Simulating Autonomous Vehicles for XAI

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Arran Scaife

Spring, 2022

Technical Project Team Members
Victor Luo
Kayla Boggess

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Lu Feng, Department of Computer Science

Simulating Autonomous Vehicles for XAI

Arran Scaife
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA
ams2uxk@virginia.edu

Victor Luo
College of Arts and Science
University of Virginia
Charlottesville, VA
vzl7ka@virginia.edu

Abstract—In this paper, we detail our experimentation with the SafeBench reinforcement learning tool and the CARLA self-driving simulator. Our goal is to lay the groundwork for future research in developing a model that can be used to provide substantial evidence in court cases involving autonomous vehicles. To do this, we first implemented CARLA to accurately simulate and generate crash data. We then set up SafeBench to dynamically generate crash data based on a few base crashes, using reinforcement learning to train the CARLA agent. We observed the agent's performance across different scenarios and generated several figures to visualize the improvement in performance over time. We also discuss the limitations of our approach and potential areas for future work.

Overall, our research advances the understanding and use of explainable AI in autonomous vehicle accident cases and provides valuable insights for lawyers and other professionals in the field. The ability to uncover the key factors influencing AI decisions in traffic accidents can help lawyers more effectively defend clients with self-driving cars in tort litigation and provide a more thorough and persuasive explanation for traffic accident cases. Furthermore, our work serves as the foundation for future research in this area, enabling the development of more comprehensive and accurate models for explaining the decision-making process of autonomous vehicles.

Index Terms—SafeBench, CARLA, XAI, explainable AI, soft actor-critic

I. Introduction

AI is becoming increasingly prevalent in our everyday lives and is expected to become a major factor in the coming decades. Self-driving cars are among the most popular applications of AI, and as adoption of these cars increases, so does the number of legal cases involving them. In these cases, lawyers are often tasked with defending clients with self-driving cars and are required to bring on expert witnesses to testify for the car's decision process. However, due to AI being a black box, this results in a naive understanding of the AI's decision process and juries ruling against clients with self-driving cars. Additionally, as self-driving cars become more prevalent, lawyers will need more expert witnesses to testify for their clients, which will become an increasing problem as the demand for AI experts outgrows their supply.

In this paper, we detail out our experimentation with SafeBench through the CARLA self-driving simulator. We will show the gradual improvement in performance as a result of SafeBench's reinforcement learning, and detail how that lays the groundwork for future research. We will also elaborate on future research to develop a model that determines a list of factors for car crashes and can be used to provide substantial evidence in a court case, allowing lawyers to reference these primary causes and provide substantial evidence that can be used in court cases.

II. Related Work

The application of explainable AI in tort litigation has been explored in the past. For example, [2] proposed a model that uses the "black box" of a self-driving car to uncover the primary causes of a crash.

In addition to the work mentioned in the previous paragraph, there have been several other notable efforts to apply explainable AI in the context of self-driving car accidents. For instance, [3] evaluates a method for generating human-readable explanations of the decisions made by a self-driving car, using natural language generation techniques. This approach allows for greater transparency and accountability in the decision-making process of autonomous vehicles. Another example is [3], which presents a framework for explaining the behavior of self-driving cars. This framework allows for the identification of the key factors that led to a particular decision or action taken by the vehicle, providing a more comprehensive understanding of the underlying reasons behind the car's behavior.

Overall, the use of explainable AI in the context of self-driving car accidents has garnered significant attention in recent years, with researchers proposing

various methods and approaches for generating human-readable explanations of the decisions and actions taken by autonomous vehicles. These efforts aim to improve the transparency and accountability of self-driving cars, ultimately leading to greater trust and acceptance of this technology.

III. Methodology

To solve the lawyer's problem, we first set up CARLA version 9.11.0 to be able to accurately simulate crashes. An earlier version of this software is shown in Fig. 1. Then, we implement SafeBench into CARLA and verify that it can dynamically generate crash data based on a few base crashes. We will then use 10 of SafeBench's trials to train the CARLA agent using each of its given 8 scenarios, shown in Fig. 2 and observe the rate of improvement for these different scenarios. This will aid later papers by providing the most complicated scenarios for self-driving agents that could correlate to likelihood of crashes.

Fig. 1. Rudimentary visual of a CARLA simulation involving an autonomous vehicle and a pedestrian

Further, the SafeBench architecture is structured as seen in Fig. 3, which required us to clone the GitHub repository, initialize the docker container,

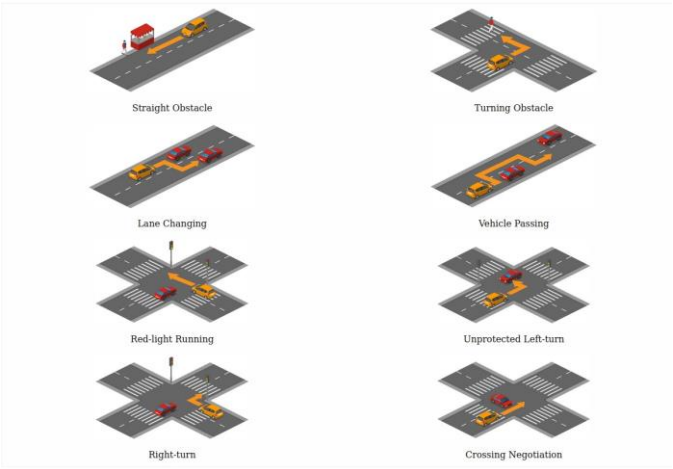

Fig. 2. 8 scenarios to be trained on

and repeatedly output the results using the overall score provided, which was transformed to Excel and plotted. This overall score was a normalized linear combination of 10 metrics: collision rate, frequency of running red lights/stop signs, average distance driven off road, route-following stability, average percentage of route completion, average time to complete route, average acceleration, average yaw velocity, and lane invasion frequency. Here, a higher score near 1 implies better performance on average between these metrics.

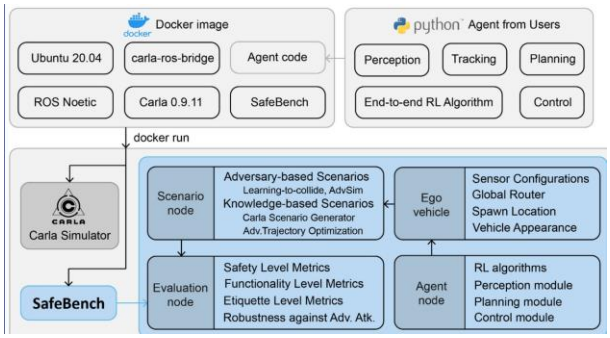

Fig. 3. Wireframe of the SafeBench architecture

After the results and conclusions of this paper, we plan on implementing a novel research paper's out-of-distribution detection algorithm on said generated crashes, which will clearly highlight the most influential factors determining the AI's decision and form a prioritized list of substantial factors lawyers can use to defend clients with self-driving cars. With this, we can run this out-of-distribution model on all the simulations provided by SafeBench to prove the

model's effectiveness in detecting out-of-distribution data points.

Expanding further, we will train CARLA across additional scenarios, using weather conditions or anomaly traffic situations as independent variables. By providing additional training, our model can better explain the reasons behind traffic accidents, such as where training may have been inadequate or where certain scenarios were overlooked. This can help provide more accurate and comprehensive explanations for traffic accident cases in tort litigation.

IV. Experiments

The purpose of this paper was to analyze the results of training a reinforcement learning agent using the CARLA autonomous driving simulator and SafeBench. In particular, we examined how the agent's performance improved over a series of training steps as it was evaluated in a variety of different scenarios. The policy used for the RL agent was the Soft Actor-Critic Policy (SAC), which is an RL algorithm that uses 2 Q-Tables and Entropy Regularization alongside stochastic experiments to learn an optimal policy via off-policy learning. The mechanism for this policy is shown in Fig. 5. To this end, we generated a graph showing the evaluation results across 10 training steps for 8 different scenarios.

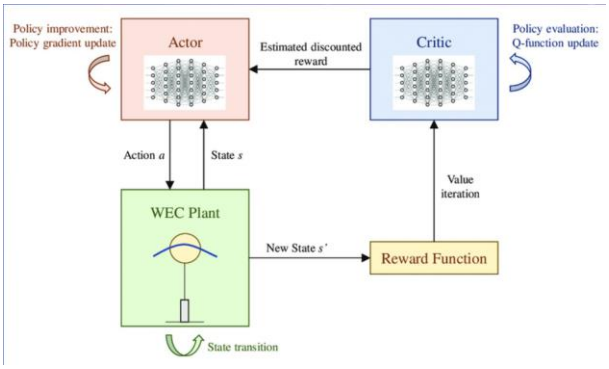

Fig. 4. General diagram for the SAC policy

We began by outlining the process for testing the agent in various scenarios. Each of these scenarios was composed of a set of conditions that the agent had to navigate in order to reach the goal.

For example, in the "following-distance" scenario the agent had to maintain a safe following distance from the lead vehicle. Similarly, in the "speed-limit" scenario the agent had to adhere to the speed limit. Once the agent had been trained and evaluated in each of the scenarios, the evaluation results were recorded and plotted against the number of training steps. We expect the graph to show a general improvement in the evaluation results as the number of training steps increase due to the fact that the agent should be able to learn from its experience and improve its performance as it is exposed to more scenarios per the SAC policy. This graph will allow us to compare the performance of the agent in different scenarios over the same number of training steps, enabling us to determine which scenarios require the least learning effort for the agent and which scenarios are the most difficult for the agent to master.

V. Results

The purpose of this paper was to analyze the results of training a reinforcement learning (RL) agent using the CARLA autonomous driving simulator and SafeBench. Looking at the 8 scenarios over each training iteration, we generated a graph of each scenario's score relative to its training iteration, as shown in Fig. 2. We show the exact results in Table. I.

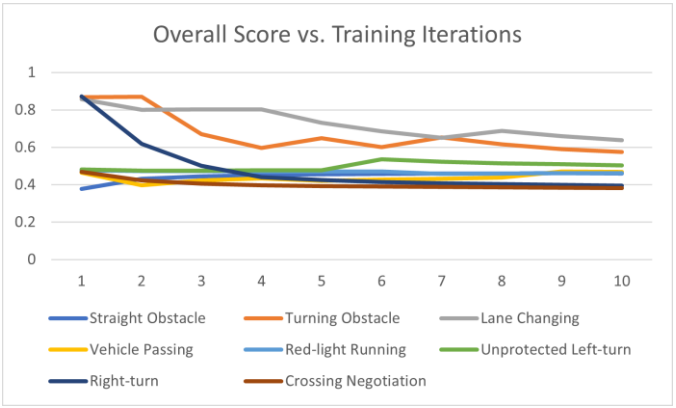

Fig. 5. Line graph demonstrating each scenario's overall score after each training iteration

Over all the training scenarios, although learning did happen, only Straight Obstacle, Vehicle Passing, and Unprotected Left-turn showed improvement in their overall scores while Turning Obstacle, Lane Changing, Red-light Running, Right-Turn, and Crossing Negotiation actually showing diminishing

TABLE I
OVERALL SCORE VS. TRAINING ITERATIONS FOR EACH SCENARIO

Training Step	Scenario		
	SO	*TO*	*LC*
1	0.376876624	0.868103329	0.856828999
2	0.431565177	0.870692585	0.799507216
3	0.445271435	0.670208128	0.802436532
4	0.451001413	0.596069807	0.802199332
5	0.455020163	0.648180217	0.731885346
6	0.458233252	0.601176536	0.684743822
7	0.460197287	0.653649608	0.651535023
8	0.460860893	0.616977412	0.688004034
9	0.462543392	0.589065219	0.660084564
10	0.459812311	0.574435773	0.638893596
Training Step	Scenario		
	VP	*RLR*	*ULT*
1	0.464569348	0.476318166	0.481521147
2	0.396119809	0.471867902	0.47445008
3	0.423071856	0.473161848	0.475923114
4	0.433044515	0.469725752	0.476951438
5	0.421499477	0.470582824	0.476389389
6	0.426616784	0.471415227	0.53533472
7	0.43191347	0.457288705	0.523362227
8	0.437359938	0.459814367	0.514210217
9	0.470187107	0.462266234	0.50875195
10	0.467678597	0.462385488	0.50229386

Training Step	Scenario	
	RT	*CN*
1	0.871080867	0.467912458
2	0.618150985	0.423924052
3	0.501134174	0.404738676
4	0.440438987	0.397253887
5	0.424542083	0.392140833
6	0.414551142	0.389964999
7	0.407432544	0.387567732
8	0.402527609	0.385936266
9	0.398767104	0.383699162
10	0.395740418	0.382588017

scores as training increased. Despite this, we assume that overall learning did happen as the SAC policy is designed to minimize the probability of failure over each successive stochastic training step. We notice that the successful scenarios that did improve generally start with a score below 0.40. This indicates to us that these were scenarios that were difficult to learn and usually crashed on the first run, minimizing the overall score at start. Additionally, since we see considerable improvements at each run, we also conclude that these scenarios are easily avoidable scenarios the model quickly learned from. This makes sense considering these scenarios are Straight Obstacle, Vehicle Passing, and Unprotected Left-turn, which only require minor adjustments to avoid the crash. On the other hand, the most notable offenders of our expected results is shown in Turning Obstacle, Lane Changing, and Right-turn. These scenarios start with an overall score above 0.80, indicating that the agent performed nearly optimally given the 10 evaluation factors. Despite this, the score quickly dropped, which lets us conclude that the initial score was a lucky pass and that the agent didn't truly learn the optimal strategy. This also makes sense in the context of being initially avoided, and then diminished score as the model experiments towards finding an objective solution, as minimal adjustments are not either necessary or enough to find an optimal path. For these 3 scenarios, the car is required to slow down as the only solution to avoid a crash. Additionally, pedestrians are especially difficult to see, so we are not surprised the car had such a difficulty with this scenario. Finally, Lane Changing, which has both cars in front of the ego vehicle, is likely difficult for the car to control as it wants to speed up to maximize its objective function. Thus, it takes more iterations to learn that speeding up will always lead to crashes. Finally, we notice that the best final score across all 8 scenarios was the Lane Changing scenario while the worst performing scenario was the Crossing Negotiation scenario. Lane Changing started with a high overall score and ended with the highest overall score, indicating that this scenario was not necessarily a lucky guess, but that the optimal policy involving its initial actions was well chosen. Since the 2 cars are relatively obvious to the ego vehicle, we weren't necessarily surprised by this. The ego vehicle has plenty of time to evaluate and react to the situation, unlike many of the other scenarios, allowing it more time and confidence to find the optimal solution. On the other hand, the Crossing Negotiation is more difficult to interpret for the car, as the agent needs to evaluate whether the opposing car will cross the intersection or not. This is known to be a difficult scenario for AI agents to evaluate, so we are not surprised that this

scenario not only had the lowest overall score, but also had a low overall starting score.

VI. Conclusion

For simple scenarios, the results of our analysis demonstrated that the agent was able to learn from its experience and improve its performance as it was exposed to more trials. We observed a general improvement in the evaluation results as the number of training steps increased, with some scenarios requiring the agent to learn more complex tasks and showing a slower, or diminishing, rate of improvement relative to the 10 scores the overall score evaluates. We were also able to compare the performance of the agent in different scenarios over the same number of training steps, enabling us to determine which scenarios require the most learning effort for the agent and which scenarios are the most difficult for the agent to master.

Thus, our project successfully set up the infrastructure necessary to run further experiments to reveal a model to provide additional explainability to self-driving vehicle AI decision-making in tort litigation. By training the agent using SafeBench's trials on the CARLA simulator, we revealed a gradual improvement in the agent's ability to remain safe on the road. As adoption of self-driving cars continues to grow within the coming decades, leveraging AI to determine factors can streamline the judicial process by bypassing the need to wait for an expert witness to be available. This will prove to be frontier work in the space of explainable AI used in tort litigation.

In the future, we plan to continue our research on the primary factors behind car crashes involving self-driving cars with the goal of improving the explainability of AI systems for use in tort litigation. We plan to implement the novel research papers and to accurately determine the primary factors behind car crashes. In particular, we will use the work of [1] to explore the use of out-of-detection models to detect anomalies in training that could lead to car accidents.

From this, we plan on developing a model that would provide a prioritized list of substantial factors that lawyers can use to defend clients with self driving cars, helping explain to judges and juries the AI's decision process more thoroughly without the need for expert AI witnesses. Our model would also provide a prioritized list of substantial factors that lawyers can use to defend clients with self-driving cars, helping explain to judges and juries the AI's decision process more thoroughly without the need for expert AI witnesses. We believe that this model will provide valuable evidence in a court case by allowing lawyers to point to primary and substantial factors behind a self-driving car crash. Additionally, this model will shed light on the reasons for why the AI agent was potentially hindered, allowing juries to better understand the inner workings of AI agents. This also loosens the burden on AI expert witnesses that will have to come in and stand trial in self-driving car accidents.

Furthermore, we plan to expand the scope of our research to include other factors that may impact the safety of self-driving cars, such as weather conditions and road infrastructure. This analysis would enable us to understand the relationship between the number of training steps and the evaluation results for various scenarios and to identify scenarios in which the agent required the most learning effort. We believe that by considering these additional factors, we can provide a more comprehensive analysis of the causes of car crashes and develop more effective strategies for mitigating these risks.

References

[1] Aaqib Parvez Mohammed and Matias Valdenegro-Toro, "Benchmark for Out-of-Distribution Detection in Deep Reinforcement Learning," in CoRR, vol. 2112, 2021.

[2] Atakishiyev, Shahin & Salameh, Mohammad & Yao, Hengshuai & Goebel, Randy, "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions," 2021.

[3] Eloi Zablocki, Hedi Ben-Younes, Patrick Perez, Matthieu Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," 2022.