

CuRLA: Curriculum Learning Based Deep Reinforcement Learning for Autonomous Driving

Bhargava Uppuluri, Anjel Patel, Neil Mehta, Sridhar Kamath and Pratyush Chakraborty

Birla Institute of Technology & Science, Pilani, Hyderabad Campus, Jawahar Nagar, Kapra Mandal,

Medchal District - 500078, Telangana, India

{f20191279h, f20190126h, f20190177h}@alumni.bits-pilani.ac.in,

Keywords: Computer Vision, Deep Reinforcement Learning, Variational Autoencoder, Proximal Policy Optimization, Curriculum Learning, Autonomous Driving.

Abstract: In autonomous driving, traditional Computer Vision (CV) agents often struggle in unfamiliar situations due to biases in the training data. Deep Reinforcement Learning (DRL) agents address this by learning from experience and maximizing rewards, which helps them adapt to dynamic environments. However, ensuring their generalization remains challenging, especially with static training environments. Additionally, DRL models lack transparency, making it difficult to guarantee safety in all scenarios, particularly those not seen during training. To tackle these issues, we propose a method that combines DRL with Curriculum Learning for autonomous driving. Our approach uses a Proximal Policy Optimization (PPO) agent and a Variational Autoencoder (VAE) to learn safe driving in the CARLA simulator. The agent is trained using two-fold curriculum learning, progressively increasing environment difficulty and incorporating a collision penalty in the reward function to promote safety. This method improves the agent's adaptability and reliability in complex environments, and understand the nuances of balancing multiple reward components from different feedback signals in a single scalar reward function.

1 INTRODUCTION

The quest for safer, more efficient, and accessible transportation has made autonomous vehicles (AVs) a key technological innovation. Rule-based systems showed some promise (Moravec, 1990) but lacked the adaptability and robustness to tackle real-world scenarios. In the late 1900s, neural network-based supervised machine learning algorithms were trained on labeled datasets to predict steering angles, braking, and other control actions based on sensor data (e.g., ALVINN (Pomerleau, 1988)). In the early 2000s SLAM techniques such as LiDAR and Radar improved understanding of vehicle position and surroundings, improving vehicle navigation accuracy (Thrun et al., 2005), even in dynamic lighting and weather conditions - (Li and Ibanez-Guzman, 2020). The DARPA Grand Challenge in 2005 - (Wikipedia contributors, 2024) which saw Stanley emerge as the winner, marked significant

milestones in autonomous driving. Stanley utilized machine learning techniques to navigate the unstructured environment, thus recognizing machine learning and artificial intelligence as essential components of autonomous driving technology. (Grigorescu et al., 2019)

Eventually, with the advancement in camera technology, vision-based navigation became an area of research, with algorithms processing visual data to identify lanes, obstacles, and traffic signals. Although this approach worked very well in controlled environments, it faced challenges in dynamic scenarios and varying light conditions (Dickmanns and Zapp, 1987). Combining the advancements in Deep Learning and Computer Vision, CNNs were used to extract feature maps from visual data from the camera mounted on the vehicle. This led to breakthroughs in object detection, lane understanding, and road sign recognition (e.g., ImageNet) (Krizhevsky et al., 2012). Despite progress, fully automating decision-making and vehicle control in dynamic environments remains challenging, requiring significant manual feature engineering in ML and DL methods.

Here is where reinforcement learning comes into

P. Chakraborty and S. Kamath are with the Birla Institute of Technology & Science (BITS) Pilani, Hyderabad Campus. B. Uppuluri, A. Patel, and N. Mehta were with BITS Pilani, Hyderabad Campus, and are now alumni.

play, showing great potential for decision-making and controlling tasks. Combining DL with RL, where agents learn through trial and error in simulated environments, has significantly improved AV decision-making - (Mnih et al., 2013). Unlike rule-based systems, DRL agents can learn to navigate through diverse scenarios by trial and error behavior. This kind of learning also allows the agents to master intricate maneuvers and handle unexpected situations. (Mnih et al., 2013) used DQN to showcase driving policy learning from raw pixel inputs in austere video game environments. However, a simple DQN (Mnih et al., 2013) will not work well in real-life applications, such as driving, as the action space is not discrete but continuous.

Deep Deterministic Policy Gradient (DDPG) is an on-policy actor-critic algorithm specifically designed to handle continuous action spaces by directly parameterizing the action-value function (Lillicrap et al., 2019). The authors of (Kendall et al., 2018) used this algorithm to drive a full-sized autonomous vehicle. The system was first trained in simulation before being introduced in real-time using onboard computers. It used an actor-critic model to output the steering angle and speed control with the help of an input image. They suggested having a less sparse reward function and using Variational Auto Encoders (VAEs) (Kingma and Welling, 2022) for better state representation. DDPG also requires many interactions with the environment, making the training slow, especially in a high-dimensional and complex environment. DDPG training can also be unstable, especially in sparse reward or non-stationary environments. This instability can manifest as oscillations or divergence during training (Mnih et al., 2016). Proximal Policy Algorithm (PPO) (Schulman et al., 2017b) was developed to address these issues.

In our work, we use PPO (Schulman et al., 2017b) and Curriculum Learning (Bengio et al., 2009) for the self-driving task. To obtain a better representation of the state, we have used Variational Auto Encoders (VAE) (Kingma and Welling, 2022) to encode the current scene from CARLA (Dosovitskiy et al., 2017), the urban driving simulator (Town 7). Our paper builds upon the foundational work about accelerated training of DRL-based autonomous driving agents presented in (Vergara, 2019). Salient features of our work:

- Introduction of Curriculum learning in the training process allows the agent to learn the easier tasks like moving forward initially and as difficulty increases to make it learn more difficult tasks, like maneuvering in traffic or avoiding high-speed collisions.

- We have introduced a refined reward function that gives a higher reward to the agent to travel at higher speeds. This is important to increase the average speed, reducing travel time.
- Unlike our base paper, our reward function takes into account the collision penalty as well as other rewards like angle, centering, and speed reward. This is crucial to make the reward function less sparse and aid a smoother driving experience.
- The combination of the curriculum learning approach, involving increasing traffic density and augmenting the reward function, as well as the modified reward function, helps in the faster training process and better average speed of the agent.

We name this method CuRLA - Curriculum Learning Based Reinforcement Learning for Autonomous Driving, as curriculum learning is integral to the features in our work. These features and the improvements they bring about will be discussed in the paper in further detail.

2 PRELIMINARIES

In this section, we provide the foundational tools, concepts, and definitions necessary to understand the subsequent content of this paper. We begin by introducing the environment used for training, followed by a brief introduction to Policy Gradient RL algorithms. Next, we discuss the concept of curriculum learning, and finally, we detail the encoder used in our approach.

2.1 CARLA Driving Simulator

The rise of autonomous driving systems in recent years owes much to the emergence of sophisticated simulation environments. CARLA (Dosovitskiy et al., 2017) is a critical resource for researchers and developers in autonomous driving, providing an open-source, high-fidelity simulator. Its capabilities include realistic vehicle dynamics, sensor emulation (such as LiDAR, radar, and cameras), and dynamic weather and lighting conditions. Moreover, CARLA's scalability enables us to simulate large-scale scenarios involving multiple vehicles and pedestrians interacting in complex urban environments. Its standardized metrics and scenarios facilitate fair comparisons between different self-driving approaches in the field. We particularly chose the CARLA simulator as it also provides a collision intensity whenever a vehicle collides with another object in the environment, and we use this in the reward function design.

2.2 Policy Gradient Methods

Policy gradient methods (Sutton et al., 1999) are pivotal in reinforcement learning for continuous action spaces, directly parameterizing policies, enabling the learning of complex behaviors. Unlike value-based approaches that estimate state/action values, these methods learn a probabilistic mapping from states to actions, enhancing adaptability in stochastic environments. These methods optimize policy parameters " θ " in continuous action spaces through on-policy gradient ascent on the performance objective $J(\pi_\theta)$.

Trust Region Policy Optimization (TRPO) (Schulman et al., 2017a) is a type of policy gradient method that stabilizes policy updates by imposing a KL divergence constraint (Kullback and Leibler, 1951), preventing large updates. However, TRPO's complex implementation and incompatibility with models sharing parameters or containing noise are drawbacks.

Proximal Policy Optimization (PPO) (Schulman et al., 2017b) is an on-policy algorithm suited for complex environments with continuous action and state spaces. It builds upon TRPO (Schulman et al., 2017a) by using a clipped objective function for gradient descent, simplifying implementation with first-order optimization while maintaining data efficiency and reliable performance. In later sections, we will see how this has been implemented in our work.

2.3 Curriculum Learning

Curriculum Learning (Bengio et al., 2009) is a strategy aimed at enhancing the efficiency of an agent's learning process by optimizing the sequence in which it gains experience. By strategically organizing the learning trajectory, either performance or training speed on a predefined set of ultimate tasks can be improved. By quickly acquiring knowledge in simpler tasks, the agent can leverage this understanding to reduce the need for extensive exploration to tackle more complex tasks. (Narvekar et al., 2020)

2.4 Variational Autoencoder

Autoencoders (Rumelhart et al., 1986) are essentially generative neural networks that comprise an encoder followed by a decoder, whose objective is to transform input to output with the least possible distortions (Baldi, 2011).

VAEs (Kingma and Welling, 2022) excel in reinforcement learning by producing varied and structured latent representations, enhancing exploration strategies, and adapting to novel states. Moreover,

their probabilistic framework enhances resilience to uncertainty, which is crucial for adept decision-making in dynamic settings.

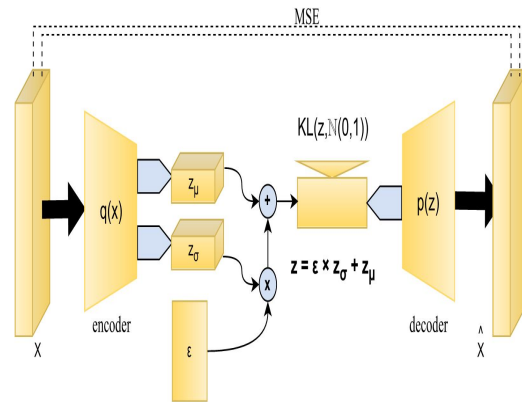


Figure 1: Variational autoencoder architecture.

Figure (1) shows a variational autoencoder architecture. The encoder operates on the input vector, yielding two vectors, z_μ and z_σ . Then, a sample z is drawn from the distribution $N(z_\mu; z_\sigma)$, which is fed into the decoder p , producing a reconstructed signal.



Figure 2: Top down view of the lap in Town 7.

3 EXPERIMENTAL STUDY AND RESULT ANALYSIS

In this section, we present the experimental setup, methodology, and results of our study. We start by describing the experimental environment, followed by a detailed explanation of the evaluation metrics and the baseline methods for comparison. We then present the results of our experiments and finally discuss the implications of our findings and compare our results with existing work in the field.

3.1 Proposed Model and Experimental Study

Our model is named as CuRLA and the model used in the base paper is named as Self-Centering Agent (SCA) for further reference. We also perform experiments using only curriculum learning for the optimized reward function to compare its performance to the two-fold curriculum learning method that is implemented in CuRLA. This model is named One-Fold CL. We use Town 7 from the list of Towns in CARLA Lap Environment (Dosovitskiy et al., 2017) as shown in figure (2). The reason for using this specific town is that it provides a highway environment without any exits, thus making our job of end-to-end driving easier. In our experiments, curriculum learning is employed by gradually increasing traffic volume and introducing functionalities to the agent in its reward function in stages, allowing it to quickly grasp the basics of the environment and efficiently tackle more complex tasks.

Table 1: VAE Parameters.

Hyperparameter	Value
z_{dim}	64
Architecture	CNN
Learning rate α	$1e-4$
β	1
Batch size N	100
Loss	BCE

The VAE (Kingma and Welling, 2022) we have used serves as feature extractors by compressing high-dimensional observations into a lower-dimensional latent space. This aids the learning process by providing a more manageable representation for the agent. Variational Autoencoder (VAE) is chosen as it is used for learning probabilistic representations of the input in the latent space, unlike regular autoencoders (Rumelhart et al., 1986), which learn deterministic representations. We use the VAE architecture from (Vergara, 2019) to encode the state of the environment, utilizing the pre-trained VAE from the same study to replicate the encodings. The dataset used to train the VAE contains 9000 training and 1000 validation examples (total of 10,000 160×80 RGB images), all collected manually from the driving simulator environment.

We have made use of the state-of-the-art PPO-Clip (Schulman et al., 2017b) algorithm to optimize our agent's policy and to arrive at an optimal policy that maximizes the return. PPO-Clip clips the policy to ensure that the new policy does not diverge far away from the old policy. The policy update equation

of PPO-Clip is given as:

$$\theta_{k+1} = \arg \max_{\theta} E_{s,a \sim \theta_k} [L(s, a, \theta_k, \theta)] \quad (1)$$

Here, L is given as:

$$L(s, a, \theta_k, \theta) = \min \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}, \text{clip} \left(\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)}, 1 + \epsilon, 1 - \epsilon \right) A^{\pi_{\theta_k}} \right) \quad (2)$$

Here, θ refers to the policy parameters being updated, and θ_k is the policy parameters currently being used in the k^{th} iteration to get the next iterations parameters, θ_{k+1} . Also, ϵ is the clipping hyperparameter defining how much the new policy can diverge from the old one. The min in the objective computes the minimum of the un-clipped and clipped objectives.

The overall graphical representation of the PPO+VAE training process we have used in our paper is shown in the figure (3). In the figure (3), the external variables are acceleration, steering angle, and speed from top to bottom.

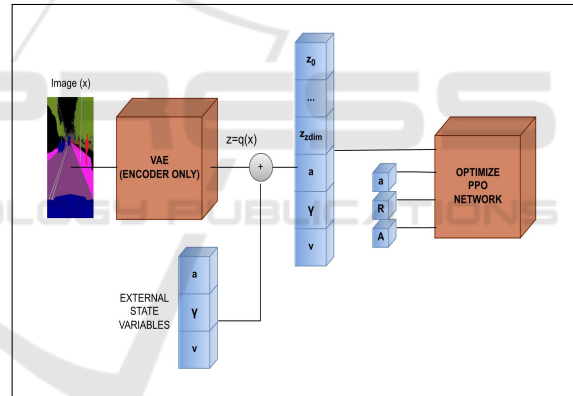


Figure 3: PPO+VAE training architecture.

The models are trained with the same parameters mentioned in (Vergara, 2019). Table 2 shows the complete list of hyperparameters used in the experiments. All models are trained for 3500 episodes, and the evaluation is done once every 10 episodes. Curriculum learning is implemented in both CuRLA and One-Fold CL. After 1500 episodes, traffic and a collision penalty are introduced in CuRLA. In One-Fold CL, only the collision penalty is introduced after 1500 episodes, and traffic is present from the first episode itself. An episode ends when the agent finishes three laps, drifts off the center of the lane by more than 3 metres, or has a speed of less than 1 km/hr for 5 seconds or more. A collision will not lead to the end of an episode (unless the agent collides and maintains a low speed under 1 km/hr for 5 seconds and more due

to the collision). The results we obtain are shown in Figs [6,8,7,9].

Table 2: Model Parameters.

Hyperparameters	Value
Batch size M	32
Entropy loss scale β	0.01
GAE parameter λ	0.95
Initial noise σ_{init}	0.4
Epochs K	3
Discount γ	0.99
Value loss scale α	1.0
Learning rate	$1e-4$
PPO-Clip parameter ϵ	0.2
Horizon T	128

3.2 Improvements Made

3.2.1 Reward Function Optimization

Keeping the architecture of the VAE and the RL algorithm the same as the original paper, we chose to change the original reward function. The original reward function (3) has 3 components: the angle reward, the centering reward, and the speed reward. Our revised reward function (4) has 4 components, the three original components, and we have also added a collision penalty, to encourage the agent against unsafe driving behaviour.

$$r = r_\alpha \cdot r_d \cdot r_v \quad (3)$$

$$r' = r_\alpha \cdot r_d \cdot r_{v'} + r_c \quad (4)$$

- 1. Angle Reward:** The angle reward component ensures that the agent is aligned with the road. Angle α , determines the angle difference between the forward vector of the vehicle (the direction the vehicle is heading in) and the forward vector of the vehicle waypoint (the direction the waypoint is aligned with). Using α , the angle reward r_α is defined by

$$r_\alpha = \max\left(1 - \left|\frac{\alpha}{\alpha_{max}}\right|, 0\right) \quad (5)$$

where $\alpha_{max} = 20^\circ$ ($\pi/9$ radians).

This ensures that the angle reward $r_\alpha \in [0, 1]$, and that the angle reward linearly decreases from 1 (perfect alignment) to 0 (when the deviation is 20° or more).

- 2. Centering Reward.** The centering reward factor ensures that the driving agent stays to the center of the lane. The usage of a high-fidelity simulator like CARLA enables us to have a precise measurement of distance between objects in the environment. To reward the agent for staying within

the center of the lane, we use the distance d between the center of the car and the center of the lane to define the centering reward by

$$r_d = \left(1 - \frac{d}{d_{max}}\right) \quad (6)$$

where $d_{max} = 3$.

This ensures that the centering reward component $r_d \in [0, 1]$ (as the episode terminates when the agent goes off center by more than 3 metres) and that r_d is inversely proportional to d , encouraging more centering for the agent while driving.

- 3. Speed Reward.** While keeping the angle and centering reward the same for all three agents, we change the speed reward component. The minimum speed v_{min} , maximum speed v_{max} , and target speed v_{target} are taken as 15 Km/hr, 105 Km/hr and 60 Km/hr respectively, and are kept as the same for both the original and updated speed reward components. The original speed reward component r_v and the new speed reward component $r_{v'}$ are then defined by

$$r_v = \begin{cases} \frac{v}{v_{min}}, & v < v_{min}, \\ 1, & v_{min} \leq v \leq v_{target}, \\ 1 - \frac{v - v_{target}}{v_{max} - v_{target}}, & v_{target} < v \leq v_{max} \end{cases} \quad (7)$$

$$r_{v'} = \begin{cases} 0.5 \cdot \frac{v}{v_{min}}, & v < v_{min}, \\ 1 - \frac{v_{target} - v}{v_{target} - v_{min}}, & v_{min} \leq v \leq v_{target}, \\ \frac{v_{max} - v}{v_{max} - v_{target}}, & v_{target} < v \leq v_{max} \end{cases} \quad (8)$$

where v is the current speed of the agent.

As seen in the graphs (Fig.4 & Fig.5), in the original speed reward function (Fig.4), the graph is constant in the range $[v_{min}, v_{target}]$ and receives the highest reward of 1. This is misleading as the agent can get confused, deciphering the minimum speed as the target speed itself, as it is getting a constant reward of 1 at either speed. This ensures that the agent will drive slower, as ensuring a higher centering and angle reward at a lower speed is much easier than at a higher speed. To rectify this, we have replaced the constant graph in $[v_{min}, v_{target}]$ with an increasing function (see Fig.5) to prioritize getting as close to the target speed as possible without losing too much performance on the angle and centering reward components. Both CuRLA and One-Fold CL use this revised reward function, whereas SCA uses the original reward function.

- 4. Collision Penalty.** A collision penalty factor was introduced for both One-Fold CL and CuRLA to

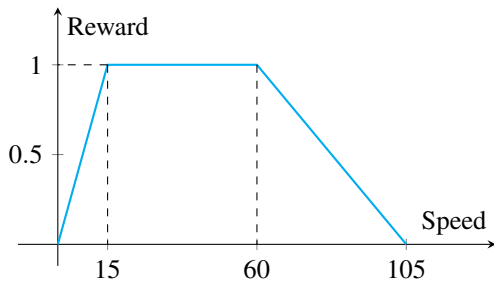


Figure 4: SCA Reward Function.

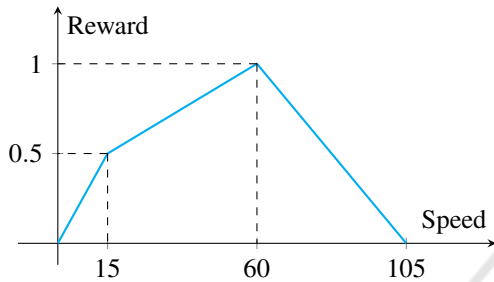


Figure 5: CuRLA & One-Fold CL Reward Function.

ensure the agent explicitly learns the behaviour of safe driving, avoiding collisions with other objects and vehicles in the environment. The advantage of using a simulator like CARLA also enables us to get a collision intensity (I_c) value between the agent and other objects in the environment, which we use to devise the collision penalty. The collision penalty r_c is defined by

$$r_c = \max(-1, -\log_{10}(\max(1, I_c))) \quad (9)$$

This penalty ensures $r_c \in [-1, 0]$, thus ensuring $r' \in [-1, 1]$.

3.2.2 Curriculum Learning

In CuRLA, curriculum learning is implemented in a two-fold manner. Firstly, we gradually increase the traffic volume of our simulation environment as the number of episodes increases. The agent gets to focus on learning to drive in a traffic-free environment initially and then slowly navigate through traffic in the later epochs. Secondly, the functionalities of the agent are gradually added. As the agent learns to drive in a traffic-free environment, we introduce a collision penalty while increasing the volume of traffic to teach it to avoid collisions with other vehicles. The reasoning behind adding this collision penalty is that while the current reward function accounts for smooth driving, it does not punish the agent enough for colliding with the other vehicles. All agents have been trained with traffic included in the roads, making a collision

penalty extremely important. This method of training helped the agent to learn the basics of the environment quickly and enabled it to learn harder tasks efficiently while also keeping safety as a factor.

3.3 Metrics and Result Analysis

The models are compared on two metrics - distance travelled and average speed. These metrics are chosen as they capture an accurate representation of the trade-off between efficiency and performance in autonomous driving scenarios. The distance traveled metric emphasizes the model's capability to maximize the length of the path it traverses, reflecting its ability to sustain long journeys without interruption. Conversely, the average speed metric accounts for both the distance covered and the time taken to complete the journey, offering a thorough assessment of the model's performance in terms of both speed and efficiency. Evaluating these metrics provides a detailed understanding of how each model balances speed and distance, which are essential factors in assessing the overall effectiveness of autonomous driving systems. The evaluation of our models against the base model provided insightful findings. (All graphs have been considered with a smoothing factor of 0.999)

In our assessment, for the metric of Distance Traveled, we measure it by the number of laps traveled by the vehicle, where 100% is considered as one lap finished. As it can be seen in Fig.6 & Fig.7, CuRLA and SCA have similar performance, completing approximately 1.5 laps on average in training and 0.3 laps in evaluation. Meanwhile, One-Fold CL did worse than both CuRLA and SCA, completing approximately 1.25 laps in training and 0.25 laps in evaluation. CuRLA does perform slightly better in training compared to SCA, which performs slightly better in evaluation. However, from the graph it can be seen that CuRLA had a consistent learning curve while training compared to SCA, which dropped in performance and then picked back up. One-Fold CL also performed comparably to both models, albeit slightly worse. This attests to the improvement in performance that two-fold curriculum learning shows over simple curriculum learning during training and training without curriculum learning.

The assessment of the metric of Average Speed highlighted the strengths of our assumption in changing the underlying speed reward component. During training, as can be seen in Fig.8, both CuRLA and One-Fold CL significantly outperform SCA reaching average speeds of 22 Km/hr and 20 Km/hr in training, compared to SCA's 14 Km/hr. The difference is

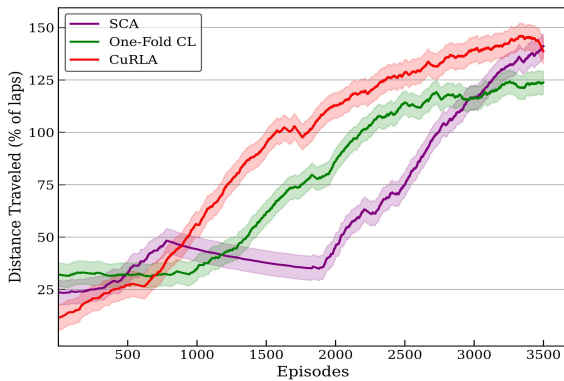


Figure 6: Training Metric: Distance Traveled.

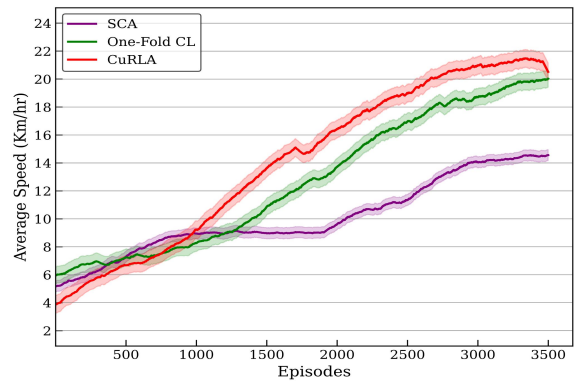


Figure 8: Training Metric: Average Speed.

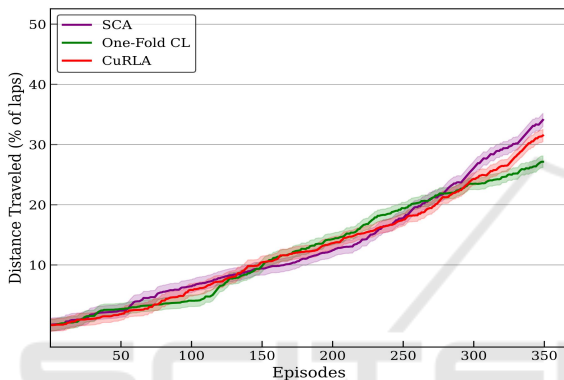


Figure 7: Evaluation Metric: Distance Traveled.

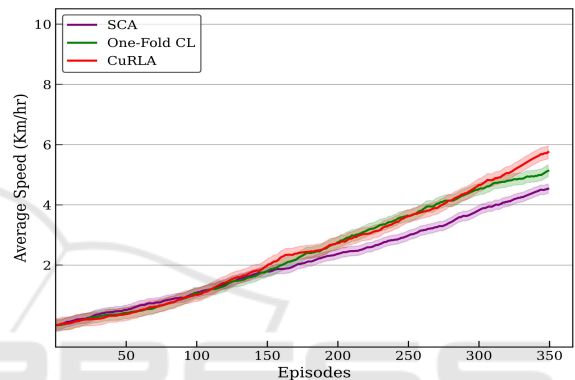


Figure 9: Evaluation Metric: Average Speed.

not as much during evaluation (Fig.9), but CuRLA and One-Fold CL still outperform SCA here as well, with CuRLA reaching an average speed of 6 Km/hr and One-Fold CL reaching an average speed of 5 Km/hr, compared to SCA's 4 Km/hr. This superior performance of CuRLA and One-Fold CL (on the revised reward function) compared to the SCA agent (using the original reward function) underscores our reward function's efficiency in optimizing speed-related aspects.

4 CONCLUSIONS

In this paper, we presented a model (CuRLA) that used a PPO+VAE architecture and two-fold curriculum learning along with a reward function tuned to accelerate the training process and achieve higher average speeds in autonomous driving scenarios. We show the performance of two-fold curriculum learning against simple curriculum learning (One-Fold CL agent), as well as the performance of agents on the revised reward function compared to the base reward function. While CuRLA and One-Fold CL perform comparably to the base agent (SCA) in the distance

traveled metric (with CuRLA performing slightly better and One-Fold CL being slightly worse), a significant improvement in average speed is observed. This prioritization of speed was a deliberate design choice. The distance traveled metric solely optimizes for maximizing the traversed path length. Conversely, the average speed metric inherently optimizes for both distance and the time taken to complete the journey, effectively accounting for two performance factors within a single measure. The performance of CuRLA and One-Fold CL agents, when compared to SCA, also attest to the benefits of using curriculum learning while training, and how decomposing the tasks in the autonomous driving problem helps agents to learn better and faster. Integrating multiple objectives into a single scalar reward function often leads to suboptimal agent performance. However, by employing curriculum learning during training, we can enable agents to master the nuances of each reward component more effectively. This approach facilitates better understanding of the environment and objectives, and ultimately enhances overall performance.

Future research will focus on enhancing performance by updating the architecture and algorithms. One area of interest is investigating vision-based

transformers and advanced transformer-based reinforcement learning methods for autonomous driving control. This entails replacing the current Variational Autoencoder with architectures like Vision Transformers (ViT, Swin Transformer, ConvNeXT) tailored for raw visual data. Furthermore, newer techniques such as Decision Transformers or Trajectory Transformers could replace the Proximal Policy Optimization (PPO) algorithm to potentially enhance decision-making capabilities. Another promising area for future research is Multi-Objective Reinforcement Learning (MORL) (Van Moffaert and Nowé, 2014; Hayes et al., 2021; Liu et al., 2015), where an agent optimizes multiple reward functions, each representing different objectives. Evaluating these advancements through simulated testing may lead to substantial performance improvements.

REFERENCES

- Baldi, P. (2011). Autoencoders, unsupervised learning, and deep architectures. In *ICML Unsupervised and Transfer Learning*.
- Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. volume 60, page 6.
- Dickmanns, E. D. and Zapp, B. (1987). An integrated dynamic scene analysis system for autonomous road vehicles. In *Intelligent Vehicles '87*, pages 157–164. IEEE.
- Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). Carla: An open urban driving simulator.
- Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. (2019). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386.
- Hayes, C. F., Radulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., de Oliveira Ramos, G., Restelli, M., Vamplew, P., and Roijers, D. M. (2021). A practical guide to multi-objective reinforcement learning and planning. *CoRR*, abs/2103.09568.
- Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J.-M., Lam, V.-D., Bewley, A., and Shah, A. (2018). Learning to drive in a day.
- Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Li, Y. and Ibanez-Guzman, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4):50–61.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2019). Continuous control with deep reinforcement learning.
- Liu, C., Xu, X., and Hu, D. (2015). Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(3):385–398.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning.
- Moravec, H. (1990). Sensor fusion in autonomous vehicles. In *Sensor Fusion*, pages 125–153. Springer, Boston, MA.
- Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., and Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey.
- Pomerleau, D. A. (1988). Alvin: An autonomous land vehicle in a neural network. In Touretzky, D., editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation.
- Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2017a). Trust region policy optimization.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017b). Proximal policy optimization algorithms.
- Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.
- Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*.
- Van Moffaert, K. and Nowé, A. (2014). Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512.
- Vergara, M. L. (2019). Accelerating training of deep reinforcement learning-based autonomous driving agents through comparative study of agent and environment designs. Master thesis, NTNU.
- Wikipedia contributors (2024). Darpa grand challenge (2005) — Wikipedia, the free encyclopedia.