Autonomous Highway Driving using Deep Reinforcement Learning

Subramanya Nageshrao[1] and Eric Tseng[2] and Dimitar Filev[2]

Abstract— The operational space of an autonomous vehicle (AV) can be diverse and vary significantly. This may lead to a scenario that was not postulated in the design phase. Due to this, formulating a rule based decision maker for selecting maneuvers may not be ideal. Similarly, it may not be effective to design an a-priori cost function and then solve the optimal control problem in real-time. In order to address these issues and to avoid peculiar behaviors when encountering unforeseen scenario, we propose a reinforcement learning (RL) based method, where the ego car, i.e., an autonomous vehicle, learns to make decisions by directly interacting with simulated traffic. The decision maker for AV is implemented as a deep neural network providing an action choice for a given system state.

In a critical application such as driving, an RL agent without explicit notion of safety may not converge or it may need extremely large number of samples before finding a reliable policy. To best address the issue, this paper incorporates reinforcement learning with an additional short horizon safety check (SC). In a critical scenario, the safety check will also provide an alternate safe action to the agent provided if it exists. This leads to two novel contributions. First, it generalizes the states that could lead to undesirable "near-misses" or "collisions ". Second, inclusion of safety check can provide a safe and stable training environment. This significantly enhances learning efficiency without inhibiting meaningful exploration to ensure safe and optimal learned behavior.

We demonstrate the performance of the developed algorithm in highway driving scenario where the trained AV encounters varying traffic density in a highway setting.

I. INTRODUCTION

Reinforcement learning (RL), is a branch of artificial intelligence where an agent learns an optimal control strategy by interacting with the environment. Over the last few years there are many monumental success stories in RL. The most spectacular and well-known results involve state-of-the-art performance in many challenging tasks such as video games [1], decision making in an extremely complex environments such as game of Go [2], and continuous control of physical systems [3], [4].

In recent years autonomous driving has been a subject of great interest among both researchers and general public. Autonomous vehicles (AVs) have enormous potential not only economically but also for enhancing mobility options, and potentially reducing carbon footprint [5]. As in any robotic system, autonomous driving also involves decision making. A typical implementation of AV decision making is done by solving a number of path planning problems. They are formulated as a set of optimization problems, subject to various system and environmental constraints. The

solution corresponding to the lowest cost is often chosen as reference to the low-level controller [6]. This approach is computationally intensive, additionally due to the real time constraints on the solver, the obtained reference path may not result in a comfortable vehicle motion. Compare this to human driving, where a driver perceives the surrounding traffic and makes a maneuver decision, such as change lane, maintain speed, or brake etc. Based on this decision, the driver then proceeds with steering and throttle/brake pedal actuation to perform the maneuver with a smooth path. This approach fits well within the realm of the RL framework.

In this work we develop a Deep Reinforcement Learning (DRL) agent for highway driving. The DRL agent is designed by training an ego vehicle (EV) to learn a driving policy by directly interacting with diverse simulated traffic. The foundation of our DRL agent is a modified version of the double deep Q-network (DDQN) algorithm that was first discussed in [7], [8]. DDQN is considered as one of the state-of-the art RL algorithm for discrete state and continuous action space optimal problems. As detailed in [7], DDQN not only rectifies overestimation problem of DQN [1] but also improves performance across many benchmark problems. In this work we customize DDQN from [7] to include the following distinctive features.

- First, we augment the DDQN learning agent with a short-horizon safety check. In a critical situation, if the original DDQN action choice was deemed unsafe then the safety check will replace it with an alternate safe action, provided there is a safe alternative. For training, Q-learning often relies on an exploration policy, such as ϵ-greedy, as we have experienced and as one may anticipate, arbitrary exploration without constraints could frequently lead to scenarios involving collisions, and in turn, simulation resets. The safety check augmentation significantly improves the learning efficiency in avoiding frequent resets during training. Additionally during inference phase the safety check can be used as an override authority whenever a non-safe maneuver was chosen by the algorithm. This rare situation may occur for example, due to the use of function approximation such as neural networks. The safety rules is based on kinematic constraints so as to minimize the collision risks, they are elaborated in section III and are similar to the ones proposed in [9].
- Second, we keep the situations that invoke the safety violation in a collision buffer. As such, the safety check facilitates a constrained exploration during learning which enhances learning efficiency and achieves stable

[1] Ford Greenfield Labs, 3251 Hillview Ave, Palo Alto, CA 94304, USA `snageshr@ford.com`.
[2] Ford Research and Innovation Center, 2101 Village Road, Dearborn, MI 48124, USA, `htseng@ford.com`, `dfilev@ford.com`.

learning [10], [11]. Use of two buffers is a simpler version of prioritized experience replay as discussed in [12].

- Third, the inclusion of *safety controller* during inference phase leads to an interesting scenario, where by using new information the learned agent can be re-trained thus resulting in continuous adaptation. The new information can be potentially obtained from real-world data by deploying the initially trained agent in multiple vehicles.

These changes were found to be essential for both stability and efficiency of the DRL agent and will be discussed in detail in Section. IV of the paper.

We demonstrate the usability of the developed algorithm for highway driving of an AV with varying traffic density in a simulation environment. In summary, the key contributions of this work are:

1) We develop a DRL agent and a corresponding control architecture for highway autonomous driving using deep Q-network as a decision maker in conjunction with standard state feedback controllers.
2) We integrate the semantic information of the rules of the road in the form of a *short-horizon safety check* complementing with the statistical inference mechanism used in the DDQN algorithm.
3) We demonstrated the capability of the developed DRL agent in an adaptive framework in presence of new data.

The rest of paper is organized as follows. In Section II, a preliminary introduction on RL for decision making and control is presented. Following that, in Section. III, the proposed DDQN based algorithm along with the DRL based control architecture is explained. The implementation of the DRL agent for highway driving of an autonomous vehicle in an varying traffic density is given in Section. IV. Finally, Section. V concludes the paper with a note on our future research.

II. PRELIMINARIES: DEEP LEARNING FOR CONTROL

In this section we provide a brief theoretical background on decision making and control using deep reinforcement learning. Additionally, we elaborate on the main motivation for the control architecture presented in this article.

A. Reinforcement learning

In reinforcement learning (RL) an agent learns an optimal behavior with respect to a cost function by directly interacting with the system. Generally, RL problem is formulated as a Markov decision process (MDP) which is formally defined by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} \rangle$. Where $\mathcal{S} \in \mathbb{R}^n$ is the state-space, $\mathcal{A}$ is the action space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the state transformation function, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the user defined reward function.

At each time step t the learning agent selects an action a_t according to some policy π based on the current system state s_t, i.e., $a_t = \pi(s_t)$. For discrete action space a_t will be from a set of feasible actions in $\mathcal{A}$. On applying this action the system's state s_t transitions to a new state s_{t+1}, along

with this the environment provides a scalar reward r_{t+1}. This process is repeated either until N_s samples or the agent reaches a terminal state, this is called a learning episode. After the termination, the learning episode is restarted, this step is repeated either until N_e episodes or convergence. In RL, it is assumed that the system under consideration satisfies the Markovian property [13].

The goal in RL is to learn a policy π so as to maximize the total accumulated reward termed as return $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ where the scalar constant $\gamma \in (0, 1]$ is called the discount factor. For discrete action space the optimal policy is obtained by solving for the optimal Q function, called action value function. An optimal Q function satisfies the Bellman equation:

$$Q^{\pi_{\text{opt}}}(s, a) = Q^*(s, a) = \mathbb{E}_{s'} \left[r + \gamma \max_{a'} Q^*(s', a') | (s, a) \right].$$
(1)

Any action a that satisfies (1) in state s will be an optimal action in that state. Since last few years RL has had a significant success [1], [2], [14], [4], [3]. This can be mainly attributed to augmenting RL with deep learning techniques [15].

B. Decision making and low-level actuation

In this section, inspired by an example from [13] (see Chapter 1.2), we elaborate on the difference between decision making ($\mathcal{D}$'s) and low level actuation ($\mathcal{C}$'s) : Phil wants to have breakfast, he may chose between having cereal or a bagel ($\mathcal{D}_1$), after deciding he will either walk to the cupboard or to the counter ($\mathcal{C}_1$). Based on $\mathcal{D}_1$ he will either pick a choice of cereal ($\mathcal{D}_2$) then walk to fridge ($\mathcal{C}_2$) to get milk or chose a bagel ($\mathcal{D}_3$) and walk to the toaster ($\mathcal{C}_3$), etc. This entire process involves a series of decision making followed by low-level actuation. A high-level decisions (i.e., $\mathcal{D}$'s) are decided by reward function such as pleasure or getting enough nutrition etc. Whereas low level actuation is characterized by a series of hand-eye coordination.

Any RL methodology that tries to solve both decision making and actuation simultaneously may require a large amount of training data. In this work we simplify this process by having a clear hierarchy between high-level decision making and low level actuation. Whereas RL is used to solve the decision making problem, classical feedback control method such as PID is used for low-level actuation. In a way the methodology presented here has considerable overlap with hierarchical RL where both decision making and control are learned simultaneously subject to a set of related but different cost functions [16]. For training the decision making agent we use a modified DDQN algorithm from [7].

C. Need for safety

A key distinguishing component of RL when compared with other forms of machine learning is the trade-off between exploration and exploitation [13]. In order to achieve better reward an agent must try actions that it has not tried prior. Hence during the initial learning phase the agent will try all viable actions, unfortunately this curiosity may be fatal

and can become expensive particularly when learning on a physical platform such as a robotic platform [17]. For example while training a highway driving agent, use of unguided exploration could frequently lead to collision or near-miss scenarios, which may result in simulation resets, thus slowing the learning process. Additionally even after convergence, due to the function approximation by the Q-network, the trained agent may chose a non-safe maneuver. In order to address these issues, we augment the DDQN decision maker with an explicit *short-horizon safety check* that is used both while learning and also during the implementation phase. Safe exploration or safe-RL is an active research topic, for a detailed survey on inclusion of various safety mechanisms during learning see [10].

A standard RL approach, such as DDQN, may require a lot of samples before learning that few actions could be potentially dangerous in certain states. For example, consider a highway where one of the lane has been barricaded by concrete blocks, see Fig. 1a. A standard epsilon-greedy algorithm may need to collide multiple times before learning that the action *change lane to right* when in the right-most lane to be catastrophic. This may result in significant waste of learning effort as the agent will spend considerable time exploring irrelevant regions of the state and action space. Additionally, by virtue of function approximation there is a small probability, where the trained DDQN agent may still chose *change lane to right* leading to catastrophic outcome. This can be avoided by including an explicit short-horizon safety check that evaluates the action choice by the learning agent and provides an alternative safe action whenever it is feasible [11].

In the RL algorithm presented in this work (see Section III) we use a simple safety check based on common sense road rules. This forces the agent to avoid non-safe actions in dangerous situations resulting in faster learning. Our approach is similar to a teacher who provides corrective action when it is necessary [18]. Note that the safety controller may not be optimal, i.e., an expert in the task under consideration. Additionally, due to explicit safety check, new data can be obtained even in the implementation phase which can then be used for continuous adaption of the learned network, this is further elaborated in Section III.

## III. SAFE AND ADAPTIVE DECISION MAKING FOR AV

The DRL architecture used in this work for autonomous highway driving is given in Fig. 2.

Unlike the mediated perception method that relies on complete reconstruction of scene prior [19], [20], we use the concept of affordance indicators, i.e., state variables based on the direct perception approach from [21]. For a three lane highway scenario, the ego vehicle (EV) can be surrounded by up to six traffic vehicles (TV), see Fig. 1b. To define the traffic vehicle's variable we use the following notation

$$\mathrm{State}_{\mathrm{Lane}_{\mathrm{Location\&Direction}}}$$
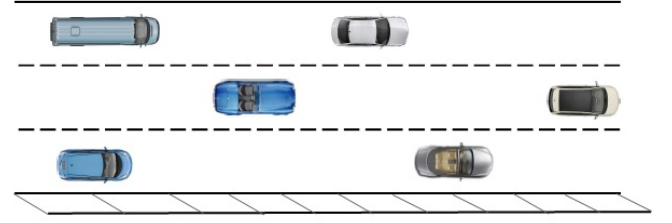
(a) Highway with barricaded right lane.
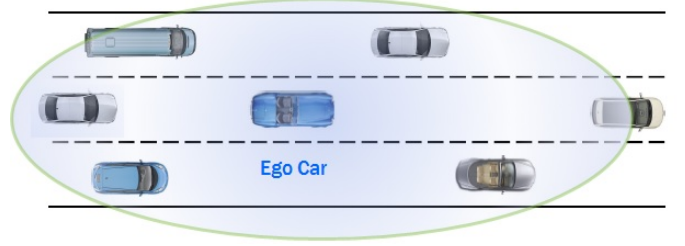
(b) A three lane highway scenario.

Fig. 1: An example for common sense road rule along with the ego vehicle perspective considered in this work.

where State can be distance d or velocity v, Lane is either right lane rl, center lane cl, or left lane ll. Location is either front f or rear r and Direction is either longitudinal x or lateral y.

In total the following 24 indicators are used represent the spatio-temporal information of the six nearest traffic vehicles (see Fig. 1b), they are formulated from the ego vehicle's perspective as:

- The relative distance and velocity, in longitudinal and lateral direction, to the closest front car in center lane, i.e., $d_{\mathrm{cl_{fx}}}$, $v_{\mathrm{cl_{fx}}}$, $d_{\mathrm{cl_{fy}}}$, and $v_{\mathrm{cl_{fy}}}$. Similarly to the closest rear car in the center lane, i.e., $d_{\mathrm{cl_{rx}}}$, $v_{\mathrm{cl_{rx}}}$, $d_{\mathrm{cl_{ry}}}$, and $v_{\mathrm{cl_{ry}}}$.
- The relative distance and velocity, in longitudinal and lateral direction, to the closest front car in the right lane, i.e., $d_{\mathrm{rl_{fx}}}$, $v_{\mathrm{rl_{fx}}}$, $d_{\mathrm{rl_{fy}}}$, and $v_{\mathrm{rl_{fy}}}$. Similarly to the closest rear car in the right lane, i.e., $d_{\mathrm{rl_{rx}}}$, $v_{\mathrm{rl_{rx}}}$, $d_{\mathrm{rl_{ry}}}$, and $v_{\mathrm{rl_{ry}}}$.
- The relative distance and velocity, in longitudinal and lateral direction, to the closest front car in the left lane, i.e., $d_{\mathrm{ll_{fx}}}$, $v_{\mathrm{ll_{fx}}}$, $d_{\mathrm{ll_{fy}}}$, and $v_{\mathrm{ll_{fy}}}$. Similarly to the closest rear car in the left lane, i.e., $d_{\mathrm{ll_{rx}}}$, $v_{\mathrm{ll_{rx}}}$, $d_{\mathrm{ll_{ry}}}$, and $v_{\mathrm{ll_{ry}}}$.

Here the lane occupancy of a TV and the closest car to EV are obtained using the methodology presented in [22]. In addition to the 24 traffic vehicle states we use longitudinal velocity $v_{\mathrm{e_x}}$, lateral position $d_{\mathrm{e_y}}$, and lateral velocity $v_{\mathrm{e_y}}$, of the ego vehicle e. Since the affordance indicators are formulated w.r.t. ego vehicle we do not need a state corresponding to longitudinal position of the ego vehicle. These variables are minimal requirement for highway driving, however they are not sufficient for all highway driving tasks such as use of restricted lane, on-ramp to enter the highway, off-ramp to exit, etc.

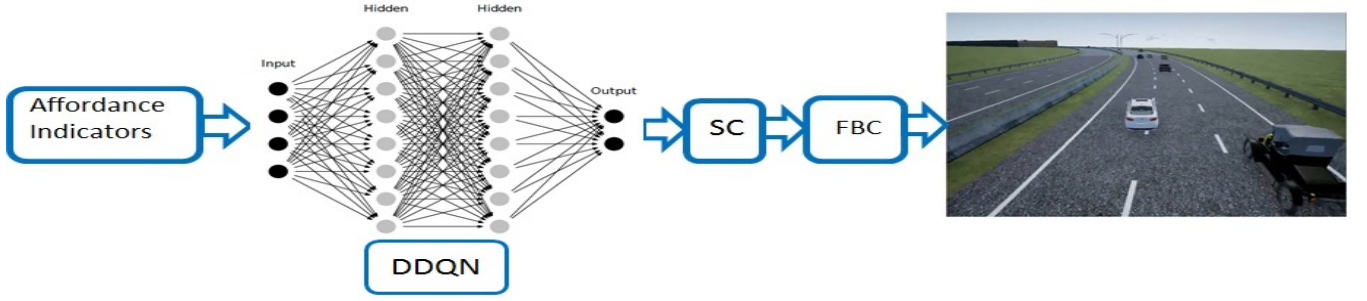A total of 27 affordance indicators, i.e., $s_t \in \mathbb{R}^{27}$ are

Fig. 2: DRL agent control architecture: *SC* is the short-horizon safety check and *FBC* is the low-level feedback controller.

used as an input to the deep Q-network. The Q-network is trained by using a customized variant of the deep double Q-learning algorithm from [7], see Algorithm 1. For decision making of an AV in a highway driving scenario, we consider four action choices along longitudinal direction, namely, *maintain*, *accelerate*, *brake*, and *hard brake*. Whereas for lateral direction we assume three action choices, one for *lane keep*, *change lane to right*, and *change lane to left*, respectively. This results in 12 unique action choices. For each of these action choices a numerical value can be assigned as in [23], for example for longitudinal acceleration four different discrete choices $\{a_1, 0, -a_1, -a_2\}$ may be considered. Alternatively, one can obtain the reference for the throttle or brake controller either using the intelligent driver model (IDM) [24] or adaptive cruise control [25], etc. The reference for the steering controller is self-evident, it is either stay in lane or change lane to right or left. To obtain the steering command we use a simple feedback controller [26]

$$\kappa_{\mathrm{cmd}} = f(k_{\mathrm{road}}, e_{\mathrm{y_{off}}}, e_\psi, T_{\mathrm{LC}}, v_{\mathrm{e_x}}) \tag{2}$$

where k_{road} is the road curvature, e_ψ is the heading angle offset, T_{LC} is the desired time to complete a lane change, and the reference $e_{\mathrm{y_{off}}}$ is the lateral offset to the desired position. When the DDQN agent decides to perform a lane change, the absolute value of the lane offset $e_{\mathrm{y_{off}}}$ will be set to the lane width. The inputs to the steering feedback controller (2) can be considered as additional affordance indicators which can be obtained from the perception module.

As elaborated in Section II-C, we use an explicit short-horizon safety check to validate the action choice by DDQN. For the current action choice, the safety controller verifies common sense but well known rules of the road such as ensuring a minimum relative gap to a TV based on its relative velocity,

$$d_{\mathrm{TV}} - T_{\min} * v_{\mathrm{TV}} > d_{\mathrm{TV_{min}}} \tag{3}$$

where d_{TV}, v_{TV} are the relative distance and relative velocity to a traffic vehicle, $T_{\min}$ is the minimum time to collision, $d_{\mathrm{TV_{min}}}$ is the minimum gap which must be ensured before executing the action choice by the DDQN. If this condition is not satisfied and when it is feasible, an alternate safe action will be provided by the short-horizon safety controller. In this example we use a simple variant of the intelligent driver model (IDM) [24] to provide safe alternative longitudinal

action, and is formulated as

$$a_{\mathrm{s}} = \begin{cases} \text{Hard brake} & \text{if } T_{\mathrm{C}} \leq T_{\mathrm{HB}} \\ \text{Brake} & \text{if } T_{\mathrm{HB}} < T_{\mathrm{C}} \leq T_{\mathrm{B}} \\ \text{Maintain} & \text{if } T_{\mathrm{B}} < T_{\mathrm{C}} \end{cases} \tag{4}$$

where a_{s} is the safe action, T_{C} is the calculated time to collision [27], T_{HB} and T_{B} are the thresholds above which the decision made by the DDQN agent is considered to be safe.

In particular we use the following safety check prior to performing an action given by DDQN:

1) Instead of the in-lane longitudinal action by DDQN, chose a safe action using (4) if (3) is not satisfied and the ego vehicle is faster than the preceding vehicle.
2) If the ego vehicle is in left most lane then *change lane to left* is not valid, similarly for the right lane.
3) For *change lane to left* continuously monitor (3) for preceding car, the front and the rear car in the target lane. If condition (3) fails then lane change is either not initiated or aborted, similarly for the *change lane to right*. Lane abort is initiated by switching to an appropriate action which is opposite to the current lateral direction.

Generally in many RL applications after learning, the trained agent is frozen and used as a feedback controller. However, in reality the agent may encounter new information, additionally there can be a considerable variation between the training environment and the real-world experience. Also due of function approximation there can be a small probability of choosing an unsafe action even by the trained agent, this can happen even after convergence and in the absence of any explicit exploration. In order to address these issues, in the implementation phase we augment the trained DDQN agent with the short-horizon safety check that was used during learning. Any new safety violation data will be added to the collision buffer $\mathrm{Buf_C}$, by using the training part of the Algorithm 1 (line 13 until 15) the learned agent can be re-trained or adapted in a continuous manner. In the following section we apply the developed DRL based decision making Algorithm 1 for autonomous highway driving.

Algorithm 1 A DRL based safe decision maker for autonomous highway driving

---

1: Initialize: Buf_S, Buf_C, $Q(\theta)$ and target network $\hat{Q}(\hat{\theta})$ with $\hat{\theta} = \theta$
2: **for** episode $=1,N_e$, **do**
3:    Initialize: $\{1, \cdots, N_T\}$ cars randomly, obtain affordance indicator s_0
4:    **for** samples $t = 1, N_S$, or Collision, **do**
5:       With ϵ select random action a_t, else $a_t = \arg\max_a Q\left((s_t, a, \theta_t)\right)$
6:       For ego car: **If** a_t is not safe **Then** store $(s_t, a_t, *, r_{\text{col}})$ in Buf_C and replace a_t by safe action a_s
7:       Apply action, observe s_{t+1} and obtain $r_{t+1} = \rho(s_t, s_{t+1}, a_t)$
8:       **if** Collision **then**
9:          Store transition $(s_t, a_t, *, r_{\text{col}})$ in collision buffer Buf_C
10:       **else**
11:          Store transition $(s_t, a_t, s_{t+1}, r_{t+1})$ in safe buffer Buf_S
12:       **end if**
13:       Sample random minibatch $(s_j, a_j, s_{j+1}, r_{j+1})$ from Buf_S and Buf_C
14:       Set
$$y_j = \begin{cases} r_{j+1} & \text{if sample is from } \text{Buf}_C \\ r_{j+1} + \gamma\hat{Q}\Big(s_{j+1}, \arg\max_a Q\left(s_{j+1}, a, \theta_t\right) \\ \qquad\qquad , \hat{\theta}_t\Big) & \text{if sample is from } \text{Buf}_S \end{cases}$$
15:       Perform gradient descent on $(y_j - Q(s_j, a_j, \theta_t))^2$ w.r.t. θ
16:       Every N_C episodes set $\hat{Q} = Q$
17:    **end for**
18: **end for**

IV. Use of DRL for decision making

In this section we will show the usability of our DRL based decision making Algorithm 1 for autonomous highway driving. First, we introduce the vehicle dynamical model that was used training, following this we will elaborate on the training environment and evaluate the learned policy.

A. Vehicle dynamics

Each vehicle is modeled using a computationally efficient point-mass model. For longitudinal equations of motion we use a discrete-time double integrator,

$$x(t+1) = x(t) + v_x(t)\Delta t$$
$$v_x(t+1) = v_x(t) + a_x(t)\Delta t. \tag{5}$$

where t is the time index, Δt is the sampling time, $x \in \mathbb{R}$ is the longitudinal position, and $v_x \in \mathbb{R}$ is the longitudinal velocity of the vehicle. For the lateral motion we assume a simple kinematic model

$$y(t+1) = y(t) + v_y(t)\Delta t \tag{6}$$

where $y \in \mathbb{R}$ is the lateral position of the car. In (5) and (6), the external control inputs $a_x(t)$ and $v_y(t)$ represents the longitudinal acceleration and lateral velocity of the vehicle, respectively.

We assume $a_x(t)$ to vary from nominal acceleration, to hard brake and is discretized into four values, i.e., $a_x = \{a_1, 0, -a_1, -a_2\}$, with $a_1 = 2\,m/s^2$ and $a_2 = 4\,m/s^2$. Only in case of emergency hard braking of $a_x = -a_2$ is applied. The lateral velocity $v_y(t)$ provides a reference lane for the vehicle, we assume a lane change action requires 5 seconds to complete [28], with an option of aborting the lane change maneuver during each sampling instance. In this work we use $1\,Hz$ sampling.

B. Simulation environment

A schematic of the simulation environment used for training is given in Fig. 3. It is a three lane circular loop and is used to approximate an infinite stretch of straight highway. At the beginning of an episode, anywhere between $\{1, \cdots, N_T\}$ number of cars are placed randomly within a distance of $250\,m$ from the ego car. In this example we chose N_T to be 30.

During learning stage the ego car (for e.g., white Fusion in Fig. 3) uses an ϵ-greedy policy to make decisions, whereas for the traffic vehicles a combination of controllers from [23] and [22] are used along with an IDM controller. Additionally the traffic vehicles can randomly chose to perform lane change. For the traffic vehicles, the system parameters such as maximum velocity are randomly chosen. This is to ensure a diverse traffic scenario in training and evaluation. We assume that all the traffic vehicles take into account the relative distance and velocity to preceding vehicle before making a decision, i.e., they will not rear end the preceding car in the same lane. We use Algorithm 1 to train an agent for decision making.
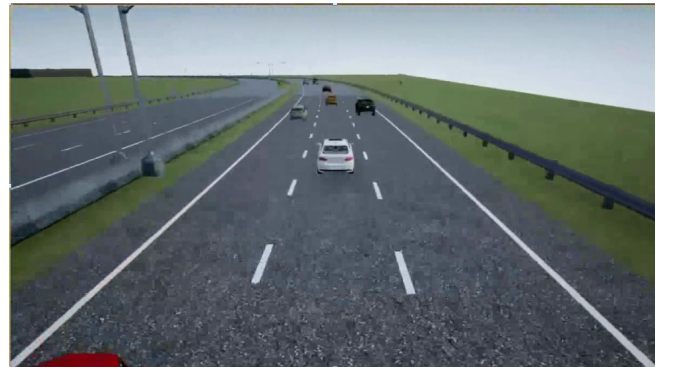

Fig. 3: A schematic of the simulation environment used for training.

C. Decision making for ego vehicle

In order to train the policy π we use a reward function ρ that consists of a set driving goals for the ego car. It is formulated as a function of

- Desired traveling speed subject to traffic condition (7),

- Desired lane and lane offset subject to traffic condition (8),
- Relative distance to the preceding car based on relative velocity (9),

$$r_v = e^{-\frac{(v_{e_x} - v_{des})^2}{10}} - 1, \tag{7}$$

$$r_y = e^{-\frac{\left(d_{e_y} - y_{des}\right)^2}{10}} - 1, \tag{8}$$

$$r_x = \begin{cases} e^{-\frac{(d_{lead} - d_{safe})^2}{10 d_{safe}}} - 1 & \text{if } e_x < d_{safe} \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where v_{e_x}, d_{e_y}, and d_{lead} are the ego velocity, lateral position, and the longitudinal distance to the lead vehicle respectively. Similarly, v_{des}, y_{des}, and d_{safe} are the desired speed, lane position, and safe longitudinal distance to the lead vehicle respectively.

Fig. 4 gives an indicative plot of the reward functions (7)-(9), it is formulated assuming $v_{des} = 30\ m/s$ which can be achieved in the center lane i.e., $y_{des} = 3.8m$ with a minimum safe distance $d_{safe} = 40m$. The desired values are based on the traffic condition and can change depending on the scenario. For slow/fast moving traffic the peak in Fig. 4c will be adjusted based on the traffic condition. In this work we penalize the ego vehicle if it cannot maintain a minimum time headway of at least 1.3 seconds.
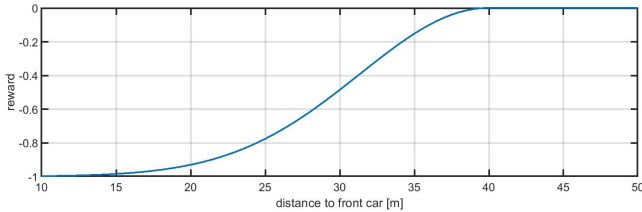
(a) Reward for desired relative distance.
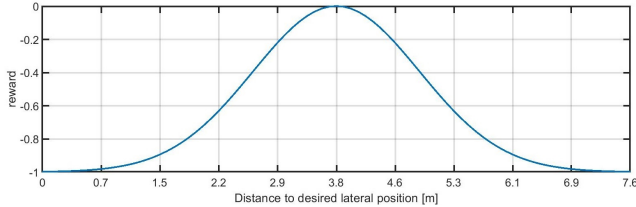
(b) Reward for desired lateral position.
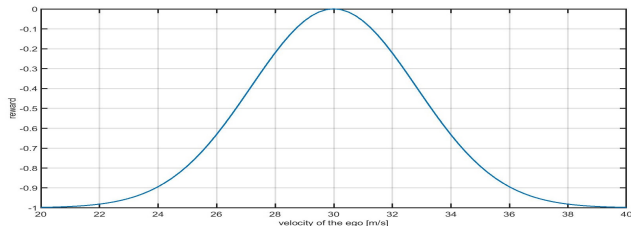
(c) Reward for desired ego speed.

Fig. 4: Reward for the ego car based on traffic condition, sub goals are weighted equally when calculating the final reward.

During learning, we evaluate the (partially) trained DRL controller every 100th episode. Fig. 5 shows the average reward per decision during the training phase. It takes nearly 2000 episodes for the agent to converge. We train the DRL agent for a total of 10000 episodes. Where each episode lasts until 200 samples or collision, whichever is earlier. Exploration is continuously annealed from 1 to 0.2 over first 7000 episodes and then kept constant for the remaining duration of learning. The Q-network is a deep neural network with 2 hidden layers each having 100 fully connected leaky ReLU's [29]. We train the network using Adam optimizer [30] with a fixed learning rate of $1e - 4$.

For the highway driving task, the safety controller was found to be a key component for learning a meaningful policy. Fig. 5 shows the mean and confidence bound for training with and without safety controllers over 200 training iterations of Algorithm 1. Training a standard DDQN agent without explicit safety check could not learn a decent policy and always resulted in collision. Whereas DDQN with explicit safety check was able to converge to an optimal policy. Based on (7) - (9), the maximum reward an agent can receive is zero per decision, the average reward per decision obtained by our trained DDQN agent with safety check is around -0.025.
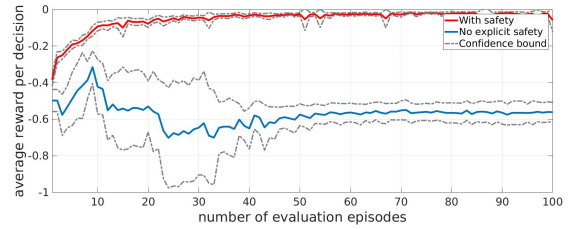

Fig. 5: Average learning curve with confidence bound for with and without short horizon safety check in Algorithm 1.

In Fig. 6 We evaluate our trained DDQN agent to obtain average velocity with increase in traffic density. We compare this against modified safety controller from (4), the modification provides an acceleration command when the calculated time to collision T_C is higher than T_A. This is referred as IDM in Fig. 6. It must be noted IDM controller from (4) cannot initiate lane change, in order to address this we integrate IDM with lane change decision making from [31] and [32]. Fig. 6, clearly demonstrates advantage of RL for high level decision making when compared to model-based approaches. With the increase in traffic density both the trained DDQN agent and the model-based lane change controller converges to IDM controller. This is anticipated since lane change is neither safe nor advantageous in higher traffic density.

Use of two explicit buffers namely Buf_S and Buf_C in Algorithm 1 to store safe and non-safe transitions is simplified version of prioritized experience reply (PER) from [12]. Fig. 8 shows the mean and confidence bound for training with two buffers and PER over 200 training iterations of Algorithm 1. For the highway driving example using two explicit buffers provides marginally better policy
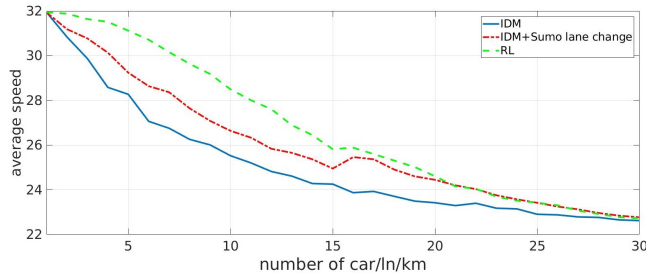
Fig. 6: Average speed for simple IDM controller, with lane change, and trained RL agent.

when compared to PER. This can be due to clear bifurcation of safe and non-safe transitions.
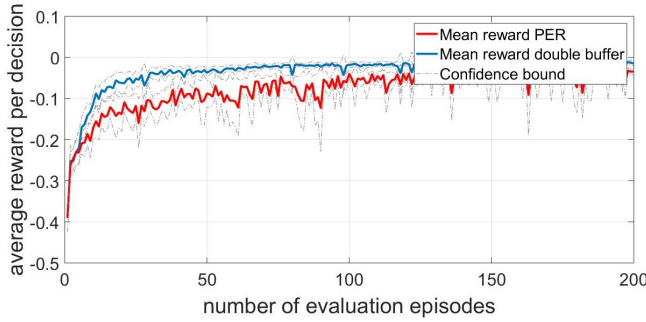

Fig. 7: Mean learning curve with confidence bound for Algorithm 1 and prioritized experience reply [12]. In this work we used the PER implementation from [33].

D. Continuous adaptation

During the implementation phase, we replace the ϵ-greedy policy, i.e., π_ϵ in Algorithm 1 line 5 by the learned policy π. Whenever the control decision by DDQN fails the short-horizon safety check, buffer Buf_C is updated with additional data. Using a lower learning rate than the one used for training, Q-network can be retrained (line 13 until 16). Fig. 8 shows the continuous adaptation result over 30K episodes and is obtained by averaging the data over 10k episodes using a moving average filter. Because of filtering, the mean number of safety trigger increases over first 10k episodes and stays constant for no adaptation scenario whereas it monotonically decreases to a smaller value thanks to continuous adaptation. Even with continuous adaptation the mean number safety trigger never converges to zero, this may be due to

1) Use of function approximation where a trained NN can potentially chose a non-safe action,
2) Use of rigid and static safety rules.

In our future work we plan to address this issue.

V. Conclusions

In this paper, we presented a control architecture based deep RL framework for safe decision making in autonomous driving. The Q-network is augmented with a short-horizon safety control thus embedding few common rules of the road
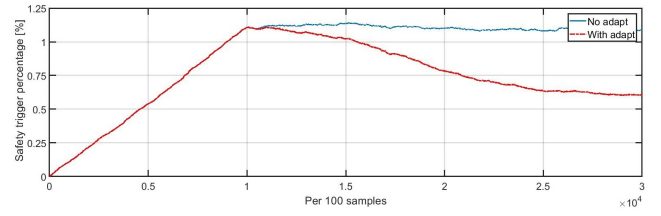

Fig. 8: Comparison of number of safety trigger after learning with and without continuous adaptation.

Symbol	Description	Value
γ	Discount factor	0.9
Δt	Sampling time	1 sec
T_{LC}	Lane change duration	5 sec
$T_{\min}$	Minimum time to collision	3 sec
T_{HB}	Hard break threshold	2 sec
T_{B}	Nominal break threshold	3 sec
T_{A}	Nominal acceleration threshold	8 sec
$d_{\mathrm{TV}_{\min}}$	Minimum gap	15 m

TABLE I: Parameter values used in this work.

into AV decision making. By using well known feedback controller, high level decision by DDQN is converted into a low-level actuation vis-à-vis, throttle, brake, and steering. We have evaluated the learned controller under varying traffic density, the results demonstrate the superior capabilities of the learned DRL agent. Finally, by modifying the learning algorithm we have demonstrated the continuous adaptation framework that has shown to reduce the number of safety triggers. In our future work we plan to extend this work to achieve mandatory lane change, and to perform on-ramp to off-ramp highway driving.

Appendix

Table I lists the parameter values used in this work.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[3] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[4] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*, 2016, pp. 1928–1937.

[5] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, 2016.

[6] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, no. 0, 2018.

[7] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning." in *AAAI*, vol. 16, 2016, pp. 2094–2100.

[8] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[10] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[11] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu, "Safe reinforcement learning via shielding," *arXiv preprint arXiv:1708.08611*, 2017.

[12] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[14] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," *arXiv preprint arXiv:1710.02298*, 2017.

[15] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[16] K. Frans, J. Ho, X. Chen, P. Abbeel, and J. Schulman, "Meta learning shared hierarchies," *arXiv preprint arXiv:1710.09767*, 2017.

[17] J. Garcia and F. Fernández, "Safe exploration of state and action spaces in reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 45, pp. 515–564, 2012.

[18] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.

[19] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2174–2182.

[20] S. Hecker, D. Dai, and L. Van Gool, "End-to-end learning of driving models with surround-view cameras and route planners," in *European Conference on Computer Vision (ECCV)*, 2018.

[21] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2722–2730.

[22] M. Zhang, N. Li, A. Girard, and I. Kolmanovsky, "A finite state machine based automated driving controller and its stochastic optimization," in *ASME 2017 Dynamic Systems and Control Conference*. American Society of Mechanical Engineers, 2017, pp. V002T07A002–V002T07A002.

[23] N. Li, D. W. Oyler, M. Zhang, Y. Yildiz, I. Kolmanovsky, and A. R. Girard, "Game theoretic modeling of driver and vehicle interactions for verification and validation of autonomous vehicle control systems," *IEEE Transactions on control systems technology*, 2017.

[24] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[25] A. Vahidi and A. Eskandarian, "Research advances in intelligent collision avoidance and adaptive cruise control," *IEEE transactions on intelligent transportation systems*, vol. 4, no. 3, pp. 143–153, 2003.

[26] V. Ivanovic, E. Tseng, M. Hafner, and P. Madhavan, "Analytical and experimental study of automated lane change control based on lane centering algorithms," *Internal technical report, Ford Motor Company*, 2017.

[27] M. M. Minderhoud and P. H. Bovy, "Extended time-to-collision measures for road traffic safety assessment," *Accident Analysis & Prevention*, vol. 33, no. 1, pp. 89–97, 2001.

[28] T. Toledo and D. Zohar, "Modeling duration of lane changes," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1999, pp. 71–78, 2007.

[29] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.

[32] J. Erdmann, "Lane-changing model in sumo," *Proceedings of the SUMO2014 modeling mobility with open data*, vol. 24, pp. 77–88, 2014.

[33] A. Hill, A. Raffin, M. Ernestus, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, "Stable baselines," https://github.com/hill-a/stable-baselines, 2018.