Reinforcement Learning

Polaka Surendra Kumar Reddy

Department of Computer Science, Bioengineering, Robotics and System Engineering (DIBRIS)

University of Genova

Supervisor

Armando.tacchella , Stefano.demarchi

In partial fulfillment of the requirements for the degree of

Master of Science in Robotics Engineering

March, 2024

Acknowledgements

First and foremost, I would like to take this opportunity to express my sincere gratitude to my Professor Armando.tacchella and supervisor stefano.demarchi for his valuable advice, guidance and willingness during my thesis. Both expertise and feedback have been invaluable for the success of the work I would like to thanks to all the professors during my course for sharing the knowledge. My close friends have shown all their support and encouragement for the work I am doing. Last and before all, I would like to express my sincere gratitude to my parents who gave me love and hope to succeed

" This Thesis dedicated to my Parents "

Abstract

This thesis provides a comprehensive exploration of Reinforcement Learning (RL) from fundamental principles to advanced tools and experimental applications. Beginning with motivations, Objectives. The transition from tabular methods to neural networks, specifically Convolutional Neural Networks and the Actor-Critic architecture, is highlighted, with a focus on the architectural elegance and the introduction of Soft Actor Critic (SAC).

The thesis then delves into practical implementations through tools and frameworks like Open AI Gym, PyTorch, TensorFlow, and the Never2 Tool (CoCoNet). The Never2 Tool's architectural design, installation process, and procedures for building models, defining properties, and handling models through a command-line interface are outlined. The tool's functionalities extend to training networks, verification strategies, and output visualization.

Experimental results in the Classic control environment are detailed, evaluating different methods and neural network approaches. The network verification process is emphasized, ensuring the robustness of the tool. The thesis concludes by contributing a holistic perspective on RL, bridging theoretical foundations with practical applications, and paving the way for future advancements in RL research and real-world implementations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning (ML) stands as a pivotal branch of artificial intelligence, distinguishing itself by its focus on the development of models and algorithms capable of learning from data and enhancing their performance through iterative processes. This transformative technology has found its significance across diverse industries, including finance, healthcare, marketing, and manufacturing, reshaping the landscape of problem-solving and decision-making.

At the heart of machine learning lies its intrinsic ability to learn from experience, a characteristic that allows it to iteratively refine its understanding as it encounters more data. In this introductory phase, we delve into fundamental concepts that serve as the building blocks of machine learning. This exploration includes a comprehensive understanding of supervised and unsupervised learning, unraveling the intricate machine learning workflow, and familiarizing ourselves with key terminologies such as features and labels. As we embark on this journey into the realm of machine learning, our goal is to illuminate its applications across various industries, shedding light on its transformative potential while acknowledging the challenges that come with harnessing this powerful technology for real-world solutions.

The transformative power of machine learning becomes evident as we examine its applications in diverse sectors. In finance, for instance, machine learning algorithms play a crucial role in risk assessment, fraud detection, and algorithmic trading. These systems can analyze vast datasets to identify patterns indicative of potential risks or fraudulent activities, enabling financial institutions to make informed decisions and safeguard their operations. In healthcare, machine learning contributes to diagnostic processes, drug discovery, and personalized medicine, offering insights that can significantly improve patient outcomes. Marketing efforts are also revolutionized through ML, with personalized recommendations and targeted advertising driven by algorithms that understand consumer preferences and behaviors.

The machine learning workflow encapsulates the iterative process of model development, training, evaluation, and refinement. It begins with defining the problem and collecting relevant data, followed by the crucial step of data preprocessing. This involves cleaning the data, handling missing values, and transforming features to make them suitable for the chosen algorithm. Feature engineering is a critical aspect, where the algorithm's performance can be greatly influenced by the selection and transformation of input features.

Once the data is prepared, it is divided into training and testing sets. The training set is used to teach the algorithm, and the testing set evaluates its performance on new, unseen data. Model training involves adjusting the algorithm's parameters to minimize errors and improve accuracy. Evaluation metrics, such as accuracy, precision, and recall, provide quantitative measures of the model's performance.The iterative nature of machine learning involves refining the model based on evaluation results. This may include adjusting hyperparameters, selecting different algorithms, or incorporating additional features. Continuous refinement is essential to ensure that the model generalizes well to new data and

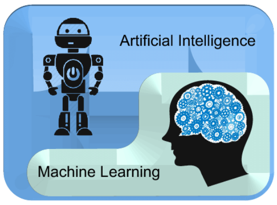


Figure 1.1: Machine Learning is subset of Artificial Intelligence

maintains its accuracy over time.

Key terminologies in machine learning, such as features and labels, are integral to understanding how algorithms operate. Features are the input variables that the algorithm uses to make predictions or classifications. They represent the characteristics or attributes of the data, and their selection and quality greatly impact the model's performance. Labels, on the other hand, are the desired outputs for a given set of inputs in supervised learning. The algorithm learns to associate features with corresponding labels during the training process.

Manufacturing processes benefit from machine learning through predictive maintenance and quality control. By analyzing sensor data and production metrics, ML algorithms can predict when equipment is likely to fail, allowing for timely maintenance that minimizes disruptions to production. Quality control processes are also optimized through ML, ensuring that products meet stringent standards by identifying and addressing potential defects.

machine learning stands as a transformative force with the potential to revolutionize diverse industries. Its ability to process large volumes of data quickly and accurately empowers organizations to make data-driven decisions, leading to improved performance, increased efficiency, and reduced costs. From finance and healthcare to marketing and manufacturing, machine learning applications continue to reshape how we approach problem-solving and decision-making.

As we navigate the intricacies of machine learning, it is crucial to acknowledge not only its potential benefits but also the challenges it presents. The need for high-quality data, interpretability, ethical considerations, and the demand for continuous learning underscore the dynamic nature of this field. By addressing these challenges thoughtfully, we can harness the full potential of machine learning and ensure that it contributes positively to our ever-evolving technological landscape.

There are three main types of machine learning:

(1)Supervised Learning: In supervised learning, the model is trained on a labeled dataset, where the input data is paired with the correct output. The model then uses this labeled data to learn the patterns and relationships between the input and output variables. The goal of supervised learning is to use the trained model to make accurate predictions on new, unseen data.

(2)Unsupervised Learning: In unsupervised learning, the model is trained on an unlabeled dataset, where there is no given output or target variable. The model then learns the underlying structure and patterns within the data without any specific guidance. The goal of unsupervised learning is to uncover hidden patterns or insights in the data that can be useful for further analysis.

(3)Reinforcement Learning: In reinforcement learning, the model learns by interacting with an environment and receiving feedback in the form of rewards or penalties. The goal of reinforcement learning is to maximize the rewards over time by learning from trial and error.

Reinforcement learning (RL) is a key area of machine learning that is particularly useful for solving problems where decisions must be made based on trial and error. In RL, The agent, the brain of reinforcement learning, learns to make decisions according to the information it receives from the surrounding environment and from the positive (or negative) reward it receive. The goal of the agent is to learn a policy that maximizes its expected reward over time. Unlike supervised learning, where the agent is provided with labeled examples of input-output pairs, RL involves learning through trial and error.

RL is particularly useful in situations where there is no existing labeled data or where the optimal policy is not known in advance. For example, RL has been used to train robots to perform complex tasks, to optimize traffic flow in transportation systems, and to develop intelligent agents for video games. One of the key challenges in RL is the exploration-exploitation tradeoff. In order to maximize the expected reward, the agent must explore different actions and learn from the resulting feedback. However, it also needs to exploit its current knowledge to maximize its expected reward in the short term.

There are various RL algorithms, including Q-learning, SARSA, and policy gradient methods. These algorithms differ in their approach to estimating the optimal policy and updating the agent's policy. Overall, RL is an important area of machine learning that has many practical applications. As the field continues to develop, we can expect to see even more exciting and innovative uses of RL in a wide range of fields such as UAV(Unmanned Aerial Vehicles). Unmanned Aerial Vehicles (UAVs), commonly known as drones, have a wide range of applications across various fields due to their versatility and capabilities.

1.1 Motivations

The allure of delving into the realm of reinforcement learning (RL) within the framework of Gymnasium environments is rooted in the captivating convergence of theoretical depth and practical impact. This dynamic subfield of machine learning presents an intellectually stimulating arena, motivating researchers and practitioners to explore its multifaceted dimensions.At the core of this motivation is the dynamic learning paradigm that RL encapsulates. Unlike traditional supervised learning, RL propels agents into a world where learning unfolds through direct interaction with an environment.

The agents, or learners, make sequential decisions based on the feedback received, leading to a continual process of adaptation and optimization.Real-world applicability is a primary motivator for shifting focus to RL. It is especially interesting because RL can address difficult decision-making challenges. RL enables agents to make informed decisions in dynamic and uncertain situations, mimicking the obstacles experienced in real-world scenarios in applications ranging from robotics and autonomous systems to finance and healthcare. The variety and standardization provided by Gymnasium environments, as demonstrated by platforms such as OpenAI Gym, increase the attraction of RL research.

These environments function as dynamic playgrounds, providing a wide range of activities ranging from classic control problems to intricate simulated situations. Researchers use Gymnasium to rigorously benchmark and compare RL algorithms, establishing a culture of healthy competition and collaborative advancement The continual learning and exploration inherent in RL correspond to the ever-changing character of many real-world issues. In a world where optimal solutions may alter due to changing conditions, RL shines by allowing agents to explore, learn from experiences, and adapt their methods dynamically over time.

This adaptability is critical in settings where the environment is not static, necessitating the use of intelligent systems to negotiate uncertainties and unexpected complications.In conjunction with RL, Gymnasium environments offer benchmarking and evaluation. The open-source nature of RL libraries and the availability of Gymnasium resources democratize the learning curve, allowing researchers, students, and practitioners to explore with RL models without incurring the cost of developing specialized settings. This accessibility hastens knowledge diffusion and builds a collaborative community.

The interdisciplinary character of RL adds another dimension of interest to its investigation. RL necessitates an interdisciplinary approach to problem resolution, drawing on concepts from computer science, mathematics, neurobiology, and control theory. This idea exchange not only enriches the learning experience, but also opens the door to collaborative thoughts from various sectors.It represents a journey into a realm where theoretical brilliance meets tangible impact, and where innovation thrives at the intersection of exploration, adaptability, and continuous learning

Beyond the theoretical realm, RL within Gymnasium environments has profound implications for innovation in artificial intelligence. Breakthroughs in RL reverberate across the broader landscape of AI, influencing advancements in autonomous systems, natural language processing, and beyond. Researchers working at this intersection contribute to the ongoing evolution of intelligent systems, pushing the boundaries of what is achievable in the field of machine learning.

1.2 Objectives and Contributions

The Primary objective of the thesis consists of three steps :

1. Create RL-based UAV control policies: We intend to construct Neural Networks (NNs) that can autonomously execute control tasks required for UAVs using RL algorithms such as Proximal Policy Optimization (PPO) and Deep Deterministic Policy Gradients (DDPG). These responsibilities include navigation, obstacle avoidance, and trajectory optimization. The emphasis is on using RL's learning capabilities to adapt to diverse and dynamic settings, giving UAVs the agility required for real-world applications.

2. Evaluating Robustness in Simulation and Real-World Scenarios: The trained NNs will be empirically tested in both simulated and real-world settings.

Simulations provide a controlled environment for preliminary evaluations, allowing us to examine the performance of learned control strategies under a variety of scenarios. Following that, real UAVs will be used to test the robustness of these strategies in complicated and dynamic circumstances. The purpose of this empirical evaluation is to identify potential obstacles, strengths, and flaws of learnt control strategies when applied to actual UAVs

3. Integrate Formal Verification approaches: To evaluate and verify the trained NNs, we will use cutting-edge NN verification approaches, including methodologies pioneered in our lab. Formal verification is a methodical way to finding potential flaws, vulnerabilities, or violations of control policy safety restrictions. The goal is to improve the reliability of UAV control systems by proactively addressing these concerns. Formal verification will provide an additional degree of confidence by confirming that control rules correspond to stated specifications and safety standards

This work has important implications for the development and deployment of UAVs in real-world situations. The goal is to contribute to the creation of a reliable and robust UAV control system by integrating RL methods, empirical tests, and formal verification. This work mainly focuses on the growing need for special and safe technologies, especially in applications where mistakes have high consequences

1.3 Overview of the Thesis

The intersection of reinforcement learning (RL), PyTorch, and specialized control tools such as Never2 in a new era of transformative development in artificial intelligence (AI). It provides a comprehensive framework for understanding development and challenges. and ensuring the security of neural networks. This approach leverages the dynamic computer graphics capabilities of PyTorch, formal verification techniques built into Never2, and interactive interfaces for property specification, providing a comprehensive approach with deep implications across fields.

Reinforcement learning, a central part of this paradigm, has gained significant traction due to its ability to train agents with the environment through trial and error, resulting in decision-making skills spanning applications from games to robotics. RL algorithms such as deep Q networks (DQN) and political gradient methods have shown significant achievements in forming the backbone of intelligent systems that can adapt and learn from their experiences.

Neural networks, the basis of many modern AI applications, introduce a layer of complexity and nonlinearity, especially in the context of deep learning, which involves several hidden layers. These networks, inspired by the human brain, have proven to be very effective in tasks such as image recognition, natural language processing and learning. However, the transparency of deep neural networks presents challenges to understand their decision-making processes, which requires the development of tools and methods for verification and transparency.

PyTorch, an open source deep learning framework, emerged in this landscape. Known for its dynamic computation graph and user-friendly interface, PyTorch facilitates the implementation of complex neural network architectures. Its flexibility has made it the best choice for researchers and professionals involved in developing cutting-edge AI applications. PyTorch's seamless integration with reinforcement learning algorithms has simplified the process of building and testing RL models, enabling rapid prototyping and iterative development.

Never2, a dedicated neural network tool, plays a critical role in addressing growing concerns about the reliability and security of neural networks. In this context, verification requires that neural network models conform to certain properties, constraints, or security guidelines. Never2 uses formal verification methods to mathematically prove the correctness of these models and provides a robust platform for verifying user-defined properties. Its interactive interface allows users to visualize the verification process, explore counterexamples, and gain insight into the decision limits of the neural network. Importantly, Never2 is designed to seamlessly integrate with PyTorch, simplifying the verification process and making it more widely accessible to researchers and practitioners.

The integration of these components has far-reaching effects in various fields. In safety critical applications such as autonomous vehicles and healthcare, the ability to formally verify neural network models becomes paramount to ensure the reliability and ethical operation of AI systems. In addition, the transparency provided by tools like Never2, combined with the flexibility and learnability of PyTorch, accelerates the development of sustainable and verifiable AI solutions, promoting a future where intelligent machines positively influence society.

In conclusion, the convergence of reinforcement learning, PyTorch, and neural network Tool , exemplified by Never2, represents an important step forward in the field of artificial intelligence. This integrated approach not only meets the challenges of developing and understanding complex neural networks, but also lays the foundation for building reliable and transparent AI systems with applications that span different sectors and shape the future of AI development and deployment.

Chapter 2

Reinforcement Learning

Reinforcement Learning (RL) has emerged as a vibrant field in machine learning, propelled by recent strides in deep learning (DL), which paved the way for the evolution of deep reinforcement learning. Positioned as the third paradigm in machine learning, alongside supervised and unsupervised learning, RL introduces a novel approach to decision-making problems. The core concept revolves around the agent, the central figure in RL, engaging in a continuous cycle of trial and error. Within this dynamic, the agent discerns valuable decisions from penalizing ones by leveraging information derived from a reward signal, mirroring the trial-and-error process inherent in human and animal behavior.

To comprehend the current state-of-the-art in RL, this chapter embarks on a journey through the theoretical underpinnings of traditional RL, establishing the notation used. It then seamlessly transitions towards the realm of deep RL, providing an introductory exploration into deep learning fundamentals. The discussion delves into essential algorithms, with a keen focus on those integral to the thesis project. The ultimate section aims to paint a vivid picture of the contemporary landscape of deep RL as applied to autonomous systems and real-world robotic tasks, setting the stage for the subsequent exploration.

Before delving into the thesis results, it is paramount to grasp the intricacies of this paradigm. The fusion of RL with deep learning not only enhances function approximation but also reshapes the landscape of decision-making processes. This synthesis captures the essence of learning through sequential experimentation, enriching the agent's ability to navigate complex environments. As we unravel the chapters that follow, the convergence of RL and deep learning unfolds as a potent force, driving innovation in autonomous systems and real-world robotic applications.

2.1 Fundamentals of Reinforcement Learning

Reinforcement Learning (RL) is a paradigm in machine learning where an agent learns to make decisions by interacting with an environment. The agent is the central entity in this framework, responsible for decision-making. This could manifest as a physical robot, a software algorithm, or any system with the capability to take actions within a given environment. Understanding the components of RL and the relationship between the agent, environment, actions, and rewards is essential to grasp the dynamics of this learning approach. The agent's [1]primary objective is to maximize its cumulative reward over time by learning a strategy or policy that maps environmental states to actions. States represent the current situation or configuration of the environment, actions are the decisions made by the agent, and rewards provide feedback on the desirability of those actions. The environment is the external system with which the agent interacts, and it responds to the actions taken by the agent by transitioning to new states and providing corresponding rewards.

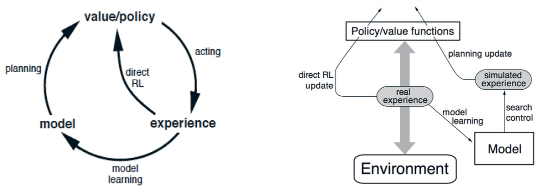


Figure 2.1: Overview of the different components in the Reinforcement Learning

2.1.1 Core Components of Reinforcement Learning:

Agent : The agent encapsulates the decision-making entity within the RL system. It can take various forms, from a physical robot navigating a real-world environment to a software algorithm playing a game. The agent's decision-making process is guided by a policy, a mapping from states to actions. The goal is to learn the optimal policy to improve and maximize the cumulative reward.

Environment: The environment is the external context in which the agent operates. It could be a physical space, a simulated world, or any system that the agent interacts with. The environment has a state, which represents the current situation. The agent's actions influence the environment, causing it to transition to new states

Actions:Actions are the decisions or moves made by the agent within the environment. The set of possible actions is defined by the task at hand and the capabilities of the agent. The agent's goal is to learn a policy that selects actions leading to favorable outcomes, i.e., actions that result in high cumulative rewards.

Rewards: Rewards are numerical values that provide feedback to the agent about the desirability of its actions. The agent's main goal is to improve and maximize the cumulative reward over time. Positive rewards encourage the agent to repeat actions that lead to favorable outcomes, while negative rewards or punishments discourage undesirable actions.

The concept of return

The core objective of an agent in reinforcement learning is to maximize its cumulative reward, termed the return (gt). The return is the sum of discounted rewards from a given timestep (t) onward, represented by the geometric series defined in equation (2.1), where is the discount factor. This factor is crucial not only due to the observed tendency in animal and human behavior to favor immediate rewards over future ones but also for mathematical reasons. Without discounting (= 1), an infinite-horizon sum of rewards may not converge to a finite value, making the agent's learning process mathematically challenging.

$$gt = rt + 1 + art + 2 + \; = \O krt + k + 1, [0, 1) \tag{2.1}$$

The formulation of the return function, a geometric series, ensures convergence when [0,1). In this range, the series converges to a finite value, specifically 1/(1). This mathematical convergence is essential for a meaningful interpretation of the cumulative reward. Notably, the case where = 1 is rationalized only in the context of a finite-horizon cumulative discounted reward, underlining the importance of discounting in aligning mathematical rigor with practical learning scenarios. Overall, the discount factor balances immediate rewards with convergence requirements, shaping the agent's decision-making process in reinforcement learning.

states and Observations

The information provided by the environment in reinforcement learning includes observations (ot), which are directly linked to the state (st). Observations serve as a condensed representation of information used by the agent to determine its next action. The state, denoted as st, is a function of the history, encompassing the sequence of observations, actions, and rewards at timestep t, as expressed in equation (2.2).

$$ht = o1, r1, a1, ..., at1, ot, rt, st = f(ht)$$

(2.2)

Additionally, the state can be categorized into the agent state (sa) and the private state of the environment (se). This differentiation is particularly relevant for distinguishing between fully observable environments (where o = se = sa) and partially observable environments (where se sa). In fully observable scenarios, the agent can directly observe the environment state, while in partially observable cases, the agent has access to only partial information about the environment state.

$$P[st + 1|st] = P[st + 1|s1, ..., st] \tag{2.3}$$

Despite the focus of this chapter on fully observable environments, the distinction between state and observation is often ambiguous. Conventionally, the input for the agent is considered to be composed of both the reward and the state, as illustrated in figure 2.1.

Moreover, a state is termed an informational state or Markov state when it encapsulates all relevant data and information about its history. Formally, a state is considered a Markov state if and only if it satisfies equation (2.3), which asserts that the probability of the next state (st+1) given the current state (st) is independent of the entire history and depends only on the current state. In simpler terms, a Markov state contains all the necessary information for decision-making, rendering the entire history redundant.

The concept of an informational state is fundamental in reinforcement learning as it signifies that the agent's decision-making process relies solely on the current state and not on the entire historical sequence. Importantly, the envi-

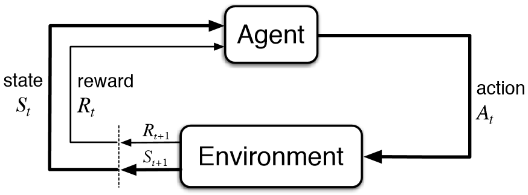


Figure 2.2: Interaction loop between Agent and Environment.

ronment state (se) is inherently a Markov state, aligning with the principle that it contains all essential information for decision-making. This distinction aids in understanding the nature of the information available to the agent and influences the formulation of learning algorithms.

In summary, the environment provides the agent with observations, which are condensed representations of the state. The state is a function of the historical sequence of observations, actions, and rewards. This chapter predominantly addresses fully observable environments but acknowledges the conventional input composition of rewards and states. The informational state, or Markov state, is crucial as it encapsulates all relevant information for decision-making, emphasizing the significance of the current state over the entire historical sequence. The distinction between agent state and environment state is pivotal, particularly in discerning fully observable and partially observable environments.

The Markov Decision Problem

A Markov Decision Problem (MDP) is a formal mathematical framework for modeling decision-making problems involving uncertainty over time. It is characterized by a set of states, actions, transition probabilities, rewards, and a discount factor. In addition to these components, MDPs involve the concepts of policies, models, and value functions, which play crucial roles in finding optimal strategies for decision-making.

1. Policies: A policy in the context of an MDP is a strategy or a rule that defines the agent's behavior. It maps states to actions, specifying the action the

agent should take in each possible state. Policies can be deterministic, prescribing a single action for each state, or stochastic, providing a probability distribution over actions. The goal is to find an optimal policy that maximizes the expected cumulative reward over time.

2. Models: The transition model, often denoted by P P, describes the dynamics of the MDP. It defines the probabilities of transitioning from one state to another based on the agent's actions. In a Markovian setting, the future state depends only on the current state and action, not on the history of events leading to that state. The transition model is a fundamental component for predicting the evolution of the system and is crucial for planning under uncertainty.

3. Value Functions: Value functions are central to solving MDPs and evaluating the desirability of different states and actions. There are two main types of value functions: state value function (V (s) V(s)) and action value function (Q (s , a) Q(s,a)). State Value Function (V (s) V(s)): The state value function represents the expected cumulative reward starting from a given state and following a particular policy. It quantifies the desirability of being in a particular state under a specific policy. Action Value Function (Q (s , a) Q(s,a)): The action value function, also known as the Q-function, represents the expected cumulative reward starting from a given state, taking a specific action, and then following a particular policy. It evaluates the desirability of taking a particular action in a specific state under a given policy. Solving the MDP.

An MDP is defined by where

$$< S, A, P, R, >$$

S is a finite set of states

A finite set of actions

P a state transition probability matrix

R is a rewardfunction Ra =E[r —s =s,a =a]
(2.4)

The primary objective in solving an MDP[5] is to find the optimal policy that maximizes the expected cumulative reward. Dynamic programming methods, such as the Bellman equations, iteratively update the value functions until convergence, guiding the determination of the optimal policy. Reinforcement learning algorithms, which involve learning from interactions with the environment, are also widely employed to find optimal policies when the system dynamics are unknown. In summary, within the framework of a Markov Decision Problem, policies prescribe actions, models describe system dynamics, and value functions assess the desirability of states and actions. The interplay between these components forms the foundation for addressing decision-making problems in environments with uncertainty and sequential interactions, making MDPs a versatile and widely applicable tool in fields such as artificial intelligence, operations research, and control theory.

2.1.2 Bellman Equations

The recursive relationships between state values play a crucial role in understanding the dynamics of Markov Decision Processes (MDPs). These equations capture the dependence of a state's value on the values of its successor states. The Bellman equations, depicted in (2.7) further emphasize this by indicating that the next state is sampled from the environment, denoted as st+1 E.

$$V(s) = maxa(R(s,a) + V(s')) \tag{2.5}$$

State(s): current state where the agent is in the environment

Next State(s'): After taking action(a) at state(s) the agent reaches s'

Value(V): Numeric representation of a state which helps the agent to find its path.V(s) here means the value of the state s.

Reward(R): treat which the agent gets after performing an action(a).

R(s): reward for being in the state s
R(s,a): reward for being in the state and performing an action a
R(s,a,s'): reward for being in a state s, taking an action a and ending up in s'
Example Good reward can be +1, Bad reward can be -1, No reward can be 0.

The essence of value functions lies in their ability to establish a partial ordering among policies. Specifically, for policies and ', if V is greater than or equal to V' for every state in the state space S, then is considered better than or equal to ', denoted as '. This ordering sets the stage for the sanity theorem, which posits that for any MDP, there exists an optimal policy, denoted as , surpassing or equal to all other policies, i.e., for any policy . Moreover, all optimal policies share the achievement of the optimal state-value function and the optimal action-value function.

Despite the significance of the Bellman optimality equation in characterizing optimal policies, its solution is non-linear, lacking a closed-form expression. . These iterative approaches provide computational tools to converge towards the optimal policy and value functions, offering a practical means to navigate the challenges posed by the non-linearity of the Bellman optimality equation

Policy Iteration is a key DP strategy aimed at discovering the optimal policy by directly manipulating the starting policy. However, before embarking on this process, a thorough evaluation of the current policy is essential. This evaluation follows an iterative procedure, as outlined in algorithm A.1, where the parameter defines the accuracy of the evaluation. A lower value indicates a more precise

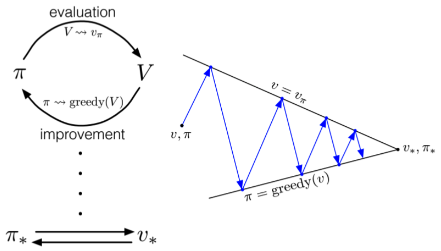


Figure 2.3: Generalised Policy Iteration Value and Policy functions.

evaluation.

2.1.3 Model-Free Approaches in Reinforcement Learning: Monte Carlo and Temporal

Temporal Difference Learning: Temporal Difference (TD) learning represents an amalgamation of ideas from Monte Carlo methods and dynamic programming. Like Monte Carlo, TD is a model-free method but incorporates bootstrapping for updates, akin to dynamic programming. TD methods diverge by calculating a temporal error instead of relying on the total accumulated reward. This error, the difference between the new and old value function estimates, is computed considering the reward received at the current time step, enabling TD methods to operate in continuing (non-terminating) environments.

The update equation for the value function in TD is expressed as Eq. (2.6), emphasizing the TD error and TD target. This equation underlines the balance between reduced variance compared to Monte Carlo methods and increased bias due to bootstrapping.

$$V(st)\text{¸}V(st) + [rt + 1 + V(st + 1)V(st)] \tag{2.6}$$

Two prominent TD algorithms for RL control are SARSA (State-Action-

Reward-State-Action) and Q-Learning. SARSA is an on-policy algorithm focusing on learning an action-value function, emphasizing transitions and state-action pairs rather than specific state values. Eq. (2.7) represents the SARSA update function, where the algorithm aims to refine action-value estimates.

$$Q(st, at)¸Q(st, at) + [rt + 1 + Q(st + 1, at + 1)Q(st, at)] \tag{2.7}$$

Q-Learning: Q-Learning, an off-policy TD control algorithm, stands as a milestone in the evolution of reinforcement learning. It diverges from SARSA by estimating state-action values directly, marking a significant advancement in RL. Q-Learning's impact on the field highlights its ability to learn optimal policies while being off-policy.

2.1.4 Exploring Model-Based Approaches in Reinforcement Learning

While the preceding sections delved into model-free methods, crucial in the context of this thesis, it is pertinent to provide an overview of model-based approaches. These methods, grounded in the acquisition of knowledge about the environment, aim to facilitate planning and enhance algorithmic sample efficiency.

Model-based learning hinges on two fundamental principles. The first involves constructing a model based on prior knowledge and leveraging it to calculate both the policy and the value function. However, the accuracy of prior knowledge poses a potential limitation, leading to sub-optimal results. The second principle involves inferring the model directly from the environment through sampling experiences. This approach is favored as it mitigates reliance on potentially inaccurate prior knowledge

Model-based approaches are distinguished by their enhanced sample efficiency compared to model-free methods. They require fewer data to learn a policy, a valuable trait when resources are limited. However, it comes at the cost of increased computational complexity. The algorithm not only needs to learn the policy but also the model, introducing two distinct sources of approximation errors.

2.2 Evolution from Tabular Methods to Neural Networks in Reinforcement Learning

Reinforcement learning strategies designed for systems with well-defined states and actions often rely on lookup tables. In this paradigm, the state-value function V V and action-value function Q Q have entries for each state and state-action pair, respectively. However, this approach faces significant challenges when scaling up to large Markov Decision Processes (MDPs). Issues related to memory constraints, slow individual state learning, and the impracticality of linear lookup in continuous action and state spaces become apparent.

Function approximators offer a compelling solution to overcome these challenges. By utilizing a parameter vector = (1 , 2 , . . . , n) T =(1 , 2 ,..., n) T , these approximators estimate V V and Q Q functions, enabling generalization from observed to unseen states. Equation (2.14) showcases this estimation, with V (s ,) V(s,) approximating V (s) V (s) and Q (s , a ,) Q(s,a,) approximating Q (s , a) Q (s,a). Function approximators act as a mapping from the vector to the value function, reducing the number of parameters to learn and enhancing generalization with fewer training samples.

In the contemporary landscape of research, neural networks have emerged as the most intuitive option for function approximation. Their widespread use is driven by their ability to reduce training time for high-dimensional systems

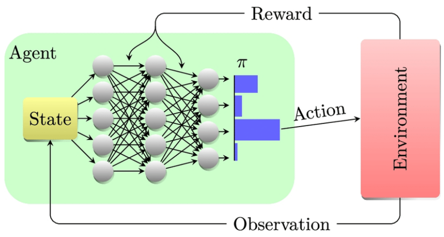


Figure 2.4: Scientific Diagram of Neural Network

and their efficient memory utilization. This integration serves as a crucial bridge between traditional reinforcement learning methods and recent advancements in deep learning theory. The enthusiasm surrounding deep learning over the last decade has established neural networks as fundamental tools for developing deep reinforcement learning (Deep RL), yielding remarkable results.

DeepMind's seminal papers [44] and [45] mark a pivotal step toward Deep RL and general artificial intelligence. These contributions demonstrate the broad applicability of AI across various environments. Given the focus of this work on model-free algorithms, the ensuing section explores the state-of-the-art theories underpinning the Deep RL framework. Additionally, it provides an overview of deep learning, culminating in the presentation of two deep actor-critic algorithms employed in the thesis experiments: Deep Deterministic Policy Gradient (DDPG) and Soft Actor-Critic (SAC). This comprehensive exploration encapsulates the evolution from tabular methods to the pivotal role of neural networks in contemporary reinforcement learning landscapes.

2.2.1 Architectural Elegance in Convolutional Neural Networks

Sensory reception, a fundamental aspect of how humans and animals react to changes, involves the use of sensors that process input data and respond to specific stimuli. This concept serves as inspiration for the architecture of Convolutional

Neural Networks (CNNs), designed to efficiently handle significant input data, particularly finding applications in computer vision.

A notable representation of CNN [2]architecture is LeNet-5 [40], showcased in Figure 2.6 on the following page. LeNet-5 is adept at recognizing digits in images, making it a quintessential example of a standard convolutional neural network. This architecture typically comprises a sequence of convolutional layers followed by a subsampling pooling layer. In the convolutional stack's culmination, the values map into the final hidden layers of the network, ultimately computing the low-dimensional output. These final layers often consist of fully-connected layers.

The key strength of CNNs lies in their ability to hierarchically learn features from input data. It is conceivable that the initial layers focus on learning low-level features, such as edges and textures, from the input data. As one progresses through the network, the subsequent layers then combine these low-level features to form more abstract and high-level representations. This hierarchical feature learning makes CNNs particularly effective in tasks like image recognition.

The convolutional layers in a CNN perform the vital task of convolving filters or kernels over the input data. This operation helps detect local patterns, allowing the network to recognize features in various spatial locations. The subsampling pooling layer contributes to spatial invariance, enhancing the network's robustness to translations and distortions in the input data.

In summary, Convolutional Neural Networks draw inspiration from sensory reception systems, leveraging a hierarchical architecture to efficiently process substantial input data. The example of LeNet-5 demonstrates the standard structure of a CNN, emphasizing the role of convolutional and pooling layers in learning hierarchical features. This understanding of CNNs underscores their significance, especially in the realm of computer vision applications, where they excel in tasks such as image recognition

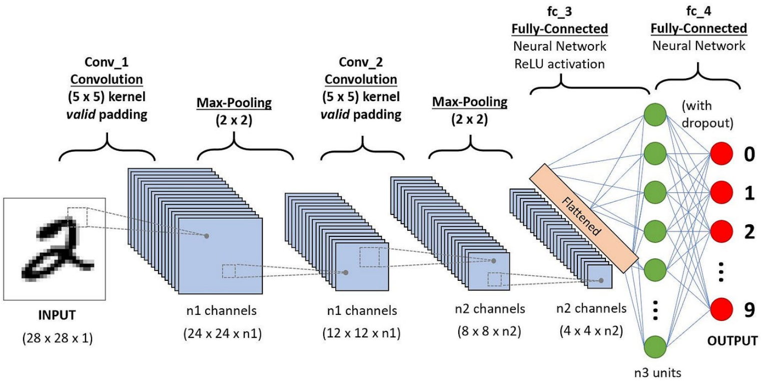


Figure 2.5: Schematic Diagram of Convolutional Neural Networks

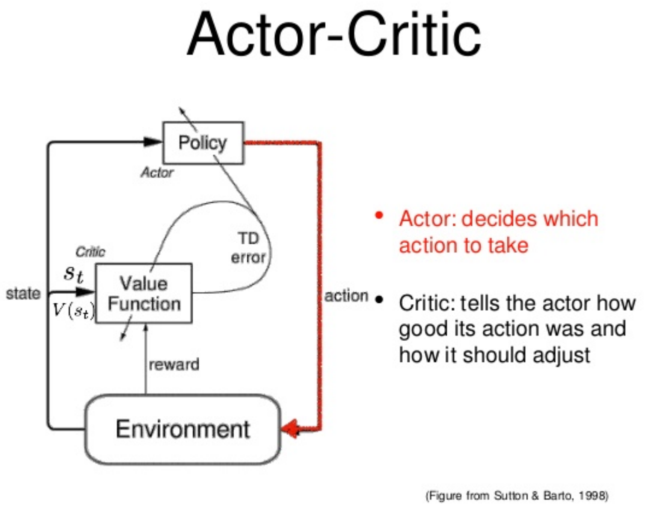
2.2.2 Actor Critic Architecture

The Actor-Critic architecture, illustrated in Figure 2.6 on the subsequent page, stands as the juncture where value-based approaches intersect with policy gradient methods. Essentially policy gradient methods and actor-critic methods utilize the value function to learn the parameters

of the policy. This approach involves two distinct components, the actor, and the critic, forming a symbiotic relationship in the pursuit of effective reinforcement learning.

In this tandem framework, the actor pertains to the policy, dictating the agent's actions in the environment. Conversely, the critic is responsible for estimating a value function, often in the form of a Q-value function. Deep reinforcement learning integrates neural networks as function approximators [43] to represent both the actor and the critic. The actor leverages gradients derived from the policy gradient theorem to adjust policy parameters, while the critic estimates the approximate value function corresponding to the current policy .

A common strategy in Actor-Critic architectures is to update both networks using the Temporal Difference (TD) Error, as discussed in Section 2.1.5 on page 15. The critic's estimation plays a pivotal role in determining the contribution



Actor-Critic

- Actor: decides which action to take
- Critic: tells the actor how good its action was and how it should adjust

(Figure from Sutton & Barto, 1998)

Figure 2.6: Sample Diagram of Actor-Critic Architecture

that expected values of the current and next state provide to the TD-error. Essentially, the output of the critic becomes instrumental in the update of the actor's parameters, fostering a dynamic interplay between policy improvement and value estimation.

The synergy between the actor and critic in the Actor-Critic architecture enhances the efficiency of reinforcement learning algorithms. The actor learns to improve decision-making policies by incorporating feedback from the critic, which, in turn, refines its estimates based on the observed performance of the actor. This dual-learning process contributes to the stability and effectiveness of the overall system, making Actor-Critic architectures a popular choice in the realm of deep reinforcement learning

2.3 Soft Actor Critic(SAC)

Soft Actor-Critic (SAC) innovatively combines the off-policy actor-critic setup with a stochastic policy, creating a link between stochastic policy optimization and DDPG-style approaches. This proves especially valuable in scenarios with continuous action spaces[6], showcasing SAC's model-free capabilities. Unlike DDPG, SAC addresses the challenges associated with stabilizing and tuning hyperparameters, providing a robust alternative.

DDPG's Achilles' heel lies in the interplay between the deterministic actor network[3] and the Q-function, resulting in instability and sensitivity to tuning. The learned Q-function tends to overestimate Q-values, leading to policy breakdown by exploiting errors in the Q-function. SAC mitigates this by adopting Clipped Double-Q Learning, a technique also employed by Twin Delayed DDPG (TD3). SAC employs two Q-functions, using the smaller Q-value to formulate targets in the Bellman error loss functions, enhancing stability.

Entropy regularization is another standout feature of SAC. The policy is trained to optimize a trade-off between expected return and entropy, a measure of policy randomness. This characteristic directly addresses the exploration-exploitation trade-off, where increased entropy facilitates more exploration, accelerating learning while preventing premature convergence to local optima.

In SAC, five neural networks come into play: the local stochastic policy network with parameter , two local Q-Networks with parameters 1 and 2, and two target Q-Networks with parameters 1 and 2. The behavior mirrors that of DDPG target networks, updating through the algorithm's specified equations. This ensemble of networks contributes to SAC's effectiveness in handling complex reinforcement learning tasks.

Entropy-regularized reinforcement learning introduces the concept of entropy, which quantifies the average rate at which a stochastic data source produces

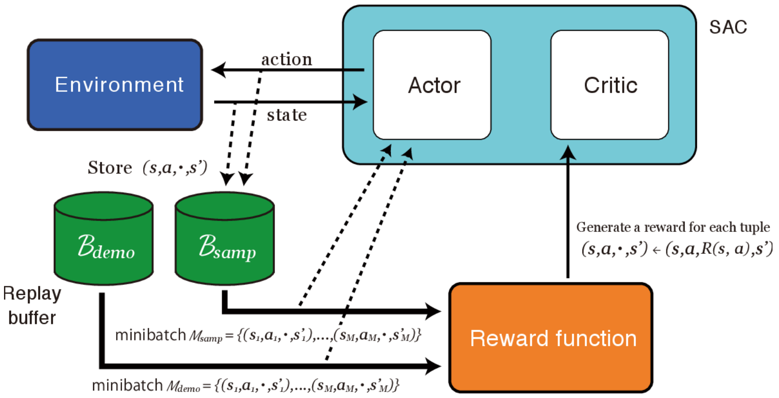

SAC
Environment
action
state
Actor
Critic
Store $(s,a,\cdot,s')$
$\mathcal{B}_{demo}$
$\mathcal{B}_{samp}$
Replay buffer
Generate a reward for each tuple
$(s,a,\cdot,s') \leftarrow (s,a,R(s,a),s')$
Reward function
minibatch $M_{samp} = \{(s_1,a_1,\cdot,s'_1),\ldots,(s_M,a_M,\cdot,s'_M)\}$
minibatch $M_{demo} = \{(s_1,a_1,\cdot,s'_1),\ldots,(s_M,a_M,\cdot,s'_M)\}$

Figure 2.7: Schematic Diagram of Soft Actor-Critic extensive Reward Function

information. In simpler terms, entropy measures the randomness of a random variable. The formula for calculating the entropy (H) of a random variable x with probability mass or density function P is given by eq. (2.8):

$$H(P) = ExP[logP(x)] \tag{2.8}$$

This equation captures the essence of entropy as a measure of information content, emphasizing that low-probability events convey more information than high-probability ones.

In the context of reinforcement learning (RL), entropy regularization modifies the standard RL objective by incorporating entropy[7]. The agent receives a bonus reward at each time step proportional to the entropy of the policy at that timestep. In an infinite-horizon discounted setting, the augmented RL problem is expressed in eq. (2.9), where (¿ 0) serves as the temperature parameter determining the relative importance of the entropy term, thus controlling the stochasticity of the optimal policy:

$$\pi^* = \arg\max_{\pi} \sum_{t} \mathbb{E}_{\theta}[\gamma^t R(s,a,s') + \alpha H(\pi(\cdot|s))] \tag{2.9}$$

Notably, as approaches 0, the standard maximum expected return is recovered.

From eq. (2.9) It is possible to derive a state-value function (V(s)) and action action-value function can be derived by eq (2.10) and eq(2.11)

$$V_\phi(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[Q_\phi(s,a) - \alpha \log(\pi_\theta(a|s))], \tag{2.10}$$

$$Q_\phi(s,a) = \mathbb{E}_{\epsilon\sim\mathcal{N}}[Q_\phi^{\text{main}}(s,a+\epsilon) - \alpha \log(\pi_\theta(a|s))] \tag{2.11}$$

where ϕ and θ are the parameters of the Q-function and policy, respectively.

These equations reveal the interrelation between state-value and action-value functions and introduce the Bellman equation (eq. 2.12):

$$\begin{aligned} Q_\phi(s,a) = \mathbb{E}_{r,s'\sim\mathcal{E}}\big[r + \gamma\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}\big[&Q_\phi(s',a') \\ &- \alpha\log(\pi_\theta(a'|s'))\big]\big], \end{aligned} \tag{2.12}$$

Entropy regularization thus extends the RL framework by incorporating the trade-off between expected return and policy entropy, offering a principled approach to balancing exploration and exploitation. This approach aligns with the overarching goal of reinforcement learning to discover optimal policies in uncertain environments

Learning Q-Functions with Target Soft Q-Network

n the Soft Actor-Critic (SAC) algorithm, Q-functions are learned through Mean Squared Bellman Error (MSBE) minimization. The update involves a target value network, utilizing the Bellman backups computed using Equation (2.12). The state-value function is implicitly parameterized by eq 2.10, and the learning process incorporates a target soft Q-function with parameters i i . These parameters are determined as an exponentially moving average of the soft Q-function parameters. The optimization involves stochastic gradients.

Learning the Policy through KL-Divergence Minimization

SAC adopts a policy learning approach derived from soft policy iteration. The policy learning process minimizes the expected KL-divergence, as demonstrated in Equation (2.13):

$$J_\pi(\theta) = \mathbb{E}_{s,t\sim D}\left[\mathbb{E}_{a_t\sim\pi}[\alpha \log \pi_\theta(a_t|s_t) - Q_\phi(s_t, a_t)]\right] \tag{2.13}$$

To optimize J(), a common strategy involves using the reparameterization trick. The policy is reparametrized using a neural network transformation, as expressed in Equation (2.14),

$$a_t = f_\theta(\hat{O}_t; s_t) \tag{2.14}$$

This allows rewriting into an expectation over noise, as shown in Equation (2.15):

$$J_\pi(\theta) = \mathbb{E}_{s,t\sim D,\hat{O}_t\sim\mathcal{N}}[\alpha \log \pi_\theta(f_\theta(\hat{O}_t; s_t)|s_t) - Q_\phi(s_t, f_\theta(\hat{O}_t; s_t))] \tag{2.15}$$

Exploration vs. Exploitation with Entropy Regularization

The SAC algorithm incorporates entropy regularization to train a stochastic policy, controlled by the entropy regularization coefficient . This parameter explicitly governs the exploration-exploitation trade-off, with higher values encouraging more exploration and lower values promoting exploitation. Selecting the optimal is a crucial task that requires careful tuning for stable and high-reward learning

The gradients for entropy regularization are computed using Equation (2.16)

$$J(\alpha) = \mathbb{E}_{a_t\sim\pi}[-\alpha \log \pi(a_t|s_t) - \alpha\bar{H}] \tag{2.16}$$

Chapter 3

Tools and Frameworks

This chapter delves into the pivotal tools and frameworks employed in the development of the thesis project, aimed at advancing the field of reinforcement learning. The journey begins with the OpenAI Gym toolkit[8], a fundamental component for developing and comparing reinforcement learning algorithms. OpenAI Gym provides a standardized environment for testing and benchmarking various algorithms, allowing for a systematic evaluation of performance across different scenarios. Its role in the thesis project is critical, as it sets the stage for experimentation and analysis, providing a consistent platform to measure the efficacy of the developed algorithms.

Moving forward, the exploration delves into the PyTorch framework, a powerful deep learning library that has gained immense popularity in both research and industry. PyTorch's dynamic computational graph and intuitive interface make it an ideal choice for implementing complex neural network architectures, a necessity when dealing with reinforcement learning tasks. The chapter highlights the significance of PyTorch in enabling the seamless integration of neural networks into the project, fostering the development of sophisticated models that can learn and adapt in dynamic environments.

The narrative then extends to TensorFlow, another prominent deep learning framework that has played a pivotal role in shaping the landscape of artificial intelligence. TensorFlow's strengths lie in its scalability and flexibility, making it suitable for a wide range of applications. In the context of the thesis project, TensorFlow complements PyTorch by providing an alternative framework for experimentation and comparison. The chapter sheds light on the unique features of TensorFlow that contribute to the project's overarching goals.

As the exploration reaches its zenith, attention is directed towards PyTorch Networks, a specialized extension of PyTorch designed for reinforcement learning tasks. This framework goes beyond the standard capabilities of PyTorch, offering tailored functionalities that cater specifically to the nuances of reinforcement learning algorithms. The chapter underscores the importance of PyTorch Networks in fine-tuning models and optimizing their performance within the context of reinforcement learning challenges.

To gauge the effectiveness and efficiency of the developed algorithms, the Never2 Tool emerges as a key element in the concluding sections of the chapter. This measurement tool provides quantitative insights into the performance metrics of the reinforcement learning models, allowing for a meticulous evaluation of their learning capabilities and adaptability. The chapter culminates with a comprehensive discussion on the insights derived from the Never2 Tool, providing a holistic view of the impact and contributions made by the implemented tools and frameworks in advancing the field of reinforcement learning.

In essence, this chapter serves as a roadmap through the intricacies of the selected tools and frameworks, highlighting their individual significance and collective synergy in shaping the trajectory of the thesis project. Each component contributes to the overarching goal of advancing reinforcement learning, laying the groundwork for innovative solutions and pushing the boundaries of what is achievable in this dynamic and evolving field

3.1 Open AI Gym ToolKit

OpenAI Gym, introduced in 2016 during its public beta, has evolved into one of the most influential toolkits and frameworks in the realm of reinforcement learning. This discussion explores the motivations behind the creation of OpenAI Gym, delving into the challenges within the reinforcement learning landscape and how the framework addresses them.

Reinforcement Learning Landscape and Challenges:

Reinforcement learning, a subset of machine learning, is dedicated to the study of decision-making and motor control. Researchers aim to understand how an agent can learn and improve to achieve specific goals in complex, often unknown environments. Its broad applicability, ranging from robotics to business decisions and financial trading, has made reinforcement learning an attractive area for both academia and industry.

However, the progress in reinforcement learning faced hurdles, primarily due to the absence of robust benchmarks. Unlike supervised learning, which flourished with datasets like ImageNet, reinforcement learning lacked equivalent standardized benchmarks. Additionally, the lack of standardization in the design of environments presented a challenge. Minor differences in problem definitions, reward functions, or action spaces could significantly impact the difficulty of tasks, impeding reproducibility and hindering the comparison of results across different studies.

Motivations for OpenAI Gym: The need to address these challenges was the driving force behind the development of OpenAI Gym. The framework aimed to provide a solution to the dearth of benchmarks and the lack of standardization in reinforcement learning experiments. It envisioned a platform that would serve as a standardized interface for environments, allowing researchers and developers to focus on the core of reinforcement learning—agent design—without being constrained by predefined interfaces.

3.1.1 Environments of open AI Gym

In reinforcement learning, the key components are the agent and the environment. OpenAI Gym made a strategic choice to emphasize the abstraction of environments rather than agents. Instead of imposing pre-defined agent interfaces, the framework provides a standard environment interface. This decision empowers developers to design agents independently, fostering creativity and innovation in the core aspects of reinforcement learning.

The significance of this approach lies in the versatility it affords. Agents implemented with OpenAI Gym can seamlessly interact with any environment within the framework, promoting adaptability and ease of experimentation. Developers can tailor environments to suit specific experiments, enabling personalized testing scenarios that cater to the unique requirements of diverse research endeavors.

OpenAI Gym encompasses various environments categorized into distinct types:

Algorithms: This category focuses on tasks involving the imitation of computations, such as copying or reversing symbols from an input tape. Task difficulty can be adjusted by varying the sequence length.

Atari: Emulating the Atari 2600 video games, this section of environments relies on the Arcade Learning Environment (ALE). It provides over 100 environments offering observations in the form of raw pixel images or RAM.

Box2D: Tasks in this group involve continuous control in a simple 2D simulator, featuring challenges like BipedalWalker, CarRacing, and LunarLander.

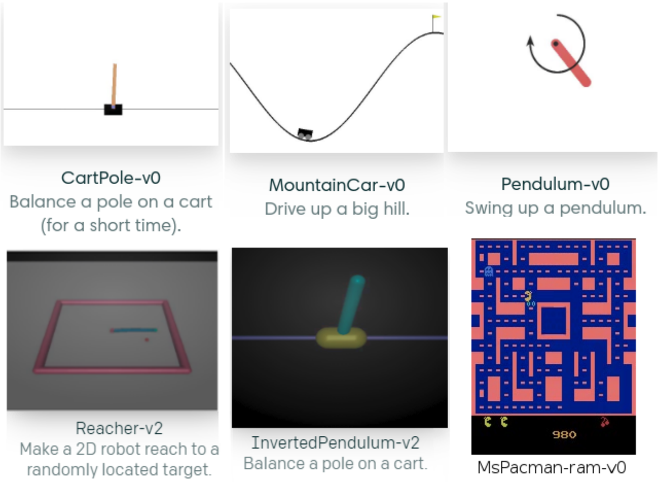


Figure 3.1: open AI gym Environments

Classic Control: Derived from control theory, this class includes problems widely used in classic reinforcement learning literature. Examples include balancing a pole on a cart or swinging up a pendulum.

MuJoCo: This collection introduces continuous control tasks within a fast physics 3D simulator, known as MuJoCo. It serves as a valuable resource for research and development in robotics, biomechanics, graphics, and animation.

Robotics: OpenAI Gym's Robotics category presents eight environments with more complex manipulation tasks than MuJoCo. Notable examples include Fetch, a robotic arm for object manipulation, and ShadowHand, a robotic hand for intricate object manipulation.

These environments collectively offer a rich and diverse set of challenges for reinforcement learning algorithms, spanning various domains and complexities

Interface functions

Exploring OpenAI Gym, it is essential to focus on the most crucial interface func- tions that the agent will exploit to interact with the environment. The functions which constitute the skeleton of an OpenAI Gym environment are the following:

- **def step(self, action)**: through this function, the agent can communi- cate the action it wants to take. The input data depends on the type and number of variables in the actions space (e.g. discrete or continuous). As will be discussed in section 3.1 on the following page, the values returned by this function represent the environment state after the manipulation caused by the agent action. Thanks to these data, the agent will be able to select the next action following the reinforcement learning loop.

- **def reset(self)**: during the episode, internal variables of the environment changes, influenced by the action taken previously. This function allows the agent to restart the initial situation of the environment. This procedure is particularly helpful when an episode finishes and the agent has to restart the next learning episode in a brand new copy of the environment.

- **def render**(self, mode='human', close=False): this function is mainly used in simulated environments. It enables the visual render (if available) of the environment.

- **def close(self)**: the final function to close the environment after the end of all experiments and episodes.

3.2 Pytorch

PyTorch, an open-source machine learning and deep learning library developed by Facebook's AI Research Lab, was released to the public in October 2016. It aims to provide an intuitive and straightforward framework for artificial intelligence projects, with a particular focus on computer vision and natural language processing.

Developed using Python, C++, and CUDA, PyTorch leverages CUDA-enabled GPUs for general-purpose processing. While the primary interface is in Python,

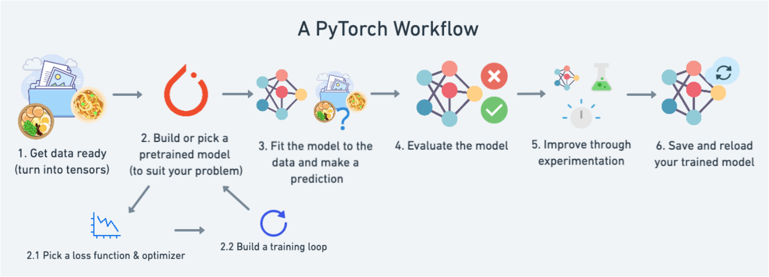

A PyTorch Workflow
1. Get data ready (turn into tensors)
2. Build or pick a pretrained model (to suit your problem)
3. Fit the model to the data and make a prediction
4. Evaluate the model
5. Improve through experimentation
6. Save and reload your trained model
2.1 Pick a loss function & optimizer
2.2 Build a training loop

Figure 3.2: Pytorch workflow

there is also a C++ interface, showcasing the versatility of the library. The components of PyTorch include:

torch: A tensor library with robust GPU support, implementing interfaces similar to NumPy. It includes data structures for multi-dimensional tensors and mathematical operations, offering utilities for efficient serialization of tensors and arbitrary types.

torch.autograd: A tape-based automatic differentiation library supporting every differentiable operation on tensors available in torch.

torch.jit: A compilation stack that uses TorchScript to create serializable and optimizable models from PyTorch code. This allows training in PyTorch using Python and exporting the model to production environments where Python might be less advantageous for performance reasons.

torch.nn: A neural networks library compatible with autograd and designed for flexibility. torch.multiprocessing: Based on the Python multiprocessing library, it implements memory sharing of torch tensors across processes.

torch.utils: Contains utility functions to better exploit the features of PyTorch.

PyTorch provides a NumPy-like experience for interacting and manipulating data structures suitable for GPU computation, known as Tensors. Tensors can be used on both CPU and GPU, accelerating computations with functions explicitly designed for scientific computation needs.

Unlike frameworks that are primarily complex C++ bindings, PyTorch prioritizes Python, providing a natural user experience. The design emphasizes intuitiveness and linearity, making PyTorch synchronous for improved debugging experiences. Developers aimed to create a product that is easy to use, and this intention is reflected in the library's design.

One distinctive feature of PyTorch is its tape-based automatic differentiation, offering a single way to build neural networks. While other frameworks like TensorFlow or Theano utilize a static approach in graph creation, PyTorch employs Tape-Based Autograd. This approach, based on reverse-mode automatic differentiation, allows users to change the network structure dynamically without lag or overhead. It relies on the properties of the chain rule, making it possible to calculate derivatives efficiently.

PyTorch's commitment to providing a user-friendly, flexible, and efficient deep learning platform has made it a popular choice in the machine learning community. Its seamless integration with Python and emphasis on dynamic computation graph creation set it apart from other frameworks, contributing to its widespread adoption in both research and industrial applications

3.3 Tensorflow

TensorFlow is a powerful open-source machine learning library extensively used in reinforcement learning (RL), a paradigm where an agent learns by interacting with an environment to maximize cumulative rewards. TensorFlow's flexibility and efficiency make it well-suited for building and training neural networks, a key component in many RL algorithms.

In RL, TensorFlow is employed to define and train agents, which are typically

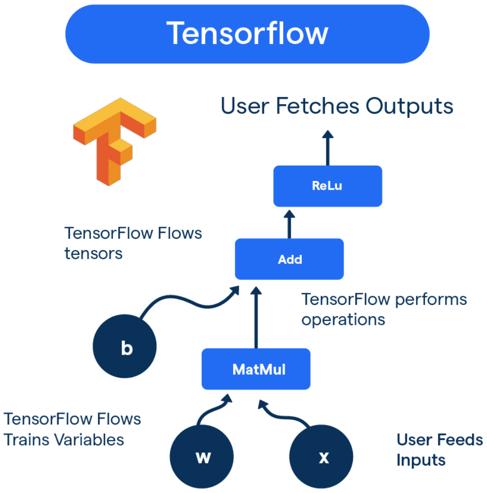


Figure 3.3: Tensor flow Key features

neural network models. Using TensorFlow's high-level API, Keras, users can construct and optimize complex neural networks that represent an agent's policy or value function. The policy is a strategy guiding the agent's decisions based on observed states, and the value function estimates the expected cumulative reward for a given state-action pair. Training RL agents involves iterative interactions with the environment. TensorFlow facilitates this process by providing optimization algorithms and tools for efficient neural network training. Experience replay, a technique enhancing training stability, is also supported by TensorFlow.

Integration with OpenAI Gym, a popular RL toolkit, is seamless. TensorFlow users can leverage OpenAI Gym environments to simulate and test various RL algorithms. TensorBoard, a visualization tool integrated with TensorFlow, assists in monitoring training progress, evaluating performance, and diagnosing issues in real time.

A common RL example using TensorFlow is the implementation of a deep Q-network (DQN). TensorFlow's model-building capabilities, optimization algorithms, and gradient computation simplify the DQN training process. In a typical

DQN implementation, a neural network approximates the Q-values, representing the expected cumulative rewards for state-action pairs. Training involves adjusting the network's parameters to minimize the difference between predicted and target Q-values.

TensorFlow's model-saving functionality enables users to store trained models, facilitating deployment for decision-making in real or simulated environments. This feature is crucial for practical applications where RL models transition from training to serving stages.

3.3.1 Comparision between Tensorflow and Pytorch

n the post-deep learning era, the development of neural network architectures and frameworks has become a focal point for many companies. Two prominent frameworks, TensorFlow by Google and PyTorch by Facebook , have emerged as leaders in this field. Despite serving the same purpose, implementing a neural network in these frameworks can yield different results due to the inherent distinctions in their training processes and underlying technologies.

One significant difference lies in the construction of computational graphs, an abstraction representing the computation process. TensorFlow adopts a static approach, defining computational graphs before code execution, allowing for parallelism and driving scheduling. This method communicates with the external world via tensors, which are later substituted by input data during runtime. In contrast, PyTorch employs a dynamic computational graph, constructed incrementally at runtime without placeholders. This flexibility supports on-the-fly changes to the computational graph, making PyTorch more adaptable and Pythonic.

Distributed training and data parallelism are crucial features, with PyTorch offering native support for asynchronous execution from Python, potentially improving performance. TensorFlow, while also supporting these capabilities, requires more developer effort to fine-tune computations for specific devices. Although both frameworks provide opportunities for distributed training, PyTorch requires less effort for seamless integration.

Visualization tools play a vital role in machine learning research, and here TensorFlow leverages TensorBoard, offering extensive features for visualizing and tracking the training process. PyTorch, on the other hand, relies on Visdom, developed by Facebook researchers, which provides minimalistic features compared to TensorBoard. However, TensorBoard can be used with PyTorch through the TensorBoardX library, bridging the gap in visualization capabilities.

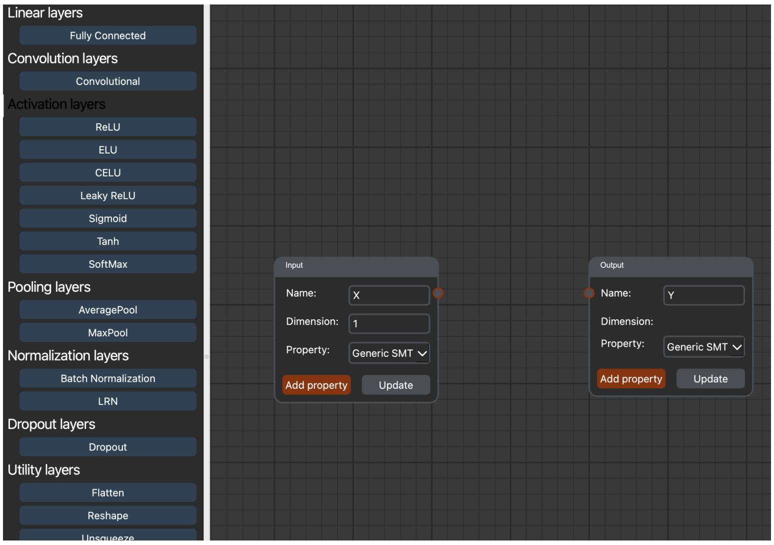
In terms of production deployment, TensorFlow shines with TensorFlow Serving, a framework that facilitates REST Client API usage for deploying trained models. PyTorch has made strides in deployment, but it currently lacks a dedicated framework for web deployment. Developers must resort to Flask or Django as backend servers to create the necessary environment for model exploitation

3.4 Never2 Tool

NeVer2, a powerful Python tool at the forefront of neural network development, stands as a testament to innovation and versatility. With its robust foundation and seamless integration with the pyNeVer API, this tool has been meticulously crafted to address the intricate challenges of learning and verifying neural networks. NeVer2 caters to a broad spectrum of users, from seasoned researchers to practitioners, offering a comprehensive solution for tasks ranging from training and testing to the validation of neural networks.

Key Features:

PyNeVer API Integration:



Input

Name: X

Dimension: 1

Property: Generic SMT

Output

Name: Y

Dimension:

Property: Generic SMT

Figure 3.4: Sample Diagram of Never2 Tool

NeVer2 leverages the pyNeVer[9] API, which serves as the backbone for its functionality. This integration allows users to harness the capabilities of pyNeVer seamlessly within a Python environment, fostering ease of use and extensibility.

Neural Network Learning:

NeVer2 facilitates the training of neural networks through a streamlined process. Users can define and configure neural network architectures using popular frameworks like TensorFlow or PyTorch. The tool supports dynamic adjustments to computational graphs, reflecting the flexibility and adaptability demanded by diverse research projects.

Verification Capabilities: One of NeVer2's standout features is its emphasis on neural network verification. The tool incorporates advanced techniques to assess the reliability and correctness of trained models. This includes methods for robustness testing, adversarial example analysis, and formal verification approaches. Researchers can employ NeVer2 to gain insights into a model's vulnerability to perturbations and potential adversarial attacks.

Chapter 4

Architectural Design of Never2 Tool(CoCoNet) API

NeVer2 stands as a versatile tool, combining a graphical user interface (GUI) with a command-line interface (CLI) to facilitate the seamless creation, training, and validation of neural networks. This innovative platform streamlines the complex process of managing neural networks within a unified environment, providing users with an integrated solution for their machine-learning endeavors. Moreover,[10] NeVer2 extends its functionality to encompass the verification of Very Deep Neural Network Library (VNN-LIB) properties on Open Neural Network Exchange (ONNX) models, offering a command-line interface for users who prefer a more streamlined and efficient approach.

One of NeVer2's distinguishing features lies in its commitment to user-friendly implementation. The tool is designed to be accessible to users at varying levels of expertise, ensuring that the complexities of deep learning are made more manageable. This commitment is exemplified through the incorporation of the pyNeVer API, providing a solid and extensible framework that underpins the tool's seamless functionality.

NeVer2's proficiency extends across the entire lifecycle of neural network development, emphasizing a holistic approach to model refinement. Its support for dynamic computational graphs, reminiscent of the PyTorch philosophy, is a key feature that sets it apart. This dynamic nature empowers users to make real-time adjustments to the computational graph, offering unparalleled flexibility during runtime. For projects requiring on-the-fly modifications to the neural network architecture, NeVer2 emerges as a flexible and adaptive solution.

The tool's prowess is further showcased in its adept handling of distributed training, a critical component of large-scale deep learning endeavors. NeVer2 leverages PyTorch's capabilities to streamline the complexities associated with asynchronous execution from Python. This capability not only enhances performance but also ensures scalability in scenarios where significant computational resources are required. The seamless integration of distributed training capabilities underscores NeVer2's commitment to providing users with a comprehensive and efficient neural network development environment.

In the realm of neural network verification, NeVer2 excels by providing a command-line interface (CLI) for validating Very Deep Neural Network Library (VNN-LIB) properties on Open Neural Network Exchange (ONNX) models. This versatile CLI allows users to execute verification procedures outlined in Satisfiability Modulo Theories (SMT) files on specified neural networks in the ONNX format. This dual-interface approach caters to diverse user preferences, allowing for both graphical and command-line interactions, depending on the user's workflow and requirements.

The heart of NeVer2's verification capabilities lies in its diverse range of strategies. Users can opt for the 'complete' strategy, leveraging the exact algorithm suitable for small-sized networks. Alternatively, the 'approximate' strategy employs an over-approximate algorithm, balancing accuracy and computational efficiency. The ’mixed1’ and ’mixed2’ strategies introduce a nuanced approach, refining one or two neurons per layer, respectively. This adaptability ensures that users can tailor the verification process to the specific characteristics and requirements of their neural network, striking an optimal balance between precision and computational resources.

NeVer2 stands as a comprehensive and adaptable tool that seamlessly integrates into the Python ecosystem. Its user-friendly design, support for dynamic computational graphs, and adept handling of distributed training make it a valuable asset for both researchers and practitioners in the field of deep learning. Whether through its graphical user interface or command-line capabilities, NeVer2 caters to diverse user preferences, ensuring a flexible and efficient neural network development environment. As the landscape of machine learning continues to evolve, NeVer2 remains at the forefront, empowering users with the tools they need to navigate the complexities of neural network development with confidence and ease.

NeVer2 emerges as a robust and flexible tool that addresses various facets of neural network development and verification. Its integration with the pyNeVer API, dynamic computational graph support, and emphasis on verification techniques position it as a valuable asset for researchers and practitioners. The tool’s commitment to user-friendly implementation, distributed training support, and visualization capabilities contributes to its versatility

4.1 Installation

The installation process for NeVer2 involves setting up the required packages and dependencies to enable seamless operation of the neural network tool. NeVer2 relies on two main components: the pyNeVer API and the PyQt6 framework.

These can be easily installed using the Python package manager, PIP.

To initiate the installation, execute the following command in your terminal or command prompt:

```
pip install pynever PyQt6
```

This command fetches and installs the pyNeVer API and PyQt6 framework, ensuring that NeVer2 has the essential components to function properly.Once the packages are successfully installed, NeVer2 can be launched from the root directory using the following command:

```
python NeVer2/never2.py
```

This command activates NeVer2, allowing users to access the graphical user interface (GUI) and leverage its capabilities for neural network development, training, and verification.

ARM BASED MAC OS

For users on ARM-based Mac OS, additional considerations come into play due to compatibility issues with the default Python distribution for ARM platforms. To address this, it is recommended to install miniforge for arm64 (Apple Silicon), a distribution that is compatible with the architecture.Creating a Python virtual environment is the next step in the ARM-based Mac OS installation process. This involves using Conda, a package manager, to set up a specific environment named 'myenv' with Python version 3.9.5:

```
conda create -n myenv python=3.9.5
conda activate myenv
```

This ensures that the subsequent installations are specific to the newly created environment. The next step involves installing additional dependencies for TensorFlow using Conda:

```
conda install -c apple tensorflow-deps
pip install tensorflow-macos tensorflow-metal
pip install pynever PyQt6
```

Now, NeVer2 is ready to be executed on an ARM-based Mac OS with the following command

```
python NeVer2/never2.py
```

It's important to note that each time NeVer2 is to be run, the Conda environment must be activated using:

```
conda activate myenv
```

This ensures that NeVer2 operates within the specific environment with the correct dependencies, providing a smooth and consistent experience for users on ARM-based Mac OS. In summary, the installation process for NeVer2 involves acquiring the necessary packages, addressing compatibility concerns on specific platforms, and setting up an isolated environment to ensure a reliable and efficient operation of the neural network tool

4.2 software Architecture

The Never2 Tool design is depicted in Figure 1 using a UML class diagram, focusing on its fundamental structure rather than a comprehensive software architecture overview. The internal representation of neural networks is managed by pyNeVer, a [11]Python API offering learning and verification capabilities. The core component is the Project class, encapsulating functionalities for network design, interaction with pyNeVer, and procedures for opening and saving neural network files.

For the graphical interface, PyQt's Graphics View framework is employed. This framework facilitates the creation and interaction with 2D graphical items, supporting zooming and rotation. The event propagation architecture and Binary Space Partitioning (BSP) tree enable real-time visualization of large scenes. The GraphicsScene class is used for creating or destroying objects and setting global parameters, while the Scene class serves as a container for all application objects.

Concrete instances of the abstract class Block are displayed in the scene. The LayerBlock represents network layers, the FunctionalBlock defines input and output, and the PropertyBlock represents VNN-LIB properties. The Block classes are designed to support multiple inputs and outputs, allowing future extensions to accommodate architectures beyond feed-forward neural networks, such as ResNets and recurrent neural networks.

Leveraging pyNeVer, the tool enables direct import of neural network models in ONNX or PyTorch formats for visualization, property addition, or conversion. The generic interface of pyNeVer facilitates the creation of custom model conversion for various file formats, expanding the tool's capabilities and supporting additional benchmarks. The flexibility and extensibility of CoCoNet are highlighted, emphasizing its adaptability to evolving neural network architectures and file formats

4.3 Procedure

In NeVer2, building a neural network for the ACC application involves creating layers and defining input/output using the LayerBlock and Functional Block classes. The network is exported to PyTorch for training. Upon completion, it's re-imported into CoCoNet, demonstrating the tool's bidirectional capability. This allows users to seamlessly transition between NeVer2 and PyTorch for design

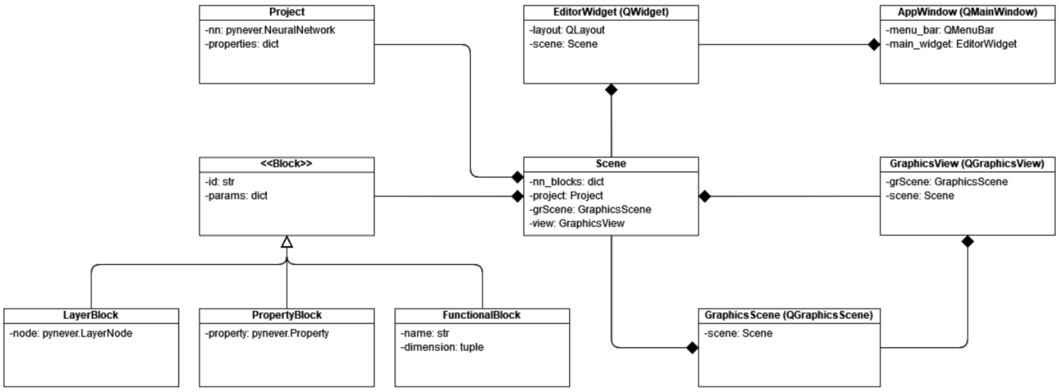
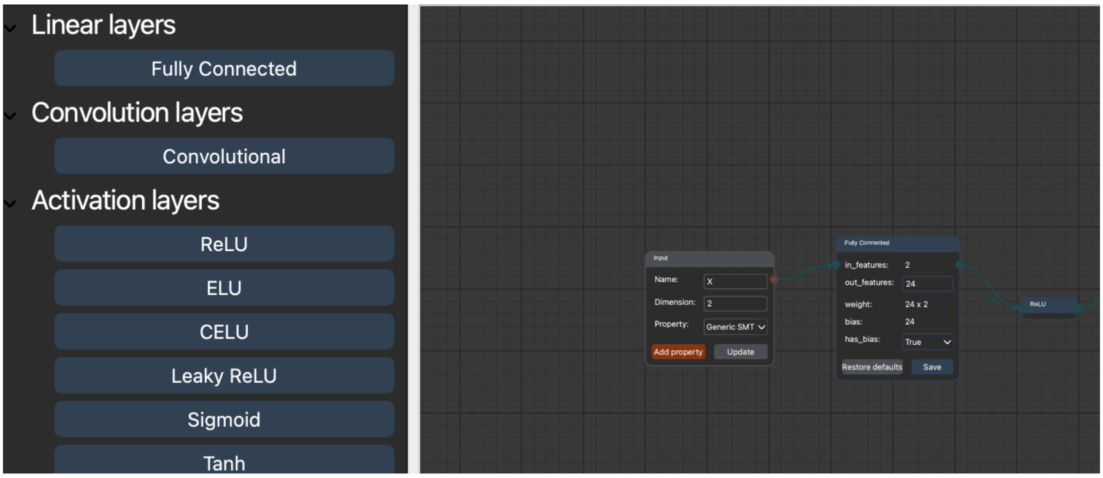


Figure 4.1: UML Diagram of Never2 Tool

and training, reinforcing the tool's versatility in supporting diverse workflows.

4.3.1 Building the model

The initial screen of Never2 showcases two Functional Blocks, enabling users to define network input and corresponding labels. Sequentially, the first fully connected layer with 24 neurons, followed by a ReLU activation, is defined within the tool. To maintain order, each layer is added and updated individually, incorporating the input block to specify dimensions. The Save button on each block facilitates parameter updates, ensuring configuration accuracy. For user convenience, the Restore defaults option resets values to default settings without overwriting, enhancing flexibility in experimentation. This side-by-side interface design streamlines the process of constructing neural networks within NeVer2, providing a user-friendly experience. The automatic sequential addition of layers aligns with default settings, allowing users to progressively build and refine their network architecture. This step-by-step approach, coupled with intuitive controls, exemplifies NeVer2 commitment to simplicity and precision in network design for applications like the ACC scenario.



Name: X
Dimension: 2
Property: Generic SMT

in_features: 2
out_features: 24
weight: 24 x 2
bias: 24
has_bias: True

Figure 4.2: Interface of Never2 Tool after adding defining Properties

Activation layers(ReLU)

Rectified Linear Unit, or ReLU, is a widely used activation function in artificial neural networks, playing a crucial role in introducing non-linearity to the model's decision-making process. It operates on an input x, producing an output f(x) defined as the maximum of zero and the input itself, expressed as f(x) = max(0, x).

One of the primary reasons for ReLU's[12] popularity is its simplicity and efficiency. By allowing positive inputs to pass through unchanged and setting negative inputs to zero, ReLU aids in the model's ability to learn complex patterns in the data. This simplicity also contributes to faster convergence during the training process, as the linearity of the function eases gradient computation.

ReLU effectively addresses the vanishing gradient problem encountered in traditional activation functions like sigmoid or hyperbolic tangent. During backpropagation, gradients can diminish as they are propagated through layers, hindering the learning process. ReLU's derivative is either zero or one, preventing the vanishing gradient issue and facilitating more effective learning in deep networks.

However, ReLU is not without challenges. One notable issue is the "dying

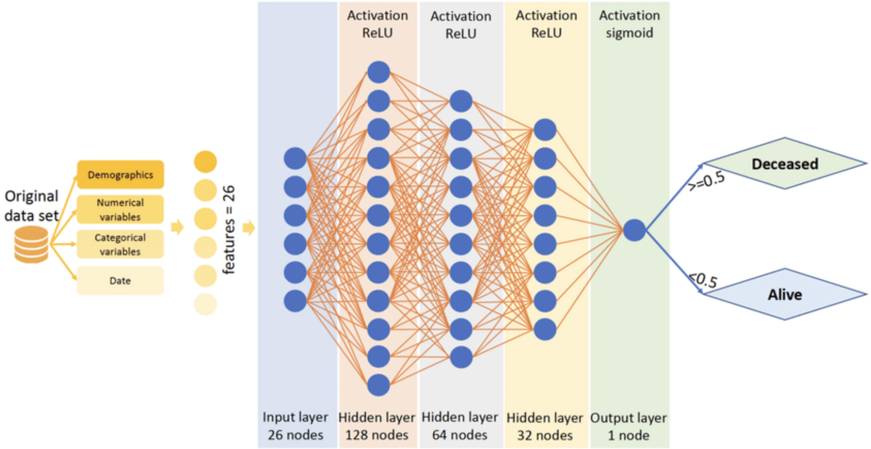


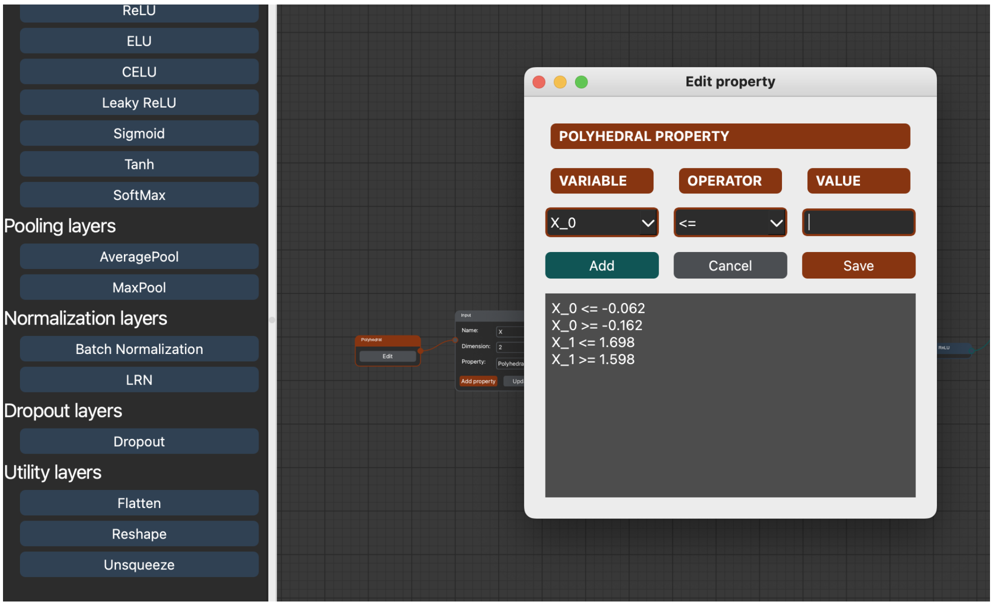
Figure 4.3: ReLU Neural network architecture

ReLU" problem, where neurons can become inactive during training and cease to update their weights. If a large gradient flows through a ReLU unit, it can cause the weights to update in such a way that the unit will always output zero. This can limit the model's capacity to learn and adapt.

To address the dying ReLU problem, variations like Leaky ReLU have been introduced. Leaky ReLU allows a small, non-zero gradient when the input is negative, ensuring that even neurons with negative inputs can contribute to the learning process.ReLU is a foundational activation function in deep learning, offering simplicity, efficient computation, and mitigation of the vanishing gradient problem. While it has challenges like the dying ReLU problem, researchers have introduced variations to enhance its performance and adaptability in training deep neural networks.

4.3.2 Defining the Property

In the finalization phase of configuring a neural network, a pivotal step involves defining VNN-LIB properties related to both input and output. Drawing guidance from the description provided in [8], the emphasis is on bounding input



X_0

<=

X_0 <= -0.062
X_0 >= -0.162
X_1 <= 1.698
X_1 >= 1.598

Figure 4.4: sample Diagram of Defining Polyhedral Property

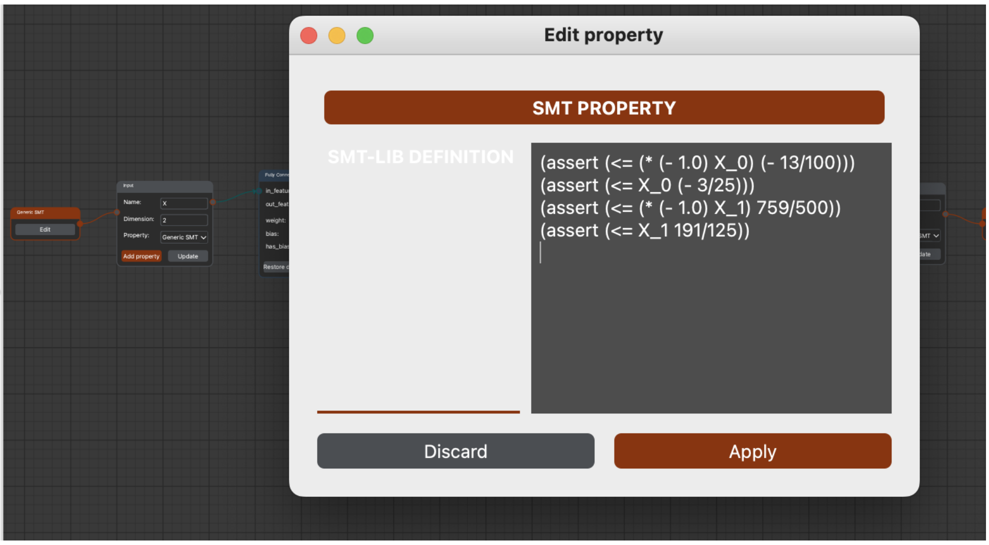
variables. Within this context, the OutBounds description recommends two distinctive approaches: Polyhedral Properties and Generic SMT Properties.

Polyhedral Property:

provides an insightful look into the interface where a Polyhedral Property is defined. This approach leverages the property selector in the input block, offering a controlled and efficient environment for bounding variables without the need for manually crafting SMT (Satisfiability Modulo Theory) statements. The Polyhedral Property method streamlines the process, allowing users to set constraints on input variables within a well-defined and controlled framework. As depicted in the figure, this property configuration aligns with the values outlined in Section 3.2, ensuring a harmonious integration of input bounding mechanisms.

SMT Property:

Conversely, the Generic SMT Properties approach involves the direct crafting of SMT expressions. While providing more flexibility, this method demands a deeper understanding of SMT syntax. Unlike the automated nature of Polyhe-



```
(assert (<= (* (- 1.0) X_0) (- 13/100)))
(assert (<= X_0 (- 3/25)))
(assert (<= (* (- 1.0) X_1) 759/500))
(assert (<= X_1 191/125))
```

Figure 4.5: Sample Diagram of Defining smt property

dral Properties, SMT Properties require explicit SMT statements for bounding variables. This approach allows for a more intricate and customized specification of constraints, affording users finer control over the bounding configurations.

In the Polyhedral Property method, the emphasis is on a user-friendly interface, minimizing the need for users to delve into the complexities of SMT expressions directly. Instead, the property selector in the input block becomes the conduit for bounding variable constraints. This approach simplifies the user experience while maintaining robust bounding mechanisms for input variables.

4.3.3 Load and Save Models

NeVer2 simplifies the integration of neural networks into its framework by supporting direct loading of models in ONNX and PyTorch formats. It also streamlines the linking of properties to networks when they share identical input and output identifiers, enhancing its utility for VNNCOMP benchmark creation.To generate a VNNCOMP-compatible benchmark using NeVer2, users import a neural network model in ONNX or PyTorch format, ensuring compatibility. Subsequently, users link a property to the imported network, verifying alignment in input and output identifiers.

Once the neural network and its corresponding property are integrated in NeVer2, users create benchmark files easily. In the menu, selecting ”Save as...” with the VNN-LIB entry results in two distinct files: a .onnx file representing the neural network model and a .smt2 file encapsulating the defined property. This structured process streamlines benchmark creation, ensuring compatibility with VNNCOMP standards. NeVer2, with its capacity to seamlessly link properties to networks and export in standardized file formats, serves as a versatile tool for researchers participating in VNNCOMP or working within the VNN-LIB framework. Its commitment to simplifying the bench-marking process underscores NeVer2's significance in the realm of neural network verification.

4.3.4 Command-line interface

NeVer2 Tool offers command-line functionality through the options check model and -convert model, extending its capabilities beyond the graphical interface. These command-line features empower users to efficiently incorporate NeVer2's functionalities into automated workflows or scripts. The -check option facilitates the quick validation of an ONNX model's compliance with the VNN-LIB standard. Users can easily determine whether a given ONNX model adheres to the specifications outlined by VNN-LIB, streamlining the validation process through the command line.

Similarly, the -convert option enables the transformation of PyTorch models to the ONNX[13] format using NeVer2, contingent upon the compatibility of operators with the VNN-LIB standard. This command-line functionality ensures a seamless conversion process, provided the necessary operators are supported. These command-line options enhance NeVer2 Tool's versatility, allowing users to integrate its features into automated processes. Whether validating compliance or converting between model formats, the command-line interface provides a flexible means of leveraging NeVer2 Tool's capabilities for neural network analysis and verification in a programmatic fashion.

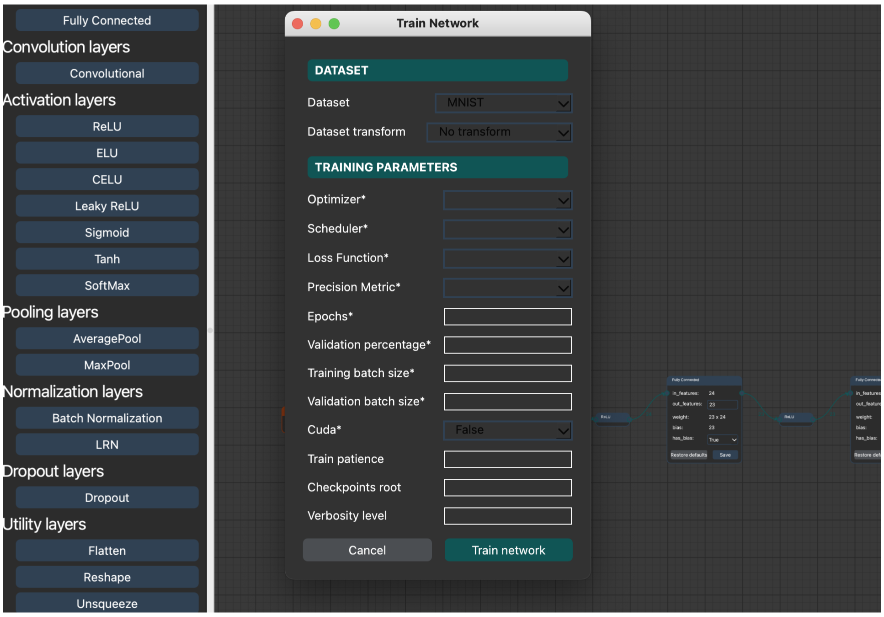
```
python pynever.py complete single -s /Users/surendrakumarreddypolaka
/Desktop/Thesis/NeVer2-main/dqn_network.onnx > /Users
/surendrakumarreddypolaka/Desktop/Thesis/NeVer2-main/output1.smt2
```

4.4 Training the Network

Upon completing the network construction, the next step involves training. Navigate to the menu bar and select "Learn..." -¿ "Train" to access the training window. In this interface, users can choose the dataset and configure various learning parameters.

NeVer2 provides default access to both MNIST and fMNIST datasets due to their widespread popularity. If these datasets are not already present, they will be downloaded and stored in the NeVer2 working directory upon the first selection. The inclusion of such widely used datasets ensures user convenience and accessibility.

For the training process, NeVer2 [14]offers a pre-defined dataset transform tailored for both convolutional and linear MNIST and fMNIST networks. This transformation consists of 2 or 3 steps: pilToTensor and Normalize(1, 0.5) are common to both cases, and an additional Flatten transform is included exclusively for the linear MNIST network. These transformations are crucial for preparing the data in a format suitable for training the neural network.



Train Network

DATASET

Dataset: MNIST

Dataset transform: No transform

TRAINING PARAMETERS

Optimizer*

Scheduler*

Loss Function*

Precision Metric*

Epochs*

Validation percentage*

Training batch size*

Validation batch size*

Cuda*: False

Train patience

Checkpoints root

Verbosity level

Figure 4.6: sample diagram of parameters how to train the Network

In summary, NeVer2 simplifies the training process by providing a user-friendly interface for dataset selection and learning parameter configuration. The availability of default datasets, along with pre-defined transformations optimized for various network architectures, enhances the efficiency and accessibility of the training workflow within the NeVer2 tool.For efficient training in NeVer2, users can fine-tune learning parameters through a user-friendly interface:

Optimizer: Select the "Adam" optimizer, an acclaimed gradient-based optimization algorithm known for its effectiveness. Users can further customize related parameters once the optimizer is chosen.

Learning Rate Scheduler: Currently, NeVer2 supports the "ReduceLROnPlateau" learning rate scheduler. This scheduler adjusts the learning rate when a plateau in model performance is detected, providing adaptability during training.

Loss Function: Choose between "Cross Entropy" and "MSE Loss" based on the neural network's structure. This selection influences the calculation of the error during training.

Precision Metric: Select either "Inaccuracy" or "MSE Loss" to define the precision metric, guiding the evaluation of the model's performance during training.

Epochs: Define the number of training epochs, representing the iterations through the entire dataset during training.

Validation Percentage: Set a value between 0 and 1 to indicate the percentage of the dataset allocated for validation purposes.

Training and Validation Batch Size: Define the dimensions of training and validation batches, influencing the granularity of data processed during each iteration.

CUDA: Enable this option to leverage NVidia GPU architecture for accelerated computation, optimizing training speed.

Train Patience (Optional): Specify the number of epochs with no loss decrease before triggering early stopping in the training process.

Checkpoints Root (Optional): Designate the directory for storing training strategy checkpoints. The default location is the NeVer2 working directory.

Verbosity Level (Optional): Set the frequency of log prints during training, controlling the level of detail displayed. The default is set to print logs after each training batch.

These configurable parameters empower users to tailor the training process to their specific needs, balancing customization and user-friendly design within the NeVer2 environment.But in this Thesis we mainly focus on Verification tool rather than training the Network.

4.5 Verification Strategy

NeVer2 is a versatile tool designed to simplify the neural network development process. It features a user-friendly GUI that streamlines building, training, and verifying neural networks within a unified environment. The graphical interface offers an intuitive platform for users to navigate the complexities of machine learning seamlessly.

In addition to the GUI, NeVer2 caters to advanced users by providing a Command-Line Interface (CLI). This allows for efficient verification of VNN-LIB properties on ONNX models without the need for the graphical interface. Users can execute commands like ”python NeVer2/never2.py” to verify specific properties specified in SMT files on ONNX networks. This CLI functionality enhances flexibility, enabling users to interact with NeVer2 in a more streamlined manner, especially when dealing with specific verification tasks.

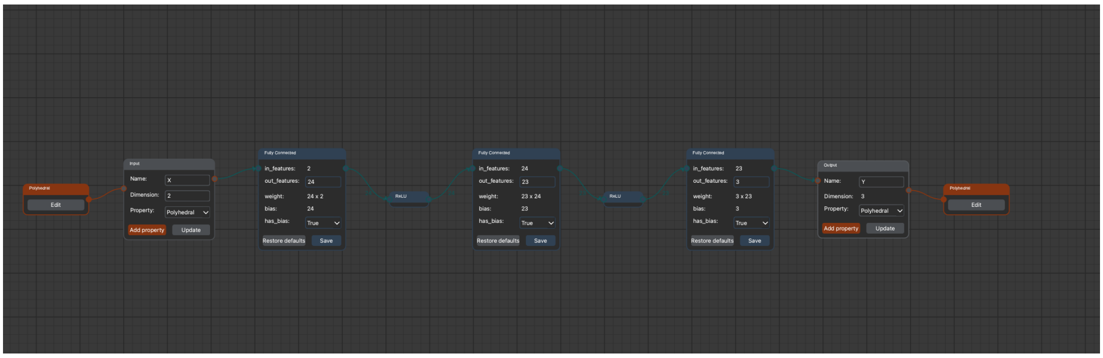
```
python NeVer2/never2.py -verify <property>.smt2
<network>.onnx [complete | approximate | mixed1 | mixed2]
```

verification procedure for the property specified in the SMT file on the network specified in the ONNX file. The verification strategy is one among the following:

- complete: uses the exact algorithm (for small-sized networks)
- approximate: uses the over-approximate algorithm
- mixed1: uses the mixed algorithm refining 1 neuron per layer
- mixed2: uses the mixed algorithm refining 2 neurons per layer

4.6 Output Visualization

In NeVer2, extending the neural network involves selecting corresponding blocks from the left toolbar. Adding ReLU activation functions is straightforward, re-



Polyhedral

Input
Name: X
Dimension: 2
Property: Polyhedral

Fully Connected
in_features: 2
out_features: 24
weight: 24 x 2
bias: 24
has_bias: True

ReLU

Fully Connected
in_features: 24
out_features: 23
weight: 23 x 24
bias: 23
has_bias: True

ReLU

Fully Connected
in_features: 23
out_features: 3
weight: 3 x 23
bias: 3
has_bias: True

Output
Name: Y
Dimension: 3
Property: Polyhedral

Polyhedral

Figure 4.7: Output Visualization to verify the Network

quiring no additional parameters. Fully Connected layers can be seamlessly incorporated directly. The presented example illustrates the construction of a complete network, showcasing the simplicity of layer addition within NeVer2.

Seamless Layer Addition and Output Visualization in Neural Network Construction. Explore the simplicity of integrating ReLU activations and Fully Connected layers. Witness the clear computation of the final output, enhancing user understanding. NeVer2 provides an intuitive interface for efficient and visual network design

In this example, particular emphasis is given to the final layer, which performs the computation yielding the network's output. The output is visually presented in the output block, offering users a comprehensive view of the entire network's architecture and its functional progression.[15] NeVer2's intuitive interface facilitates a user-friendly experience, allowing for a clear visualization of the neural network's structure. This streamlined process enhances users' ability to comprehend the intricacies of each layer, fostering a deeper understanding of how the network computes the final output. As a result, NeVer2 provides a powerful tool for both novice and experienced users to efficiently design and analyze neural networks within a visually accessible environment

Chapter 5

Experimental Results

The experimental results section delves into the application of classic control reinforcement learning algorithms on two widely recognized problems, Mountain Car and Pendulum. These problems serve as benchmarks to assess the performance of three distinct types of algorithms: Basic Q-learning, Q-learning with a neural network, and Q-learning with a neural network implemented using the PyTorch framework. The emphasis is on measuring the output of the system through various metrics and employing the Never2 Tool to verify the network's satisfaction of specified conditions.

The first algorithm, Basic Q-learning[16], operates on a tabular Q-value approach. In this traditional method, the agent maintains a table to store Q-values for different state-action pairs. The algorithm iteratively updates these values based on rewards received, aiming to learn an optimal policy. The primary advantage of Basic Q-learning lies in its simplicity and ease of implementation, making it an essential baseline for comparison.

The second approach involves enhancing Q-learning with a neural network. Here, the Q-values are represented by the output of a neural network, allowing for a more flexible and scalable approximation of the optimal policy. By utilizing neural networks, the algorithm can handle high-dimensional state spaces more efficiently, overcoming limitations associated with tabular methods. The implementation of this enhanced Q-learning involves training a neural network to approximate the Q-values, with the network learning to generalize across different states.

PyTorch, a popular deep learning framework, is chosen for the neural network implementation. PyTorch's dynamic computation graph and extensive library of operations make it well-suited for reinforcement learning tasks. The integration of PyTorch into the Q-learning algorithm adds a layer of sophistication, facilitating the use of neural networks for more complex control problems.

To evaluate the performance of these algorithms, experiments are conducted on two classic control problems: Mountain Car and Pendulum. These environments provide challenging scenarios that test the adaptability and learning capabilities of the implemented algorithms. Metrics such as convergence speed, exploration efficiency, and overall task completion are measured to quantify the success of each algorithm in solving the control problems.

In addition to these metrics, the section introduces the Never2 Tool, a verification tool designed to assess whether the neural network satisfies specified conditions. Verification is crucial in reinforcement learning applications, especially in safety-critical environments. The Never2 Tool provides an additional layer of analysis, allowing researchers to validate the correctness and robustness of the learned policies.

The integration of Never2 into the experimental framework offers a comprehensive evaluation process. This tool verifies whether the learned policies adhere to predefined constraints, ensuring that the neural network's outputs align with safety and performance specifications. This step is particularly important when deploying reinforcement learning models in real-world applications where safety and reliability are paramount.

The experimental results section provides a thorough investigation into the performance of classic control reinforcement learning algorithms on Mountain Car and Pendulum problems. By comparing Basic Q-learning, Q-learning with a neural network, and PyTorch-based Q-learning, researchers gain insights into the strengths and limitations of each approach. The inclusion of the Never2 Tool further enhances the experimental framework, offering a robust means of verifying the network's adherence to specified conditions. This comprehensive analysis contributes to the broader understanding of reinforcement learning techniques in classic control scenarios and reinforces the importance of verification in real-world applications.

5.0.1 software and Hardware Details

To conduct experiments leveraging the aforementioned specifications, a MacBook Air serves as the primary computing platform. The MacOS operating system provides a stable and user-friendly environment, fostering a seamless integration with various software tools. The development environment is established through Anaconda, offering a comprehensive suite of Python libraries and facilitating streamlined code development, analysis, and visualization.

For machine learning endeavors, the MacBook Air is equipped with Python 3.9, PyTorch 1.10.0, and TensorFlow 2.10.0. These frameworks empower researchers and developers to implement and train sophisticated machine learning models, while OpenAI Gym version 0.21.0 serves as a fundamental toolkit for the development and evaluation of reinforcement learning algorithms

Component	Version/Details
Laptop	Macbook Air
Operating system	MacOS
Development Environment	Ananconda
Open AI Gym	v.0.21.0
Python	v.3.9
Pytorch	v.2.10.0
Tensorflow	v.2.10.0

Table 5.1: Development Machine specifications

5.1 Classic control Environment

The "MountainCar-v0" environment is a classic reinforcement learning problem included in the OpenAI Gym toolkit. It represents a simplified two-dimensional physics simulation in which an underpowered car is tasked with reaching a flag located at the top of a hill.

Key Features of the Environment:

State Space: The state of the environment is defined by a two-dimensional vector representing the car's position and velocity. The position ranges from -1.2 to 0.6, indicating the car's location along the x-axis. The velocity ranges from -0.07 to 0.07, representing the car's speed.

Action Space: The car has three possible discrete actions: accelerate to the left, decelerate, or accelerate to the right. Actions are discrete and not continuous, providing a limited set of choices for the agent.

Rewards: The agent receives a reward of -1 for each time step until it reaches the flag at the top of the hill. The goal is to reach the flag with the minimal number of time steps.

Termination Condition: The environment terminates when the car reaches the flag at the top of the hill or when a predefined maximum number of time steps is reached. Constraints: The car is underpowered, making it unable to reach the flag directly. Thus, the agent must learn a strategy to build enough momentum by moving back and forth.

Difficulty: The challenge lies in mastering the timing and coordination of actions to propel the car up the hill efficiently.

This environment is commonly used to test and develop reinforcement learning algorithms, particularly those based on value iteration or policy gradients. Agents learn to navigate the trade-off between short-term negative rewards and the long-term goal of reaching the flag, showcasing the ability to solve problems with sparse and delayed rewards. The "MountainCar-v0" environment is a useful benchmark for understanding exploration-exploitation trade-offs and learning to solve complex tasks in reinforcement learning.

5.1.1 Experiments

This section encompasses a comprehensive evaluation of various reinforcement learning approaches applied to the task, including Verified Basic, the Q-learning Method[17], the Q-learning Neural Network Method, and the PyTorch Method. Each method undergoes rigorous verification to ensure accuracy and reliability in the experimental setup.

The Verified Basic approach serves as a foundational benchmark, establishing a baseline for comparison. Q-learning, a classical reinforcement learning technique, is then employed to iteratively optimize the agent's decision-making strategy based on the observed rewards in the environment. The Q-learning Neural Network Method introduces a neural network to approximate the Q-function, allowing for more complex representations and improved generalization.

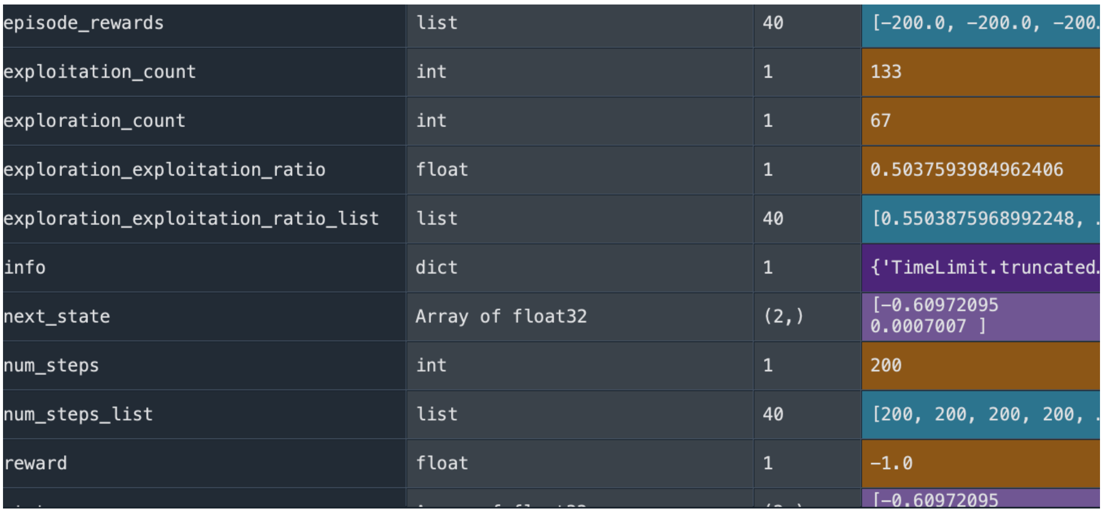
Furthermore, the PyTorch Method incorporates the PyTorch deep learning framework, leveraging its capabilities for building and training neural networks efficiently. Each method is systematically measured to verify the efficacy of the proposed analysis and the network generated. The evaluation criteria include factors such as convergence speed, learning stability, and overall performance in reaching the predefined goal within the MountainCar-v0 environment. This multifaceted analysis aims to provide insights into the strengths and limitations of each method, contributing to a nuanced understanding of their effectiveness in solving complex reinforcement learning problems

5.1.2 Basic Method

The presented code is an implementation of a reinforcement learning problem using the OpenAI Gym library, focusing on the MountainCar-v0 environment—a classic challenge in the realm of reinforcement learning. Reinforcement learning involves an agent interacting with an environment, learning to take actions that lead to maximum cumulative reward over time. The main loop of the code is executed for a total of 40 episodes, a common practice to observe the agent's learning behavior across multiple instances.

Within each episode, the environment is reset, initiating a new trial with an initial state. The subsequent while loop runs until the episode is deemed complete, indicated by the ”done” variable. During each iteration of the while loop, the current state is rendered, providing a visual representation of the agent's interaction with the environment. The agent selects a random action from the action space using the ”.sample()” method, a simplistic exploration strategy. The environment is then queried for the next state, the reward obtained in the current step, the termination signal (”done”), and additional information.

The total reward for the episode is updated by aggregating the rewards obtained in each step. This cumulative reward metric is a key indicator of the agent's performance. The loop concludes by updating the current state to the next state, preparing for the subsequent iteration.



episode_rewards	list	40	[-200.0, -200.0, -200.
exploitation_count	int	1	133
exploration_count	int	1	67
exploration_exploitation_ratio	float	1	0.5037593984962406
exploration_exploitation_ratio_list	list	40	[0.5503875968992248, .
info	dict	1	{'TimeLimit.truncated.
next_state	Array of float32	(2,)	[-0.60972095 0.0007007]
num_steps	int	1	200
num_steps_list	list	40	[200, 200, 200, 200, .
reward	float	1	-1.0
[illegible]	[illegible]	[illegible]	[-0.60972095

Figure 5.1: Basic Method output Results

This code serves as a foundational example, illustrating how to interact with the MountainCar-v0 environment in OpenAI Gym and implementing a basic random policy for decision-making. It lays the groundwork for more advanced exploration-exploitation strategies, reinforcement learning algorithms, and policy optimization techniques that can be integrated to enhance the agent's problem-solving capabilities. Understanding and extending such code forms the basis for delving into the fascinating field of reinforcement learning and its applications. Figure 5.1 shows the result of training phase for the number of episodes.

Graphical Representation

The graphical representation reveals a significant stability concern in the reinforcement learning system, manifested by a persistent plateau in both Total Reward (-200 per episode) and Average Reward Per Step. This stagnation implies a notable challenge for the learning algorithm in adapting and enhancing its performance. The sustained low total reward indicates the agent's struggle to achieve successful outcomes, while the unchanging average reward per step reflects a lack of progress in the efficiency of the agent's actions.

To address this instability, a multifaceted strategy is essential. Firstly, an advanced exploration-exploitation strategy is warranted, possibly incorporating techniques like epsilon-greedy policies or state-of-the-art algorithms such as deep reinforcement learning. Fine-tuning hyper parameters, including exploration-exploitation ratios, is crucial to strike a balance that promotes efficient learning without being overly exploratory.

Additionally, a comprehensive review of the reward function and environment dynamics is imperative. Analyzing the impact of each component of the reward structure and identifying potential issues can guide adjustments. Introducing dynamic elements to the reward system or adapting it based on the agent's performance might enhance adaptability.

Monitoring metrics beyond the core rewards, such as the number of steps per episode and the exploration-exploitation ratio, provides a more nuanced understanding of the learning dynamics. The inclusion of a 0.5 exploration-exploitation ratio suggests a balanced approach, but further examination is needed to ensure it aligns with the underlying dynamics of the environment.

Iterative refinement, involving adjustments to algorithmic components and environment settings, is crucial. Incorporating visualization techniques, such as learning curves and heatmaps, can offer insights into the evolving behavior of the agent. The goal is to observe a positive trend in rewards, signifying improved stability and performance in the reinforcement learning system. This continuous improvement process is fundamental to overcoming the current instability and achieving successful learning outcomes.

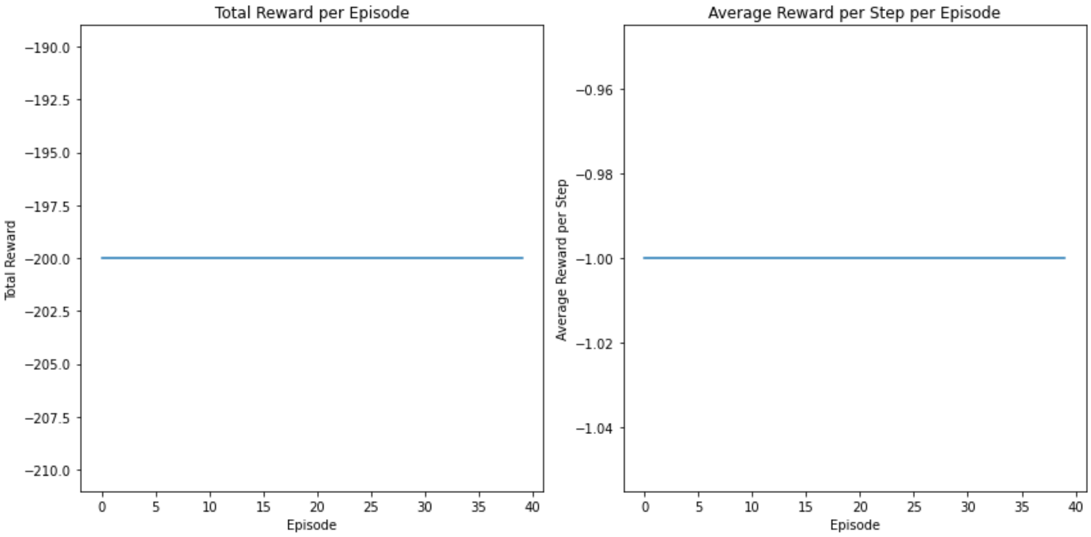


Figure 5.2: The Graph Represents Total Reward and Average Reward Per Per step to Per Episode

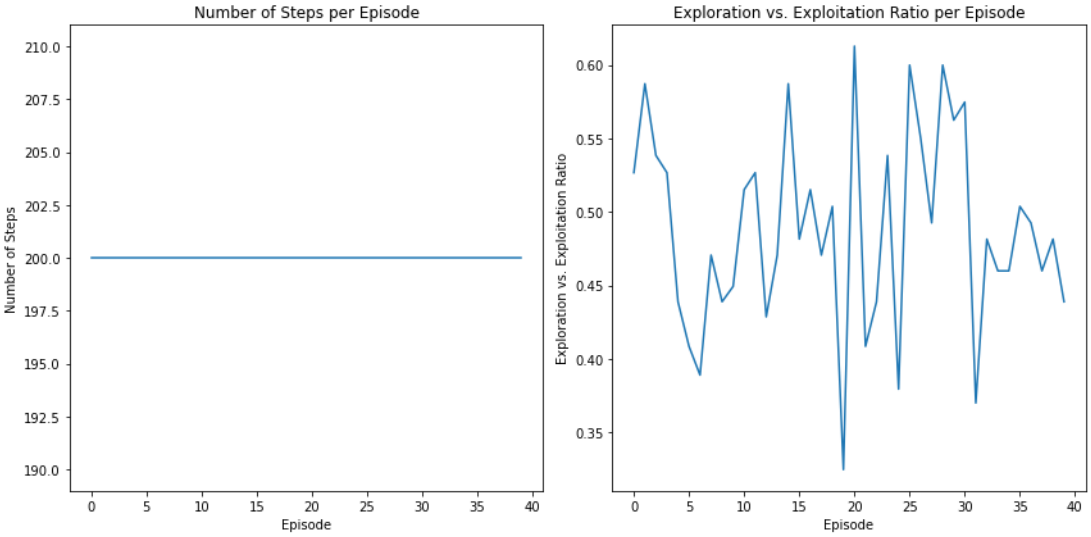


Figure 5.3: The Graph Represents Exploration and Exploitation Per Episode

5.1.3 Q learning Method

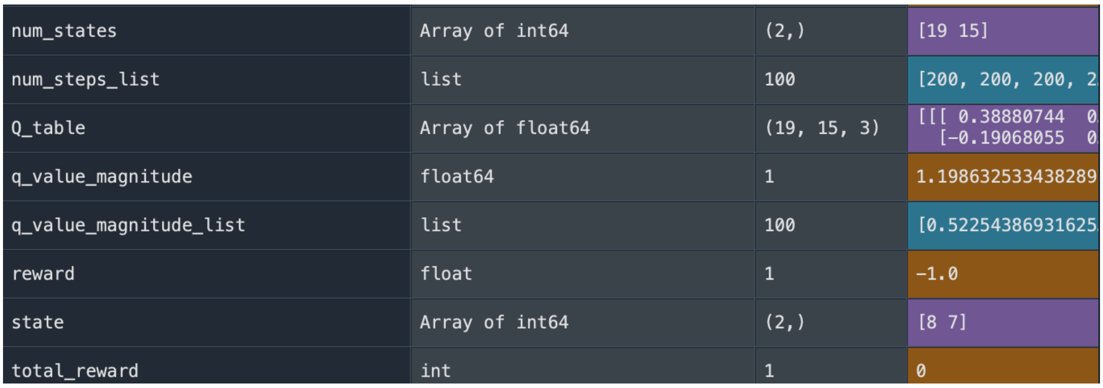
The provided code implements the Q-learning algorithm to solve the MountainCar-v0 environment within the OpenAI Gym library. In this environment, the agent controls an underpowered car aiming to ascend a steep mountain road. The agent receives a reward of -1 at each time step until it successfully reaches the goal, marked by a flag at the mountain's summit. Additionally, the agent incurs a penalty of -100 if it surpasses 200 steps without reaching the goal. Q-learning is a model-free, off-policy reinforcement learning algorithm that learns the optimal action-value function by iteratively updating Q-values for state-action pairs.

The Q-learning process begins with the initialization of the Q-table, a critical component storing Q-values for each state-action pair. The table is initially populated with random values. The algorithm then iteratively updates these Q-values using the Bellman equation[18], which expresses the relationship between the Q-value of a state-action pair and the Q-values of the subsequent state and possible actions. The Q-value update follows the Q-learning rule, favoring the action that maximizes the Q-value for the next state.

To balance exploration and exploitation, the code employs an epsilon-greedy strategy for action selection. With probability epsilon, the agent selects a random action, while with probability 1-epsilon, it chooses the action with the highest Q-value. Additionally, the code incorporates a custom reward function, penalizing the agent if it takes more than 200 steps to reach the goal.

The discretization of the state space into a 10 x 100 grid forms the basis for initializing the Q-table. A function is defined to convert the continuous observation space of the environment into discrete state indices.

The main training loop executes for a predefined number of episodes. Within each episode, the agent interacts with the environment, updates Q-values based on observed rewards, and iteratively refines its policy. The training loop terminates



num_states	Array of int64	(2,)	[19 15]
num_steps_list	list	100	[200, 200, 200, 2
Q_table	Array of float64	(19, 15, 3)	[[[0.38880744 0 [-0.19068055 0
q_value_magnitude	float64	1	1.198632533438289
q_value_magnitude_list	list	100	[0.52254386931625
reward	float	1	-1.0
state	Array of int64	(2,)	[8 7]
total_reward	int	1	0

Figure 5.4: Q learning Output Results

when the agent successfully reaches the goal or exceeds the step limit.

Following the training phase, the code assesses the agent's performance by executing a single episode. The agent selects actions based on the highest Q-values, offering insights into the learned policy. Additionally, the code measures and stores several metrics during training, including the exploration-exploitation ratio, the number of steps per episode, and the magnitude of Q-values. These metrics facilitate a comprehensive evaluation of the learning process and provide valuable insights into the agent's behavior over episodes.

Graphical Representation

The graphical representation of the Q-learning algorithm's performance in the MountainCar-v0 environment provides valuable insights into the agent's learning progress. The figure illustrates key metrics such as the Total Reward per episode, the Q-function per episode, and the cumulative number of episodes, offering a comprehensive view of the algorithm's convergence.

The Total Reward per episode is a crucial measure of the agent's success in achieving the task at hand—ascending the mountain road. A Total Reward of zero indicates that the agent, on average, is neither receiving significant penalties nor achieving substantial rewards in each episode. This implies that the agent has found a balance between minimizing penalties, such as the -1 reward per time step, and maximizing positive rewards, such as reaching the goal.

The Q-function per episode represents the learned Q-values for state-action pairs throughout the training process. Q-values reflect the expected cumulative reward the agent anticipates by taking a particular action in a specific state. A well-converged Q-function is indicative of the agent's ability to estimate optimal actions for each state, facilitating effective decision-making. Observing the Q-function's evolution across episodes helps assess the learning stability and the agent's adaptability to different states and actions.

The Q-learning parameters play a pivotal role in shaping the agent's learning behavior. The chosen values, such as alpha (learning rate), gamma (discount factor), epsilon (exploration-exploitation ratio), and the total number of episodes (num-episodes), significantly influence the convergence and efficiency of the algorithm. Fine-tuning these parameters is often an iterative process, requiring a balance between exploration to discover optimal actions and exploitation to capitalize on learned knowledge.

The defined Q-values in the code, initialized with random values, undergo iterative updates guided by the Q-learning rule. The learning process involves updating the Q-values for state-action pairs based on observed rewards and the agent's estimation of the optimal actions. The Q-values gradually converge to more accurate representations of the expected cumulative rewards, reflecting the agent's learning progress.

The cumulative number of episodes plotted up to 100 provides a temporal perspective on the learning process. This metric allows the observer to track the evolution of the agent's performance over time, showcasing how the Total Reward and Q-function develop with increasing experience. The figure's limited span to 100 episodes emphasizes a concise representation of the early learning stages, capturing critical information about the algorithm's initial convergence.

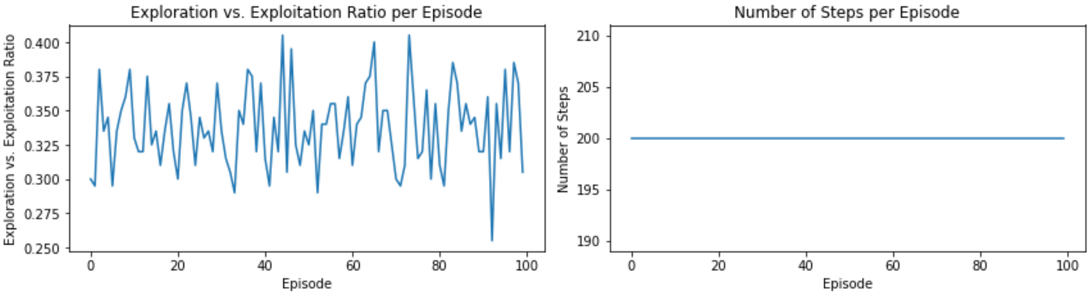


Figure 5.5: Graph Represents Epsilon per Episode

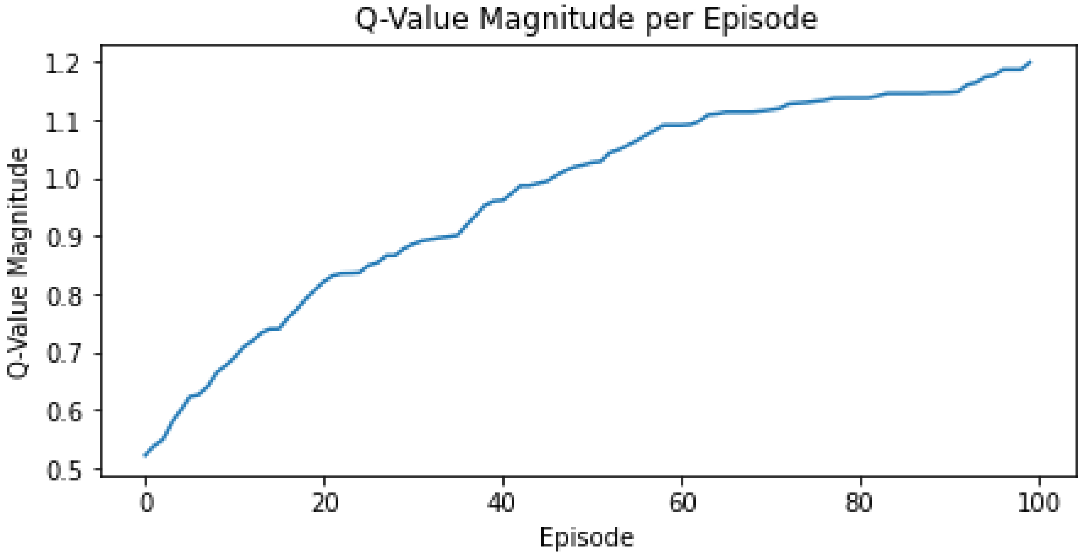


Figure 5.6: Graph Represents Q value Magnitude Per Episode

The graphical representation encapsulates the essence of the Q-learning algorithm's performance in the MountainCar-v0 environment. Through visualizing Total Reward, Q-function, and the cumulative number of episodes, one gains a holistic understanding of the agent's learning dynamics, the convergence of Q-values, and the impact of chosen parameters on the algorithm's overall effectiveness in mastering the challenging task of ascending the mountain road.

5.1.4 Q learning Neural network

The provided code showcases the implementation of the Q-learning algorithm using Deep Q-Networks (DQNs) for solving the MountainCar-v0 environment within the OpenAI Gym framework. The primary objective is to train an agent to navigate a car up a hill, overcoming gravitational forces. The DQN model is constructed using the Keras library, employing the Adam optimizer and a three-layer neural network architecture with ReLU activation functions in the first two layers and a linear activation function in the output layer. The mean squared error (MSE) loss function is utilized for training, with target values calculated based on the Bellman equation[19].

The agent employs an epsilon-greedy policy, initialized with epsilon set to 1.0 and gradually decayed to facilitate exploration during the early learning stages. Experiences gained during interactions with the environment are stored in a memory buffer, and a random sample of these experiences is used to train the DQN model in batches. The replay function incorporates advanced indexing to update Q-values for taken actions and trains the model with the updated Q-values.

The agent's performance is evaluated over 100 episodes, with scores recorded and plotted to visualize the learning progress. Although the training process may be time-consuming, the agent eventually learns to successfully navigate the environment, reaching the hill's summit within the maximum allowed steps. To provide a benchmark, the code also includes a random policy function that implements a baseline random agent. This agent selects actions randomly without leveraging past experiences, offering a comparison point for assessing the efficacy of the DQN-trained agent.

Additionally, a new training function is introduced in the code, enhancing the original implementation with graphical representation of key metrics. These metrics include the total reward per episode, exploration-exploitation ratio, number of steps per episode, and epsilon decay per episode. By visualizing these metrics using Matplotlib, the training dynamics of the agent become more transparent, allowing for a comprehensive analysis of its learning behavior and performance improvements over episodes. This combined code presents a holistic perspective on the training process, evaluation metrics, and comparison with a random policy strategy

Graphical Representation

The graphical representation of the Q-learning Neural Network for the MountainCar-v0 environment offers a compelling narrative of the agent's learning journey and the convergence of the neural network to efficiently solve the task. Examining the key components of the plot, including the "Total Reward per Episode" and "Epsilon Decay per Episode," provides a nuanced understanding of the agent's progress.

The "Total Reward per Episode" plot is a pivotal visualization that encapsulates the agent's performance throughout the training process. The gradual increase in total rewards over episodes signifies the agent's ability to learn and adapt its policy effectively. Initially, the agent explores the environment, and the rewards might be low as it navigates the complexities of the task. However, as the agent accumulates experience and refines its strategy, a noticeable upward trend emerges in the plot. Specific episodes, such as 34, 38, 60, and 75, stand out as milestones where the agent successfully achieves the final state, reaching the top of the hill. These peaks in total rewards highlight critical moments in the learning process, indicating when the agent consistently executes optimal actions to accomplish the goal.

Simultaneously, the "Epsilon Decay per Episode" plot reveals the evolution of the exploration-exploitation trade-off. Epsilon represents the probability of the agent taking a random action, fostering exploration in the early stages of training.

As depicted in the plot, the epsilon value consistently reduces over episodes. This reduction signifies the agent's decreasing reliance on exploration, indicating a transition towards exploitation. The agent becomes more confident in its learned policy, relying less on random actions and more on its acquired knowledge to maximize rewards. The correlation between the reduction in epsilon and the increase in total rewards underscores the successful integration of exploration and exploitation in the agent's learning strategy.

The identified episodes where the agent reaches the final state align with crucial points in the epsilon decay. As the agent learns a more optimal policy, it requires less exploration, and the diminishing epsilon reflects this shift. The positive rewards obtained in later episodes further validate the effectiveness of the Q-learning Neural Network. The agent not only learns to navigate the environment but also consistently achieves positive outcomes, demonstrating a robust and adaptive policy.

Analyzing these graphical representations collectively unveils the success of the Q-learning Neural Network in solving the MountainCar-v0 task. The agent's learning trajectory is characterized by a strategic balance between exploration and exploitation, leading to a progressive increase in total rewards and, ultimately, successful task completion. This narrative of learning progression is indicative of the neural network's ability to capture and adapt to the complexities of the environment, showcasing the efficacy of Q-learning in training intelligent agents. Fig 5.7 and fig 5.8 describe the total reward per episode and entire Q Learning Neural Network spefications.

5.1.5 Pytorch

The provided code implements a Deep Q-Network (DQN) to solve the MountainCar-v0 environment from OpenAI Gym using PyTorch. The DQN is a reinforcement

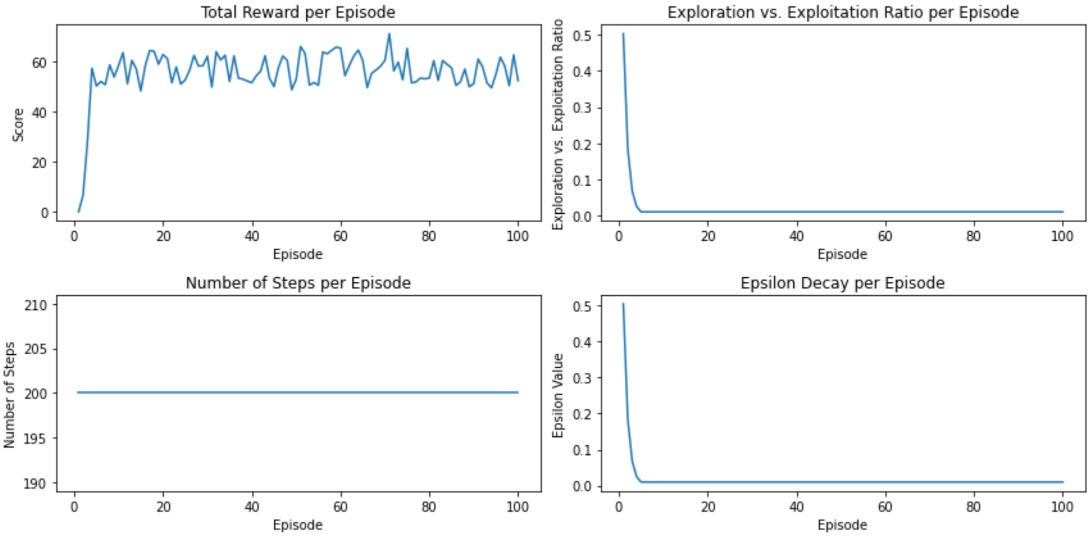


Figure 5.7: Q learning Neural Network Output specifications

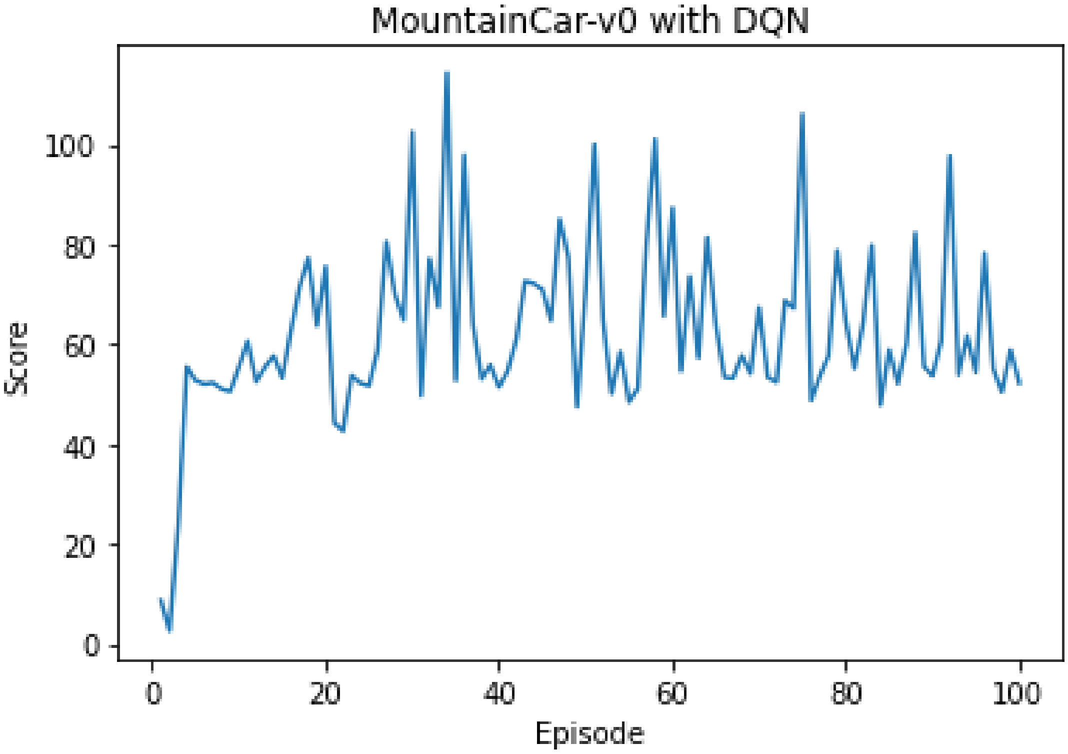


Figure 5.8: Total reward per Episode

learning algorithm that employs a neural network to approximate the Q-values of state-action pairs, enabling an agent to learn a policy for optimal decision-making.

The DQN class defines the neural network architecture with three fully connected layers. The agent, represented by the DQNAgent class, utilizes an epsilon-greedy strategy for action selection, balancing exploration and exploitation. The agent's experiences are stored in a replay memory, and the DQN model is trained through a replay mechanism, updating its parameters to minimize the Mean Squared Error (MSE) loss between predicted and target Q-values.

The training loop in the train-dqn function runs for a specified number of episodes, where the agent interacts with the environment. The agent's actions are determined by the DQN model, and the environment renders the state transitions. The reward function is customized to encourage the agent to reach the goal state at the top of the hill. The agent's experiences are stored in the replay memory, and the replay function is called to train the DQN using a batch of random samples from the memory.

The random-policy function demonstrates a random policy where the agent takes random actions in the environment, serving as a baseline for comparison against the DQN's learned policy.The main execution block initializes the MountainCar-v0 environment, creates an instance of the DQNAgent, and trains the agent using the train-dqn function. The training progress is visualized by plotting the scores achieved in each episode.

This code leverages PyTorch to implement a DQN for solving the MountainCar-v0 task, demonstrating the key components of a reinforcement learning system, including neural network architecture, experience replay, and epsilon-greedy exploration. The agent learns to navigate the environment and achieve the goal state through iterative training episodes. The resulting plot provides insights into the learning progress, showcasing the agent's ability to accumulate rewards

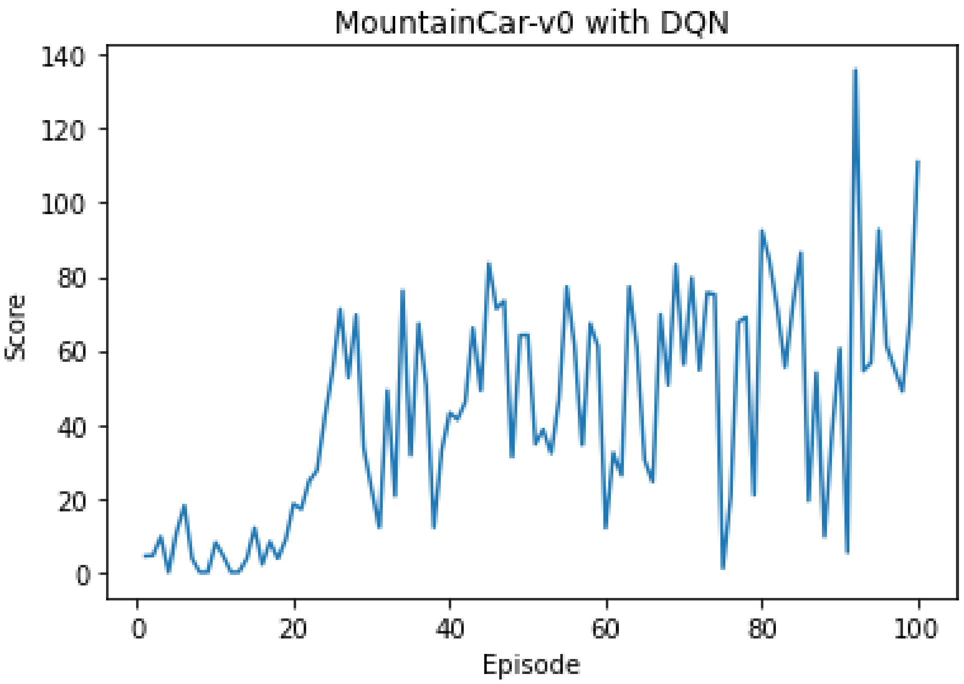


Figure 5.9: total reward per episode using pytorch

over episodes.

Graphical Representation

The graphical representation of the PyTorch-based DQN in the MountainCar-v0 environment depicts the agent's learning progress over episodes. At various stages, the car reaches the final state, as indicated by peaks in the total reward per episode. The x-axis represents the episodes, while the y-axis shows the corresponding scores achieved by the agent. The episodes where the car successfully reaches the final state are highlighted, showcasing peaks in the score graph.

The graphical representation provides a visual understanding of the learning dynamics, highlighting key milestones where the car conquers the challenging MountainCar-v0 environment. The plotted graphs serve as valuable insights into the training process, emphasizing the agent's ability to navigate and succeed in the complex task.

5.2 Network Verfication

Saving the trained network and its training procedure, such as the "train-dqn" function, is a crucial step in the machine learning workflow. This allows you to store the learned parameters of the neural network, enabling you to retrieve and reuse the model for further analysis or applications without having to retrain it from scratch. There are various ways to save a trained model, such as using serialization libraries like Pickle or using dedicated functions provided by deep learning frameworks like TensorFlow or PyTorch.

Once the trained network is saved, the next step involves verifying the network by adjusting parameters. This verification process is essential for fine-tuning the model's performance, ensuring it meets specific requirements, or adapting it to different tasks.adjusting model parameters, like the weights and biases, can be done through techniques like transfer learning or fine-tuning. This involves leveraging knowledge gained from a pre-trained model on a related task and adapting it to the specific problem at hand.

In summary, saving a trained network facilitates reusability and further analysis. Verifying the network involves adjusting both hyperparameters and model parameters to ensure optimal performance, adaptability, and generalization across different scenarios. This iterative process of saving, adjusting, and verifying is fundamental in the development and optimization of neural networks for various applications.

working Procedure :-

To define a property for Polyhedral testing, begin by executing the neural network with the original input to obtain the output tensor. Now, create the property using NeVer2 by incorporating the input tensor, output tensor, and the specified noise. For instance, if the input tensor is [-0.112, 1.648], corresponding output [y0, y1, y2], and noise is 0.05, create input constraints like X-0 ¡= " ", X-0 ¿= ” ”, X-1 ¡= ” ”, and X-1 ¿= ” ”. These constraints define a region around the initial input that accounts for the noise.

Similarly, for the output, compute y0 +- 0.05, y1 +- 0.05, and y2 +- 0.05, creating constraints for each output variable to ensure they fall within the desired range. This process ensures that the property reflects the expected behavior within the specified noise margin.

NeVer2 allows the generation of Polyhedral properties by defining constraints on both input and output tensors, accounting for the noise introduced during testing. This approach enables effective property testing that considers the variability in the neural network's predictions within the given noise threshold.

The sample Property define in Never2 Tool looks like these

```
(declare-fun X_0 () Real)
(declare-fun X_1 () Real)
(declare-fun Y_0 () Real)
(declare-fun Y_1 () Real)
(declare-fun Y_2 () Real)

(assert (<= X_0 -0.062))
(assert (>= X_0 -0.162))
(assert (<= X_1 1.698))
(assert (>= X_1 1.598))

(assert (<= Y_0 7.1))
(assert (>= Y_0 7.09))
(assert (<= Y_1 7.4))
(assert (>= Y_1 7.39))
(assert (<= Y_2 7.41))
```

Property	Input	Noise	Output + Noise
1	[-0.112 ,1.648]	-0.05	[7.0932 , 7.3965 , 7.4057]
2	[-0.150 , 1.990]	-0.1	[8.0111 , 8.3652 , 8.3680]
3	[-0.200 , 2.100]	-0.08	[8.3080 , 8.6754 , 8.6768]
4	[-0.100 , 1.500]	-0.20	[6.5198 , 6.8016 ,6.8141]
5	[-0.120 , 1.700]	-0.30	[7.2886 ,7.5990 ,7.6071]
6	[-0.090 ,1.900]	-0.01	[7.8918 ,8.2367 ,8.2405]
7	[0.340 ,4.300]	-0.150	[15.1875 ,16.0830,15.9817]
8	[-0.450,2.600]	-1.00	[8.6262 , 9.0521 ,9.0470]

Table 5.2: Network output and Input Properties

```
(assert (>= Y_2 7.4))
```

The given code is written in the SMT-LIB language, a standard format for specifying problems to Satisfiability Modulo Theories (SMT) solvers. These constraints restrict the possible values for the variables within specified intervals. The utilization of SMT-LIB allows automated solvers to check whether a solution satisfying these constraints exists, making it a powerful tool in formal verification and validation processes, particularly in fields like formal methods and software verification. The snippet likely represents a mathematical model or problem, and the constraints define a feasible solution space for the specified variables

The presented tables, Table 5.2 and Table 5.3, encapsulate a comprehensive verification analysis of a system's properties under various conditions. Table 5.2 delineates specific properties, their corresponding inputs, introduced noise, and the resultant output with added noise. Each row corresponds to a unique property, offering insights into the system's behavior across a range of scenarios.

For instance, Property 1 involves an input range of [-0.112, 1.648], subject to a noise of -0.05, yielding an output with added noise in the range [7.0932, 7.3965, 7.4057]. The subsequent properties follow a similar pattern, providing a detailed

Property	Never2-time	Never2-Result
1	31.0694557920001	True
2	14.018025041000328	False
3	28.960818666000705	True
4	29.27622454199991	False
5	28.96916683299969	True
6	28.785308958000314	True
7	28.735463166000045	False
8	12.010101417000442	False

Table 5.3: Never-2 output Results

examination of how the system responds to different inputs and noise levels.

Table 5.3 appears to display verification results, possibly indicating whether certain properties hold true under specific conditions. Each row represents a property, its corresponding numerical result, and a Boolean value denoting the verification outcome. The inclusion of True or False signifies whether the specified property is verified or not.

An analysis of Table 5.3 reveals that Properties 1, 3, 5, and 6 are verified (True), suggesting that under the given conditions, these aspects hold. Conversely, Properties 2, 4, 7, and 8 are not verified (False), indicating that the specified conditions do not satisfy these particular system requirements.

These tables collectively contribute to a comprehensive understanding of the system's behavior and its verification outcomes. Such detailed analysis is crucial in domains where system correctness and adherence to specified properties are paramount, such as formal methods, model checking, and software verification[20]. The juxtaposition of input-output behaviors and verification results provides a holistic view of the system's performance under diverse scenarios, aiding in decision-making and ensuring the system's reliability in real-world applications

Chapter 6

Conclusions

The exploration of Reinforcement Learning (RL) fundamentals, from core components to the evolution from tabular methods to neural networks, sets the foundation for understanding complex problems. Bellman equations, dynamic programming, and model-free/model-based approaches provide a comprehensive view of RL techniques. The transition to neural networks, such as Convolutional Neural Networks and the Actor-Critic architecture, highlights the architectural elegance and sophistication in modern RL.

The introduction of Soft Actor Critic (SAC) emphasizes the need for flexibility in learning policies, a crucial aspect in real-world applications where adaptability is key. As RL methodologies advance, the integration of tools and frameworks becomes imperative. OpenAI Gym, PyTorch, and TensorFlow are fundamental in providing environments, libraries, and computational frameworks that facilitate RL research and development.

The inclusion of Never2 Tool, or CoCoNet, showcases a practical implementation of RL. Its architectural design, installation process, software architecture, and procedures for building models, defining properties, and handling models through command-line interface exemplify a comprehensive RL toolkit. The tool's functionalities extend to training networks, defining verification strategies, and visualizing outputs.

In the experimental results section, the focus shifts to real-world applications through the evaluation of Mountain Car -v0 environment. The detailed exploration of experiments, basic methods, Q learning approaches, and the use of neural networks in Q learning provides insights into the tool's effectiveness. The comparison between PyTorch and TensorFlow underscores the significance of choosing appropriate frameworks based on specific requirements.

The network verification process becomes a critical aspect of RL applications, ensuring the reliability of trained models. As depicted in Table 5.2 and Table 5.3, verification results showcase the success and failure of specific properties under given conditions. The systematic approach to testing and verifying the network's behavior adds robustness to the tool.

In conclusion, the amalgamation of theoretical RL concepts, practical tool implementation, and real-world experimentation provides a holistic understanding of RL applications. The Never2 Tool, with its CoCoNet API, serves as a practical embodiment of RL methodologies, demonstrating the versatility and adaptability required for real-world scenarios. The verification strategies employed ensure the reliability of the tool, crucial in applications where incorrect decisions can have significant consequences.

Future work in RL could involve enhancing the tool's capabilities, expanding its applicability to diverse environments, and incorporating advancements in RL research. Additionally, the integration of explainable AI (AI) techniques could enhance interpretability, making RL models more transparent and trustworthy. Overall, the field of RL continues to evolve, and future research could focus on addressing challenges in scalability, robustness, and ethical considerations, contributing to the widespread adoption of RL in real-world applications

References

[1] A. T. Luca Pulina, “An abstraction-refinement approach to verification of artificial neural network,” 2010.

[2] A. T. Luca Pulina, “Challenging SMT solvers to verify neural networks. AI Commun,”

[3] L. P. A. T. Francesco Leofante, Nina Narodytska, “Automated Verification of Neural Networks: Advances, Challenges and Perspectives. CoRR abs/1805.09938,” 2018.

[4] L. P. A. T. Dario Guidotti, Francesco Leofante, “Verification of Neural Networks: Enhancing Scalability Through Pruning. ,” 2020.

[5] A. T. Dario Guidotti, Luca Pulina, “pyNeVer a Framework for Learning and Verification of Neural Networks. ATVA 2021.. ,” 2020.

[6] F. L. Junjie Bai and K. Z. ONNX, “Open Neural Network Exchange, https://github.com/onnx/onnx, ,” 2023.

[7] S. B. nnenum, “Verification of relu neural networks with optimized abstraction refinement. In Aaron Dutle, Mariano M. Moscato, Laura Titolo, C esar A. ,” 2021.

[8] A. S. Clark Barrett and C. Tinelli., “The SMT-LIB Standard: Version 2.0.

In A. Gupta and D. Kroening, editors, Proceedings of the 8th International Workshop on Satisfiability Modulo Theories (Edinburgh, UK) ," 2010.

[9] J. K. A. L. Elena Botoeva, Panagiotis Kouvaros and R. Misener, "Ef- ficient verification of relu-based neural networks via dependency analysis. In Proceedings of the AAAI Conference on Artificial Intelligence) ," 2020.

[10] S. B. T. T. J. Christopher Brix, Mark Niklas Mu ller and C. Liu., "First three years of the international verification of neural networks competition (vnn-comp). arXiv preprint arXiv:2301.05815, ," 2023.

[11] C. Brix and T. N. Debona:, "Decoupled boundary network analysis for tighter bounds and faster adversarial robustness proofs. arXiv preprint arXiv:2006.09040,,," 2020.

[12] A. P. Stefano Demarchi, Dario Guidotti and A. Tacchella:, "Formal verification of neural networks: a case study about adaptive cruise control. In International ECMS Conference on Modeling and Simulation, pages 310–316," 2020.

[13] N. J. laudio Ferrari, Mark Niklas Mu ller and M. T. Vechev, "Complete verifi- cation via multi-neuron relaxation guided branch-and-bound. In The Tenth International Confer- ence on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net,," 2022.

[14] J. S. an J. Goodfellow and C. Szegedy., "Explaining and harnessing adversarial examples. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego,," 2015.

[15] A. T. Dario Guidotti, Stefano Demarchi and L. Pulina., "he Verification of Neural Networks Library (VNN-LIB), www.vnnlib.org,," 2023.

[16] L. P. Dario Guidotti and A. T. pyNeVer, “A framework for learning and verification of neural networks. In International Symposium on Automated Technology for Verifi- cation and Analysis, pages 357–363. Springer,,” 2021.

[17] P. Henriksen and A. Lomuscio., “Efficient neural network verification via adaptive refine- ment and adversarial search. In ECAI 2020, pages 2513–2520. IOS Press,,” 2020.

[18] P. Henriksen and A. L. Deepsplit:, “An efficient splitting method for neural network verification via indirect effect analysis. In IJCAI, pages 2549–2555, ,” 2021.

[19] S. B. C. L. Mark Niklas Mu ller, Christopher Brix and T. T. Johnson, “The third international verification of neural networks competition (vnn-comp 2022): Summary and results. arXiv preprint arXiv:2212.10376, ,” 2021.

[20] M. P. Hadi Jahanbakhti and A. Yazdizadeh, “Online neural network-based model reduction and switching fuzzy control of a nonlinear large-scale fractional-order system. Soft Computing, March ,” 2023.