자 결정을 내릴 수 있다. 예를 들어, "슈팅 스타" 패턴은 일반적으로 상승 추세가 반전됨을 나타낸다. 캔들 스틱 차트는 수치 데이터인 재무 데이터를 투자자들에게 의미 있는 패턴으로 전달 가능하다. 다시 말하자면 이 차트는 실제 가치 계산보다는 패턴 인식을 선호하는 투자자에게 강력한 예측 방법을 제공한다.

2.3 변동성 지수

1993년 미국 시카고 옵션 거래소(Chicago board options exchange, CBOE)에서 듀크 대학의 Robert Whaley교수가 제안한 지수이다. CBOE에서 거래되는 S&P500 지수가 향후 30일간 얼마나 움직일 지에 대한 시장의 예상치를 나타내게 된다. 보통 주식시장이 급락하거나 불안할 수록 수치가 올라 '공포지수' 라고도 불린다. 변동성 지수가 20미만일 땐 투자자들의 시장에 대한 기대는 낙관적, 20이상일때는 시장에 두려움이 생기기 시작했다고 볼 수 있다. 2008년 ~ 2009년 서브 프라임 사태가 발생하였을 때 변동성 지수는 50보다 훨씬 높은 수준으로 급등하였고,

1987년 10월 '검은 월요일'엔 172.29로 이날 하루 만에 S&P500지수는 20%가량 폭락하였다.

이러한 변동성 지수의 특성으로 많은 연구들이 변동성 지수와 시장의 감성의 상관관계를 알아보는 연구를 진행하였다. 대표적으로 Liu et al.(2017) 연구팀은 시장 감성을 이용하여 변동성 지수를 예측하였다. 약 1년간(2015년 9월 ~ 2016년 9월) 주식 포럼에서 포스팅된 게시물들을 바탕으로 감성 점수를 구하고 그 결과로 변동성 지수를 예측하였고 약 73%의 정확도로 변동성 지수를 예측할 수 있었다. 이로 미루어 볼 때, 변동성 지수는 시장의 감정 상태를 잘 반영하고 있다고 볼 수 있고 본 연구에서도 시장의 감정 상태를 반영하기 위해 변동성 지수를 사용하려 한다.

2.4 딥 러닝 기반 주가 예측

최근 컴퓨터 공학이 발전하며 그에 맞춰 딥 러닝 기법들에 대한 연구들이 활발해졌다. 이에 많은 사람들의 관심을 받던 주식 시장 예측 방법론 개발에 딥 러닝 기법들을 적용하고자 하는 시도가 늘어나게 되었다.

딥 러닝을 적용된 주가 예측은 크게 3가지 분류로 나눌 수 있다. 주식시장의 기본 주가데이터를 활용한 연구, 시장의 감성을 분석하여 주식시장을 예측하는 연구, 기술적 분석를 기반으로 한 연구 등이 그것이다. 초기에는 컴퓨팅 파워 등 여러가지 한계 때문에 Niaki & Hoseinzade(2013)의 작업과 같이 단순 주식시장 변수들을 입력 값으로 하여 인공 신경망

(artificial neural network, ANN)으로 예측하는 것이 주를 이루었다. 이후 CNN을 사용하거나 순환 신경망(recurrent neural network, RNN)을 사용하여 주가 예측을 시도하는 연구들이 나오기 시작하였고, 과거 딥 러닝 기반 주가 예측 연구에 사용되었던 모델에 전통적인 피처 엔지니어링 방식을 섞는 방법 또한 제안되었다(Persio & Honchar, 2016; Song, 2018; Zhong & Enke, 2017).

자연어 처리 분석 기법이 발달하면서 사람들이 쓴 글로 감정 상태를 예측 할 수 있게 되었다. 이에 사람들의 감정 상태를 분석하여 주식 시장 예측을 시도하는 연구들이 늘어났는데, Ding et al.(2015)은 경제 신문의 내용에 있는 사건들과 해당 기업들을 하나의 튜플(tuple)로 묶어 이벤트 임베딩(event embedding)으로 벡터화하고 이 전처리 된 기사들을 딥 러닝 모델들을 적용하여 다음날 주가의 등락을 예측하였다. SNS에 올라오는 글들의 감성 점수를 이용하여 주가 예측을 시도한 연구도 진행되었는데, 대표적으로 Nguyen et al.(2015)은 트위터의 게시물들을 잠재 디리클레 할당(latent dirichlet allocation, LDA)으로 분석하여 나온 결과에 감성 점수를 매겨 글 전체의 분위기를 보던 기존 연구들과는 다르게 특정 키워드에 집중함으로써 기존 연구들보다 더 좋은 결과를 얻었다고 주장했다. Huynh et al.(2017)은 기존 시장 참여자들의 감정 상태와 주식 시장의 관계를 보는 연구들이 고려하지 않았던 시계열 측면을 추가로 고려하여 좋은 결과를 얻었다.

기술적 분석에 딥 러닝을 접목시킨 연구는 비교적 최근에 각광받기 시작하였다. Bao et al.(2017)은 웨이블릿 변환(wavelet transform), 오토 인코더(auto encoder, AE), LSTM(long-short term memory)을 사용하여 캔들 스틱 차트와 여러 기술적 지표를 입력 값으로 주가를 예측하였다. 캔

들 스틱 차트와 어떤 딥 러닝 모델을 같이 사용하여야 좋은 성능 나오는지 확인한 연구가 진행되었고(H. Liu & Song, 2018), 캔들 스틱 차트와 캔들 스틱 차트 구성 데이터를 입력 값으로 CNN과 LSTM을 사용하여 주가를 예측한 연구도 존재하였다(Kim & Kim, 2019). <표 1>은 딥 러닝 기반 주식 시장 예측 연구를 요약한 표이다.

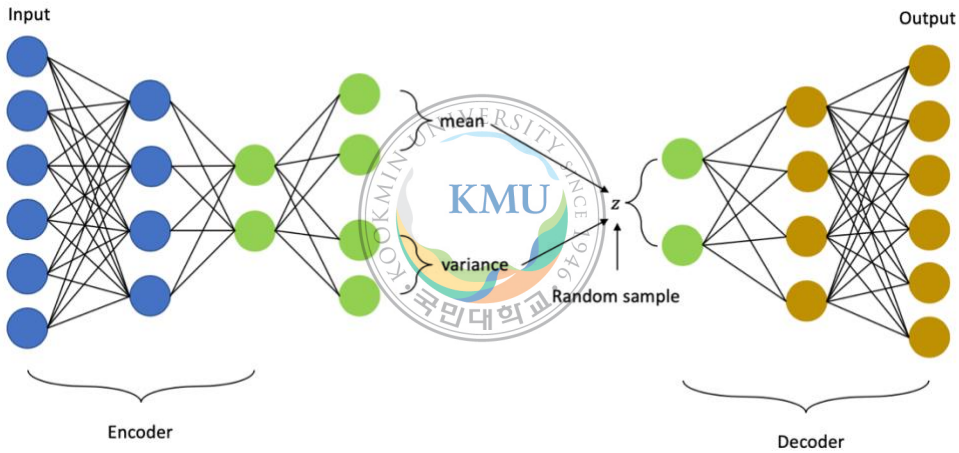
<표 1> 기존 연구

		저자	데이터셋		사용 모델	결과 양식	분석 기간
정형	기본	Niaki & Hoseinzade, 2013	S&P 500	Technical indexes	ANN	이진 분류	1994.3 ~ 2008.6
		Persio & Honchar, 2016	S&P 500	Technical indexes	ANN, WCNN, RNN	이진 분류	2006 ~ 2016
		Zhong & Enke, 2017	S&P 500	Technical indexes	ANN, PCA	이진 분류	2003.6 ~ 2013.5
		Song, 2018	S&P 500	Technical indexes	RNN, SVM, XGBoost	이진 분류	2010. 1 ~ 2017. 12
비정형	감성	Ding et al., 2015	S&P 500	Finance news	EB, CNN	이진 분류	2006.10 ~ 2013.11
		Nguyen et al,. 2015	Twitter, price(OHLC)		LDA	이진 분류	2012.7 ~ 2013.7
		Huynh et al., 2017	S&P 500	Reuters News dataset	GRU	이진 분류	2006.10 ~ 2013.12
	이미지	Bao et al., 2017	S&P 500 Hang Seng index CSI 300 Nikkei 225 Nifty 50	Candlestick chart, Technical indexes	AE+ LSTM	가격 예측	2008.7 ~ 2016.9
		Guo et al., 2018	TAIFEX	Candlestick chart	AE + CNN	이진 분류	1998 ~ 2016
		Liu & Song, 2018	10 China mkt index stock 10 White Horse stock	Candlestick chart, MA	SVM, CNN, DNN	모델 비교	2006.1 ~ 2017.8
		Kim & Kim , 2019	S&P 500	Candlestick chart	CNN, LSTM	가격 예측	2016.10 ~ 2017.10
	감성+ 이미지	본 연구	S&P 500	Candlestick chart, VIX chart	VAE, Attention model	이진 분류	1993. 7 ~ 2019. 7

WCNN : wavelet CNN
PCA : principal component analysis
EB : event embedding

2.5 변이형 오토 인코더(variational autoencoder)

Kingma & Welling(2013)이 제안한 변이형 오토 인코더는 비지도 기반 생성 모델로 AE와 베이지안 추론을 바탕으로 데이터의 분포를 추정하는 방법이다. AE의 잠재 변수 z는 인코더에 의해서 압축된 입력 데이터의 특징이라고 할 수 있는 반면에, 변이형 오토 인코더의 잠재 변수 z는 특정한 분포를 따르는 입력 데이터의 특징이라고 할 수 있다. 변이형 오토 인코더의 대략적인 모형은 <그림 2>와 같다.



<그림 2> Variational Autoencoder

AE의 손실 함수(loss function)는 입력 데이터 X와 잠재 변수 z가 디코더를 거쳐 나온 출력 데이터 X'간의 차이를 보게 되는데 다음의 식 (1)과 같이 나타낼 수 있다. AE는 이 손실 함수를 최소화 하도록 학습하게 되며 잠재 변수 z는 입력 데이터 X의 특징을 보존하고 있으므로 입력 값의 유사도를 구하거나 분류 문제에 사용되게 된다.

$$binary\ cross\ entropy(X, X') = -\frac{1}{N}\sum_i (X_i \log(X_i') + (1 - X_i)\log(1 - X_i')) \quad (1)$$

변이형 오토 인코더는 잠재 변수 z가 확률 분포를 따르게 하기 위해 쿨백-라이블러 발산(Kullback-Leibler divergence, KLD)을 사용한다. KLD란 두 확률 분포의 차이를 계산하는 데에 사용하는 함수로 식 (2)로 나타낼 수 있는데, 어떤 이상적인 분포에 대해 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면 발생할 수 있는 정보 엔트로피 차이를 계산한다. 변이형 오토인코더는 손실 함수에 잠재 변수 z가 정규분포를 따르도록 다음과 같은 KLD값을 추가한다. 따라서 변이형 오토 인코더의 손실 함수는 아래의 식 (3)과 같은 형태를 띄게 된다.

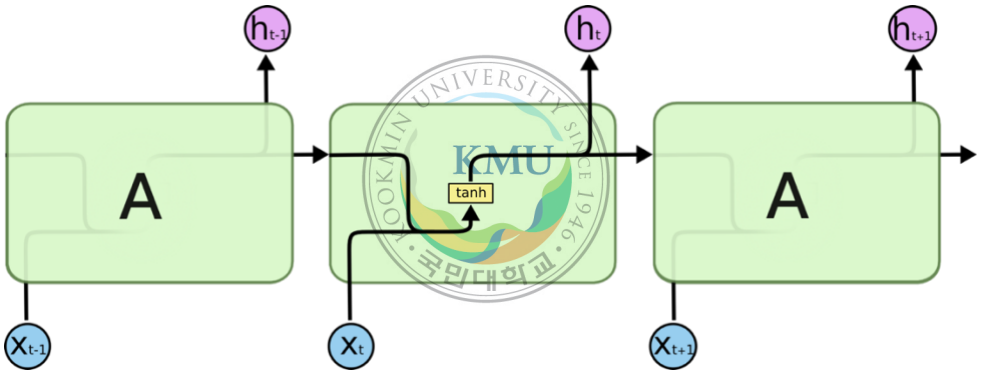
$$KL[N(\mu,\sigma^2)\,\|\,N(0,1)] = -\frac{1}{2}(1+\log\sigma-\mu^2-\sigma^2) \quad (2)$$

$$Total\ Loss = binary\ cross\ entropy(X,X') + KL[N(\mu,\sigma^2)\|N(0,1)] \quad (3)$$

기존 캔들 스틱 차트를 이용한 주가 예측 연구들은 특성 추출(feature extraction)과 차원 축소(dimension reduction)을 위해 AE를 사용하였다. 하지만 앞서 언급했다시피, AE의 목적은 재구축 비용을 최소화 하는 것이기 때문에 재구축에 필요하지 않은 특징까지 학습하는 과적합(overfitting)의 위험이 존재하였다. 변이형 오토 인코더는 새로운 데이터 생성을 위한 목적으로 잠재 변수 z가 확률 분포를 따르게 하였는데, 이 방법이 표준화(standardization)의 효과를 내 AE와 비교하여 더 우수하게 입력 데이터 X의 특성을 추출 할 수 있게 하였다. 이에 본 연구에서는 기존 AE의 방식의 한계점을 극복하기 위해 변이형 오토 인코더를 사용하려고 한다.

2.6 양방향 LSTM

Hochreiter & Schmidhuber(1997)은 기존 RNN의 문제점인 기울기 소실(vanishing gradient)을 극복하기 위해 LSTM을 고안하였다. RNN은 시계열 데이터를 처리하기 적절한 인공 신경망 모델인데 데이터가 순차적으로 입력됨으로써 모델이 순서를 고려할 수 있게 되기 때문이다. t시점의 출력 값의 예측을 위해 t시점의 입력 값 뿐만 아니라 t-1시점의 입력 값을 바탕으로 생성된 t-1시점의 은닉 계층(hidden state) 값을 고려하게 되는데 대략적인 RNN의 모형은 <그림 3>과 같다.



<그림 3> Recurrent Neural Network(Olah, 2015)

모형의 예측 값이 실제 데이터와 얼마나 정확한지 평가하기 위해 cross entropy 함수를 비용 함수(cost function)로 설정한다. 이 비용 함수를 최소화시키는 가중치(weight)와 편향(bias)을 최적화 알고리즘을 통해 찾는데, 최적화 알고리즘은 경사하강법을 기반으로 하는 확률적 경사하강 알고리즘을 사용한다. 단순 경사하강법같은 경우 경사를 이용하여 비용 함수의 최소값을 찾는 방법이지만, 모든 데이터를 사용하기 때문에 연산량이 많아져 속도가 느리지만 확률적 경사하강법은 일부 데이터만을 사용하여 최소값을 구하기 때문에 최소값을 빠르게 찾게 된다. 여기에 추

가하여 출력층에서부터 경사를 계산하여 가중치와 편향을 학습하는 역전파(back-propagation) 알고리즘을 사용한다.

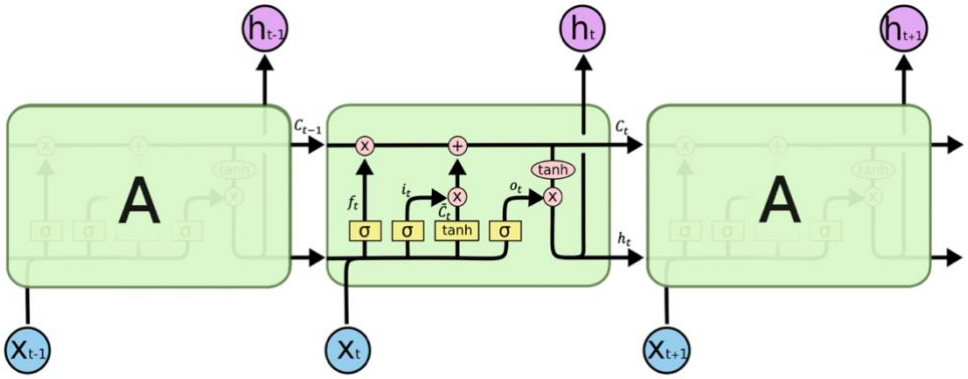
역전파 알고리즘을 사용하여 모델을 학습시킬 때, 모델이 고려하는 시계열 기간이 너무 길면 비용 함수의 편미분 계수가 0에 가까워 지거나 매우 커지는 문제가 발생(Bengio et al., 1993)하는데 이 경우 모델의 학습이 제대로 이루어지지 않게 된다. 앞서 언급했듯이 RNN은 이러한 문제점에서 자유로울 수 없었고 이를 해결하기 위해 LSTM모델이 제안되었다. LSTM의 경우 RNN과 비슷한 구조를 가지지만 은닉 계층이 메모리 셀이라는 구조를 가진다(Hochreiter & Schmidhuber, 1997). 이는 오래된 시점의 출력 값과 현 시점의 입력 값을 어느정도 반영할지 세 가지 게이트(gate)를 이용해 학습하는 구조라고 할 수 있다. 관련 식을 나타내면 다음과 같다.

$$\tilde{C}_t = \tanh(x_t U^c + h_{t-1} W^c) \quad (4)$$

$$C_t = \tilde{C}_t \circ i_t + C_{t-1} \circ f_t \quad (5)$$

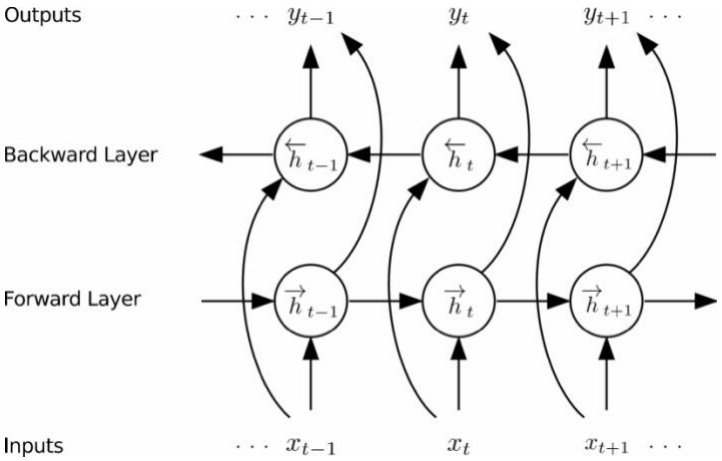
t시점의 입력 값(x_t)과 $t-1$시점의 은닉 노드의 출력 값(h_{t-1})의 선형결합을 활성화 함수(σ)를 통해 은닉 노드 후보 값($\tilde{C}_t$)을 계산한다(식 (4)). 이후 입력 게이트(input gate)의 출력 값(i_t)과 은닉 노드 후보 값($\tilde{C}_t$)을 결합하여 얼마나 다음 노드에 반영할 지 결정한다. 망각 게이트 출력 값 f_t는 이전 시점 은닉 값 C_{t-1}와 결합하여 얼만큼 이전 은닉 값을 반영할 지 결정 후 $\tilde{C}_t \cdot i_t$와 결합하여 t시점의 은닉 값(C_t)을 생성한다(식 (5)). t시점의 은닉 값은 내부 활성화 함수(tanh)를 통과하여 출력 게이트 출력 값(o_t)과 결합하게 되는데 이 값이 다음 t+1시점의 은닉 값에 얼마만큼 반영되는지 결정하는 값이 된다. 식에서 사용한 U, W 는 가중치를 나타

내는 행렬이며 위첨자 c, f, i, o는 은닉 노드 후보 값, 망각 게이트, 입력 게이트, 출력 게이트에서의 계산을 의미한다. ∘는 벡터간 요소의 곱이며 σ는 활성화 함수이다. LSTM의 대략적인 모형은 <그림 4>와 같다.



<그림 4> LSTM(Olah, 2015)

하지만 기존 RNN베이스 모델들은 다음 시점의 데이터를 가지고 있음에도 이전까지의 데이터만을 가지고 예측을 해야하는 문제점이 있었다. 이를 해결하기 위해 양방향 RNN이 제안되었다(Graves & Schmidhuber, 2005; Schuster & Paliwal, 1997). 기존 모델들이 과거에서 현재로의 흐름만을 고려한다면 양방향 RNN 모델은 미래에서 현재로의 흐름을 같이 보게 된다. 양방향 RNN 모델의 모형은 <그림 5>와 같다.



<그림 5> 양방향 RNN(Graves et al., 2013)

입력 값(x)은 정방향(forward layer)과 역방향(backward layer)의 은닉 노드 모두에 입력되고 양방향 히든 레이어의 출력 값이 서로 결합하여 최종 출력 값(y)이 나오게 된다. 이 구조를 식으로 나타내면 식 (6)과 같다.

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \quad (6)$$

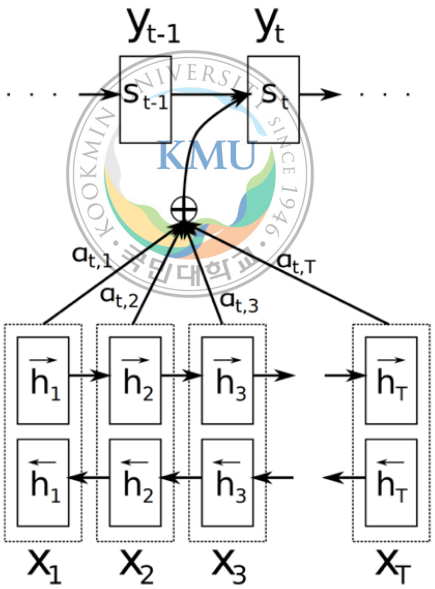
$$\vec{h}_t = H\left(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}\right) \quad (7)$$

$$\overleftarrow{h}_t = H\left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right) \quad (8)$$

식 (7)의 $\vec{h}$는 정방향의 은닉 노드를 나타내고 식 (8)의 $\overleftarrow{h}$는 역방향의 은닉 노드를 의미한다. H, W, b, t는 각각 활성화 함수, 가중치, 편향, 현재 시점의 의미를 갖는다. Graves & Schmidhuber(2005)가 고안한 양방향 LSTM의 경우 은닉 노드가 LSTM의 셀로 바뀌는 점을 제외하면 형태가 같다. 본 연구에서는 양방향 LSTM모형을 사용하였다.

2.7 어텐션 메커니즘(attention mechanism)

RNN은 입력 시퀀스가 길어질수록 장기 의존성(long term dependencies)문제가 발생하였다. LSTM은 이를 잘 처리한다고 알려졌지만 여전히 문제가 발생하는 실정이다. 어텐션 메커니즘(Bahdanau et al., 2016) 은 출력을 예측하는 매 시점마다 기존에 들어온 입력 전체를 다시 한번 참고하는데 전체를 동일한 비율로 참고하기보단 모델로 하여금 예측에 중요한 부분을 참고하게 한다. 양방향 LSTM에 어텐션 메커니즘을 적용했을 때의 대략적인 모형은 <그림 6>과 같다.



<그림 6> Attention mechanism(Bahdanau et al., 2016)

t시점의 어텐션 층의 은닉 노드값(s_t)을 구하기 위해선 t-1시점의 은닉값(s_{t-1})과 이전의 출력 값(y_{t-1}), 그리고 t시점의 출력을 예측하기 위한 어텐션 값(a_t)이 필요하다. 식으로 나타내면 식 (9)와 같다.

$$s_t = f(s_{t-1},\ y_{t-1},\ a_t) \quad (9)$$

어텐션 값을 구하기 위해 t시점의 출력 값(y_t)을 구하기 위해 입력 값의 은닉 노드(h) 각각이 어텐션 레이어의 t-1시점의 은닉 값과 얼마나 유사 한지 판단하는 어텐션 스코어(attention score)를 구하게 된다. 어텐션 스코어 값을 구하기 위해 어텐션 레이어의 t-1시점의 은닉 값을 전치(transpose)하여 입력 값의 은닉 노드 각각과 내적(dot product)을 수행한다(식 (10)). 이렇게 얻은 어텐션 스코어값(식 (11))에 소프트맥스(softmax) 함수를 적용하여 어텐션 분포(attetion distribution)을 구하게 되는데 이를 식으로 표현하면 식 (12)와 같이 나타낼 수 있다.

$$score(s_{t-1}^T, h_i) = s_{t-1}^T h_i \quad (10)$$

$$e^t = [s_{t-1}^T h_i,\ \cdots, s_{t-1}^T h_N] \quad (11)$$

$$\alpha^t = softmax(e^t) \quad (12)$$

이렇게 구한 어텐션 분포(α^t)를 입력 값의 은닉 노드와 곱하여 모두 더한 값이 어텐션 값이 된다. 이렇게 구한 어텐션 값을 어텐션 레이어의 t-1시점의 은닉 값과 결합(concatenate)하게 되는데 식 (13)과 식 (14)는 이를 식으로 표현한 것이다.

$$a_t = \sum_{i=1}^{N} \alpha^t h_i \quad (13)$$

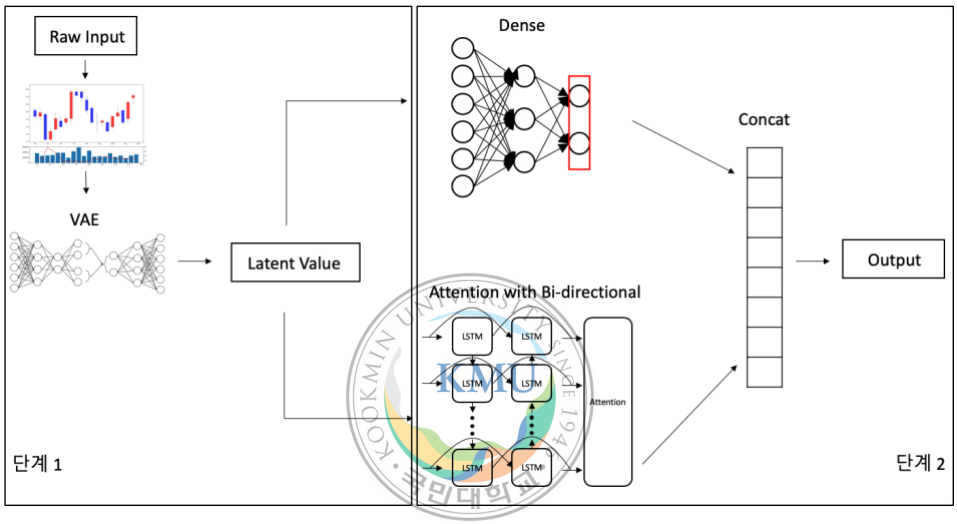
$$s_t = f(v_t,\ y_{t-1}) \quad (14)$$

어텐션 메커니즘은 많은 연구에서 좋은 성능을 보였으며 현재 다양한 분야에서 활용되고 있다(Rush et al., 2015; Yao et al. 2015). 본 연구에서

는 캔들 스틱 차트의 시계열적 특성을 더 잘 반영하기 위해 어텐션 메커니즘을 사용하였다.

제 3장 연구 방법론

3.1 연구 모형



<그림 7> 연구 모형

<그림 7>는 본 연구의 연구 모델이다. 만들어진 캔들 스틱 차트를 변이형 오토 인코더에 입력 값으로 주고 재구축 손실이 적어지도록 학습을 시키게 된다. 이렇게 학습된 변이형 오토 인코더를 이용하여 캔들 스틱 차트의 잠재(latent) 값을 구하게 된다. 이 잠재 값들은 양방향 LSTM-어텐션 모델과 dense 레이어로 각각 넘어가게 되고, 각 신경망의 출력물들은 이후 결합(concatenate)되어 마지막 dense 레이어의 입력 값이 된다. 소프트맥스(softmax) 함수를 통과한 마지막 dense 레이어의 익일 종가(현재가)가 금일의 종가보다 오를 것인지 이진 예측을 하게 된다. 본 연구에서 제안된 모델을 VAT(Variational autoencoder with Attention

Technic)으로 명명하였다.

3.2 단계 1 : 데이터 셋 구성

단계 1에서는 대상이 되는 차트의 특징을 변이형 오토 인코더로 추출하고 추출된 특징을 바탕으로 단계 2에 입력이 되는 데이터 셋을 만들게 된다. 본 연구는 주가 예측을 그 목적으로 하므로 주가 데이터를 이용하여 캔들 스틱 차트를 만들게 된다. <그림 9>는 본 연구에서 사용될 차트의 예시이다. 하나의 이미지에 많은 수의 차트를 넣게 되면 더 좋은 성능을 보인다는 연구 결과(Kim & Kim, 2019)에 따라 본 연구에서도 캔들 스틱 차트와 거래량 차트, 그리고 변동성 지수 차트를 하나의 이미지로 구성하였다.

<그림 8> 차트 예시

이렇게 만들어진 차트는 변이형 오토 인코더의 입력 데이터가 되고 변이형 오토 인코더에서 단계 2에서 사용될 잠재 값들을 추출하게 된다. 변이형 오토 인코더에 입력되는 이미지의 크기는 넓이와 높이 모두 128로 설정하였고 변이형 오토 인코더의 네트워크 구조는 다음과 같이 구성하였다. 인코더 부분에 다섯 층의 컨볼루전(convolution) 네트워크를 쌓고 각 컨볼루전 레이어의 필터의 수는 입력 레이어에 가까운 순서대로 128, 256, 256, 512, 1024개이다. 필터의 크기는 3, stride는 모두 2로 설정하였다. 컨볼루전 레이어를 거쳐 나온 결과에 평탄화 레이어 층을 적용한 후 dense 레이어를 연결하였고 dense 레이어의 노드 수는 30으로 설정하고 변이형 오토 인코더의 잠재 차원(latent dimension)도 30으로 설정하였다. 디코더 층은 인코더 층의 역순으로 전치 컨볼루전(transpose convolution)층을 쌓았고 각 필터의 개수는 1024, 512, 256, 256, 128개가 되게 된다. 사용된 모든 층의 활성화 함수는 Relu를 사용하였고 최적화 알고리즘은 Adam 방식을 사용하였다.

이 과정을 거쳐 나온 잠재 값들은 dense에 입력되는 데이터 셋과 양방향 LSTM에 입력되는 데이터 셋 두 가지로 구성되게 된다. <표 2>는 구성된 양방향 LSTM의 데이터 셋의 예시이다.

<표 2> 구성된 양방향 LSTM 데이터 셋

	0	1	…	329
0	-106.03342	2.5770082	⋯	1.3704153
1	11.75793	-0.1898792	⋯	0.9307256
2	-0.7532625	0.81035775	⋯	0.9168761
⋮	⋮	⋮	⋮	⋮

3.3 단계 2 : 예측 단계

단계 2에서는 단계 1에서 생성된 데이터 셋을 바탕으로 모델을 학습시키고 예측을 하게 된다. 단계 2의 위쪽 dense 레이어 부분은 시계열적 측면을 고려하지 않고 20일치 데이터로 구성된 캔들 스틱 차트 이미지 한 장에 집중하여 익일 주가의 등락을 예측하기 때문에 해당하는 차트의 이미지 특징 잠재 값을 입력으로 받게 된다. 따라서 입력 레이어는 30개의 노드, 히든 레이어의 노드의 개수는 10개와 2개로 구성을 하였다. 단계 2의 아래쪽 어텐션 메커니즘 부분은 하나의 양방향 LSTM 레이어와 하나의 어텐션 레이어로 이루어져 있고, 각 레이어의 셀은 lookback기간과 동일한 11개의 셀을 가지고 있다. 이후 어텐션 메커니즘 부분은 2개의 레이어를 가진 dense 레이어로 연결되고 dense 레이어의 각 노드의 개수는 30개, 10개로 구성되어 있다. 이후 도식의 dense 레이어 부분과 어텐션 메커니즘 부분을 결합하여 익일 주가의 등락을 예측한다.

본 연구에서는 익일의 주가를 상승 혹은 하락으로 이진 예측을 그 목적으로 한다. P_t는 금일 종가, P_{t+1}는 익일 종가로 익일의 종가가 금일의 종가보다 높거나 같을 시 1, 낮을 시 0으로 원 핫 인코딩 하였다. 식으로 나타내면 식 (15)와 같다.

$$y = \begin{cases} 1, & P_t \leq P_{t+1} \\ 0, & P_t > P_{t+1} \end{cases} \quad (15)$$

제 4장 실험 방법

4.1 사용 데이터

본 연구에서는 S&P 500 구성 종목 중 임의로 50개의 기업을 선정해 1993. 07. 01 ~ 2019. 07. 31까지 약 25년간의 OHLC(시가, 최고가, 최저가, 종가)와 거래량, 변동성 지수를 기반으로 하여 그린 캔들 스틱 차트를 데이터로 사용한다. 해당 기간 동안의 수집된 기업들의 주가 등락은 48:52의 비율을 가지고 있다. 주식데이터의 수집은 Investing.com에서 진행하였으며 기업당 약 6400일의 데이터, 전체 약 32만건의 데이터를 수집하였다. 수집된 데이터는 <표 3>와 같다.

<표 3> 수집된 데이터

날짜	현재가	오픈	고가	저가	거래량	변동%
2019 년 07 월 22 일	207.22	203.65	207.23	203.61	22.28M	2.29%
2019 년 07 월 23 일	208.84	208.46	208.91	207.29	18.36M	0.78%
2019 년 07 월 24 일	208.67	207.67	209.15	207.17	14.99M	-0.08%
2019 년 07 월 25 일	207.02	207.02	209.24	206.73	13.91M	-0.79%

각 캔들 스틱 차트는 Bao et al.(2017) 의 연구결과를 바탕으로 20일간의 데이터로 구성되었으며, LSTM의 timestep은 관련 분야에서의 선행연구들을(Chen et al., 2016; Persio & Honchar, 2016) 바탕으로 하여 30일로 구성하였다. Sliding window는 1일로 설정하였다. 파이썬 3.6버전과 'mpl_finance'라이브러리를 사용하여 캔들 스틱 차트를 그렸으며, 관련

연구(Kim & Kim, 2019)를 바탕으로 캔들 스틱 차트 아래 거래량 막대그래프와 꺾은선 그래프로 구성된 변동성 지수를 그렸다. 약 31만개 가량의 그래프를 생성하였다.

4.2 평가 방법

연구 모델의 성능을 평가하기 위해 AUC(area under the curve)값을 사용하였다. AUC는 ROC(receiver operating characteristic)곡선 아래의 면적을 말하는데 ROC곡선이란 가로축을 특이도(specificity), 세로축을 민감도(sensitive)로 하여 시각화한 그래프이다. specificity와 sensitive의 정의는 식 (16), 식 (17)과 같다.

$$specificity = \frac{TN}{TN+FP} \quad (16)$$

$$sensitive = \frac{TP}{P} \quad (17)$$

TN: *실제 값이 부정인데 부정으로 예측한 수*

TP : *실제 값이 긍정인데 긍정으로 예측한 수*

FP: *실제 값이 부정인데 긍정으로 예측한 수*

AUC의 장점 중 하나는 분류 문제에 있어서 클래스 불균형에 대한 견고성이다. Borovkova & Tsiamas(2019)가 언급하였듯이 익일 종가가 오름을 표시하는 비중은 전체 데이터의 약 46% 정도이다. 모델이 한 가지 값만 예측할 가능성이 있기 때문에, 단순 정확도(accuracy)로는 모델이 54%의 정확도를 보였다고 해서 실제 예측 상황에서도 그 정도의 성능을 보여줄 것이라 생각하기 힘들다. 하지만 AUC같은 경우 예측된 긍정

(positive) 값 중 실제 긍정 값, 예측된 부정(negative)값 중 실제 부정 값을 고려하기 때문에 이러한 문제에서 자유로울 수 있다. F1 score도 AUC와 비슷한 장점이 있지만 F1 score는 계산 과정에서 TN이 고려되지 않고, 특정 임계 값(threshold)에 영향을 받기 때문에 본 연구에서는 모델 간의 비교에 AUC를 사용하고자 한다.

제 5장 연구 결과

5.1 실험 종류

본 연구에서 제안하는 모델의 성능을 정확히 알기 위해 총 6가지 실험을 진행하였다. 캔들 스틱 차트기반 예측 모델의 성능을 보기 위해 차트를 만들기 전 데이터를 어텐션 메커니즘이 적용된 양방향 LSTM에 입력값으로 주고 예측을 시도하였다. 이후 본 연구에서 제안된 모델인 VAT의 성능을 비교하기 위해 CNN, VAE, VAE+ CNN, 단순 VAE + Attention의 실험을 진행하였다. <표 4>은 진행된 실험들을 정리한 표이다.

<표 4> 실험 모델

사용 데이터	사용 모형	최적화 알고리즘	비고
주식 데이터	Attention	rmsprop	
캔들 스틱 차트	CNN	rmsprop	CNN 을 사용하여 예측
	VAE	rmsprop	잠재 값을 Dense 레이어로 예측
	VAE, CNN	rmsprop	VAE, CNN 방식을 결합
	VAE+ Attention	rmsprop	잠재 값을 Attention 에 넣고 예측
	VAT	rmsprop	VAE 방식과 위의 방식을 결합

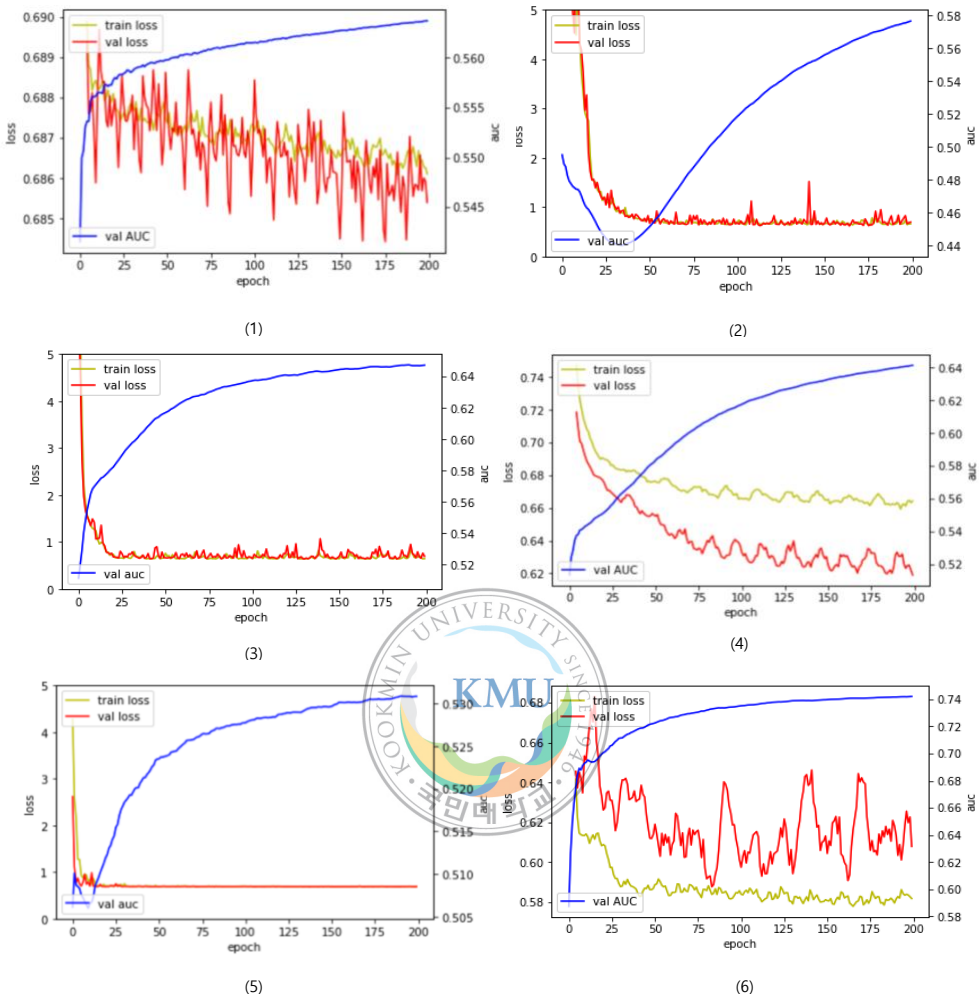
5.2 변이형 오토 인코더의 잠재 값

변이형 오토 인코더에 캔들 스틱 차트를 입력 값으로 주고 잠재 값을 얻게 된다. 변이형 오토 인코더의 잠재 값의 차원은 앞서 언급하였던 것처럼 30으로 설정하였다. 하나의 캔들 스틱 차트는 20일치의 데이터를 담고 있고 sliding window는 1일로 설정 하였으므로 30일의 데이터를 고려하기 위해서는 11개의 잠재 값이 필요하였고, LSTM모델에 입력 값으로 주기 위해 30일의 데이터들을 하나의 행으로 묶어 CSV 파일로 만들게 된다.

5.3 AUC 및 손실

<그림 9>는 5.1절에서도 언급하였다시피 본 연구에서 실험한 딥 러닝 모델들의 결과를 그래프로 표현한 것이다. 학습은 파이썬 3.6 버전에서 텐서플로우와 케라스로 진행하였고 대부분의 모델의 손실이 더 이상 하락하지 않는 200epoch 동안 진행하였다.

<그림 9>의 (1) 은 단순 주식 데이터에 어텐션 메커니즘을 적용한 학습 결과이다. (2)는 20일치의 캔들 스틱 차트를 CNN으로 예측하는 모델의 학습 결과이고, (3)은 변이형 오토 인코더의 잠재값을 dense 레이어에 입력 값으로 주고 학습 시킨 모델, (4)는 CNN방식과 단순 변이형 오토 인코더방식을 결합한 방식이다. (5)는 변이형 오토 인코더의 잠재 값을 어텐션 모델에 넣고 예측하는 모델을 학습한 결과이고 (6)은 (3)방식과 (5)의 방식을 결합한 모델의 학습 결과이다.



<그림 9> 모델 별 손실과 AUC

200epoch 학습 후 (1)방식의 경우 평가 손실은 0.6863 AUC는 0.5637, (2)방식의 경우 평가 손실은 0.6966 AUC는 0.5766, (3)방식의 평가 손실은 0.6929, AUC는 0.6469, (4)방식의 평가 손실은 0.6087, AUC는 0.6413, (5)방식의 평가 손실은 0.6871, AUC는 0.5344, 본 연구에서 제안한 방법인 (6)방식의 평가 손실 0.5609, AUC는 0.7417이 나왔다. <표 5>는 결과를 정리한 표이다.

<표 5>모델 별 결과표

	손실	정확도	AUC
Attention	0.6863	0.5506	0.5637
CNN	0.6966	0.5856	0.5766
VAE	0.6929	0.5789	0.6469
VAE, CNN	0.6087	0.7031	0.6786
VAE + Attention	0.6871	0.5359	0.5344
VAT	**0.5609**	**0.7114**	**0.7417**

본 연구에서 사용된 데이터가 주가 상승이 전체 날짜의 약 48%, 주가 하락이 전체 날짜의 약 52%로 구성되어 있는 것으로 볼 때, 좋은 성능을 보인다고 할 수 있다. 또한 제안된 모델의 성능을 객관적으로 평가하기 위해 관련 연구에서의 결과와 비교하였다. Niaki & Hoseinzade(2013), Persio & Honchar(2016), Zhong & Enke(2017), Song(2018)의 연구들은 일반적인 주식데이터를 이용하여 익일 주가의 등락을 예측한 연구들이다. Ding et al.(2015), Nguyen et al.(2015), Huynh et al.(2017)은 시장의 감성을 기반으로 익일 주가의 등락을 예측한 연구들이다. 경제관련 뉴스나 트위터의 게시물을 기반으로 시장의 감성을 추출하였다. 마지막으로 Guo et al.(2018)의 연구는 본 연구와 마찬가지로 캔들 스틱 차트를 기반으로 익일 주가의 등락을 예측하는 방법을 사용하였다. <표 6>는 각 연구에서 제안된 모델 성능을 정리한 표이다.

<표 6> 성능 비교

		저자	사용 데이터 셋		사용 방법	분석 기간	정확도
정형	기본	Niaki & Hoseinzade, 2013	S&P 500	Technical indexes	ANN	1994.3 ~ 2008.6	51.78%
		Persio & Honchar, 2016	S&P 500	Technical indexes	ANN, WCNN, RNN	2006 ~ 2016	56.9%
		Zhong & Enke, 2017	S&P 500	Technical indexes	ANN, PCA	2003.6 ~ 2013.5	59.2%
		Song, 2018	S&P 500	Technical indexes	RNN, SVM, XGBoost	2010. 1 ~ 2017. 12	64.33%
비정형	감성	Ding et al., 2015	S&P 500	Finance news	EB, CNN	2006.10 ~ 2013.11	65.48%
		Nguyen et al,. 2015	Twitter, price(OHLC)		Various	2012.7 ~ 2013.7	56%
		Huynh et al., 2017	S&P 500	Reuters News dataset	GRU	2006.10 ~ 2013.12	59.98%
	이미지	Guo et al., 2018	TAIFEX	Candlestick chart	AE, CNN	1998 ~ 2016	69.11%
	감성+ 이미지	본 연구	S&P 500	Candlestick chart, VIX chart	VAE , Attention	1993. 7 ~ 2019. 7	**71.14%**

5.4 추가 실험

본 연구에서 제안한 모델의 실용성을 보이기 위해 추가 실험을 진행하였다. 2011년 10월 14일부터 2019년 7월 31일까지의 기간 동안 무작위로 선별된 5가지 주식을 대상으로 시뮬레이션을 진행하였다. 벤치마크 수익률은 Buy & hold 방식을 사용하여 2011년 10월 14일에 주식을 산 후 계속 보유하여 2019년 7월 31일에 판매하는 것으로 가정하였다. 대조군은 본 연구의 모델이 예측하는 등락에 따라 그 날 종가에 주식을 판매하고 구매하는 것으로 설정하였다. 대상 기업은 시뮬레이션 기간 동안 주가가 10% 이상 상승한 기업, 주가가 10% 이상 하락한 기업, 주가의 변동이 ±10% 미만인 기업 3가지로 나누어 진행하였다. 대부분의 경우 본 연구에서 제안된 모델이 더 우수한 성능을 보였고 특히 시뮬레이션 기간 중 주가가 하락한 기업들의 경우 제안된 모델이 특히 우수한 성능을 보였음을 알 수 있다. 다음 장의 <표 7>은 시뮬레이션의 결과를 정리한 것이다.

<표 7> 시뮬레이션 결과

	매수 후 보유	본 연구
10% 이상 상승 기업		
H&R 블록	86%	**170%**
풋 락커	89.8%	**230.2%**
듀폰	115.9%	**167%**
10% 이상 하락 기업		
포드	-17.5%	**81%**
제프리 파이낸셜 그룹	-13.8%	**140%**
리미티드 브랜즈	-35.1%	**77%**
변동 10% 이하 기업		
Helmerich & Payne	2.3%	**91%**
켈로그	6.1%	**122.6%**
뉴웰 브랜드	**8.4%**	0.1%

10%이상 주가가 상승한 기업은 H&R블록, 풋 락커, 듀폰 등이 있었고 본 연구의 모형이 매수 후 보유 전략보다 약 1.5배에서 3배 가까이 되는 수익을 얻는 것을 확인하였다. 10%이상 주가가 하락한 기업에서도 본 연구가 더 좋은 성능을 보였다. 주가의 변동이 10%이하인 3개의 기업에서 Helmerich & Payne, 켈로그에선 VAT모델이 더 좋은 수익을 얻었지만 뉴웰 브랜드에선 더 낮은 수익을 거두었다. 토이저러스(Toys R Us)의 파산의 여파가 최근 2년간 뉴웰 브랜드의 주가 폭락에 영향을 주고 있는 사실을 고려해 볼 때, 개별 기업의 특정 사건으로 인한 영향은 VAT모형이 반응하기 힘든 것으로 보인다.

제 6장 결론

6.1 논문 요약

본 연구에서는 기존 캔들 스틱 차트 주가 등락 예측의 연구의 여러 문제점을 해결하고자 변동성 지수를 추가하고 변이형 오토 인코더와 어텐션 메커니즘을 적용하는 방법을 제안하였다. 30일치의 캔들 스틱 차트와 변동성 지수 그래프를 데이터셋으로 사용하였으며 기존 연구들과 성능을 비교하였다. 총 여섯 가지 실험을 진행하여 정확도, 손실, AUC값을 보았다.

실험 결과, CNN, VAE, VAE + LSTM등이 55~64%의 정확도와 0.64이하의 AUC값을 보여주는데 비해 본 연구에서 제안하는 VAT 모델은 최대 71%의 정확도와 0.74의 AUC값을 보여주어 가장 성능이 우수한 것을 확인했으며, 다른 연구들과 비교하였을 때도 좋은 성능을 보였다. 추가로, 2011년 ~ 2019년의 데이터를 토대로 시뮬레이션 투자를 해 본 결과 Buy & Hold 전략을 채택한 벤치 마크와 비교하여 본 연구의 모델은 더 우수한 성과를 거두는 것을 확인하였다.

6.2 연구의 시사점

본 연구는 다음과 같은 항목에서 의의가 있다. 첫째, 기존 캔들 스틱

차트연구들이 고려하지 않았던 시장의 감정상태를 변동성 지수라는 시장 지표를 추가하여 더 좋은 분석 결과를 도출하였다. 앞서 언급하였던 것처럼, 변동성 지수는 시장 참여자들의 시장에 대한 전망을 나타내는 지표이다. 시장 상태에 영향을 크게 받는 업종의 기업이나, 상황이 좋지 않은 기업 등에서 특히 더 좋은 모의 투자 결과를 보인 것으로 볼 때, 본 연구의 방법론에 대한 유용성을 반증한다고 할 수 있다.

둘째, 기업의 수와 분석 기간을 크게 늘려 자료의 대표성을 얻었다. 기존 주가 예측 연구들은 약 1년에서 10년 가량의 기간만을 데이터 셋으로 사용하여 경기변동에 따른 주식 시장의 움직임을 파악하기 힘들었지만 본 연구에서는 30년 가량의 데이터셋으로 분석을 실시하여 그 단점을 극복하였다. 1993년부터 2019년까지를 분석기간으로 잡아 90년대 후반, 서브 프라임 사태 등 원인이 다른 경기 침체기를 고려할 수 있었기 때문에 좋은 성능을 보일 수 있었던 것으로 생각된다.

셋째, 과거 연구들은 시계열데이터인 주가를 예측하는 것에 그 특성을 고려하지 않거나 결함이 있는 모델을 사용하였다. 본 연구에서는 이 문제점을 양방향 LSTM에 어텐션 메커니즘을 적용하는 방식으로 해결하였다. 해당 방법론을 적용하면서 본 연구에서 사용된 데이터 셋을 효율적으로 활용할 수 있었고, 중요 부분에 집중하여 예측하는 효과를 얻을 수 있었다.

넷째, 특징 추출에 변이형 오토 인코더를 사용함으로서 차원 축소 효과를 얻을 수 있었다. Investing.com에서 가져온 데이터 셋은 시가, 종가, 최고가, 최저가, 거래량으로 이루어져 있고 여기에 해당 날짜의 변동성 지수가 추가로 들어가 있는 형태이다. 이를 본 연구에서 사용한 20일치

의 그래프로 구성한다면 140개의 차원으로 구성된다. LSTM모형에 넣기 위해 30일의 기간을 고려한다면 140개의 차원으로 구성된 11개의 데이터 셋이 수평으로 연결되고 고려할 변수가 많아지면 많아질 수록 데이터 셋의 용량이나 모형의 성능에 큰 영향을 미치게 된다. 본 연구에서는 20일치의 이미지를 30개의 차원으로 줄여 컴퓨팅파워의 절약 뿐만 아니라 고차원의 데이터 셋에서 발생할 수 있는 잠재적인 문제점을 해결하였다.

6.3 한계점 및 향후 연구 계획

한계점으로는 첫째, 본 연구에서는 학습 시 주식의 매수와 매도에 발생하는 거래 비용을 고려하지 않았다. 모델을 학습 할 때, 주식을 거래할 때 발생하는 비용을 고려할 수 있다면 조금 더 실용적인 주가 예측 모델이 될 수 있을 것이라 생각된다.

둘째, 본 연구에서 사용된 데이터 셋에서 기업의 특성을 나타낼 만한 지표는 각 기업의 캔들 스틱 차트 뿐이었다. 추가 실험 부문에서도 언급하였지만, 제안된 모델에서는 각 기업별 특이사항을 반영할 수 있는 방법이 없기 때문에 비체계적 위험(unsystematic risk)에 노출된 기업의 주가 예측은 정확하지 않을 수 있다. 시장 전체의 상황보다 기업에 대한 의견에 집중하는 감성 분석 기반 주가예측 방법론을 본 연구에서 제안된 모델에 추가한다면 특성을 고려하지 않는 VAT모델의 단점을 해결할 수 있을 것으로 예상한다.

본 연구에서는 모든 차트를 하나의 이미지로 합쳐 분석을 실시하였는

데 여러 장의 차트를 따로 입력하여 분석하는 방식은 시도해 보지 않았다. 더 많은 종류의 차트를 추가하고 각 차트를 따로 입력 값으로 주어 분석을 한다면 더 좋은 결과를 보일 가능성이 있다.

참고 문헌

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv:1409.0473 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1409.0473

Bao, W., Yue, J., & Rao, Y. (2017). A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory. *PLOS ONE*, *12*(7), e0180944.

Bengio, Y., Frasconi, P., & Simard, P. (1993). The Problem of Learning Long-Term Dependencies in Recurrent Networks. *IEEE International Conference on Neural Networks*, 1183-1188.

Borovkova, S., & Tsiamas, I. (2019). An Ensemble of LSTM Neural Networks for High-Frequency Stock Market Classification. *Journal of Forecasting*, *38*(6), 600-619.

Chen, J.-F., Chen, W.-L., Huang, C.-P., Huang, S.-H., & Chen, A.-P. (2016). Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 87-92.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep Learning for Event-Driven Stock Prediction. *International Joint Conference on Artificial Intelligence*, 2327-2333.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*, *38*(1), 34-105.

Ghoshal, S., & Roberts, S. J. (2018). Thresholded ConvNet Ensembles: Neural Networks for Technical Forecasting. *ArXiv:1807.03192 [q-Fin]*. Retrieved from http://arxiv.org/abs/1807.03192

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. *ArXiv:1303.5778 [Cs]*. Retrieved from http://arxiv.org/abs/1303.5778

Graves, A., & Schmidhuber, J. (2005). Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, *18*(5), 602-610.

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

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780.

Huynh, H. D., Dang, L. M., & Duong, D. (2017). A New Model for Stock Price Movements Prediction Using Deep Neural Network. *Proceedings of the Eighth International Symposium on Information and Communication Technology - SoICT 2017*, 57-62.

Kim, T., & Kim, H. Y. (2019). Forecasting Stock Prices with A Feature Fusion LSTM-CNN Model Using Different Representations of The Same Data. *PLOS ONE*, *14*(2), e0212320.

Kingma, D. P., & Welling, M. (2013). Stochastic Gradient VB and the Variational Auto-Encoder. *Second International Conference on Learning Representations*, 1-14.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541-551.

Liu, H., & Song, B. (2018). Stock Price Trend Prediction Model Based on Deep Residual Network and Stock Price Graph. *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, *02*, 328-331.

Liu, Y., Qin, Z., Li, P., & Wan, T. (2017). Stock Volatility Prediction Using Recurrent Neural Networks with Sentiment Analysis. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems,* 192-201.

Maney, K. (2017). *Goldman Sacked: How Artificial Intelligence Will Transform Wall Street*. Retrieved February 26, 2017, from https://www.newsweek.com/how-artificial-intelligence-transform-wall-street-560637

Marshall, B. R., Young, M. R., & Rose, L. C. (2006). Candlestick Technical Trading Strategies: Can They Create Value For Investors? *Journal of Banking & Finance*, *30*(8), 2303-2323.

Murphy, J. J. (2011). *Intermarket Analysis: Profiting from Global Market Relationships*. Toronto: John Wiley & Sons.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment Analysis on Social Media for Stock Movement Prediction. *Expert Systems with Applications*, *42*(24), 9603-9611.

Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 Index Using Artificial Neural Networks and Design of Experiments. *Journal of Industrial Engineering International*, *9*(1), 1-9.

Olah, C. (2015). Understanding LSTM Networks. Retrieved August 27, 2015, from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Persio, L. D., & Honchar, O. (2016). *Artificial Neural Networks Architectures for Stock Price Prediction: Comparisons and Applications*. International Journal of Circuits, Systems and Signal Processing. *10*, 403-413.

Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. *ArXiv:1509.00685 [Cs]*. Retrieved from http://arxiv.org/abs/1509.00685

Schuster, M., & Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681.

Song, Y. (2018). *Stock Trend Prediction: Based on Machine Learning Methods* (UCLA). Retrieved from https://escholarship.org/uc/item/0cp1x8th

Yao, K., Zweig, G., & Peng, B. (2015). Attention with Intention for A Neural Network Conversation Model. *ArXiv:1510.08565 [Cs]*. Retrieved from http://arxiv.org/abs/1510.08565

Zhipeng, J., & Chao, L. (2019). Financial Time Series Forecasting Based on Characterized Candlestick and the Support Vector Classification with Cooperative Coevolution. *Journal of Computers*, *14*(3), 195-209.

Zhong, X., & Enke, D. (2017). Forecasting Daily Stock Market Return Using Dimensionality Reduction. *Expert Systems with Applications*, *67*, 126-139.

Abstract

Stock Prediction Using VAE and Attention Mechanism:
Focused on Candlestick Chart and VIX Indexes

by SangHyun Bae

Department of Data Science
Graduate School, Kookmin University
Seoul, Korea

Predictions of the stock market have long been the subject of much attention. Although many studies have been conducted, the stock market has a characteristic of random walk, so it was not easy to predict the future stock market. In order to overcome this, various analysis techniques have appeared, but due to various limitations, it was difficult to achieve satisfactory results.

Recent advances in deep learning techniques are increasingly attempting to apply these methodologies to financial markets. Therefore, this study proposes a method to predict the fluctuation of stock price next day using Variational Auto Encoder and Attention Model. To do this, we randomly extract 50 companies from the S & P 500 stocks and draw a candlestick chart with the market price, upper and lower price, closing price data and trading volume of each company. After that, the VIX index, which is also called the emotional score of

the market, is taken and added to the candlestick chart. The data set is extracted using VAE, and this feature is used as an input value to the attention model, and the next day's share price is predicted.

The performance of the model was evaluated by loss and AUC, and the proposed model showed better performance than the simple CNN or LSTM model, and also compared with previous studies. In addition, simulation investments were made using the trained model, and the model of this study outperformed the benchmark.