딥 러닝을 이용한 주택가격 예측에 관한 연구

A Study on Prediction of Housing Price Using Deep Learning

저자
(Authors)

전해정, 양혜선
Chun, Hae Jung, Yang, Hye Seon

출처
(Source)

주거환경 17(2), 2019.6, 37-49(13 pages)
RESIDENTIAL ENVIRONMENT : JOURNAL OF THE RESIDENTIAL ENVIRONMENT INSTITUTE OF KOREA 17(2), 2019.6, 37-49(13 pages)

발행처
(Publisher)

한국주거환경학회
Residential Environment Institute Of Korea

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08746811

APA Style

전해정, 양혜선 (2019). 딥 러닝을 이용한 주택가격 예측에 관한 연구. 주거환경, 17(2), 37-49

pISSN 1738-0316, eISSN 2508-3961
住居環境 | 韓國住居環境學會誌 | 第 17卷 第 2號 (通卷 第 44號) | pp.37~49.
접수일: 2019.03.14. 1차심사일자: 2019.04.08. 2차심사일자: 2019.05.23. 게재확정일자: 2019.06.13.
http://dx.doi.org/10.22313/reik.2019.17.2.37

딥 러닝을 이용한 주택가격 예측에 관한 연구

전 해 정* • 양 혜 선**

A Study on Prediction of Housing Price Using Deep Learning

Chun, Hae Jung • Yang, Hye Seon

Abstract

The purpose of this study is to estimate housing prices using deep running. The simple RNN, LSTM, and GRU models, which are evaluated to be suitable for time series forecasting, are based on the time series data of apartment real price index, interest rate, household loan, building permit area and consumer price index. As a result of the empirical analysis, it is confirmed that the prediction power of the GRU model is superior to that of the learning data by evaluating the performance of forecasting power on apartment real price index based on the RMSE value. On the other hand, in the verification data, it is confirmed that the prediction power of the RNN model is excellent. Also, if the performance of the deep running model is evaluated with accuracy, the accuracy of the RNN model and the GRU model is the highest. As a result of this study, the government needs to build and develop a system that can predict and diagnose the housing market by using the deep learning technique that combines artificial neural network and big data to advance the housing market.

키 워 드: 주택매매가격, 비선형, 예측, 딥 러닝, 순환신경망(RNN)
Key words: Housing Price, Nonlinear, Prediction, Deep Learning, Recurrent Neural Network

* 상명대학교 경영대학원 글로벌부동산학과 조교수, 주저자
** 중앙대학교 도시계획・부동산학과 시간강사, 교신저자

Ⅰ. 서 론

우리나라의 가계자산 중 가장 높은 비중을 차지하는 것은 주택이다. 통계청의 가계금융 · 복지조사 결과에 따르면 2018년 3월 기준 가계의 평균 자산은 약 4억 1,573만 원으로 그 중 부동산이 차지하는 비중은 약 2억 9,177만 원으로 약 70.2%이다. 특히, 거주주택이 약 1억 6,895만 원으로 전체 가계 자산 규모의 절반에 가까운 약 40.6%를 차지한다. 이처럼 가계의 자산 구조가 주택에 편중되어 있어 주택가격의 급격한 변동이 발생하게 되면 국민 경제에 큰 영향을 미치게 된다.

주택가격의 상승은 주택소유자의 자산가치를 증가시켜 소비가 늘어나고 건설경기가 활성화되며 내수경기가 회복되는 순기능이 있다. 그러나 주택가격의 급격한 상승은 물가상승을 유발하여 국민총생산을 감소시키는 역기능도 있다(정원구 · 이상엽, 2007). 반면 주택가격이 급격하게 하락하게 되면 자산가치가 감소하여 역자산효과의 발생으로 소비와 투자가 위축되어 자산디플레이션으로 연결될 우려가 있다. 특히 최근과 같이 가계의 부채비율이 높아진 상태에서 자산디플레이션이 발생하게 되면 금융기관의 부실화, 기업의 파산 등을 통해 대공황과 같은 복합불황으로 이어질 가능성이 존재한다(손경환 외, 1998).

이로 인해 주택가격의 상승과 하락은 국민과 국가의 주요 관심사항이며, 주택가격 변동에 대비하기 위해 다양한 방법을 이용한 주택가격 예측연구가 계속 진행되고 있다. 주택정책이 가격 변동과 같은 문제에 적절히 대응하면서 예측 가능한 주택시장을 만들어 나가려면 주택가격 예측모형의 개발이 필요하기 때문이다. 기존 연구의 주택가격 예측은 대체로 자기회귀이동평균모형(Autoregressive Integrated Moving Average Model, ARIMA)이나 벡터자기회귀모형(Vector Autoregressive Model, VAR) 등 시계열 분석을 활용하여 주택가격을 예측하는 방식으로 이루어져왔다. 하지만 시계열 분석은 현실과는 다소 동떨어진 선형모형(Linear Model)을 가정하기 때문에 실제적으로 나타나는 주택가격의 비선형적인(Non-linear) 움직임을 반영하기에는 한계가 있다는 비판이 있다(이창로 · 박기호, 2016).

최근 공학 분야에서 주목받고 있는 딥 러닝(Deep Learning)은 비선형 추정기법으로 함수 형태, 변수나 오차항의 분포, 특성 간의 상관관계 등에 대한 별도의 가정 없이 활용될 수 있어 기존 방법보다 유연성이 있다(Mester, 1997). 이로 인해 딥 러닝 기법은 주가 예측 등 미래 예측 분야에서 활발한 연구와 우수한 성과를 보여주고 있으며 주택가격 예측과 관련해서도 그 활용 가능성이 높을 것으로 기대된다. 딥 러닝은 머신 러닝(Machine Learning)의 한 연구 분야로 인공신경망(Artificial Neural Network, ANN)과 빅 데이터(Big Data)를 기반으로 한 분석방법이다. 즉, 딥 러닝은 머신 러닝의 알고리즘 집합체로 학습데이터의 종류가 많고 크기가 클수록 예측력이 우수해지는 특성이 있다(최정원 외, 2017 ; 민성욱, 2017). 과거의 국내 주택가격 예측연구에서는 거시경제나 주택시장에 대한 빅 데이터가 구축되어 있지 않았다. 또한 딥 러닝 연구 분야에서는 학습데이터 유형에 따라 학습과정에서 문제가 발생하면서 응용이 어려워 연구가 위축되었다. 그러나 최근 컴퓨팅의 기술력이 빠르게 발전하면서 딥 러닝 연구에서 학습자료 유형에 대한 응용성과를 높였다. 비선형성을 반영하여 주택가격을 정확하게 파악할 수 있는 딥 러닝을 이용한 예측 연구는 앞서 언급한 관련 자료의 빅데이터 구축 문제와 컴퓨팅 기술력 문제 등으로 인해 최근에서야 연구가 이루어지고 있는 상황이다. 이런 현실을 감안할 때 딥 러닝을 이용한 한국 주택가격 예측모형 개발이 필요한 상황이다.

본 연구의 목적은 인공신경망 기술과 빅 데이터를 접목한 딥 러닝을 이용해 주택가격을 예측하는 것이다. 딥 러닝 모형은 알고리즘의 구조에 따라 다양하게 분류되며, 대표적으로 합성곱 신경망(Convolutional Neural Network, CNN), 심층 신

경망(Deep Neural Network, DNN)과 순환신경망(Recurrent Neural Network, RNN) 등이 있다. 본 연구에서는 시계열 예측에 적합한 딥 러닝의 알고리즘으로 평가받고 있는 RNN, LSTM(Long Short Term Memory Network), GRU(Gated Recurrent Unit) 모형의 예측결과를 비교분석해 예측력이 우수한 모형을 판별하고자 한다. 본 연구의 종속변수는 전국 아파트 실거래가격지수이며, 독립변수로는 주택가격의 영향을 고려하여 CD금리, 가계대출금, 건축허가면적, 소비자물가지수를 선정하였다.

본 연구의 구성은 다음과 같다. 2장은 딥 러닝을 이용한 주택가격 예측과 관련된 선행연구를 고찰한다. 3장은 분석모형인 딥 러닝에 대해 살펴보고자 한다. 4장은 실증분석으로 자료의 기초통계량을 살펴보고 딥 러닝을 이용해 나타낸 분석결과를 기술한다. 마지막 5장은 결론으로 연구결과를 요약하고 연구 결과에 따르는 시사점을 제시하고자 한다.

II. 선행연구 고찰

국내 주택가격과 부동산 가격에 대한 예측 연구는 오랜 기간 동안 다양한 분석모형을 이용하여 연구되어 왔다(서승환, 1994; 박헌수, 2003; 정규일, 2006; 김세완, 김은미, 2009; 이준용・손재영, 2010; 김문성・배형, 2013; 전해정, 2017). 그러나 최근 공학계열에서 이용하기 시작한 인공신경망 분야의 딥 러닝 분석방법을 주택・부동산 시계열 자료에 활용한 국내 연구는 미미한 상황이다.

조유나・김수현・송규원(2016)은 2016년 12월 총 4,628개의 서울시 아파트 시세로 딥 러닝 기술을 이용한 주택가격 예측을 하였다. 주택가격의 설명변수는 특성요인, 지역요인, 건설사, 난방방식을 선정하였다. 분석 결과, 딥 러닝의 DNN모형이 헤도닉 가격모형보다 설명력(R^2)과 예측력(RMSE)에서 더 우수한 것으로 확인되었다.

민성욱(2017)은 2006년부터 2016년까지 서울시 아파트 실거래가지수와 거시경제변수로 딥 러닝(DNN)과 인공 신경망의 다층 퍼셉트론, 머신 러닝의 SVM, RF을 이용해 주택가격 예측을 하였다. 서울시 아파트 실거래가 지수에 영향을 미치는 연속형 자료의 설명변수는 수익증권, 경제심리지수, 기업대출 연체율로 분석되었다. 분석결과, 딥 러닝의 DNN모형보다는 인공 신경망의 다층 퍼셉트론모형 예측력이 가장 우수하다는 것을 확인하였다.

배성완・유정석(2017)은 2006년부터 2016년까지 서울시의 아파트 매매실거래가격지수, 아파트 매매가격지수, 아파트 전세가격지수, 지가지수로 딥 러닝(DNN, LSTM)을 이용해 부동산가격지수에 대한 단변량 예측 연구를 하였다. 분석결과, 딥 러닝 모형이 ARIMA모형보다 우수한 예측력이 있다고 확인하였다. 딥 러닝 모형 중에서는 DNN모형이 LSTM모형보다 더 우수한 예측력을 보였으며, 부동산가격지수 중에서는 상대적으로 변동이 적은 지가지수의 예측력이 가장 우수하고 변동이 심한 아파트 매매실거래가지수가 가장 부족하다는 분석결과를 제시하였다. 지표의 변동성에 대한 특성이나 과거 값에 대한 의존도에 따라 예측력이 다르게 나타난다고 하였다.

이창로・김세형(2018)은 2012년부터 2015년까지 서울시 강남구, 경상남도 김해시, 전주시 덕진구, 전라남도 해남군의 단독주택 거래가격으로 딥 러닝(DNN)을 이용하여 주택가격을 예측하였다. 지리적 위치의 비선형 효과를 주택가격에 포함하기 위해 단독주택을 분석대상으로 선정하였다. 개별보다는 단지의 위치와 품질이 더 중요한 아파트와 달리 단독주택은 위치와 건물의 특성이 각각 다르기 때문이다. 분석결과, 지리적 좌표로 인해 비선형성을 포착할 수 있는 DNN의 예측력이 선형 회귀모형보다 더 우수한 것으로 확인되었다. 특히, DNN 모형의 예측 성능은 단독주택의 특성이 다양하고 복잡할수록 더 높아지는 것으로 분석되었다.

배성완・유정석(2018)은 2006년부터 2017년까

지 서울시 아파트 매매실거래가격지수로 머신 러닝 방법을 이용해 부동산가격지수를 예측하였다. 부동산가격지수의 설명변수는 회사채수익률, 소비자물가지수, 통화량, 광공업지수로 선정하였다. 시장상황에 따른 예측력을 비교하기 위해 안정적 상승기(2016년 8월~2017년 8월)와 구조적 변화기(2008년 월~2009년 8월)를 검증 데이터로 설정하여 분석하였다. 분석결과, 안정적인 시장과 불안정적인 시장에서 모두 머신 러닝모형(SVM, RF, GBRT, DNN, LSTM)의 예측력이 시계열분석 모형(ARIMA, VAR, BVAR)보다 더 우수하다고 확인하였다. 특히, 시계열분석 모형은 선형모형으로 가정되어 있어 급격한 변화가 포함된 시장 추세를 예측하기에는 한계가 있다고 보았다.

이태형・전명진(2018)은 2006년부터 2017년까지 서울시 중대형과 대형아파트 실거래가격지수를 이용하여 시계열분석 모형인 VAR모형과 딥 러닝의 RNN, LSTM 알고리즘으로 서울시 주택가격을 예측하였다. 서울시 주택가격을 예측하는 설명변수는 주가지수, 대출금리, 기대인플레이션율, 소비자물가지수, 주택전세가격지수, 실업률로 선정되었다. 분석결과, RMSE 값 기준으로 시계열분석 모형보다. 딥 러닝모형이 더 우수한 예측력을 보인다는 것을 확인하였다. RNN과 LSTM의 RMSE값 차이는 크지 않으나 LSTM의 표준편차가 상대적으로 작아 LSTM의 예측력이 더 높다고 하였다.

본 연구의 차별성은 시계열 자료를 이용한 딥 러닝 모형으로 국내 주택가격에 대한 예측력을 실증적 비교분석하는 것에 있다. 즉, 시계열 자료를 이용한 딥 러닝 기법 중 예측력이 높은 모형인 RNN 알고리즘인 simple RNN, LSTM과 GRU 모형의 분석결과를 제시하고 이를 비교 평가하는 해 예측력이 높은 모형을 실증적으로 알아내는 것에 차별성이 있다. 또한 주택시장 수요와 공급이론과 선행연구에 기초하여 변수를 체계적으로 선정함에 차별성이 있다.

Ⅲ. 분석모형

1. 딥 러닝(Deep Learning; 심층 학습)

딥 러닝은 컴퓨터의 학습능력을 키우기 위해 사람의 사고방식을 가르치는 머신러닝(Machine Learning; 기계 학습)의 연구 분야로, 데이터 특징을 학습 기반으로 한다(이요섭・문필주, 2017). 이 연구의 목적은 컴퓨터가 스스로 여러 비선형 변환기법의 조합을 이용하여 다량의 복잡한 학습 데이터들을 높은 수준의 추상화된 정보로 추출하거나 분류하는 것이다. 이러한 딥 러닝 기술은 사람이 만든 프로그램 없이 기존 자료의 학습된 알려진 속성을 바탕으로 새로운 자료를 정확하게 처리할 수 있다. 즉, 딥 러닝의 핵심적인 역할은 학습에 의한 예측이며 최근에는 주가 예측이나 기업 부도 예측 등 금융시장 예측 분야에서 우수한 성과를 거두고 있다(최근우・송기선・강요셉, 2016).

딥 러닝의 구조는 인간의 뇌를 모방한 인공 신경망(ANN; Artificial Neural Network)에서 비롯되었다. 가장 기본적인 구조는 여러 개의 은닉 레이어(Layer)를 가진 다층 퍼셉트론(Multilayer Perceptron)이 있다. 즉, 일반적인 인공 신경망은 입력 레이어와 출력 레이어 사이의 중간층인 은닉 레이어가 1개인 반면 딥 러닝은 은닉 레이어가 여러 개로 이루어질 수 있다. 예를 들어, 은닉 레이어가 2개인 경우 은닉 레이어1에서 받은 입력신호를 가공하여 도출된 출력을 은닉 레이어2로 보내는데 이때 은닉 레이어1에서 도출된 출력은 은닉 레이어2의 입력이 된다. 이러한 다중 처리 계층으로 구성된 딥 러닝은 여러 수준의 추상화를 통해 자료의 표현을 정확하게 학습할 수 있다.

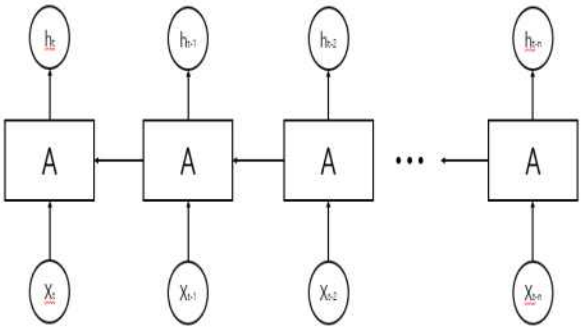
다른 머신 러닝 방법과 마찬가지로, 딥 러닝 방법은 지도 학습과 비지도 학습으로 구별되는데 학습 프레임워크에 따라 학습 모델이 다르다. 지도

학습(Supervised Learning)은 입출력의 쌍으로 구성된 학습 예제들로부터 입력을 출력으로 사상(morphism)하는 함수를 학습하는 과정이다. 비지도 학습(Unsupervised Learning)의 경우는 학습 예로 입력은 주어지나 대응되는 출력이 없으며 이 경우 입력 패턴들에 공통적인 특성을 파악하는 것이 학습의 목적이다(장병탁, 2007). 예를 들어 지도학습에서의 심층 머신러닝 모델은 컨볼루션신경망(Convolutional Neural Networks)이 있으며, 비지도 학습에서의 머신러닝 모델은 DBN(Deep Belief Nets)이 있다.

딥 러닝은 컴퓨터가 데이터를 학습하고 실행할 수 있도록 하는 머신 러닝의 알고리즘 집합이다. 이러한 알고리즘은 다양한 인공 신경망 구조에 따라 분류되는데, 대표적으로 DNN(Deep Neural Network; 심층 신경망), CNN(Convolutional Neural Network; 합성곱 신경망), RNN(Recurrent Neural Network; 순환 신경망) 등이 있다. 본 연구에서는 주택가격의 예측이 연구의 목적이므로 시계열 예측에서 우수한 성과를 보이고 있는 순환 신경망(simple RNN)과 LSTM, GRU를 중심으로 실증 분석하였다(신동하 • 최광호 • 김창복, 2017).

2. 순환신경망(RNN)

RNN은 현재 출력데이터와 과거 정보와의 관계를 밝히는데 쓰인다. 기존의 전연결 신경망(fully connected layers)과 CNN에서는 입력 레이어에서 은닉 레이어 그리고 출력 레이어를 거친 과정을 거쳐야 한다. 각 레이어 사이는 완전 연결되거나 부분 연결되어 있다. 하지만 현재의 학습과 과거의 학습 간의 연결이 없으므로 순차적인 데이터 학습에 한계가 있다. 반면 RNN은 과거 정보를 기억하고 이를 이용해서 현재의 출력값을 예측한다. 따라서 RNN의 은닉 레이어들은 연결되어 있으며 현재 입력 레이어가 포함되어있을 뿐만 아니라 과거 은닉 레이어의 출력값까지 포함되어 있다. 이러한 특징으로 인해 순환 신경망(RNN)은 시간을 통한 역전파 알고리즘과정(BPTT)을 통해 훈련하기 때문에 주로 순서가 있는 서열 데이터를 처리하거나 시계열 자료를 효과적으로 모델링하는 알고리즘이다(배성완 • 유정석, 2018).



〈그림 1〉 RNN의 구성도

위에서 언급한 것과 같이 simple RNN은 과거의 관측값에 의존하는 구조로 RNN에 의한 예측은 긴 시간 동안 나타나는 입력 데이터의 장기적인 의존성(Long Tern Dependency)이 문제가 된다. 학습과정에서 입력값의 길이가 늘어나게 되면 이전 은닉층의 개별 기울기가 1보다 작으면 가중치를 추정하기 위한 목적함수의 기울기가 소멸(Vanishing Gradient)하고 이전 은닉층의 개별 기울기가 1보다 크게 되면 목적함수의 기울기가 폭발(Exploding Gradient)하게 된다[1]. 따라서 simple RNN은 장기적인 패턴이 있는 자료를 학습하는데 한계가 있다(이태형 • 전명진, 2018).

전형적인 RNN모형의 구조도는 아래와 식1과 같다. RNN에서 가장 중요한 개념은 자료의 순서인데 RNN은 각 시점의 입력을 현재 모형의 상태와 결합하여 출력하는 알고리즘이다. 이러한 특성이 있어 RNN은 특히 음성 식별, 언어 모델링, 번역, 이

1) 목적함수의 기울기 값이 0에 근접하게 되면 목적함수의 최적화를 위한 모수벡터의 진행 방향을 찾기 어렵고 목적함수의 기울기 값이 과도하게 크게 되면 딥 러닝에서의 학습이 불안정해지는 문제가 있다(안성만 외, 2017).

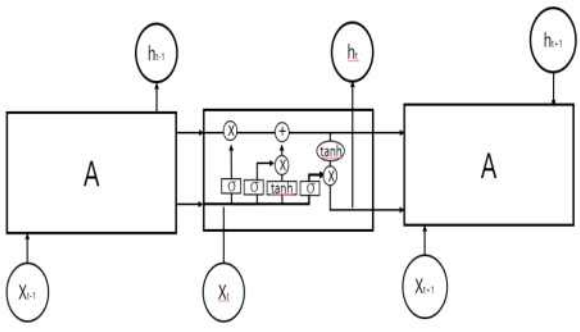
미지 인식 분야에서 주로 사용되고 있다.

RNN의 은닉 레이어를 수식으로 표현하면 아래와 같다. x_t는 t번째 term의 입력을 의미하고 W는 학습을 통해 값이 결정되는 가중치 행렬(Weight Matrix)들이며, b는 각 편차를 의미하며, σ는 sigmoid함수를 뜻한다. h_t는 은닉 레이어의 출력을 뜻한다(최경호, 2016:32).

$$h_t = \sigma(W \bullet [h_{t-1}, x_t] + b_h) \qquad \text{(식1)}$$

3. LSTM(Long Short Term Memory)

LSTM은 Hochreiter & Schmidhuber (1997)에 의해 제기되었으며 장기 의존성 문제를 해결할 수 있는 특수 형태의 RNN 모형이다. 위에서 설명한 것과 같이, RNN은 순차적인 데이터가 길어지면서 이전의 먼 거리에 있는 학습이 현재의 결과에 미치는 영향이 미미해진다는 단점이 있다. 반면 LSTM은 기존 입력값을 기억할 수 있는 기억소자(Memory Cell)이라고 불리는 구조를 사용하고 있어 이러한 장기 의존성 문제를 해결할 수 있다. 따라서 LSTM의 경우 데이터의 길이가 긴 작업에서 상대적으로 좋은 성능을 보일 수 있다(김양훈 외, 2016).



〈그림 2〉 LSTM의 구성도

모든 RNN에는 체인 형태의 반복적인 신경망 모듈이 있으며 그 구조가 단순한 형태로 존재한다. LSTM 또한 이와 같은 구조지만 내부의 반복 모듈은 상대적으로 다른 구조를 갖고 있다. 단일 신경망 계층과는 달리 LSTM에서는 네 가지 종류의 모듈이 있는 방식으로 상호 작용한다. <그림 2>를 보면 LSTM은 세 개의 '게이트'가 있는 특수 네트워크 구조이다.

LSTM의 '게이트'은 정보를 각 시점별 상태에 선택적으로 영향을 주는 중요한 역할을 한다. 이는 sigmoid 활성화 함수를 사용하는 전연결 신경망(Fully Connected Network) 계층이 0과 1사이의 값을 출력하기 때문에 문이 열리면 (sigmoid 출력이 1) 모든 정보를 전달하고 문을 닫으면 (sigmoid 출력이 0) 아무 정보도 전달하지 않은 구조이다.

위의 LSTM 구성도를 수식으로 표현하면 다음 식 2와 같이 표현할 수 있다. i 와 o 는 각각 현재 시점의 입력 정보를 사용할지를 결정하는 입력 게이트(input gate)와 현재 시점에서 유닛의 상태를 출력할지를 결정하는 출력 게이트(Output Gate)이다. c는 기억소자(Memory Cell)로 해당 유닛의 현재 시점에서의 상태를 뜻한다. f는 망각 게이트(Forget Gate)로 메모리 셀이 유닛의 이전 시점의 상태를 기억하여 현재 시점의 시퀀스에 적용할지를 정한다(최경호, 2016). tanh는 쌍곡선탄젠트 활성화 함수를 뜻한다.

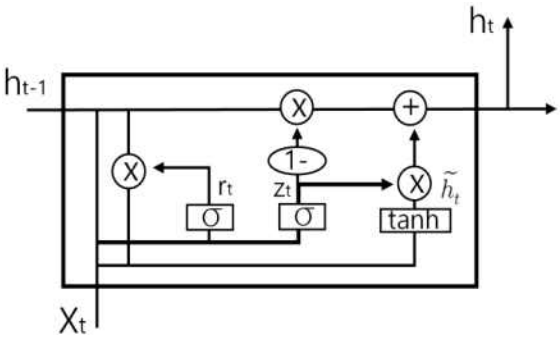
$$\begin{aligned} i_t &= \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \\ f_t &= \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \\ c_t &= f_t \odot c_{t-1} + i_t \odot tahn(W_i \bullet [h_{t-1}, x_t] + b_i) \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

(식2)

4. GRU(Gated Recurrent Unit)

GRU는 LSTM의 변형으로 LSTM 구조에서의 망각 게이트와 입력 게이트 대신 GRU는 업데이트 게이트를 가지고 있다. 셀 상태(Cell State)와 은닉 상태(Hidden State) ht를 결합하여, 현재 시점에서의 새로운 정보를 계산하는 방법은 LSTM과 다르다. 다음 <그림 3>은 GRU가 ht를 업데이트하는 과정을 보여준다.

위 GRU 구성도를 수식으로 표현하면 다음 (식 3)과 같다. r은 과거 정보의 은닉층 출력을 유닛의 입력에 고려하는 비율을 설정하는 리셋 게이트(Reset Gate)이고, z는 업데이트 게이트로 유닛의 출력에 과거 정보의 은닉층 출력을 그대로 반영하는 정도를 결정한다.



〈그림 3〉 GRU의 구성도

리셋 게이트는 LSTM의 망각 게이트처럼 과거 정보의 상태를 현재 정보에 적용할지 결정하는 역할을 하고, 업데이트 게이트는 현재 정보에서의 유닛의 상태를 출력에 고려하는 비율을 설정한다. 특히 업데이트 게이트는 과거 정보에서의 유닛의 출력을 다음 정보로 전해주는 bypass를 생성할 수 있어 멀리 떨어진 간격의 정보를 잘 전달할 수 있다(최경호, 2016).

$$
\begin{aligned}
r_t &= \sigma(W_r X_t + U_r h_{t-1} + b_r) \\
z_t &= \sigma(W_z X_t + U_z h_{t-1} + b_z) \\
\widehat{h}_t &= tahn(W_h X_t + r_t U_h h_{t-1} + b) \\
h_t &= (1 - z_t)\widehat{h}_t + z_t h_{t-1}
\end{aligned}
\qquad \text{(식3)}
$$

LSTM은 세 개의 게이트를 가지지만, GRU는 리셋 게이트(Reset Gate)과 업데이트 게이트(Update Gate) 두 개의 게이트만 가지고 있으며, LSTM의 출력 게이트(Output Gate)는 없다. 출력 게이트가 없는 상태에서 LSTM이 구현하고자 하는 정보의 전달과 조합 방식을 다르게 구현한 것이 GRU라고 말할 수 있다. 직관적으로 보면, 리셋 게이트는 새로운 입력을 과거 메모리와 어떻게 조합하는지를 결정하며, 업데이트 게이트는 이전 메모리 정보를 어느 정도만 유지하여 새로운 상태(State)를 계산해 내는지를 결정한다.[2)]

Ⅳ. 실증 분석

1. 자 료

본 연구에서 딥 러닝을 이용해 주택가격을 예측하는 연구로 사용 변수는 선행연구와 관련 이론[3)]을 참조해 아파트가격, CD금리, 가계대출금, 건축허가면적과 소비자물가지수를 설정하였다. 시간적 범위는 2006년 1월[4)]부터 2018년 3월까지로 월별 자료를 이용하였고 공간적 범위는 전국으로 하였다. 표1과 같이 아파트가격은 한국감정원에서 제공하는 전국 아파트 실거래가격지수를 이용하였고 CD금리, 가계대출금, 건축허가면적과 소비자물가

2) https://jay.tech.blog/2016/12/08/lstmlong-short-term-memory-rnn/

3) D-W 4분면 모형(1996)에 따르면 주택가격은 수요와 공급이 일치하는 점에서 결정된다. 주택수요함수의 영향요인은 이자율, 물가, 유동성 등이 있다. 본 연구에서 이자율의 대리변수로 CD금리를 선정하였는데, 이는 CD금리가 주택담보대출금리와 연동되어 시중금리를 잘 반영하기 때문이다(이용만, 이상한, 2004). 물가의 경우 인플레이션이 발생하게 되면 실물자산에 대한 선호현상으로 주택수요가 급증하게 된다(허재완, 1991). 이에 본 연구에서는 물가의 대리변수로 소비자물가지수를 선정하였다. 또한 유동성이 확대되면 주택수요가 증가해 자산가격이 상승하는 효과가 있다. 본 연구에서는 금융기관의 자산측면에서 유동성을 살펴보기 위해 가계대출금(예금취급기관)을 유동성의 대표변수로 선정하였다(정규일, 2006). 주택공급함수의 영향요인으로 신축 주택량, 기존주택의 재고량, 주택건설비 등이 고려될 수 있다. 신규 주택량의 증가로 주택수요량보다 주택공급이 많아지게 되면 주택가격은 하락하게 된다. 이에 따라 본 연구에서는 신규 주택량의 대리변수로 건축허가면적현황을 선정하였다(배영균, 2012).

4) 본 연구의 분석기간 시작점은 2006년 1월부터로 설정하였는데 이는 해당기간부터 한국감정원에서 아파트실거래가격지수를 제공하고 있기 때문이다.

지수는 한국은행 자료를 이용하였다.

〈표 1〉 변수

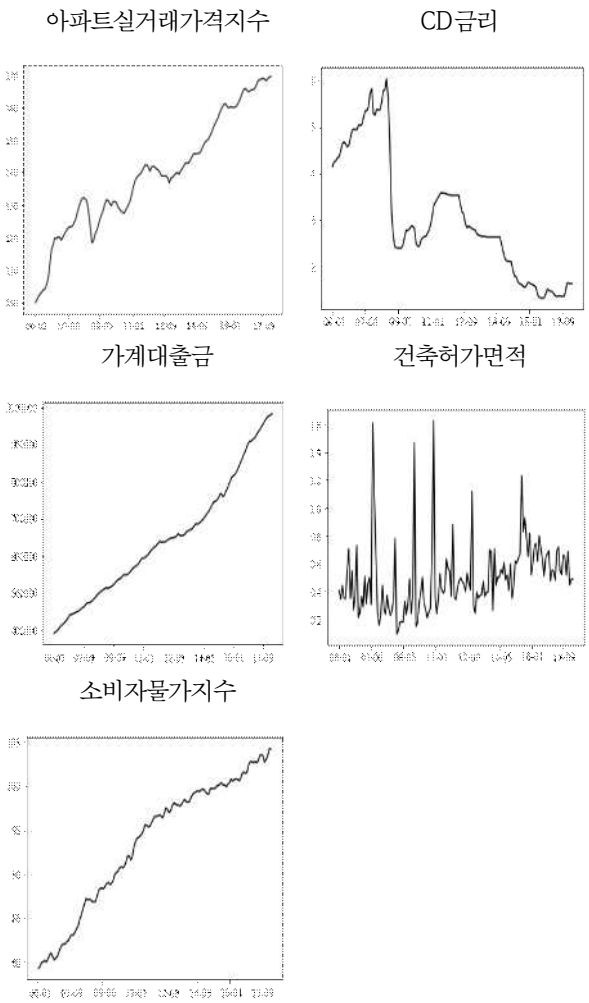
변수명	설명	단위	출처
아파트가격 (y)	아파트 실거래가격지수	지수	한국감정원
CD금리 (x1)	시장금리	연%	한국은행
가계대출금 (x2)	예금취급기관의 가계대출(월별)	십억원	한국은행
건축허가면적 (x3)	건축허가현황 연면적(주거용)	㎡	한국은행
소비자물가지수 (x4)	소비자물가지수	지수	한국은행

수집된 자료에 대해 각각 RNN, LSTM과 GRU 세 가지 딥 러닝 기법을 사용해 분석하였으며 각 변수에 대한 기초통계량은 표 2에 제시되어 있다. 아파트실거래가격지수 평균은 139.42이고 CD금리는 평균 3.06%, 가계대출금은 평균 6,465,837억원, 건축허가면적은 평균 4,940,438㎡, 소비자물가지수는 평균 93.61로 나타났다.

〈표 2〉 기초통계량

변수명	Mean	Std. Dev.	Min	Max
아파트 가격 (y)	139.42	17.68	100.00	169.60
cd금리 (x1)	3.06	1.31	1.34	6.03
가계 대출금 (x2)	646583.70	161223.60	391955.40	983476.50
건축허가 면적 (x3)	4940438	2534217	904537	16300000
소비자 물가지수 (x4)	93.61	7.45	79.31	104.26

본 연구에 사용된 변수들은 Min-Max 정규화 가공 과정을 거쳤는데 정규화를 통해서 딥 러닝 학습 모델에 최소, 최대의 기댓값을 한정지어 줌으로써 컴퓨터가 조금 더 빠르게 데이터를 학습하도록 할 수 있다. 변수의 변화추이는 〈그림 4〉와 같다.



〈그림 4〉 변수 변동추이

2. 모형설계

1) Basic RNN Cell (세팅값)

RNN을 최적화하기 위해 결정해야 하는 사항들은 활성화 함수(Activation Function), 검증 횟수(Epochs), 중간층(Hidden Layer) 개수, 노드(Node) 개수, 최적화 방식(Optimizer), 배치(Batch) 등이 있다(배성완・유정석, 2017). 모형들의 최종 초모수 설정값은 〈표 3〉에 요약되어 있다.

본 연구에서는 은닉층은 2개, 테스트횟수는 200

회, 배치사이즈는 10, 활성화 함수는 렐루 함수(Relu Function), 최적화(Optimizer)방법은 아담 (ADAM) 알고리즘, 초기화(Initialization)방법은 기본값인 0상태(Zero State)방법을 사용하였다. RNN모형은 노드 수 100개의 경우 MSE와 RMSE가 최소가 되었다.

〈표 3〉 초모수의 설정

구 분	RNN	LSTM	GRU
입력변수	5	5	5
은닉 레이어수	2	3	1
신경망수	100	300	300
Epochs	200	300	300
Batch Size	10	10	10
활성화 함수	reLU	reLU	reLU
최적화 방법	ADAM	ADAM	ADAM
초기화 방법	zero state	He	He

2) LSTM

LSTM은 DNN과 마찬가지로 모형을 최적화하기 위한 초모수를 결정해야 한다. 본 연구에서는 은닉층은 3개, 테스트횟수는 300회, 배치사이즈는 10, 활성화 함수는 렐루 함수(Relu Function), 최적화(Optimizer)방식은 아담 (ADAM)알고리즘, 초기화(Initialization)방식은 He et al.(2015)이 주장한 방법을 사용하였다. LSTM모형은 노드 수 300개의 경우 MSE와 RMSE가 최소가 되었다.

3) GRU

본 연구에서 사용한 GRU모형의 은닉층은 1개, 활성화 함수는 렐루 함수(Relu Function), 최적화(Optimizer)방식은 아담 (ADAM)알고리즘, 테스트횟수는 300회, 배치사이즈는 10, 초기화(Initialization)방식은 He et al.(2015)이 연구한 방법을 사용하였다. GRU모형은 노드 수 300, 신경망수 경우 MSE와 RMSE가 최소가 되었다.

3. 분석결과

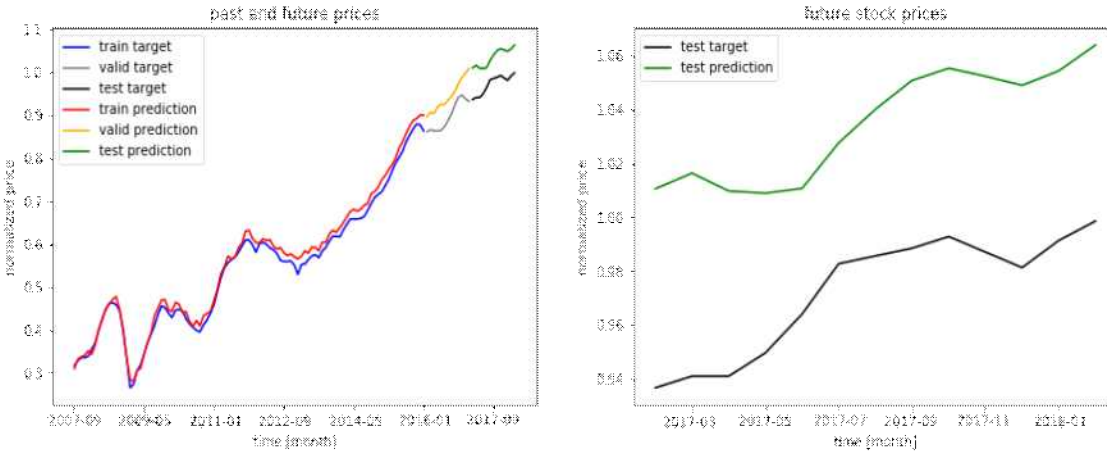
실험에 사용한 자료는 학습 데이터(Training), 평가 데이터(Validation)와 검증 데이터(Testing)로 각각 8:1:1의 비율로 나눴으며 결과는 <표 4>에 제시되어 있다. 학습 데이터에서는 GRU모형의 예측력이 가장 우수한 것으로 나타났다. 하지만 검증 데이터에서는 RNN모형, LSTM모형, GRU모형의 순으로 예측력이 가장 우수하였다. <그림 5>를 보면 검증 데이터에서 예측값과 실제 데이터가 다소 차이를 보이고 있으나 하락과 상승 추세를 보면 방향성이 거의 일치한 것으로 보인다. 특히, RNN모형과 GRU모형의 경우 LSTM모형보다 예측 성능이 우수하다는 것을 확인할 수 있다.

〈표 4〉 머신러닝 결과

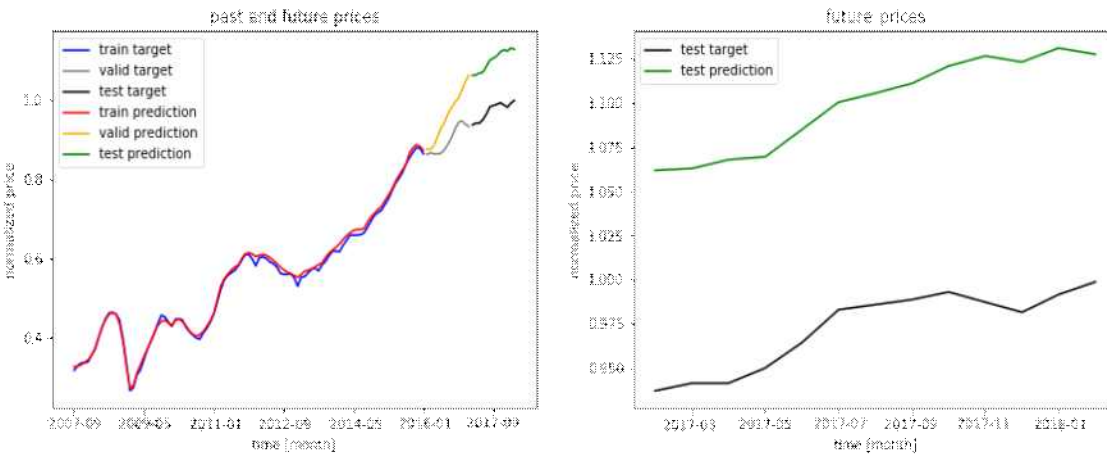
구분		Training Data	Test Data
RNN	RMSE	0.012727	**0.013370**
	MSE	0.000162	**0.000179**
LSTM	RMSE	0.015719	0.014728
	MSE	0.000247	0.000217
GRU	RMSE	**0.012146**	0.020958
	MSE	**0.000148**	0.000439

머신 러닝 알고리즘의 성능을 평가할 때 RMSE와 MSE 통계량을 사용할 수도 있지만 정량화 척도인 민감도(Sensitivity), 특이성(Specificity), 정밀도(Precision), 정확도(Accuracy)를 사용하기도 한다. 민감도는 모형이 실제 가격 상승을 가격 상승이라고 얼마나 맞게 예측하는지를 나타내며, 특이성은 모형이 실제 가격 하락을 가격 하락이라고 얼마나 맞게 예측하는지를 나타낸다. 정밀도는 모형의 예측이 얼마나 일정한지를 나타내며 정확도는 모형의 예측이 얼마나 실제와 가까운 지를 나타낸다(서윤범, 2018). 이와 같은 평가 기준들은 범주형

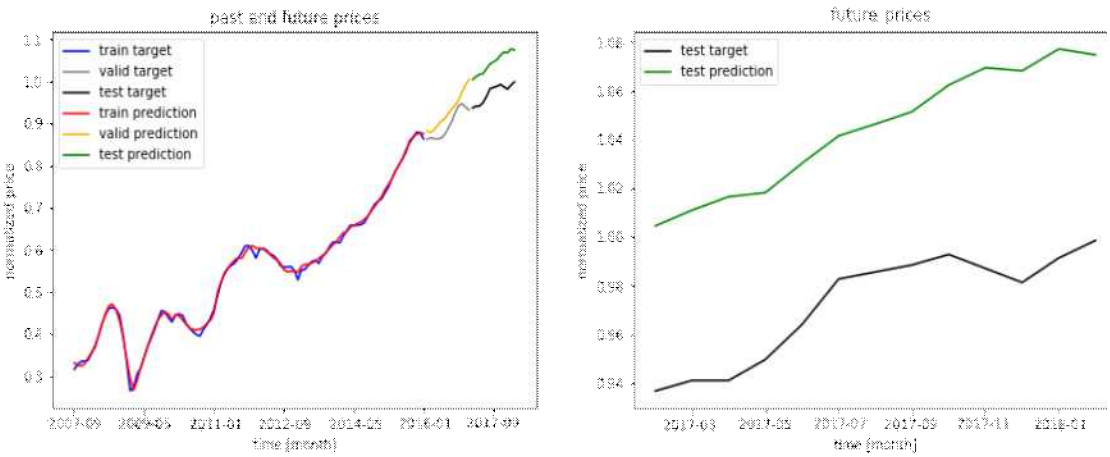
RNN 예측결과



LSTM 예측결과



GRU 예측결과



〈그림 5〉 모형 예측결과

자료를 목적으로 분류하는 머신러닝 알고리즘의 경우 많이 사용하지만 본 연구에서는 RNN, LSTM, GRU 세 개의 딥 러닝 모형으로 예측한 결과의 방향성이 실제 데이터와 비교할 때 어느 정도 정확한지 확인하기 위해 사용하였다.

모형의 성능을 평가하는 판단 기준으로는 정확도가 높을수록 모형의 예측력이 우수하다고 할 수 있다(이우식, 2017). 이에 따라 정확도를 기준으로 RNN, GRU > LSTM 순으로 높은 수치로 나타났다. 이를 통해 RNN모형과 GRU모형은 모두 LSTM 모형보다 더 나은 예측 정확도를 확인할 수 있다. RNN모형과 GRU모형을 비교하면 민감도, 즉 가격 상승을 예측할 때 GRU모형이 RNN모형보다 약간 우세한 것으로 나타났다. 다만, 검증용 데이터 구간에서 실제 가격 하락이 3건에 불과하여 정밀도 지표에서는 다소 낮게 나타난 점이 아쉬웠다. 향후 연구에서는 보다 많은 양의 데이터를 확보하여 이와 같은 한계점을 보완할 필요가 있다.

〈표 5〉 실험결과 평가지표

구분	Sensitivity (민감도)	Specificity (특이성)	Precision (정밀도)	Accuracy (정확도)
RNN	77.78%	25.00%	77.78%	66.67%
LSTM	44.44%	14.29%	66.67%	41.67%
GRU	88.89%	0.00%	72.73%	66.67%

Ⅴ. 결 론

본 연구는 빅 데이터를 활용하여 주택가격 예측을 딥 러닝 기법으로 적용하는데 의의가 있다. 또한, 시계열 예측에 적합한 딥 러닝의 알고리즘으로 평가받고 있는 simple RNN, LSTM과 GRU모형을 이용해 아파트 실거래가격지수와 금리, 가계대출금, 건축허가면적과 소비자물가지수로 구성된 시계열 자료를 이용하여 각 모형들 간의 예측력을 비교 평가하고 예측력이 우수한 모형을 판별하였다.

실증분석결과, 아파트 실거래가격지수에 대한 예측력의 성능을 RMSE 값 기준으로 평가하게 되면 학습 데이터의 경우 GRU모형의 예측력이 우수한 것으로 확인되었다. 반면 검증 데이터에서는 미미한 차이로 RNN모형의 예측력이 우수한 것으로 확인되었다. 또한, 딥 러닝 모형의 성능을 정확도로 평가하게 되면 RNN모형과 GRU모형의 정확도가 가장 높아 우수한 예측력을 확인하였다. 본 연구결과, 딥 러닝이 주택시장을 예측하는데 유용한 분석도구로 사용될 수 있을 것으로 기대된다.

정확한 주택가격의 예측은 정부, 개인과 기업의 입장에서 매우 중요한 사항이다. 정부는 주택시장의 동향을 정확히 파악해 시의적절한 주택정책을 수립 집행해 정책의 효율성을 높일 수 있고 개인과 기업의 입장에서는 시장의 상황을 파악하면서 각각의 상황에 맞는 합리적인 투자를 할 수 있기 때문이다.

딥 러닝은 비선형 알고리즘의 조합이기 때문에 함수 형태, 변수나 오차항의 분포, 변수 간의 상관관계 등의 어떠한 가정도 없이 자유롭게 적용될 수 있어 전통적인 회귀분석보다 적용범위에 유연성이 있고 예측력이 높다(배성완, 유정석, 2017). 또한 딥 러닝 모형은 입력변수가 불안전하거나 변동 폭이 넓은 경우에도 해석이 가능하고 데이터 수가 적거나 불규칙하더라도 반복학습을 통해 정밀한 산정이 가능하다. 그러나 딥 러닝 모형은 예측력이 우수하다는 장점 있으나, 출력변수와 입력변수 간의 관계를 명확하게 설명하지 못한다는 단점이 있다(정원구, 이상엽, 2007). 딥 러닝 모형의 예측력이 초모수 값에 민감하게 반응하나 최적 모형에 대한 객관적인 기준이 없어 최적 모형 선택에 작위성이 있을 가능성이 있다(이태형, 전명진, 2018). 또한 분석 자료와 변수 설정에 따라 딥 러닝 모형의 분석결과가 달라질 수 있다는 한계점이 있다(배성완,

유정석, 2018).

본 연구의 결과로부터 도출된 시사점은 딥 러닝을 적용한 예측력이 우수한 주택가격 예측모형이 개발되면 주택 가격변동에 적절히 대응하면서 예측 가능한 주택시장을 만들어 나갈 수 있는 바, 정부는 주택시장을 선진화하기 위해 인공신경망과 빅데이터를 접목한 딥 러닝 기법을 이용해 주택시장을 예측하고 진단할 수 있는 시스템 구축 및 개발을 할 필요성이 있다.

본 연구는 전국 주택가격 자료를 이용하였으나 지역별 주택자료와 더욱 다양한 거시경제변수를 이용해 연구를 확장하는 것은 추후 연구과제로 남긴다.

참고문헌

1. 김양훈 · 황용근 · 강태관 · 정교민 (2016) LSTM 언어모델 기반 한국어 문장 생성, 한국통신학회 논문지, 41권, 5호, pp.592-601.
2. 민성욱 (2017) 딥러닝(Deep Learning)을 이용한 주택가격 예측모형 연구, 강남대학교 대학원 박사학위 논문.
3. 배성완 · 유정석 (2017) 딥 러닝을 이용한 부동산 가격지수 예측, 부동산연구, 27권, 3호, pp.71-86.
4. 배영균 (2012) 주택공급의 가격탄력성과 주택가격 변동성, 대한부동산학회지, 30권, 1호, pp.67-84.
5. 서윤범 (2018) 비트코인 가격 등락 예측을 위한 딥러닝 모델 연구, 단국대학교 대학원 석사학위 논문.
6. 손경환 외 8 (1998) 금융위기하의 자산디플레이션 및 주택부문 현안과 대책 세미나, 국토, 통권199호, pp.51-59.
7. 신동하 · 최광호 · 김창복 (2017) RNN과 LSTM 을 이용한 주가 예측율 향상을 위한 딥러닝 모델, 한국정보기술학회논문지, 15권, 10호, pp.9-16.
8. 안성만 · 정여진 · 이재준 · 양지헌 (2017) 한국어 음소 단위 LSTM 언어모델을 이용한 문장 생성, 지능정보연구, 23권 2호, pp.71-88.
9. 오세경 · 최정원 · 장재원 (2017) 빅데이터를 이용한 딥러닝 기반의기업 부도예측 연구, KIF working paper, pp.1-113.
10. 이요섭 · 문필주 (2017) 딥 러닝 프레임워크의 비교 및 분석, 한국전자통신학회 논문지, 12권, 1호, pp.115-122.
11. 이용만 · 이상한 (2004) 강남지역의 주택가격이 주변지역의 주택가격을 결정하는가?, 국토계획, 39권, 1호, pp.73-91.
12. 이창로 · 김세형 (2018) 딥러닝 방식에 기초한 부동산 가격평가, 한국지역개발학회지, 30권, 4호, pp.179-201.
13. 이창로 · 박기호 (2016) 단독주택가격 추정을 위한 기계학습 모형의 응용, 대한지리학회지, 51권, 2호, pp.219-233.
14. 이태형 · 전명진 (2018) 딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구, 주택도시연구, 8권, 2호, pp.39-56.
15. 장병탁 (2007) 차세대 기계학습 기술, 정보과학회지, 25권, 3호, pp.96-107.
16. 전해정 (2017) 데이터 마이닝을 이용한 주택가격 결정요인에 관한 연구, 주거환경, 15권, 3호, pp.35-46.
17. 정원구 · 이상엽 (2007) 인공신경망을 이용한 공동주택 가격지수 예측에 관한 연구-서울지역을 중심으로, 주택연구, 15권, pp.39-64.
18. 조유나 · 김수현 · 송규원 (2017) 헤도닉 가격모형과 딥러닝을 이용한 주택가격예측 비교, 한국정보과학회 학술발표논문집, pp.1890-1892.
19. 최경호 (2016) Recurrent Neural Network를 이용한 자연언어처리, 강원대학교 대학원 석사학위 논문.
20. 허재완 (1991) 주택가격 상승률의 결정요인에 관한 실증분석, 국토계획, 26권, 2호, pp.141-151.
21. DiPasquale, D. · Wheaton, W. C., (1996) Urban economics and real estate markets, Englewood

Cliffs, NJ: Prentice Hall.
22. Mester, L. J., 1997, “What’s the point of credit scoring?”, Business review, 3:3-16.
23. 한국감정원 http://www.kab.co.kr/
24. 한국은행 경제통계시스템 http://ecos.bok.or.kr/