Deep Learning-Based Sign Detection in Car Surveillance Videos

Felipe Inagaki de Oliveira (fd2264)
Suemy Inagaki de Oliveira (si2324)
Carlos David Rodrigues Melo (cdr9659)
New York University

Introduction

With the increasing availability of surveillance videos captured by vehicles, there is a growing need for automated systems to detect and recognize traffic signs, such as stop signs and pedestrian signs, and to enhance road safety and navigation assistance.

Traditional methods of sign detection often rely on handcrafted features and may struggle with variations in lighting conditions, occlusions, and sign degradation (de la Escalera et al. 1997),(Houben et al. 2013). In this way, deep learning techniques have shown promising results in various computer vision tasks, making them an attractive approach for sign detection in car surveillance videos (Zhu and Yan 2022).

The goal of this project was to develop a deep learning-based system for accurately detecting and recognizing Brazilian traffic signs in car surveillance videos, such as dashcam recordings. Specifically, we used state-of-the-art deep learning architectures, such as YOLOv8 (Reis et al. 2023), and DEtection TRansformers (DETRs) (Zhao et al. 2024), for robust sign detection.

That way, we finetuned three versions of YOLOv8 and one of DETR models on a dataset containing 4127 categorized Brazilian traffic signs (Cimirro 2013) to learn discriminative features for different types of signs.

However, these models were initially trained to detect objects in images, not videos. Therefore, they cannot differentiate whether a new object has appeared in the video frame, or whether it is the same object that is still present in the scene. To avoid this, we also used a Re-Identification model. In particular, we used Bot-SORT (Aharon, Orfaig, and Bobrovsky 2022), where we pass as input the output data of the models above, and the output is the result of identifying and counting the signs on the streets, avoiding duplicates. Finally, we evaluated and compared the system's performances on a 17-minute surveillance video captured on a Brazilian road.

Literature Review

(Zhu and Yan 2022) presented several convolutional neural network architectures that are used for traffic sign detection, such as Faster R-CNN, YOLO, SSD, among others. And some of the challenges they presented involve detecting small objects, tracking moving objects, and integrating data from multiple sensors into vehicle surveillance systems. According to them, the use of Transformers (Vaswani et al. 2023) for object detection can improve the effectiveness of vehicle surveillance systems.

Models Architectures

YOLOv8

YOLOv8 is a convolutional neural network architecture developed for real-time object detection in images. It uses a single-stage approach, that is, it detects objects directly in a single pass through the network, instead of dividing the process into several steps like other architectures. In addition, YOLOv8 combines advanced neural network techniques, such as residual layers and Feature Pyramid Network (FPN), to improve detection results and speed. It is an evolution of the YOLO family, seeking to improve accuracy and computational efficiency in relation to previous versions (Reis et al. 2023). Its architecture is demonstrated in figure 1

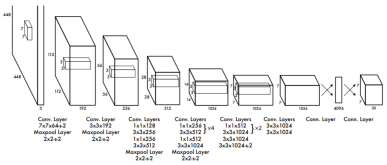

Figure 1: YOLO family architecture extracted from (Reis et al. 2023)

DEtection TRansformer (DETR)

The DETR model utilizes an encoder-decoder transformer architecture with a convolutional backbone. It incorporates two heads atop the decoder outputs to facilitate object detection: a linear layer for class labels and a MLP (multilayer perceptron) for bounding boxes. Object queries are employed by the model to identify objects within an image, with each query dedicated to detecting a specific object. (Zhao et al. 2024). Its architecture is demonstrated in figure 2

Bot-SORT

BoT-SORT introduces a novel approach for robust tracking of multiple objects in surveillance environments. This in-

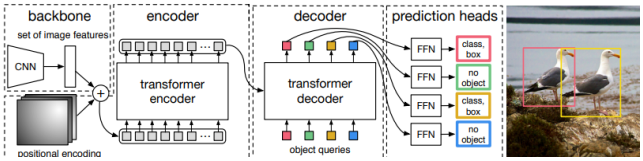

Figure 2: Architecture of DEtection TRansformer (DETR) model

novative methodology tackles common challenges in object tracking systems, such as occlusion, interaction between objects, lighting variations, and rapid movements, in a unique and effective way (Aharon, Orfaig, and Bobrovsky 2022).

Their approach uses the SORT initial object tracking. It then uses a robust re-identification module that leverages learned visual representations in a supervised manner to correctly associate object detections over time, even in challenging situations.

Its methodology consists of Object Detection, Initial Tracking with SORT, Object Re-Identification and Tracking Update. This methodology is specifically designed to handle partial occlusions, temporal jitters, and other variations in the environment that can degrade the performance of traditional tracking systems.

Dataset

We use a publicly available dataset called The Brazilian Traffic Signs in the Wild (BTSW) (Cimirro 2013). This dataset consists of 3613 images, divided into training and validation sets, totaling 4127 annotated traffic signs. Figure 3 shows examples of the data. It was collected by a GoPro mounted on a vehicle traveling along Brazilian roads in Rio Grande do Sul and Bahia. The dataset includes three classes of signs provided for in Brazilian legislation: *Regulatory, Warning, and Directional*.

For the test, we used a 17-minute video collected on the roads in Ceará, Brazil.

(a) (b)

(c) (d)

Figure 3: Images extracted from the Dataset that we use to train the models. Figure (a) and (b) show scenarios with variation in lighting, while (c) and (d) show a sunnier location.

Method

We first trained the three pre-trained YOLOv8 models: small model (YOLOv8s), nano model (YOLOv8n) and medium model (YOLOv8m) and the pre-trained DETR model with the data we had for training, as explained in the section Dataset. Then, we pass the output of these tracking models to the pre-trained Bot-SORT model for the traffic sign re-identification task. This way, our final output contains the class to which each traffic signal belongs and its ID. Therefore, different signs have different IDs.

The Table 1 shows the number of parameters for each model we trained.

Model	Params (M)
YOLOv8n	3.2
YOLOv8s	11.2
YOLOv8m	25.9
RT-DETR-l	42

Table 1: Number of parameters for each model extracted from (Reis et al. 2023) and (Zhao et al. 2024)

And for each model we used the hyperparameters represented in the Table 2 for training.

Hyperparameter	Value
Epochs	100
Batch Size	16
Learning Rate	0.01
Weight Decay	0.0005

Table 2: Hyperparameters description

Results

All results are available on GitHub, at the end of the report. We chose to show only the results of the RT-DETR model due to the limitation in available space in this report. The confusion matrix shown in figure 4 is the most interesting among the four models. This model, despite being the most powerful and being the one that correctly identified the largest number of traffic signs, also identified many backgrounds as signs. In other words, it presented the highest number of false positives compared to the other models we trained.

This behavior of this model in identifying false positives negatively interfered with the result we received from the re-identification model (See Images 5 and 6). While models like the YOLOv8-n had ID 5, the RT-DETR-L model had ID 29 for the same sign, as shown in the Image 6. This is due to the sudden appearance of signs in the background of the video, when in fact there is no sign, as shown in image 5.

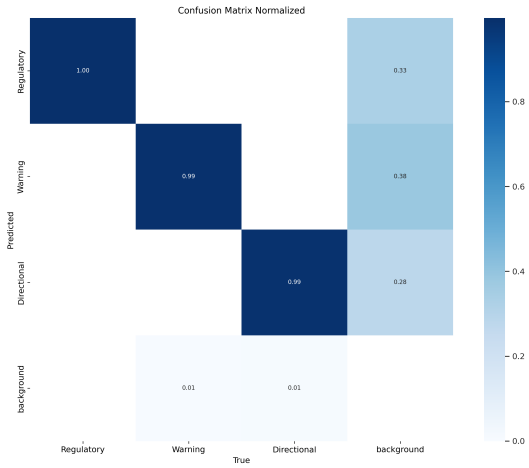

Figure 4: Confusion Matrix of RT-DETR model.

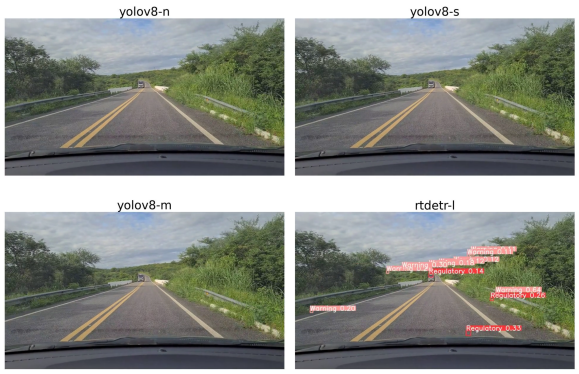

Figure 5: Result of each model in an excerpt from the test video. We can observe that the RT-DETR-L model identified several signs that do not exist in the image, while the other models did not identify any false positives in this timestamp.

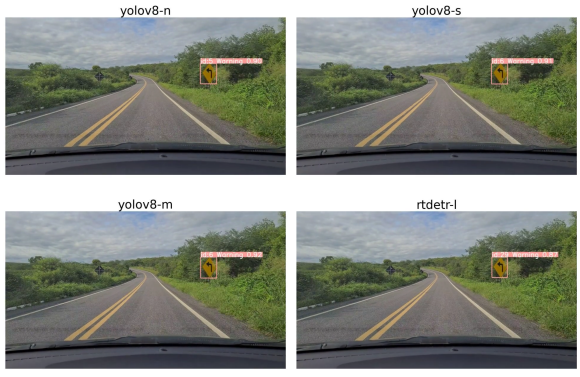

Figure 6: Result of the Re-ID model that received as input the traffic sign detections of each model. We noticed that the Re-ID result of the RT-DETR-L model output was impaired due to the false positive problem that we mentioned.

To be able to understand the difference in the models' behaviors, we show in figures 7 and 8 the precision-recall curves of the RT-DETR-L and YOLOv8-nano models, respectively.

These curves show the tradeoff between precision and recall for different thresholds. A larger area under the curve, as in the case of the YOLOv8-nano model, shown in figure 8, represents that the YOLOv8-nano handled the tradeoff between precision and recall better. While the slightly smaller area, as in the case of RT-DETR, shown in figure 7, indicates that this model dealt worse with the tradeoff between precision and recall, evidenced in the examples shown in the figure 5 and the confusion matrix 4.

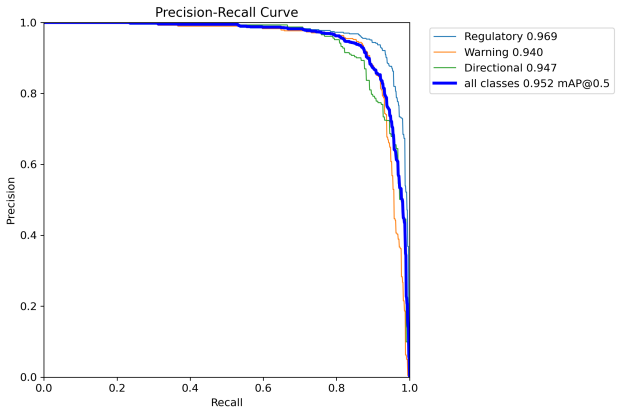

Figure 7: Precision-Recall Curve of the RT-DETR-L model

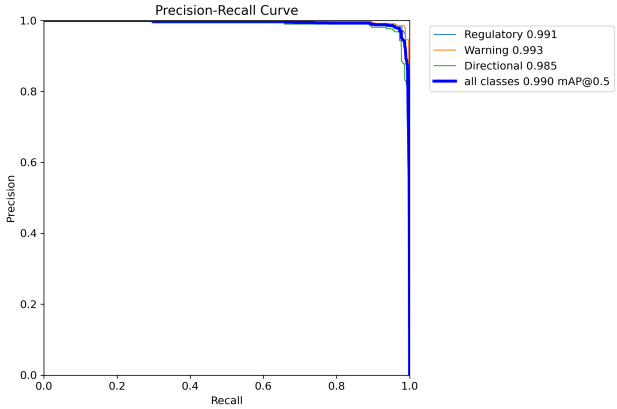

Figure 8: Precision-Recall Curve of the YOLOv8-nano model

Conclusion

In conclusion, our initial objective was to use a network capable of identifying Brazilian traffic signs. We chose three pre-trained models from the YOLOv8 family and one RT-DETR model. We fine-tuned these models with data that was publicly available, containing around 4000 Brazilian road signs.

In the second step, we used the Bot-SORT (Aharon, Orfaig, and Bobrovsky 2022) model for the traffic sign re-identification task. This way, the final result of the evaluation did not contain a track of repeated signs, instead, it contained the ID related to the sign.

We evaluated the performance of our deep learning system using a 17-minute video captured on Brazilian roads. We noticed that the most interesting model was the RT-DETR model, because even though it had a very high number of correct predictions when classifying the signs, it identified several false positives, as illustrated in the results section.

GitHub

Our code is available in the GitHub repository https://github.com/felipeinagaki/traffic_sign_detection.
In this repository we make available the outputs of each model, as well as their metrics and the documentation.

References

Aharon, N.; Orfaig, R.; and Bobrovsky, B.-Z. 2022. BoT-SORT: Robust Associations Multi-Pedestrian Tracking.

Cimirro, J. 2013. Dataset-Placas-Transito. https://github.com/jean2612/Dataset-Placas-Transito.

de la Escalera, A.; Moreno, L.; Salichs, M.; and Armingol, J. 1997. Road traffic sign detection and classification. *IEEE Transactions on Industrial Electronics*, 44(6): 848–859.

Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; and Igel, C. 2013. Detection of traffic signs in real-world images: The German traffic sign detection benchmark. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Reis, D.; Kupec, J.; Hong, J.; and Daoudi, A. 2023. Real-Time Flying Object Detection with YOLOv8. arXiv:2305.09972.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. DETRs Beat YOLOs on Real-time Object Detection. arXiv:2304.08069.

Zhu, Y.; and Yan, W. 2022. Traffic sign recognition based on deep learning. *Multimedia Tools and Applications*, 81: 1–13.