Deep Reinforcement Learning in Decision-Making of Autonomous Driving: A Survey

Jizhou Cai
Huazhong University of Science and Technology, Wuhan 430074, China.

Abstract: Deep reinforcement learning (DRL) is a burgeoning sub-field in the realm of artificial intelligence that combines the benefits of deep learning (DL) and reinforcement learning (RL). By integrating these two methods, deep reinforcement learning has effectively addressed previously complex problems related to autonomous driving system (ADs) and has played a vital role in their development. Specifically, deep learning enhances reinforcement learning's ability to handle extensive high-dimensional data, which is critical for ADs. In this review, we mainly concentrate on the application of DRL algorithms in ADs, focusing primarily on decision-making processes. The review will be- gin by introducing deep learning and reinforcement learning independently before delving into the current applications, future prospects, and challenges facing deep reinforcement learning in this field. Finally, we will conclude with a summary of this review.

Keywords: Autonomous Driving; Deep Learning; Reinforcement Learning

1. Introduction

Over the past few decades, the swift advancement of industrial and scientific techniques in manufacturing has resulted in the ubiquitous incorporation of automobiles within society. As asserted by Carlier & Mathilde (2023), despite the repercussions of COVID-19, it is projected that the quantity of automobile purchases will attain an estimated 67.2 million by the year 2022.[1]

Unfortunately, the surge in the number of cars has also led to a high number of injuries or deaths caused by traffic accidents every year.

According to a research conducted by Caroselli (2022), it was found that a significant majority of approximately 95% of traffic accidents can be attributed to human error.[2] Consequently, there is an increasing focus on autonomous vehicles as a means to mitigate the incidence of accidents and fatalities. This phenomenon arises from a recognition that such technology holds promise in enhancing safety performance on roads.

Autonomous driving system is a comprehensive high-tech complex, which uses the technologies of perception, information fusion, decision-making, implementation, etc. How complex an environment the decision-making part can handle is one of the core indicators to measure and evaluate the ability of autonomous driving system. So, this research will focus on the decision-making part of automatic driving.

2. Background

2.1 Reinforcement learning

Usually, machine learning (ML) algorithms are categorized into several diverse fields: unsupervised learning, supervised learning, and reinforcement learning. RL is a subdiscipline of ML that focuses on the development of algorithmic models capable of acquiring knowledge and making decisions through interplaying with complex environments. The successful implementation of this discipline across various industries, such as robotics, gaming, finance, and autonomous driving, has attracted a great deal of attention from both the society and academic circles recently.

2.2 Deep learning

DL utilizes artificial neural networks comprising multiple interconnected layers. These layers are comprised of nodes that play a critical role in processing and analyzing data for decision-making purposes. Each node performs a simple operation and passes the output to the next layer. The input data is fed into the first layer, and then it passes through multiple layers until it reaches the output layer. The network can learn from the data by adjusting the weights of the connections between the nodes.

2.3 Autonomous driving

Autonomous driving, also known as self-driving vehicles, is a rapidly growing technology that promises to revolutionize the transportation industry. Up to the present time, a substantial body of scholarship has been conducted within this domain, encompassing a plethora of investigations pertaining to the multifaceted dimensions of self-driving vehicles, comprising their advantages and shortcomings. For more literature on the field of autonomous driving, please refer to Chen & Chen (2015).[3]

2.4 Deep reinforcement learning

DRL is a fusion of two powerful fields of artificial intelligence: DL and RL. DL is used to process multi-dimensional incoming data, such as videos or natural language, while reinforcement learning is used to optimize behaviors through trial and error.

In recent years, several DRL algorithms have been developed that utilize deep neural networks, such as Deep Q-Networks (DQN), which was mentioned earlier. Besides, Schulman proposed the Proximal Policy Optimization algorithm, which uses a policy network to directly output actions and updates its parameters using a trust region optimization method. Other important DRL algorithms include policy gradient methods, actor-critic methods, and their variants. For instance, the Trust Region Policy Optimization algorithm stabilizes policy gradient methods by constraining the policy update step size. Moreover, the Asynchronous Advantage Actor-Critic algorithm, which accelerated the learning process by utilizing multiple actors and learners was presented by Mnih.

3. Previous studies

3.1 Deep q-network (DQN)

The essential difference between DQN and Q-Learning is that DQN uses a neural network to fit a particular function. During training, the DQN samples mini-batches of go through tuples from the replay buffer in order to update the Q-network. The Q-network is trained to reduce the difference between the predicted and target values, which are computed using a target network and the Bellman equation. The target network refers to an independent duplication of the Q-network, which is utilized for the computation of target Q-values, and is updated periodically to match the current Q-network. By using a replay buffer and a separate target network, DQN can improve stability and convergence compared to standard Q-learning, which can be prone to instability due to the correlations between consecutive experience tuples.

3.2 Soft actor-critic (SAC)

SAC combines the benefits of maximum entropy RL and deep Q-networks. The purpose of the SAC algorithm is to obtain an policy for an agent in an environment through trial-and-error experience and interactions. The neural network used by the algorithm represents the agent's policy and value function. The agent's policy represents the actions taken by the agent in response to specific states of the environment. The value function estimates the expected benefits the agent will gain by following a particular policy in a state. Finally, the value function assesses the expected return from taking a specific action in a state under a specific policy.

SAC's key innovation is to employ the maximum entropy framework, which means that the agent's policy has some stochasticity. This approach encourages exploration, as the agent will continue to try new things even when it discovers policies that work well. Additionally, the agent's policy is optimized not only to increase the expected reward but also to increase its entropy. By doing so, the agent can explore with greater certainty about the potential outcomes of its decisions. Other similar researches about SAC in decision-making of autonomous driving can be found in Cheng & Song (2020).[4]

3.3 Proximal policy optimization (PPO)

PPO is used to optimize policies in environments with MDPs, and belongs to the class of actor-critic algorithms. The main idea behind PPO is to prevent the policy from changing too quickly during training, as this can lead to unstable behavior. To achieve this, a proximal objective function is employed to restrict the degree of variation between the new and old policies. The objective function includes a clipping parameter that limits how much the policy update can change the probability distribution over actions.

PPO also uses two different optimization objectives: a clipped surrogate objective and a value function objective. These objectives are combined to provide a single objective function that can be optimized using techniques such as stochastic gradient descent. Another key feature of PPO is the use of a mini-batch approach, where multiple trajectories are collected and used to update the policy and value function parameters. This helps to reduce variance in the update process and improves the stability of the algorithm.

3.4 Deep deterministic policy gradient (DDPG)

DDPG has gained significant attention due to its capability of solving challenging continuous control problems. DDPG learn both the Q-function and the policy function simultaneously, hence enhancing the training efficiency. DDPG employs a two-fold neural network architecture consisting of an actor and a critic network which work hand-in-hand during the learning process. Although the actor network processes the present state as an input and yields the subsequent action to execute, the critic network employs both the current state and action inputs to approximate the value of the corresponding state-action pair.

Throughout the training process, DDPG employs a variant of gradient descent algorithm to modify the weights of both actor and critic networks. Specifically, the temporal difference error (TD-error) is utilized to update the critic network, while updating of the actor network is attained via the gradient of the expected rewards concerning the actor parameters.

4. Challenges and future research directions

Despite some notable achievements in the development of autonomous driving decision-making processes, further testing is required to evaluate the robustness of DRL algorithms in real-world situations. The primary objective of introducing DRL into transportation systems is to mitigate the frequency of traffic accidents, however, given the stochastic factors such as pedestrians, bicycles, and other vehicles, guaranteeing safety remains a challenging task. In rare conditions where prior knowledge may be absent, it is crucial that ADs possess the ability to make informed decisions. One additional challenge confronting DRL algorithms pertains to reducing the latency experienced when processing high dimensional inputs without compromising accuracy. It is critical that AD systems make optimal decisions promptly when faced with emergent accidents.

References

[1] Carlier, & Mathilde. (2023). Topic: Automotive industry worldwide. Retrieved from https://www.statista.com/topics/1487/automotive-industry/#topicOverview.

[2] Caroselli. (2022, Jun). What percentage of car accidents are caused by human error?: Pittsburgh law blog. Caroselli, Beachler amp; Coleman, LLC. Retrieved from https://www.cbmclaw. com/what- percentage-of-car-accidents-are-caused-by-human-error/.

[3] Chen, X., & Chen, S. (2015). Autonomous vehicles: A review. International Journal of Automotive Engineering, 6 (1).

[4] Cheng, Y., & Song, Y. (2020). Autonomous decision-making generation of uav based on soft actor-critic algorithm. In 2020 39th Chinese control conference (ccc) (p. 7350-7355).