Object Detection, Recognition & Classification for Autonomous Vehicles

Neel Haria
Department of Electrical & Computer Engineering
Stevens Institute of Technology
Hoboken, USA
nharia@stevens.edu

Project Advisor: Prof. Hong Man

Meghana D Masudi
Department of Electrical & Computer Engineering
Stevens Institute of Technology
Hoboken, USA
mmasudi@stevens.edu

***Abstract* - To function safely, the detection, recognition & classification of objects around an autonomous vehicle is crucial. For e.g., Tesla, best known for its electric cars have had several accidents in the past few years related to failure in recognizing the heavy vehicles when the autopilot was activated. It was evident that Tesla's Autopilot may have trouble recognizing the other heavy vehicles that were stationary and the ones crossing the roads. Hence, we also explore semantic segmentation in the later stage of our research which is a critical aspect in autonomous vehicles as it is necessary for the models to understand the context of the environment in which they are operating. To do that, we use YOLOv3 which uses a new network for performing feature extraction. The new network is a hybrid approach between the network used in YOLOv2(Darknet-19), and residual network.**

Keywords – YOLOv3, Convolutional Neural Network, Object detection

I. INTRODUCTION

Self-driving technology presents a rare opportunity to improve the quality of life in many of our communities. Avoidable collisions, single-occupant commuters, and vehicle emissions are choking cities, while infrastructure strains under rapid urban growth. Autonomous vehicles are expected to redefine transportation and unlock a myriad of societal, environmental, and economic benefits.

Both car manufacturers and IT companies are competitively investing to self-driving field. Companies such as Google, Uber, Ford, and BMW have already been testing their self-driving vehicles on the track. An integral part of autonomous cars is optical vision. Accurate real-time object detection, such as vehicles, pedestrians, animals, and road signs, could accelerate the pace of building a self-driving car as safe as human drivers. For decades, image understanding has been a challenging task. Objects in the physical world, unlike geometric figures, are often irregular figures. Besides, depictions of objects in the environment of the real world are variant to illumination, rotation, scale, and occlusion that makes the task of object detection more complex and challenging. Large improvements are made in the recent years in object detection using Convolutional Neural Network. We are therefore inspired to apply YOLOv3 to the autonomous driving for the task of object detection, recognition, and classification.

II. RELATED WORK

A. YOLOv3

YOLOv3 is the latest variant of a popular object detection algorithm YOLO – You Only Look Once and an open source method of object detection. It recognizes 80 different objects in images and videos, but most importantly it is super-fast and nearly as accurate as Single Shot Detector (SSD). The detection is done by applying 1x1 detection kernels on feature maps of three different sizes at three different places in the network. YOLO v3 makes prediction at three scales, which are precisely given by down sampling the dimensions of the input image by 32, 16 and 8 respectively as shown in the below architecture.

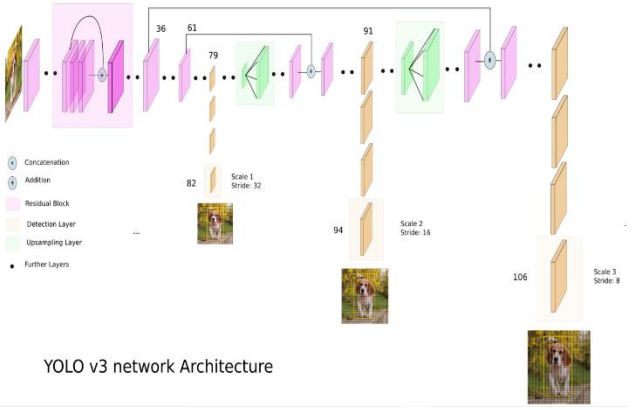


Image credit: https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b

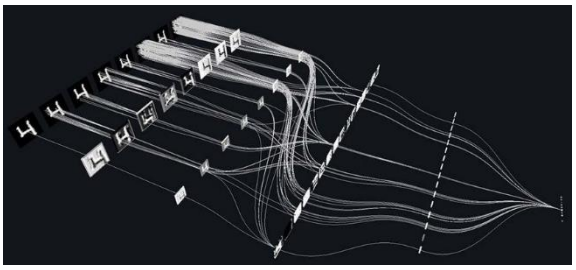
B. DarkNet

YOLOv3 uses a variant of Darknet, which originally has 53-layer network trained on ImageNet. For the task of detection, 53 more layers are stacked onto it, giving us a 106 layer fully convolutional underlying architecture for YOLOv3.

C. CNN

A convolutional neural network performs convolution operations on images to classify them. By convoluting the kernel on the input image, CNN computes the feature map corresponding to the kernel. The method of convolution and the method of pooling are repeatedly applied for extracting the feature map. Input to fully connected layers is the extracted feature map, and each class's probability is finally output.

CNN training is done by modifying the network parameters using the back-propagation process. The CNN parameters refer to the kernel of the convolutional layer and the weights of all coupled layers. CNN can perform not only image classification but also object detection and semantic segmentation by designing the output layer according to each task of image recognition. In object detection using CNN, object proposal regions with different aspect are detected by CNN, and multiclass object detection is possible using the Region Proposal approach that performs multiclass classification with CNN for each detected region.



Convolutional Neural Network

D. Pytorch

PyTorch is an open source machine learning framework that provides a complete end-to-end research framework which comes with the most common building blocks for carrying out every day deep learning research. It allows chaining of high-level neural network modules because it supports Keras.

E. Semantic Segmentation (Scene Understanding)

Semantic segmentation is the task of clustering parts of images together which belong to the same object class. Object detection, in comparison to semantic segmentation, must distinguish different instances of the same object. While having a semantic segmentation is certainly a big advantage when trying to get object instances, there are a couple of problems: neighboring pixels of the same class might belong to different object instances and regions which are not connected may belong to the same object instance. For example, a tree in front of a car which visually divides the car into two parts.

III. CHALLENGES

The challenges faced by us during the implementation of YOLOv3 on the dataset chosen are as follows:

- Data Collection- It was difficult to obtain appropriate images for our model training. For object detection applications, it is important to obtain an image data set such that the object to be localized and classified is labelled precisely. An unlabeled data set is not useful as our model would not know what to localize or classify. The data set should be appropriately classified into two sections, one for training and one for testing. Usually the norm/ratio followed for the same is a 4:6 ratio i.e. 40% of the dataset is used for training and the next 60% is used for testing.
- Speed of Localization- For applications in autonomous driving, speed of localization is extremely important or else it could lead to disastrous outcomes. Along with that, it is also important to correctly classify objects in its vision. Hence, to improve accuracy in both, it is important to test various algorithms, which goes back to the point that we provide special attention on accurately training our model and preparing our dataset correctly.
- Classification Problem- One major problem faced in autonomous driving applications are classification problems. The algorithm should correctly classify between many same types of objects. Such as the algorithm should be able to identify multiple similar type of objects in the same frame.

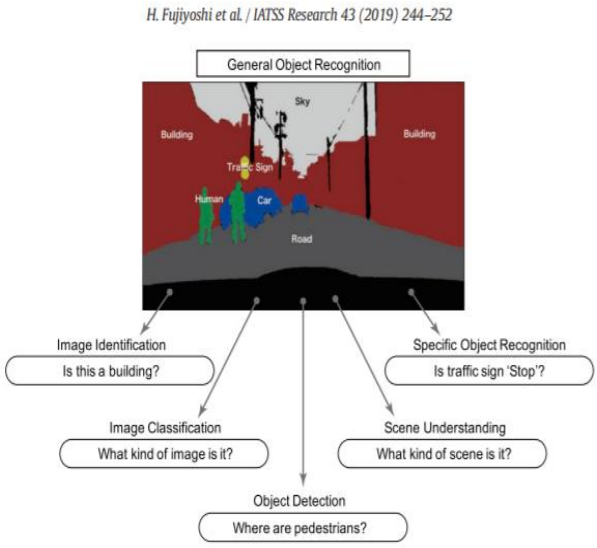
These are all the challenges we are currently working on, on a deeper level to solve other problems and obtain a workable solution.

IV. FORMAL DEFINITION

Object Detection- Object detection is the issue of identifying an object's location of a certain category in the image. In this task, face detection and pedestrian detection are also included. Multiclass object detection targeting multiple categories can be done with one network in deep learning-based object detection.

Object Recognition- Object recognition is the issue of recognizing a particular object. It is a computer technology related to computer vision and image processing that deals with **detecting** instances of semantic **objects** of a certain class (such as humans, buildings, or cars) in digital images and videos.

Object Classification- Object classification is a problem to find out the category to which an object in an image belongs to, among predefined categories.



Segmentation of general object recognition

V. FORMULATION OF THE PROBLEM

There are loopholes in the methods adopted for object detection since years. We try to test the already existing algorithms to find out those loopholes and ways to solve them by applying different methodologies with the technologies in trend.

Detecting a stationary vehicle and lane detection are the two most important topics of our research project. We got inspired by three Tesla incidents that took place where Tesla Model S crashed into a stopped firetruck, one of which was in San Jose, California. It is a matter of concern as to how is it possible that one of the most advanced driving systems on planet does not see a huge fire truck, dead ahead of it.

Fig 1. A fire truck parked in the parking lot hit by Tesla Model S

Seems like Volvo's semiautonomous system, Pilot Assist, has the same shortcoming. Say the car in front of the Volvo changes lanes or turns off the road, leaving nothing between the Volvo and a stopped car. "Pilot Assist will ignore the stationary vehicle and instead accelerate to the stored speed". The driver must then intervene and apply the brakes. In other words, your Volvo won't brake to avoid hitting a stopped car that suddenly appears up ahead. It might even accelerate towards it.

Hence, along with object detection and recognition we believe by researching and learning about semantic segmentation and unlocking its potential we could understand what kind of scene it is thereby to detect if a vehicle is stationary or moving. That will also help us further to think about autonomous parking of vehicles.

VI. DESCRIPTION OF THE SOLUTIONS/DESIGNS

Since it is the beginning stage of our research. By this point of time we have learned how to train our model on the dataset that we extracted from the site roboflow called Vehicles-OpenImages.v1-416x416.darknet which is a labelled dataset.

We inputted the image shown in Fig 1. which was passed through the Convolutional Neural Network and the output we received was a matrix of bounding box predictions as shown in Fig 2.

Fig 1. The input image given to the Convolutional Neural Network (CNN)

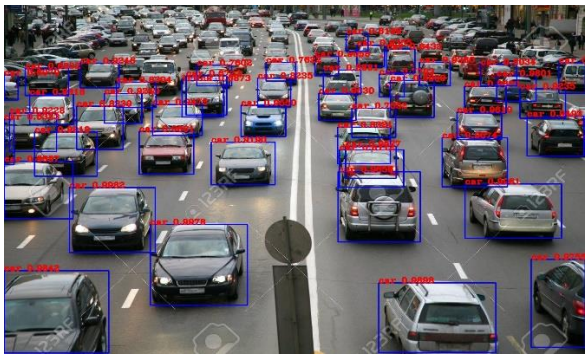


Fig 2. The output image obtained with the bounding box predictions with a class label for each bounding box

We observed that YOLO v3 predicts boxes at 3 different scales. For the same image of 416 x 416, the number of predicted boxes are 10,647. Hence, this also means that YOLO v3 predicts 10x the number of boxes predicted by YOLO v2.

VII. FUTURE PLANS

Going further we plan on studying and exploring already existing algorithms to dissect what kind of different scenarios if considered, they do not work the way it is expected to work. We will finally try to present our own improvements on the existing algorithms that could be made for them to work better than they do presently.

We expect the outcomes or improvements provided by us to work the way we desire them to at the end of our research project.

REFERENCES

[1] Gene Lewis, Stanford University, "Object Detection for Autonomous Vehicles,".

[2] Nilesh J.Uke and Ravindra C.Thool, " Moving Vehicle Detection for Measuring Traffic Count Using OpenCV," Journal of Automation and Control Engineering. Vol. 1, No. 4, December 2013

[3] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, Karol Zieba, VisualBackProp: Visualizing CNNs for Autonomous Driving, arXiv preprint, arXiv:abs/1611.05418 2016.

[4] M. Bojarski, D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba, End to End Learning for Self-Driving Cars, arXiv preprint, arXiv:abs/1604.07316 2016.

[5] Kongming Liang, Yuhong Guo, Hong Chang and Xilin Chen, "Visual Relationship Detection with Deep Structural Ranking," The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).

[6] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, European Conference on Computer Vision 2018, pp. 563–578.

[7] Hironobu Fujiyoshi*, Tsubasa Hirakawa and Takayoshi Yamashita, "Deep-learning based image recognition for autonomous driving," IATSS Research. Volume 43, Issue 4, December 2019.

[8] Y. Mori, H. Fukui, T. Hirakawa, N. Jo, T. Yamashita, H. Fujiyoshi, Attention neural baby talk: captioning of risk factors while driving, IEEE International Conference on Intelligent Transportation Systems, 2019.