Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

Article

Attributation Analysis of Reinforcement Learning Based Highway Driver

Nikodem Pankiewicz [1,2] and Paweł Kowalczyk [1,3]

[1] Aptiv Services Poland S.A., ul. Podgórki Tynieckie 2, 30-399 Cracow, Poland
[2] AGH Universitu of Science and Technology, nikodem.pankiewicz@agh.edu.pl
[3] Silesian University of Technology ; pawel.2.kowalczyk@aptiv.com
* Correspondence: nikodem.pankiewicz@agh.edu.pl;

Abstract: While machine learning models are powering more and more everyday devices, there is a growing need for explaining them. This especially applies to the use of Deep Reinforcement Learning in solutions that require security, such as vehicle motion planning. In this paper, we propose a method of understanding what the RL agent's decision is based on. The method relies on conducting statistical analysis on a massive set of state-decisions samples. It indicates which input features have an impact on the agent's decision and the relationships between decisions, the significance of the input features, and their values. The method allows us for determining whether the process of making a decision by the agent is coherent with human intuition and what contradicts it. We applied the proposed method to the RL motion planning agent which is supposed to drive a vehicle safely and efficiently on a highway. We find out that making such analysis allows for a better understanding agent's decisions, inspecting its behavior, debugging the ANN model, and verifying the correctness of input values, which increases its credibility.

Keywords: Autonomous Vehicles; Reinforcement Learning; Explainable Reinforcement Learning; XRL

1. Introduction

1.1. Motivation

Machine learning is increasingly applied in everyday devices and computer applications. Beyond making popular applications more attractive with AI, researchers are trying to use it to solve real-world complex problems. One such challenge is to plan the motion of the automated vehicle on the highway in a safe and effective manner. A promising approach to this problem is the application of Deep Reinforcement Learning (RL) [1] methods which use Artificial Neural Networks (ANN) to train the decision-making agents. However, the use of ANN-based methods introduces the black box factor, which makes agents' decisions unpredictable and therefore increases operational risk. Such a factor is ineligible in the application which safety must be verified and proved. Therefore the utilization of ANN-based methods to plan the vehicle motion on the road, without understanding the ANN decisions, may be risky for the system's end-user.

Knowing this threat, we propose the evaluation method of RL agents based on Interpretable Machine Learning (IML) techniques combined with statistical analysis. The presented solution is intended to decipher the black-box model by analyzing the neural activations in the distribution of possible inputs with respect to agent decisions. Our method allows investigating whether the agent's decisions are consistent with the assumptions and ANN decision process matches human intuition. Additionally, it enables debugging the model itself and detecting data or model corruption. The proposed method is created for inspecting RL-driven applications whose decisions are critical for safety and confirmation of proper functioning is required.

© 2022 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

1.2. Contribution

In this work, we present a novel method of evaluation of two DRL agents which are designated to plan the behavior to achieve a safe and effective highway driving experience. The first agent (Maneuver Agent) selects the appropriate discrete maneuvers (Follow Lane, Prepare For Lane Change (Left/Right), Lane Change (Left/Right), Abort) and the second one (ACC agent) controls the continuous value of acceleration. On the basis of these two trained agents, we propose an evaluation method based on Integrated Gradient [2] and further statistical analysis. The analysis consists of ANOVA, t-test, and examination of linear (Pearson [3]) and monotonic (Spearman Rho [4]) correlation. We describe our experiments and show the results of analyzes of agents operating in discrete and continuous action space. Additionally, we specify the applicability and relevance of such methods.

2. Related Work

2.1. RL in AV

Over the past few years, there has been an increasing interest in the use of RL in the motion planning of automated vehicles. In the literature, we can find multiple examples of applications of RL for typical driving scenarios such as lane-keeping, lane changing, ramp merging, overtaking, and more. For example, [5] proposed to train a driving policy with DQN algorithm [6], to decide whether it is worthwhile to change lanes to the left or right or to keep the lane. The training took place in a simulated three lanes highway environment. The agent's objective was to drive safely, smoothly and maintain efficiency. A similar solution was proposed in [7] where authors considered comparable environment and action space. Additionally, the work emphasized safety assurance, integrating the RL methodology with the Responsible Sensitive Safety framework [8], which guarantees at least not to cause a collision.

A more challenging environment was solved in [9], where authors focused on training agents to handle unsignalized intersections. To successfully navigate through the junction, the agent had to learn other drivers' intentions and predict their movement. It is supposed to drive to the destination as fast as possible and avoid a collision. The agent got a positive reward for achieving the target lane, a large negative reward for collision, and small punishment for each step of the simulation.

Another work [10] introduced a novel solution based on Reinforcement Learning combined together with a classical A* algorithm. The authors presented a model-based RL algorithm which depends on tree search where the heuristic is learned with DQN algorithm. Such an approach allows for increased control and understanding of the algorithm. A more detailed overview of the works on the application of RL in motion planning can be found in [11] and [12].

2.2. Explainable RL

As the application of machine learning becomes more popular, the demand for its interpretability has increased. Initially, a field of Interpretable Machine Learning (IML) has been developed, partially focused on the interpretation of Neural Networks activation. The interpretation relies on calculating how the output of the ANN was impacted by each element of the given part of the network. For example by the input features as in the case of Primary Attributation [13–16]. In the case of Layer Attributation [17–19] it regards the impact of each neural layer and each single neuron activation in the case of Neuron Attributation [14,17].

However, the eXplainability of RL (XRL) goes beyond understanding single neural activation. That is because of temporal dependency between consecutive states and the agent's actions which induce the next visited states. A sequence of transitions may be used to interpret the agent's action concerning the long-term goal. Additionally, it is also important that the objective of agent training is maximizing the sum of collected rewards, rather than mapping the inputs to the ground truth label as in the case of Supervised

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

Learning. These additional features allow explaining the behavior of RL agents in an introspective, causal, and contrasting way.

The recent advances in XRL were categorized in [20] into two major groups: *transparent algorithms* and *post-hoc explainability*. The group of transparent algorithms includes those whose models are built to support their interpretability. Such an approach is implemented in *hierarchical RL* [21,22] where the major task is decomposed for sub-tasks with a trained higher-level agent and lower-level agents. The hierarchical structure is designed to provide an understanding of the agent's decision-making processes. Another approach is *simultaneous learning* which learns both policy and explanation at the same time. An example is work [23] which proposed to learn multiple Q-functions one for each meaningful part of the reward to understand predictions about future rewards. The last type of transparent learning is *representation learning* which involves learning latent features to facilitate the extraction of meaningful information by the agent models. The representative work [24] proposes to reconstruct the observation with autoencoders, training model to predict the next state, or train inverse model to predict action from the previous state.

However, DRL algorithms are not natively transparent therefore *post-hoc explainability* is more common and debated in this paper. It relies on an analysis of states and neural activations of transitions executed with an already trained agent.

One of the post-hoc methods is *saliency maps* [19,25] which may be applied to Convolutional Neural Networks (CNN) with images as an input. This method generates a heatmap that highlights the most relevant information for CNN on the image. Another interesting work is [26] which proposed 3 step analysis of agent transitions in order to classify interesting agent interactions and present them in a visual form. However, from our perspective, understanding individual decisions is not enough to interpret the general behavior of an agent.

3. Preliminaries

3.1. Reinforcement Learning Agents

Our experiment intends to develop a method of interpreting RL agent decisions, adequate for discrete and continuous action space. For this purpose, we train two separate agents. The first one (Maneuver Agent) is responsible for planning appropriate maneuvers to be executed. Agent's action space is discrete and contains six items: Follow Lane, Prepare For Lane Change (Right, Left), Lane Change (Right, Left), and Abort Maneuver. The objective of the agent is to navigate in the most efficient way while preserving the gentleness desired on the roads. Expected behaviors are for example changing to the faster lane if the ego's velocity is lower than the speed limit, or returning to the right lane when it is possible and worthwhile.

The second agent (ACC Agent) is responsible for planning the continuous value of acceleration when Follow Lane maneuver is selected by the higher-level agent. Reward function incentives the agent to drive as fast as possible in terms of the speed limit, keep a safe distance to the vehicle ahead, increase comfort by minimizing jerks and avoid collisions.

The training uses a simulation [27] of a highway environment in which parameters such as the number of lanes, traffic flow intensity, characteristics of other drivers' behavior, and vehicle model dynamics are randomized providing diverse traffic scenarios.

Agents take the form of Feed Forward Neural Network that are trained with Proximal Policy Optimization (PPO) algorithm [28]. As an input, they consume information about the ego's vehicle, percepted vehicles around (position, speed, acceleration, dimensions), and information about the road geometry. Additionally, the Maneuver agent consumes a list of maneuvers that are available from the safety perspective according to rules defined in [8]. As an output, Maneuver Agent returns categorical distribution parameters which are the probabilities of selecting maneuvers. ACC Agent outputs the parameters of Normal distribution (mean, and log standard deviation). From that values, the actual agent's action is sampled with respect to corresponding distributions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

3.2. Integrated Gradients

Integrated Gradients (IG) [13] is an example of the Primary Attributation method which aims at explaining the relationship between a models' output with respect to the input features by calculating the importance of each feature for the model's prediction. For calculation, IG needs baseline input x' which is composed arbitrarily and should be neutral for the model. For example, if the model consumes images, the typical baseline is an image which contains all black or white pixels. IG, firstly, in small steps α generates a set of inputs by linear interpolation between the baseline and the processed input x. Then it computes gradients between interpolated inputs and model outputs (eq. 1) to approximate the integral with the Riemann Trapezoid rule.

$$IntegratedGradients_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\delta F(x' + \alpha \times (x - x'))}{\delta x_i} d\alpha \tag{1}$$

where i = feature; x = input; x' = baseline; α = interpolation constant

4. Description of experiment

4.1. Collecting Neural activations

We train the Maneuver and ACC Agents (sec. 3.1) with PPO algorithm [28]. The training lasts until the mean of episodes sum of rewards has reached the target value. Afterward, we select the best model checkpoints and run an evaluation of agents on test scenarios generating 5h driving experience for Maneuver Agent and 3.5h for ACC agent. The samples consist of state inputs and agent decisions - action value for ACC agent and probabilities of selecting particular action in case of Maneuver Agent. In the case of Maneuver Agent, the input vector consists of 372 float values and 162 accordingly for ACC Agent. Based on that data we calculate the attributation of each input value using the Integrated Gradients method 3.2. As a baseline input, we select a feature vector that represents 3 lanes highway with no other vehicles besides the ego in its default state (max legal velocity, 0 acceleration). For calculation, we use Captum library [29] (BSD licensed) which provides an implementation of a number of IML methods 2.2 for PyTorch models. The results of attributations calculation with associated input features and ANN's decisions are further inspected with statistical analysis.

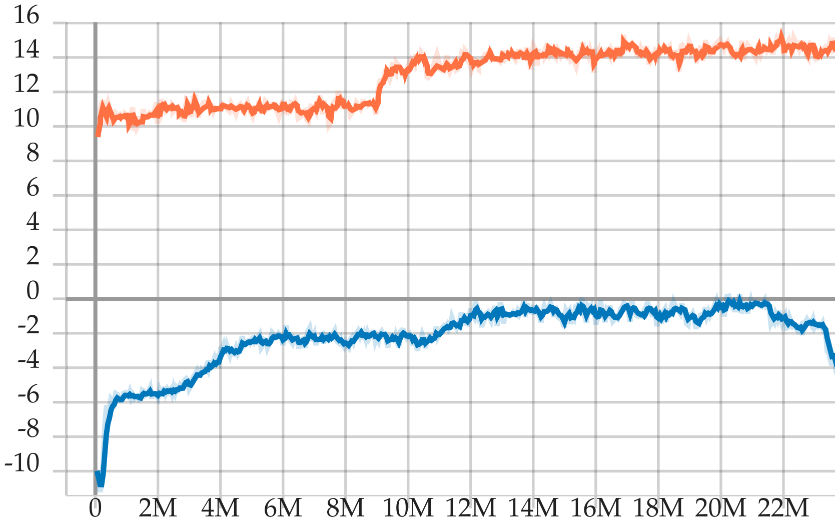

Figure 1. The graph shows the accumulated average sum of rewards (Maneuver Agent - Orange and ACC Agent - Blue) during the episodes between optimization steps. The training typically achieves its best performance and starts collapsing. Then to get the best agent we choose the checkpoint with the highest mean sum of rewards. The difference between the reward levels is due to the different definitions of the reward function.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

4.2. Statistical analysis

The statistical analysis consists of two parts. For all calculations, we use the Minitab software [30]. The first part focuses on the examination of the level of significance of the attributation values and analysis of their distribution. The second one studies the relationships between attributation values, values of input features, and probabilities of selecting maneuvers in the case of Maneuver Planning agent.

4.2.1. ANOVA and t-tests

The first step of statistical analysis of attributation is to identify parameters with statistically significant parameters of attributation distribution regarding the selected item from action space for Maneuver Agent and overall distribution for ACC Agent. The next step is to perform an analysis of variance for the set of parameters determined in the first step. To do so we divide attributation data according to the type of maneuver into six groups. Attributation that regards objects and roads are summed up according to each one of the characteristic parameters for those aspects. Then we perform t-test for every parameter with Null hypothesis $H_0 : \mu = 0.03$ and alternative hypothesis $H_1 : \mu > 0.03$. We assume the significance level of all tests as $\alpha = 0.05$. Based on those results we decide which distributions of parameters have a significantly higher mean value than 0.03 distinguishing between different maneuvers. Finally, we perform Welch's ANOVA [31] for results that are significantly based on the t-test which gives us information about which parameters were significantly more important than others regarding available maneuver. Samples were divided into groups with additional post-hoc test (Games Howell [32]). To visualize distinguished results, we calculate the standard deviation for those samples and 95%-confidence intervals for their means which gives us 95% assurance that the expected value is within those intervals regarding the dispersion of data.

4.2.2. Correlation tests

The second part of the analysis relies on the examination of the linear and monotonic relationship between feature attributation and the probability of selecting a given maneuver. We apply a Pearson correlation [3] to study linear correlation and Spearman's rank correlation coefficient Rho [4] to examine a monotonic correlation. Correlations are calculated for the attributation of all input features concerning the probability of selecting a particular maneuver.

An analysis based on a Pearson correlation begins with the calculation of the p-value and identification of whether the correlation is significant at 0.05 α-level. The p-value indicates whether the correlation coefficient is significantly different from 0. If the coefficient effectively equals 0 it indicates that there is no linear relationship in the population of compared samples. Afterward, we interpret the Pearson correlation coefficient itself to determine the strength and direction of the correlation. The correlation coefficient value ranges from -1 to $+1$. The larger the absolute value of the coefficient, the stronger the linear relationship between the samples. We take the convention that the absolute value of a correlation coefficient lower than 0.4 is a weak correlation, the absolute value of a correlation coefficient between 0.4 and 0.8 is a moderate linear correlation, and if the absolute value of the Pearson coefficient is higher than 0.8 the strength of the correlation is large. The sign of the coefficient indicates the direction of the dependency. If the coefficient is positive variables increase or decrease together and the line that represents the correlation slopes upward. A negative coefficient means that one variable tends to increase while the other decreases and the correlation line slopes downward.

The fact that an insignificant or low Pearson correlation coefficient does not mean that no relationship exists between the variables because the variables may have a nonlinear relationship. Considering that we utilize Spearman's rank correlation coefficient Rho [4] to examine the monotonic relationship between samples. In a monotonic relationship, the variables tend to move in the same relative direction, but not necessarily at a constant rate. To calculate the Spearman correlation, we have to rank the raw data and then calculate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

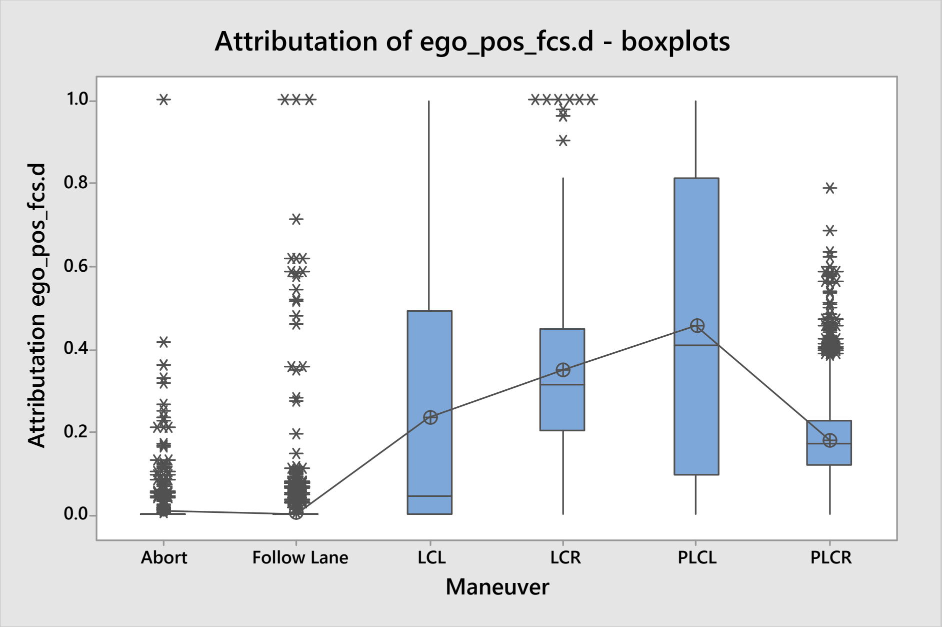

Figure 2. Distributions of attributation values for one of the ego parameters (distance from the center of lane) for all maneuver types.

its correlation. Test also consists of the significance test; the Spearman Rho correlation coefficient describes the direction and strength of the monotonic relationship. The value is interpreted analogously as the Pearson values. To visualize results and look for other types of relationships we created scatterplots for different pairs of samples.

5. Results

We inspect the results of statistical analysis in the following way. Firstly we examine the boxplots which visualize the distribution of attributation for a particular maneuver for each input signal. From the plots, we can easily see how much a given feature contributes to choosing a given maneuver. For example figure 2 presents the distribution of attributation of the feature ego_pos_fcs.d which indicates the ego's lateral distance from the driving lane center. We can see that the middle 50% of distributions (box), mean (dot), and median (horizontal line in the box) of attributation values lie much higher for maneuvers connected to lane change. For Follow Lane and Abort maneuvers attributation higher than 0 is considered an outlier (star). This behavior is in line with the driver's intuition and proves to us that the neural network works as intended, at least in this individual field. Next, we examine the correlation between attributations and values of input features. We check this in two directions. Firstly we look at the strong correlations and compare them with human intuition. For example, we notice that the agent, while considering selecting Follow Lane maneuver, pays less attention to the value of longitudinal velocity (vel_s) while the velocity grows. Although, it is more attentive to the parameter which informs about fulfilling velocity limit (vel_s_limit). This attitude is shown by Spearman Rho correlation, however, the Pearson does not reveal it. Additionally, we confirm that by inspecting the scatterplots of vel_s and vel_s_limit attributations presented in figure 3. We believe that such behavior is similar to human drivers, because they, while speeding up, stop thinking about absolute speed and start concerning if they drive with legal velocity, comparing their velocity with the speed limit.

As regards ACC agent, we identify a medium-strength correlation between attributation of acceleration and value of acceleration action. It means that the agent pays more attention to the value of acceleration when it increases. This is desired correlation but in our opinion, the values of attributation should be higher. In the figure 4 we notice that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

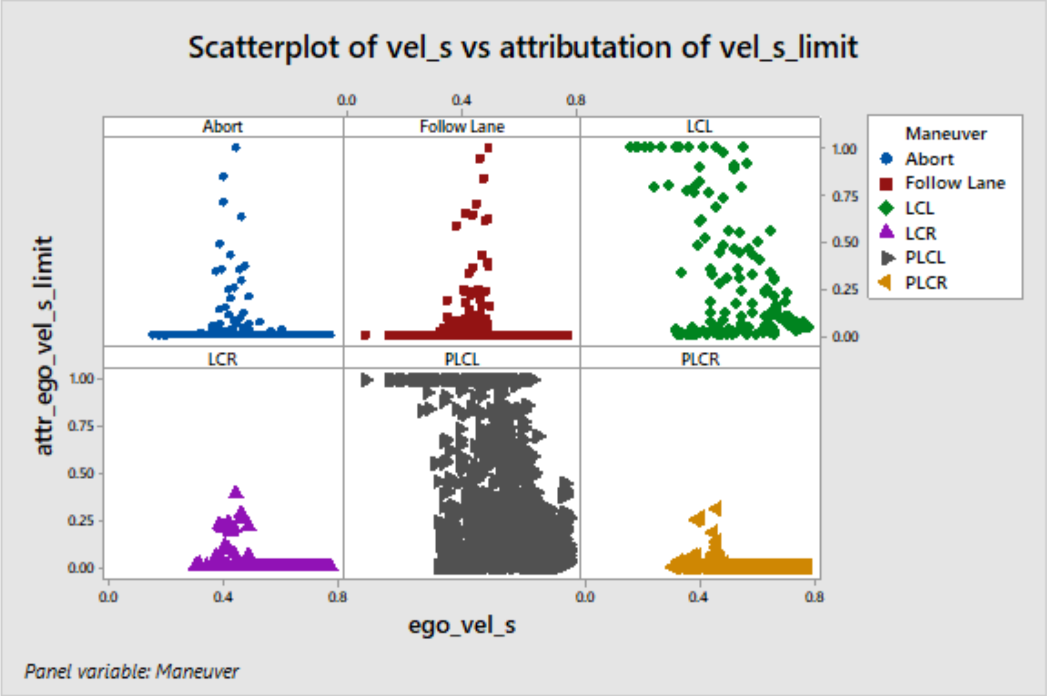

Figure 3. Scatterplot shows comparison between ego's longitudinal velocity (vel_s) and attributation values of vel_s_limit (velocity normalized to speed limit) for all maneuvers

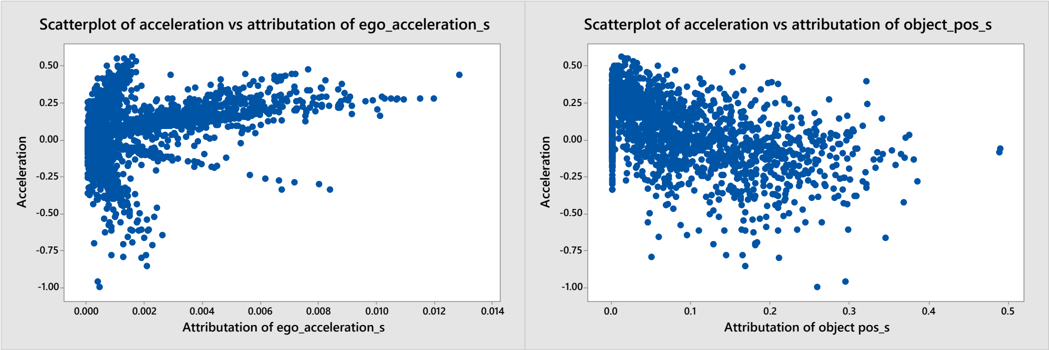

Figure 4. Scatterplots presents the correlation between acceleration of continuous agent and attributation of its acceleration value and position of other objects.

there exist at least 3 different patterns of correlations. They are connected to factors not analyzed in this experiment and it is probably beneficial to investigate further.

Also, we find out that the agent focuses more on other vehicles' positions when it is braking (see fig. 4).

Secondly, we deliberate where the strong correlation should occur to match human intelligence. For example, we assume that driver should compare the longitudinal distance to the target vehicle to its velocity. Therefore the correlation between attributation of objects' position with respect to the longitudinal velocity should be strong. The analysis indicates only weak strength of correlation thus contrary to assumptions.

Additionally, results inspection allows us to detect two types of errors in our model. During looking at the scatterplots (ex. 3, which demonstrates the value of attributation with respect to the input feature values, we easily detected that one's feature (lateral position) is normalized to the range (-2,0) instead of (-1,1). This allows us for fixing the implementation of the agent's observations.

The second finding regards the ANN architecture. The lack of attribution for every sample in one region of input features made us aware of the problem of vanishing gradients

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

Table 1. Table with values of Pearson correlation for attributation of vel_s_limitation with respect to values of ego's features. Red highlights strong correlation, and yellow - medium strength of correlation.

PEARSON						
ego	Follow Lane	PLCL	PLCR	LCL	LCR	Abort
pos_fcs.d	0.008	-0.017	-0.036	-0.862	-0.707	-0.001
vel_s	0.024	-0.221	0.08	0.04	0.255	0.162
vel_s_limi	-0.125	0.334	0.031	-0.071	0.096	0.113
vel_d	0.002	0.018	-0.003	-0.915	-0.677	-0.006
acc_s	-0.013	0.053	-0.035		-0.1	0.029
acc_d	-0.008	-0.02	0.024	0.309	-0.104	-0.025
rot_fcs	0.002	0.01	0	-0.893	-0.668	-0.003

Table 2. Table with values of Spearman Rho corellation for attributation of vel_s_limitation with respect to values of features which describe ego state. Red color means strong correlation, yellow - medium strength of correlation.

Spearman Rho						
ego	Follow Lane	PLCL	PLCR	LCL	LCR	Abort
pos_fcs.d	-0.019	0.012	-0.019	-0.681	-0.823	-0.116
vel_s	0.065	-0.217	0.038	0.114	0.076	0.182
vel_s_limi	-0.003	0.497	0.036	-0.107	0.188	0.135
vel_d	0.014	0.015	0.013	-0.706	-0.775	-0.081
acc_s	-0.022	-0.107	-0.014	-0.025	-0.037	0.171
acc_d	0.005	0.029	0.017	0.191	-0.341	-0.042
rot_fcs	0.012	0.013	0.016	-0.696	-0.753	-0.084

in our model. The wrong implementation of tensors concatenation does not pass the gradients through the model and deprives the agent to use part of the input knowledge.

6. Application

The presented method may contribute to a better understanding of the behavior of Reinforcement Learning agents whose consecutive decisions came from sampling from the distribution generated by ANN. First of all, it allows for identifying which input features influence the agent's decisions the most and inspecting the correlation between the importance of a given input feature to its value. It enables checking whether the ANN decision process matches human intuition (ex. the faster the agent drives the more it pays attention to the value of acceleration). Besides that such analysis enables detecting errors present in the model itself (ex. vanishing gradients - important information is ignored) or in input data (ex. the charts shows the wrong data distribution caused by the incorrect implementation). In our opinion application of the presented method increases the safety and predictability of the entire system. In the case of AV motion planning, it may lead to an increase in the reliability of RL applications, in the opinion of OEMs and consumers.

7. Discussion

Knowing what the agent is paying attention to can be calming for the end-user and build confidence in the model. This, in addition to the evaluation performed with Key Performance Indicators (KPI), may increase the model's reliability. However, knowing on what basis an agent makes decisions does not explain *why* agent makes it. It still does not solve the problem of RL explainability. We may only assume that examination of behavior in a significant number of situations expounds us the agent's character.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

One more thing to discuss is a situation when an agent's evaluation metrics (KPI) are high but analysis results contradict it. This may indicate that either KPI definitions are wrong or the model uncovers correlations in feature inputs that are not obvious to humans but still correct. Nevertheless, such a situation may decrease reliability in ANN-based models and discourage their application.

8. Conclusions

In this paper, we present the method for detailed inspection of the ANN model of the RL agent. The statistical methods applied to collected samples of agent decisions allow for recognition of agent's behavior patterns by looking globally at overall behavior and not at individual action. This is achieved by analysis of attributation distribution differentiated by considered maneuver and juxtaposed with values of other parameters describing the situation. By inspecting the analysis results we can seek confirmation that ANN concentrates on input features which are also important for a human driver. With the examination of the correlation between attributation and feature values, we find a pattern which match human intuition and that which is contrary to it. This knowledge helps us improve the model by changing model architecture, enhancing the training process, and ensuring that decisions are made in accordance with environment evaluation that prioritizes safety and effectiveness.

References

1. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; A Bradford Book: Cambridge, MA, USA, 2018.
2. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *34th International Conference on Machine Learning, ICML 2017* **2017**, *7*, 5109–5118.
3. Freedman, D.; Pisani, R.; Purves, R. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* **2007**.
4. Zar, J.H. Spearman rank correlation. *Encyclopedia of Biostatistics* **2005**, *7*.
5. Wang, P.; Chan, C.; de La Fortelle, A. A Reinforcement Learning Based Approach for Automated Lane Change Maneuvers. *CoRR* **2018**, *abs/1804.07871*, [1804.07871].
6. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M.A. Playing Atari with Deep Reinforcement Learning. *CoRR* **2013**, *abs/1312.5602*, [1312.5602].
7. Orłowski, M.; Wrona, T.; Pankiewicz, N.; Turlej, W. Safe and Goal-Based Highway Maneuver Planning with Reinforcement Learning. In Proceedings of the Advanced, Contemporary Control; Bartoszewicz, A.; Kabziński, J.; Kacprzyk, J., Eds.; Springer International Publishing: Cham, 2020; pp. 1261–1274.
8. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. On a Formal Model of Safe and Scalable Self-driving Cars. *CoRR* **2017**, *abs/1708.06374*, [1708.06374].
9. Isele, D.; Cosgun, A.; Subramanian, K.; Fujimura, K. Navigating Intersections with Autonomous Vehicles using Deep Reinforcement Learning. *CoRR* **2017**, *abs/1705.01196*, [1705.01196].
10. Keselman, A.; Ten, S.; Ghazali, A.; Jubeh, M. Reinforcement Learning with A* and a Deep Heuristic. *CoRR* **2018**, *abs/1811.07745*, [1811.07745].
11. Aradi, S. Survey of Deep Reinforcement Learning for Motion Planning of Autonomous Vehicles. *CoRR* **2020**, *abs/2001.11231*, [2001.11231].
12. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Sallab, A.A.A.; Yogamani, S.K.; Pérez, P. Deep Reinforcement Learning for Autonomous Driving: A Survey. *CoRR* **2020**, *abs/2002.00444*, [2002.00444].
13. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *34th International Conference on Machine Learning, ICML 2017* **2017**, *7*, 5109–5118. Integrated <b>Gradients</b>.
14. Lundberg, S.M.; Allen, P.G.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* **2017**, *30*. Gradient SHAP.
15. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. *34th International Conference on Machine Learning, ICML 2017* **2017**, *7*, 4844–4866. DeepLIFT.
16. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Saliency.
17. Dhamdhere, K.; Yan, Q.; Sundararajan, M. How Important Is a Neuron? *7th International Conference on Learning Representations, ICLR 2019* **2018**. Layer Conductance.
18. Leino, K.; Sen, S.; Datta, A.; Fredrikson, M.; Li, L. Influence-Directed Explanations for Deep Convolutional Networks. Internal Influence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0196.v1

19. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* **2016**, *128*, 336–359. GradCAM, https://doi.org/10.1007/s11263-019-01228-7.
20. Heuillet, A.; Couthouis, F.; Díaz-Rodríguez, N. Explainability in deep reinforcement learning. *Knowledge-Based Systems* **2021**, *214*, 106685. https://doi.org/https://doi.org/10.1016/j.knosys.2020.106685.
21. van Seijen, H.; Fatemi, M.; Romoff, J.; Laroche, R.; Barnes, T.; Tsang, J. Hybrid Reward Architecture for Reinforcement Learning. *CoRR* **2017**, *abs/1706.04208*, [1706.04208].
22. Kawano, H. Hierarchical sub-task decomposition for reinforcement learning of multi-robot delivery mission. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, 2013, pp. 828–835. https://doi.org/10.1109/ICRA.2013.6630669.
23. Juozapaitis, Z.; Koul, A.; Fern, A.; Erwig, M.; Doshi-Velez, F. Explainable Reinforcement Learning via Reward Decomposition. In Proceedings of the in proceedings at the International Joint Conference on Artificial Intelligence. A Workshop on Explainable Artificial Intelligence., 2019.
24. Raffin, A.; Hill, A.; Traoré, R.; Lesort, T.; Rodríguez, N.D.; Filliat, D. S-RL Toolbox: Environments, Datasets and Evaluation Metrics for State Representation Learning. *CoRR* **2018**, *abs/1809.09369*, [1809.09369].
25. Mundhenk, T.N.; Chen, B.Y.; Friedland, G. Efficient Saliency Maps for Explainable AI **2019**.
26. Sequeira, P.; Gervasio, M. Interestingness Elements for Explainable Reinforcement Learning: Understanding Agents' Capabilities and Limitations. *Artificial Intelligence* **2019**, *288*. https://doi.org/10.1016/j.artint.2020.103367.
27. Traffic AI - Simteract - simteract.com/traffic-ai/.
28. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *CoRR* **2017**, *abs/1707.06347*, [1707.06347].
29. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. Captum: A unified and generic model interpretability library for PyTorch, 2020, [arXiv:cs.LG/2009.07896].
30. Minitab, LLC - version 18 - https://www.minitab.com.
31. Liu, H. *Comparing Welch's ANOVA, a Kruskal-Wallis Test, and Traditional ANOVA in Case of Heterogeneity of Variance*; Virginia Commonwealth University, 2015.
32. Sauder, D.C.; DeMars, C.E. An Updated Recommendation for Multiple Comparisons. *Advances in Methods and Practices in Psychological Science* **2019**, *2*, 26–44, [https://doi.org/10.1177/2515245918808784]. https://doi.org/10.1177/2515245918808784.