Generating Apartment Rent Predictions from Google Street View Images

Piero Orderique
MIT
Cambridge, MA
porderiq@mit.edu

Chuqing Zhao
Harvard
Cambridge, MA
chuqingzhao@fas.harvard.edu

Abstract

Understanding house price could be critical for placed-based makings and urban dynamics. In this study, we propose a framework that combines street view image data as well as house attributes, locations, transportation, local socio-economic well being, and schools. In detail, we examine more than 1,800 apartments in the Boston Area by extracting deep features from street view images using CNN models. Combining image features with other attributes, we plan to build a price prediction model to help place-based decision makings.

1. Introduction

Boston consistently ranks in the top most expensive cities to rent a 1 bedroom apartment in the US. Understanding the house price not only provides insights for individual buyers but also could explore the social dynamics around this region. Regional-level measurements in urbanization are important for evaluating economic well-being, policy choices, and individual decision-making.

Moving beyond the substantive question, we hope to make contributions to research methods using deep learning models trained on street view images as well as interior images to capture the economic dynamics of house prices, gentrification, etc. Further, a model could help residents better understand what surrounding factors might influence the price of an apartment in the area. Although traditional methods like surveys and field studies could provide insights into societal changes, data collection is quite expensive, and analyzing images by humans is less efficient.

2. Related Work

Urban planners, economists, sociologists and architects have developed theories that connect the neighborhood's physical appearance and it environment as well as human dynamics. Schelling [11] and Grodzins [3] argues that neighborhoods in poor physical environments would get worse and reinforce the segregation between social groups. From the perspective of human capital, sociological theories such as invasion theory of Burgess [1] hypothesize that improvement in cities' appearance would become a concentric zone.

Street-level visual data and spatial image have been utilized in a variety of social scientific inquiries, such as gentrification [7], urban demographics and political ideology in neighborhoods [2], and urban housing [8]. To measure urbanization on a large scale, field surveys and observing the street-level images have been applied. Previous studies about urbanization research mainly use qualitative approaches by manually examining the neighborhood images and annotating the images based on human judgment [4].

More recent studies have adopted computer vision techniques to quantify urbanization through street-level image data. Ratti et al (2021) [8] extract deep features from the images and house photos and merge these features at two spatial scales: fine-scale point level and aggregated neighborhood level. Li et al (2017) [2] have an implementation for determining demographic makeup using Google Street View images, but its main focus is on basing predictions by focusing on extracting the types of cars shown in the image. Car detection requires an amount of human annotation, and Google street view images might blur the car labels leading to inaccurate estimation. Hystad and Larkin (2019) [9] collected street-level images to analyze public health; they used deep learning methods to capture green space environments.

3. Data

We plan to combine three separate sets of input data: housing attributes, street image data, and house interior image data around the Boston area. For street view images, we collect data using Google Street map images of several residential areas (using Google's Street View Static API); the house interior images have been collected from Trulia websites with additional census data including housing price by zip code, crime rate, demographics dataset that includes location information.

3.1. Street View Images

We have collected 1,876 images of residential areas in the Boston area using the Google Street View API. These were stored as 800x800 images with fixed camera perspective parameters of fov=100, heading=0, and pitch=0, so that the building perspectives remain consistent throughout the data. The Boston residential street names come from web scraping multiple sites and querying government databases on rent and crime statistics. This dataset was adapted from another project.

Figure 1. An example of Google Street View image used

3.2. House Interior Images

1,641 apartment interior images have been downloaded from the Trulia website - an important website for house renting. House interior images are important parts of the house information, which were uploaded by house wanderers to show the appearance from the inside.

Figure 2. An example of interior images scraped

3.3. House Attributes

This project also has included structural house attributes, such as rent price, location of property, the numbers of bath rooms and bed rooms, the crime rate around zip code areas, the commute convenience (i.e. percentage of nearby residents that commute by car), the number of schools (elementary schools, middle schools, and high schools), crime rates, and the number of shops (restaurants, groceries, and nightlife clubs). We incorporate those structural attributes that have been widely used in traditional hedonic pricing models [10]. Figure 3 shows a summary of the Rent column, which is the target number we are trying to predict.

	Mean	Std	Min	Max
Rent Price	2,930.50	1,278.43	650.00	24,000.00
Bed_num	2.42	1.12	1.00	9.00
Bath_num	1.54	1.35	1.00	9.00
Commute	43.15	19.20	15.00	83.00
School_num	9.88	5.69	1.00	31.00
Shop_num	577.09	480.59	16.00	905.25

Table 1. Summary statistics

3.4. Data Processing

The housing dataset has included some outliers and textual information. Categorical information was one-hot encoded and rent prices were discretized into bucket ranges. To address this issue, we first normalize the house attribute data by computing the minimum and maximum values, standard deviation of each variable for houses in the dataset. We subtracts the original value of each data point from its maximum value and divides the result by the range; therefore, each value is between zero to one.

4. Methodology

We propose a model which predicts the house price from three sources of data: street view images, house interior images, and house attributes. The project was divided into four stages: (1) visual representation and model creation, (2) model evaluation + feedback, (3) deployment + final analysis.

4.1. Understanding the Visual Feature

We experimented with a Convolutional Neural Network (CNN) architecture, which has achieved state-of-the-art performance on image tasks, such as object detection, image recognition and generative models. Pre-trained models, such as ResNet18 [5] and ResNet50 [6] have been used for producing feature maps to help the model identify objects in the scene and pass it through some connected layers to produce a final prediction. ResNet18 and ResNet50 are CNN-based architectures that have been used in various vision tasks. It can extract high-dimensional visual features on street view images based on pre-trained models. The

earlier layers capture the edges while the later layers capture more complex shapes in images. We take the values at the last convolutional layers as the output of image features. Here our preliminary results adopt the weights from the last layer in ResNet50:

Figure 3. An Example of Feature Map on First Layer ResNet50

4.2. Price Estimate Network

The first few models will perform poorly, so we iterate back to the model creation step (and even possibly the data collection step using data augmentation) to improve model accuracy and other metrics. In the initial model, we only include the street image view and use PCA to reduce high-dimensional visual representation into a 2-dimension vector in a linear regression model; however, PCA only could explain less than 10% variation. This motivated us to incorporate more data and move to neural networks that could capture non-linear relationships.

We trained the CNN-based model for the task of estimating apartment rental price for a training set of house interior photos, exterior photos and its attributes. The architecture for the price estimation network is shown in the figure 4. The rental price has been used as a ground-truth to train the model. Interior and exterior pictures of each room have been extracted the visual features from a pre-trained model, and then normalized into image vectors. House attributes data has also been concatenated with feature vectors, and passed to 4 fully connected dense layers. In this way, we obtain a representative vector for each apartment with both image data and metadata.

To evaluate our model, we have standard evaluation metrics like MAE and MSE as well as summary statistics (mean and standard deviation). The input is the representative vector of each apartment while the output is the estimated price of each house. We then compare the estimated prices with actual prices as the ground-truth. Losses are presented as MSE – the difference between predicted prices and estimated prices.

4.3. Deployment

Once acceptable metrics had been reached, the model was deployed as a web API for users to call on their own street view images around Boston.

5. Experiments

5.1. Ablation Studies

5.1.1 Hedonic Perceptron Model

We first estimate the house price using only the house attribute data and ignore all visual images. We train a fully connected neural network with 2 hidden layers to predict house prices on the basis of normalized house attributes. The first fully connected layer has 128 hidden nodes, and the second layer has 64 hidden nodes. The learning process has been optimized with stochastic gradient descent optimiser with learning rate of The loss function is computed by the mean squared error. The resulting mean squared error rate is 0.2892 and mean average error rate is 0.5179.

5.1.2 Incorporate Image data

In the next experiment, we train 3-layer CNN model with over 1,600 apartment data, and we concatenate the CNN output to a hedonic perceptron model that consists of 4 hidden layers. As some apartments do not have street view image or interior image, we remove the apartments with missing values. We split the data 70% for training and 30% is used for testing; the model's input are only images and the output is predicted house price. To compare which aspects of data would generate better predictions, there are two separate experiments: first we test on interior image data and house attribute data, and then we test on exterior image data and house attribute data. As result shows in Table2, MSE significantly decrease from 0.2892 to 0.004 after we incorporate the image data. The reduction in test MSE indicate that image data as inputs could help model make more accurate prediction of apartment price.

Compared to the exterior images, interior images could help the model make better predictions because the MSE for using exterior data is 0.0043 while the MSE for using interior data is 0.0046. One possible explanation could be that exterior images could contain more relevant features to predict prices, such as landscaping, overall condition of the exterior buildings, which could provide valuable insights into house price prediction. Further, the interior images could be more subjective because they could be modified and self-uploaded by house-hold owners.

5.1.3 Incorporate Interior and Exterior data

We consider a 3-layer CNN model by combining interior images, exterior images, and house attributes data. The

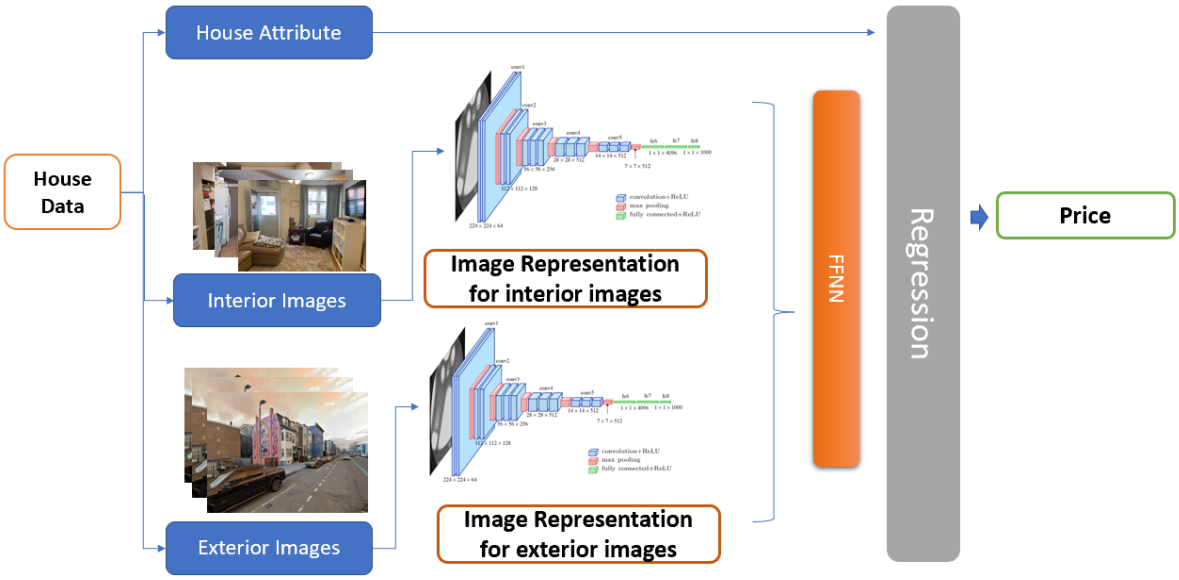

Figure 4. Model architecture for price prediction by taking in image and structural data

Method	Dataset	Test MSE	Test MAE
Fully connected Neural Network	Only Attributes	0.2892	0.5179
3-layer CNN	Interior data + Attributes	0.0046	0.0389
3-layer CNN	Exterior data + Attributes	0.0043	0.0374
3-layer CNN	Exterior data + Interior data + Attributes	0.0041	0.0383
3-layer CNN + ResNet18	Interior data + Attributes	0.0030	0.0371
3-layer CNN + ResNet18	Exterior data + Attributes	0.0029	0.0398
3-layer CNN + ResNet18	Exterior data + Interior data + Attributes	0.0029	0.0367

Table 2. Prediction results on the test set

model architecture has been kept as consistent with the previous section. However, the visual representation vectors for both interior and exterior images are obtained from 3 CNN layers; they are concatenated and then passed to in a fully connected layer. The experiment result shows that when the model incorporates various sources of data, the model's performance shows a slight improvement with test MSE of 0.0041 and MAE 0.0383. Note that the decrease of MSE while the increase of MAE indicates that although the model results in more accuracy overall, the predicted prices are slightly larger than actual prices with less outliers.

5.1.4 Transfer Learning: Using pre-trained Resnet

Instead of training the network built by ourselves, we pre-train the data on ResNet18. The final output dense layer has been removed while the convolutional layers remained intact with the original, using pre-trained weights on ImageNet to extract features from ResNet18 CNN layers. It results in increasing performance in predicting house price in all sets of data. For example, using 3-layer CNN with ResNet18 decreases the test MSE from 0.0041 to 0.0029, which indicates that a pre-trained model like ResNet18 could help model learning important visual cues in both exterior and interior images.

5.2. Qualitative Studies

As qualitative and validation examples, we have shown some data instances and compared the ground-truth results in Figure 5 and Figure 6. One example5 at 484 E 4th Street, MA around south Boston shows high accuracy of our method while the model gives lower estimated value for apartments that are unfurnished 6.

6. Further Study

Suggestions for further work include making use of Google's recently launched maps feature: Immersive View.

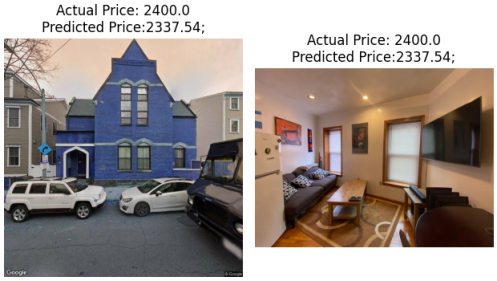

Figure 5. Sample Images with value predicted by our model

Figure 6. Sample Images with value predicted by our model

While satellite shots have been considered in previous papers, there has not been an exploration of multiple aerial shots from a close distance, which Immersive View may be able to capture.

Further, interior image data could be classified to different rooms using a place database such as MIT Indoor67 dataset. As we have evaluated whether and how the model would benefit from incorporating interior and exterior image data, the project could collect various types of image data in the future.

7. Conclusion

Estimating the rent of an apartment can be difficult without knowing the patterns and trends that can affect prices. While others have attempted to build robust models to accomplish this goal, many have overlooked taking into account image data or overlook the expressiveness of interior images. This paper attempts to fill this gap in the research by creating a double-layered CNN architecture for interior and exterior images that also takes into account housing attributes to create a model that can make more informed predictions.

8. Individual Contributions

8.1. Piero Orderique

(1) Setup Google Street API and write scripts for generating the exterior dataset. (2) Scrape and help clean the rent dataset to use for residential street locations. (3) Setup transfer learning architecture for base usage. (4) Design and implement initial baseline CNN models and evaluate finalized architecture. (5) Written and presentation work.

8.2. Chuqing Zhao

(1) Scrape and create interior images dataset. (2) Literature review and replicated methods in previous research. (3) Design and implement final CNN architecture as detailed in Figure 4. (4) Run and evaluate all dataset subsets of [Attributes, Interior Data, and Exterior Data]. (5) Written and presentation work

References

[1] Ernest W Burgess. *The growth of the city: an introduction to a research project*. Springer, 2008. 1

[2] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017. 1

[3] Morton Grodzins. Metropolitan segregation. *Scientific American*, 197(4):33–41, 1957. 1

[4] Daniel J. Hammel and Elvin K. Wyly. A model for identifying gentrified areas with census data. *Urban Geography*, 26(5):382–403, 2005. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision*, pages 630–645, 2016. 2

[7] Tianyuan Huang, Timothy Dai, Zhecheng Wang, Hesu Yoon, Hao Sheng, Andrew Y. Ng, Ram Rajagopal, and Jackelyn Hwang. Detecting neighborhood gentrification at scale via street-level visual data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–22, 2018. 1

[8] Yuhao Kang, Fan Zhang, Wenzhe Peng, Song Gao, Jinmeng Rao, Fabio Duarte, and Carlo Ratti. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 111:104919, 2021. 1

[9] Andrew Larkin and Perry Hystad. Evaluating street view exposure measures of visible green space for health research. *Journal of exposure science & environmental epidemiology*, 29(4):447–456, 2019. 1

[10] Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55, 1974. 2

[11] Thomas C Schelling. Models of segregation. *The American economic review*, 59(2):488–493, 1969. 1