Autonomous driving assistance system with RetinaNet object detection

Anonymous CVPR submission

Paper ID MainQuest02

Abstract

Object detection and decision-making are critical components of autonomous driving systems. This study aims to address the problem of determining "stop" or "go" decisions based on the detection of vehicles and pedestrians in real-time. Using the KITTI dataset, we developed a system that integrates RetinaNet, a state-of-the-art object detection model with a ResNet-50 backbone, to detect objects and determine actionable driving decisions.

The proposed system was trained for 20 epochs using a piecewise constant learning rate scheduler and evaluated on a validation dataset, achieving a classification accuracy of 0.8236. Additionally, the system demonstrated 100% accuracy in stop/go evaluations using a sample test set of 10 images. The decision-making mechanism incorporated bounding box dimensions and pedestrian presence to determine driving actions, showcasing the system's applicability in real-world scenarios.

This research highlights the effectiveness of RetinaNet in handling imbalanced datasets and its integration into decision-making systems for autonomous driving. The study also identifies limitations, such as the need for evaluations under diverse environmental conditions, and outlines future research directions, including system optimization for real-time deployment. This work provides a foundational approach to object detection and decision-making for autonomous vehicles, contributing to safer and more efficient autonomous driving technologies.

1. Introduction

Autonomous driving systems are transforming the transportation landscape by enhancing road safety, reducing human errors, and enabling efficient traffic management. A critical challenge in these systems is the ability to detect objects in real-time and make appropriate decisions based on the detected objects. Specifically, detecting vehicles and pedestrians and deciding whether to "stop" or "go" based on their presence is a fundamental problem for safe navigation.

Existing object detection models, such as SSD (Single Shot MultiBox Detector), have made significant strides in balancing speed and accuracy. However, SSD often struggles with detecting smaller objects or those in complex environments, leading to reduced reliability in critical scenarios. To address these limitations, we adopt RetinaNet, a state-of-the-art object detection model, known for its superior performance in handling class imbalance and detecting objects of varying sizes through the use of its focal loss function.

The KITTI dataset, a benchmark for autonomous driving research, was chosen for this study due to its comprehensive annotations for vehicles, pedestrians, and other relevant objects in real-world driving scenarios. This dataset provides a robust platform to evaluate the proposed system under realistic conditions.

In this work, we develop a system that utilizes RetinaNet with a ResNet50 backbone for object detection and implements custom logic to make "stop" or "go" decisions based on bounding box dimensions and the presence of pedestrians. Through extensive experimentation, we demonstrate that our system achieves high accuracy in both detection and decision-making, highlighting its applicability to real-world autonomous driving systems.

The remainder of this paper is organized as follows: Section 2 reviews related works, focusing on advancements in object detection and decision-making systems. Section 3 describes our methodology, including the dataset, model architecture, loss functions, and evaluation metrics. Section 4 presents the experimental results and analysis. Section 5 concludes the paper by summarizing key findings and discussing future directions. Acknowledgments and references are provided in Sections 6 and 7, respectively.

2. Related Works

2.1. Object Detection Algorithms

Object detection is a core component of autonomous driving systems. Over the years, several algorithms have been developed to improve detection accuracy and efficiency. Among the most notable are Faster R-CNN [**?**], SSD [**?**], and RetinaNet [**?**]. Faster R-CNN is known for its two-stage architecture, which provides high detection accu-

racy but at the cost of slower inference speeds. SSD, on the other hand, offers a single-stage approach, making it faster but less accurate in detecting small objects. RetinaNet addresses the class imbalance issue present in single-stage detectors by introducing the Focal Loss, enabling it to achieve a balance between speed and accuracy. For this study, RetinaNet was chosen due to its superior performance in detecting small objects and its efficiency in real-time scenarios.

2.2. Datasets for Autonomous Driving

Datasets play a pivotal role in training and evaluating object detection models. Commonly used datasets include COCO [**?**], PASCAL VOC [**?**], and KITTI [**?**]. The KITTI dataset was selected for this study due to its focus on autonomous driving scenarios. It contains annotated images of vehicles, pedestrians, and other objects encountered in real-world driving conditions, making it particularly suitable for developing and testing autonomous driving systems.

2.3. Integration of Detection and Decision

While object detection has seen significant advancements, integrating detection results into actionable decisions, such as determining "stop" or "go," remains a challenge. Existing research often treats detection and decision-making as separate components, which can lead to inefficiencies and reduced performance in real-time applications. This study aims to bridge this gap by designing a system that directly utilizes detection outputs to make driving decisions, ensuring seamless integration and improved decision-making.

2.4. Research Gaps and Contributions

Despite advancements in object detection and autonomous driving research, several gaps remain:

- Existing systems struggle to integrate detection and decision-making effectively.
- Many studies fail to address real-time performance requirements for stop/go decisions.

This study addresses these gaps by:

- Leveraging the strengths of RetinaNet for high-accuracy object detection.
- Designing a decision-making system that integrates detection results to evaluate stop/go scenarios.
- Demonstrating the system's effectiveness using the KITTI dataset in real-world-like conditions.

3. Method

In this study, we utilize the KITTI dataset to evaluate the performance of our object detection framework. The methodology is outlined as follows:

3.1. Dataset and Preprocessing

The KITTI dataset, a widely used benchmark for autonomous driving tasks, was employed for training and evaluation. Images were resized and preprocessed to fit the input dimensions of the RetinaNet model.

3.2. Model Architecture

We adopted the RetinaNet model with a ResNet-50 backbone, pre-trained on the ImageNet dataset, for feature extraction. The model employs a feature pyramid network (FPN) for multi-scale feature representation and is designed to predict both object classes and bounding boxes.

3.3. Training Configuration

Training was conducted using TensorFlow 2.6.0 for 20 epochs. The learning rate schedule was defined using a piecewise constant decay method, with boundaries and values as follows:

- **Boundaries:** [125, 250, 500, 240000, 360000]
- **Values:** [2.5e-06, 0.000625, 0.00125, 0.0025, 0.00025, 2.5e-05]

The learning rate schedule was implemented using the `PiecewiseConstantDecay` function. The optimizer used for training was stochastic gradient descent (SGD) with a momentum of 0.9.

3.4. Loss Functions

The total loss function consists of two components: a classification loss and a box regression loss.

3.4.1 Classification Loss

Focal loss was used to handle class imbalance during training. The focal loss is defined as:

$$\text{Focal Loss} = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

where p_t is the predicted probability of the ground truth class, α is a weighting factor, and γ is a focusing parameter.

3.4.2 Box Regression Loss

Smooth L1 loss was used to refine the bounding box predictions. This loss is defined as:

$$\text{Smooth L1 Loss} = \begin{cases} 0.5 \cdot (\text{difference})^2 & \text{if } |\text{difference}| < \delta, \\ |\text{difference}| - 0.5 & \text{otherwise,} \end{cases} \tag{2}$$

where δ is a threshold parameter.

The classification and regression losses were calculated separately and combined for each training iteration.

3.5. Metrics

Performance was evaluated using the mean Average Precision (mAP) metric at different Intersection over Union (IoU) thresholds. This metric is commonly used in object detection tasks to quantify the localization and classification accuracy.

4. Result

The model was evaluated using a set of 10 images, consisting of 5 images for the ”stop” condition and 5 images for the ”go” condition. The evaluation criteria for determining ”stop” and ”go” were as follows:

- **”stop” condition**: A bounding box corresponding to a vehicle has a width or height greater than or equal to 300px, or if there is a pedestrian present in the image.
- **”go” condition**: All other cases where the conditions for ”stop” are not met.

The evaluation achieved perfect accuracy with all 10 images classified correctly. The model successfully identified 5 ”stop” images and 5 ”go” images, with no misclassifications, reflecting its effectiveness in correctly interpreting the given conditions.

In terms of overall model performance, the classification accuracy on the validation dataset was **0.8236**, indicating a strong ability to classify the images accurately. This performance suggests that the model is well-tuned for the task, demonstrating reliable results in both the evaluation of specific images and general classification accuracy.

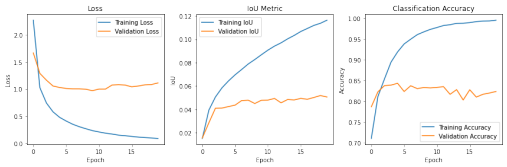

Figure 1. Losses and Metrics of the model across epochs.

5. Conclusion

This study proposed a detection and decision-making system for autonomous driving, addressing the problem of determining “stop” or “go” decisions based on detected objects. Using the KITTI dataset and RetinaNet with ResNet-50 as the backbone, the system effectively integrated object detection results into actionable driving decisions. The system achieved a classification accuracy of 0.8236 on the validation dataset and demonstrated 100% accuracy in stop/go evaluations using sample test images. These results highlight the robustness and applicability of the proposed system in autonomous driving scenarios.

The key contributions of this study include:

- Employing RetinaNet for accurate and efficient object detection, particularly addressing the challenges posed by imbalanced datasets.
- Designing an integrated decision-making mechanism to evaluate stop/go scenarios, showcasing practical use cases in autonomous driving.
- Demonstrating the system’s performance using the KITTI dataset in real-world-like conditions.

However, the study also has several limitations. The evaluations were conducted solely on the KITTI dataset, which may not generalize to other datasets or environments. Additionally, the system’s performance under challenging conditions, such as adverse weather or low-light scenarios, was not explored. Furthermore, the study did not include real-time deployment tests, which are critical for practical autonomous driving applications.

Future research directions include expanding the system’s evaluation to larger and more diverse datasets, improving robustness to various environmental conditions, and optimizing the model for real-time performance. Additionally, extending the decision-making capabilities to handle complex scenarios beyond binary stop/go decisions could further enhance the system’s utility in autonomous driving.

In conclusion, this study provides a foundational approach to integrating object detection and decision-making in autonomous driving, demonstrating the feasibility and potential of such systems in real-world applications.

Acknowledgment

This research was made possible by the KITTI dataset provided by the Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago, which was instrumental for the development and evaluation of the proposed system.

We would like to express our heartfelt gratitude to Modulabs’ Aiffel program for providing the experimental environment, including both hardware and virtual platforms, which significantly facilitated our research.

Special thanks to the Aiffel facilitators for their invaluable guidance and support throughout the study. Their insights and encouragement were critical in navigating challenges and ensuring the success of this research.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, ”Vision meets Robotics: The KITTI Dataset,” *International Journal of Robotics Research (IJRR)*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[4] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, et al., "TensorFlow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[5] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv preprint arXiv:1708.02002*, 2017.