Journal of the Brazilian Computer Society, 2024, 30:1, doi: 10.5753/jbcs.2024.3678

Recognition of Brazilian vertical traffic signs and lights from a car using Single Shot Multi box Detector

Monhel Maudoony Pierre [**Federal University of Uberlandia** | ***monhel.pierre@ufu.br***]
Henrique Fernandes [**Federal University of Uberlandia/ IVHM Centre, School of Aerospace, Transport and Manufacturing, Cranfield University** | ***henrique.fernandes@ufu.br***]

Faculty of Computing, Universidade Federal de Uberlândia, Av. João Naves de Ávila, 2121 - Santa Mônica, Uberlândia - MG, 38400-902, Brazil.

Received: 14 September 2023 • **Accepted:** 19 December 2023 • **Published:** 24 July 2024

Abstract This work presents an automated system for recognizing Brazilian vertical traffic signs and lights using artificial intelligence. The main objective of the system is to contribute to road safety by alerting drivers to potential risks such as speeding, alcohol consumption, and cell phone use, which could lead to severe accidents. The system's core contribution lies in its ability to accurately recognize various traffic signs and lights, providing crucial warnings to drivers. To achieve this, the system utilizes a light version of the single shot multi box detector as its detection algorithm and experiments with three Mobilenet versions as base networks. The optimal Mobilenet version is selected based on a mean average precision higher than 80%, which guarantees reliable detection results. The dataset used for training and evaluation comprises images extracted from YouTube traffic videos, each annotated to create the necessary labels for training. Through this extensive experimentation, the system demonstrates its efficacy in achieving accurate and efficient detection. The results of the experiments are compared with other existing approaches and our work significantly advances the field by providing a tailored dataset, an optimized model, and also valuable insights into traffic sign and light recognition, collectively contributing to the improvement of road safety.

Keywords: Artificial Intelligence, MobileNet, SSD, Traffic Signs, Traffic Lights

1 Introduction

In the relentless pursuit of enhancing road safety and optimizing transportation systems, the integration of cutting-edge technologies is not just advantageous but imperative [Dong, 2018]. Computer vision algorithms, like the single shot multi box detector (SSD), are like smart tools that help cars quickly understand and recognize the objects around us.

The complexity of Brazil's expansive and diverse road network, coupled with its substantial population [Story, 2021], underscores the challenges faced by self-driving cars and driver assistance systems. To navigate through Brazil's diverse landscape, it's crucial to customize these technologies to understand the specific characteristics of Brazilian traffic signs and lights.

According to World Health Organization [2022], traffic accident results in the death or disability of approximately 1.3 million people every year in the world. Speeding, driving under influence of alcohol or any other psychoactive substances, distracted drivers using mobile phones are some of the main reasons why those accidents occur. Brazil was the third country with the most traffic deaths in the world in 2016.

According to Souza *et al*. [2023], Brazil is currently navigating a surge in its vehicle count, an upward trajectory that unfortunately corresponds with a troubling increase in morbidity and mortality rates resulting from traffic accidents.

With those mentioned motivations we noticed that many accidents take place because of the lack of focus from the drivers, so the objective of this work is to develop a system for the recognition of vertical traffic signs and traffic lights by adapting it to the federal regulation of colors and shapes in Brazil. Figure 1 and Figure 2 illustrate the types of Brazilian vertical traffic signs and the three states of traffic lights invented by Garrett Morgan [DeLuca, 2021], respectively.

Figure 1. Brazilian vertical traffic signs. Extracted from [Palmieri, 2021].

The remaining sections of this paper are structured as follows: Section 2 which is about some related works, Section 3 which is the methodology used, Section 4, the experiments, Section 5, the results and discussion and finally, Section 6 which is the conclusions.

Figure 2. Brazilian traffic lights. Extracted from [JDV, 2023].

2 Related works

This section contains a brief review of the literature on traffic signs detection and the detection process.

Alghmgham *et al*. [2019] presented a study on vertical Arabia Saudi traffic sign classification using deep convolutional neural network and the different angles and including other parameters and conditions. The dataset contains a total of 2718 images. The images were collected from three different cities in Arabia Saudi. Theses images were then transformed into grayscale with a dimension of 30×30 pixels. After training process the authors obtained an accuracy of 100% within 150 epochs in 16 different experiments on different number of epochs and batch size numbers.

Pon *et al*. [2018] proposed a hierarchical model built upon the ResNet-50 version of region-based convolutional neural networks (R-CNN). The proposed model is part of the two-stage model detection algorithm and the images of the dataset are exclusively from the United States. After the experiments 54% as accuracy was obtained.

Hoelscher [2017] presented a study on vertical traffic sign recognition techniques in images of complex traffic scenarios. The author tested two approaches for image segmentation and selection of regions of interest. The first one is a color thresholding with fourier descriptors but was not satisfactory and the second one is a color filtering using fuzzy logic together with and algorithm that select stable regions in different shades of gray. The model was used with a Brazilian dataset by the author and along with the German dataset to obtain 93% as extraction accuracy and 95% as classification accuracy.

Santos *et al*. [2020] proposed a real-time traffic sign detection and recognition algorithm using neural network. For the network architecture they used a faster region-based convolutional neural networks (F-RCNN) model with VGG-16 and Inception V4 which had the best result as for feature extraction network and 128×128 pixel as the input of the images. Those images were collected from videos that they took using a camera in front of a car and the frames of the video were extracted to later be augmented using different degrees of rotation on those images where 90% were used for training and 10% for validation. 82% is the accuracy obtained by the best architecture they used.

Bhatt *et al*. [2022] proposed a model for traffic sign detection and recognition using deep learning with convolutional neural networks and a hybrid dataset that includes a reference dataset for German traffic sign recognition from Kaggle and a self-generated Indian traffic sign dataset with an hybrid dataset from the previous ones. For the experiments, 50 epochs were used to train on the German dataset, 15 for the Indian created dataset and 25 epochs on the hybrid dataset which results an accuracy of 95.45% for the hybrid datasets, 91.08% for the Indian dataset, and 99.85% for the German dataset.

Yoneda *et al*. [2020] proposed an algorithm exclusively about traffic lights and the arrow lights, where the method achieved 91.8% and 56.7% as accuracy for traffic lights and arrow lights respectively. YOLOv3, which is one of the one-stage model detection, was used with different processes to be able to detect the arrow lights from the images and an f-value of 91.8% for the traffic lights was obtained.

William *et al*. [2019] proposed an effective solution for real-time traffic sign detection and recognition, specifically addressing challenges related to weather conditions, illumination, and visibility. To achieve this, the authors explored advanced multi-object detection systems, such as faster region-based convolutional neural networks (F-RCNN) and single shot multi box detector (SSD). They also explored various feature extractors including MobileNet v1, Inception v2, and Tiny-YOLOv2. The focus, however, was on evaluating the performance of F-RCNN Inception v2 and Tiny YOLO v2, as they demonstrated the most promising results.

Dalborgo *et al*. [2023] focused on traffic sign recognition systems enabled by embedded systems with internet connections. The implementation of traffic sign recognition systems using convolutional neural networks and datasets for AI training is discussed. The datasets included a new class for traffic sign recognition called vegetation occlusion. The results demonstrated that this approach facilitates faster traffic sign maintenance by utilizing vehicles as moving sensors. The proposed technique enables the identification of irregularities in traffic signs, allowing for timely reporting and fixing of issues, ultimately enhancing traffic safety. The authors also evaluated the performance of various YOLO models based on case studies.

Zhu and Yan [2022] presented an experiment evaluating the performance of the latest version of YOLOv5, a deep learning model, for traffic sign recognition using a dataset created by the authors. The objective was to demonstrate the suitability of deep learning models for traffic sign recognition by comparing YOLOv5 with SSD, another popular object detection algorithm. The experiments utilized the authors' custom dataset. The experimental results showed that YOLOv5 achieved a mean average precision (mAP) of 97.70% for all classes at a threshold of 0.5, whereas SSD achieves a mAP of 90.14% under the same conditions. Furthermore, YOLOv5 demonstrated superior recognition speed compared to SSD.

Zhang *et al*. [2019] focused on the development of lightweight neural networks for traffic sign recognition, specifically designed for resource-constrained environments. The authors proposed two novel lightweight networks that achieve higher recognition precision while minimizing the number of trainable parameters. They utilized knowledge distillation to transfer knowledge from a larger trained model called teacher network to a smaller model called student network. Additionally, the authors pruned redundant channels from the student network by identifying insignificant channels based on the values of batch normalization scaling factors. This resulted in a compact model with comparable accuracy to more complex models. The teacher network achieved an accuracy rate of 93.16% on the CIFAR-10 gen-

eral dataset. Using the knowledge from the teacher network, the student network was trained on the GTSRB and BTSC traffic sign datasets, achieving high accuracy rates of 99.61% and 99.13% respectively, with only 0.8 million parameters.

These studies primarily focused on the detection of traffic signs, traffic lights, and arrow lights. However, it is worth noting that most of these works were limited to the use of two-stage model detection tasks and did not consider the simultaneous detection of traffic signs and lights unlike the work of [Pon *et al*., 2018]. Table 1 showcases a more detailed comparison between some characteristics of theses related works and our proposal.

Table 1. Comparison between the related works and our proposal.

Paper	Classification	Recognition	Traffic signs	Traffic lights	Two-stage	One-stage	CNN model	Brazilian dataset
Hoelscher [2017]		×			×			
Pon *et al*. [2018]			×	×	×			
Alghmgham *et al*. [2019]	×		×				×	
Zhang *et al*. [2019]		×	×				×	
William *et al*. [2019]		×	×		×	×		
Yoneda *et al*. [2020]		×		×	×			
Santos *et al*. [2020]		×	×		×			
Bhatt *et al*. [2022]		×	×				×	
Zhu and Yan [2022]		×	×			×		
Dalborgo *et al*. [2023]		×	×			×		×
Ours		×	×	×		×		×

Given the aforementioned gaps in the existing literature, we aim to contribute by developing a novel system that combines the recognition of Brazilian vertical traffic signs and lights. Additionally, a created dataset of these annotated objects will be presented.

3 Methodology

This section outlines the step-by-step process we followed in selecting our model, placing a primary emphasis on reducing computational demands. To initiate this process, we conducted an in-depth exploration of existing studies, focusing specifically on the single shot multi box detector. Subsequently, after careful consideration, we chose to implement SSD-Lite due to its distinct advantage in requiring less computational power.

The subsequent phase involved the creation of a dataset tailored to our research objectives. This dataset was thoughtfully curated to include images that uniquely featured Brazilian vertical traffic signs and lights, aligning closely with the contextual nuances of our study.

Simultaneously, we engaged in a comprehensive review of widely adopted evaluation metrics utilized in related works. This exploration aimed at establishing a robust framework for objectively assessing the performance of object detection models, ensuring that our evaluation process was well-grounded in established practices.

In the conclusive segment of this methodology, we delve into the specifics of our proposed approach. This includes providing a detailed perspective on the methodologies employed throughout our research. Central to this discussion is the customized implementation of SSD-Lite within our dataset and the nuanced adaptations made to amplify the overall performance of our model. These adaptations were particularly designed to address the unique challenges associated with the detection of Brazilian traffic signs, offering a comprehensive view of our research methodology.

3.1 Dataset creation & augmentation

Our database consists of images extracted from videos, like in the work of [Santos *et al*., 2020] but from a specific YouTube channel[1] . These videos showcase various traffic scenes from different cities in Brazil. To extract the frames from the videos, we utilized the OpenCV library.

Subsequently, the extracted frames were labeled using the labelImg[2] tool. We filtered out the useful images, resulting in a total of 1,363 images. It's worth to be noted that this dataset[3] was specifically created and already used for a master's thesis having the same title than this research paper by the same authors. It served as a foundational resource for our experiments and analyses, contributing significantly to the outcomes presented in this study.

To enhance the diversity of our dataset, we performed manual augmentation on these images. Several techniques were employed, including contrast adjustment, noise addition, linear and sigmoid contrast transformations, channel shuffling, image solarization, and invert the image color. By applying these augmentation methods, we generated a grand total of 55,276 images. Each of these augmented images has an associated annotation file following the pascal visual object classes (Pascal VOC) format.

In our work, we focused on 16 specific types of objects that appeared most frequently in the daily traffic life. These objects were selected for further analysis and classification, resulting in a total of 16 classes within our dataset. Additionally, all input images were standardized to a single color mode and given a consistent size.

Figure 3 showcases the 16 classes used in our work, along with their corresponding descriptions in English and Portuguese. We also demonstrate the visual effects of applying the seven aforementioned augmentation methods on an image in Figure 4 while Figure 5 illustrates our dataset class distribution.

[1] https://www.youtube.com/@DrivinginBrazil
[2] https://pypi.org/project/labelImg
[3] https://data.mendeley.com/datasets/jbpsr4fvg9/2

#	ENGLISH NAME	PORTUGUESE NAME	LABEL	SIGN IMAGE	# OF OBJECTS	#	ENGLISH NAME	PORTUGUESE NAME	LABEL	SIGN IMAGE	# OF OBJECTS
1	Stop sign	Parada obrigatória	000		2.461	9	Road hump	Lombada	025		4.018
2	Give way	Dê a preferência	001		1.518	10	Direction of the way circulation	Sentido de circulação da via/pista	028		9.499
3	No left turn	Proíbido virar à esquerda	003		2.415	11	Trucks keep right	Ônibus, caminhões e veículos de grande porte mantenham-se à direita	035		20.705
4	No right turn	Proíbido virar à direita	004		1.155	12	Bus route	Circulação exclusiva de ônibus	040		1.590
5	No parking	Estacionameno proíbido	007		6.809	12	Cycling	Circulação exclusiva de bicicleta	042		1.505
6	Regular parking	Estacionamento regulamentado	008		2.581	14	Yellow light	Atenção veículos	051		1.550
7	No park and stop	Proíbido parar e estacionar	009		5.401	15	Red light	Parada para veículos	052		4.967
8	Speed limit	Velocidade máxima permitida	023		20.703	16	Green light	Veículos podem seguir	053		16.376
TOTAL: 85.253 objects for 16 classes											

Figure 3. The classes of the created dataset. Source: [Pierre and Fernandes, 2023].

Figure 4. Example of the augmentation methods on one image from the dataset. Source: [Pierre and Fernandes, 2023].

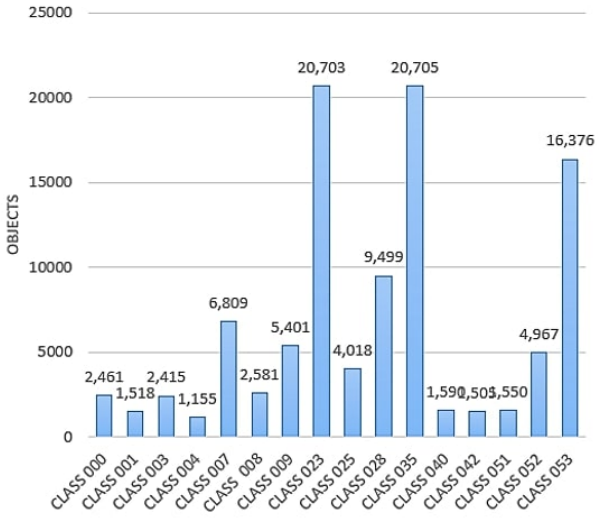

Figure 5. Dataset class distribution. Source: [Pierre and Fernandes, 2023].

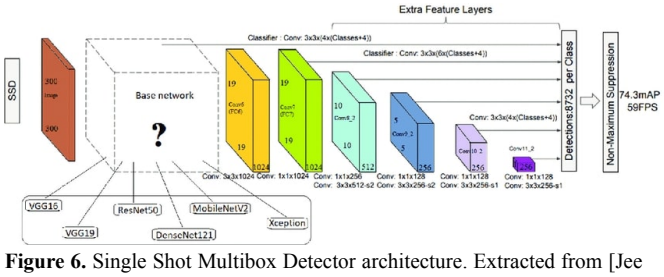

Figure 6. Single Shot Multibox Detector architecture. Extracted from [Jee *et al.*, 2021].

3.2 The model selection

The core architecture employed in our work is the SSD, which consists of three main components, as depicted in Figure 6.

The first component is responsible for extracting features from the input images. That first part can adopt VGG-16 network like in the original paper without dropout layer, FC8 and soft-max classification layers. It replaces the fully connected layers FC6 and FC7 in the ordinary VGG network with convolutional layers Conv6 and Conv7.

The second component is the detection heads, which are responsible for generating bounding boxes and class confidence scores. To create a lighter version of SSD, known as SSD-Lite, certain layers were removed from this component. This optimization allows for a more efficient and streamlined detection process. In that second part, four convolutional layers of Conv8, Conv9, Conv10, and Conv11 have been newly added. Each convolutional layer utilizes a 1×1 convolution kernel for dimensionality reduction and then makes use of a 3×3 convolution kernel for feature extraction. The loss function of the SSD model consists of two parts: The localization loss (Lloc) and the confidence loss (Lconf). The entire loss function is a weighted sum of the localization and confidence losses.

The final component is crucial for eliminating redundant detections and ensuring the best predictions for each object. It employs a mechanism to remove duplicate detections and retain only the most accurate and relevant results. This is achieved by applying a predefined threshold value, typically set at 0.5 or 0.7, depending on the specific dataset and requirements. This architecture forms the foundation of our system where its detection heads and its extra layers are illustrated in Table 2 and Table 3 respectively.

Table 2. SSD-Lite regression/classification heads.

Layer	In	Out	Kernel	Stride
Conv2d	576	576	(3, 3)	(1, 1)
BatchNorm2d	576	—	—	—
RELU6	—	—	—	—
Conv2d	576	68	(1, 1)	(1, 1)
Conv2d	1280	1280	(3, 3)	(1, 1)
BatchNorm2d	576	—	—	—
RELU6	—	—	—	—
Conv2d	1280	102	(1, 1)	(1, 1)
Conv2d	512	512	(3, 3)	(1, 1)
BatchNorm2d	576	—	—	—
RELU6	—	—	—	—
Conv2d	512	102	(1, 1)	(1, 1)
Conv2d	256	256	(3, 3)	(1, 1)
BatchNorm2d	576	—	—	—
RELU6	—	—	—	—
Conv2d	256	68	(1, 1)	(1, 1)
Conv2d	256	256	(3, 3)	(1, 1)
BatchNorm2d	576	—	—	—
RELU6	—	—	—	—
Conv2d	256	68	(1, 1)	(1, 1)
Conv2d	128	128	(3, 3)	(1, 1)
BatchNorm2d	576	—	—	—
RELU6	—	—	—	—
Conv2d	128	68	(1, 1)	(1, 1)

It's to be noted that the first Conv2d layers from every module are in a group of ConvBNReLU layers where they are followed by a BatchNorm2d and ReLu6 layers, except for the last Conv2d layer which is apart from any other group.

Table 3. SSD-Lite extra layers.

Layer	In	Out	Kernel	Stride
Conv2d	1280	256	(1,1)	(1,1)
Conv2d	256	256	(3,3)	(2,2)
Conv2d	256	512	(1,1)	(1,1)
Conv2d	512	128	(1,1)	(1,1)
Conv2d	128	128	(3,3)	(1,1)
Conv2d	128	256	(1,1)	(1,1)
Conv2d	256	128	(1,1)	(1,1)
Conv2d	128	128	(3,3)	(2,2)
Conv2d	128	256	(1,1)	(1,1)
Conv2d	256	64	(1,1)	(1,1)
Conv2d	64	64	(3,3)	(2,2)
Conv2d	64	128	(1,1)	(1,1)

3.3 SSD-Lite base network

The base network of the SSD-Lite architecture is MobileNet. In our implementation, we utilized three different models from the MobileNet family: MobileNet v2, MobileNet v3 (small), and MobileNet v3 (large). These models are employed to extract meaningful features from the images, enabling subsequent detection. It is specifically optimized for efficient inference on devices with low power and memory constraints. Table 4 shows some layers taken out from that Mobilenet v2 for feature extractions but it is composed of 19 modules.

Table 4. Some layers from the Mobilenet v2 SSD-Lite extraction feature.

Layer	In	Out	Kernel	Stride
Conv2d	3	32	(3,3)	(2,2)
Conv2d	32	32	(3,3)	(1,1)
Conv2d	32	16	(1,1)	(1,1)
Conv2d	16	96	(1,1)	(1,1)
Conv2d	96	96	(3,3)	(2,2)
Conv2d	96	24	(1,1)	(1,1)
Conv2d	24	144	(1,1)	(1,1)
Conv2d	144	144	(3,3)	(1,1)
Conv2d	144	24	(1,1)	(1,1)
——-	—-	—-	—-	—-
——-	—-	—-	—-	—-
Conv2d	96	576	(1,1)	(1,1)
Conv2d	576	576	(3,3)	(2,2)
Conv2d	576	160	(1,1)	(1,1)
Conv2d	160	960	(1,1)	(1,1)
Conv2d	960	960	(3,3)	(1,1)
Conv2d	960	160	(1,1)	(1,1)
Conv2d	160	960	(1,1)	(1,1)
Conv2d	960	960	(3,3)	(1,1)
Conv2d	960	160	(1,1)	(1,1)
Conv2d	160	960	(1,1)	(1,1)
Conv2d	960	960	(3,3)	(1,1)
Conv2d	960	320	(1,1)	(1,1)
Conv2d	320	1280	(1,1)	(1,1)

It's to be noted that, like the extra layers of the SSD-Lite, the first Conv2d layers of the mobilenet v2 described in Table 3 are in a group but that time it's a Conv2dNorm activa-

tion layers where every Conv2d layer is followed by a BatchNorm2d and ReLu6 layers, and except for the first and last module where we can find the Conv2d layers apart from any other group.

3.4 Evaluation metrics

Evaluation metrics play a crucial role in assessing the performance of object detection models. Among these metrics, average precision (AP) and mean average precision (mAP) are widely used to evaluate the effectiveness of various object detection models, including faster region-based convolutional neural networks (F-RCNN), mask region-based convolutional neural networks (Mask-RCNN), SSD, YOLO, and others. To understand these metrics, below are the definitions of some terms:

- True Positive (TP) — Correct detection made by the model.
- False Positive (FP) — Incorrect detection made by the detector.
- False Negative (FN) — A Ground-truth missed (not detected) by the object detector.
- True Negative (TN) — This is the background region correctly not detected by the model. This metric is not used in object detection because such regions are not explicitly annotated when preparing the annotations.

After defining the aforementioned terms, there are several other metrics used to assess the performance of a model on data. These metrics provide additional insights into the model's effectiveness in object detection:

- Precision is a metric that quantifies the accuracy or exactness of a model in correctly identifying relevant objects. It is calculated as the ratio of true positives to the total number of detections made by the model. Precision focuses on minimizing false positives, meaning it measures how many of the model's predicted positives are actually true positives.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

- On the other hand, recall measures the model's ability to detect all relevant objects or ground truths. It is calculated as the ratio of true positives to the total number of ground truths. Recall aims to minimize false negatives, indicating how well the model captures all the positives in the dataset.

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

In summary, precision evaluates the model's accuracy in making correct positive predictions, while recall assesses the model's ability to capture all positive instances in the dataset. Both metrics are important for assessing the performance of object detection models.

3.5 Mean average precision

AP@α refers to the area under the precision-recall curve evaluated at the alpha intersection over union threshold. It quantifies the performance of object detection models by measuring the precision and recall trade-off at a specific IoU threshold. A higher value of area under the precision-recall curve indicates higher precision and recall rates.

The precision-recall curve typically exhibits a zig-zag pattern, as it is not necessarily monotonically decreasing. Average precision is calculated individually for each class, resulting in as many AP values as there are classes. These average precision values are then averaged to obtain the mean average precision metric. The mAP provides an overall assessment of the model's performance by taking into account the average precision values across all classes.

Equation 3 and Equation 4 give the formulas related to the AP and the mAP respectively.

$$\text{AP} = \sum_{i=0}^{N-1} [R_i - R_{(i+1)}] * P_{(i)} \tag{3}$$

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{4}$$

3.6 Intersection over union

Commonly called IoU, it is a used metric in object detection to measure the degree of overlap between a predicted bounding box and the ground-truth bounding box, which is manually annotated. It helps evaluate the accuracy of object detection by quantifying the similarity between the predicted and ground-truth bounding boxes. The IoU value ranges from 0 to 1, where a value of 0 indicates no overlap between the boxes, and a value of 1 represents a perfect overlap. A higher IoU value indicates a better alignment between the predicted and ground-truth bounding boxes.

When considering an IoU threshold of α, a true positive refers to a detection where the IoU (ground-truth, predicted) $> \alpha$. A false positive occurs when the IoU (ground-truth, predicted) $< \alpha$ where the model incorrectly predicts the presence of a class. A false negative is a ground truth that was missed when the IoU (ground-truth, predicted) $\leq \alpha$. The formula for calculating IoU is shown in Figure 7, which represents the ratio of the intersection area of the predicted and ground-truth bounding boxes to the union area of the two boxes.

In summary, IoU provides a quantitative measure of the overlap between predicted and ground-truth bounding boxes, assisting in evaluating the accuracy and correctness of object detection models.

$$IOU = \frac{\text{area of overlap}}{\text{area of union}} =$$

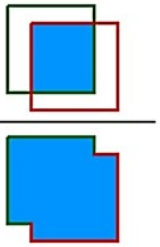

Figure 7. Intersection over union. Extracted from [Padilla *et al*., 2020].

3.7 Proposal

The SSD is a popular object detection framework, and the SSD-Lite model is a lightweight version from it. Figure 8 illustrates the functioning of our system that combines efficiency and accuracy, making it suitable for real-time applications on resource-constrained devices. By leveraging the MobileNet backbone and the SSD framework, it achieves effective object detection for traffic signs and other objects in images or video streams.

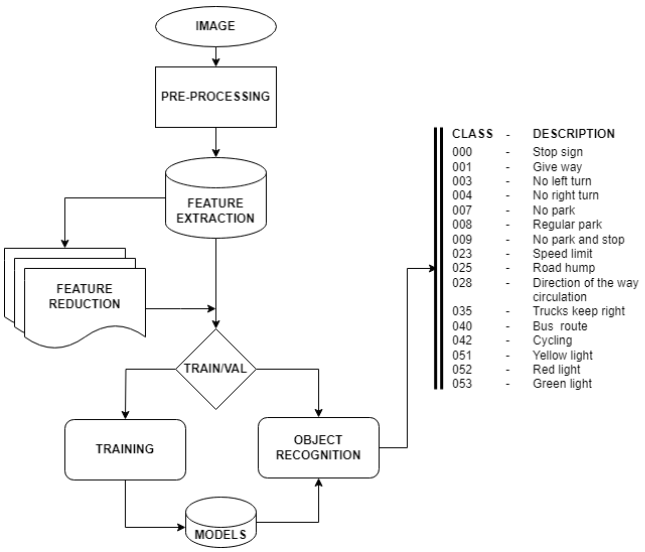

Figure 8. System overview. Adapted from [Alghmgham *et al*., 2019].

4 Experiments

For the training process 70% of the data were used for training, 20% for validating and 10% for testing. Five experiments with different input sizes for the following hyperparameters in Table 5 have been taken out and Table 6 shows the number of parameters for each model.

Table 5. Hyperparameter list.

Hyperparameter	Value
Learning rate	0.001
Batch size	32
Optimizer	SGD
Number of epochs	25
Weight decay	0.00004
Gamma	0.1

Table 6. Number of parameters for each model.

Model	Parameters
Mobilenet v2 SSD-Lite	3.286.326
Mobilenet v3 small SSD-Lite	1.304.522
Mobilenet v3 large SSD-Lite	3.881.522

5 Results and discussion

Our methodology for model evaluation utilized the mean average precision metric, offering a nuanced understanding of our object detection model's performance. The evaluation process highlighted MobileNet v2 with a 320×320 pixel input size as the optimal model, showcasing its superior performance across various classes in the test set, as delineated in Table 7. It is crucial to underscore that our evaluation strategy involved continuous assessments during each training epoch, culminating in the selection of the best-performing model. This dynamic process ensured that the retained model represented the pinnacle of accuracy throughout the training iterations.

Table 7. Training results for every input size.

			VAL DATA
Input	**Base network**	**Day**	**mAP@0.5**
128×128	Mobilenet v2	1.49	0.64
	v3 small	1.11	0.11
	v3 large	1.38	0.54
320×320	Mobilenet v2	3.89	***0.87**
	v3 small	7.15	0.46
	v3 large	5.35	0.84
320×320	Mobilenet v2	3.77	0.79
320×320	Mobilenet v2	3.30	0.78
512×512	Mobilenet v2	8.9	0.77
	v3 small	5.7	0.54

In the practical application of our model to real-world scenarios, particularly using the testing of videos, from the referenced YouTube channel in Section 3.1, our evaluation was a post-training assessment, focusing on the model's performance after each epoch. The best model, identified through these evaluations, was then saved and Figure 9 and Figure 10 show, for that best model, the loss and accuracy graphs respectively.

Throughout our experimentation, an intersection over union threshold of 0.5 was applied, resulting in a mAP of 0.87. Notably, MobileNet v2 with a 320×320 input size demonstrated superiority in accuracy. However, models with larger input sizes, such as 512×512 pixels, introduced challenges, including heightened computational time and slower object detection in videos and images. Consequently, only two experiments were conducted using the 512×512 input size, emphasizing the careful consideration of computational efficiency in our methodology.

On the other hand, the 128×128 input size showcased faster object detection, yet this advantage came with a trade-off—a decrease in the mAP due to the loss of certain features during training. A distinctive aspect of our work is the comprehensive consideration of the classification and localization of Brazilian traffic signs and lights in images, filling a gap in prior studies. For a visual exposition, refer to Figure 11, illustrating examples from the test set, and Figure 12, presenting selected frames from the referenced videos.

In comparison to existing approaches outlined in Section 2, our automated system for recognizing Brazilian vertical traffic signs and lights using artificial intelligence presents no-

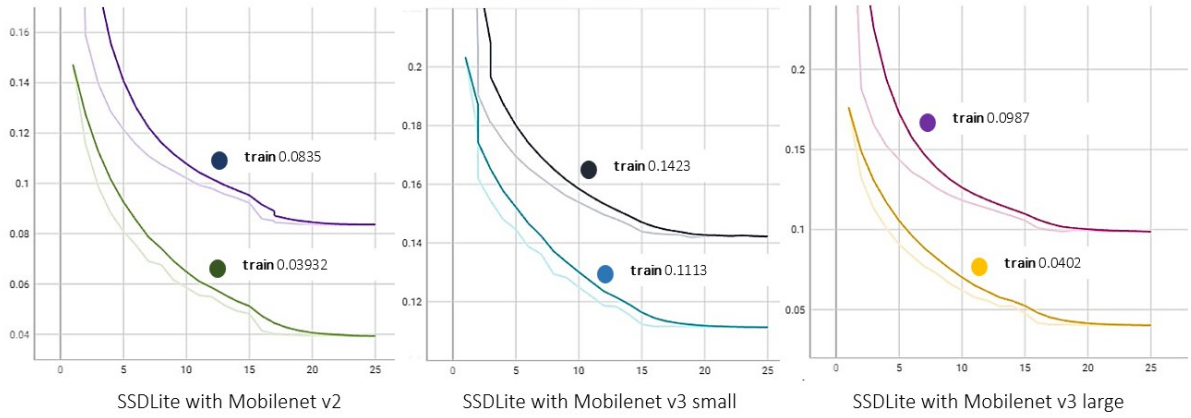

Figure 9. Loss for input size 320. Source: [Pierre and Fernandes, 2023].

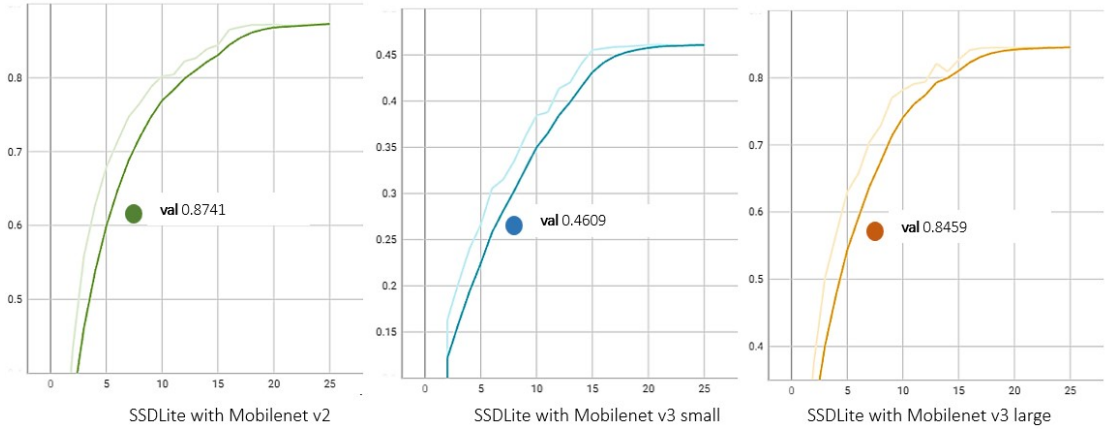

Figure 10. Accuracy for input size 320. Source: [Pierre and Fernandes, 2023].

table advancements. Our proposal, as determined by a mean average precision exceeding 80%, showcases its superior performance across various traffic sign and light classes. This surpasses the achievements of several previous studies. For instance, Alghmgham *et al*. [2019] achieved 100% accuracy in vertical Arabia Saudi traffic sign classification, but our approach, tailored to Brazilian regulations, extends its effectiveness within less epochs to a more diverse set of classes while being able to recognize several objects in just one image. The same case for the works of [Zhu and Yan, 2022], [Zhang *et al*., 2019] and [Yoneda *et al*., 2020]. Similarly, our model outperforms the work of [Pon *et al*., 2018], and [Santos *et al*., 2020] in terms of accuracy where they are less than 83%.

The consideration of multiple input sizes, including the optimal 320×320, larger 512×512, and smaller 128×128, emphasizes our methodology's flexibility and sensitivity to computational efficiency. This stands out in contrast to some works, like [Yoneda *et al*., 2020], which achieved lower accuracies for specific classes.

Post-testing, an insightful observation surfaced—lower accuracy, especially for the 'No right turn' class in Table 8. Subsequent investigation unveiled that this discrepancy was influenced by an increased presence of noisy images in the test data compared to the training and validation datasets. Despite encountering challenges, such as a discrepancy in accuracy for specific classes influenced by noisy test data, our model displayed robustness in handling real-world complexities.

Table 8. Accuracy of every class on the test data.

	TEST DATA	
Class name	**Label**	**AP@0.5**
Stop sign	000	0.80
Give way	001	0.76
No left turn	003	0.78
No right turn	004	0.60
No park	007	0.79
Regular park	008	0.79
No park and stop	009	0.79
Speed limit	023	0.79
Road hump	025	0.80
Direction of the way circulation	028	0.79
Trucks keep right	035	0.79
Bus route	040	0.79
Cycling	042	0.79
Yellow light	051	0.80
Red light	052	0.76
Green light	053	0.79
	mAP@0.5	0.78

The best model has also been tested with some videos from the earlier mentioned YouTube channel in Section 3.1. It's to be noted that we do not have an accuracy for the training

process here but, in fact, the model was evaluated after each epoch and only the best model after the evaluation with its information was saved.

6 Conclusions

In this research paper, our primary focus was on developing a robust system for the recognition of vertical traffic signs and traffic lights.

To achieve our goal, we conducted a series of experiments using three different base networks in combination with the lightweight version of Single Shot Multi box Detector. After rigorous experimentation and analysis, we obtained remarkable results.

Our system achieved an impressive accuracy rate of 87.4% in recognizing Brazilian vertical traffic signs and traffic lights. This achievement can be attributed to the selection of the second version of Mobilenet as the best-performing base network, combined with an input image size of 320×320 pixels.

In conclusion, our research presents a highly accurate and efficient system for the recognition of Brazilian vertical traffic signs and lights where it achieves a processing speed of 30 frames per second (FPS) as we can see in Figure 12, indicating that each frame is processed in approximately 0.0333 seconds. We believe that our findings contribute significantly to the field of object detection and can pave the way for improved traffic management systems, ensuring safer and more efficient road transportation.

6.1 Contributions

Our research has yielded impactful contributions by addressing key challenges in traffic sign and light detection for the Brazilian context.

Foremost, we introduced a crafted dataset dedicated to this specific domain, significantly enriching the available resources for object detection in Brazil.

Beyond dataset creation, our work presents a noteworthy advancement with the introduction of a lightweight model based on the Single Shot Multi box Detector. This model, tailored for the nuances of traffic sign and light detection, strikes a balance between accuracy and computational efficiency. Leveraging a one-stage detection approach, we successfully mitigated computational memory requirements while ensuring reliable detection performance.

Furthermore, our experiments not only validated the effectiveness of our model but also provided a deeper insight into the intricacies of traffic signs, lights, and the detection processes employed. This enhanced understanding stands as a valuable contribution, with potential implications for future advancements in the broader domain of object detection, particularly within the realm of traffic management and safety.

While our research has made significant strides in addressing critical challenges in traffic sign and light detection for the Brazilian context, it is crucial to acknowledge certain limitations. Despite the crafting of our dedicated dataset, the challenges of generalizing the model to diverse real-world scenarios, encompassing varied environmental conditions and potential data biases, remain pertinent.

Additionally, our lightweight model, while achieving a balance between accuracy and computational efficiency, may encounter constraints in handling certain classes or specific environmental nuances. The one-stage detection approach, although successful in mitigating computational memory requirements, might pose challenges in scenarios with increased complexity. These limitations underscore the need for ongoing research to refine and expand our model's capabilities, especially in addressing specific class-related challenges and ensuring robust performance in diverse real-world conditions.

6.2 Future work

In terms of future work, there are several potential implementations that align with the theme of our research.

Firstly, it would be valuable to conduct a comparative analysis with RetinaNet, another SSD-based model, to assess its performance as a classification technique for traffic sign and light recognition. Such a comparison could provide insights into the strengths and weaknesses of different SSD base models and potentially lead to improvements in accuracy and efficiency. The use of another type of optimizer like the adaptive moment estimation (ADAM) and f-value as another metric to better understand how the model learns could also be considered.

Expanding the scope of the dataset to include a broader range of traffic signs and the new type of traffic lights with four colors while augmenting the number of objects for certain classes would also be a worthwhile endeavor.

For practical uses, it would be advantageous to transform the trained model into the open neural network exchange (ONNX) format. According to Choudhary [2023], it is a freely accessible format tailor-made for deep learning models, facilitating seamless transfer between various frameworks with minimal preparation and without the necessity of rewriting the models. That would enable its integration into mobile applications for real-life testing facilitating the deployment of the system in a practical setting, allowing for validation and performance evaluation under real-world conditions.

Lastly, incorporating text-to-speech functionality in the system could enhance driver safety and attention on the road. By providing auditory descriptions of detected signs, drivers can focus on the road instead of constantly looking at the screen. This feature can contribute to a more user-friendly and distraction-free experience.

Declarations

Acknowledgements

The authors are grateful to the Brazil PAEC OEA-GCUB scholarship program for granting a master scholarship to the first author; to the Federal University of Uberlândia and to the team at the Faculty of Computer Science. The authors also thank the reviewers of this article and the editor of the Journal of Brazil Computer Science, for

Figure 11. Example of detection on the test set. Source: [Pierre and Fernandes, 2023].

Figure 12. Some detection examples from video. Source: [Pierre and Fernandes, 2023].

their criticisms and suggestions throughout the submission, review and publication processes.

Funding

The development of the project under the title *"Recognition of Brazilian vertical traffic signs and lights from a car using Single Shot Multi box Detector"* was totally financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and also by the Fondation Connaissance et Liberté - Haiti (FOKAL[4]).

Authors' Contributions

- Monhel : Conducted the research, designed and performed the experiments, generated the results, created the dataset, selected the model, and wrote the original draft of the document. - Henrique : Advised on the type of model to use. Guided on where and how to obtain dataset images and provided instructions on creating the dataset. Reviewed and corrected the document for consistency and accuracy.

Competing interests

The authors declare that they have no competing interests

Availability of data and materials

Data can be made available upon request

References

Alghmgham, D. A., Latif, G., Alghazo, J., and Alzubaidi, L. (2019). Autonomous traffic sign (ATSR) detection and recognition using deep CNN. *Procedia Computer Science*, 163:266–274. DOI: 10.1016/j.procs.2019.12.108.

Bhatt, N., Laldas, P., and Lobo, V. B. (2022). A real-time traffic sign detection and recognition system on hybrid dataset using cnn. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*. IEEE. DOI: 10.1109/icces54183.2022.9835954.

Choudhary, A. S. (2023). ONNX Model | Open Neural Network Exchange. Available at:https://www.analyticsvidhya.com/blog/2023/07/onnx-model-open-neural-network-exchange/. Accessed 27-11-2023.

Dalborgo, V., Murari, T. B., Madureira, V. S., Moraes, J. G. L., Bezerra, V. M. O. S., Santos, F. Q., Silva, A., and Monteiro, R. L. S. (2023). Traffic sign recognition with deep learning: Vegetation occlusion detection in brazilian environments. *Sensors*, 23(13):5919. DOI: 10.3390/s23135919.

DeLuca, L. (2021). Black Inventor Garrett Morgan Saved Countless Lives with Gas Mask and Improved Traffic Lights — scientificamerican.com. Available at:https://www.scientificamerican.com/article/black-inventor-garrett-morgan-saved-countless-lives-with-gas-mask-and-improved-traffic-lights/. Accessed 11-04-2024.

Dong, X. (2018). *Research on Road Transportation Safety Management*, page 160–166. Springer International Publishing. DOI: 10.1007/978-3-030-00214-5_21.

Hoelscher, I. G. (2017). Detecção e classificação de sinalização vertical de trânsito em cenários complexos. Master's thesis, Universidade Federal do Rio Grande do Sul. Available at:https://lume.ufrgs.br/handle/10183/163777.

JDV (2023). O que significa as cores do semáforo. Available at: https://www.jdv.com.br/wp-content/uploads/2023/04/o-que-significa-as-cores-do-semaforo.webp. Accessed on: 16 July 2024.

Jee, G., Gm, H., Gourisaria, M. K., Singh, V., Rautaray, S. S., and Pandey, M. (2021). Efficacy determination of various base networks in single shot detector for automatic mask localisation in a post covid setup. *Journal of Experimental amp; Theoretical Artificial Intelligence*, 35(3):345–364. DOI: 10.1080/0952813x.2021.1960638.

Padilla, R., Netto, S. L., and da Silva, E. A. B. (2020). A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE. DOI: 10.1109/iwssip48289.2020.9145130.

Palmieri, N. (2021). Sinais de Trânsito que todo motorista precisa conhecer. Available at:https://www.despachantedok.com.br/blog/multas-de-transito/sinais-de-transito-que-todo-motorista-precisa-conhecer/. Accessed 31-01-2023.

Pierre, M. M. and Fernandes, H. C. (2023). Recognition of Brazilian vertical traffic signs and lights using Single Shot Multi box Detector. Master's thesis, Universidade Federal de Uberlândia. Available at: https://repositorio.ufu.br/handle/123456789/39298.

Pon, A., Adrienko, O., Harakeh, A., and Waslander, S. L. (2018). A hierarchical deep architecture and mini-batch selection method for joint traffic sign and light detection. In *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE. DOI: 10.1109/crv.2018.00024.

Santos, D. C., Silva, F. A. d., Pereira, D. R., Almeida, L. L. d., Artero, A. O., Piteri, M. A., and Albuquerque, V. H. (2020). Real-time traffic sign detection and recognition using cnn. *IEEE Latin America Transactions*, 18(03):522–529. DOI: 10.1109/tla.2020.9082723.

Souza, S. S., Santos, M. F., and Souza, G. M. S. (2023). Incapacidade em motociclistas envolvidos em acidente de trânsito. *Research, Society and Development*, 12(4):e12112441047. DOI: 10.33448/rsd-v12i4.41047.

Story, E. (2021). Road building in brazil. *Oxford Research Encyclopedia of Latin American History*. DOI: 10.1093/acrefore/9780199366439.013.992.

William, M. M., Zaki, P. S., Soliman, B. K., Alexsan, K. G., Mansour, M., El-Moursy, M., and Khalil, K. (2019). Traffic signs detection and recognition system using deep learning. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE. DOI: 10.1109/icicis46948.2019.9014763.

World Health Organization (2022). Road traffic injuries. Available at: https://www.who.int/news-room/

[4]https://www.fokal.org/

fact-sheets/detail/road-traffic-injuries. Accessed 31-03-2023.

Yoneda, K., Kuramoto, A., Suganuma, N., Asaka, T., Aldibaja, M., and Yanase, R. (2020). Robust traffic light and arrow detection using digital map with spatial prior information for automated driving. *Sensors*, 20(4):1181. DOI: 10.3390/s20041181.

Zhang, J., Wang, W., Lu, C., Wang, J., and Sangaiah, A. K. (2019). Lightweight deep network for traffic sign classification. *Annals of Telecommunications*, 75(7–8):369–379. DOI: 10.1007/s12243-019-00731-9.

Zhu, Y. and Yan, W. Q. (2022). Traffic sign recognition based on deep learning. *Multimedia Tools and Applications*, 81(13):17779–17791. DOI: 10.1007/s11042-022-12163-0.