Real-Time Object Recognition in Urban Environments Using YOLO

Hsiaochen Huang
hsiaochen.huang@sofia.edu

Jiacheng Weng
jiacheng.weng@sofia.edu

1 Motivation and Problem Statement

Nowadays, the continuing evolution of autonomous vehicles aims to deliver even greater safety benefits. One day, we can handle all kinds of tasks of driving when we don't want to or can't do it ourselves.

Considering the safety factor, the first challenge that stuck out to us is how to tell the difference between people and signs. It's a hugely important, but typically very simple, distinction that you would make reflexively. However, autonomous vehicles can't do this effortlessly. Therefore, we will help vehicles how to classify objects.

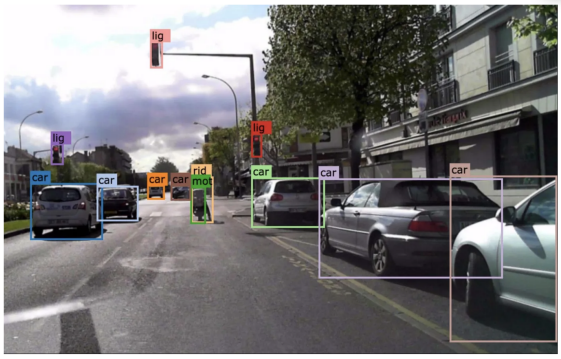

Figure 1: Example of urban scene object recognition [4]

2 Approach and Model Development

Though there has been much work in image segmentation [1], in this project, we will take advantage of those networks and observe different responses by tuning the hyperparameters such as the number of layers and the size of receptive fields. We are going to implement Convolutional Neural Networks (CNN), which are powerful visual models that yield hierarchies of features [2], to achieve our goal. By using a contemporary classification model, like YOLO, trained with image-level labels for the task.[3]

3 Dataset

The model will be trained and evaluated with Berkeley DeepDrive 100K (BDD100K) dataset[4]. We will predict objects of different movable objects appearing in the view of a car camera.

The BDD100K dataset comprises 100,000 video clips collected from over 50,000 rides across New York, the San Francisco Bay Area, and other regions, showcasing diverse scenes such as city streets, residential areas, and highways, recorded in various weather conditions and times of the day. Split into training (70K), validation (10K), and testing (20K) sets, each video is 40 seconds long with 720p resolution and a frame rate of 30fps, with annotations provided for image classification, detection, and segmentation tasks for the frame at the 10th second of each video.

4 Computing Platform

Since the availability of the powerful Graphics Processing Units (GPUs) will be a crucial key to train a deep convolutional neural network. We will use the compute engine with one NVIDIA GPU. To build networks, we choose to use Python as our programming language and Pytorch as our library due to the coding convince.

5 Evaluation Metrics

We will be using two methods, Intersection over Union(IoU) and Mean Average Percision(mAP). IoU measures the overlap between the predicted bounding boxes and the ground truth bounding boxes. It is calculated as the ratio of the intersection area to the union area of the two bounding boxes. A threshold IoU value (e.g., 0.5) is commonly used to determine whether a detection is considered correct or not. mAP is the mean of the average precisions for all classes. It gives an overall performance measure of the model across different object classes.

6 Expected Result

We anticipate that we can help autonomous vehicles to identify objects immediately on the road. In our project, we will find-tune the hyperparameters to look for the lowest error rate, which is lower than 10% (expected). To give autonomous vehicles the ability to segment the scene instantly, the first challenge we must tackle is to do object recognition precisely in images. Since object recognition is a crucial and fundamental concept for video segmentation, it is useful for us to go through this obstacle. Finally, we can obtain our final goal to make autonomous vehicles safe and feasible.

References

[1] J. M. Alvarez, T. Gevers, Y. LeCun, and A. Lopez. Road scene segmentation from a single image. In *European Conference on Computer Vision*, pages 376–389, 2012.

[2] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.

[3] C. Wang, I. Yeh, and H. M. Liao. Yolov9: Learning what you want to learn using programmable gradient information. *CoRR*, abs/2402.13616, 2024.

[4] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2018.