Report for ROB 599 Perception: Team 2 - Knights

Problem Statement: Estimate the number of cars (a non-negative integer) in a scene given a snapshot (a collection of one image, LiDAR points, and the projection matrix)

Approach: We are using a network based on the YOLO architecture from the paper You Only Look Once: Unified, Real-Time Object Detection [1]. The architecture consists of 22 convolutional layers, 5 max pooling layers to decrease the size and one classification/detection layer that gives us the output. The output that we get are the 2-D bounding boxes with the dimensions relative to the image size/resolution and the corresponding classes.

Training Procedure: We used transfer learning on the Yolo pre-trained weights for ImageNet [3]. Our training happened in two stages. Firstly, the above mentioned pretrained weights were used to train the network for detecting one class "car" for the GTA dataset from [2]. The dataset has 200K images from GTA along with the annotations in the PASCAL-VOC format. The data was preprocessed to get the labels and the 2D bounding boxes. This dataset has only one label "car". We trained using this dataset to improve the localization performance of the network for better detection of vehicles. However, we needed a dataset that could classify the detections into 23 classes. For this, we used the 6K image dataset and annotations provided for the competition. We preprocessed the 3-D annotations to get the 2D bounding boxes for detection and classification into the 23 different classes. Using these datasets we trained three networks using our two stage training methodology:

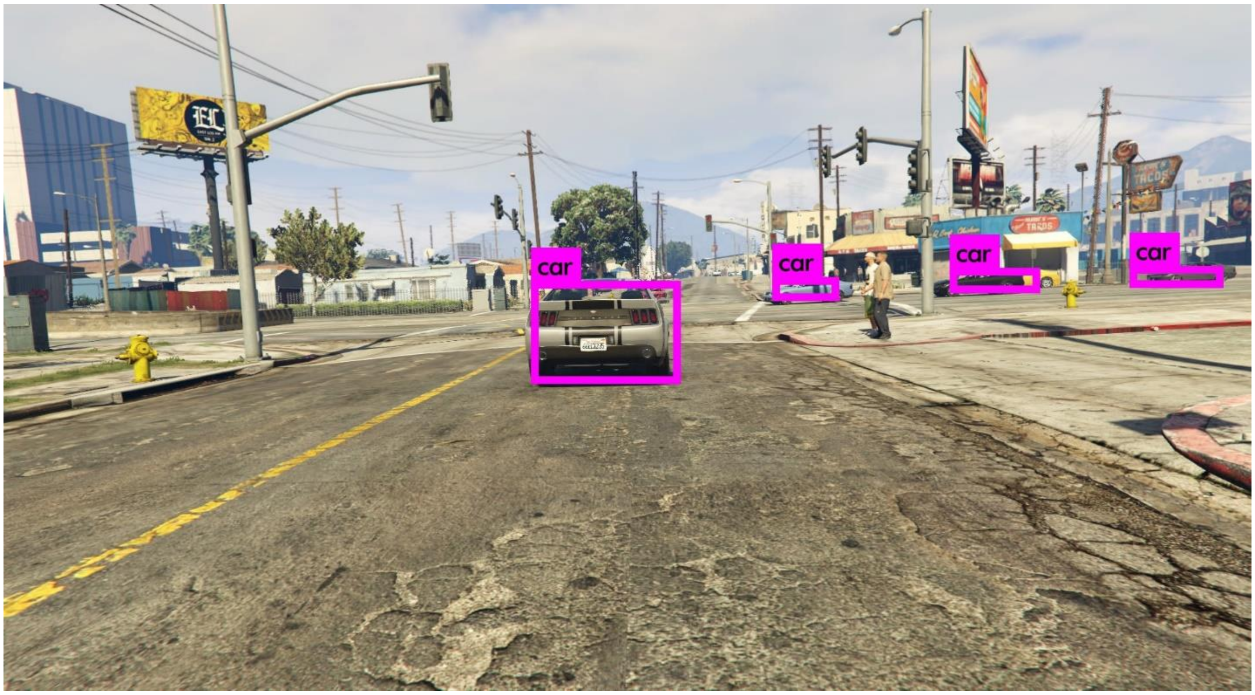
1. Network with a resized input image resolution of 608 pixels x 608 pixels trained for 100000 iterations. This network trained on the 200K dataset is now trained on the 6K dataset for 50000 iterations to detect and classify the images into the 23 different classes that we were supposed to classify in. The classification confidence for the network depends on the training which is one of the hyper parameters we needed to tweak to get better results. The learning rate was 0.0001 and number of output classes of the final network was 23.
2. Network with a resized input image resolution of 416 pixels x 416 pixels trained for 100000 iterations on the 200K dataset. The performance of the network was subpar compared to the network with 608 pixel x 608 pixel resolution network due to which we decided not to move ahead with this network.
3. Network trained on 6K dataset with 416 pixel input resolution initially for 10000 iterations to classify in 23 classes. The prediction for the network was not good due to which we trained the network on 200K dataset for 100,000 iterations to improve the localization performance. For the second stage we trained this network on 6k image dataset for 80000 iterations. The learning rate for the final stage was 0.0001 and the output classes were 23.

After getting the 2-D bounding boxes coordinates and their classifications, we used the given camera projection matrices and the cloud data to calculate the 3-D bounding box coordinates. The 3-D bounding box coordinates and the class labels were used in our evaluation script to meet the given evaluation criteria. After observing the results of all the three networks, we decided to go with the first network as it was giving us the lowest mean absolute error.

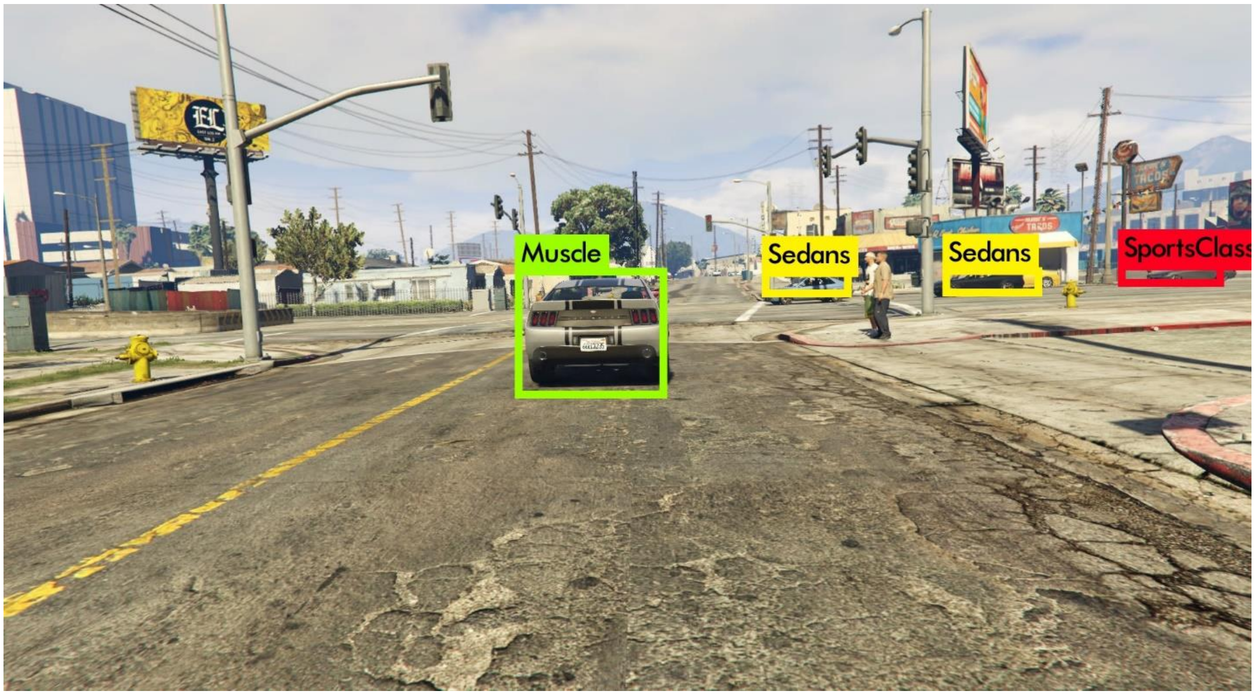
Links for all the datasets, and weights are provided in the Readme.txt file.

Results:

1) Predictions after stage 1 of training – Trained on 200k dataset for 1 class (“car”).



2) Predictions after stage 2 (final) of training – Trained on 6k dataset for 23 class.



References:

1. Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
2. M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, Karl Rosaen and R. Vasudevan, “Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?,” in IEEE International Conference on Robotics and Automation, pp. 1–8, 2017.
3. https://pjreddie.com/media/files/darknet19_448.conv.23