YOLO 9000: 2D Object Detection from images

Abstract:
The main aim of the project is to be able to understand the Neural network architecture, and the working methodology behind the YOLO object detection and customise the implementation to training on a custom dataset. For achieving the goal of object detection two datasets are developed for training the YOLO deep neural network architecture and performing the inference. The first dataset is aimed at person detection in a indoor environment scenario and the second dataset is aimed at detecting Persons, Machinery, Critical emergency response systems in a mine environment. For this the raw data from the specific environment under study is processed into training and test set and then fed to the neural network for training and inference learning. Further, the trained networks with some customisation to the architure of the neural network are deployed onto the Jetson TX2 system to understand the capability of the embedded AI module for real time object detection. The learnings from the study will be helpful in assessing the benefits of deploying a embedded module on a mobile robotic platform.

This report discuss the specific details of how this task of object detection using a deep neural network architecture.

Introduction:

Object detection forms one of the important part of the current mobile robotics systems. For abling the perception capability to robots many vision techniques are in use which enable the systems to understand the environment of the robot. The input to the vision technique comes from the vision sensors on the mobile robot. Number of sensors are made available to the mobile robots to provide great information on the environment and its envisioning. Some of the widely used sensors are the high resolution cameras, RGBD, stereo, LIDAR and thermal sensors. Deep Neural Network (DNN) based object detection techniques are currently the state of the art in performing the task of object detection. The neural networks prior to the deployment for particular task are trained to perform the inference. This neural networks are composed of tens of layers with millions of the training weights which gets trained on the input training examples.

Methodology:

You only look once (YOLO) is one of the state of the art DNN methodology for object detection. The YOLO neural network is composed of 24 neural network layers and 2 fully connected layers. The neural network layers perform convolution on the

input training images with the filters whose weights are trained using a optimisation function. The output of the convolution operation in each layer is subjected to non linear activation functions, and max pooling. The max pooled output from the previous layer is passed onto the next layer and the convolution process is carried on layer by layer throughout the depth of the neural network.

The YOLO Neural Network is based upon unified detection where in the the neural network will be able to predict bounding boxes for all classes in the image simultaneously. Further, the architecture of the neural network consists of 20 Classification layers and 4 detection layers. The neural network is trained end to end to perform the task of object detection. This particular methodology in which YOLO generates the region proposals and train the millions of parameters in the neural network gives the neural network an upper hand in speed and almost an equal mean average precision (mAP) in relative to the current state of the art deep neural network architecture.

The neural network architecture forms the backbone of the YOLO implementation on which the training process is carried out. In this method of training the input image is divided into grid cells. Each of the grid cells in the image are responsible for detecting five bounding boxes. Along with the bounding boxes the methodology includes the confidence scores for the bounding boxes detected in the image. The bounding box and the confidence score help the network learn to predict the regions which have a higher chance of having a object. In addition to the above each bounding box also predicts the class probability. The bounding box, confidence scores and the class probability values obtained for each training instance are validated and the loss calculated and is back propagated through the network by updating the weights to minimise the loss. This cycle of forward and backward propagation of training images is repeated until a desired level of accuracy in predicting the bounding box and the class probability is achieved.

In the first stage of the project the YOLO neural network is trained to detect legs of persons in the images and in the second stage the object detection is extended to detecting people, Machinery and critical emergency response systems in the mine environment. In addition to the YOLO implementation the Tiny YOLO which is a small version of the original YOLO is trained and tested for deployment onto a mobile robot for real time object detection using the Jetson TX2 which is a embedded AI system.

For this, firstly The network architecture is loaded with pretrained weights which are obtained by training the neural network on the Imagenet classification dataset. Then

the neural network is trained end to end for detection using the custom training data. Currently, the deep neural network deployed onto the mobile robot enabled with Jetson TX2 is able to achieve detection frame rates of 15. The Deep neural network is trained for detection of persons, hardhats, fire extinguisher,reflection jacket, and valves. The results of the inference of the trained neural network for detections can be seen in fig 1

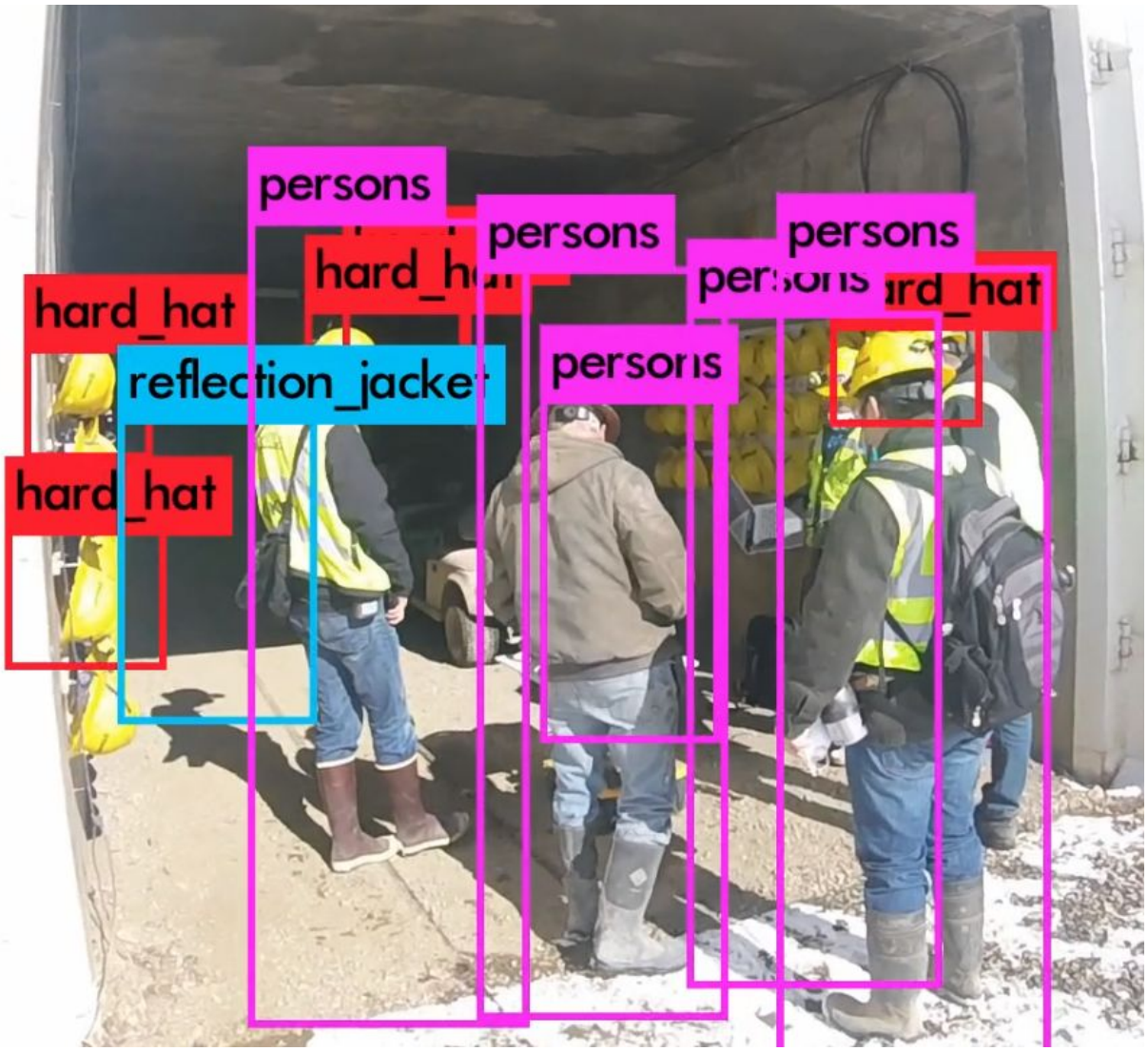

Figure 1: The object detection inference on YOLO neural network trained on persons, mine equipment.

In addition to the classes mentioned afore the neural network will be trained for detections on other Critical emergency response systems, and objects making the Deep neural network more robust and reliable object detection tool.

Conclusion:

In conclusion the end goal of customising the YOLO neural network architecture and designing custom datasets to train the Deep neural network for performing detections of objects of interest in the environment under study is achieved. The learnings in creating the dataset and customising YOLO are documented and made into a step by step guide for developing custom datasets from the scratch.