Room Layout Estimation with Pretrained Convolutional Neural Networks and Color Augmentation

Derek Chung
Computer Science and Music Student
Stanford University
dchung22@stanford.edu

Abstract

With companies such as Tesla rising in value, so too does the use of computer vision. Demand for autonomous machines may be in the near future, which means models need to identify and navigate through cluttered environments. A particular subset of Computer Vision that deals with these matters is called scene understanding - the features and properties computers are able to extract from images, videos, or streams. Scene understanding itself can be divided into subsets, such as layout estimation, scene classification, and object detection. This paper will focus on layout estimation, where I will attempt to estimate the layout of a given room given a single image. My method consists of a fully convolutional deep neural network. Using the LSUN room layout dataset, I will generate a room layout and apply an algorithm to smooth the edges. The main purpose of this paper is to compare changing image qualities (saturation, contrast, hue, brightness) to data augmentation techniques used in other papers. My model achieves state of the art results against previous benchmarks.

1 Introduction

One of the most prominent trends that have grown in the past decade is computer-reality interaction. Alexa and Siri, for example, are two programs that interact with language and the human voice. Another example can be seen in devices such as the Apple Watch; machine learning methods are used to construe the meaning of human motions, such as identifying a single step based on input received from different sensors. The most relevant example to this paper is Tesla, using a multitude of different computer vision practices to identify properties - streets, pedestrians, and obstructions - using various different instruments in order to produce safe autonomous vehicles.

Room estimation layout is a subset of computer vision where one tries to estimate the key geometric points of the images of a room. Given an input image of the interior of a room, my model outputs a matrix M, where each entry in the matrix corresponds to a pixel in the input image. The elements in M will be such that for all $m \in M$,

$$m \in 0, 1, 2, 3, 4 \tag{1}$$

such that each number corresponds to a specific label, such as the center wall, top wall, bottom wall, right wall, or left wall. Figure 1 denotes the input and output in a graphical sense.

2 Related Works

Scene understanding is important for tasks such as robot navigation and augmented reality. Before Convolutional Neural Networks, there existed some methods for indoor scene estimation. Some of the more prominent methods include line segment estimation[2] by Lawrence Gilman Roberts and geometric context estimation[3] by Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Roberts devised an algorithm that finds points in an image that are likely to be on a line, and then uses a

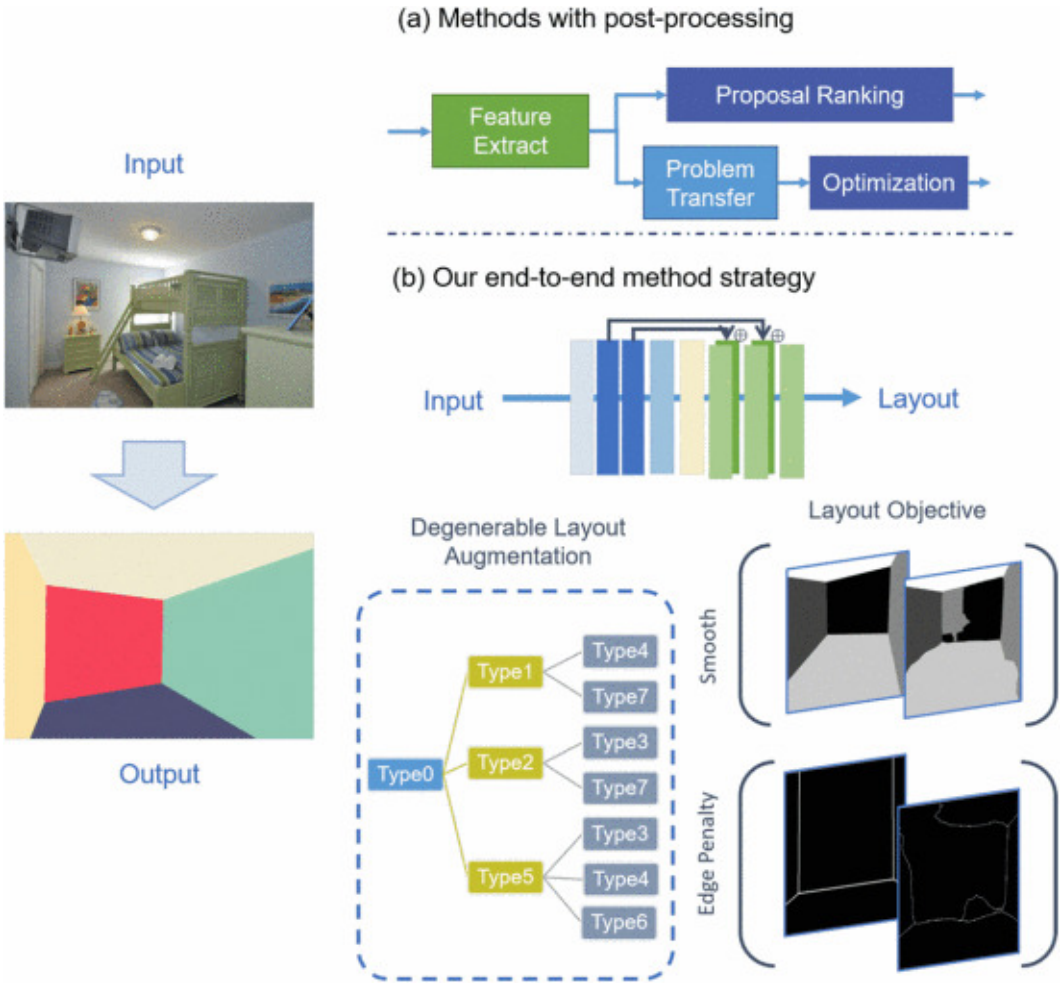

Figure 1: Model for Indoor Scene Estimation from a Single Image[1]. Each region of the output can be represented as an integer value in a matrix.

mean squared error to construct a line representing a geometric edge. The points are determined by converting the image into a differential picture. Felzenszwalb and Huttenlocher separates an image into regions using a graph-based representation and a segmentation algorithm. Neither of these methods utilize the deep learning I go into later on. Rather, mathematical properties of graphs and images nullify the need to train a model. Furthermore, Convolutional Neural Networks weren't developed during Roberts's time.

Machine Learning began to pave the way for many computer vision problems. Structured learning[4] models the environment structure using information from local features. However, this requires one to extract several hand crafted features. The winner of the LSUN Room Layout Challenge 2015[5] used a deep neural network to identify regions and a vanishing point algorithm to smooth out edges. This method, called DeLay, inspired multiple other two staged approaches involving a deep learning network followed by post-processing optimization. Instead of vanishing point optimization, a deconvolution network[6] applies guassian blur to the regions between the layout map and estimates the boundaries between the surfaces with a sigmoid activation function.

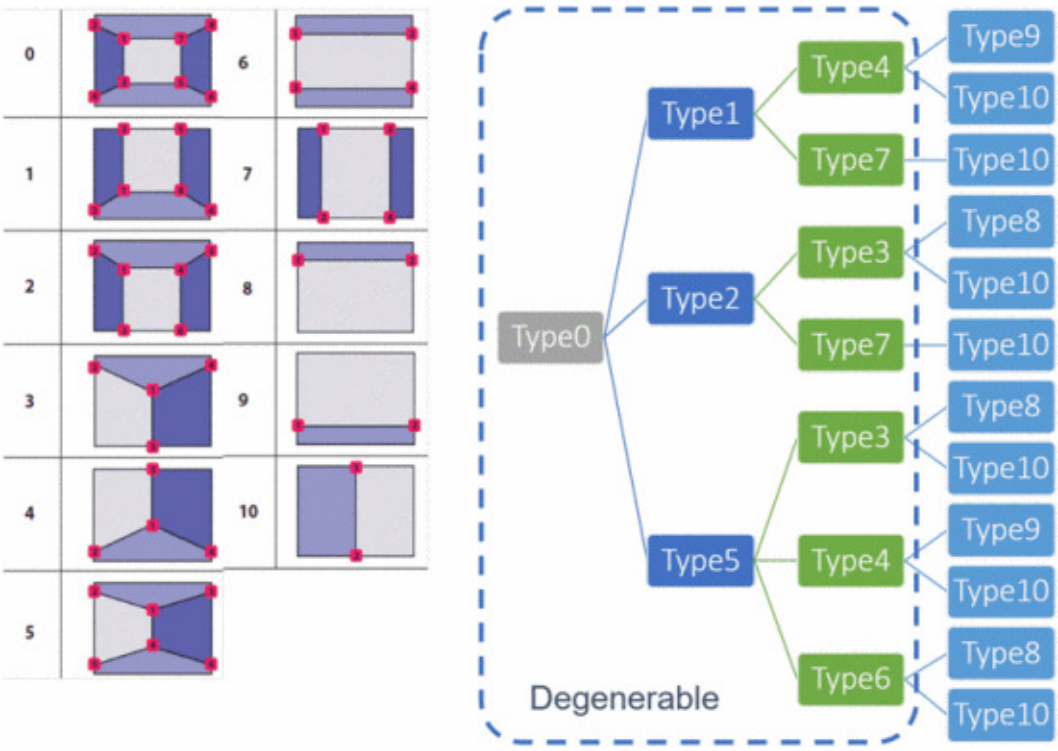

Figure 2: Room layout types and layout degradation pathways. Figure taken from other publication[1].

3 Dataset

My model uses the LSUN Room Layout Dataset, containing 4000 training examples, 394 validation examples, and 1000 testing examples. The images are scaled to be 321x321, and labels are reordered to remain consistent.

Only the training and validation sets from the LSUN dataset are used. For the testing set, I use the dataset published by Hedau[7] containing 104 testing images. The dataset also contains 209 training images that are not used.

3.1 Data Augmentation

One way to organize the dataset is by the type of room, given in Figure 2. However, there is an imbalance of room types; in the training set, there are 1808 rooms of type 5, while there are only 2 rooms of type 2. A previous publication[1] uses two types of data augmentation to address this issue, both of which I will implement. First, a training image is flipped across the vertical axis with a probability of 0.5. Second, layout degradation is applied to rooms of a certain type. Figure 2 shows how rooms of different types can be converted into other rooms by removing a single surface. Layout degradation increases the distribution of room types, leading to higher accuracy for room types with less data.

In addition to the above protocols, I will apply a third type of augmentation. With a probability of 0.5, I will alter the brightness, contrast, saturation, and hue by a random amount. Restricting augmentation by a probability will preserve the content of the original training set while expanding the capability of the final model.

4 Methods

I train my model based on the Manhattan World Assumption[8], which models each image as a layout of planes, allowing me to restrict the room types to those denoted in Figure 2. These planes refer to the different surfaces within a room, including the front wall, floor, ceiling, right wall, and left wall. The entries in the output matrix given in Equation 1 refer to one of these surfaces.

The model I will be using is a ResNet-101 Fully Convolutional Neural Network pretrained on the 1000-class ImageNet classification dataset. Afterward, the model runs through an additional max-pooling and fully convolutional layer, followed by three transposed convolutional-layers to extract segmentation from the features found by the ResNet framework. The loss function is modeled as a combination of two terms, given in Equation 2.

$$L = L_{seg} + \lambda_s L_s + \lambda_e L_e \tag{2}$$

, where L_{seg} is the cross entropy loss, L_s is the L2 norm of the diference between the ground truth and the predicted output map, and L_e is the cross-entropy loss of the edges between the ground truth and predicted output. The edges are calculated from the output map, and are thicker at the beginning of training to allow for more error. As training progresses, the edges in the edge map become thinner as the model continues to improve.

5 Results

My model was able to achieve a total net loss of 0.55, with $L_{seg} = 0.282$, $L_s = 0.798$, and $L_e = 0.349$, averaged across all training examples. For the testing set, the total loss was 0.7376 averaged across all images.

Method	Pixel Error
Hedau[7]	21.20
Mallya[9]	12.83
Zhang[10]	12.70
DeLay[5]	9.73
My model	**8.13**

Table 1: Pixel error of various benchmarks

Here is an example of an input image, followed by the output my model predicts.

Input	Output
	

Table 2: Example of a single input-output image ran through the model

6 Conclusion

My model is able to sufficiently predict the layout of a room given an image of its interior cuboid structure. Based on Table 2, my model is able to account for 'clutter' within a room, and Table 1 shows that the model performs better than several previous benchmarks.

Plenty of work is to be done regarding room layout estimation with clutter. I tinkered with hyper parameters to augment the data, but didn't add significant post-processing mechanics such as the vanishing point algorithm delineated in DeLay[5]. While a vanishing point algorithm would take computational resources, an effective implementation could reduce the error and smooth out the edges presented in Table 2.

References

[1] Hung Jin Lin, Sheng-Wei Huang, Shang-Hong Lai, and Chen-Kuo Chiang. Indoor scene layout estimation from a single image. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 842–847, 2018.

[2] Lawrence Roberts. Machine perception of three-dimensional solids. 01 1963.

[3] D. Hoiem, A.A. Efros, and M. Hebert. Geometric context from a single image. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 654–661 Vol. 1, 2005.

[4] Sebastian Nowozin and Christoph H. Lampert. Structured learning and prediction in computer vision. volume 6, pages 185–365, 2011.

[5] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 616–624, 2016.

[6] Weidong Zhang, Wei Zhang, Kan Liu, and Jason Gu. Learning to predict high-quality edge maps for room layout estimation. volume 19, pages 935–943, 2017.

[7] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *2009 IEEE 12th international conference on computer vision*, pages 1849–1856. IEEE, 2009.

[8] James M Coughlan and Alan L Yuille. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *NIPS*, volume 2, page 3, 2000.

[9] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 936–944, 2015.

[10] Weidong Zhang, Wei Zhang, Kan Liu, and Jason Gu. Learning to predict high-quality edge maps for room layout estimation. *IEEE Transactions on Multimedia*, 19(5):935–943, 2016.