Journal of the Korea Industrial Information Systems Research Vol. 22 No.6, Dec. 2017 :23-29
http://dx.doi.org./10.9723/jksiis.2017.22.6.023 ISSN:1229-3741

순환 신경망 기술을 이용한 코스피 200 지수에 대한 예측 모델 개발 및 성능 분석 연구

(Development and Performance Analysis of Predictive Model for KOSPI 200 Index using Recurrent Neural Networks)

김 성 수[1], 홍 광 진[2]*
(Kim Sung Soo and Hong Kwang Jin)

요 약 Wealthfront, Betterment 등의 성공에 힘입어 전세계적으로 알고리즘을 통한 자동적인 자산분배 시스템인 로보어드바이저에 대한 관심이 증가하고 있다. 로보 어드바이저는 자산을 관리하는데 있어 사람의 개입을 최소화 하기 때문에 서비스를 이용하는데 드는 비용을 줄일 수 있으며 사람의 심리적 요인을 배제할 수 있다는 장점을 지닌다. 본 논문에서는 기존의 기술적 분석 기법을 대체하기 위하여 딥러닝 기술을 이용한 코스피 200 선물지수 예측 모델을 개발하고 그 성능을 분석하였다. 모델의 성능 분석 결과 제안하는 모델은 보합세에 놓인 종목의 방향성과 주가를 예측하는 문제에 활용 될 수 있음을 확인하였고, 향후 본 연구에서 제안하는 모델을 기존의 기술적 분석과 결합하여 로보어드바이저 서비스에 적용할 수 있음을 확인하였다.

핵심주제어 : 순환신경망, 코스피 200, 로보 어드바이저

Abstract Due to the success of Wealthfront, Betterment, etc., there is a growing interest in RoboAdvisor that is an automated asset allocation methodology globally. RoboAdvisor minimizes human involvement in managing assets, thereby reducing the costs of using services and eliminating human psychological factors. In this paper, we developed a predictive model for the KOSPI 200 Futures Index using deep learning, in order to replace the existing technical analysis technique. And the proposed model confirmed that When the KOSPI 200 Gift Index is small, it can be used to predict direction and price of index. In combination with the existing technical analysis, It is confirmed that the proposed models combining with existing technical analyses and can be applied to the RoboAdvisor Service in the future.

Key Words : Recurrent Neural Networks, kospi 200, RoboAdvisor

* Corresponding Author : hongmsz@gmail.com
+ 이 논문은 2017년 한국학술진흥재단의 연구비 지원에 의해 연구되었음(NRF-2017R1C1B5017187)
Manuscript receivred October 20, 2017 / revised December 10 ,2017 / accepted December 20, 2017
1) 숭실대학교 글로벌미디어학부, 제1저자
2) 숭실대학교 글로벌미디어학부, 교신저자

1. 서 론

최근 전 세계적인 저성장 추세와 그에 대응하여 진행되는 전 세계적인 금리인하 기조로 인하여 2017년 한국의 기준금리는 사상 최저치인 1.25%를 기록하고 있고, 2016년 한국의 물가상승률은 1%로 예금을 통해서는 목돈을 만들거나 노후를 대비하기 어려워 금융투자에 대한 수요가 증가하고 있다.

금융 투자방법은 크게 기본적 분석과 기술적 분석 2가지로 나눌 수 있다. 기본적 분석이란 주식 시장에서 증권의 가격이 증권의 내재가치와 일치하지 않을 수 있다는 전제하에서 증권의 내재가치를 분석하는 방법으로 내재가치보다 시장가격이 과소평가되었다면 매수를 과대평가라면 매도를 하는 방법이다. 기술적 분석은 과거의 증권가격 및 거래량의 추세와 변동패턴에 대한 정보를 이용하여 미래 증권가격의 변화를 예측하는 분석 기법이다.

2010년대 딥러닝 기술의 발견을 기점으로 하여 많은 IT 회사들이 IT와 타 분야와의 결합을 시도하고 있으며 Wealthfront, Betterment 등의 회사를 필두로 하는 로보어드바이저와 금융과 IT를 결합한 핀테크에 대한 관심과 수요가 증가하고 있다. 2016년 국내에서도 이세돌과 알파고의 대국으로 인하여 인공지능에 대한 관심이 증가되었고 금융권에서도 인공지능을 금융에 접목시키기 위한 모습들을 보이고 있으며 그에 대표적인 것이 로보어드바이저이다. 로보어드바이저는 앞에서 언급한 금융 분석 방법 중 기술적 분석을 통해 자산을 배분하는 서비스를 제공한다. 2016년 한국과학기술정보연구원에서 2021년 경 1조 9000억 원대 자산을 로봇이 운영할 것으로 예상하고 있으며 신한은행의 경우 2016년 11월 엠폴리오 운용을 시작한 이후 2017년 상반기 모바일 펀드 판매액이 46% 증가했고 모바일 펀드 가입 고객의 77%가 로보어드바이저를 통한 포트폴리오를 사용하였다. 현재 대다수 로보어드바이저 회사들의 투자 방법은 금융 공학적인 수치 계산에 따라 미래 증권가격 변화를 예상하는 방법으로 진행된다.

2012년 김유신 등[1]은 뉴스에 대한 긍정/부정 의견과 주가의 상관관계 뉴스의 유형과 주가의 상관관계를 분석하는 연구를 진행하였다. 뉴스와 주가 사이의 연관관계를 뉴스의 긍정부정 의견과 주가의 관계, 뉴스의 긍정부정 비율과 주가의 관계, 뉴스의 유형에 따른 주가의 관계로 나누어서 뉴스와 주가 간의 연관성을 찾았으며 뉴스의 긍정부정 의견과 주가 간의 연관성, 뉴스 유형과 주가의 유의미한 연관성을 확인하였다. 2010년 김선웅 등[2]은 기술적 지표를 사용하여 주가를 예측하는 기존의 논문과 달리 비가격 지표들을 연동하여 주가 예측을 시도하였다. 실험 결과 높은 정확도와 더불어 비가격 지표들의 시장 예측에 대한 효용성을 확인하였다. 그러나 앞의 연구들은 종합주가지수에 대한 예측과 달리 개별 기업종목의 경우, 뉴스 분석으로는 지수 예측이 어렵다는 한계를 가진다. 2017년 이우식[3]은 비지도 학습방법 중 하나인 은닉노드에 핵심 특성에 대한 압축된 표현을 저장하는 오토인코더를 통하여 코스피 지수를 예측하는 연구를 진행하였으며 79%의 예측 정밀도를 보였다. 2004년 김유일 등[4]은 SVM과 신경망을 통하여 코스피 200 지수를 예측하는 연구를 진행하였다. 코스피 200 데이터를 입력 값으로 받고 일주일 후의 값을 예측하는 모델을 만들었으며 신경망, SVM 모두 예측률이 50%를 넘겼다. 2002년 김헌수 등[5]은 인공신경망을 이용하여 코스피 200 주가지수 선물의 가격결정성을 실증 분석하는 연구를 하였다. 이 연구는 입력 변수의 개수와 신경망의 성능간 상관관계를 분석 하였으며 입력 변수가 7개인 경우 인공 신경망 모형이 일반 모형보다 좋은 성능을 보이는 것을 확인 하였다. 그러나 이들 연구의 경우 일반 다층 신경망을 사용한 한계로 성능 향상 효과는 그리 크지 않다는 한계를 가진다.

본 논문에서 우리는 딥러닝을 이용한 시계열 분석 모델을 사용하여 코스피 200 선물 지수를 예측하는 연구를 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 선물 시장 index 예측 시스템에 대해서 설명하고, 3장에서는 구현된 예측 시스템을 이용하여 수행한 실험에 대한 내용과 그 결과를 설명하고, 4장에서는 결론 및 향후 연구 진행 방향에 대해 설명한다.

2. 선물시장 index 예측 시스템

2.1 시스템 구성

본 논문은 코스피 200 선물 시장에서 거래할 때 다음날 선물 가격의 변화 방향성을 예측하여 수익을 창출하는 것을 목적으로 한다. 따라서 본 논문에서 우리는 실제 선물 가격 데이터와 인공 신경망을 통해 예측된 값에 따라 거래 하였을 때 각각의 모델에 대하여 어느 정도 수익을 낼 수 있는지 확인하는 것에 목적을 두고 실험을 진행한다.

주가 데이터는 현재의 값이 과거의 값들에 영향을 받는 시계열 데이터의 성질을 지닌다. 본 연구에서는 시계열 데이터의 일종인 선물 데이터를 사용하기 때문에 시계열 분석에 적합한 RNN을 사용한다. RNN은 이전 정보를 기억하여 현재의 정보에 반영을 할 수 있는 장점을 지닌다. 기존의 ANN 알고리즘에선 특정 사건이 발생 할 때 이전에 일어났던 사건을 이용하여 문제를 해결하는 것이 어려웠다. RNN은 신경망 내부에 루프가 들어 있어 과거의 데이터가 미래에 영향을 줄 수 있는 구조를 지닌다. 하지만 전통적인 RNN알고리즘을 사용하여 시계열 데이터를 처리할 경우, 신경망 층이 깊어질수록 backpropagation을 통한 계산 과정에서 값들이 점점 작아지고 결국 소멸되는 Vanishing Gradient 문제가 존재한다. 따라서 본 논문에서는 RNN 모델 중 이러한 장기 메모리 손실 문제를 해결한 LSTM(Long Short-Term Memory) 모델[6] 을 사용한다. 본 연구의 모델을 만들기 위하여 Tensorflow의 LSTM 기능을 사용하였다.

본 논문에서 사용한 데이터는 다음과 같이 구성된다. 코스피 200 주가지수 선물 데이터는 1999년 12월 28일부터 2017년 5월 30일까지의 데이터를 사용하였으며 데이터의 입력 변수는 각 거래일당 시가, 저가, 고가, 종가, 거래량 등의 5개 열로 구성된다.

본 연구의 실험 모델은 코스피 200 데이터 전체를 0이상 1 이하의 값으로 만들기 위하여 minmax scaling을 사용하여 데이터를 전처리하였으며 전처리된 데이터를 training data와 test data로 나누고 각각을 이용하여 학습과 테스트를 수행하였다. 본 연구는 사전예측구간의 길이를 일정하게 유지한 상태로 시작점을 일정하게 변경하면서 예측하는 실제 금융 공학의 자산분배 모델을 만들 때 많이 사용되는 기법인 rolling window[3] 방식('Fig. 1')을 사용한다.

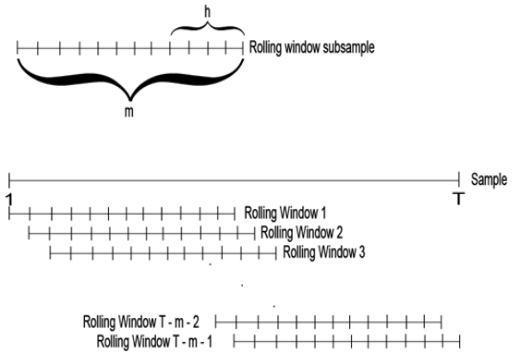

Fig. 1 Rolling window

앞서 언급한 것과 같이 예측 시스템에 사용되는 신경망 모델은 LSTM 모델('Fig. 2')을 사용한다. LSTM 모델의 각 cell은 10개의 은닉 유닛을 가지고 있으며 각 셀에 대한 입력 값은 (나뉜 데이터의 총 길이, 과거 반영하는 데이터의 길이, 데이터의 차원)으로 구성 되어 있으며 본 연구에서는 한 셀에 일반적인 주당 거래일인 5일의 주가 데이터를 넣었다. 또한 데이터의 차원은 5차원(시가, 고가, 저가, 종가, 거래량)의 형태를 가지고 있다.

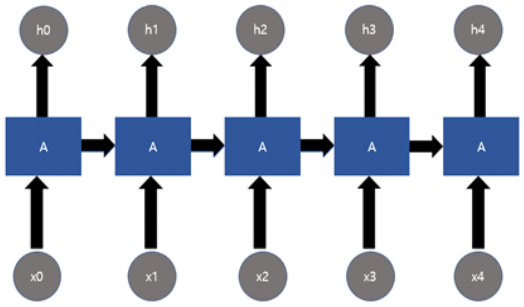

Fig. 2 LSTM

3) https://kr.mathworks.com/help/econ/rolling-window-estimation-of-state-space-models.html

2.2 분석기간 및 자료

본 논문에서 사용하는 실험 데이터는 1999년 12월 28일부터 2017년 5월 30일까지 약 17년간의 코스피 200 선물 지수 가격 데이터를 사용하였으며 데이터는 구글 finance를 통해 수집하였다.

우리는 400거래일부터 1000거래일까지 100 거래일 씩 시작점이 증가되는 기간들로 나뉜 데이터에서 60%부터 90%까지 10%씩 증가하는 분할 비율로 학습 데이터와 테스트 데이터로 나누고 학습을 통해 신경망의 가중치와 편향값을 최적화 시킨 후, 최적화된 가중치와 편향값으로 테스트 데이터를 예측한다. 실험에 사용되는 데이터 셋은 직전 5거래일의 가격 정보를 통하여 다음날의 가격을 예측하도록 구성되어 있다. 전일가 대비 예측값의 가격 변동 방향과 실제 값이 방향이 같으면 1, 아니면 0으로 표기하여 저장한 데이터 프레임을 출력한다.

3. 실험결과

3.1 모형의 성과 측정 방법

본 연구에서 각 모형들이 잘 학습 되었는지 판단하는 지표로 통계적 측정의 정확도에 대한 질적 척도로 많이 사용되는 RMSE(Root Mean Square Deviation)과 금융 공학에서 예측기의 성능 평가를 할 때 사용하는 지표인 Winrate, 만들어진 모델을 사용하였을 때 거래 기간에 평균적으로 발생하는 수익을 나타내는 Mean return을 사용하였다.

RMSE는 예측 값이 실제 값과 얼마나 유사한지를 확인할 때 흔히 사용하는 측도이며 식(1)에서와 같이 실제 가격과 예측 가격의 차이를 제곱한 결과들의 평균에 제곱근을 씌운 값이다.

$$\sqrt{\sum_{i}^{n}\frac{(\widehat{y_i}-y_i)^2}{N}} \qquad \text{식(1)}$$

Winrate은 금융 공학에서 모델이 얼마나 가격의 증감 방향성을 잘 예측 하는지를 나타내는 지표이며 식(2)에서와 같이 전체에서 맞게 예측된 값이 얼마만큼 있는 지를 나타낸다.

$$\frac{True}{True+False} \qquad \text{식(2)}$$

Mean return은 본 실험에서 제안한 시스템을 통해 실제 운용을 하였을 때 얼마만큼의 평균 수익이 발생하는지를 나타내는 값이며 식(3)에서와 같이 도출되는 전체 수익을 합산한 값에 거래일 만큼 나눈 값이다.

$$\sum_{i}\frac{y_i-y_{i-1}}{N} \qquad \text{식(3)}$$

3.2 실험 결과

본 논문에서 우리는 모델의 실험과정에서 과거 5 거래일의 데이터를 학습하여 다음날 하루를 예측하고 다음날은 그 직전 5 거래일까지의 데이터를 학습하여 값을 예측하는 당일 예측방법을 사용하였다. 이러한 방법으로 테스트 데이터를 만들어 놓고 각 데이터의 평균 RMSE, 평균 Winrate, 평균 누적수익을 이용하여 모형을 검증하였다.

본 연구에서는 제안한 시스템을 운용할 시 얼마만큼의 rolling window 방식의 사전예측구간을 설정하는지, 얼마만큼의 training, test 데이터의 분할 비율을 사용할 때 수익이 최대가 나오는지를 확인하기 위해 데이터를 400, 500, 600, 700, 800, 900, 1000의 기간 값으로 나누며 60%, 70%, 80%, 90%의 분할 비율을 가지고 실험을 진행하였다.

모델의 평균 Winrate는 모델을 통해 나온 전날 대비 예측가격의 증감 방향과 실제 데이터의 전날대비 증감 방향을 비교하여 두 값이 일치하면 1, 다르면 0이 되도록 하여 전체 데이터 셋의 길이에 대한 1의 개수의 비율을 구하였다.

제안하는 모델로 시뮬레이션할 때, 구매 방법은 일반적인 투자 방법인 종목을 직접 사서 보유

하는 long position과 해당 종목이 떨어질 것을 예측하여 해당 종목을 빌려서 파는 short position을 혼합하여 사용하였다. 수익은 현재의 코스피 200 선물의 거래승수에 따라 1 포인트당 25만원으로 가정한다. 위의 규칙에 따라 본 연구의 모델을 데이터의 기간만큼 수행하고 그 수익들의 누적 합으로 총 수익을 구한다. 데이터의 총 수익으로 성능을 비교 할 시 기간 값을 짧게 잡을수록 거래일 수가 증가하여 수익이 크게 나타나 모델의 성능과 총 수익 간의 일치성에 위배될 수 있으므로 구해진 총 수익에 총 거래일을 나눠 평균 수익을 구한다.

RMSE는 식(1)을 이용하여 도출하였으며 Winrate은 식(2)를, 평균 수익은 식(3)을 이용하여 도출하였고, 평균 수익은 총 수익을 총 거래일로 나눈 값으로 추정하였다. 'Table 1'에서 보는 것처럼 RMSE를 기준으로 분류하였을 때 500 기간 값과 0.7 학습, 테스트 데이터 셋 분할 비율을 가진 경우 가장 낮은 평균 RMSE 값을 가졌으며, 그래프로 표현하면 'Fig. 3'과 같다.

Table 1 top 5 test by RMSE

Term	Train_rate	Average RMSE
500	0.7	0.006463
600	0.7	0.007162
400	0.7	0.007217
800	0.6	0.007428
900	0.6	0.007554

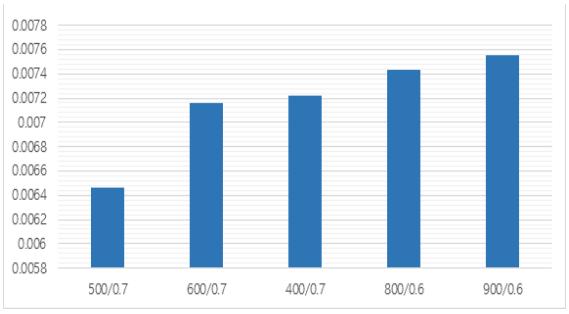

Fig. 3 chart of Table 1

'Table 2'에서와 같이 실험을 진행한 결과 평균 Winrate을 기준으로 분류하였을 때 900일 기준과 60%의 training 데이터 비중을 가진 경우가 약63.92%로 가장 높은 승률을 보였으며 그래프로는 'Fig. 4'와 같이 표현된다.

Table 2 top 5 test by Winrate

Term	Train_rate	Average Winrate
900	0.7	0.006463
600	0.7	0.007162
400	0.7	0.007217
800	0.6	0.007428
900	0.6	0.007554

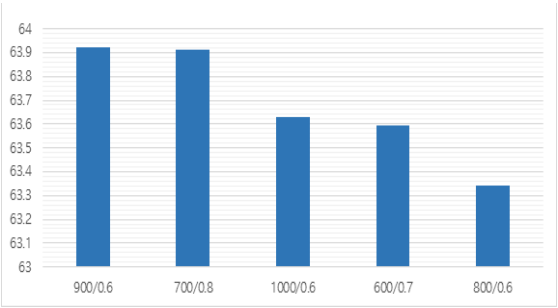

Fig. 4 chart of Table 2

또한 'Table 3'과 같이 평균 수익을 기준으로 분류하였을 때 800일 기준과 90%의 training 데이터 비중을 가진 경우가 거래일당 수익의 평균이 약 0.54 포인트로 코스피 200 선물의 거래승수로 환산하였을 경우 평균 13,5000원의 가장 높은 포인트 당 수익을 보였다. 이를 그래프로 그렸을 때 'Fig. 5'와 같이 표현된다.

Table 3 top 5 test by return

Term	Train_rate	Average mean return
800	0.9	0.543859
400	0.8	0.511233
1000	0.9	0.500202
600	0.9	0.484512
400	0.9	0.476794

실험 결과 제안한 시스템을 통해 얻어진 RMSE 값은 기존 연구 중 일반 다층 신경망을 사용하여 예측된 결과와 비교하여 약 4배의 정확도를 보였고, Winrate의 경우, 금융 공학 분야에서 통상적으로 이야기하는 50% 보다 높은 결과

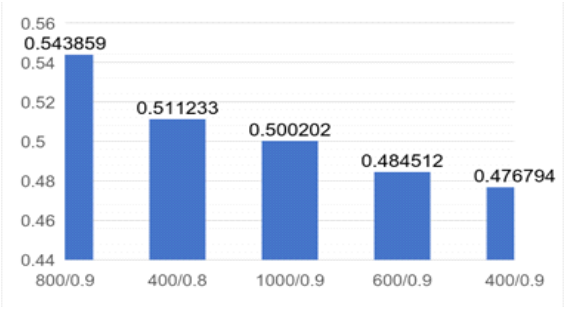

Fig. 5 chart of Table 3

를 얻음으로써, 기존 연구와 비교하여 좋은 성능을 얻을 수 있음을 확인하였다. 또한 mean return이 0보다 크게 나옴으로써 제안한 시스템을 통해 실제 운용 시 수익을 얻는 것이 가능함도 확인하였다.

4. 결 론

2010년대 이후로 전세계적으로 딥러닝에 대한 관심과 연구가 활성화되고 있으며 다양한 분야에 인공지능과의 접목을 하려는 경향을 보인다. Wealthfront, Betterment의 성공에 힘입어 현재 한국 금융업계에서도 쿼터백, 디셈버 등의 여러 로보어드바이저 서비스 회사들이 등장하였고 지속적으로 관련 회사들이 증가하는 추세이다. 본 논문에서 우리는 기술적 주가분석 방법으로 기존의 금융 공학적인 방법이 아닌 LSTM 모형을 활용한 방법을 제안하고 한국 코스피 200 선물지수의 증감 방향성 예측에 대해 분석하였다. 실험 결과를 통해 우리는 다음의 결과를 확인하였다. 60일 변동성 및 배당락 지수 등 다른 지표들을 사용하지 않고 코스피 200 지수의 가격만을 변수로 사용하여도 코스피 200 주시의 방향을 예측하는 것이 가능함을 확인하였다. 제안하는 모형을 이용하여 일정 기간 동안 실제 투자 시뮬레이션을 돌렸을 때 수익이 발생함을 확인하였다.

향후 연구는 다음과 같이 진행할 예정이다. 제안하는 모델의 성능 개선을 위해 코스피 200 선물지수의 가격과 밀접한 상관계수가 있는 다른 지표들을 연동한 개선된 모델을 개발하고, 2007년 금융위기, 2011년 그리스발 유로존 위기와 같이 외부적 요인으로 인한 심한 변동성을 가진 경우 이를 빠르게 감지하여 거래에 적용시키는 방안을 연구할 예정이다.

References

[1] Kim Y., Kim N., Jeong S., “Stock-Index Invest Model using News Big Data Opinion Mining,” Journal of Intelligence and Information System, Vol. 28, No. 2, pp. 143-156, 2012.

[2] Kim S. and Ahn H., “Development of an Intelligent Trading System Using Support Vector Machines and Genetic Algorithms,” Journal of Intelligence and Information System, Vol. 16, No. 1, pp. 71-92, 2010.

[3] Lee W., “A deep learning analysis of the KOSPI’s directions,” Journal of the Korean Data and Information Science Society, Vol. 28, No. 2, pp. 287-295, 2017.

[4] Kim Y., Shin E., Hong T., “Comparison of Stock Price Index Prediction Performance Using Neural Networks and Support Vector Machine,” The Journal of Internet Electronic Commerce Resarch, Vol. 4, No. 3, pp. 221-243, 2004.

[5] Kim H., Kim K., Jeong D., “A Study on the Price Determination of KOSPI 200 Futures using Artificial Neural Network Model,” Korea Insurance Research Institute, Insurance Financial Research, Vol. 13, No. 3, pp. 155-176, 2003.

[6] Hochreiter S. and Schmidhuber. J., “Long Short-Term Memory,” Neural Computation, pp. 1735-1780, 1997.

[7] Ban J., Kim M., Jeon Y., “Search Frequency in Internet Portal Site and the Expected Stock Returns,” Journal of the Korea Industrial Information Systems Research, Vol. 21, No. 5, pp. 73-83, 2016.

[8] Hwang R., Kim S., Lee D., Nam D., “A

Directional Distance Function Approach on the Efficiency of Chinese Commercial Banks," Journal of the Korea Industrial Information Systems Research, Vol. 17, No. 2, pp. 81-94, 2017.

[9] Ahn H., "A Study on Compression of Connections in Deep Artificial Neural Networks," Journal of the Korea Industrial Information Systems Research, Vol. 22, No. 5, pp. 17-24 ,2017.

김 성 수 (Kim Sung Soo)

- 학생회원
- 숭실대학교 글로벌미디어학부 학사과정
- 관심분야 : 인공지능, 금융공학, 컴퓨터 비전

홍 광 진 (Hong Kwang Jin)

- 정회원
- 숭실대학교 컴퓨터학부 학사
- 숭실대학교 미디어학과 공학석사
- 숭실대학교 미디어학과 공학박사
- 숭실대학교 IT대학 글로벌미디어학부 조교수
- 관심분야 : 인공지능, 영상처리, 컴퓨터 비전