International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

In Door Target-Driven Visual Robot Navigation Using Deep Reinforcement Learning (DRL) Approaches

Pooja Upadhyay[1], Dr. Bappaditya Jana[2]

[1]Master of Technology in Computer Science and Engineering, Sheat College of Engineering, Varanasi, India

[2]Faculty of Computer Science and Engineering, Dr. APJ Abdul Kalam Technical University, Lucknow, India

ARTICLEINFO

Article History:

Accepted: 05 Sep 2023

Published: 16 Sep 2023

Publication Issue

Volume 9, Issue 5

September-October-2023

Page Number

216-235

ABSTRACT

There have been several successful implementations of deep RL in game-like settings. Deep reinforcement learning (RL) has great potential, but it is difficult to apply it to visual navigation in realistic 3D settings. To guide an agent to an image-based goal, we present a unique learning architecture. We improved the efficiency of visual navigation by including additional tasks into the batched advantage actor-critic (A2C) algorithm. For the prediction of the depth map, the segmentation of the observation picture, and the segmentation of the target image, we propose three new auxiliary tasks. By doing these tasks, supervised learning may be used to pre-train a significant portion of the network, cutting down on the total number of training iterations. Gradually increasing the environment's complexity over time may further increase training performance. An effective neural network architecture is described that can generalize across numerous goals and settings. Our approach outperforms the best goal-oriented visual navigation algorithms in the literature on the AI2-THOR environment simulator, and it works in continuous state spaces.

Keywords: Deep Reinforcement Learning, AI2-THOR, Advantage Actor-Critic

I. INTRODUCTION

The area of robotics relies heavily on autonomous navigation. Because of its cheap initial cost, small footprint, and wealth of data, visual navigation has received a lot of attention and research in recent decades. Conventional visual navigation techniques [1–5] often use rule-based approaches to pinpoint the robot's location in service of planning and control. These rule-based techniques may be effective, but they are labor- and time-intensive to develop manually and computationally.

The expansion of cellular networks has kept pace with the speed of scientific and technological progress. Wireless communication technologies are frequently employed in practice; for instance, 5G networks have become widespread in many major cities to facilitate easier communication. It is not yet powerful enough to

fulfill user capacity, and it does nothing to address the root causes of the disconnect between people's communication requirements and the availability of spectrum [3]. As a result, academics are now focusing on how to make the most use of spectrum resources for wireless communication via strategic planning and implementation.

Radio's electromagnetic spectrum resources are among the most valuable parts of the wireless communication network, and the government has distributed them wisely. Even if there are more and more IoT terminals, the congestion problem has not been solved. However, this does not imply that everything has been used up. According to research, the usage rate of various spectrum resources is as low as 15% in some locations of the United States. This study provides conclusive evidence that existing national spectrum resources have not been exploited efficiently in many nations [5]. So, this problem has been handled by the National Spectrum Supervision Bureau. Only those who have been granted permission to utilize the allocated frequency range may do so using the original electrostatic allocation procedure. Unlicensed users and legal users may coexist peacefully in the same frequency range because to dynamic spectrum resource allocation. As a promising new technique, reinforcement learning offers considerable promise in addressing challenging issues related to the allocation of dynamic resources [7]. The notion of deep reinforcement learning (DRL) has been introduced in response to the rising profile of deep learning algorithms in the world of computing. The first is that it prioritizes the system's long-term profitability above its short-term profitability while pursuing improvement. The second is that the process of application doesn't need any background knowledge of the environment to provide near-optimal results, with the added bonus of being able to explore and make decisions optimally on its own [10]. That's why incorporating DRL into wireless communication technologies is so crucial.

Numerous academic investigations of its effectiveness and enhancements have been undertaken. Three simple network procedures were also developed for use in assessing these networks' efficacy. Finally, excellent transmission performance was observed between VLC at varying speeds and FSO in five representative air quality circumstances, validating the viability [11]. Nayak et al. [12] made it easier to regulate the velocities, amplitudes, and directions of waves, which improved the quality of radiation along the receiving line. The antenna was discovered to rely mostly on a flat construction, which allowed for easier integration and minimization of mobile terminals [12]. After comparing several energy-efficient communication strategies for WSNs, Sopara et al. [13] concluded that their suggested system had the best energy-saving effects and offered a solid experimental foundation for advancing wireless communication technologies.

People living in the Internet of Things age deal with an overwhelming quantity of data and information on a daily basis. As a result, data processing that takes use of intelligence is crucial. Application of deep learning, which is utilized for the intelligent extraction of information characteristics, is expanding rapidly throughout all sectors of society. Eventually, Ohsugi et al. [14] in the area of materials medicine used deep learning to identify RRD in ultra-wide field fundus pictures. For early detection of RRD and avoidance of blindness, they observed that ultra-wide field fundus considerably improved diagnostic accuracy [14. The potential of deep neural networks (DNNs) to filter out background noise was investigated by Chen et al. [18]. The anti-noise capabilities of DNNs has been bolstered by the proposal of a novel activation function called rand-softplus (RSP) to model the response process [18]. For their regression models, Joy et al. [19] turned to DNNs. Standardization at the discourse level was shown to be possible using the DNN-based technique after training and optimizing the model. Liu and Wang [20] turned to DNNs. In addition, two distinct DNN training methodologies were developed, therefore

extending the use of DNNs beyond the realms of language and signal processing [20]. To optimize the DNNs' architecture.

In conclusion, there are several studies on both DRL and wireless communication, but very few that combine the two.Both rule-based and learning-based approaches may be used for visual navigation in robots.

An overview of IL can be found in [13], and a subset of IL techniques refers to the work as behavior cloning (BC), a kind of specialized supervised learning. NVIDIA [14] introduced a standard end-to-end visual navigation technique for autonomous driving. Yang [15] suggested an IL-based multi-task architecture that uses picture sequences to anticipate both the vehicle's speed and its steering angles. Wang [16] suggested a unique angle-branch network for autonomous driving, including sequential pictures, vehicle speed, and subgoal angle as inputs. Visual navigation techniques based on imitation learning have their uses, however they are vulnerable to overfitting when employing just IL.

The capacity of RL, particularly DRL, to interact with its surroundings has led to its recent use in robot navigation challenges. One such DRL approach is DeepMind's proposed deep Q-learning algorithm (DQN), which has already helped robots acquire human-level control strategies]. Many techniques for enhancing the DQN network model have now been described in rapid succession, with promising outcomes in a variety of contexts [20–23]. Including 3D VizDoom, and a feedforward architecture is described in [27] for learning a deep successor representation. It was also suggested that navigation challenges may be solved using the AutoRL approach [28], which combines DRL with gradient-free hyperparametric optimization. However, the RL-required reward function is notoriously hard to build and often fails to ensure optimum performance. To address the whole scope of visual navigation issues, Zhu [12] presented a target-driven DRL (TD-DRL) architecture, which has shown promising results whereas our proposed method is specifically tailored to improve sample efficiency in visual navigation tasks.

In multi-agent deep reinforcement learning (MADRL), agents (or decision makers) work together or against one another in a given setting to attain a common objective. By incorporating deep learning (DL), the most current AI development, MADRL improves upon the capabilities of classical RL and multi-agent RL (MARL). This allows the agents to work together[4] to improve system performance.

In recent years, MADRL has made significant progress because of its capacity to address challenging real-world situations, which conventional RL has struggled to address. Numerous agents collaborate or compete with one another, has shown to be insufficient [6]. As interest in MADRL has grown, researchers have conducted several surveys to learn more about it from various angles. The collaborative nature of MADRL is discussed by Oroojlooyjadid and Hajinezhad [9]. Knowledge reuse in MADRL is discussed by Da Silva et al. [10]. Gronauer and Diepold [12] and Zhang et al. [11] explored the technical issues of MADRL from a mathematical viewpoint, while other research focused on theoretical assessments.

This study contributes to the current literature by establishing a taxonomy of MADRL aspects such as aims, characteristics, problems, applications, and performance measurements, and by surveying MADRL algorithms applied to different state-of-the-art applications. Based on the taxonomy, the MADRL algorithms are categorized, examined, and debated, and their unresolved problems are elucidated. Our interpretations from these angles are new to the literature to the best of our knowledge. Focus, method, and intended multi-agent setting are summarized across recent MADRL surveys in Table 1. Information regarding our article is included in the table as well. This paper's overarching goal is to provide an overview of current and emerging MADRL application research fields and to inspire readers to go further into the topic.

When it comes to human navigation, a specific location isn't necessary; rather, we only have to figure

out where to travel based on the sights we've seen. This kind of conduct has sparked a rise in curiosity in comprehensive visual navigation strategies that let users go straight from viewing a picture to performing an action. picture classification [6, 7], object identification [8, 9], and picture segmentation [10, 11] have all made significant strides thanks to deep learning's (DL) ability to directly extract valuable features from pixels. Achieving a greater degree of intelligence. First, the observation is assumed to be four consecutive photos; in reality, the robot has access to additional observations from its immediate surroundings. Second, the halt action is ignored despite the fact that it may be automatically identified in a simulation setting by comparing the current state's ID number to the target state's ID number; however, this method is inappropriate in the real world because there is no ID number and only photos are available for comparison. Finally, DRL needs a lot of training data to develop an effective navigation policy, which increases training time and decreases efficiency. The imitation learning (IL) approach is an option to enhance sample efficiency during training; it can correctly mimic expert experience, and it needs less training time to build a navigation model. Overfitting is a potential issue when employing IL instead of DRL. Several effective applications of deep RL in game-like contexts have been developed. Although deep RL shows promise, it is challenging to apply to visual navigation in realistic 3D environments. We provide a novel learning architecture for directing an agent toward an image-based objective. By adding more work to the batched advantage actor-critic (A2C) algorithm, we were able to increase the effectiveness of visual navigation. We propose three new ancillary tasks: depth map prediction, observation picture segmentation, and target image segmentation. By doing these actions, supervised learning may be utilized to pre-train a sizable chunk of the network, reducing the need for iterations during actual training. We offer a powerful neural network design that can generalize over a wide range of objectives and environments.

II. THE STRUCTURE OF THE SYSTEM

One of the fundamental requirements for an autonomous agent to carry out a wide range of activities in complicated situations is the ability to navigate visually. This feature refers to an agent's capacity to comprehend its immediate surroundings and safely travel to a predetermined destination using data collected from its own on-board visual sensors. There are wokey aspects to this. To begin, the agent has to be able to evaluate the current observation and deduce the factors that are most important to the goal. Second, the agent has to know how its navigational behaviors influence the way it looks at its environment.

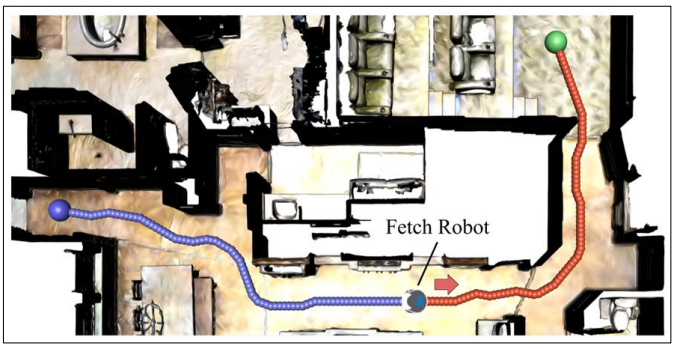

Fig. 1.1: Visual Guidance System Diagram

III. SYSTEM OVERVIEW

In contrast to [3], which employs ResNet features, our method has the agent learn navigation only from the observed raw pictures, which allows the agent to acquire valuable traits even when there are no informative rewards available. We employ depth-maps and picture segmentations as training objectives for the auxiliary tasks while training the deep neural network. We also provide a technique for pre-training the neural network before to using the RL algorithm. This is achieved by acquiring knowledge in one context and applying it to a more complicated one. Finally, we offer a unique, efficient, and compact neural network design

to solve the partial observability issue. We test our approach in a variety of interior settings representative of real-world use, as in [3] and [8].

IV. RELATED WORK

The expansion of cellular networks has kept pace with the speed of scientific and technological progress. Wireless communication technologies are frequently employed in practice; for instance, 5G networks have become widespread in many major cities to facilitate easier communication. It is not yet powerful enough to fulfill user capacity, and it does nothing to address the root causes of the disconnect between people's communication requirements and the availability of spectrum [3]. As a result, academics are now focusing on how to make the most use of spectrum resources for wireless communication via strategic planning and implementation. Radio's electromagnetic spectrum resources are among the most valuable parts of the wireless communication network, and the government has distributed them wisely. Even if there are more and more IoT terminals, the congestion problem has not been solved. However, this does not imply that everything has been used up. According to research, the usage rate of various spectrum resources is as low as 15% in some locations of the United States. This study provides conclusive evidence that existing national spectrum resources have not been exploited efficiently in many nations [5]. So, this problem has been handled by the National Spectrum Supervision Bureau. Only those who have been granted permission to utilize the allocated frequency range may do so using the original electrostatic allocation procedure. Unlicensed users and legal users may coexist peacefully in the same frequency range because to dynamic spectrum resource allocation. As a promising new technique, reinforcement learning offers considerable promise in addressing challenging issues related to the allocation of dynamic resources [7]. The notion of DRL has been introduced in response to the rising profile of deep learning algorithms in the world of computing. By seeing high-dimensional raw data, DRL is able to train the resulting agents to learn the behaviors in photos or videos, which is not possible characteristics [8,9]. There are two main features of DRL. The first is that it prioritizes the system's long-term profitability above its short-term profitability while pursuing improvement. The second is that the process of application doesn't need any background knowledge of the environment to provide near-optimal results, with the added bonus of being able to explore and make decisions optimally on its own [10]. That's why incorporating DRL into wireless communication technologies is so crucial. Huang et al. [11] suggested interconnection, particularly in a radio frequency-sensitive or safety-required setting. Three simple network procedures were also developed for use in assessing these networks' efficacy. Finally, excellent transmission performance was observed between VLC at varying speeds and FSO in five representative air quality circumstances, validating the viability of this HOW network [11]. Nayak et al. [12] made it easier to regulate the velocities, amplitudes, and directions of waves, which improved the quality of radiation along the receiving line. Some designs of the receiving equipment were envisioned by incorporating the existing state of advancements in telecommunication frameworks and radio lines. After comparing several energy-efficient communication strategies for WSNs, Sopara et al. [13] concluded that their suggested system had the best energy-saving effects and offered a solid experimental foundation for advancing wireless communication technologies. People living in the Internet of Things age deal with an overwhelming quantity of data and information on a daily basis. As a result, data processing that takes use of intelligence is crucial. Application of deep learning, is expanding rapidly throughout all sectors of society. Eventually, Ohsugi et al. [14] in the area of materials medicine used deep learning to identify RRD in ultra-wide field fundus pictures. For early detection of RRD and avoidance of blindness, they observed that ultra-wide field fundus considerably improved diagnostic

accuracy [14]. It was also suggested that navigation challenges may be solved using the AutoRL approach [28], which combines DRL with gradient-free hyperparametric optimization. However, the RL-required reward function is notoriously hard to build and often fails to ensure optimum performance. To address the whole scope of visual navigation issues, Zhu [12] presented a target-driven DRL (TD-DRL) architecture, which has shown promising results. these methods typically rely on sequence data of trajectory for IL, whereas our proposed method is specifically tailored to improve sample efficiency in visual navigation tasks. In multi-agent deep reinforcement learning (MADRL), agents (or decision makers) work together or against one another in a given setting to attain a common objective. This allows the agents to work together [4] to improve system performance [5]. In recent years, MADRL has made significant progress because of its capacity to address challenging real-world situations, which conventional RL has struggled to address. As interest in MADRL has grown, researchers have conducted several surveys to learn more about it from various angles. The collaborative nature of MADRL is discussed by Oroojlooyjadid and Hajinezhad [9]. Zhang et al. [11] explored the technical issues of MADRL from a mathematical viewpoint, while other research focused on theoretical assessments. This study contributes to the current literature by establishing a taxonomy of MADRL aspects such as aims, characteristics, problems, applications, and performance measurements. Based on the taxonomy, the MADRL algorithms are categorized, examined, and debated, and their unresolved problems are elucidated. Our interpretations from these angles are new to the literature to the best of our knowledge. Focus, method, and intended multi-agent setting are summarized across recent MADRL surveys in Table 1. Information regarding our article is included in the table as well. This paper's overarching goal is to provide an overview of current and emerging MADRL application research fields and to inspire readers to go further into the topic.

There are two main approaches to visual navigation for robots: rule-based and learning-based. Because of our interest in the latter, we will discuss the related field of learning-based navigation, introducing concepts like IL and DRL as they pertain to visual orientation.

An overview of IL can be found in [13], and it is defined as the process of learning a behavior policy from a set of demonstrations. Some IL techniques approach the problem by treating it as a kind of specialized supervised learning, or behavior cloning (BC). NVIDIA [14] introduced an end-to-end visual navigation approach based on BC that accumulates a huge number of expert samples from three cameras and is employed in unmanned driving regions. Visual navigation techniques based on imitation learning have their uses, however they are vulnerable to overfitting when employing just IL.

Due to its capacity to adapt to its surroundings, RL, and specifically DRL, has lately found use in robot navigation challenges. The deep Q-learning algorithm (DQN) introduced by DeepMind is an example of a common DRL approach; it has allowed robots to learn human-level control strategies[19]. Many techniques for enhancing the DQN network model have now been described in rapid succession, with promising outcomes in a variety of contexts [20–23]. DDPG [24] and the A3C [25] are two further examples of policy gradient-based approaches. Using the MazeBase gridworld as a testing ground, the authors of [27] offer a feedforward architecture that is adaptable to shifting incentives and is trained on 3D VizDoom to acquire a deep successor representation. It was also suggested that navigation challenges may be solved using the AutoRL approach [28], which combines DRL with gradient-free hyperparametric optimization. However, the RL-required reward function is notoriously hard to build and often fails to ensure optimum performance. To address the whole scope of visual navigation issues, Zhu [12] presented a target-driven DRL (TD-DRL) architecture, which has shown promising results. The authors in [29] highlighted nine difficulties in RL, one of which being DRL's low-sample efficiency, which

means it requires large samples to train and is hence unsuitable for many real-world systems and tasks. While methods combining DRL and IL have been proposed for a wide variety of tasks, including grasping [30], MuJoCo [31, 32], and traffic control [33], the sequence data of a trajectory is required for these methods, whereas the sample efficiency for visual navigation tasks is the focus of our proposed method.

One possibility is curiosity-driven exploration35, which draws its inspiration from biological systems and uses intrinsic motivation to steer discoveries. Several theorists from other fields have suggested the following pattern of intrinsic motivation: novelty[38], surprise[37], and empowerment[36]. Based on the evidence presented by the novelty hypothesis, we design the reward function to encourage the animal to seek out unique experiences. In addition, our reward function includes not one but two distinct reward categories, both of which are related to episodic memory.

Similar models, like ours, infer the uniqueness of a state-action combination from how often the agent accesses that pair. Our reward function contains a count-based first component. Using ow(owW) waypoints and the TC-network TC:SW to discretize the state space is fundamental to our strategy. States are given waypoints so that their occurrence rates may be calculated. When a fresh observation is made in a region of the environment that has not yet been mapped, however, how will the reward be calculated if the mapping does not exist? Now let's go on to the next question we'll be asking. The idea may be formalized by rewarding researchers for gathering data in parts of the environment that have not yet been studied. A waypoints buffer is useful because it can recall past events and be maintained up to date when new areas of the globe are explored.

V. Research Methodology

REINFORCEMENT LEARNING FOUNDATION

As was previously said, both the value-based approach and the policy-based approach aim to get policies, but they go about doing it in different ways, each with their own set of benefits and drawbacks. The actor-critic (AC)32 algorithm is developed to incorporate the benefits of both approaches. Advantage estimate A(st,at)=Q(st,a)V(st) is used to scale the policy gradient, where the actor and critic are represented by policy and value function V(st), respectively. Therefore, universities are increasingly emphasizing the need of designing and implementing strategies to make the most efficient use of spectrum resources for wireless communication. The government has done a good job allocating the electromagnetic spectrum resources used by radio, which are among the most important components of the wireless communication network. Congestion has not been resolved, even with the proliferation of IoT terminals. This does not, however, mean that there is nothing left. Research shows that in certain parts of the United States, just 15% of available spectrum is being used. This analysis presents compelling evidence that many countries are failing to fully take use of their current national spectrum resources [5]. The National Spectrum Supervision Bureau has thus addressed this issue. authority to use the given frequency range through the initial electrostatic allocation mechanism is restricted to those who have been granted such authority. Due to dynamic spectrum resource allocation, licensed and unlicensed users may live harmoniously in the same frequency band. Reinforcement learning is a promising new approach that shows promise in solving difficult problems involving the allocation of dynamic resources [7]. As interest in deep learning algorithms grows, the concept of DRL has been introduced to the field of computer science. DRL can train the resultant agents to learn the behaviors in photographs or videos, which is not achievable with other methods because to the high-dimensionality of the raw data [8,9]. DRL is

distinguished by two key characteristics. The first is that it seeks to enhance the system over the long run rather than only in the near term. The second is that exploration and decision-making may be performed optimally without any prior knowledge of the environment, leading to near-optimal outcomes [10]. That's why it's so important for DRL to be integrated into wireless communication systems. In conclusion. Fig. 3.1 is a flowchart depicting the iterative optimization process that is the AC algorithm.

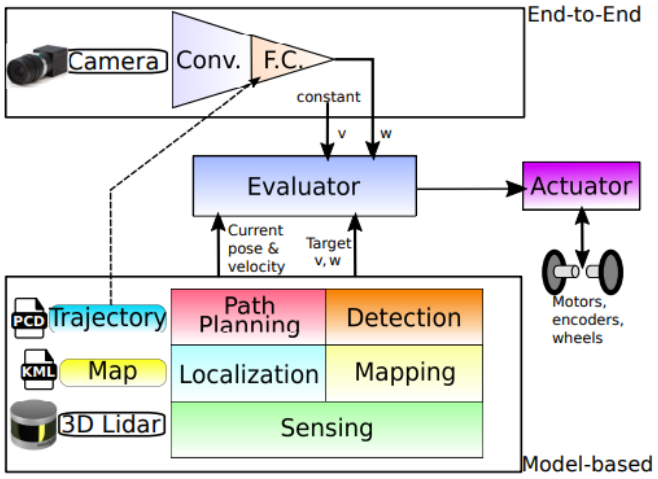

Fig. 3.1. : Implemented navigation system

Go to: 2. Research in Existing Works The development of wireless communication systems has been accelerated by the rise of mobile network applications. There have been a plethora of scholarly studies conducted on its efficacy and ways to improve it. In a radio frequency sensitive or safety-required environment, Huang et al. [11] proposed interconnectivity as a future air-ground-sea integrated communication architecture. Additionally, three simple network processes were created for evaluating the performance of these systems. Lastly, in five sample air quality situations, good transmission performance was recorded between VLC at different speeds and FSO [11], proving the practicality of this hybrid optical wireless network. Nayak et al. [12] simplified the process of controlling wave speeds, amplitudes, and orientations, which enhanced radiation quality at the receiver end. Incorporating the current state of the art in telecommunications frameworks and radio lines, certain designs of the receiving equipment were envisioned. Sopara et al. [13] compared a number of different energy-efficient communication techniques for WSNs and came to the conclusion that the approach they proposed had the greatest impact on reducing energy consumption and provided a strong experimental basis for developing wireless communication technologies. In today's Internet of Things era, information and data may easily become overwhelming for the average person. Therefore, it is essential to use intelligence in the data processing process. Deep learning, which is used for the intelligent extraction of information features, is quickly increasing across all societal domains. One group eventually employed deep learning to identify RRD in ultra-wide field fundus images was the materials medicine group led by Ohsugi et al. [14]. They found that ultra-wide field fundus significantly increased diagnostic accuracy for early identification of RRD and prevention of blindness [14]. The incentive function necessary for RL, however, is notoriously difficult to construct and typically fails to guarantee optimal performance. Zhu [12] introduced a target-driven DRL (TD-DRL) architecture that has demonstrated promising results in addressing the whole spectrum of visual navigation difficulties. Its application in many tasks is limited due to its low-sample efficiency, which means it takes large samples to train. Our proposed method is tailored to improve sample efficiency in visual navigation tasks, while other methods combining DRL and IL have been proposed for tasks such as grasping [30], the MuJoCo task [31, 32], and traffic control [33]. To achieve a goal in multi-agent deep reinforcement learning (MADRL), agents (or decision makers) may cooperate or compete with one another in a specific environment. This permits the agents to coordinate [4] and boost [5] overall system efficiency. MADRL has made great strides in recent years because to its ability to tackle difficult real-world scenarios, something that traditional RL has had trouble doing. Researchers have performed many surveys to get a better understanding

of MADRL from a variety of perspectives as interest in the topic has increased. In [9], Oroojlooyjadid and Hajinezhad highlight the cooperative aspect of MADRL. While previous studies concentrated on theoretical evaluations, Zhang et al. [11] investigated the mathematical underpinnings of MADRL's technological difficulties. By creating a classification system for MADRL features such goals, characteristics, challenges, applications, and performance metrics, that we bring to the literature with our interpretations. Table 1 summarizes the focus, methodology, and intended multi-agent context of current MADRL surveys. The table also contains data relevant to the article we wrote. The primary purpose of this work is to introduce readers to existing and potential MADRL application research topics and to encourage their further exploration.

When it comes to visual navigation, robots may use either a rule-based or a learning-based approach. Our focus on the latter motivates us to talk about learning-based navigation, where we will introduce ideas like internal locus (IL) and dynamic representation learning (DRL) as they apply to visual orientation.

According to [13], some IL methods tackle the issue by framing it as a subset of supervised learning, often known as behavior cloning (BC). NVIDIA [14] created a full-stack BC-based visual navigation method that uses three cameras to collect a large number of expert samples and is used in autonomous driving environments. To predict both the vehicle's speed and its steering angle, Yang [15] proposed an IL-based multi-task architecture that leverages image sequences. To get around the fact that most existing IL methods need a well calibrated actuation setup in order to train on a dataset, Xu [18] presented a novel framework for doing so. While IL-based visual navigation systems have their merits, they are susceptible to overfitting if just IL is used.

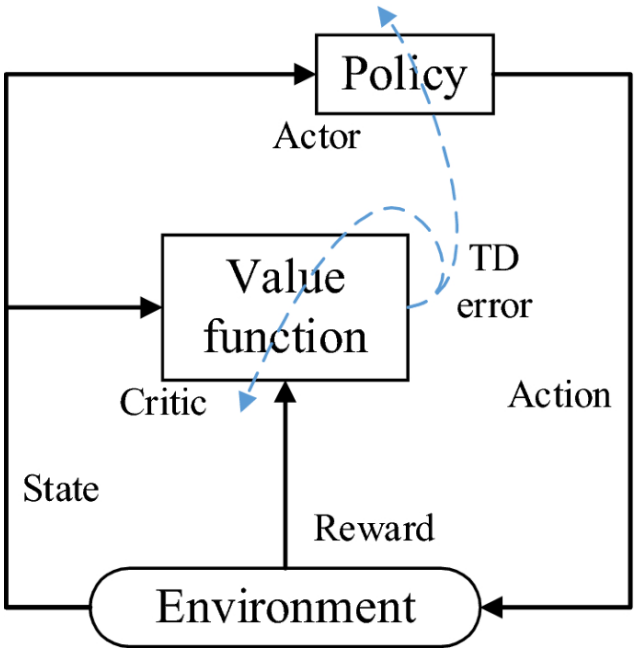

Fig. 3.2: Flowchart of the AC algorithm

VI. Experiment and Result

4.1 EXPERIMENT SETUP & ENVIRONMENT

Here, we compare our technique to relevant baselines and assess its performance on exploration and goal-attainment tasks. In DMlab44, we put our method to the test against relevant baselines throughout a number of mazes, and Fig. 1 depicts an agent making its way through the environment to reach its destination. 12. The agent has first-person perception in this 3D simulated world, along with access to extra environmental data like inertial and local depth. Six distinct actions—forward/backward, left/right, and left/right + forward—make up the action space, which is discrete while yet allowing for precise control. Reaching the apple (+1 point) and the objective (+10 points) in the environment grants extrinsic rewards, and the environment refreshes at a frame rate of 60 frames per second. If the objective is completed, the agent is reborn in a different initial location, and the episode will continue for a predetermined length of time. Ubuntu 18.04 is used as the operating system, and a DELL T7920 workstation equipped with 64 GB of RAM, an Intel Xeon Gold 5118 processor, and two Nvidia RTX 2080TI graphics cards will serve as the

hardware environment for this experiment. Python was used for all of the experiment's programming needs.

(a) (b) (c) (d)

Fig. 4.1: Simulation environment. (a) Go forwards. (b) Apple. (c) Goal. (d) Door."

4.2 PRINCIPLES

We contrasted our approach to a control group whose agents were also given an innate drive to explore in a series of studies in which agents were instructed to create exploration behavior. Trust region policy optimization (TRPO) [45], the simplest RL method, uses the heuristic greedy way of promoting intrepid discovery. We then utilize VIME[19], a comparative tool based on a Bayesian neural network (BNN), to detect environmental changes in real time and derive an exploration strategy by optimizing the information obtained from these observations. Third, we have EX2[20], which is a classifier-centric baseline that investigates novelty detection using just a discriminatively trained exemplar model. Finally, we also duplicate the current state-of-the-art curiosity approach ICM[21], as a sanity check. We use three DRL-enabled models as our starting points for achieving-goal experiments. There is the previously described Nav A3C3, the popular feedforward model DQN[29], and the recurrent model Deep Recurrent Q Network (DRQN)[46]. Experiments also report on a more advanced form of Nav A3C called Nav A3C + D2L, which learns supplementary tasks in tandem with the primary aim.

4.3 METHOD APPLICATION

Here we discuss the architecture of our learning model. In the first convolutional layer, 16 feature maps and 44 stride filters are used. In the second convolutional layer, 32 feature maps and 44 stride filters are used. This is followed by a 256-unit fully-connected layer, and finally a ReLU nonlinearity unit is placed after all three layers. The CNN-encoded observations, actions, and rewards are then fed into a 256-unit long short-term memory (LSTM) layer, which produces linear projections for the policy and value function. Each of the TC-network and L-network's inputs is a 512-dimensional feature vector generated by the ResNet-18 encoder from two observations. The TC-network predicts if the two observations are next to one other by first concatenating these characteristics and then placing them in a fully connected network with 4 hidden layers, each with 512 units and a ReLU nonlinearity unit. Similarly, the L-network takes these properties and processes them in parallel. All of the agent's possible actions are represented by 6 outputs from the softmax layer, which is part of the fully-connected portion.

4.4 HYPERPARAMETERS

For this research, we employ the widely-adopted A3C algorithm as our foundational RL strategy with a sample size of 84 x 84. Input data consists of RGB images captured every three frames, with each action repeated four times. Eight employees, each with their own decentralized RMSProp, labor in tandem to shape their surroundings. A log-uniform distribution between 0.0001 and 0.005 is used to sample the learning rates, while a log-uniform distribution between 0.0005 and 0.01 is used to sample the entropy costs. All training data was created by the agents themselves, and their only inputs are two RGB pictures with a resolution of 160x120 pixels to the TC-network and L-network, respectively.

4.4.1 Experiment to Determine Parameter Values

Although we're interested in agents that can explore and encode the environment on their own accord, we can't evaluate our method's efficacy until we've determined a few key parameters. In the labyrinth shown in Fig. 1, we isolate the parameters relevant to the training details of the TC-network and L-network,

as well as the main elements of the reward function. 13. Fig. Location for Choosing Parameters (Section 4.2).

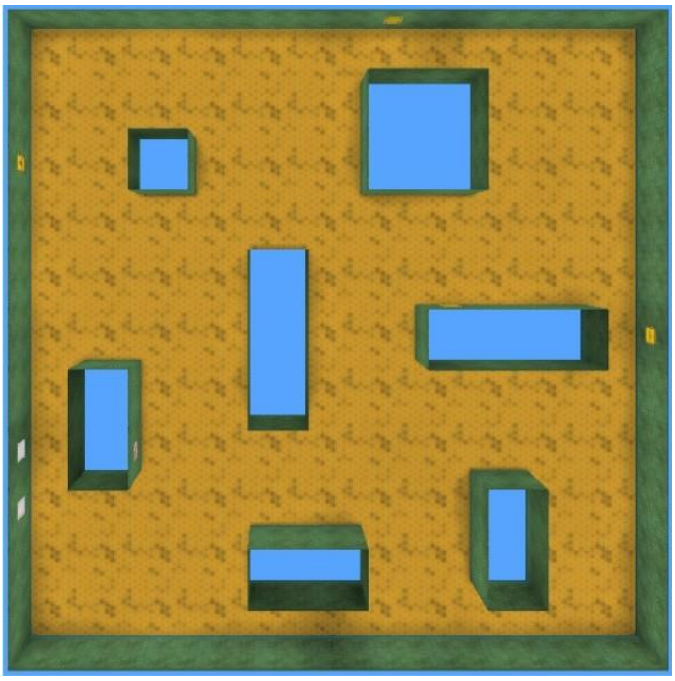

Fig. 4.2: Parameter selection environment.

4.4.2 Parameter for Separating Samples

As with the L-network's training samples, the TC-network's threshold k is used to categorize training samples as either positive or negative. Therefore, we undertake an experiment where we illustrate the consequences of altering k from 1 to 10. The percentage of waypoints is determined using the associated TC-network and 30 random observation sequences, with training results from both the TC-network and L-network (1.5 M interaction quantity) averaged across the top 5 random hyperparameters. As can be seen in Table 4.1, the difference between positive and negative samples is strongly correlated with the training impact of the TC-network. Initial TC-network accuracy suffers from sample-to-sample consistency issues. Then, when k grows, the threshold rises, but the accuracy drops below a certain point once again.

Table 4.1 The experimental results of the sample separation parameter.

Threshold k	TC-network (%)	L-network (%)	Waypoints proportion (%)
1	88.68	95.74	31.25
2	91.53	95.26	22.34
3	93.06	93.42	16.27
4	92.32	91.04	12.51
5	90.87	86.35	11.63
7	86.59	80.73	12.48
10	81.93	75.68	12.65

The performance of the L-network steadily declines with increasing k, particularly when k>4. This is in stark contrast to the TC-network. Finally, the number of waypoints reduces with increasing k, however as we've indicated, the predictive capability of the TC-network approaches a bottleneck when k is bigger than certain value, causing an increase in the number of waypoints. Sample separation parameter testing results are listed in Table 4.1.

4.4.3 Parameter of interaction volume

Pretraining parameters include not just the threshold k but also the degree of environmental interaction. Our approach divides the sample's difficulty into two stages: pre- and post-learning. The sample size does not affect the exploratory behavior of online learning, but the number of samples does affect the pretraining impact. The top 5 random hyperparameters are aggregated to highlight the correlation between interaction volume and network performance in Table 2.

Table 4.2: The experimental results of the interaction volume parameter.

Sample size	TC-network (%)	L-network (%)
300 K	80.35	82.91
500 K	82.42	86.57
1 M	87.95	90.83
2.5 M	92.63	93.78
5 M	91.02	93.94

Table 2 demonstrates that as the training data is increased, the TC-network's accuracy improves, but accuracy declines after overfitting. Similarly, the L-network's prediction accuracy improves with an increase in interaction volume, however the rate of improvement slows down with time.

As shown in Table 4.2, the interaction volume parameter was measured experimentally. In conclusion, both the TC-network and the L-network are capable of learning effective controllers from the trajectories of randomly behaving agents, and can then apply these controllers to problems in visual perception and local navigation. This lack of generalizability is a consequence of the fact that all samples in the pretraining phase come from the same context. Therefore, we gather data from new locations to train these networks twice throughout the future exploration.

4.4.4 Parameter of the reward function

Our reward function is a hybrid one, consisting of both traditional and unique incentives. We demonstrate two main results—the episode reward (novelty rewards achieved by the agent within 1800 time steps) and the number of interactions required to encode the environment—to compare the effects of different parameter sets that are set +1 and sampled within the same interval (0.1). The findings are presented in Fig. by taking an average across the top 5 random hyperparameters. 14 after data standardization (using the minimum value as the standard). For example, The agent may develop a wide range of exploratory behaviors by depending on a novelty reward of the form,=(0.0,1.0) or,=(1.0,0.0). They need more contact with the environment to encode it, but their exploration is less efficient than agents that employ both novelty incentives. The experimental findings may also be understood in terms of the components of the reward function. First, we use a count-based method to calculate novelty rewards for previously explored environments and push the agent toward infrequently visited waypoints; second, we use a temporal distance method to calculate novelty rewards for previously unexplored state spaces and try to nudge the agent toward more remote locations. These two benefits might serve as excellent indicators of where to go next in your investigation. The parameter of the reward function, as determined experimentally, is shown in Figure 4.3. In the next experiment, we choose the agent equipped with the parameter sets,=(0.2,0.8), since it demonstrates the highest exploration efficiency and needs the least amount of human intervention while encoding the surroundings. In addition, at the fine-tuning step, agents no longer behave randomly but instead learn the exploration policy in the environment.

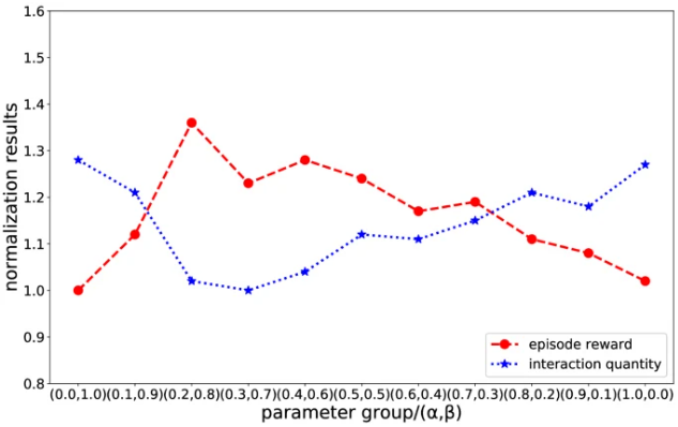

Fig. 4.3: Experimental results on the reward function parameter.

4.5 EXPERIMENTAL EXPLORATION METHOD

The experiment's goal is to demonstrate the impact of various learning strategies and training patterns on encoding efficiency by providing a quantitative evaluation of their exploration performance. Fig. displays the test setups. 15; Maze-1 has three routes of varying lengths, while Maze-2 has a center corridor and six limbs; both are based on tests with rodents and their spatial cognition. The third maze is a standard kind, with a variety of traps and many possible exits. These mazes do not have any external incentives (such a destination or a fruit, for example).

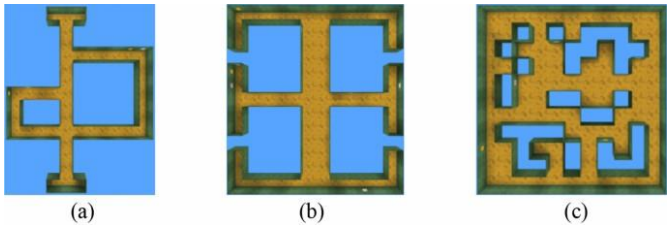

Fig. 4.4: Test maze seen from above (a) Maze-1. (b) Maze-2. (c) Maze-3.

An episode-wide reward based on the agent's region of exploration is used to compare the effectiveness of the various approaches. An episodic reward/training step diagram is used to explain how the agent learns to

successfully complete a task under a time constraint of 7200 steps (2 minutes). The agent dies at the conclusion of each episode and is reborn in a different area, forcing him or her to start exploring from scratch.

4.1 Adaptive tuning using internal drives

Figure below demonstrates that when ICM and our method are trained in conjunction with the fine-tuning method, random exploration can be terminated; However, there is an interesting trend in the learning curve. The exploration strategy can be learned rapidly from start because of Maze-1's straightforward design. Some mismatches occur early on in training when using the fine-tuning strategy, since wall walking and obstacle avoidance are not explicitly taught. While the fine-tuning strategy greatly improves ICM's training efficiency and helps the policy converge faster in Maze-2 than when starting from scratch, its overall contribution to our method is little. Maze-3 provides a clearer demonstration of the importance of fine-tuning, which is reflected in two main ways: the improvement in the ICM module's performance, where exploration efficiency rises once more following the first policy stabilization, and the use of fine-tuning to further decrease the required number of interactions in the maze's encoding. Precision tuning with intrinsic motivation: experimental findings, number

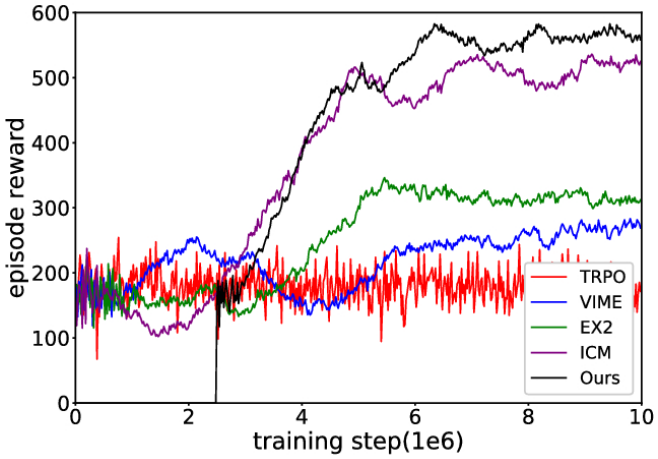

Fig. 4.6: Results from the Maze-1 experiment, The results of the second version of the Maze experiment). The outcomes of the Maze-3 study, in (c). Importantly, the experiments demonstrate that the fine-tuning method does not always help and might sometimes hinder learning, particularly in easy settings. Fine-tuning, on the other hand, excels in complicated contexts because it can take a previously-trained policy as input and utilize it to help the agent better adapt to the new setting.

4.2 Using monetary incentives for fine-tuning

It is important for the reader to bear in mind that the current experiment employed extrinsic incentives as drivers to direct investigation. The placements of the apples (Fig. 12c, value + 1) and the objectives (Fig. 12b, value + 10) were consistent within an episode but shuffled between episodes to provide extrinsic incentives. The agent's success in each way is still judged by the uniform reward it gets during an episode (the area traversed was estimated by the count-based approach), but once the objective is accomplished, the agent respawns to a new starting place and must explore the labyrinth again. Particular findings are highlighted in Fig. See also Table 6 (average outcomes over the top 5 best performances in the learning process) and Table 18 (average results over the top 5 random hyperparameters).

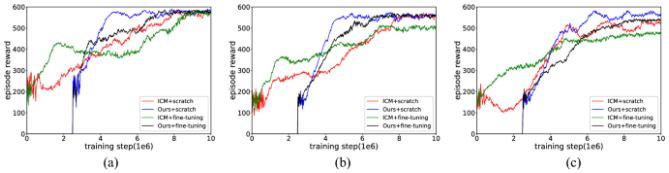

Fig. 4.7: The experimental results of fine-tuning with extrinsic reward. (a) Maze-1 experiment results. (b) Maze-2 experiment results. (c) Maze-3 experiment results.

Using extrinsic incentives to perform fine-tuning has a greater detrimental impact than the previous strategy, since it not only slows down training but also muddles the distinction between exploration and navigation. Because reaching the objective is often seen as gaining a substantial intrinsic reward during exploration, the state holding the extrinsic appealing reward is desirable, and the agent desires to attain it consistently. This performance degradation is the core reason of the policy shift. Furthermore, it seems that the aim of fine-

tuning is to discover the goal rather than to explore the environment, since the agent is reset to a new starting location upon achieving the goal. Because of this, agents in Maze-1 and Maze-2 may earn substantial rewards rapidly during the first stages of training, but they still need further involvement to finish the investigation. The worst of it takes place in Maze-3, where the whole area is too large to explore in a single episode. Extrinsic reward fine-tuning experiment findings are shown in Figure 4.7. Results from the Maze-1 experiment, (a). The results of the second version of the Maze experiment (part b). The outcomes of the Maze-3 study, in (c). Consequently, agents are more effectively motivated to act in accordance with their goals when they are motivated by extrinsic incentives, which take the form of discrete objects in the environment. This Noisy-TV trial is an experiment As shown above, the ICM approach beats the other baselines and comes close to matching the performance of our approach in the first two test mazes. This prediction-based curiosity strategy has yet to solve the couch potato issue, as shown by the noisy-TV experiment. This experiment is designed to help us determine whether our approach, which uses agents' observation and memory to direct exploration, is more resilient to stochastic objects. This is how we really did our loud TV experiment. The TV was always shown on the agent's primary display, and its location within an episode was always the same but changed randomly between episodes. Each time the agent does an action, the TV screen displays an unrelated, 21x21-pixel picture in one of the four quadrants the agent is monitoring, with each pixel evenly sampled from the range [0,255]. The data from the experiments shown in Fig. As can be seen in figure (average results over the top 5 best performances in the learning process) and figure (average results over the top 5 random hyperparameters), the addition of the randomness source has a negative impact on the performance of both the ICM and our technique. While the fine-tuning technique may ease exploration to some amount and help ICM keep learning when starting from fresh, the resultant policy is still unsatisfying. It's clear that not all of the state space can be modeled, such the motion of leaves in the wind or the noise in the television set. The ICM method gets sucked into the curiosity trap and devolves into undesirable behavior because its prediction mistakes stay high and exhibit an attractive allure to the agent. The agent's continued interest in the rustling of the leaves and the flickering of the TV screen serves no useful purpose. That's why we used memory to look for what we were curious about rather than making any assumptions. The agent gets over the couch potato issue by making use of historical similarities to eliminate its curiosity for seemingly unrelated items. Furthermore, in our future study, we will elaborate on additional potential origins of environmental unpredictability.

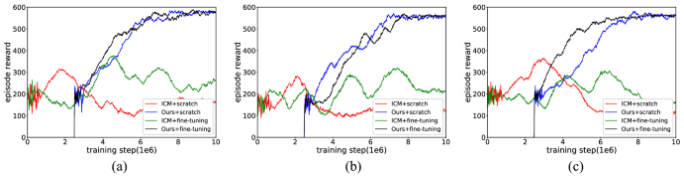

Fig. 4.8: The experimental results of Noisy-TV. Results from the Maze-1 experiment, (a). The results of the second version of the Maze experiment (part b). The outcomes of the Maze-3 study, in (c).

4.6 Purposeful Experiment

We evaluate our method against a variety of DRL-supported navigation models (DQN, DRQN, Nav A3C, and Nav A3C + D2L) on a variety of navigation tasks. Exploration-based topological learning may be immediately applied to the task at hand. We trained these models using the same training procedures as the exploratory policy and then saved the trained models as baselines so that we could fairly compare them. These models were trained and additional experiments aimed at achieving the goals were conducted using the environment described in the Fine-tuning with extrinsic rewards section. We compared the percentage of time spent moving toward the objective at various time steps to the extrinsic reward gained by the agent throughout the course of an episode (5000 time steps). The Static Maze Procedure In this test, we

guarantee the correctness of all operations by performing them in a static environment where the only variables are the locations of the agent and the target. Once an episode has begun, the only remaining computational activity is the agent's self-localization, since target localization is done only once. For example, In order to make up for the DQN's inability to recall its previous visits to the target, the DRQN model incorporates a long short-term memory (LSTM) to speed up subsequent visits to the site during subsequent episodes. The Nav A3C model incorporates more data (relative agent velocity, action, and reward) to further enhance navigational efficacy. In the first two test mazes, its performance is comparable to ours when paired with a ground-truth depth map and loop closure. Maze-3 is different, however, as its structure is more complex, depth information contributes less to action selection, and Nav A3C + D2L's navigation efficiency shows a declining trend, all while our method still maintains an efficient navigation policy. The experimental outcomes of goal attainment in stationary mazes are shown in Figure 20. Results from the Maze-1 experiment, (a). The results of the second version of the Maze experiment (part b). The outcomes of the Maze-3 study, in (c). To begin, navigation relies heavily on the memory function, and its absence greatly affects the frequency with which an agent accomplishes a goal and the rewards it obtains during an episode. This finding is most obvious in memory-intensive settings like Maze-2, where the DRQN model offers nearly twice as much reward as the DQN model. Second, although the DRQN model is superior than the DQN model in all test mazes, it still takes a long time for the DRQN model to reach the endpoint. We attribute this to the agent's lack of observant clarity, which calls for more inputs and depth information to solidify the mapping link between states and actions. Finally, the results show that the map-less strategy is a viable alternative to the map-based approach in basic situations, but that its performance degrades in Maze-3, suggesting that the map-based approach is better able to deal with the increasing complexity of the state space. A Study of Dynamic Blocking Here, we compare and contrast how well these various approaches do when faced with dynamic obstructions in state space. The test environments are replicas of Maze-1 with the addition of obstacles at the places shown in Fig. 21 (the labyrinth still has accessible routes). Next, we perform tests in

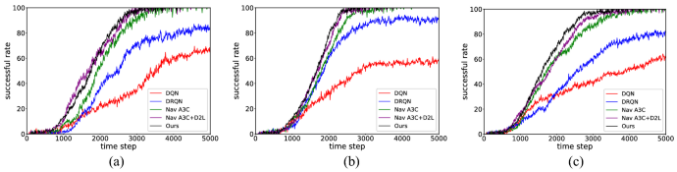

Fig. 4.9: The goal reaching experimental results in static mazes.

Fig. 4.9's surroundings using the space topological cognition and navigation model we developed in Maze-1. 21. In fact, if you look at Fig. 22 (the findings are an average of 30 iterations), the navigation efficiency of all techniques is impacted by the addition of dynamic obstructions to Maze-1. However, as compared to map-based approaches, the effects of obstructions are catastrophic for map-less approaches. Both DQN and DRQN rely heavily on reactive behavior while searching for the objective. Consequently, the success rate plummets if the agent is blocked from reaching the target until the conclusion of the episode. Even if they perform very well in the non-blocking state space, Nav A3C and Nav A3C + D2L fail to ensure arrival at the destination when the blockage emerges at point A or B. Since there are now two obstructions in the path between the agent and the target, the performance drop is more pronounced in the final environment. Our method also has a lower success rate since it takes more time steps to redirect the agent around the obstacle.

4.7 Conclusion

As obstacles are introduced, the success rate and reward achieved by all approaches declines, but the approaches without maps suffer the most. When barriers are put at either site A or B, the incentives for DQN and DRQN drop by about 30%, and they drop by

nearly 50% when there are two barriers in the environment. Furthermore, the presence of blockages reveals the limitations of Nav A3C and Nav A3C + D2L, particularly in the case where the agent and the target are placed at both ends of the blockage, where the agents constantly attempt to break through the blockage, significantly reducing the frequency with which they reach the target. Our method assures the agent can identify a viable route to the objective by relying on the agent's recollection of the environment structure and a dynamic path planning mechanism. By updating the topological memory, the bot may navigate around obstacles A and B. However, our method typically needs the whole of the episode to achieve the objective, which results in fewer payouts owing to the additional navigation distance.

VII. CONCLUSION

The capacity to navigate visually is crucial for an autonomous agent to perform a broad variety of tasks in challenging environments. Using information gathered from its own internal visual sensors, an agent is able to understand its immediate surroundings and safely navigate to a predefined location. There are two main points to consider. The first step is for the agent to assess the present observation and infer the aspects that are crucial to the end result. Second, the agent must be aware of how its navigational behaviors shape its perceptions of its surroundings. In this research, we offer a new navigation architecture that combines topological knowledge of space with the discovery of intrinsic motivation. Our method is two-pronged, with the first part of the method tasked with exploring the environment and the second part tasked with encoding it. Analyses and findings from experiments demonstrate the significance of reward function and training patterns in the acquisition of an exploratory stance. We also looked at how space topologically cognizant bots fared in both static and semi-dynamic settings when it came to navigation. The cognitive mechanisms of animals serve as inspiration for our method because of their ability to simultaneously investigate and encode the structure of their surroundings. We utilized DRL as the foundational learning framework and gave AI agents the freedom to design their own incentives in order to achieve spontaneous exploration from unprocessed visual inputs. Unlike prediction-based exploration approaches, our reward function is calculated using episode memory and has two distinct sorts of non-traditional rewards. Exploration waypoints are a common feature of both space topological cognition and episode memory. Such spatial cognition may be employed as a planning module for the navigation system and to progressively cover the surroundings by merging exploration sequences.

Our method efficiently learns an exploration strategy inside the end-to-end DRL framework; but, the 1-layer LSTM's limited memory might be overstretched in very vast settings. Increasing the LSTM size or including external memory will be crucial in the future if we want to make our learnt model more powerful. In addition, when more ground is covered, our spatial awareness expands proportionally. Again, this may be problematic when trying to find your way across really expansive areas. One approach is secondary sampling, in which only the most useful or distinguishable landmarks are retained. Finally, we foresee further development in taking our technology outside and contrasting it with vision-based SLAM techniques.

By switching from our present four-layer fully-connected model to a pre-trained ResNet, we can significantly increase our accuracy. Success rates may be increased by retraining segments of the ResNet model, introducing new Dagger, and expanding training data to include additional scenes and targets.

VIII. REFERENCES

[1]. Li Y, “Deep reinforcement learning”, In: ICASSP 2018—2018 IEEE international conference on acoustics, speech and signal processing

(ICASSP), Calgary, AB, Canada, 15–20, April 2018.

[2]. Sun ZJ, Xue L, Xu YM, et al, "Overview of deep learning", Appl Res Comput 2012, 12, pp. 2806–2810.

[3]. Sutton RS and Barto AG, "Reinforcement learning: an introduction", IEEE Transactions on Neural Networks, 2005.

[4]. Hosu I-A and Rebedea T, "Playing Atari games with deep reinforcement learning and human checkpoint replay", 2016. ArXiv, abs/1607.05077.

[5]. Lillicrap TP, Hunt JJ, Pritzel A, et al, "Continuous control with deep reinforcement learning", Comput Sci 2015, 8(6): A187.

[6]. Caicedo JC and Lazebnik S, "Active object localization with deep reinforcement learning", In: Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 2015, pp. 2488–2496.

[7]. Meganathan RR, Kasi AA, and Jagannath S, "Computer vision based novel steering angle calculation for autonomous vehicles", In: IEEE international conference on robotic computing, Laguna Hills, CA, USA, 31 January–2 February, 2018.

[8]. Gupta S, Tolani V, Davidson J, et al, "Cognitive mapping and planning for visual navigation", In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 7272–7281.

[9]. Zhu Y, Mottaghi R, Kolve E, et al, "Target-driven visual navigation in indoor scenes using deep reinforcement learning", In: 2017 IEEE international conference on robotics and automation (ICRA), Stockholm, 16–21 March 2016, pp. 3357–3364.

[10]. S. Amarjyoti, "Deep reinforcement learning for robotic manipulation-the state of the art", Bull. Transilv. Univ. Braşov, vol. 10, no. 2, 2017.

[11]. A. V. Bernstein, E. Burnaev, and O. Kachan, "Reinforcement learning for computer vision and robot navigation", in Proc. International Conference on Machine Learning and Data Mining in Pattern Recognition, 2018, pp. 258-272: Springer.

[12]. V. Matt and N. Aran, "Deep reinforcement learning approach to autonomous driving", ed: arXiv, 2017.

[13]. X. Da and J. Grizzle, "Combining trajectory optimization, supervised machine learning, and model structure for mitigating the curse of dimensionality in the control of bipedal robots", Int. J. Rob. Res., vol. 38, no. 9, pp. 1063–1097, 2019.

[14]. I. Zamora, N. G. Lopez, V. M. Vilches, and A. H. Cordero, "Extending the openai gym for robotics: A toolkit for reinforcement learning using ros and gazebo", arXiv preprint arXiv:1608.05742, 2016.

[15]. H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning", The International Journal of Robotics Research, vol. 35, no. 11, pp. 1289-1307, 2016.

[16]. L. Tai and M. Liu, "A robot exploration strategy based on qlearning network", in Proc. 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR), 2016, pp. 57-62.

[17]. L. Tai, G. Paolo, and M. Liu, "Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation", in Proc. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 31-36.

[18]. Mnih. V, Kavukcuoglu. K, Silver. D, Rusu. A.A, Veness. J, Bellemare. M.G, Graves. A, Riedmiller. M, Fidjeland. A.K, Ostrovski. G, et al, "Human-level control through deep reinforcement learning", Nature 2015, pp. 518-529.

[19]. Van Hasselt. H, Guez. A, Silver. D, "Deep Reinforcement Learning with Double Q-

Learning", AAAI: Phoenix, AZ, USA, 2016; Volume 2, p. 5.

[20]. Wang. Z, Schaul. T, Hessel. M, Van Hasselt. H, Lanctot. M, De Freitas. N, "Dueling network architectures for deep reinforcement learning" arXiv 2015 arXiv:1511.06581. Avaliable online: https://arxiv.org/pdf/1511. 06581.pdf (accessed on 12 September 2018).

[21]. Diederik P, Kingma and Jimmy Ba, "Adam: A method for stochastic optimization", CoRR, abs/1412.6980, 2015.

[22]. Dr.V.V.Narendra Kumar, T.Satish Kumar, "Smarter Artificial Intelligence with Deep Learning" SSRG International Journal of Computer Science and Engineering Vol-5,Iss-6,2018. A

[23]. Oudeyer, P.Y. Computational theories of curiosity-driven learning. arXiv:1802.10546 (2018).

[24]. Tolman, E. C. Cognitive maps in rats and men. Psychol. Rev. 55(4), 189–208 (1948).

[25]. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A.J., Deil, M., Goroshin, R., Sifre,L., Kavukcuoglu, K., Kumaran, D., & Hadsell, R. Learning to navigate in complex environments. arXiv:1611.03673 (2017).

[26]. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521(7553), 436–444 (2015).

[27]. Oh, J., Chockalingam, V., Singh, S. P., & Lee, H. Control of memory, active perception, and action in Minecraft. arXiv:1605.09128 (2016).

[28]. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., & Farhadi, A. Target-driven visual navigation in indoor scenes using deep reinforcement learning, in 2017 IEEE International Conference on Robotics and Automation (ICRA) 3357–3364 (2016).

[29]. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T.P., Harley, T., Sliver, D., & Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. arXiv:1602.01783 (2016).

[30]. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with LSTM. Neural Comput. 12(10), 2451–2471 (2000).

[31]. Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Sliver, D., & Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. arXiv:1611.05397 (2016).

[32]. Ye, X., Lin, Z., Li. H., Zheng, S., & Yang, Y. Active object perceiver: Recognition-guided policy learning for object searching on mobile robots. arXiv:1807.11174v1 (2018).

[33]. Yang, W., Wang, X., Farhadi, A., Gupta, A., & Mottaghi, R. Visual semantic navigation using scene priors. arXiv:1810.06543 (2018).

[34]. Devo, A., Mezzetti, G., Costante, G., Fravolini, M. L. & Valigi, P. Towards generalization in target-driven visual navigation by using deep reinforcement learning. IEEE Trans. Robot. 36(5), 1546–1561 (2020).

[35]. Berlyne, D. E. Conflict, Arousal and Curiosity 38–54 (McGraw-Hill Book Company, 1960).

[36]. Harlow, F. H. Learning and satiation of response in intrinsically motivated complex puzzle performances by monkeys. J. Comp. Physiol. Psychol. 43, 289–294 (1950).

[37]. Sylva, K., Bruner, J. S., & Jolly, A. Play: Its role in development and evolution 279–292 (Penguin Books Ltd, 2017).

[38]. Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., & Munos, R. Unifying count-based exploration and intrinsic motivation, in NIPS (2016).

[39]. Ostrovski, G., Bellemare, M.G., Oord, A. V. D., & Munos, R. Count-based exploration with neural density models. arXiv:1703.01310 (2017).

[40]. Tang, H., Houthooft, R., Foote, D., Stooke, A., Chen, X., Duan, Y., Schulman, J., Turck, F. D., & Abbeel, P. Exploration: A study of count-based exploration for deep reinforcement learning, in NIPS (2017).

[41]. Houthooft, R., Chen, X., Duan, Y., Schulman, J., Turck, F. D., & Abbeel, P. Vime: Variational information maximizing exploration, in NIPS (2016).

[42]. Fu, J., Co-Reyes, J. D., & Levine, S.: EX2: Exploration with exemplar models for deep reinforcement learning, in NIPS (2017).

[43]. Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. Curiosity-driven exploration by self-supervised prediction. arXiv:1705.05363 (2017).

[44]. Pritzel, A., Uria, B., Srinivasan, S., Puigdomenech, A., Vinyals, O., Hassabis, D., Wierstra, D., & Blundell, C. Neural episode control. arXiv:1703.01988 (2017).

[45]. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., & Levine, S. Time-contrastive network: Self-supervised learning from video. arXiv:1704.06888 (2018).

[46]. Aytar, Y., Pfaff, T., Budden, D., Paine, T. L., & Wang, Z. Playing hard exploration games by watching youtube. arXiv:1805.11592 (2018).

[47]. Cadena, C. et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Trans. Robot. 32(6), 1309–1332 (2016).

[48]. Bhatti, S., Desmaison, A., Miksikm, O., Nardelli, N., Siddharth, N., & Torr, P. H. S. Playing doom with SLAM-augmented deep reinforcement learning. arXiv:1612.00380 (2016).

[49]. Parisotto, E., & Salakhutdinov, R. Neural map: Structured memory for deep reinforcement learning. arXiv:1702.08360 (2017).

[50]. Gupta, S., Tolani, V., Davidson, J., Levine, S., Sukthankar, R., & Malik, J. Cognitive mapping and planning for visual navigation. arXiv:1702.3920 (2019).

[51]. Mnih, V. et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015).

[52]. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. 8, 229–256 (1992).

[53]. Nachum, O., Norouzi, M., Xu, K., & Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. arXiv:1702.08892 (2017).

[54]. Sutton, R. S., & Barto, A. G. Reinforcement learning: An introduction 215–260 (The MIT Press, 1998).

[55]. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. arXiv:1512.03385 (2015).

[56]. Friston, K., Fitzgerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active inference: A process theory. Neural Comput. 29(1), 1–49 (2017).

[57]. Forestier, S., & Oudeyer, P. Y. Modular active curiosity-driven discovery of tool use, in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 3965–3972 (2016).

[58]. Salge, C., Glackin, C. & Polani, D. Changing the environment based on empowerment as intrinsic motivation. Entropy 16(5), 2789–2819 (2014).

[59]. Little, D. Y. & Sommer, F. T. Learning and exploration in action–perception loops. Front. Neural Circuits 7(37), 1–19 (2013).

[60]. Sutton, R. S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming, in Proceedings of the Seventh International Conference on Machine Learning 226–224 (1995).

[61]. Sigaud, O., & Stulp, F. Policy search in continuous action domains: An overview. arXiv:1803.04706 (2018).

[62]. Moser, E. I., Kropff, E. & Moser, M. B. Place cells, grid cells, and the brain's spatial representation system. Annu. Rev. Neurosci. 31, 69–89 (2008).

[63]. Kirichuk, V. S., Kosykh, V. P., Popov, S. A. & Shchikov, V. S. Suppression of a quasi-stationary

background in a sequence of images by means of interframe processing. Optoelectron. Instrument. Data Process. 50(2), 109–117 (2014).

[64]. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).

[65]. Cormen, T. H., Leiserson C. E., Rivest, R. L., & Stein, C. Introduction to Algorithms, 3rd ed, 658–664, 682 (The MIT Press, 2005).

[66]. Beattie, C., Leibo, J.Z., Teplyashin, D., Ward, T., Wainwright, M., Kuttler, H., Lefrancq, A., Green, S., Valdes, V., Sadik, A., Schrittwieser, J., Anderson, K., York, S., Cant, M., Cain, A., Bolton, A., Caffney, S., King, H., Hassabis, D., Legg, S., & Petersen, S. Deepmind lab. arXiv:1612.03801 (2016).

[67]. Schulman, J., Levine, S., Moritz, P., Jordan, M. I., & Abbeel, P. Trust region policy optimization. arXiv:1502.05477 (2017).

[68]. Hausknecht, M., & Stone, P. Deep recurrent Q-learning for partially observable MDPs. arXiv:1507.06527 (2017).

[69]. Kingma, D. P., & Ba, J. Adam: A method for stochastic optimization. arXiv:1412.6980 (2017).

Cite this article as :

Pooja Upadhyay, Dr. Bappaditya Jana, "In Door Target-Driven Visual Robot Navigation Using Deep Reinforcement Learning (DRL) Approaches", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 5, pp.216-235, September-October-2023.