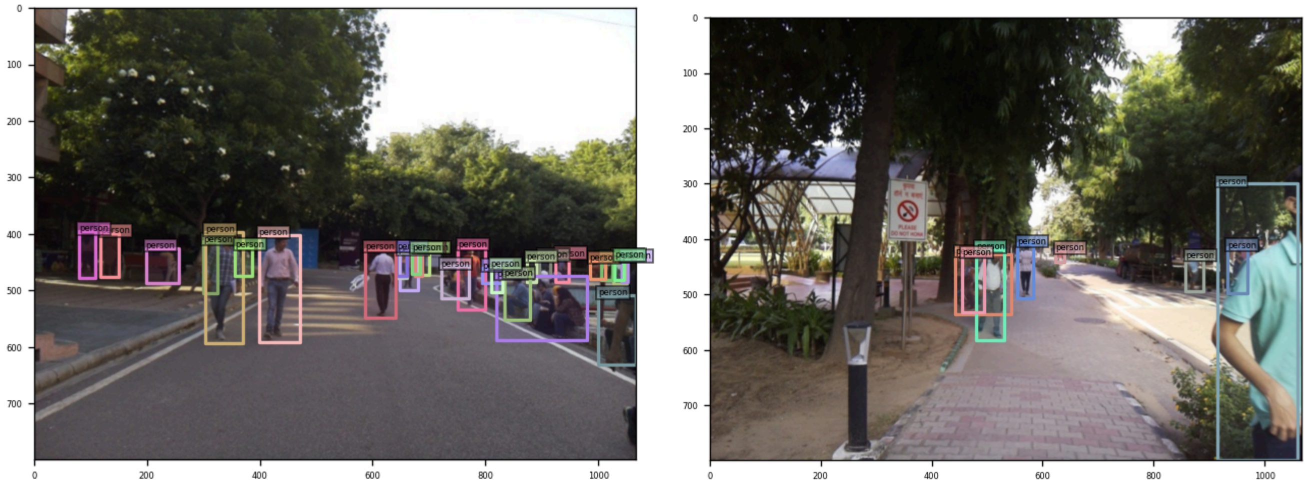
Report on DINO model evaluation and training

Aditya Kumar

This report focuses on evaluating and training the DINO model on a small subset of the IIT Delhi Pedestrian dataset. The dataset consists of 200 images belonging to the "person" class, annotated using bounding boxes. The annotations are saved in a JSON format similar to the COCO dataset. The DINO model, a transformer-based architecture, demonstrates strong performance even with a low number of training epochs.



(a) (b)

Fig 1: (a) & (b) Samples from the dataset with Original Bounding Box annotations

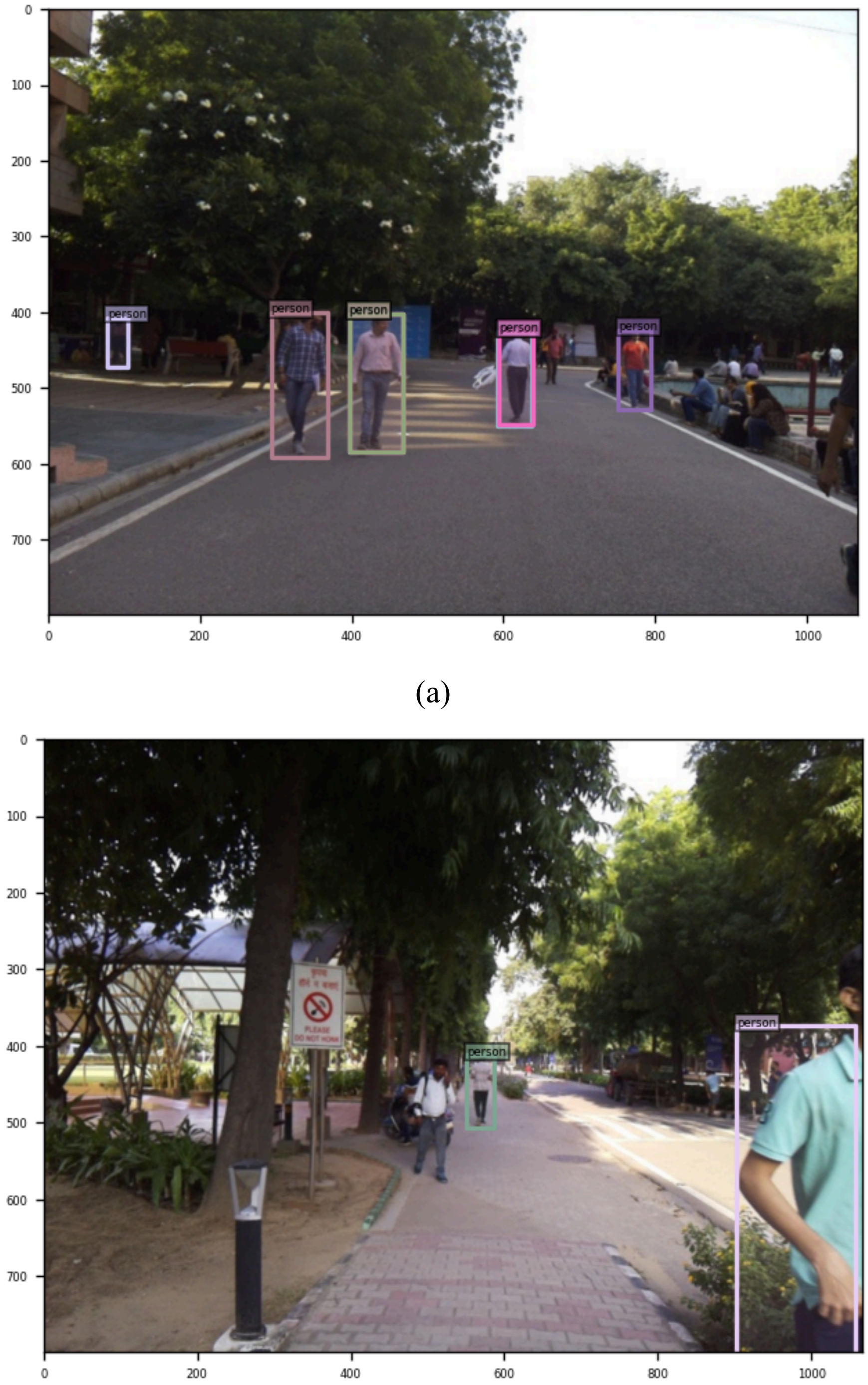
The first step in the process is to visualize the dataset using multiple image samples and their corresponding annotations. This helps in assessing the overall complexity of the dataset and the variability present in it. Following this, the dataset is preprocessed into a format that is compatible with the DINO model. The data is split into 160 training images and 40 validation images, which are stored in separate folders named train2017 and validation2017, respectively. Randomization is applied to the original dataset to ensure a fair and unbiased starting point for training, preventing any patterns in the data order from influencing the model's learning process.

The DINO model repository is cloned into a Google Colab environment for easy access. This repository contains models, tools, and scripts used for training and testing the model. The environment and all necessary dependencies are set up, including updating the cocoeval script in pycocotools, where ‘np.float’ is updated to ‘np.float64’ or ‘float’.

The original DINO model is then evaluated on the dataset, yielding an average precision of 84% at IoU50 and an average recall of 50.5% at IoU50-95. These results indicate that the original model performs well, showing high precision and accuracy in detecting pedestrians.

The DINO model is then trained on the preprocessed dataset for 12 epochs. After training, the model achieves an average precision of 56.4% at IoU50 and an average recall of 38.7% at IoU 50-95. These results highlight the challenge of working with a small dataset, as the model tends to overfit the training data, leading to lower accuracy on the validation set.

Finally, the model is used to infer results on the validation set, and the detections are compared with the actual annotations. The model is unable to detect all the objects present in the images that are annotated, demonstrating the limitations of the model due to the small dataset and possible overfitting.



(a)

(b)

Fig 2: (a) & (b) Trained DINO model inference on validation images

From Figures 1 and 2, there is a disparity between the original annotation and the prediction done by the DINO model. It is observed that some of the "person" objects in the images are being detected, while some are not. There is minimum incorrect detection, however, for reliable model performance, all the objects in the image need to be detected. This shows the need for fine-tuning the model while training to achieve more accuracy.