Object Detection in Autonomous Driving using RetinaNet

Daehyeon Ko

July 25, 2024

Abstract

This study aims to enhance the safety and accuracy of autonomous vehicles by using RetinaNet for object detection. Experiments were conducted to distinguish between humans and cars using the KITTI dataset. The two main experiments evaluated the model's object detection performance and go/stop decision-making based on human presence and vehicle size. The results showed high accuracy in the first experiment and a score of 90 in the second experiment, demonstrating the potential of real-time object recognition in autonomous driving.

1 Introduction

1.1 Background

Autonomous driving technology is at the forefront of innovative changes in the automotive industry. Autonomous vehicles (AVs) can drive without human intervention, using a combination of sensors and algorithms. This technology has the potential to reduce traffic accidents, alleviate congestion, and enhance the convenience of mobility. For AVs to become mainstream, they must accurately perceive their surroundings and respond in real-time. Object detection is a critical technology in autonomous systems, identifying and classifying vehicles, pedestrians, cyclists, and traffic signals to help determine the vehicle's path and speed.

1.2 Applications of Autonomous Driving

Autonomous driving technology is utilized in various fields:

- **Transportation**

  Autonomous taxis and buses can reduce traffic congestion and improve passenger convenience. Autonomous trucks can enhance efficiency and reduce

driver fatigue in long-distance freight transport.

- **Logistics**

  Autonomous drones and robots automate warehouse operations, increase delivery speed, and reduce costs.

- **Agriculture**

  Autonomous farm machinery automates crop harvesting and management, increasing productivity and reducing labor costs.

- **Military**

  Autonomous technology performs dangerous missions using unmanned vehicles and drones, protecting soldiers.

1.3 Advantages and Disadvantages of Autonomous Driving

Advantages

- **Safety**

  AVs can reduce human errors and decrease traffic accidents.

- **Convenience**

  Drivers can engage in other activities while the vehicle is in motion.

- **Efficiency**

  AVs can select optimal routes, reduce fuel consumption, and alleviate traffic congestion.

Disadvantages

- **Technical Limitations**

  AVs may not handle complex road situations perfectly.

- **Legal Issues**

  Liability in case of accidents involving AVs must be clarified.

- **Security Risks**

  AVs can be vulnerable to hacking and other security threats.

1.4 Related Works

1.4.1 Overview of Object Detection Technologies

Object detection is a key research area in computer vision, with various models and algorithms developed over time. This section introduces major object detection technologies and their working principles.

1.4.2 One-Stage Detectors

YOLO (You Only Look Once)

- **Working Principle**

  YOLO divides the input image into an SxS grid. Each grid cell predicts the presence of an object, its bounding box, and class probabilities. Bounding boxes with high confidence scores are selected for final object detection.

- **Advantages**

  Enables real-time object detection with fast processing speed.

- **Disadvantages**

  Performance may decrease with small objects or complex scenes.

SSD (Single Shot Multibox Detector)

- **Working Principle**

  SSD generates multiple default boxes from different feature maps and predicts whether these boxes contain objects. Each feature map detects objects at various scales and aspect ratios. SSD predicts all bounding boxes and class scores in a single network pass.

- **Advantages**

  Balances speed and accuracy effectively.

- **Disadvantages**

  May have lower detection performance for small objects.

1.4.3 Two-Stage Detectors

R-CNN (Region-Based Convolutional Neural Networks)

- **Working Principle**

  R-CNN divides the input image into multiple region proposals, transforms each proposal into a feature map using CNN, and predicts the object class and bounding box. Selective Search algorithm is used for region proposals.

- **Advantages**

  Provides high accuracy.

- **Disadvantages**

  Slow processing speed and high memory usage.

Fast R-CNN

- **Working Principle**

  Fast R-CNN processes the entire image using CNN to generate feature maps. ROI Pooling extracts fixed-size feature maps from each region proposal, which are then used to classify objects and predict bounding boxes.

- **Advantages**

  Significantly improves speed over R-CNN.

- **Disadvantages**

  Region proposal stage can still be a bottleneck.

Faster R-CNN

- **Working Principle**

  Faster R-CNN uses a separate network, the Region Proposal Network (RPN), to generate region proposals quickly. These proposals are then used for accurate object detection and classification.

- **Advantages**

  Provides high accuracy and faster speed.

- **Disadvantages**

  Complex structure makes it challenging to implement.

1.4.4 RetinaNet

RetinaNet combines the speed of one-stage detectors with the accuracy of two-stage detectors. It introduces Focal Loss to address class imbalance, significantly improving the detection of small objects.

Architecture

- **Backbone Network**

  RetinaNet uses a backbone network, typically ResNet or ResNeXt, to extract features from the input image. The backbone network generates a feature pyramid, which is used by the subsequent layers to predict objects at different scales.

- **Feature Pyramid Network (FPN)**

  The FPN enhances the backbone by creating a rich, multi-scale feature pyramid. It allows the network to detect objects of various sizes more effectively. The FPN generates multiple levels of features from high-resolution to low-resolution, capturing both fine and coarse information.

- **Classification Subnet**

  This subnet predicts the probability of objects being present at each spatial location for every anchor box and for every class. It uses a series of convolutional layers applied to each level of the feature pyramid.

- **Regression Subnet**

  This subnet predicts the offset for the bounding boxes relative to the anchor boxes. Like the classification subnet, it also applies a series of convolutional layers to each level of the feature pyramid.

Focal Loss

- RetinaNet introduces Focal Loss to address the issue of class imbalance in object detection. Focal Loss modifies the standard cross-entropy loss by adding a factor that down-weights easy examples, focusing the training on hard examples where the model struggles.

- The formula for Focal Loss is:

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

  where p_t is the model's estimated probability for the true class, α_t is a weighting factor, and γ is a focusing parameter. This loss function helps the model learn better from challenging examples and improve overall performance.

1.4.5 Object Detection in Autonomous Driving

Object detection in autonomous vehicles is crucial for real-time perception and safe navigation. Autonomous systems use multiple sensors (cameras, LiDAR, radar) to detect objects and control the vehicle's path. Object detection models identify people, vehicles, bicycles, and traffic signals from sensor data, providing real-time information for safe driving.

Key technologies in autonomous driving

- **Sensor Fusion**

  Combines data from LiDAR, cameras, and radar to accurately perceive the environment.

- **Path Planning**

  Plans optimal routes in real-time for safe and efficient driving.

- **Control Systems**

  Controls vehicle speed and direction to follow the planned path.

- **Communication Technologies**

  Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications share real-time information and respond to traffic conditions.

1.5 Research Position

This study aims to overcome the limitations of existing models and demonstrate the feasibility of real-time detection using RetinaNet in autonomous vehicles. It evaluates the ability to distinguish between humans and cars, contributing to the safety and accuracy of autonomous driving.

2 Method

2.1 Dataset

The experiments used the KITTI dataset, which reflects various road environments for autonomous vehicles. The KITTI dataset consists of images with annotations for cars, pedestrians, cyclists, and other objects.

2.2 Experiment 1: Object Detection Performance

The first experiment evaluated how well RetinaNet detects objects. The model focused on distinguishing between cars and humans, using metrics such as accuracy, precision, and recall for evaluation.

2.3 Experiment 2: Go/Stop Evaluation

The second experiment assessed the vehicle's ability to recognize and re-

act to human presence. The vehicle decided to go or stop based on the presence of at least one person or the size of the detected vehicle being greater than 300px. This experiment evaluated the real-world applicability of the model in ensuring safe driving.

2.4 Experimental Results

The following images show the results of object detection using RetinaNet in the first experiment. Each image includes detected vehicles and pedestrians, with accuracy scores displayed.

3 Conclusion

3.1 Contribution

This study demonstrates the potential of real-time object recognition using RetinaNet to improve the safety and accuracy of autonomous vehicles. Specifically, RetinaNet's Focal Loss addresses class imbalance and enhances small object detection, playing a crucial role in object recognition for autonomous vehicles. This study serves as a foundational research for implementing real-time object recognition in autonomous systems.

3.2 Current Object Detection Models in Autonomous Driving

Current object detection models used in autonomous driving include:

- **YOLO**

  Due to its real-time detection capabilities, YOLO is widely used in autonomous driving for detecting objects quickly and efficiently.

- **SSD**

  SSD is also popular in autonomous systems for its balance between speed and accuracy, making it suitable for detecting multiple objects at different scales.

- **Faster R-CNN**

  Known for its high accuracy, Faster R-CNN is used in scenarios where precision is critical, despite its slower processing speed compared to one-stage detectors.

- **RetinaNet**

  Increasingly adopted for its ability to handle class imbalance and detect small objects effectively, making it a versatile choice for autonomous driving applications.

3.3 Limitation and Future Work

However, there are limitations such as misclassification in certain situations, necessitating additional data training and model improvement. Detection performance may decrease in complex environments with many people or on high-speed highways. Future research will focus on data collection from diverse road environments and improving model performance. Integrating other sensor data, such as LiDAR, to enhance accuracy and exploring lightweight models for faster real-time processing are also necessary.

4 Acknowledgement and References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).

2. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. IEEE transactions on pattern analysis and machine intelligence.