Report

In this experiment, the primary goal was to develop and evaluate classification models capable of accurately predicting ten distinct vehicle types using YOLOv8n-cls and YOLOv11n-cls architectures. The experiment began by training these models on a baseline dataset without performing any additional data augmentation or addressing potential class imbalances. This baseline dataset consisted of images with limited class diversity and served as the foundation for the initial phase of model evaluation. Each model was trained using two different configurations: batch sizes of 16 and 32, over 200 epochs. These configurations were chosen to test the model's learning capacity under varying conditions while ensuring optimal utilization of available computational resources. After training, the models were evaluated based on their Macro F1 scores, a metric specifically chosen for its ability to fairly assess performance across all classes, especially in imbalanced datasets. This step was critical in determining the strengths and weaknesses of each model architecture on the baseline dataset.

Following the initial evaluation, it became clear that additional data was necessary to improve class diversity and address the potential underrepresentation of certain vehicle types. To enhance the dataset, public datasets from Roboflow were integrated into the training pipeline. These included Samlor-Tuktuk-Vehicles, Detect_Dum_Em, and Vehicle_Car. These datasets introduced additional samples of various vehicle types, contributing to a more balanced and representative training dataset. The inclusion of these datasets was carefully considered to ensure that the model could generalize well to unseen data while maintaining robust performance across all classes.

With the augmented dataset, the models were retrained using the same experimental configurations of batch sizes 16 and 32, and an epoch count of 200. This ensured a fair comparison between the baseline results and the results achieved with the enriched dataset. Throughout the training process, the Macro F1 score remained the primary evaluation metric, aligning with the requirements for final submission. The retrained models demonstrated improved performance, particularly in handling previously underrepresented vehicle types, validating the importance of data augmentation and diversification in achieving robust classification.

At the end of the experiment, I identified three models as the most promising candidates for final evaluation. These models were selected based on their Macro F1 scores during training and validation. On submission day, I will evaluate these three models on the provided unseen test dataset and select the best-performing model based on its results. This step ensures that the final submission reflects the most robust and well-performing model for live inference. Overall, this systematic approach, starting with a baseline evaluation and progressing to dataset enhancement, retraining, and final selection, underscores the importance of iterative experimentation, careful dataset curation, and metric-driven evaluation in achieving optimal results.