Detect Accessibility Feature of Storefront Using Deep Learning

Jiajun Chen, Dr. Hao Tang
Borough of Manhattan Community College

Abstract

Computer Vision is the branch of the science of computers and software systems which can recognize as well as understand images and scenes. Computer Vision is consisting of various aspects such as image classification, object detection, 3D reconstruction, image super-resolution and many more. Image classification is the fundamental field of study in Computer Vision and it's widely used for self-driving cars, robot, navigation and many other real-world applications.
In this project, we want to classify accessibility feature of storefront on sidewalk from street view images and it can help visually impaired people to avoid dangerous obstacles in the street and allow them to access each store. We are conducting some experiments on Faster-RCNN, a popular architecture in Object Detection. It enables us to have better understanding of how this model actually "sees" a physical object.

Introduction

Globally, at least 2.2 billion people have a vision impairment or blindness.[7] It's devasting to think in their perspective: No more vivid images in their life, only sound, touch and smell. They even have trouble hanging around with friends because of unpredictable obstacles they may face. Luckily, more and more organizations are devoted to help them. What if there is a map that records all the obstacles data precisely on sidewalk? With GPS and other technology, they can be notified in advance if they nearly crush into an obstacle. Since labeling those data requires enormous effort in person, it would be better if we can develop certain algorithm to label it automatically or semi-automatically. The key to this task is involved with a newly-developed research subject - Computer Vision.

Computer vision has gained a rapid development largely due to powerful hardware and enormous dataset. As a subset of computer vision, object detection really plays it vital role in our life. It is largely used in self-driving car, monitoring and so on. Also, it may be the eyes for visually impaired people in the future.

In this project, we are using object detection to detect storefront including doors, knob, stairs, ramps in Google street view dataset. By developing certain algorithm, machines are able to automatically or semi-automatically labels obstacles in geographical dataset, helping us to build a GIS Database that stores accessibility features of each location. So visually impaired people can hang out with friends as normal people with the help of our Database.

Research Question

How well does our deep learning algorithm perform? And how can we improve it?

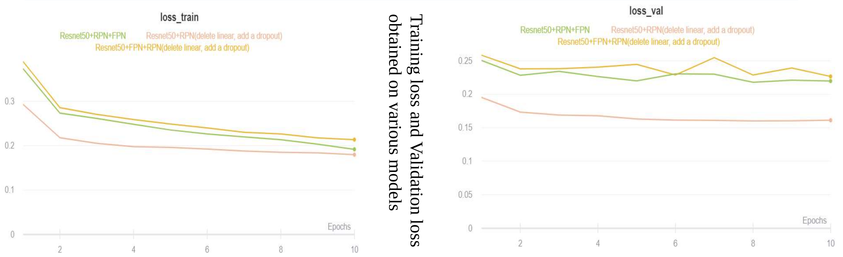
Navigation	Object Detection	Transportation	Database
Google Maps	PYTORCH	New York City Transit	Google Street View
GPU	**Object Detection**	**Organization**	**Organization**
NVIDIA	TensorFlow	BMCC Borough of Manhattan Community College	University of Central Florida 1963

Experiment

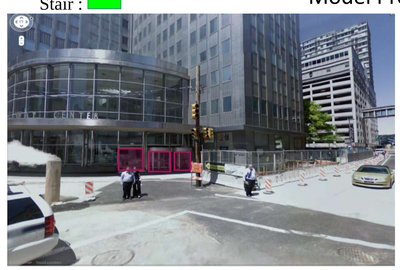
Q: We want to find out which model will yield the best performance in our validation dataset?

Methods: we collect images from UCF Google street view dataset [1], 1028 images in total and label it through Labelbox [5] website. And based on the rule "training: validation=8:1", we arrange the dataset and use 928 images for training data,100 images for validation data. And there are four kinds of object categorial that we concern: Door, Knob, Stairs, Ramp. Because our goal is evaluating how our model perform and how to modify it. Hence, we prepare fours different structure of our model based on Faster RCNN[4] architecture. Resnet50[8] + FPN (Feature Pyramid Network) + RPN (region proposal network), Resnet50[8] + RPN (a linear and add a dropout), Resnet50[8] + FPN+RPN(delete a linear layer and add a dropout)[3,4]. All the layers are non-trainable except for the RPN layer for all the above models. All the models are trained using training data (928 images) for 10 epochs, evaluating on validation data (100 images), using Adam optimizer with a 10^{-4} learning rate. Loss_train(evaluated in training set) , Loss_val(evaluated in validation set) are shown below.

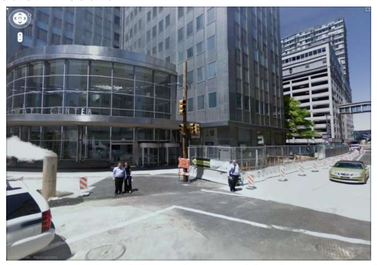
Result



Model Prediction Result

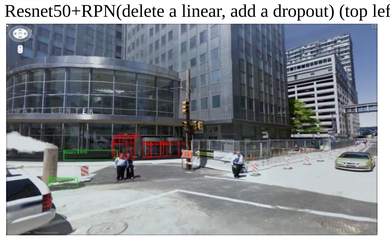


Original Annotation (top right)



Resnet50+RPN(delete a linear, add a dropout) (top left)

Resnet50+FPN+RPN(delete linear and add a dropout)(bottom left)



Resnet50+FPN+RPN (bottom right)

References

[1] Amir Roshan Zamir and Mubarak Shah, "Image Geo-localization Based on Multiple Nearest Neighbor Feature Matching using Generalized Graphs", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
[2] Tsung-Yi Lin1,2, Piotr Dollar´ 1 , Ross Girshick1. Feature Pyramid Networks for Object Detection. Apr 2017
[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In CVPR, 2017.
[5] Labelbox, "Labelbox," Online, 2020. [Online]. Available: https://labelbox.com International Conference On Internet of Things: Smart Innovation and Usages, pp. 1-5
[6] Jan Hosang, Rodrigo Benenson and Bernt Schiele. Learning non-maximum suppression. In CVPR, 2017
[7] https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
[8] Kaiming He, Xiangyu Xhang, Shaoqing Ren, Jiah Sun. Deep Residual Learning For Image Recognition. In CVPR, 2017

Acknowledgments

1. CUNY Research Scholars Program (CRSP), SSTEM Program
2. Dr. Marawan Elzoeiry – CRSP Program Coordinator
3. Labelbox, "Labelbox," Online, 2020. [Online]. Available: https://labelbox.com

Discussion

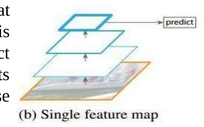
Loss: Although Resnet50 + RPN(delete linear, add a dropout) has a lower Loss_val than Resnet50 + FPN + RPN and Resnet50 + FPN + RPN (remove a linear, add a dropout), it doesn't mean that the former model will perform better than the other. If we consider a simple binary classification with binary cross entropy, the network has a very high loss if it predicts something wrong with high confidence, and lower loss for predicting something wrong with low confidence.

FPN[2]: There are four pictures on the right clearly illustrate how good their performance. As we can see, the model that doesn't include FPN is doing much worse than the other two that actually includes FPN. (Although an empty image without any prediction is shown here, doesn't mean the model without FPN can't detect any objects. But based on our result, generally it performs worse than the other two models). And based on our experiment, FPN network structure can help the base model to achieve better result in general object detection.

An visualization of the Non-FPN and FPN structure are shown below:

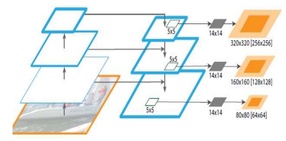
An feature map(image)'s resolution will keep decreasing when going through from the bottom to the top, while its feature will become more abstract, concentrated. (Imaging a photo has been zoom in while preserving the same scale, lots of detail will lost).

Without FPN



Right: The only information that model is using is a single, abstract feature map. Lots of detail will lose in this case.

With FPN

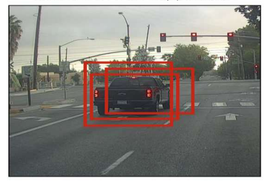


Left: the model combines multiple scale feature maps to predict object. It keeps the information for both big objects and small objects. So it has better performance.

Multiple Bounding Boxes (NMS): All the predictions in those three images shows a problem: there are so many boxes for a single object, which is not supposed to happen. An famous algorithm called Non-maximum Suppression(NMS)[6] is to solve this problem.

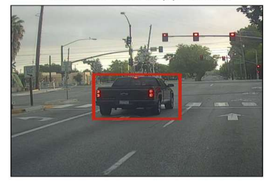
Generally, this algorithm calculate a bounding box's IoU and confidence score, and then compares it with the other nearby bounding boxes. And it only select the best bounding box for that object.

Before non-max suppression



Non-Max Suppression

After non-max suppression



Conclusion

- At current stage, This model can benefit us by classifying whether there are target objects in images. Because we found nearly 70% are Skipped images when labeling it, meaning there aren't objects that we concern. If it can classify and throw away those skipped images at first, our labelling work will be dramatically decreased.
- Through out the semester, we have gone through a lot. From building a server from ground up to building a neural network to train, not only did I learn about machine learning, but also realize the importance of writing a log as a weekly routine.

Future Work

- Eliminating multiple bounding boxes, utilizing more prediction from different models, to make a more precise prediction based on certain algorithm.
- Developing a persuasive indicator to quantitative our model performance.
- Comparing how brightness, angle of objects result in performance improvement.
- Testing our model on NYC street view and compare its performance to our current's.
- Ultimate Goal: Assisting human beings labeling accessibility facility on images more efficiently.