



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

## 석사학위논문

# 인공지능을 이용한 주가 방향성 예측 및 투자 포트폴리오 최적화

Stock Price Trends Forecasting and Investment Portfolio  
Optimization using Artificial Intelligence



국민대학교 일반대학원  
컴퓨터공학과 컴퓨터공학전공

황 석 진

2018

# 인공지능을 이용한 주가 방향성 예측 및 투자 포트폴리오 최적화

Stock Price Trends Forecasting and  
Investment Portfolio Optimization using  
Artificial Intelligence

지도교수 최 준 수

이 論文을 工學碩士學位 請求論文으로 提出함

2019 년 2 월

국민대학교 일반대학원  
컴퓨터공학과 컴퓨터공학전공

황 석 진

2018

黃錫珍의

工學碩士學位 請求論文을 認准함

2019 년 2 월



審査委員長 김 준 호 ㉠

審査委員 최 준 수 ㉠

審査委員 임 은 진 ㉠

국민대학교 일반대학원

# 차 례

|                                       |     |
|---------------------------------------|-----|
| 그림 차례 .....                           | iii |
| 표 차례 .....                            | v   |
| 국문요약 .....                            | vi  |
| 제 1장 서론 .....                         | 1   |
| 제 2장 관련 연구 .....                      | 5   |
| 2.1 딥러닝 (Deep Learning) .....         | 5   |
| 2.2 Inception-ResNet .....            | 7   |
| 2.3 유전 알고리즘 (Genetic Algorithm) ..... | 11  |
| 제 3장 딥러닝을 이용한 주가 예측 .....             | 14  |
| 3.1 데이터 전처리 .....                     | 14  |
| 3.1.1 주가 정보 분석 .....                  | 14  |
| 3.1.2 데이터 선정 및 수집 방법 .....            | 15  |
| 3.1.3 데이터 전처리 방법 .....                | 16  |
| 3.2 방향성 예측 모델 .....                   | 21  |
| 3.3 기사 제목 학습 모델 .....                 | 25  |
| 3.4 실험 .....                          | 34  |
| 제 4장 유전 알고리즘을 이용한 포트폴리오 최적화 .....     | 35  |
| 4.1 투자 종목 최적화 .....                   | 35  |
| 4.1.1 적합도 함수 .....                    | 35  |
| 4.1.2 실험 방법 .....                     | 35  |
| 4.2 투자 비율 최적화 .....                   | 36  |

|                                 |    |
|---------------------------------|----|
| 4.2.1 적합도 함수 .....              | 36 |
| 4.2.2 실험 방법 .....               | 37 |
| 4.3 투자 종목 최적화 + 투자 비율 최적화 ..... | 37 |
| 4.3.1 적합도 함수 .....              | 37 |
| 4.3.2 실험방법 .....                | 37 |
| 제 5장 결과 및 성능평가 .....            | 39 |
| 제 6장 결론 및 향후 연구 과제 .....        | 46 |
| Abstract .....                  | 50 |



## 그림 차례

|                                     |    |
|-------------------------------------|----|
| 그림 1-1 미국 내 로보 어드바이저 현황 .....       | 2  |
| 그림 1-2 국내 로보 어드바이저 시장 규모 전망 .....   | 2  |
| 그림 2-1 합성곱 신경망 .....                | 6  |
| 그림 2-2 Network in network.....      | 8  |
| 그림 2-3 Inception 모듈 .....           | 9  |
| 그림 2-4 망 깊이에 따른 오류 .....            | 10 |
| 그림 2-5 Resnet 구성 예 .....            | 10 |
| 그림 2-6 Inception-ResNet 모듈의 예 ..... | 11 |
| 그림 2-7 지역적 최적해와 전역적 최적해 .....       | 12 |
| 그림 2-8 유전 알고리즘 과정 .....             | 13 |
| 그림 3-1 1년간 주가 그래프 .....             | 15 |
| 그림 3-2 1개월간 주가 드래프 .....            | 15 |
| 그림 3-3 주가 방향성 예측 모델 .....           | 21 |
| 그림 3-4 주가 정보 학습을 위한 인셉션 모델.....     | 22 |
| 그림 3-5 기사 제목 학습을 위한 인셉션 모델.....     | 23 |
| 그림 3-6 Stem1 .....                  | 26 |
| 그림 3-7 Stem2 .....                  | 26 |
| 그림 3-8 Inception-A1 .....           | 27 |
| 그림 3-9 Inception-B1 .....           | 28 |
| 그림 3-10 Inception-C1.....           | 29 |
| 그림 3-11 Inception-A2.....           | 30 |

|                           |    |
|---------------------------|----|
| 그림 3-12 Inception-B2..... | 31 |
| 그림 3-13 Inception-C2..... | 32 |
| 그림 3-14 Reduction-A.....  | 33 |
| 그림 3-15 Reduction-B.....  | 33 |
| 그림 5-1 누적 등락률 .....       | 45 |
| 그림 6-1 누적 수익률 .....       | 46 |





## 표 차례

|  |    |
|--|----|
| 표 1-1 국내 금융회사 로보 어드바이저 서비스 추진 현황 ..... | 3  |
| 표 3-1 수집 데이터 및 학습데이터 개수 .....          | 16 |
| 표 3-3 수집데이터 샘플 .....                   | 17 |
| 표 3-4 특성에 따른 지표의 분류 .....              | 17 |
| 표 3-5 사용된 지표 및 공식 .....                | 18 |
| 표 5-1 학습 데이터 상위 15 개 종목 정확도 .....      | 39 |
| 표 5-2 학습 데이터 하위 15 개 종목 정확도 .....      | 40 |
| 표 5-3 검증 데이터 상위 15 개 종목 정확도 .....      | 41 |
| 표 5-4 검증 데이터 하위 15 개 종목 정확도 .....      | 41 |
| 표 5-5 평가 데이터 상위 15 개 종목 정확도 .....      | 42 |
| 표 5-6 평가 데이터 하위 15 개 종목 정확도 .....      | 42 |
| 표 5-7 최적화 실험 방법별 평균 등락률 비교 .....       | 43 |
| 표 5-8 월별 평균 등락률 및 전월 대비 수익률 비교 .....   | 44 |

## 국문 요약

# 인공지능을 이용한 주가 방향성 예측 및 투자 포트폴리오 최적화

국민대학교 대학원 컴퓨터공학과

황석진

주식투자는 가장 많이 알려진 재테크 방법이지만 실제 수익을 얻기는 매우 어려운 투자 방법이다. 효과적이고 안정적인 투자를 위해 주가의 변화를 예측하는 것은 경제, 수학 등 여러 분야에서 오랜 관심의 대상이었다. 하지만 과거부터 많은 노력에도 불구하고 여전히 어려운 문제로 남아 있으며, 전산학에서도 이 문제를 해결하기 위해 의사결정나무(Decision Tree), SVM(Support Vector Machine) 등의 기계학습 방법을 활용하여 노력해왔다.

본 논문에서는 딥러닝(Deep Learning)을 이용하여 다음 날 주가의 방향성을 예측하고 딥러닝 학습결과를 바탕으로 유전 알고리즘을 활용하여 투자 종목을 선정하고 투자 비율을 최적한 투자 포트폴리오 구성 방법을 제안한다. 딥러닝의 모델은 주식과 같은 시계열 데이터에 적합하다고 알려진 순환 신경망(RNN: Recurrent Neural Network)대신 Inception-Resnet-v2 모델을 기반으로 새로운 합성곱 신경망(CNN: Convolutional Neural Network)를 사용한다. 종목별로 주가 정보를 보조지표로 가공한 데이터와 경제 기사의 제목을 학습 데이터로 사용한다. 학습 후에는 유전 알고리즘을 이용하여 정밀도(Precision) / 재현율(Recall) / 정확도(Accuracy)의 결과와 당일 예측 확률을 반영하여 투자 종목을 선정하고,

선정된 종목의 투자 비율을 최적화 하는 실험을 진행한다. 그리고 벤치마크로 사용된 주가 지표와 수익성을 비교하여 본 논문에서 제안한 방법의 유용성을 증명한다.



주제어 : 주가예측, 딥러닝, 유전알고리즘

## 제 1장 서론

계속되는 금리 인하로 인해 전통적인 재테크 방법인 예금/적금보다 원금 손실 위험이 있더라도 고수익을 올릴 수 있는 재테크 방법에 대한 관심이 커지고 있다. 그중 주식투자는 가장 많이 알려진 재테크 방법으로 적은 초기 비용으로 시작할 수 있으면서 온라인 상에서 증권사를 통해 쉽게 거래를 할 수 있기 때문에 진입장벽도 비교적 낮은 재테크 방법이다.

주식 투자 방법은 기간에 따라 저평가된 주식을 매입한 후 장기간 보유하는 ‘장기 투자’와 짧은 기간 내 주식을 거래하여 수익을 올리는 ‘단기 투자’가 있다. 그리고 단기 투자 방법은 보유 시간과 거래 빈도에 따라 스캘핑(scalping), 데이트레이딩(day trading), 스윙(swing)으로 나뉜다. 스캘핑은 ‘초’나 ‘분’ 단위의 아주 짧은 시간에 매매하는 방법으로 초단타 매매라고도 하며, 데이트레이딩은 하루 안에 매수와 매도를 마무리 짓는 방법으로 장 종료 후에는 남는 주식이 없는 것이 일반적이다. 스윙은 1 ~ 7일 가량 주식을 보유하는 것으로 중단기 매매라고 볼 수 있다. 하지만 위 방법 모두 높은 위험성을 내포하고 있으며, 주식 투자를 통해 실제 수익을 내는 것은 매우 어렵다.

주식투자와 같은 위험성 높은 재테크 방법은 그 위험을 낮추기 위해 전문적인 자산 관리 서비스를 이용하는 것이 필요하지만, 높은 수수료와 최저 투자 비용으로 인해 이용이 쉽지 않다. 하지만 핀테크 산업의 성장과 인공지능 기술의 발달로 전문 자산 관리사가 아닌 프로그램을 통해 자산관리를 받는 로보 어드바이저(robo-advisor)가 나오면서 적은 금액으로 자산관리는 받는 것이 가능해졌다. 그리고 로보 어드바이저를 통한 자산 관리 시장은 글로벌 금융위기 이후 저렴한 수수료를 기반으로 빠르게 성장하고 있다. ‘MyPrivateBanking’에 의하면 전 세계 로보 어드바이저에 의한 자산 관리 규모는 2015년 200억 달러에서 2020년 4,500억 달러로

약 23배 가량 성장할 것으로 보이며 [1], 미국의 경우 총 투자액 중 로보 어드바이저로 운용되는 자산 비중이 2015년 0.5%에서 2020년 5.6%까지 커질 것으로 전망하고 있다 [2].

국내 로보 어드바이저 시장은 아직 시작 단계로 당장 미국과 비교하기에는 무리가 있다. 하지만 미국 및 해외 시장과 마찬가지로 큰 성장을 보일 것으로 판단된다. 유진투자증권에 따르면 국내 로보 어드바이저 시장은 아직 도입 및 정착기간에 있으며, 2025년 정도에는 46조 원 정도로 성장할 것으로 보고 있다 [3]. 실제로 2010년 이후부터 많은 금융사에서 로보 어드바이저 서비스를 시작하고 있다 [4].

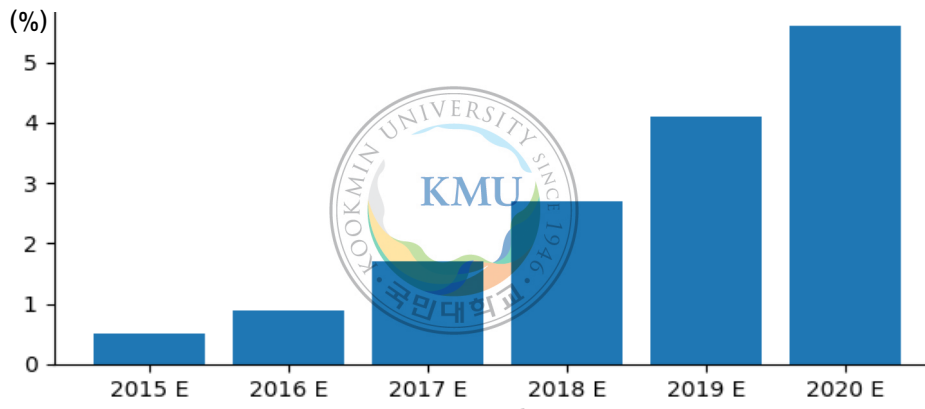


그림 1-1 미국 내 로보 어드바이저 현황

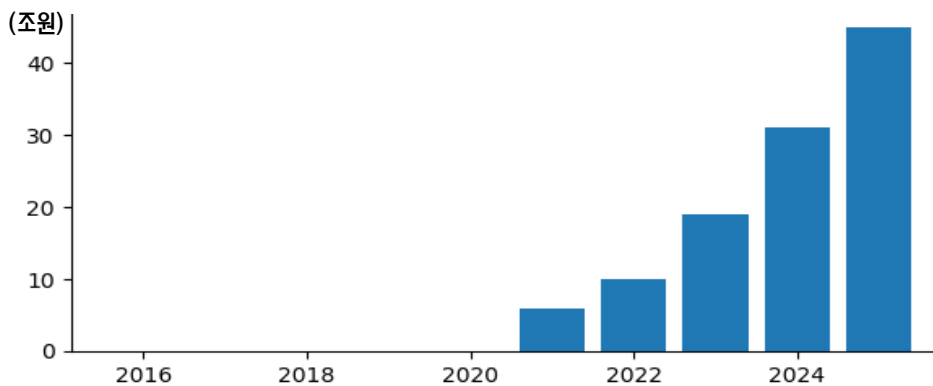


그림 1-2 국내 로보 어드바이저 시장 규모 전망

표 1-1 국내 금융회사 로보 어드바이저 서비스 추진 현황

| 유형  | 회사        | 주요내용   |
|-----|-----------|--|
| 증권사 | NH 투자증권   | 2011 년부터 운용했던 온 스마트인베스터를 기반으로 'QV 로보 어카운트'서비스 오픈                           |
|     | 삼성증권      | '투자성과 정밀검증 알고리즘 시스템' 관련 특허 출원  |
|     | 신한금융투자    | 2016 년 4 월 밸류시스템 로보어드바이저인 '아이로보'의 포트폴리오를 기반으로 운용되는 '신한명품 밸류시스템 자문형 로보랩' 출시 |
|     | 한국투자증권    | 2016 년 4 월 랩 어카운트 상품 운용 과정에 로보어드바이저의 판단이 반영된 '한국투자 로보랩' 출시                 |
|     | 현대증권      | 2016 년 2 월 로보어드바이저에 기반한 일임형 랩인 '현대 able 로보랩' 출시                            |
| 운용사 | 미래에셋자산 운용 | 2015 년 5 월부터 국내 최초로 로보어드바이저 서비스인 '글로벌 자산배분 솔루션' 제공                         |
|     | 삼성자산운용    | 2016 년에 V모멘텀 솔루션 등 퀀트분석에 기반한 ETF 자산배 분 서비스 출시 예정                           |
|     | KB 자산운용   | 2016 년 하반기 중 계열은행 증권사를 통해 로보어드바이저 투자솔루션 제공 예정                              |
| 은행권 | 우리은행      | 파운트와 제휴한 '로보어드알파' 출시, ISA 에 접목 가능  |
|     | KB 국민은행   | 은행권 최초 퀀터백투자자문과 MOU 를 체결하여 '퀀터백 R-1' 출시                                    |
|     | KEB 하나은행  | 2016 년 2 월 금융상품 포트폴리오 서비스 'Cyber PB(PrivateBank)'출시                        |
|     | 신한은행      | 2016 년 4 월 로보어드바이저 기반 펀드 추천 서비스 'S 로보 플러스' 출시                              |

[오류! 참조 원본을 찾을 수 없습니다.]와 같이 주식 투자에서도 로보 어드바이저를 활용하기 위한 시도를 하고 있다. 주식 투자에서 로보 어드바이저의 가장 중요한 역할은 효과적이고 안정적인 투자를 위해 주가의 변동을 예측하는 것으로, 이 문제는 경제학, 통계학 등 여러 분야에서 이를 해결하기 위해 노력해 왔지만, 큰 성과를 내지 못하고 여전히 어려운 문제로 남아있다. 전산학에서도 종목별 주가 변동 및 코스피 지수와 같은 주가 지수를 예측하기 위해 의사결정나무(Decision Tree) [5],

SVM(Support Vector Machine) [6], 유전 알고리즘(Genetic Algorithm) [7] 등 다양한 기계학습 방법을 활용하여 노력해왔으며, 최근에는 다양한 분야에서 좋은 결과를 내고 있는 딥러닝을 활용한 연구가 많이 진행되고 있다.

딥러닝을 활용한 주가 예측 연구는 크게 두 가지로 나뉘 볼 수 있다.

(1) 입력 정보의 다변화.

(2) 새로운 학습 모델 사용.

(1)의 경우 일반적으로 사용되는 주가 정보 외에 새로운 정보를 사용하거나 기존에 사용하던 정보를 전처리하는 과정을 다양하게 변화시킨다. 그 대표적인 예로 기사와 같은 자연어 형태의 데이터를 글자나 단어, 문장 단위로 학습데이터를 구성하는 실험 [8] [9], 주가 정보를 다양한 방법으로 전처리하여 학습 데이터로 실험하는 방법 등이 있다. (2)의 경우에는 전방향 신경망(Feedforward Neural Network) [10] [11], 장단기 기억 신경망(Long Short-Term Memory) [9] [12], 양방향 장단기 기억 신경망(Bi-directional Long Short-Term Memory) [13] 등을 다양한 딥러닝 모델을 사용하는 연구들이 있다. 본 논문과 유사한 주제를 다뤘던 대부분의 연구는 60% 안팎의 정확도를 소개하고 있지만 이후 실질적인 투자 활용 방안에 대해선 언급하지 않는 것이 대부분이다.

본 논문에서는 단순 주가 방향성 예측 모델에서 벗어나 완성된 예측 모델을 최대로 활용할 수 있도록 예측 모델을 기반으로 하는 포트폴리오 구성 방법을 함께 제안하여 데이터레이딩을 위한 투자 포트폴리오 구성에서 로보 어드바이저의 가능성을 확인하고자 한다. 본 논문의 주요 연구인 딥러닝을 이용한 주가 방향성 예측에서는 시계열 데이터를 합성곱 신경망에 적합한 데이터 형태로 변형하는 전처리 방법과 Inception-Resnet-

v2 모델을 기반으로 만들어진 주가 방향성 예측 모델, 유전 알고리즘을 이용한 포트폴리오 구성방법을 소개한다.

## 제 2장 관련 연구

본 장에서는 이번 연구에서 주가 방향성 예측에 사용된 기술인 딥러닝과 인셉션(Inception) 모델 그리고 투자 종목을 선정하고 투자 비율을 최적화하는 포트폴리오 구성에 사용된 유전 알고리즘에 대해 설명한다.

### 2.1 딥러닝 (Deep Learning)

딥 러닝은 기존 인공 신경망(Artificial Neural Network)의 발전된 형태로 인공 신경망의 한계를 극복한 다층 신경망(Multi-Layer Neural Network)을 의미한다. 과거 신경망은 학습 효율을 저하시켰던 기울기 소실(Vanishing Gradient)문제와 과적합(Overfitting)문제 그리고 하드웨어 한계로 인한 학습 속도 문제 등으로 연구가 침체됐으나 이 문제를 해결하는 방법들이 제안되고 2012년 ImageNet에서 딥 러닝을 사용한 AlexNet [14]이 압도적인 차이로 우승한 이후 많은 연구가 진행되고 있다.

다층 신경망은 기존의 인공 신경망의 입력층(input layer)과 출력층(output layer) 사이에 다수의 은닉층(hidden layer)을 포함하고 있다. 인공 신경망은 입력에 대한 출력값과 정답 데이터 사이의 오차를 줄이는 지도 학습(Supervisor Learning)을 기본으로 하며, 오차를 줄이고 가중치를 갱신하기 위해 오류 역전파 알고리즘(Backpropagation Algorithm)을 사용한다.

대표적인 인공 신경망 방법으론 합성곱 신경망(CNN : Convolution Neural Network)과 순환 신경망(RNN : Recurrent Neural Network)이 있



으며, 순환 신경망은 자연어와 같은 시계열 데이터에서 합성곱 신경망은 이미지 인식 분야에서 많은 성과를 내고 있다.

이번 연구에서 학습 데이터로 사용하는 주가 정보의 경우 시계열 데이터로 보는 것이 일반적이지만 순환 신경망을 이용한 실험에서 큰 성과를 보지 못했다. 따라서 본 논문에서는 순환 신경망을 이용한 주가 정보 분석의 한계를 해결하고자 합성곱 신경망인 인셉션 모델을 기반으로 한 모델을 사용한다.

기본적인 합성곱 신경망은 동물이 시각 정보를 인식하는 방법을 모방한 것으로, 그 과정을 보면 크게 3단계 과정으로 나뉘볼 수 있다.

- (1) 정보의 특징을 추출하는 단계
- (2) 위상 변화에 의한 영향을 줄이는 단계
- (3) 분류기 단계

합성곱 신경망은 합성곱(convolution) 계층과 풀링(pooling) 계층으로 이뤄져 있으며, 마지막 출력층은 (3) 단계인 분류를 위한 다층 신경망으로 이뤄져 있다. 합성곱 계층은 입력 정보의 특징 추출하는 (1)단계의 역할을 하고, 풀링 계층은 데이터의 크기를 줄여 연산 속도를 높이면서 필터 내 대표값을 뽑아 회전과 이동 같은 위상 변화에 적응시키는 (2)단계의 역할을 담당한다[그림 2-1].

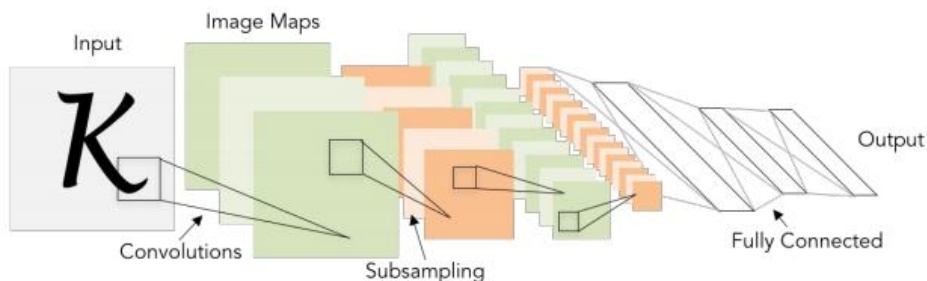


그림 2-1 합성곱 신경망

## 2.2 Inception-ResNet

### a) Inception

‘Inception’ 모델은 2014년 ‘GoogleNet’ [15]으로 처음 발표된 후 2016년 ‘Inception-v4’와 ‘Inception-ResNet-v2’ [16]까지 지속적으로 개선된 이미지 인식을 위한 합성곱 신경망의 한 종류이다. 2014년부터 합성곱 신경망에 큰 변화가 생겼으며, ‘Inception’ 모델은 그중 가장 큰 영향을 준 모델 중 하나이다.

2014년 이전의 합성곱 신경망은 망이 깊어지면 발생할 수 있는 과적합과 기울기 소실 문제 등으로 인해 10개 레이어 전후의 작은 크기로 모델을 구성하는 전형적인 형태를 취하고 있었다. 인셉션 모델에서는 아래 방법을 이용하여 이런 문제를 해결하고 깊고 넓은 망을 구성했다.

- (1) ‘ $1\times 1$ ’ 합성곱 레이어
- (2) 합성곱 레이어의 병렬적 활용
- (3) 보조 분류기 (auxiliary classifier)

인셉션 모델은 2013년 발표된 ‘Network in network’ [17]의 영향을 크게 받았다. [17]에서는 전형적인 합성곱 레이어는 지역적 특징 추출에는 효과적이지만 합성곱 레이어에 사용되는 필터가 선형적인 특징을 가지고 있기 때문에 비선형적 특징 추출에는 어려움이 있고 이를 극복하기 위해 선 특징점(feature map)의 개수를 늘려야 한다는 문제에 주목했다. 그래서 [그림 2-2]과 같이 합성곱 필터 대신 다층 신경망을 사용하여 특징점을 추출하는 방법을 제안했다.

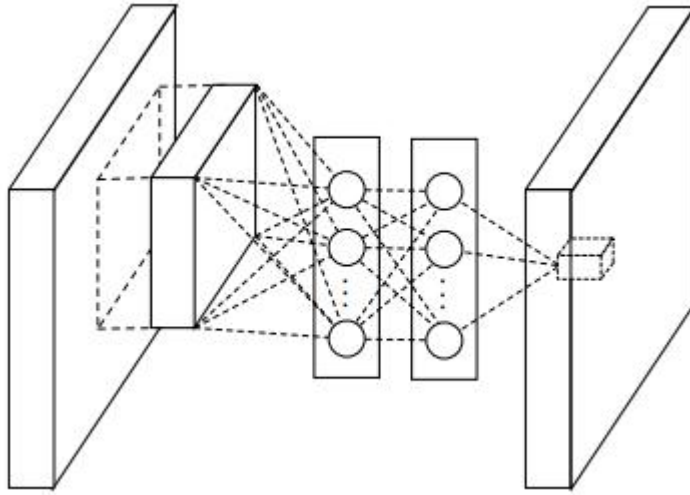


그림 2-2 Network in network

인셉션 모델에서는 [17]에서 말하는 비선형적 특징을 추출하기 위한 방법으로 ‘ $1 \times 1$ ’ 합성곱 레이어를 사용한다. 이 역시 [17]에서 소개된 것으로 본래는 차원을 줄이는 목적으로 사용되었다. 물론 이것은 인셉션 모델에서도 ‘ $1 \times 1$ ’ 합성곱 레이어를 사용하는 가장 중요한 이유 중 하나이지만, ‘ $1 \times 1$ ’ 합성곱 레이어의 연산은 다층 신경망과 같은 연산을 하기 때문에 같은 효과를 기대할 수 있다.

그리고 inception 모듈 내에서 다양한 지역적 특징을 추출하기 위해 합성곱 레이어를 병렬적으로 배치하여 활용하며, 이로 인해 문제 될 수 있는 많은 연산량은 ‘ $1 \times 1$ ’ 합성곱 레이어를 이용해 차원을 줄임으로써 해결하고 있다[그림 2-3].

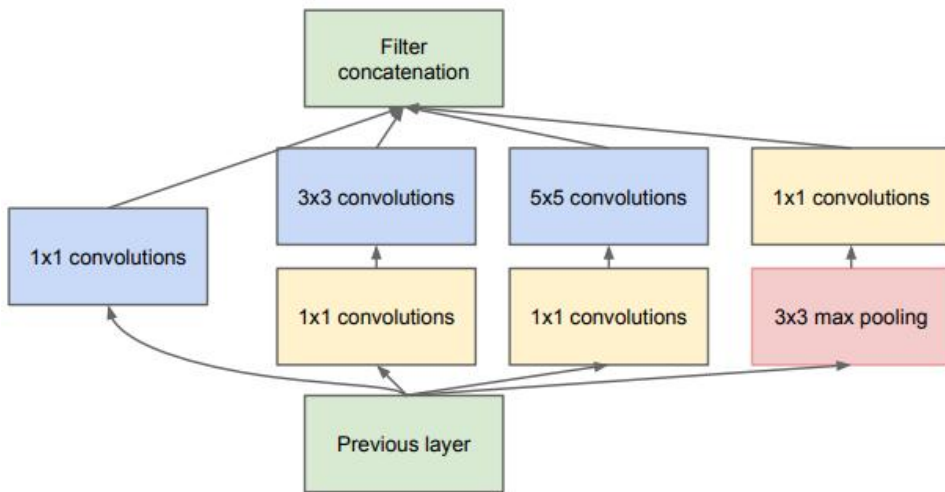


그림 2-3 Inception 모듈

그리고 망이 깊어지면서 발생하는 기울기 소실 문제를 해결하기 위해 보조 분류기를 사용하고 있다. 보조 분류기는 망 중간에 망 끝부분에 들어가는 분류 부분을 추가하는 것으로, ‘Inception’ 모델은 보조 분류기를 이용하여 기울기 소실 문제를 회피하고 빠른 수렴을 할 수 있게 되었다.

#### b) Inception-ResNet-v2

a) 에서 사용한 방법은 이후 개선된 인셉션 모델에도 그대로 유지됐다. 그리고 2016년 인셉션 모델의 4번째 버전 ‘Inception-v4’와 함께 인셉션 모델에 ‘ResNet’을 접목한 ‘Inception-ResNet’을 발표했다 [16].

‘ResNet’은 2015년 마이크로소프트에서 발표한 논문 [18]에서 소개됐으며, [그림 2-4]와 같이 망이 적정 깊이 이상의 깊어지면 오류 발생 많아지는 문제를 해결하는 방법을 제시하고 있다.

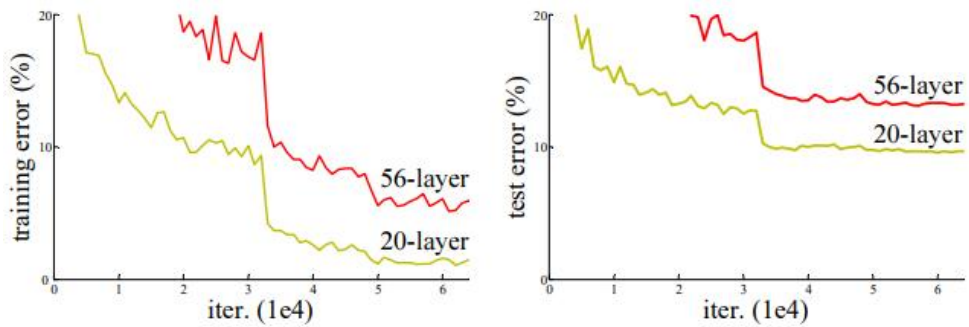


그림 2-4 망 깊이에 따른 오류

일반적인 합성곱 신경망은 [그림 2-5]의 왼쪽과 같이  $x$ 를 입력받아 은닉층(hidden layer)을 거쳐  $H(x)$ 을 출력하며, 학습을 통해 최적의  $H(x)$ 을 얻게 된다. 이러한 망을  $H(x) - x$ 를 얻는 것으로 수정하고  $F(x) = H(x) - x$ 라면,  $H(x) = F(x) + x$ 가 된다. 그러면 [그림 2-5]의 오른쪽과 같은 망으로 표현할 수 있으며, 이는 'ResNet'의 기본망이 된다. 'ResNet'은 논문에서 소개한 것과 같이 오차를 학습하는 것으로 볼 수 있지만, 과거의 특징점이 소실되지 않도록 하는 순환 신경망과 유사한 의도로도 접근할 수 있다.

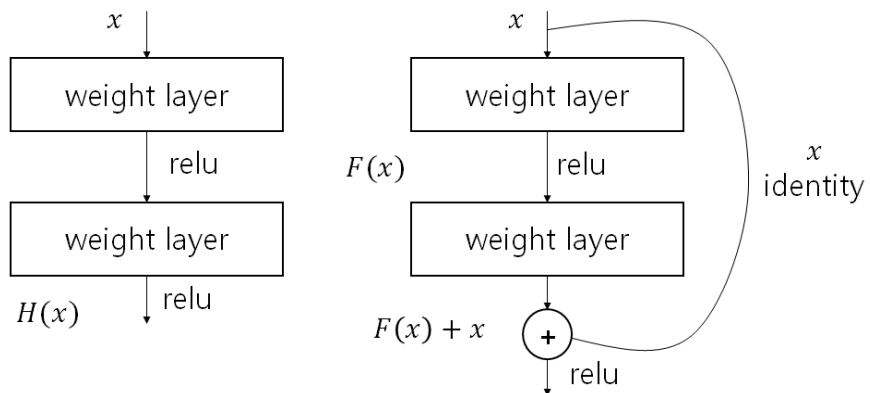


그림 2-5 Resnet 구성 예

[16]에서는 ‘ResNet’의 형태를 [그림 2-6]와 같이 inception 모듈에 접목하여 ‘Inception-ResNet’을 만들었으며, 이는 같은 논문에 소개된 ‘Inception-v4’ 모델과 비교했을 때, 최종 정확도에서는 거의 차이가 없지만, 훨씬 빠른 속도로 수렴하는 것을 보여줬다.

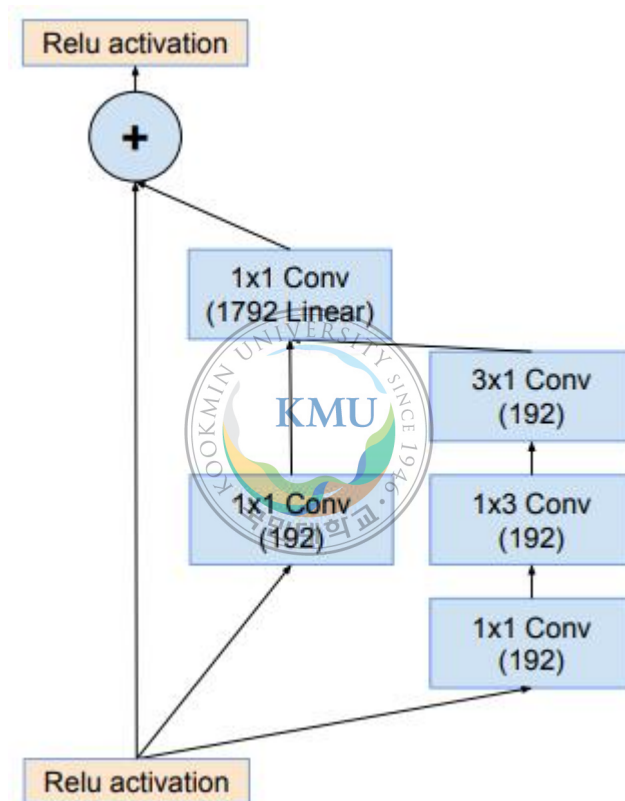


그림 2-6 Inception-ResNet 모듈의 예

## 2.3 유전 알고리즘 (Genetic Algorithm)

유전 알고리즘(Genetic Algorithm)은 1965년 Rechenberg에 의해 제안된 진화 전략(evolution strategy)에서 시작되어 1975년 John Holland

의 저서인 "Adaptation on Natural and Artificial Systems"에서 현재와 같은 형태가 완성되었다. 유전 알고리즘은 다윈의 자연선택 이론을 토대로 한 알고리즘으로, 확률적으로 해 공간을 탐색하여 복잡한 문제의 근사해를 제공하는 메타휴리스틱(metaheuristic)기법이다 [19]. 다윈의 자연선택 이론에서는 특정 환경에서 생존에 적합한 형질을 지닌 종은 그렇지 않은 종보다 생존에 더 유리하게 되며, 유전 알고리즘에서는 이것을 선택(selection), 교차(crossover), 변이(mutation)의 과정으로 표현하고 있다.

고전적인 알고리즘으로 최적해 문제를 해결하는 경우 지역적 최적해(local optimum)에 빠지는 문제가 발생할 수 있지만[그림 2-7], 유전 알고리즘에서는 이 문제를 해 공간을 확률적으로 탐색하는 방법으로 해결한다. 유전 알고리즘은 최적해를 계산하는 중 지역적 최적해에 빠지는 문제가 발생하더라도 적합도 계산 값이 원하는 결과에 못 미친다면 일부 해를 공간 내에서 확률적으로 재생산하기 때문에 지역적 최적해를 벗어나 전역적 최적해(global optimum)를 발견할 가능성을 높일 수 있다. 그리고 알고리즘을 이용하여 최적해를 구하는 방법은 도메인에 대한 깊은 이해를 바탕으로 한 수학적 해석이 필요한 반면, 유전 알고리즘은 수학적 제약이 거의 없어 여러 분야에 적용이 용이하다는 장점이 있다 [20].

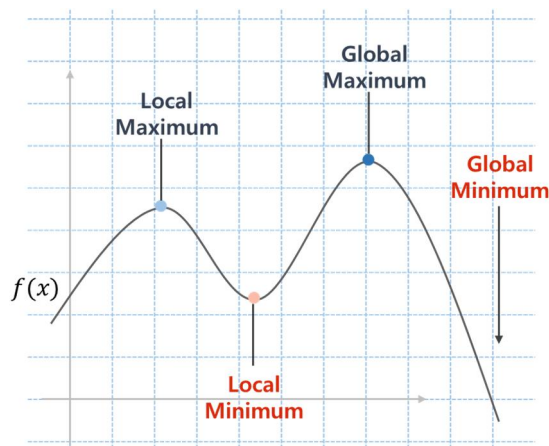


그림 2-7 지역적 최적해와 전역적 최적해

유전 알고리즘은 ‘1. 무작위로 초기집단 생성, 2. 집단 내 유전자 선택, 3. 선택된 유전자 내 교차, 4. 교차 된 유전자 내 변이, 5. 완성된 자손 집단 내 유전자들의 적합도 계산’ 순으로 진행되며, 적합도 계산 결과에 따라 2 ~ 5과정 반복하거나 알고리즘을 종료하게 된다[그림 2-8].

본 논문에서는 투자 포트폴리오를 구성 및 최적화 실험에 유전 알고리즘을 사용되며, 4장에서 실험을 위해 구현한 유전 알고리즘과 적합도 함수 설계 방법에 대해 자세히 설명한다.



그림 2-8 유전 알고리즘 과정



## 제 3장 딥러닝을 이용한 주가 예측

본 장에서는 순환 신경망 대신 합성곱 신경망을 사용하게 된 배경과 데이터 전처리 방법 그리고 Inception-Resnet-v2 모델을 기반으로 만든 주가 예측 모델에 대해 설명하고 실험 방법을 소개한다.

### 3.1 데이터 전처리

#### 3.1.1 주가 정보 분석

일반적으로 주가는 미래의 주가가 지금까지의 가격에 따라 결정되거나 영향을 받는 시계열 데이터로 보고 있다. [그림 3-1]은 모 기업의 주가를 차트로 나타낸 것으로, 이를 보면 주가는 전체적으로 일정한 경향성을 가지고 움직이는 것처럼 보인다. 하지만 [그림 3-1]은 해당 기업의 1년간 차트로, [그림 3-2]와 같이 짧은 기간의 그래프를 보면 일정한 경향성을 가진 움직임으로 보기 매우 어렵다. 이는 버튼 맬킬의 랜덤워크 이론에서 잘 설명되어 있다 [21].

랜덤워크 이론은 시장이 효율적이고 합리적이라는 가정을 토대로 한다. 랜덤워크 이론에 따르면 모든 정보는 공개되는 시점에 투자 주체들에 의해 즉각 반영되며 주체 성향에 따라 다른 판단을 하기 때문에 불규칙한 움직임을 보인다. 따라서 주식시장의 움직임은 과거 주가의 분석이나 지표를 통한 예측이 불가능하다는 것이 이 이론의 주장이다. 이 이론을 그대로 따른다면 주가 예측 연구가 무의미하게 되지만, 주가 정보를 이용하는데 있어서는 고려할만한 가치가 충분한 이론임에는 틀림없다.

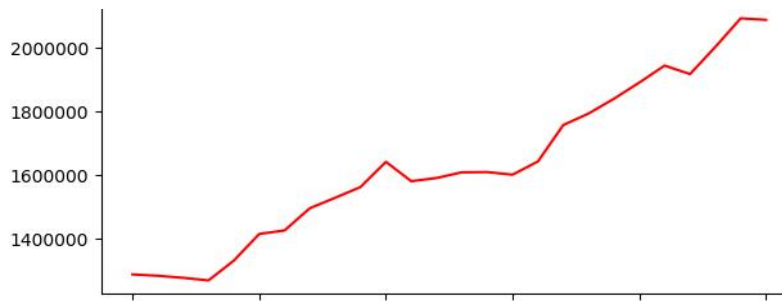


그림 3-1 1년간 주가 그래프

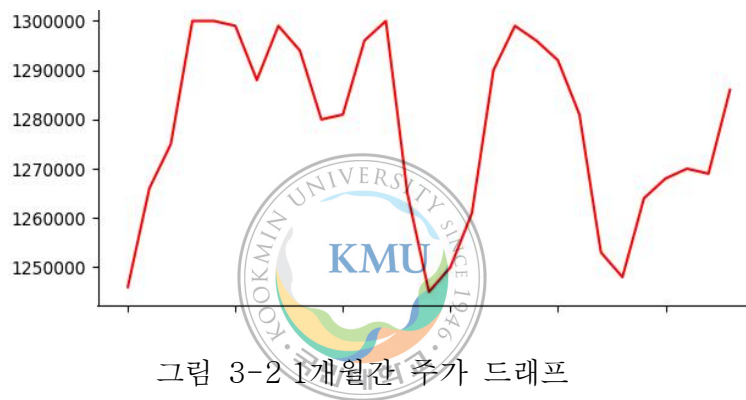


그림 3-2 1개월간 주가 그래프

본 논문에서는 랜덤워크 이론을 고려하여 주가 정보를 이미 현재까지 모든 정보를 반영하고 있는 독립된 데이터로 간주하며, 개별적인 특성을 효과적으로 학습하기 위해 순환 신경망이 아닌 합성곱 신경망을 이용한 모델을 설계한다.

### 3.1.2 데이터 선정 및 수집 방법

주가 방향성 예측 실험에 사용되는 데이터는 주가 정보와 함께 인터넷 기사 제목 두 가지를 사용한다. 기사 제목은 데이터의 다양성을 확보하고 편향된 학습과 과적합 문제를 해결하기 위한 데이터로 사용된다.

주가 정보는 2018년 7월 3일 기준 코스피 200의 회사 중 2004년 6월부터 주가 정보를 수집할 수 있는 종목 100개를 선정하고 2004년 6월 7일부터 2017년 4월 14일까지 데이터 3194개를 학습 및 검증 데이터로, 2017년 4월 21일부터 2018년 7월 3일까지 데이터 290개를 평가 데이터로 사용한다[오투! 참조 원본을 찾을 수 없습니다.].

인터넷 기사 제목은 주가에 영향을 줄 수 있는 양질의 기사 제목을 수집하기 위해 대형 신문사의 경제 기사만을 데이터로 사용하며, 주가 정보와 동일한 기간의 데이터를 학습 및 평가 데이터로 사용한다.

데이터는 Python으로 만든 크롤러를 이용하여 주가 정보는 네이버 금융에서 제공하고 있는 데이터를, 기사 제목은 중앙일보 경제 기사 제목을 수집한다.

표 3-1 수집 데이터 및 학습데이터 개수

|             | 학습 데이터        |             | 평가 데이터 |
|-------------|---------------|-------------|--------|
|             | 학습 데이터        | 검증 데이터      |        |
| 총 사용 데이터 수  | 3194          |             | 290    |
| 전처리 후 데이터 수 | 2200 (69.99%) | 994(30.01%) | 239    |

### 3.1.3 데이터 전처리 방법

#### a) 주가 정보 전처리 방법

크롤러를 이용하여 수집한 주가 정보는 종목별로 ‘날짜’, ‘시가’, ‘고가’, ‘저가’, ‘종가’, ‘거래량’의 정보를 [표 3-2]와 같은 형태의 ‘.csv’파일로 관리된다. 수집된 데이터는 주가 정보의 개별적인 특성을 부각하기 위해 수집한 정보를 그대로 사용하지 않고 기존 주식 투자에 사용되

던 보조지표로 전처리한 후 사용한다. 보조지표는 보는 관점에 따라 분류방법이 다양하지만, 본 논문에서는 ‘추세지표’, ‘시장강도지표’, ‘모멘텀지표’, ‘변동성지표’ 4종류의 지표[

표 3-3]에서 각각 2개의 보조지표를 선택하여 총 8개의 보조지표[표 3-4]를 모델의 데이터로 사용한다.

표 3-2 수집데이터 샘플

|          |       |       |       |       |        |
|----------|-------|-------|-------|-------|--------|
| 20040607 | 41000 | 41100 | 39300 | 40300 | 83237  |
| 20040608 | 40500 | 41850 | 40150 | 41000 | 133936 |
| 20040609 | 41300 | 41800 | 40300 | 40600 | 61649  |
| 20040610 | 40600 | 43800 | 40300 | 42000 | 193099 |
| 20040611 | 42600 | 42600 | 40400 | 40700 | 58240  |

표 3-3 특성에 따른 지표의 분류

| 분 류    | 특 징  |
|--------|--|
| 추세지표   | 주가의 진행방향을 분석하기 위한 지표. 주가는 과거와 유사한 방향으로 진행될 것을 가정하고 분석에 사용.<br>Ex) 이동평균선, MACD, DMI 등 |
| 시장강도지표 | 거래량을 이용하여 주가의 추세 및 변동성을 분석하는 지표.<br>Ex) 거래량 이동평균, Volume Ratio, MFI 등                |
| 모멘텀지표  | 투자 심리 등을 이용하여 추세의 변동 시점을 분석하는 지표.<br>Ex) 투자심리선, RSI, 이격도 등                           |
| 변동성지표  | 주가의 변동성을 분석하기 위한 지표. 단기 매매에 많이 사용.<br>Ex) ATR, envelope, 볼린저밴드 등                     |

표 3-4 사용된 지표 및 공식

| 분 류        | 지 표      | 공 식  |
|------------|----------|--|
| 추세지표       | MACD     | $\sum_{i=1}^{12} \alpha^i C_{t-i} - \sum_{i=1}^{26} \alpha^i C_{t-i}$  |
|            | DMI      | $\frac{abs\left(\left(\sum_{i=0}^{n-1}(+DM) - \sum_{i=0}^{n-1}(-DM)\right) / \sum_{i=0}^{n-1} TR_i\right)}{\left(\sum_{i=0}^{n-1}(+DM) + \sum_{i=0}^{n-1}(-DM)\right) / \sum_{i=0}^{n-1} TR_i} \times 100$ <p> <math>, +DM = (H_t - H_{t-1}) \text{ if } ((H_t - H_{t-1}) \wedge (H_t - H_{t-1} &gt; L_{t-1} - L_t)) \text{ else } 0)</math><br/> <math>, -DM = (L_{t-1} - L_t) \text{ if } ((L_{t-1} - L_t) \wedge (H_t - H_{t-1} &lt; L_{t-1} - L_t)) \text{ else } 0)</math><br/> <math>, TR = \max(abs(H_t - L_t), abs(C_{t-1} - H_t), abs(C_{t-1} - L_t))</math> </p> |
| 시장강도<br>지표 | VR       | $\frac{\sum_{i=0}^{n-1} UP_i + (\sum_{i=0}^{n-1} None_i \times 0.5)}{\sum_{i=0}^{n-1} DW_i + (\sum_{i=0}^{n-1} None_i \times 0.5)} \times 100$   |
|            | VMA      | $\sum_{i=0}^{n-1} V_{t-i} / n$   |
| 모멘텀<br>지표  | 투자심리선    | $ A  \times 10, A = \{x   x = C_i - C_{i-1}, 0 < i \leq 10\}$  |
|            | RSI      | $100 - \left( \frac{100}{1 + \left( \frac{\sum_{i=0}^{n-1} UP_{t-i} / n}{\sum_{i=0}^{n-1} DW_{t-i} / n} \right)} \right)$  |
| 변동성<br>지표  | ATR      | $\sum_{i=1}^{10} TR_i / 10$  |
|            | envelope | $\sum_{i=1}^{10} C_{t-i} / 10 \times 1.1$  |

전처리하는 지표별로 1일, 3일, 5일, 20일, 25일 전을 기준으로 지표를 계산한다. 지표 계산에 사용되는 주가 정보는 지표별 권장되는 기간으로 계산되며, 하루의 주가 방향성을 예측하기 위해 총 51개의 주가 정보가 사용된다.

b) 기사 제목 전처리 방법

기사 제목은 학습 시 결과가 한쪽으로 편향되는 문제와 학습 데이터에 과적합 문제를 해결하고 데이터의 다양성을 주기 위해 사용했다. 즉, 본 논문에서 기사 제목은 부가적인 학습 정보로 사용되며, 수집된 기사 제목은 오픈소스 ‘gensim’을 이용하여 문장벡터(Doc2Vec)로 전처리한 후 학습에 사용한다.

하루의 주가 방향성을 예측하기 위해 6일간 기사 제목 정보가 하루에 3개씩 총 18개가 사용되며 하루에 기사가 3개 미만인 경우 부족한 만큼 0벡터를 사용한다. 하루에 기사가 4개 이상인 경우 [알고리즘 3-1]의 방법으로 3개의 벡터를 골라 사용한다.

[알고리즘 3-1]의 경우 생성된 문장벡터가 단어벡터(word2vec)와 같이 의미에 따라 방향성을 가지고 있다는 전제로 진행되며, 다양한 의미를 학습에 사용하기 위해 서로 방향성이 가장 다른 벡터를 3개 선별한다.

---

Algorithm 1    Pseudocode for Select Doc2Vec

---

```
1: selectedVecList = [random.choice(docVecList)]
2: docVecList.remove(selectedVecList[-1])
3: while len(selectedVecList) < VecCnt :
4:   minSimilarity = 10
5:   for vec in docVecList :
6:     sumSimilarity = sum([cosSimilarity(vec, l) for l in selectedVecList])
7:     if sumSimilarity < minSimilarity :
8:       minSimilarity = sumSimilarity
9:       temp = vec
10:  selectedVecList.append(temp)
11:  docVecList.remove(temp)
```

---

알고리즘 3-1 문장벡터 선택 알고리즘

c) 목표(target) 데이터 전처리 방법

주식 거래 시 수익 여부에 상관없이 약간의 손실 비용이 발생한다.

1. 주식 거래 수수료 약 0.015% (증권사마다 상이)
2. 주식 매도 시 세금 0.3%

최종 매도 후 수익이 발생하기 위해선 위 손실 비용 이상의 수익을 얻어야 하며, [수식 3-1]을 통해  $\alpha$ 의 투자 비용에서 이익 얻기 위한 최소 주가 상승률  $x$ 는 약 1.0033배, 즉 0.34% 이상 주가가 상승해야 수익이 발생한다는 것을 알 수 있다.

$$\begin{aligned}\alpha(x-1) & \quad (1) \\ 0.00015\alpha(x+1) & \quad (2) \\ 0.003\alpha x & \quad (3) \\ \alpha(x-1) - \alpha(0.00015(x+1)) + 0.003\alpha x & > 0 \quad (1)-(2)-(3) \\ 0.99685x & > 1.00015 \\ x & > 1.00015/0.99685 \approx 1.0033\end{aligned}$$

수식 3-1 수익을 위한 최소 주가 상승률

본 실험의 목표 데이터는 주가 등락에 따라 이진 분류하며, 예측일의 시가와 종가의 등락률이 수익 발생의 최소 상승률인 0.34%보다 조금 큰 0.35%를 분류 기준으로 한다. 0.35%보다 작은 경우는 '[1, 0]', 그렇지 않은 경우를 '[0, 1]'으로 구성한다. [0.0%, 0.35%)의 등락률은 상승장이라 하더라도 수익을 낼 수 없기 때문에 하락장으로 간주한다.

### 3.2 방향성 예측 모델

주가 방향성 예측에 사용된 모델은 Inception-Resnet-v2 모델을 기반으로 만든 합성곱 모델로, 기존 모델의 컨셉은 유지하면서 입력 데이터의 특성과 크기 그리고 출력 데이터의 크기 등을 고려하여 각 모듈이 재조정되었다. 모델의 전체 구조는 [그림 3-3]와 같다.

[그림 3-3]와 같이 본 실험에서는 주가 정보와 기사 제목 각각의 데이터의 특징을 고려한 학습을 할 수 있도록, 각각의 데이터가 [그림 3-4]와 [그림 3-5]에 서로 다른 모델에 입력 데이터로 따로 들어간다. 그리고 각각의 결과를 연결(concatenate)한 후 전방 전달 신경망(Feedforward Neural Network)을 이용한 분류 단계를 통해 주가 방향성을 최종 예측한다.

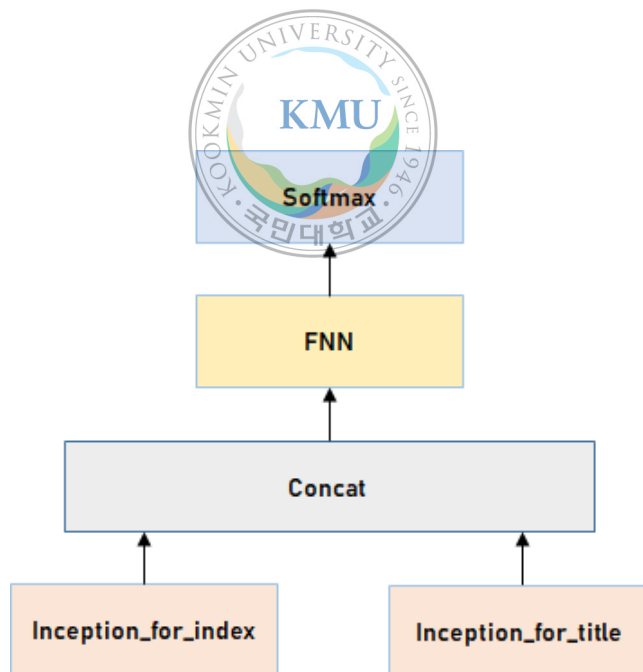


그림 3-3 주가 방향성 예측 모델



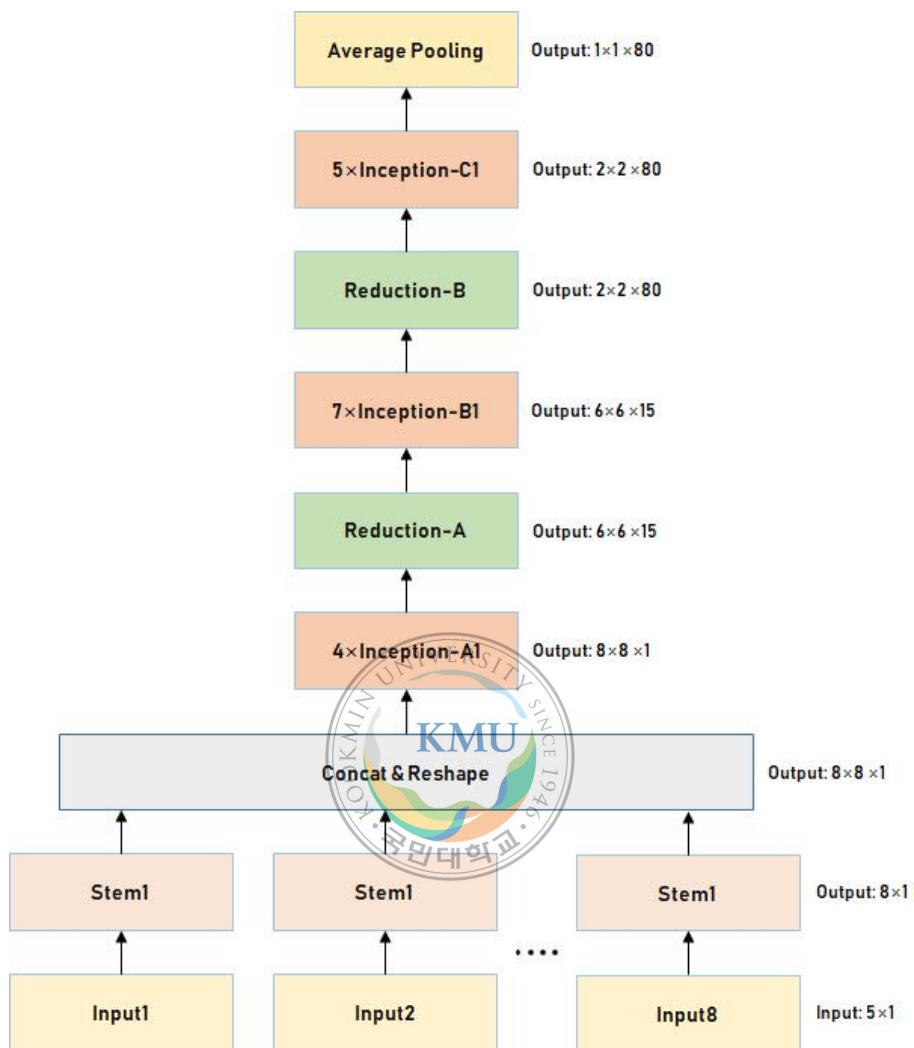


그림 3-4 주가 정보 학습을 위한 인셉션 모델

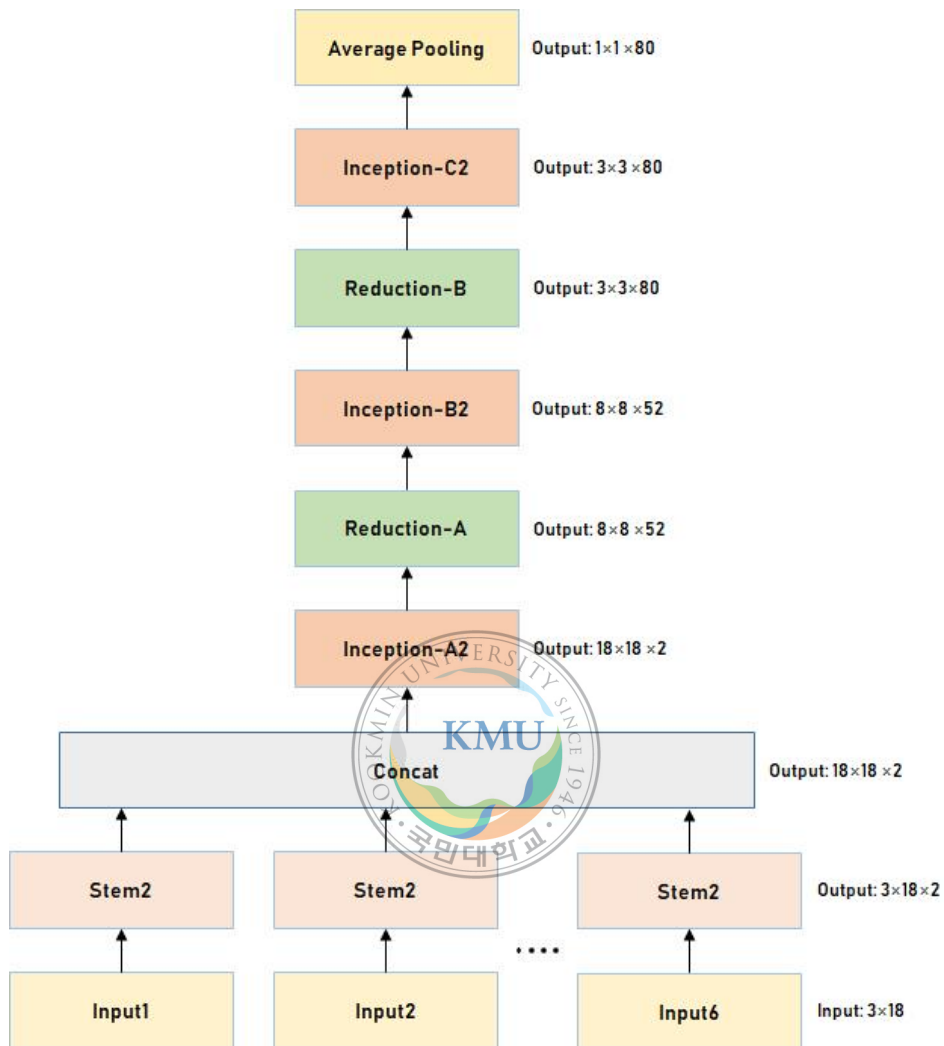


그림 3-5 기사 제목 학습을 위한 인셉션 모델

#### a) Stem

본래 인셉션 모델은 이미지 분류를 위한 모델로 매우 큰 데이터를 입력받는다. 그래서 데이터의 크기를 줄여 연산의 부하를 줄이고 특징점 (feature map)을 추출하기 위한 ‘Stem’ 단계가 존재한다.

하지만 주가 데이터의 경우 이미지와 비교하면 매우 작은 크기의 데이터가 학습에 사용되며, 데이터의 각 값이 이미 특징 있는 주기 지표로 구성되어 있다. 그렇기 때문에 본 논문에서는 데이터의 크기를 줄이고 특징점을 만들기 위한 본래의 ‘Stem’ 단계 대신 주어진 특징점에서 의미를 확장하는 의미의 ‘Stem’ 단계를 전달 신경망으로 구성하여 사용한다.

#### b) Inception

초기 인셉션 모델에서는 ‘ $3 \times 3$ ’ 크기의 합성곱 레이어와 함께 ‘ $5 \times 5$ ’와 ‘ $7 \times 7$ ’ 크기의 합성곱 레이어를 사용했지만, ‘Inception-v2’부터는 과도한 연산량으로 인한 부하를 해결하기 위해 모든 합성곱 레이어의 필터 크기를 ‘ $3 \times 3$ ’을 넘지 않게 됐다. 그러나 큰 크기의 이미지 데이터가 입력 데이터로 사용되는 인셉션 모델과 달리 본 실험의 입력 데이터로 사용되는 주가 정보는 훨씬 적은 크기이므로 ‘ $3 \times 3$ ’보다 큰 크기의 합성곱 레이어를 사용하는 것이 큰 부담이 되지 않는다. 그리고 각 정보 간의 상관관계를 보다 직접적으로 확인하기 위해 ‘ $1 \times 7$ ’부터 ‘ $5 \times 5$ ’까지 다양한 크기의 필터를 사용한다.

#### c) Reduction

Inception 단계에서는 특징점을 추출하지만 pooling을 이용한 대표점 추출과 크기 변화가 없다. 인셉션 모델에서는 이 역할을 Reduction 단계에서 수행한다. Reduction에서도 Inception 단계와 같이 ‘ $1 \times 1$ ’ 크기의 합성곱 레이어를 사용하며, max pooling과 합성곱 레이어의 stride를 2로 설정하여 크기를 줄인다. 그리고 주가 정보와 기사 제목 학습에 서로 다른 Inception 단계가 사용되는 반면, Reduction은 두 학습에 같은 모듈을 사용한다.

### 3.3 기사 제목 학습 모델

기사 제목 학습을 위한 모델도 기본적으로 주가 정보 학습을 위한 모델과 같은 구조의 모델을 사용한다. 두 모델은 Inception 모듈의 구조에서 약간 차이가 있지만, 모듈을 구성하는 기본 개념은 동일하므로 추가적인 설명은 이번 장에선 생략한다.

#### a) Stem

본 실험에서 사용되는 기사 제목은 ‘gensim’을 이용해서 만든 문장벡터로 만들어 사용하며, 하루 예측을 위해 하루에 3개씩 6일간 총 18개의 문장벡터를 사용한다. 기사제목 학습을 위한 ‘Stem’ 단계에서는 3개의 문장벡터의 특징점을 뽑아내는 단계로 사용되며, 일별로 ‘Stem’ 단계를 적용한다.



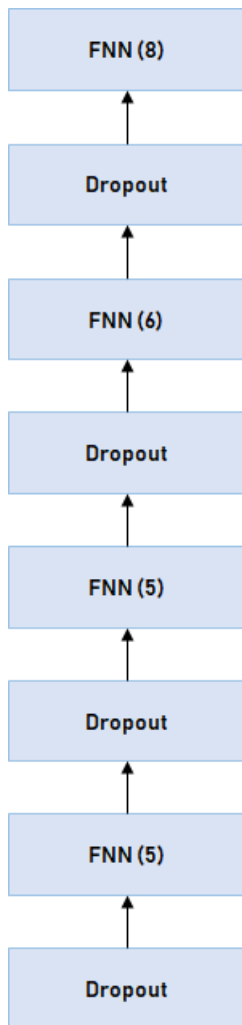


그림 3-6 Stem1

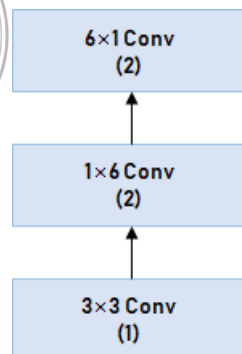


그림 3-7 Stem2

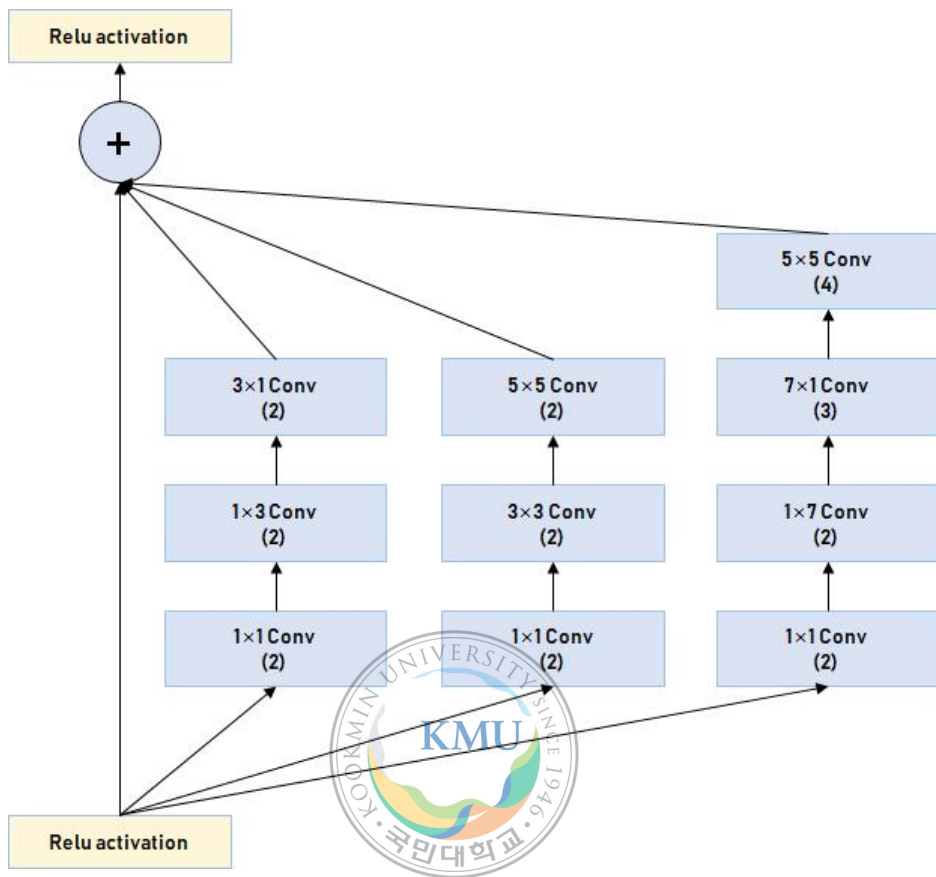


그림 3-8 Inception-A1

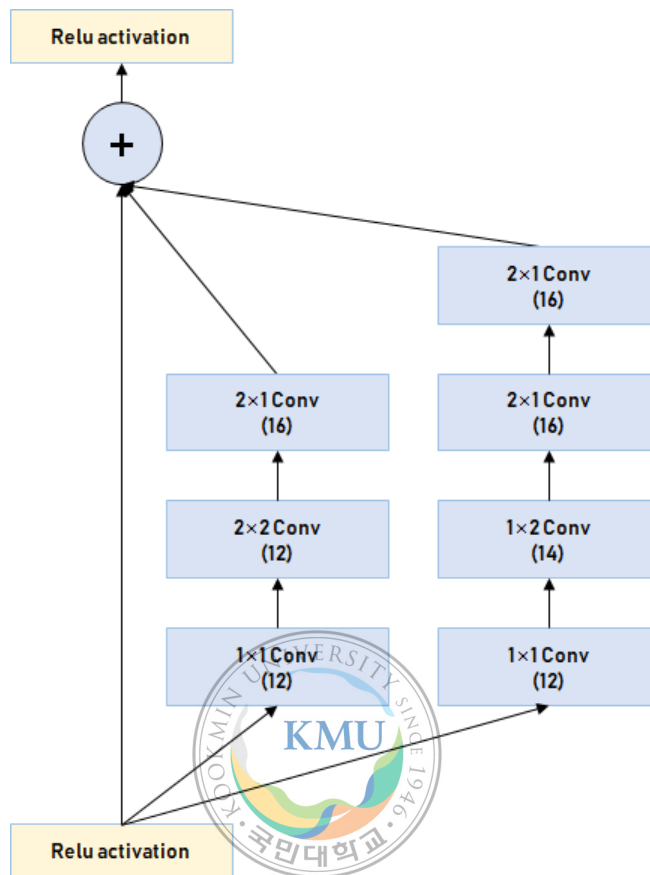


그림 3-9 Inception-B1

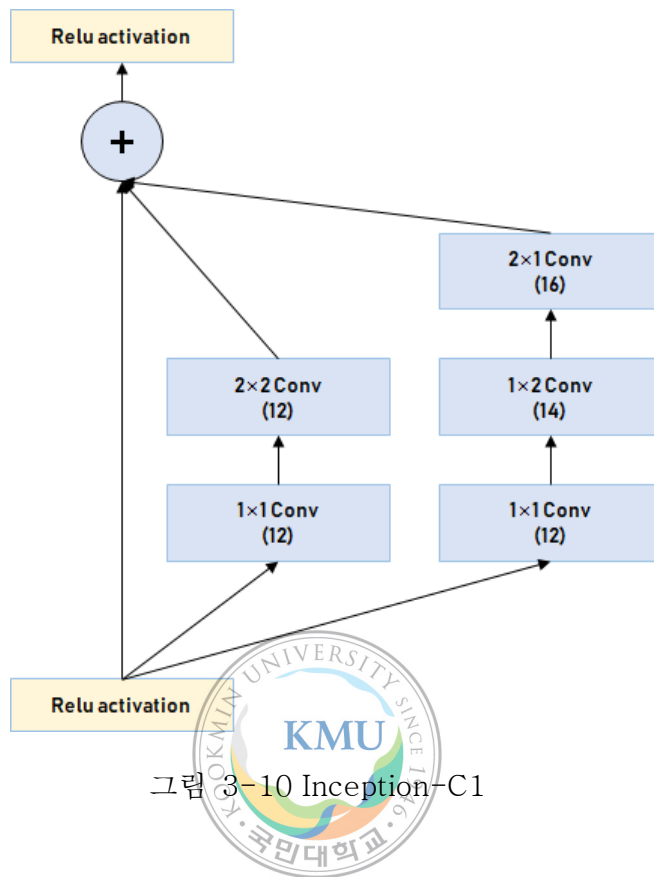


그림 3-10 Inception-C1



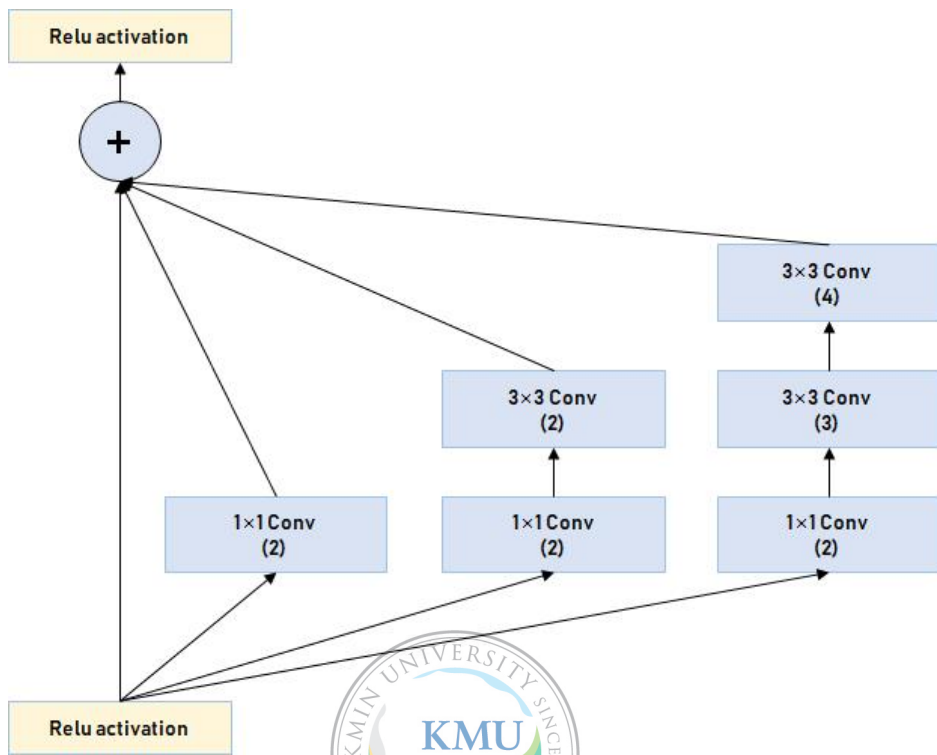


그림 3-11 Inception-A2

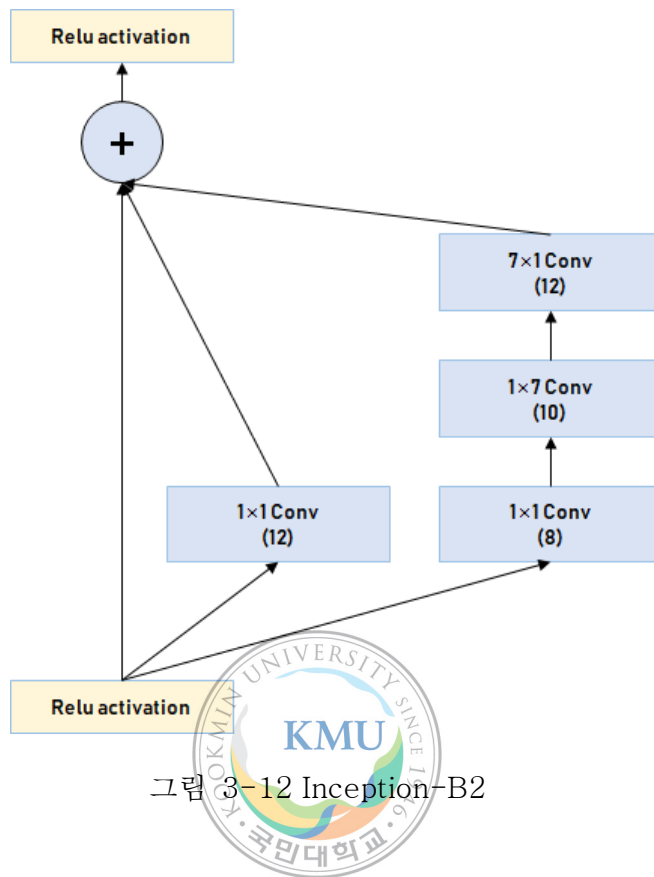


그림 3-12 Inception-B2

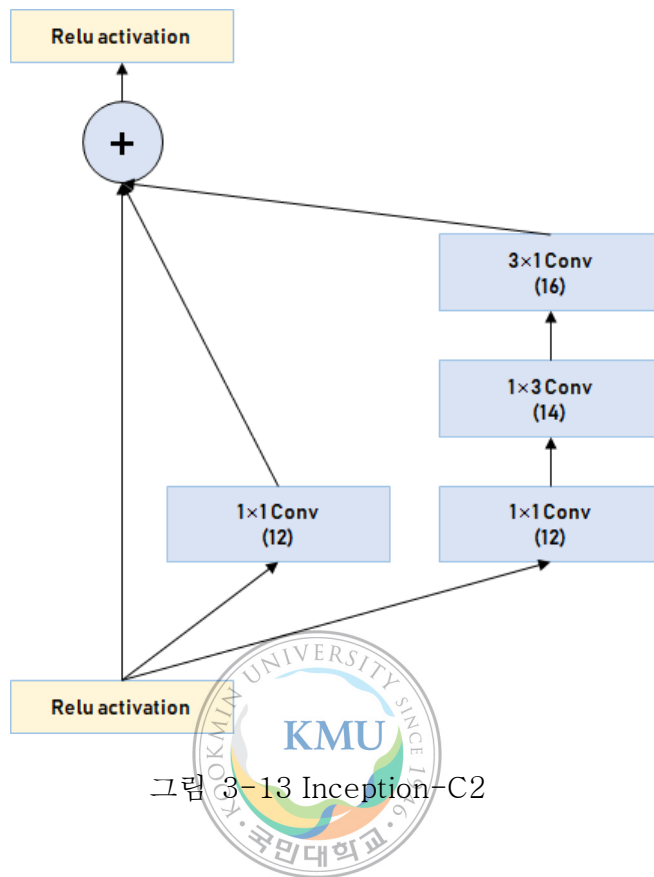


그림 3-13 Inception-C2

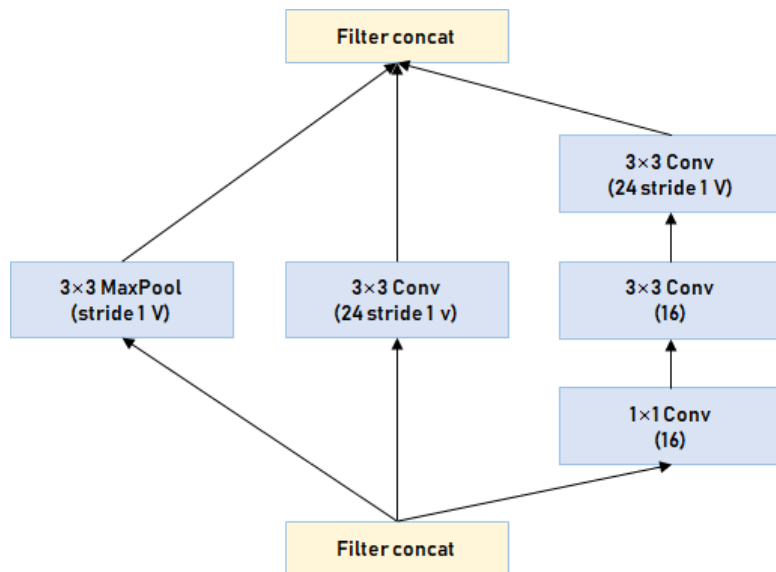


그림 3-14 Reduction-A

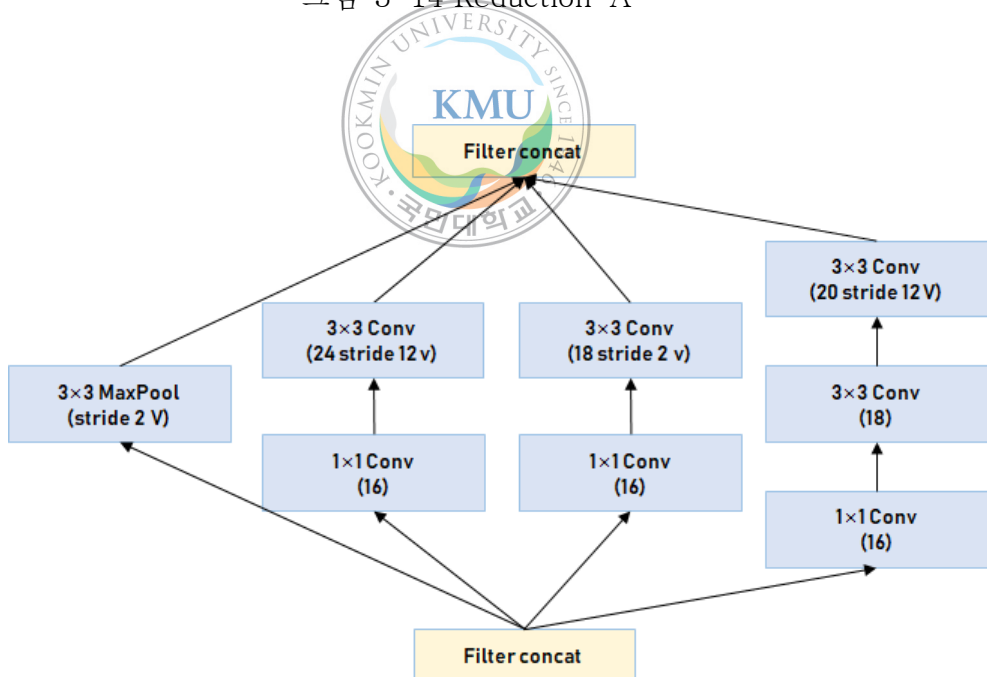


그림 3-15 Reduction-B

### 3.4 실험

모델은 Python과 Tensorflow를 이용하여 구현했다.

Loss 측정에는 cross-entropy를 optimizer에는 Adam optimizer를 사용한다. 실험에 사용된 조정값(hyperparameter)은 optimizer의 learning rate 0.002, 전담 신경망에 사용된 dropout의 keep probability 0.7로 설정한다. learning rate는 빠른 학습을 위해 3 epoch 이상 정확도 변화가 없는 겨우 반씩 내려가며 학습을 진행했다.

모델은 과적합을 예방하기 위해 학습 데이터의 정확도와 검증 데이터의 정확도가 10% 이상 차이나지 않으면서 가장 높은 정확도를 보인 모델을 저장했다.

학습은 종목별로 따로 진행했으며, 실험은 CentOS 7 운영체제에서 GeForce GTX 1070을 이용한 환경에서 진행한다.



## 제 4장 유전 알고리즘을 이용한 포트폴리오 최적화

본 장에서는 유전 알고리즘을 이용한 투자 포트폴리오를 최적화 실험을 소개한다. 투자 포트폴리오 최적화를 위한 실험은 ‘투자 종목 최적화’와 ‘투자 비율 최적화’를 두 단계로 나눠 최적화하는 실험과 두 단계를 함께 최적화하는 실험 2가지를 진행한다.

### 4.1 투자 종목 최적화

#### 4.1.1 적합도 함수

투자 종목 최적화에서는 최대 수익을 올리는 것이 목적이지만 그와 함께 고려되어야 하는 것이 투자 안정성이다. 따라서 투자의 수익성과 안정성을 모두 고려할 수 있도록 적합도 함수를 만드는 것이 중요하다. 본 실험에서 사용되는 적합도 함수는 [수식 4-1]의 (2) 수식이며, (1) 수식은 안정성을 평가하기 위한 수식이다. 이 수식은 포트폴리오 구성 종목 중 실제 수익을 낸 종목의 평균 개수(*correctCount*)와 전체 투자 중 실제 수익이 발생한 날의 비율(*successProbability* \* 10)의 합을 사용하며, 수익성을 평가는 기간 중 평균 수익률(*averageYield*)을 사용한다.

$$correctCount + successProbability * 10 \quad (1)$$

$$averageYield + (correctCount + successProbability * 10) \quad (2)$$

수식 4-1 투자 종목 최적화를 위한 적합도 계산 방법

#### 4.1.2 실험 방법

투자 종목 선정은 예측 결과를 바탕으로 점수를 매긴 후 높은 점수를 종목 6개를 선정한다. 그리고 전체 투자 종목 중 상승장으로 예측되는 종

목의 수가 15개보다 작은 경우 투자가 부적합한 위험 장이라 판단하고 투자하지 않는다.

각 종목의 점수는 주가 방향성 예측 실험의 결과로 얻는 ‘정확도(accuracy)’와 ‘재현율(Recall)’, ‘정밀도(Precision)’, ‘예측 확률’의 합으로 결정되며, 각 값의 반영비율을 최적화 하는 방법으로 유전 알고리즘 사용한다.

따라서 한 염색체(chromosome)의 유전자(gene)는 4개로 구성되며, 각 유전자는 순서대로 ‘정확도 반영비율’, ‘재현율 반영비율’, ‘정밀도 반영비율’, ‘예측 확률 반영비율’을 표현한다.

단계별로 진행되는 실험에서는 한 세대(population)를 120개의 염색체로 구성하고 적합도 함수가 최대가 되도록 학습한다. 유전자 선택 방법에는 ‘우량 유전자 우선 선별 방법(Elitist)’을 사용하고, 교차 방법은 ‘한 점 교차(one-point crossover)’를 변이에는 임의값 변이를 사용한다.

## 4.2 투자 비율 최적화

### 4.2.1 적합도 함수

위의 실험에서는 안정성을 고려하기 위해 [수식 4-1]의 (1)수식을 이용하여 안정성을 고려했다. 하지만 이번 실험에서는 이미 선정된 종목의 투자 비율을 최적화하는 것으로 손실 가능성이 있는 종목을 완전 배제하는 것은 무리가 있다. 따라서 이번 실험 적합도 함수는 [수식 4-2]와 같이 (*averageYield*)을 통해 높은 수익성을 낼 수 있도록 유도함과 동시에 손해가 발생된 투자 비율(*lossInvestmentRate*)을 통해 위험성이 있는 종목의 투자 비율을 최소화 하도록 유도한다.

$$averageYield/lossInvestmentRate$$

수식 4-2 투자 비율 최적화를 위한 적합도 계산 방법

#### 4.2.2 실험 방법

위의 결과로 선택된 6개의 종목은 첫 번째 실험에서 얻은 점수가 높은 종목이 수익성 및 안정성이 높을 것이란 가정하에 실험을 진행한다. 종목들은 첫 번째 실험에서 얻은 점수를 기준으로 오름차순으로 정렬하여 가장 왼쪽에 1순위 종목이 오도록 한다.

한 염색체의 유전자는 6개로 구성되며, 각 유전자는 순서대로 ‘1순위’ 종목 투자비율’부터 ‘6순위 종목 투자비율’까지 순서대로 표현한다.

단계별로 진행되는 실험에서는 한 세대를 120개의 염색체로 구성하고 적합도 함수가 최대가 되도록 학습한다. 유전자 선택 방법에는 ‘우량 유전자 우선 선별 방법(Elitist)’을 사용하고, 교차 방법은 ‘한 점 교차(one-point crossover)’를 변이에는 임의값 변이를 사용한다.

### 4.3 투자 종목 최적화 + 투자 비율 최적화

#### 4.3.1 적합도 함수

‘투자 종목 최적화’와 ‘투자 비율 최적화’를 한 단계로 진행하는 실험에서는 (4.2.1)의 적합도 함수를 그대로 사용한다.

#### 4.3.2 실험방법

염색체는 (4.1.2)와 (4.2.2)에서 구성했던 염색체를 그대로 가져와 연결하여 사용하며, 따라서 염색체의 구성은 1~6순위 종목의 투자 비율 유전



자 6개, ‘정확도 반영비율’, ‘재현율 반영비율’, ‘정밀도 반영비율’, ‘예측 확률 반영비율’을 최적화하기 위한 유전자 4개를 포함 총 10개의 유전자로 염색체가 구성된다.

이 실험에서는 한 세대를 120개의 염색체로 구성하고 적합도 함수가 최대가 되도록 학습한다. 유전자 선택 방법은 ‘우량 유전자 우선 선별 방법(Elitist)’을 사용하고, 교차 방법은 ‘다 점 교차(multi-point crossover)’, 변이에는 임의값 변이를 사용한다.



## 제 5장 결과 및 성능평가

학습에 사용된 100개의 종목 중 5개 종목에서 예측이 한쪽으로 편향되는 문제가 발생했다. 이는 동일 모델에서 주가 정보만을 사용했을 때 동일한 문제가 다수 발생했던 것과 비교하면 크게 줄어든 것으로, 다양한 데이터가 다량 존재할수록 편향성 문제가 줄어든다고 생각해 볼 수 있다.

[표 5-1]과 [표 5-2]는 3장에서 진행한 주가 방향성 예측 실험의 결과로, 편향성 문제가 발생한 4개의 종목을 제외한 나머지 종목 중 정확도 상위 15개와 하위 15개 대한 예측률이며, 전체 학습 데이터에 대한 예측률의 약 63.2%가 나왔다. 대부분의 종목이 학습이 반복될수록 학습 데이터의 정확도는 올라갔지만 3장에서 언급한 조건에 충족하는 모델을 저장한 결과, 63%정도의 정확도에 머물렀다.

표 5-1 학습 데이터 상위 15개 종목 정확도

| 종목명    | 예측성공수 | 학습데이터수 | 정확도    |
|--------|-------|--------|--------|
| 한화     | 1549  | 2200   | 70.41% |
| KT     | 1548  | 2200   | 70.36% |
| 농심     | 1534  | 2200   | 69.73% |
| LS     | 1502  | 2200   | 68.27% |
| 삼성증권   | 1497  | 2200   | 68.05% |
| 한국전력   | 1494  | 2200   | 67.91% |
| 대우건설   | 1490  | 2200   | 67.73% |
| LG 전자  | 1488  | 2200   | 67.64% |
| 한국가스공사 | 1486  | 2200   | 67.55% |
| 대한항공   | 1485  | 2200   | 67.50% |
| 동서     | 1476  | 2200   | 67.09% |
| 신세계    | 1471  | 2200   | 66.86% |
| 신한지주   | 1469  | 2200   | 66.77% |
| 현대제철   | 1465  | 2200   | 66.59% |
| 현대모비스  | 1462  | 2200   | 66.45% |

표 5-2 학습 데이터 하위 15개 종목 정확도

| 종목명        | 예측성공수 | 학습데이터수 | 정확도    |
|------------|-------|--------|--------|
| DB 손해보험    | 1150  | 2200   | 52.27% |
| 롯데케미칼      | 1222  | 2200   | 55.55% |
| 포스코대우      | 1226  | 2200   | 55.73% |
| 미래에셋대우     | 1227  | 2200   | 55.77% |
| HDC        | 1237  | 2200   | 56.23% |
| 한온시스템      | 1250  | 2200   | 56.82% |
| 한국타이어월드와이드 | 1252  | 2200   | 56.91% |
| DB 하이텍     | 1261  | 2200   | 57.32% |
| 한국금융지주     | 1271  | 2200   | 57.77% |
| 부광약품       | 1280  | 2200   | 58.18% |
| 카카오        | 1289  | 2200   | 58.59% |
| 롯데푸드       | 1298  | 2200   | 59.00% |
| 대한유화       | 1306  | 2200   | 59.36% |
| 동원시스템즈     | 1306  | 2200   | 59.36% |
| 한화케미칼      | 1307  | 2200   | 59.41% |

[표 5-3]와 [표 5-4]는 검증 데이터의 상/하위 15개 종목의 실험 결과, [표 5-5]와 [표 5-6]는 평가 데이터 상/하위 15개 종목의 실험 결과이며, 검증 데이터의 전체 정확도는 57.7% 평가 데이터의 전체 정확도는 57.6%가량 나왔다. 학습 결과 학습/검증/평가 데이터 상/하위 15개 종목에 겹치는 종목은 상위 15개 종목 중 'KT'외에 존재하지 않았다.

표 5-3 검증 데이터 상위 15개 종목 정확도

| 종목명        | 예측성공수 | 전체데이터수 | 정확도    |
|------------|-------|--------|--------|
| 태광산업       | 608   | 943    | 64.48% |
| 카카오        | 604   | 943    | 64.05% |
| 미래에셋대우     | 598   | 943    | 63.41% |
| 동원시스템즈     | 593   | 943    | 62.88% |
| 대우조선해양     | 590   | 943    | 62.57% |
| KT         | 589   | 943    | 62.46% |
| POSCO      | 583   | 943    | 61.82% |
| 한국타이어월드와이드 | 582   | 943    | 61.72% |
| S-Oil      | 580   | 943    | 61.51% |
| 롯데푸드       | 577   | 943    | 61.19% |
| 한섬         | 577   | 943    | 61.19% |
| 금호석유       | 574   | 943    | 60.87% |
| 현대해상       | 573   | 943    | 60.76% |
| 포스코대우      | 572   | 943    | 60.66% |
| 신한지주       | 572   | 943    | 60.66% |

표 5-4 검증 데이터 하위 15개 종목 정확도

| 종목명     | 예측성공수 | 전체데이터수 | 정확도    |
|---------|-------|--------|--------|
| 엔씨소프트   | 500   | 943    | 53.02% |
| 아모레퍼시픽  | 503   | 943    | 53.34% |
| 현대차     | 504   | 943    | 53.45% |
| 한샘      | 505   | 943    | 53.55% |
| 넥센타이어   | 505   | 943    | 53.55% |
| SKC     | 508   | 943    | 53.87% |
| 삼성전자    | 509   | 943    | 53.98% |
| LG 화학   | 510   | 943    | 54.08% |
| KT&G    | 513   | 943    | 54.40% |
| 기업은행    | 513   | 943    | 54.40% |
| SK 하이닉스 | 514   | 943    | 54.51% |
| 두산인프라코어 | 514   | 943    | 54.51% |
| CJ      | 517   | 943    | 54.83% |
| 효성      | 517   | 943    | 54.83% |
| 한미사이언스  | 517   | 943    | 54.83% |

표 5-5 평가 데이터 상위 15개 종목 정확도

| 종목명        | 예측성공수 | 전체데이터수 | 정확도    |
|------------|-------|--------|--------|
| 동원시스템즈     | 163   | 239    | 68.20% |
| 영풍         | 161   | 239    | 67.36% |
| KT         | 157   | 239    | 65.69% |
| 아이에스동서     | 156   | 239    | 65.27% |
| 한국전력       | 154   | 239    | 64.44% |
| 한국타이어월드와이드 | 153   | 239    | 64.02% |
| SK 텔레콤     | 153   | 239    | 64.02% |
| 기업은행       | 151   | 239    | 63.18% |
| LS         | 150   | 239    | 62.76% |
| LG 상사      | 150   | 239    | 62.76% |
| 현대해상       | 149   | 239    | 62.34% |
| SK 네트워크    | 149   | 239    | 62.34% |
| 세아베스틸      | 148   | 239    | 61.92% |
| 포스코대우      | 147   | 239    | 61.51% |
| 강원랜드       | 147   | 239    | 61.51% |

표 5-6 평가 데이터 하위 15개 종목 정확도

| 종목명     | 예측성공수 | 전체데이터수 | 정확도    |
|---------|-------|--------|--------|
| SK 하이닉스 | 114   | 239    | 47.70% |
| LG 전자   | 120   | 239    | 50.21% |
| LG      | 121   | 239    | 50.63% |
| 삼성전기    | 121   | 239    | 50.63% |
| LG 화학   | 123   | 239    | 51.46% |
| 한온시스템   | 123   | 239    | 51.46% |
| 삼성 SDI  | 123   | 239    | 51.46% |
| 현대중공업   | 124   | 239    | 51.88% |
| GS 건설   | 124   | 239    | 51.88% |
| 대웅제약    | 124   | 239    | 51.88% |
| 금호석유    | 126   | 239    | 52.72% |
| 아모레퍼시픽  | 127   | 239    | 53.14% |
| POSCO   | 127   | 239    | 53.14% |
| 부광약품    | 127   | 239    | 53.14% |
| OCI     | 128   | 239    | 53.56% |

포트폴리오 최적화 실험에서는 [표 5-7]와 같이 ‘투자 종목 최적화’와 ‘투자 비율 최적화’, 두 단계에 걸쳐 최적화한 실험에서 평가 데이터에 대한 등락률은 -0.15%로 떨어진 것에 반해 두 문제를 한 단계로 최적화 한 실험에서는 등락률은 0.05%로 0.2% 정도 더 좋은 결과를 보였다. 이는 두 문제를 따로 최적화할 경우 각 문제에 대한 독립성이 지나치게 높아서 생기는 문제로 판단되며, 이후 실험을 통해 구성된 포트폴리오와 벤치마크의 비교는 더 나은 결과를 보인 두 번째 실험을 기준으로 비교한다.

표 5-7 최적화 실험 방법별 평균 등락률 비교

| 실험방법                  | 평균 등락률(%) |       |
|-----------------------|-----------|-------|
|                       | 학습데이터     | 평가데이터 |
| 투자 종목 최적화             | 0.74      | -0.07 |
| 투자 비율 최적화             | 0.79      | -0.15 |
| 투자 종목 최적화 + 투자 비율 최적화 | 0.82      | 0.05  |

주식 투자 시 투자 성과를 상대적으로 평가하기 위한 벤치마크로 KOSPI 지수를 많이 사용한다. 이번 실험 평가에는 KOSPI 지수와 KOSPI 200 지수를 벤치마크로 사용하며, 동일한 시점에 동일한 금액을 투자하여 1년간 발생한 수익률을 비교한다. 수익률 계산은 거래 수수료 및 제세금을 고려하여 계산하며 그 식은 다음과 같다.

$$P_t = \frac{(C_{t-1} - O_{t-1}) - O_{t-1} * 0.00015 - C_{t-1} * 0.00015 - C_t * 0.003}{O_{t-1}} * 100$$

수식 5-1 수익률 계산 방법

[수식 5-1]에서 거래 수수료는 0.015%, 제세금은 0.3%로 설정하고 수익률을 계산하였다.

표 5-8 월별 평균 등락률 및 전월 대비 수익률 비교

| 거래연월        | 평균 등락률 (%) |           |       | 전월 대비 수익률 (%) |           |        |
|-------------|------------|-----------|-------|---------------|-----------|--------|
|             | KOSPI      | KOSPI 200 | 포트폴리오 | KOSPI         | KOSPI 200 | 포트폴리오  |
| 2017 년 07 월 | 0.06       | 0.08      | 0.13  | -4.25         | -3.95     | 0.05   |
| 2017 년 08 월 | -0.07      | -0.09     | 0.09  | -6.59         | -6.61     | 1.10   |
| 2017 년 09 월 | 0.06       | 0.12      | 0.02  | -4.12         | -2.56     | 1.11   |
| 2017 년 10 월 | 0.33       | 0.34      | 0.01  | -0.06         | 0.06      | -1.48  |
| 2017 년 11 월 | -0.08      | -0.11     | 0.15  | -4.51         | -4.59     | -0.14  |
| 2017 년 12 월 | -0.02      | 0.00      | 0.06  | -6.53         | -6.31     | -2.24  |
| 2018 년 01 월 | 0.18       | 0.12      | -0.04 | -2.49         | -3.34     | -2.88  |
| 2018 년 02 월 | -0.30      | -0.34     | -0.18 | -9.17         | -9.71     | -4.52  |
| 2018 년 03 월 | 0.04       | 0.03      | -0.04 | -5.60         | -5.29     | -3.13  |
| 2018 년 04 월 | 0.14       | 0.14      | 0.08  | -4.05         | -4.17     | -3.35  |
| 2018 년 05 월 | -0.18      | -0.2      | -0.07 | -7.20         | -6.76     | -1.52  |
| 2018 년 06 월 | -0.30      | -0.27     | -0.01 | -7.96         | -7.64     | -1.97  |
| 총합          | -0.15      | -0.19     | 0.21  | -62.53        | -60.88    | -18.96 |
| 평균          | -0.01      | -0.02     | 0.02  | -5.21         | -5.07     | -1.58  |

[표 5-8]는 포트폴리오 최적화 실험 결과 구성된 포트폴리오의 등락률과 [수식 5-1]로 계산한 전월 대비 수익률을 ‘KOSPI’, ‘KOSPI 200’과 비교한 것이다.

월별 평균 등락률을 보면 상승한 달과 하락한 달이 각각 6번씩 고르게 있지만, 상승한 달의 상승률에 비해 하락한 달의 하락률이 더 높아 1년 전체 기간을 보면 최종적으로 지수는 하락한 것을 알 수 있다. 게다가 수

익률 계산을 위해 거래 수수료 및 제세금을 고려하는 경우 더 큰 손실이 발생되 월평균 5%의 손실이 발생하는 것을 확인할 수 있다.

그에 비해 포트폴리오를 구성하는 경우 훨씬 안정적으로 자산 운용이 가능한 것을 확인 할 수 있으며, 이는 [그림 5-1]에서 더욱 확연하게 드러난다. [그림 5-1]는 평가 기간 동안 누적 등락률로, [그림 5-1] 보면 주가가 크게 하락하는 구간이 총 4개가 있음을 알 수 있다. 이때 ‘KOSPI’와 ‘KOSPI 200’는 장기간 하락장에서 벗어나지 못하는 것에 반해 실험을 통해 구성된 포트폴리오의 경우 동일기간 큰 하락 없이 상당히 안정적으로 자산이 유지되었다.

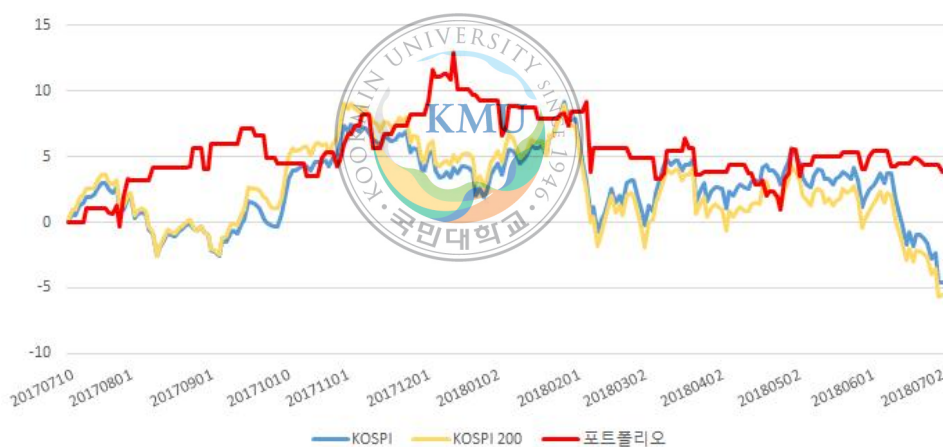


그림 5-1 누적 등락률



## 제 6장 결론 및 향후 연구 과제

주식 투자에서 로보 어드바이저에 대한 연구는 증권사 및 자산운용사를 중심으로 투자 성공 확률을 높이고 저렴한 가격에 자산 관리 서비스를 제공하기 위해 많은 연구가 진행되고 있다. 본 연구에서도 이와 같은 로보 어드바이저 시스템을 위한 주가 방향성 예측 및 포트폴리오 구성 실험을 기존의 시계열 관점이 아닌 주가 정보의 특징을 추출하는 방법으로 진행했다.

본 연구에서는 벤치마크인 ‘KOSPI’와 ‘KOSPI 200’에 비해 더욱 안정적인 투자를 할 수 있는 포트폴리오를 구성하는 데 성공했다. 하지만 최종적으로 투자 수익을 내는 것에는 좋은 결과를 보이지 못했다. [그림 6-1]는 거래 수수료와 제세금을 고려했을 때 누적 수익률로, 수익률에서도 ‘KOSPI’와 ‘KOSPI 200’보다 실험을 통해 구성한 포트폴리오가 좀 더 안정적으로 자산을 유지했다. 하지만 결과적으로 정도의 차이가 있긴 하지만 포트폴리오 역시 손실이 발생했다.

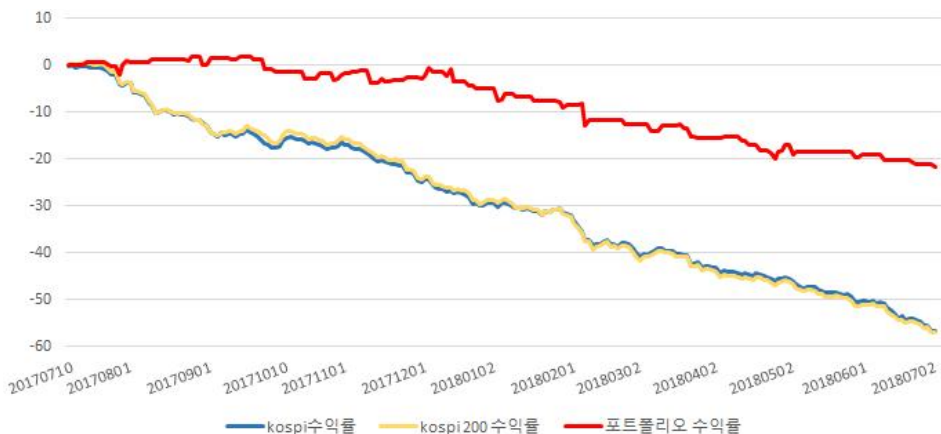


그림 6-1 누적 수익률

누적 등락률과 달리 누적 수익률에서 손실이 발생한 것은 구성된 포트폴리오의 수익률이 거래 수수료와 제세금 이상의 수익을 내지 못했기 때문이지만, 평가 데이터로 사용된 기간 동안 시장이 침체된 것도 하나의 원인이 될 수 있다.

그리고 이번 연구는 다음 날 주가 변동의 방향성을 예측하여 매수에 대한 정보는 제공하고 있다. 하지만 주가 변동의 크기, 즉, 주가가 얼마나 상승하고 하락하는지에 대한 정보는 제공하지 못하기 때문에 정확한 매도 시점을 결정하는 데 도움을 줄 수 없다는 한계가 있다. 이러한 이유로 본 논문에서도 등락률 및 수익률을 계산할 때 종가를 기준으로 계산했다.

이처럼 주가 변동의 크기를 예측하고 매도 시점을 결정하는 것은 로보 어드바이저 이후 자동으로 주식을 거래하는 오토 트레이딩(auto-trading)으로 발전하기 위해 꼭 필요한 연구로 볼 수 있다.

따라서 향후 연구로 위 문제를 보완하기 위한 두 가지 연구를 제안한다.

1. 주식 시장 경향 및 침체 예측
2. 주가 변동 크기 예측

위 두 연구가 보장된다면 보다 전문적인 로보 어드바이저 시스템 구축이 가능할 것으로 판단되며, 더 나아가 오토 트레이딩 시스템을 구축하는 것에도 큰 도움이 될 수 있을 것이다.

## 참 고 문 헌

- [1] 이근영, "국내외 로보어드바이저(RoboAdvisor) 동향 및 현황 분석," 금융보안원, 2016.
- [2] T. Epperson, B. Hedges, U. Singh and M. Gabel, "Hype vs. Reality : The Coming Waves of "Robo" Adoption," A.T. Kearney, 2015.
- [3] 서보익, "로보-어드바이저가 이끌 자산관리시장의 변화," 유진투자증권, 2016.
- [4] 김광훈, "인공지능 자산운용 시스템, 로보어드바이저," KISTI 마켓리포트, 2016.
- [5] Lai, Robert K., et al., "Evolving and clustering fuzzy decision tree for financial time series data forecasting," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3761-3773, 2009.
- [6] K. J. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1-2, pp. 307-319, 2003.
- [7] 김상호, 김동현, 한창희, & 김원일, "주가지수 관계와 유전자 알고리즘을 이용한 주식예측," *한국지능시스템학회논문지*, vol. Vol.18, no. 6, pp. 781-786, 2008.
- [8] L. & D. M. dos Santos Pinheiro, "Stock Market Prediction with Deep Learning: A Character-based Neural Language Model for Event-based Trading," *Proceedings of the Australasian Language Technology Association Workshop*, 2017.
- [9] Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K., "Deep learning for stock prediction using numerical and textual information," in *IEEE/ACIS 15th International Conference*, 2016.
- [10] Baba, N., & Kozaki, M., "An intelligent forecasting system of stock price using neural networks," in *IJCNN International Joint Conference on Neural Networks*, 1992.
- [11] 이우식, "딥러닝분석과 기술적 분석 지표를 이용한 한국 코스피주가지수 방향성 예측," *한국데이터정보과학회지*, vol. 28, no. 2, pp. 287-295, 2017.
- [12] Nelson, D. M., Pereira, A. C., & de Oliveira, R. A., "Stock market's price movement prediction with LSTM neural networks," in *Proceedings of the 23rd ACM SIGKDD*

*International Conference on Knowledge Discovery and Data Mining.*, 2017.

- [13] 류동기, 황석진, & 최준수, "양방향 장단기 기억 네트워크를 이용한 단기 추가예측," *한국정보과학회 학술발표논문집*, vol. 2017, no. 12, pp. 992-994, 2017.
- [14] Krizhevsky, A., Sutskever, I., & Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012.
- [15] Szegedy, Christian, et al., "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR), 2015*, 2015.
- [16] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A., "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI Conference on Artificial Intelligence*, 2017.
- [17] Lin, M., Chen, Q., & Yan, S., *Network In Network*, arXiv preprint arXiv:1312.4400, 2013.
- [18] He, K., Zhang, X., Ren, S., & Sun, J., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [19] 나호영, & 이상헌, "유전자 알고리즘을 이용한 도시공간형태별 입지선정 모델," *한국시물레이션학회지*, vol. 17, no. 3, pp. 1-109, 2008.
- [20] 박수영, "유전자 알고리즘을 이용한 실내 보행자 최적 대피모형", 서울시립대학교, 2012.
- [21] B. G. Malkiel, "A Random Walk Down Wall Street: Including a Life-Cycle Guide to Personal Investing", W. W. Norton & Company, 1973.

## Abstract

# Stock Price Trends Forecasting And Investment Portfolio Optimization Using Artificial Intelligent

*by Hwang, Sek Jin*

Department of Computer Science

Graduate School, Kookmin University,

Seoul, Korea



Stock investment is the most well-known investment method, but it is a very difficult investment method to get profits. Forecasting stock price fluctuations for effective and stable investment has long been a subject of interest in many areas, including economics and mathematics. However, despite many efforts from the past, it still remains a difficult problem, and computer science has tried to solve this problem by using machine learning methods such as Decision Tree and SVM (Support Vector Machine).

In this paper, we propose a method of constructing an optimal investment portfolio using the Deep Learning method, forecasting the direction of the stock price the next day, selecting investment items with the genetic algorithm using the deep learning learning results, and

optimizing the investment ratio. The Deep Learning model uses a new Convolutional Neural Network (CNN) based on the Inception-Resnet-v2 model instead of the Recurrent Neural Network (RNN), which is known to be suitable for time series data such as stocks. The data of stock price information processed as index and the title of economic article are used as learning data for each item. After the learning, the genetic algorithm is used to select the investment items reflecting the results of Precision / Recall / Accuracy and the same day's forecast probability, and the experiment is conducted to optimize the investment ratio of the selected items. We prove the usefulness of the proposed method in this paper by comparing the stock price index that used as benchmarks with our method.

