Evaluating Real Estate through Image-Based Appraisal with Mask Region Convolutional Networks

Kondragunta Rama Krishnaiah, Professor, Department of Computer Science and Engineering, R K College of Engineering, Vijayawada - 521456, Andhra Pradesh, India, email: kondraguntark@gmail.com

Abstract

Real estate appraisal is a complex process. While current appraisal applications offer acceptable accuracy in estimating real estate prices, none of them use the real estate images in the appraisal process. Ignoring real estate images may cause inaccurate appraisal. Images show the condition of the interior and exterior and indicate damage in different sections of a house. Quantifying the condition and damages in real estate images need expert evaluation which is costly and time-consuming. In addition, existing automatic image recognition systems haven't addressed this problem yet. This paper aims to develop a novel real estate appraisal system which evaluates the property's interior and exterior condition using property's images. Due to the outstanding performance of region-based CNN (R-CNN), we used an enhanced R-CNN network called Mask R-CNN to evaluate the condition of each property image. While damage in real estate images might be hard to locate, Mask R-CNN is able to capture the finely detailed objects precisely. The system is expected to be an integral module to existing real estate appraisal systems to enhance the appraisal process.

Keywords: Image-Based Evaluation, Mask R-CNN, Real Estate Appraisal, Deep Learning, Visual Recognition, Damage Detection.

1. Introduction

Real estate sector is one of the most critical economic sectors and it represents nearly fifth of the total economic activity (You, Pang, Cao, & Luo, 2017). Assessing real estate properties is a multifaceted problem, as they depend on many different parametric and non-parametric features. Real estate values are time sensitive and influenced by many factors, which makes it challenging to predict property values using predefined functions (Xu & Gade, 2017). Famous real estate applications such as Zillow, Redfin, Trulia and Realtor use automated valuation methods (AVM) to estimate properties prices by defining proprietary formulas that rely on many features such as economic index, house age, history trade and neighborhood environment. These applications are reliable estimation tools as they have direct access to Multiple Listing Services (MLSs) which contains a detailed list of all properties. (Poursaeed, Matera, & Belongie, 2018). While these automatic estimation methods give acceptable estimation accuracy with mean error rate of 8% (Poursaeed et al., 2018), sometimes they can be highly inaccurate because they do not consider the interior, exterior and cosmetic conditions of the property in their estimation. On the other hand, home buyers usually use property images first to estimate the property value which has been ignored by real estate applications in their estimation process (Di, Sundaresan, Piramuthu, & Bhardwaj, 2014). To involve images in real estate appraisal process, real estate inspectors are needed to evaluate the condition of real estate images. In addition, none of current image recognition systems tackle the problem from estimating the condition point of view. An automatic real estate image-based evaluation system would provide a clear assessment of property images conditions. It should be able to identify the different sections of a property and the basic components in each scene (e.g., bathroom section has bathtub and sink). Moreover, the system should notice fine details in walls and floor such as water damage, stains, cracks and scuff. Recently, deep learning convolutional neural networks have achieved a state of art performance in image recognition and instance segmentation.

One of the most recent convolutional networks is Mask R-CNN developed by Facebook has the ability to reach pixel level segmentation of objects. Therefore, small objects and key points could be detected easily. This paper proposes a real state images appraisal system that identifies damage in real estate properties and gives evaluation based on damage condition. The system automatically provides multi-label description for each property image. This description is obtained from expert inspector reports including the scene, exterior and interior conditions, and the severity of damage. The system uses Mask R-CNN for objects and damage detection, and Kernel Canonical Correlation Analysis (KCCA) for annotations of each damage in an image. The output of the system would help main real estate applications such as Zillow and Redfin to reach more accurate appraisal by including images in the appraisal process. In the following sections, we review prior research on real estate appraisal using different machine learning techniques. Then, an introduction to Mask R-CNN is given, followed by the description of the proposed system. The architecture of the system is described, and the algorithms used for real estate image appraisal are developed. Finally, the experiment is clarified, and future research directions are identified.

2. Background

No doubt the visual features of a property are key factors in estimating its market value. While famous real estate websites do not disclose the formulas used in price estimation, their algorithms are prone to error as they do not consider the impact of property images on its market value. However, very few research tried to include images in real estate appraisal. For example (Poursaeed et al., 2018) developed a deep learning algorithm to detect the luxury level of a property using interior and exteriors images. They also developed a framework for automating the value assessment process using images and other parameters. The authors trained a set of DenseNet networks with Maxpooling layers to estimate the luxury level of each room and to estimate the price of a property. However, the luxury level of the house is not sufficient to evaluate the real condition of a house. (You et al., 2017) proposed a framework for real estate appraisal based on neighborhood prices and exterior visual appearance.

They used LASSO price index with Deepwalk to get the price of similar houses in the neighborhood. They also used the visual features of a house from a pertained GoogleNet network and fed them into Bidirectional LSTM Recurrent Neural Network (B-LSTM RNN) for price estimation. The framework used the exterior visual similarity between houses in the same neighborhood as the main factor in estimating the price. Unfortunately, they didn’t include interior condition as a house could be severely damaged even if the exterior photos from google maps are in a perfect condition.

The price estimation of real estate is considered a regression problem as the market value of a house is the dependent variable and the house characteristics such as area are the independent variables (Meese & Wallace, 1991). Most of real estate appraisal systems are based on regression analysis and machine learning techniques. One of the traditional methods for property market valuation is the “comparable” model, a form of k-nearest neighbors regression (Pagourtzi, Assimakopoulos, Hatzichristos, & French, 2003). This method estimates the value of the property based on similar properties within the same area. Another popular method is the hedonic modeling technique which supposes that the relationship between the price and independent variables is a nonlinear logarithmic relation (Meese & Wallace, 1991; Sheppard, 1999). Before deep learning was used for real estate appraisal, Fuzzy logic was used extensively (Bagnoli & Smith, 1998) (Pagourtzi et al., 2003). It calculates the degree of membership of each house to estimate its belonging to certain group of similar houses. Additionally, Neural Networks, Genetic algorithms and Support Vector Machines were used as different regression models or as ensembles for real-estate appraisal (Lasota, Mazurkiewicz, Trawiński, & Trawiński, 2010) (Kempa, Lasota, Telec, & Trawiński, 2011). The proposed system targets adding images to the real estate appraisal process using Mask R-CNN (He, Gkioxari, Dollár, & Girshick, 2017), a powerful deep learning algorithm for instance segmentation. Mask R-CNN have been used in many image recognition applications that requires precise instance segmentation such as car collision detection (Dwivedi, 2018), robot surgery (Kong, 2018), and segmenting nuclei in microscopy images (Waleed, 2018).

3. Proposed Real Estate Image-Based Appraisal System

Scene components: Real estate properties, especially the residential ones, consist of the same sections and components. The knowledge layered graph of a property sections is shown in Figure 1. At first, the proposed system should be able to recognize the scene whether it is an exterior or an interior scene. Then, it should detect what section represents the scene (e.g., bathroom, kitchen, or closet). Also, it should recognize the main components of the section such as walls and floors. Finally, it should be able to locate any damage and estimate its severity. For example: if a kitchen wall has a damage, it should estimate how severe is that damage. Samples of different level of severity is shown in Figure 2. In addition, different types of floor, walls, heating/ cooling units must be taken into consideration (e.g., carpet or wood floors). However, the system should ignore the furniture and appliances in the estimation of house condition. Assuming that the system will be fed with images of the full property, the system output will be the estimation of damage or the severity level of each room. The real estate appraisal system would integrate the estimated level of damage severity developed by our system in the overall appraisal for a property. While the price appraisal formula is beyond the scope of the paper, it can be determined by integrated appraisal systems available from service providers such as Realtor and Zillow. As shown in Figure 2, a kitchen is a section consisting of many components. Each component in the section will be evaluated for level of severity. The system will evaluate both sections and components. For example, to determine if wall conditions in a kitchen are in critical condition or the entire kitchen is in critical condition, the system should notice the fine details and small objects from the images. In other words, the system should reach precise level of small object segmentation to locate damage.

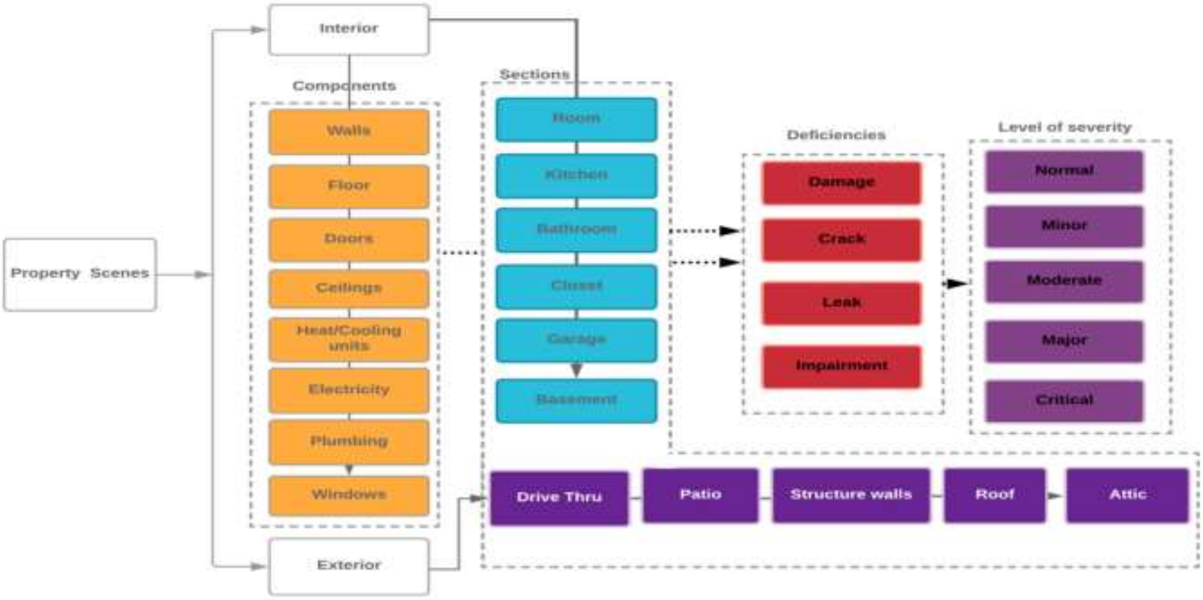

Figure 1: Knowledge layered graph of the real estate visual design

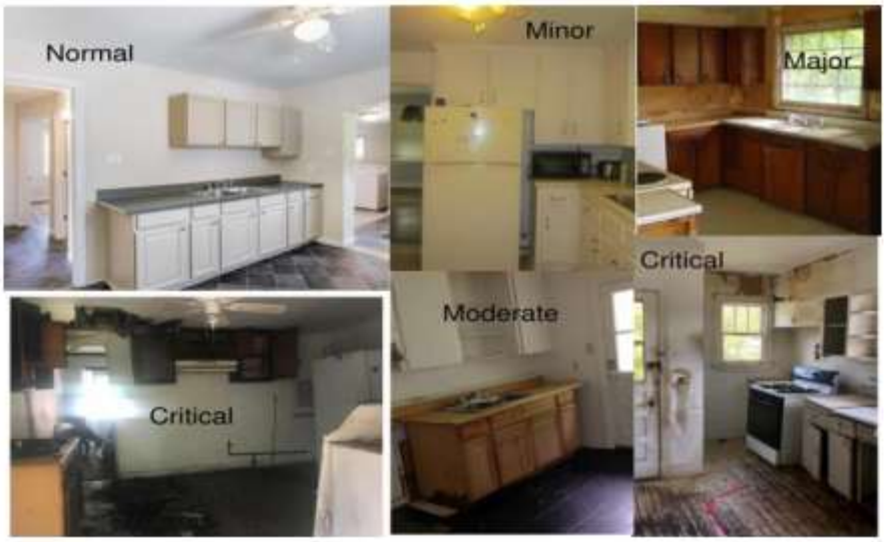

Figure 2: Kitchen in different levels of severity.

Mask R-CNN: Mask R-CNN is an extension for Faster R-CNN that is able to reach fine granularity segmentation of objects using instance segmentation. Instance segmentation is an attached binary mask of the detected object with every bounding box. The bounding box of the detected object may contain parts of other objects. Therefore, the mask allows more precise pixel wise location for objects specially if these objects are small heterogeneous shapes such as damage and cracks. The difference between bounding box and instance segmentation is shown in Figure 3. Mask R-CNN consists of a backbone network and a head network (He et al., 2017) as shown in Figure 4. The backbone network is used for feature extraction over an entire image. The backbone network is a faster Region-based CNN (R-CNN) network. The head network is for bounding-box and mask recognition (classification and regression), in other words the head will be giving labels to the identified objects and extracting the surrounding bounding box of the recognized objects. The Mask R-CNN has a parallel fully connected network for mask prediction. The backbone and the head architectures are discussed in the next section in detail.

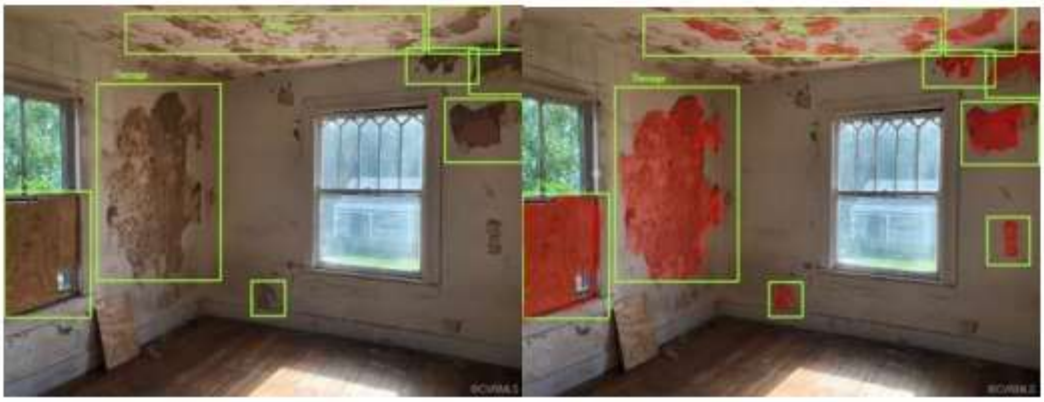

Figure 3 a: bounding box segmentation b: instance segmentation

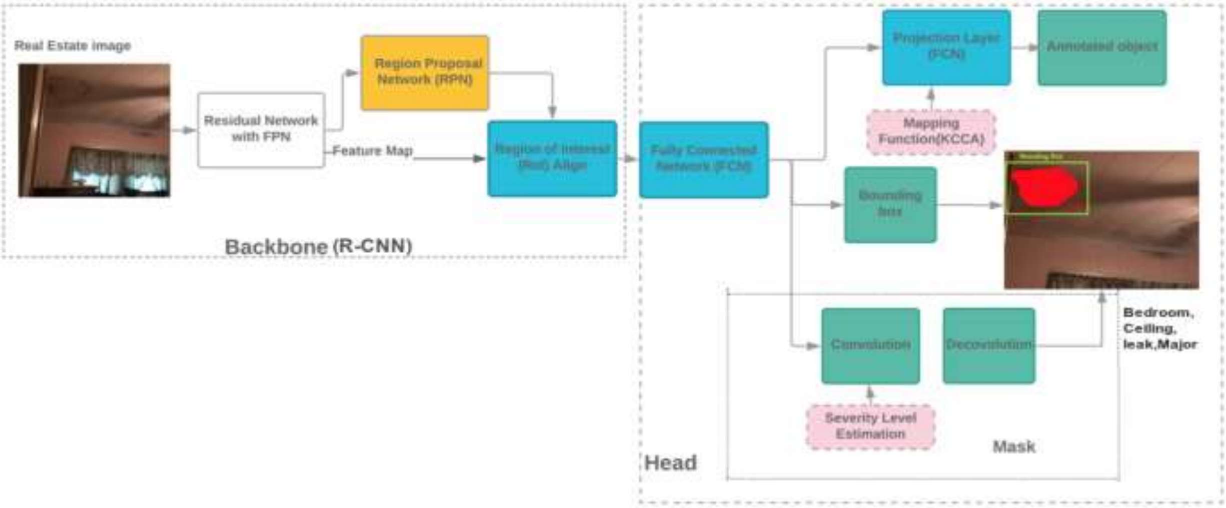
Real Estate image
Residual Network with FPN
Region Proposal Network (RPN)
Feature Map
Region of Interest (RoI) Align
Backbone (R-CNN)
Fully Connected Network (FCN)
Projection Layer (FCN)
Annotated object
Mapping Function(KCCA)
Bounding box
Bedroom, Ceiling, leak,Major
Convolution
Deconvolution
Severity Level Estimation
Mask
Head

Figure 4: System network architecture

System Backbone Architecture: The main component in the backbone network is the faster R-CNN which consists of a base Residual Convolutional Network (ResNet) and Region Proposal Network (RPN) network as shown in Figure 4. The Residual convolution networks (He, Zhang, Ren, & Sun, 2016) achieved very promising results in many image recognition applications (Dai, He, & Sun, 2016). The RPN network creates a set of regions that include the objects detected. These regions are called the Regions of Interests (RoIs). For the base residual network, we selected Feature Pyramid Network (FPN) to build a pyramid-like features architecture based on their scale (Lin et al., 2017). The FPN has a bottom-up operation for feature extraction. As layers go up, higher-level structures are detected and the semantic value of features increases. However, the image resolution decreases as layer increases. Therefore, another a top-down operation is added to reconstruct higher resolution layers from a top semantic rich layer. The output of the Residual Network with FPN (ResNet-FPN) is the feature maps of an image in a pyramid-like hierarchy. These features map along with the regions detected from RPN are fed into ROI Align module. ROI Align is used to align the detected regions with its original locations so that they remain with the same size and exact position within an image to avoid overlapping of objects.

The Head Architecture: The first component in the head architecture is a Fully Connected Network (FCN) which is known for semantic object segmentation and detection. The first output of the FCN is fed into a projection layer. This projection layer is another FCN that functions as CNN regressor for object annotation. This layer allows the mapping function to map the object detected to a word embedding vector. The second output is a boundingbox offset of the detected objects as shown in Figure 4. The bounding box is fed to the mask network which is a parallel branch of convolutional and deconvolution networks. The mask network simply creates a binary mask of pixels to locate the object in the bounding box. At first the mask network decompresses the image to 1/32th of its original size and predict the class at this level of granularity. Then, the network uses up sampling and deconvolution layers to predict the object at different level of granularity until the image returns to its original dimensions. In a nut shell, Mask RCNN combines Faster RCNN and FCN. It creates a pixel wise localization mask from the object bounding box. The Mask RCNN has three outputs. The object annotation, the bounding box, and a binary mask for the object inside the box. The loss function for the Mask R-CNN is the sum of classification loss $Lcls$, generating bounding box loss $Lbox$, and generating the mask loss $Lmask$ which is binary sigmoid in this case.

Annotation of Real Estate Images: Each image is given multiple annotation including the overall condition of each section, damage in each component, and the degree of damage severity as shown in Figure 5. Since we are using multi-labels for each image, Kernel Canonical Correlation Analysis (KCCA) was selected for annotation of images (Murthy et al. 2015). In the system head, the KCCA is embedded in the system as a mapping function in through the projection layer of the FCN to produce a set of tags for each object detected. At first, a tool called Word2Vec converts each label to a high-dimensional real-valued feature vector. The output of Word2Vec is the word embedding vector Y. Since the scope of the vocabulary used in the system is limited to real estate related terms, the size of the embedding vector Y won't exceed 100. Increasing the size of Y will increase the system complexity. So, we should compromise between accuracy and complexity (Allen 2017). Then, the visual features from the backbone network X were mapped to a high dimensional feature space Wx using a positive definite kernel function Kx. The word embedding vector Y is also mapped to a high dimensional feature space Wy using a positive definite kernel function K . KCCA aims to maximize the correlation coefficient distance between Wx and Wy projections and the solution is the top eigenvectors or tags assigned to the image (tags are also ranked according to their frequency in the training dataset).

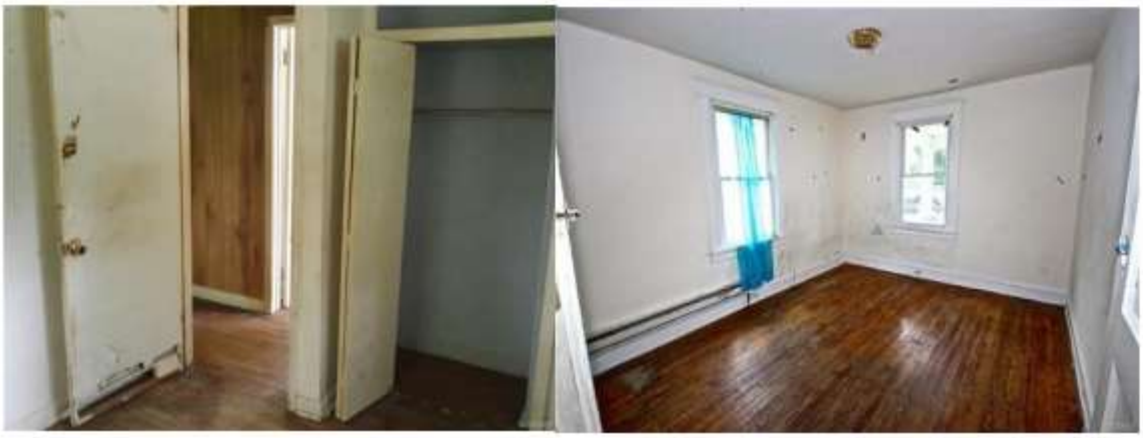

Figure 5 a: Room condition is critical, door damage critical, floor damage is major, b: Room condition is major, wall scuff moderate, floor damage major.

Level of severity estimation: In order to estimate the value of a property, the severity of damages should be estimated carefully. Critical damages in a property could reduce its value significantly even if it has a great location. However, level of severity is calculated at the mask level after the convolutional network is applied for mask detection as shown in Figure 4. The mask gives precise localization and representation for each class so a deficiency could be detected accurately. The level of severity l is calculated twice on the section level and component level. For e.g., the bathroom condition will be critical because the walls and windows are of critical conditions. Which means that if the damage in one component is critical then the entire section will be critical otherwise we sum the damage in each component to be the severity level of a section. The formula for calculating the level of severity is represented as: Assuming that each image representing a section, an importance weight of wci is assigned for each component Ci= $\{C1, \cdots ,x\ \}$. For any one or more deficiencies dij = $\{di1, \cdots , dim\}$ detected, a severity weight of $wdij$ is assigned. Additionally, $kdij$ represents the severity of each deficiency, and V the critical severity threshold which is estimated priori. The severity level l of a section is then:

$$l = \begin{cases} \max_i (wc_i \sum_{j=1}^{m} kd_{ij} . wd_{ij}), & l \geq V \\ \sum_{i=1}^{x} wc_i . \sum_{j-1}^{m} kd_{ij} . wd_j , & l < V \end{cases}$$

The section is assigned critical severity level if the damage in one component exceeds the threshold level. Otherwise, the section severity level is the sum of each component severity level. For example, if the ceiling in a room is assigned the highest importance weight wc , and it has deficiencies of high damage and leak severity $\sum kdij . wdj\ m\ j{-}1$. This would make the entire room in critical condition even if the rest of the room components are in good condition.

4. Implementation

Datasets: There are many powerful datasets for object and scene recognition such as Google ImageNet 3, OpenImageV4 (Papadopoulos, Uijlings, Keller, & Ferrari, 2017), LSUN (Yu et al., 2015), and MS-COCO (Lin et al., 2014). Some datasets are more specific to scenes and interior recognition such as place, and Houzz (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014). However, none of them are specified for damage evaluation. Therefore, we used mixed datasets in building the system training dataset. The main dataset used is the one used in (Poursaeed et al., 2018). They used more than 140k of images from Zillow and Houzz. In addition, we use property images in different conditions from MLS public records.

Preprocessing and Training: Any deep learning network needs huge number of annotated images to generalize and give acceptable recognition rate. This is challenging because the lack of annotated images for the specific purpose of evaluating house condition. To ease the training process, we will transfer learning from a pertained Mask R-CNN2 with MS-COCO dataset. COCO is used for many segmentation and captioning tasks and it has more than 1.5 millions of object instances. However, we replaced the last layer of the pertained network with MaxPooling and projection layers. The MaxPooling layer is to reduce the output dimension and projection layer to perform object annotation. Preprocessing of images is essential in order to get accurate recognition. Many images in the dataset are of low resolution. Therefore we will use waifu2x-multi1 API, a deep learning algorithm for enhancing image resolution and it is considered the best upscaling solutions for the small and noisy images. (Dong, Loy, He, & Tang, 2016). For contrast adjustment we used Contrast Limited AHE (CLAHE) (Reza, 2004), and for adjusting the brightness we adjust the level of each RGB component. All images are resized such that their scale (shorter edge) is 800 pixels as in Poursaeed et al. (2018).

5.Experiments

The experiments would be in the form of ablation analysis to observe which settings would give the best results. At the beginning, we will use backbone of residual networks with FPN since they achieve better performance than VGG and inception networks (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017). Therefore, we would use (ResNet-50-FPN) of 50 layers and then we will use 101 layers (ResNet-101-FPN) and we will watch which will achieve the best performance in terms of time required for training. For the used dataset, will use the images from MLS at first, then we will include images from both Houzz and Zillow. In addition, we will try to use the network with and without MaxPooling layer to see if reducing the dimension would affect the accuracy. In order to get better results, we will increase the batch size instead of decreasing the learning rate to decrease the time required for learning according to recent study by (Smith, Kindermans, & Le, 2017). This would be beneficial for our study as we could train the network faster.

6. Conclusion

Real Estate appraisal is a convoluted process that involve many parameters and considerations. Despite recent enhancements in automatic real estate applications, these applications do not consider the evaluation of property images in the appraisal process. A property's exterior and interior condition could greatly affect the property price. This research presents a system for evaluating property images condition using the advanced deep learning Mask R-CNN algorithm. In addition, the system uses Kernel Canonical Correlation Analysis (KCCA) to give each image multiple annotations including the scene, the section components, and the level of severity. The future work will include the instantiating the system based on the experiments setups provided earlier and fine-tuning of the model to be used as regression model to property images evaluation and finally integrate the system with popular real estate appraisals models to see if adding the image appraisal module will enhance the entire appraisal process. In addition, augmenting the system to different building codes in different countries is an open research area. Long term research objective is to develop a cognitive system that can adaptively evaluate the condition of real estate images based on different building materials used in different countries.

Reference:

[1]. Allen, G. (2017). Word Vector Size vs Vocabulary Size in word2vec. Retrieved from http://www.grega100k.com/wordvector/2016/03/17/words-vs-vocabularies.html

[2]. Bagnoli, C., & Smith, H. (1998). The theory of fuzz logic and its application to real estate valuation. Journal of Real Estate Research, 16(2), 169-200.

[3]. Dai, J., He, K., & Sun, J. (2016). Instance-aware semantic segmentation via multi-task network cascades. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[4]. Di, W., Sundaresan, N., Piramuthu, R., & Bhardwaj, A. (2014). Is a picture really worth a thousand words?:- on the role of images in e-commerce. Paper presented at the Proceedings of the 7th ACM international conference on Web search and data mining.

[5]. Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence, 38(2), 295-307.

[6]. Dwivedi, P. (2018). Ultimate Guide: Building a Mask R-CNN Model for Detecting Car Damage. Retrieved from https://www.analyticsvidhya.com/blog/2018/07/building-mask-r-cnn-model-detecting-damagecars-python/ Hardoon,

[7]. D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural computation, 16(12), 2639-2664.

[8]. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Paper presented at the Computer Vision (ICCV), 2017 IEEE International Conference on.

[9]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

[10]. Kempa, O., Lasota, T., Telec, Z., & Trawiński, B. (2011). Investigation of bagging ensembles of genetic neural networks and fuzzy systems for real estate appraisal. Paper presented at the Asian Conference on Intelligent Information and Database Systems.

[11]. Kong, C. C. (2018). Mask R-CNN for Surgery Robot. Retrieved from https://github.com/SUYEgit/SurgeryRobot-Detection-Segmentation

[12]. Lasota, T., Mazurkiewicz, J., Trawiński, B., & Trawiński, K. (2010). Comparison of data driven models for the valuation of residential premises using KEEL. International Journal of Hybrid Intelligent Systems, 7(1), 3-16.

[13]. Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., & Belongie, S. J. (2017). Feature Pyramid Networks for Object Detection. Paper presented at the CVPR.

[14]. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Paper presented at the European conference on computer vision.

[15]. Meese, R., & Wallace, N. (1991). Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices. Real Estate Economics, 19(3), 308-332.

[16]. Murthy, V. N., Maji, S., & Manmatha, R. (2015). Automatic image annotation using deep learning representations. Paper presented at the Proceedings of the 5th ACM on International Conference on Multimedia Retrieval.

[17]. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. Journal of Property Investment & Finance, 21(4), 383-401.

[18]. Papadopoulos, D. P., Uijlings, J. R., Keller, F., & Ferrari, V. (2017). Extreme clicking for efficient object annotation. Paper presented at the Computer Vision (ICCV), 2017 IEEE International Conference on.

[19]. Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. Machine Vision and Applications, 29(4), 667-676. Reza,

[20]. A. M. (2004). Realization of the contrast limited adaptive histogram equalization (CLAHE) for realtime image enhancement. Journal of VLSI signal processing systems for signal, image and video technology, 38(1), 35-44.

[21]. Sheppard, S. (1999). Hedonic analysis of housing markets. Handbook of regional and urban economics, 3, 1595-1635.

[22]. Smith, S. L., Kindermans, P.-J., & Le, Q. V. (2017). Don't Decay the Learning Rate, Increase the Batch Size. arXiv preprint arXiv:1711.00489.

[23]. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Paper presented at the AAAI.

[24]. Waleed. (2018). Mask R-CNN for Object Detection and Segmentation.

[25]. Xu, H., & Gade, A. (2017). Smart real estate assessments using structured deep neural networks. Paper presented at the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). You,

[26]. Q., Pang, R., Cao, L., & Luo, J. (2017). Image-based appraisal of real estate properties. IEEE Transactions on Multimedia, 19(12), 2751-2759.

[27]. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365.

[28]. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. Paper presented at the Advances in neural information processing systems. Allen, G. 2017. "Word Vector Size Vs Vocabulary Size in Word2vec." from http://www.grega100k.com/wordvector/2016/03/17/words-vs-vocabularies.html

[29]. Hardoon, D.R., Szedmak, S., and Shawe-Taylor, J. 2004. "Canonical Correlation Analysis: An Overview with Application to Learning Methods," Neural computation (16:12), pp. 2639-2664.

[30]. Murthy, V.N., Maji, S., and Manmatha, R. 2015. "Automatic Image Annotation Using Deep Learning Representations," Proceedings of the 5th ACM on International Conference on Multimedia Retrieval: ACM, pp. 603-606