House Price Prediction

Sedat CAN
Middle East Technical University
e246579@metu.edu.tr

Abstract

Convolutional Neural Network, a popular branch of artificial intelligence, has been extensively developing to respond to the needs of house market by offering predictive models for estimation of house prices. Among many studies, this paper evaluates a simplified version of the CNN model used in (Ahmed, et al., 2016) and its dataset to get deeper understanding contemporary applications of CNN. As many studies rely on only textual attributes of houses, the model combines the data with visual inputs. Based on performance, the paper suggests that this combined model give better scores for the prediction of house prices and has potential to offer more.

1 Introduction

In recent years, brain inspired Convolutional Neural Networks (CNN) has become fastest growing area in artificial intelligence discipline due to the increase in computing capacities of computers. As many models, tools and methods have been developed and applied in different areas, the house price prediction attracts the attention of students of CNN since it entails well-structured models for processing both visual and textual data. Moreover, businesses in this competitive market demand these kinds of tools to facilitate transactions.

In this context, this paper would apply a CNN model to housing dataset containing visual and textual data in order to make a modest contribution to the literature. To this end, it would respond to the problem raised in paper titled "House price Estimation from visual and Textual Features" by Eman H. Ahmed and Mohamed Moustafa (Ahmed, et al., 2016). The paper deploys multilayer Neural Network (NN) model that estimates the house price by using textual and visual features and suggest that it outperforms other models in its predictive capacity.

Since there is not any publicly accessible code for the model, this paper tries to reconstruct the model in (Ahmed, et al., 2016) and evaluates its performance. Nevertheless, the techniques and methods used in Ahmed et al.,' model in the paper are not explicit and some tools, such as SURF, are licensed products. Therefore, the model reconstructed is simplified model, which do not give efficiently of the original model. Yet, from the results derived from the re-constructed model, the same argument can be concluded: the CNN model that provide betters results in comparison to the models that just processes textual data.

In this framework, the paper reviews current literature on House price estimation. Then, it analyzes the dataset used in the experiments. Finally, it provides the results of the experiments.

2 Related Works

As CNN is exponentially growing area, many studies have been done on house price prediction. The primary concern of studies on the issue is that as in real function of housing market, many features have been effecting prices the suggested models and datasets do not covers the facts of market. Visual data, in this respect, draws the attention of contemporary studies.

In this context, Khamis et al in (Khamis et al., 2014) analyzes two machine-learning models, i.e., Multiple Linear Regression (MLR) and Neural Network models for the prediction of house prices and compare their performance. The paper proposes that among these models, Neural Network model provide better result: higher R square value and lower Mean Squared Error value on New York dataset. The paper uses dataset containing 1047 houses from New York., composing of features including house price, living area, bedroom and bathroom numbers, lot size and age of house.[1] Yet, as Ahmed et al in (Ahmed et al., 2016) argues this work has a limitation that predicted price is not the actual price. This deficiency result from the fact that as many factors play role in prices of houses data could not contain all related features.[2]

[1] Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin, "Comparative Study On Estimate House Price Using Statistical and Neural Network Model," International Journal of Scientific & Technology Resarch 3, no 2 (December 2014)

[2] Eman H. Ahmed, Mohamed N. Moustafa, "House price estimation from visual and textual features," in https://arxiv.org/abs/1609.08399v1

Similarly, Hong et al in (Hong et al., 2020) try to pay attention on the mismatch between datasets and real operation of housing market. For writers traditional models for house price prediction relies on the premise that the attributes and characteristics of a house determines the value of the house. Yet as these models are highly deployed for their simplicity, they are also criticized for the selection and interpretation of features, which are assumed separable and constant. For them, price in the real world snd relationship between features are more complex; Random Forest model has power to predict house price better. In this respect, RF model enables to see the effect of each attribute on price in hierarchical structure and therefore does not require assumptions. Moreover, it has the advantage of high interpretability and can be trained easily and faster.[3]

As above-mentioned works give very reliable models of price prediction, Ahmed et al give more advanced model by deploying CNN.

3 Dataset Description

The housing price dataset is publically available on the address mentioned at the footnote.[4] The dataset contains 535 house samples collected in California State in the United State of America. The dataset has two kinds of information visual and textual. Visual data has the images of bedroom, bathroom, kitchen and exterior picture for each house sample. The resolution images are between 250x187 1484x1484. Textual data consists of the textual metadata of visual data that give information on number of bedrooms, number of bathrooms, area of the house, zip code and the price. The house prices are ranging from $22,000 to $5,858,000. The houses have maximum 7 bathrooms and 10 bedrooms.

4 Proposed Baseline System

As the aim of the project is to apply a simplified version of the model provided in (Khamis et al., 2014), it uses CNN model. In this regard, it exploits the textual data of the dataset and processes them in CNN. Later, it deploys a combined model of CNN, which both use visual and textual data. Due to low score of model, different preprocessing tool are used to increase the predictability power of model. First, after the test of simple convolutional tools such as resizing, BRISK is applied. BRISK was preferred over SURF because of copyright issue.[5] This tool enables to detect key points in image, which is applied to facilitate evaluations and comparison of the images. Secondly, histogram-equalization technique was applied on the images as proposed in (Khamis et al., 2014), which is used to adjusts the contrast of an image for easy detection. After these processes, the data is normalized.

4.1 Artificial Neural Network

A neural network consists of processing units and connections (weights) between units. The structure of neural network consists input layer, hidden layers, and the output layer, which has units connected with weights. Unit in a layer is connected with each units of next layer with a weight.[6]

For the model five layers are used for visual data and two for textual data. Then, two layers are added to process these data together. To evaluate performance of the CNN, coefficient of determination (R2) and the Mean Squared Error (MSE) metrics are used.

5 Experiments and Results

For the first CNN model, only attributes of images including bathroom, bedroom, and area were processed. 271 house samples, about 75 percent of the all data, was used to train model. The rest, 91 samples were allocated for the test. The performance of the model was that

MSE = 0.0014002646197979847
R2_score = 0.6081086826405384

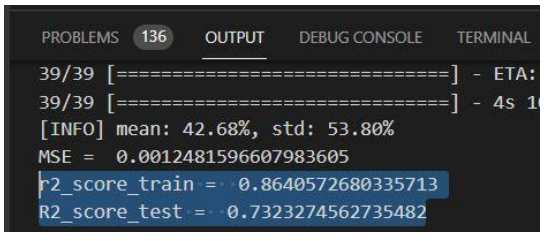
In the second experiment images were added to the data and applied to the model that processed textual data and images. The score was lower than the first model.

MSE = 0.003167823730852289
R2_score = 0.3206482609990108

In the third experiment, BRISK was added to model, which was set to detect about 200 key points in the image. Later, images were marked with flag for each detected point. The result was quite impressive. It improved score:

MSE = 0.001958306778934648
R2_score = 0.5800337301568398

In the last experiment, histogram equalization is added to the model,. The result of the experiment showed that histogram equalization improved R2 and MSE scores.
For instance, it gave results above:

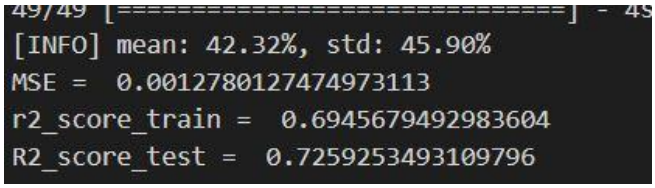


[3] Jengei Hong, Heeyoul Choi and Woo-sung Kim, "A House Price Valuation Based On The Random Forest Approach: The Mass Appraisal Of Residential Property In South Korea," International Journal of Strategic Property Management 24, no 3 (2020).

[4] https://github.com/emanhamed/Houses-dataset

[5] I did not know that SURF algorithm is licensed product.

[6] Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin, "Comparative Study On Estimate House Price Using Statistical and Neural Network Model," *International Journal of Scientific & Technology Research* 3, no 2 (December 2014)



In summary;

	First Exp.	Second Exp.	Third Exp.	Forth Exp.
MSE	0.001400	0.003167	0.001958	0.0012481 0.0012780
R2_score	0.608108	0.320648	0.320648	0.7323274 0.7259253

6 Conclusion

As the project aimed to apply model provided by Ahmed et al by using house price dataset, it dealt with many challenges such as replacing SURF[7] model with BRISK, finding a proper histogram equalization technique. Under these constraints, the usage of textual data in CNN model showed medium performance. However, adding raw visual data considerably deteriorated the performance of the model. However, after applying BRISK detector and histogram equalization, result of the model improved. As can be seen from the result of experiments, CNN model designed for textual and visual data provide better performance.

References

[Hong *et al*., 2020] Jengei Hong, Heeyoul Choi and Woosung Kim, "A House Price Valuation Based On The Random Forest Approach: The Mass Appraisal Of Residential Property In South Korea," *International Journal of Strategic Property Management* 24, no 3 (2020).

[Khamis et al., 2014] Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin, "Comparative Study On Estimate House Price Using Statistical and Neural Network Model," *International Journal of Scientific & Technology Research* 3, no 2 (December 2014)

[Ahmed et al., 2016] Eman H. Ahmed, Mohamed N. Moustafa, "House price estimation from visual and textual features"