IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 15, 2020, accepted December 24, 2020, date of publication December 28, 2020,
date of current version January 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047845

Communication Emitter Motion Behavior's Cognition Based on Deep Reinforcement Learning

YUFAN JI, JIANG WANG, WEILU WU, LUNWEN WANG, CHUANG PENG, AND HAO SHAO
College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China

Corresponding author: Lunwen Wang (wanglunwenmust@nudt.edu.com)

This work was supported by the National Natural Science Foundation of China under Grant 11975307.

ABSTRACT Considering the successful application of deep reinforcement learning (DRL) on tasks of moving objects, this paper innovatively applies deep deterministic policy gradient algorithm (DDPG) to complete the cognition task on multi-dimension and continuous communication emitter motion behavior. First, we propose a DDPG-based behavior cognition algorithm (DDPG-BC). It chooses direction, velocity, and communication frequency as state space, gains experience from interaction between network and environment and outputs deterministic cognition results. Second, under the condition of sufficient prior information such as geographic information, we further propose a novel algorithm named DDPG-based behavior cognition with Attention algorithm (DDPG+A-BC). It introduces attention mechanism into DDPG-BC which limits exploration scope and the randomness of initial state and improves the exploration efficiency and accuracy. The simulation experiments verify that DDPG-BC and DDPG+A-BC show good cognition ability on two different data set. And the algorithms are all superior to other DRL algorithm and existing cognition method with higher cognition accuracy and less time. In addition, we also discuss the influence of episode, reward function, and added attention mechanism on algorithm performance.

INDEX TERMS Communication emitter, motion behavior cognition, deep reinforcement learning, DDPG, attention mechanism.

I. INTRODUCTION

Nowadays, with the help of various positioning technologies, we can quickly obtain a large amount of moving object data. However, when mining the information carried by the moving objects, simple observation and tracking no longer meet our goals and needs. Instead, we hope to explore what happens behind the movement to enrich the information content of objects for better control and decision support. Therefore, the cognition of people, animals, vehicles and other moving objects has become a research hotspot at present.

The purpose of this paper is to analyze and cognize the motion behavior of communication emitter and its carrier in motion and find out the corresponding possible causes. For example, as shown in FIGURE 1, when communication emitter and its carrier or platform goes through this area, the original plan is to go straight (planned route). But near the interference, communication performance may degenerate. For example, obstacles like buildings will cause the signal propagation refraction and diffraction, or the place like airports will add too much noise to communication channel. The emitter may choose diversion (angle change), acceleration (velocity change), or changing the communication frequency in order to avoid interference and maintain the communication ability (actual route). The cognition process of emitter motion behavior is to discover the corresponding possible causes or determine whether interference or strike has occurred on the premise of mastering the changes of direction, velocity and communication frequency.

The data of communication emitter motion behavior are usually multi-dimension, continuous, and limited. However, recent research on motion behavior cognition prefer to classify discrete behaviors and analyze the category which each behavior belongs to. It is obviously inconsistent with the emitter's characteristics and increases the workload of preprocessing. Zitouni *et al.* [1] proposed a probabilistic formulation of different categories of socio-cognitive crowd behavior and a framework which can be considered as

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Huei Cheng.

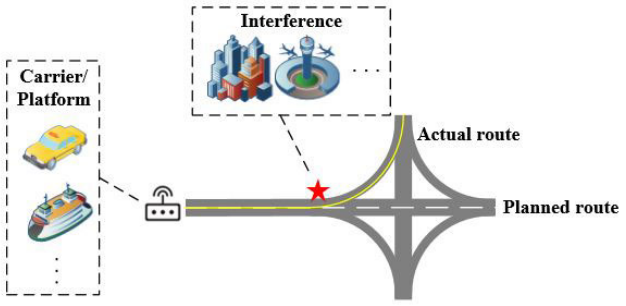


FIGURE 1. Schematic diagram.

a mid-level layer between detection and detailed semantics. But it aimed to evaluate probabilities of various socio-cognition behaviors and compared the model outputs to manually annotate ground-truth data. Goldberg *et al.* [2] used meta-cognitive model (MCM) for future hurricane evacuation with combination of past behavior and subjective confidence. [3] proposed a conceptual framework named experience-oriented intelligent things (EOIT) to extract driving behavior fingerprints which needs enough experience obtained by catching driving data in advance. At the same time, Hornischer *et al.* [4] believed the spatial information is minimal definition of a cognitive map and developed a minimal model of agents to explore environment by means of sampling trajectories. That means the formation of internal cognition are related to the spatial overlap of cognitive maps, so the introduction of geographic information can effectively support the cognitive process.

In process of cognizing communication emitter motion behavior, we hope that: 1) specific cognition results will be obtained directly rather than the probability of different results; 2) experience can be gained after interaction between network and environment of the cognition model without requirement for a lot of subjective experience; 3) the exploration efficiency of the network are able to be improved; 4) the cognitive results will be more objective, only according to the physical parameters of the emitter. By implementing the above effects, the algorithm proposed in this paper not only conforms to the characteristics of emitter motion data, but also can realize the control of cognition with the pursuit of rewards. And DDPG in DRL may become a good choice to solve the problem.

Therefore, this paper proposes a cognition method of communication emitter motion behavior based on DDPG. The main contributions are as follows:

1. Two cognition algorithms of communication emitter behavior – DDPG-BC and DDPG+A-BC. DDPG-BC cognizes communication emitter motion behavior based on DDPG algorithm. With attention mechanism introduced, it further transforms into DDPG+A-BC, which explores in attention position and results in better cognition effect.

2. Verification and related discussion on the performance and effect of DDPG-BC and DDPG + A-BC.

II. RELATED WORK

In order to realize the cognition of communication emitter motion behavior and combine with the information characteristics provided by the moving emitter, this paper aims to explore the autonomous learning ability of DRL algorithm in the cognition of emitter behavior and observe whether the introduction of attention mechanism can help improve the learning efficiency of the network.

A. BEHAVIOR COGNITION

Behavior modelling and activity interpretation are of increasing interest in the information society [5]. The research on behavior cognition mainly focuses on computer science and network and social psychology, and the research targets mainly include human [6], animal [7], traffic [8], [9] and robot [10]. The Google team proposed in 2006 that the motion behavior cognition system should be composed of four modules of "sensor-identification-transformation-controlled system (SITR)" [11]. When the sensor receives the raw data of moving objects, it will classify and process the raw data corresponding to various behaviors, then translate all kinds of data into behaviors, and finally realize the cognition and control of behaviors. Pei *et al.* [5] proposed Context Pyramid when cognize human behavior using smartphone sensors and divided it into six levels: raw sensor data, physical parameter, features/patterns, simple contextual descriptors, activity-level descriptors, and rich context.

The basic idea of motion behavior cognition is that, given a tracked feature or object, its time series should provide a descriptor that can be used in a general cognition framework [12]. Whether it is the known features or the raw physical data to be processed, correct cognition requires that the behavioral parameters we are faced with are sufficiently descriptive and will be a general element when a certain behavior occurs. [13] and [14] described human behavior using Wi-Fi channel state information (CSI) and modelled CSI data based on body movement. With the development of science and technology, it means that, as long as relevant data can be obtained, motion parameters such as velocity and direction and emitter signal parameters such as communication frequency can participate in the target's motion behavior cognition.

B. DEEP REINFORCEMENT LEARNING

It is not difficult to find that the cognition of motion behavior puts forward higher requirements for the selection of features. The features acquired by deep learning (DL) often have certain semantic features and strong discriminative ability, which can more effectively represent the behavior characteristics [15].

The generation and development of reinforcement learning (RL) are inspired by behavioral psychology. States and actions in RL network interact with each other in the environment. However, RL can only deal with low-dimensional state and action space, so the success of deep neural network

on large training data sets motivated the generation of DRL, which can be directly applied on data and process training samples by using stochastic gradient updates [16].

Mnih *et al.* [17] developed a novel agent, deep Q networks (DQN), to create a single algorithm that is able to develop a wide range of competencies on a varied range of challenging tasks. DQN interacts with the environment through a series of observations, actions, and rewards and can be used for RL tasks with discrete action. Actions are selected in a way that maximizes the accumulation of future rewards, and deep neural networks are used to approximate optimal value action functions. Although DQN algorithm has an excellent performance in various applications [18], it still has limitations such as overestimation of the model and inability to handle continuous action problems. Due to the DQN algorithm has difficulty in calculating the probability of each action or the corresponding Q values in large continuous action space, Lillicrap *et al.* [19] proposed DDPG algorithm in 2015 for applying DRL on tasks with continuous action space. DDPG algorithm is a kind of widely used DRL algorithm which can study ''end-to-end'' strategy in higher dimensional, continuous action space [20].

DDPG provides a model-free algorithm based on deterministic policy gradient (DPG), which has both Actor and Critic systems and combines two RL algorithms based on value (such as Q-learning) and action probability (such as policy gradient, PG). In addition to the Actor-Critic framework, DDPG algorithm uses the same learning algorithm, network structure, and hyperparameters as DQN. Hausknecht and Stone [21] focused on using deep neural network in structural (parameterized) continuous action spaces, represented a successful extension of DRL to the class of parameterized action space MDPs and prepared for the learning in the continuous and bound action spaces. Silver *et al.* [22] proposed an off-policy Actor-Critic algorithm that learned a deterministic target policy from an exploratory behavior policy and used DPG for RL algorithm with continuous action. DPG obtains expected gradients of action value by learning approximation of action-value function (Q function) and updates deterministic strategy via chain-rule to make the estimation more effective [23]. While DPG algorithm can solve the problem of high-dimensional continuous action space and combine the advantage of DQN which takes high-dimensional state space as input with an Actor-Critic framework, DDPG algorithm has the ability to handle continuous action control tasks.

C. ATTENTION MECHANISM

Attention mechanism is to select specific inputs which are a methodology derived from human attention. It enables practitioners to adjust attention direction and weight model according to specific task and objects to achieve the goal of reducing sequential computation costs [24]. Attention mechanism realizes via adding attention weight in the hidden layer so that the content that does not conform to the attention model will be weakened or forgotten. Attention mechanism mainly applies to learning weight distribution and task focus. Task focus is to design different network structures (or branches) through task decomposition to reduce the training difficulty of the original task. Learning weight distribution is to pay different attention to different parts of input data, which can act on the original image, spatial scale and historical features of different moments. [25] explained that the attention mechanism is to use standard back-propagation techniques and to stochastically maximize a variational lower bound, and they divided attention into two variants: ''hard'' attention mechanism and ''soft'' attention mechanism. ''Hard'' attention takes hard decisions when choosing parts of the input data, and ''soft'' attention takes the entire input into account, weighting each part of observations dynamically [26].

One of the long-standing challenges for RL agents is to deal with noisy environments [27]. Inspired by human perception, it can use two basic concepts of machine learning, attention and memory, to better cope with the noisy environment and deal with a more complex task. It is coincided with the design principle and processing power of DRL algorithm and makes the combination of DRL and attention mechanism have research and application in the field of robots and unmanned driving. Sorokin *et al.* [28] presented an extension of DQN by ''soft'' and ''hard'' attention mechanisms and proposed deep attention recurrent Q network (DARQN) to directly monitor the training process online through the built-in attention mechanism.

To sum up, in this paper, we choose to combine DDPG algorithm with attention mechanism in order to complete the motion behavior cognition task of communication emitter. We propose the motion behavior cognition algorithms DDPG-BC and DDPG+A-BC and verify the feasibility and performance of the algorithms.

III. DDPG-BASED BEHAVIOUR COGNITION FOR COMMUNICATION EMITTER

A. PROBLEM ANALYSIS

The analysis of the motion behavior of communication emitter is based on the motion trajectory and signal characteristic parameters of the emitter. These parameters of target emitter can be extracted as the raw data, from which valid physical parameters are selected, and the motion state is obtained after preprocessing. Motion state will be the input of DDPG-based behavior cognition module. When prior knowledge meets the conditions, specific attention can be added into cognition module to help the cognition process (strategy learning process) explore and learn. Finally, cognitive results are obtained to judge the working status of the emitter and its platform or carrier. The cognition process of communication emitter behavior is shown in FIGURE 2.

B. DDPG-BC

One of the main challenges of learning in continuous action spaces is exploration, and one of the great advantages of

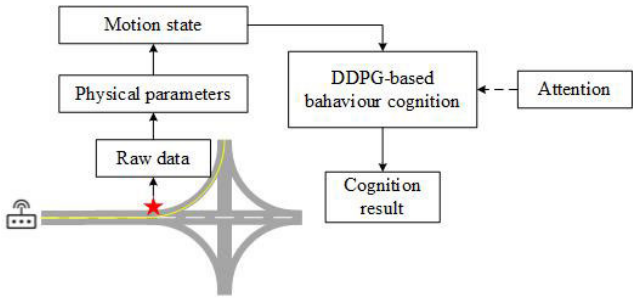


FIGURE 2. Cognition process of communication emitter.

algorithms like DDPG is that it can handle the exploration independently from the learning algorithm. DDPG algorithm is an improvement of DPG algorithm. On the basis of PG, DPG takes the state space as the algorithm model's input, but the output is no longer the probability of a certain action. Determinist action value, corresponding to a specific action, will be obtained through optimal action policy function $\mu_\theta(s)$, where s is state and θ is policy parameter.

Compared with DPG algorithm, deep neural network is added into DDPG, and DDPG takes Actor-Critic as the basic framework. DDPG imitates the idea of DQN. It uses memory tank and two neural networks with the same structure but different parameter update frequency of DDPG network to approximate policy function $\mu(s, \theta^\mu)$ and value function $Q(s, a; \theta^Q)$ respectively, which makes the learning process more effective and stable. In the function above, a is action, θ^μ is policy network parameter, and θ^Q is value network parameter. Meanwhile, Actor can easily select appropriate actions in the continuous action space, while Critic can update step by step and evaluate actions selected by learning the relationship between environment and rewards.

In addition, DDPG introduces the experience replay to remove correlation and dependency between samples when Actor interacts with the environment. Experience pool stores the state in t, action, reward, and state in $t+1$ (s_t, a_t, r_t, s_{t+1}) as experience and, each time, samples small batches of data from experience pool as training samples for policy and value network. On the one hand, let Q, μ and Q', μ' be Critic network, Actor network, and target network respectively, then target Q value can be expressed as

$$targetQ = r_t + \gamma Q'(s_{t+1}, \mu'(s_{t+1}; \theta^{\mu'}); \theta^{Q'}) \tag{1}$$

By minimizing loss function

$$Loss = \frac{1}{N}\sum (\text{targetQ} - Q(s_t, a_t; \theta^Q))^2 \tag{2}$$

Critic network will be updated via policy gradient

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum \nabla_a Q(s, a; \theta^Q)\big|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu}\mu(s; \theta^\mu)\big|_{s=s_t} \tag{3}$$

On the other hand, define target function as expectation of discount accumulative rewards

$$J_B(\mu) = E_\mu[r_1 + \gamma r_2 + \cdots + \gamma^{n-1} r_n] \tag{4}$$

and finding optimal deterministic behavior policy μ^* is equivalent to maximize target function

$$\mu^* = \text{argmax}_\mu J_\beta(\mu) \tag{5}$$

Finally, update target network

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'} \tag{6}$$

FIGURE 3 shows the Actor-Critic network structure in this paper. The Actor network has three layers and can choose the optimal action. The Critic network has four layers, including an input layer, two hidden layers and an output layer, which are used to train and generate Q values and update the Actor network. The Actor network selects action and sends into the environment, and the experience obtained after interaction is stored in the experience pool. Each time, bs training sample are sampled from the experience pool and sent to the dual network. The whole learning process is more stable and converges faster. Activation function and how modules work are shown in the diagram. The network parameters are shown in TABLE 1.

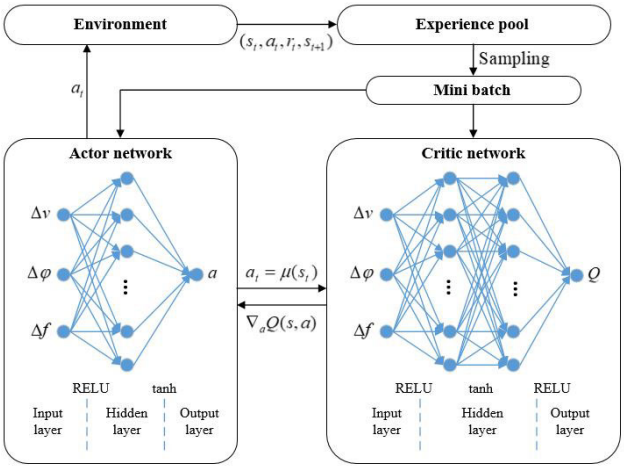


FIGURE 3. DDPG network structure for cognition.

TABLE 1. Network parameters.

Hyperparameters	Value	Description
γ	0.9	Discount factor for Q-learning
τ	0.001	Update rate of target network
LR_A	0.001	Learning rate of Actor network
LR_C	0.001	Learning rate of Critic network
mc	1000	Memory capacity of neural network
bs	32	number of samples in each training

DDPG-BC sends state space of emitter into DDPG network, explores steps times in each episode of training, learns the optimal cognition strategy, and realizes the correct cognition of communication emitter behavior. The pseudocode of DDPG-BC is summarized in TABLE 2.

TABLE 2. DDPG-BC pseudocode.

Algorithm 1 DDPG-BC

Input state/observation space from emitter data;
Input action space from environment, exploration noise N, and threshold for reward;
Initialize Actor network, Critic network, target network and experience pool;
for 1 to episodes:
  get initial state s_0;
  for 1 to steps:
    choose current action a_t (in time t);
    add randomness to action selection with N;
    execute and store transition experience $[s_t, a_t, r_t, s_{t+1}]$;
    randomly sample bs experience samples;
    calculate target Q value;
    update Actor network and Critic network;
    update target network with τ;
    $s_t \leftarrow s_{t+1}$;
  end for
  compare total reward with threshold;
  output cognition result, total reward;
end for

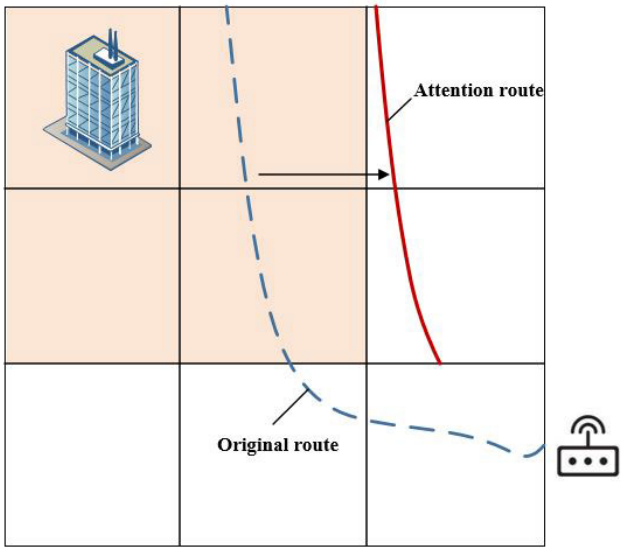


FIGURE 4. Geographic information-based attention mechanism schematic diagram.

C. DDPG+A-BC

When observing the real world, a human usually focuses on some fixation points at first glance of the scene [6]. When the prior information meets the conditions, the introduction of attention mechanism is considered as the help for solving the problem in this paper. If we know the trajectories or the geographic activity area or other relevant information of the moving communication emitter, we can build attention model, which participates in the learning process of DDPG, reduces computing cost, improves learning efficiency and accuracy, and better cognizes the motion behavior of the emitter.

According to the normal activity experience, hot spots or related areas concerned by moving objects will become a major factor affecting behavior. Because "hard" attention mechanism is generally considered a non-differentiable approach, it is not as widely used as "soft" attention. But [29] believed that feature magnitudes correlate with semantic relevance and provide a useful signal for our mechanism's attentional selection criterion. Therefore, compared with the "soft" attention mechanism, we hope to introduce an additional and explicable hyperparameter based on the "hard" attention mechanism in the training process. Then we use this built-in attention mechanism to focus on attention regions when making selections so as to improve the training speed and accuracy.

Suppose that the attention region model decides to focus on in t is M_t, which totally has L positions. If we hope model to extract features on i-th position (from L), attention position $M_{t,i}$ will be considered as the start of exploration. Schematic diagram of attention mechanism based on geographic information is demonstrated in FIGURE 4.

When a communication emitter travels along a given route (blue dotted line), it will pass over an attention region (shaded area) that will affect the normal operation of the emitter or the movement of its platform/carrier. Therefore, geographic information can be added into cognition process as the attention module in Fig. 2. It delimits the scope of network exploration and changes to focus cognition on the attention route (solid red line) for better cognition efficiency and accuracy.

It should be noted that geographic information is not the only background information that can function as an attention mechanism. Other information can also be applied to set the scope of attention and act on network learning and exploration.

After analyzing DDPG-BC, we believe that the algorithm may have some problems with cognition tasks. One is too wide selection range of actions and too strong randomness. According to the idea elaborated above, this paper proposes DDPG+A-BC. Based on DDPG -BC, attention mechanism is added before the exploration process to determine attention positions according to geographic information. We are intended to get initial state in or near attention positions and to keep exploration in attention region after any action operation. It will limit exploration scope (action selection) and improve exploration efficiency. DDPG+A-BC's pseudocode is summarized in TABLE 3.

IV. EXPERIMENTS RESULTS

Because there is no publicly available data set of communication emitter, we apply two simulation experimental data sets on verifying the performance of DDPG-BC and DDPG+A-BC, which are mainly divided into two parts: spatio-temporal data and signal parameter data. Spatio-temporal data is derived from actual data set, and we add communication frequency data for each sampling points to construct the data set for simulation. Python 3.6 and Tensorflow 2.0 are used to complete the programming implementation.

TABLE 3. DDPG+A-BC pseudocode.

Algorithm 2 DDPG+A-BC

Input state/observation space from emitter data;
Input action space from environment;
Input attention position, exploration noise N, and threshold for reward;
Initialize Actor network, Critic network, target network and experience pool;
for 1 to episodes:
  get initial state s_0 of attention positions;
  for 1 to steps:
    choose current action a_t (in time t) with N ;
    execute a_t and store transition experience $[s_t, a_t, r_t, s_{t+1}]$;
    randomly sample bs experience samples;
    calculate target Q value;
    update Actor network and Critic network;
    update target network with tau;
    $s_t \leftarrow s_{t+1}$;
  end for
  compare total reward with threshold;
  output cognition result, total reward;
end for

A. DATA AND ENVIRONMENT

We define $< \Delta v, \Delta\varphi, \Delta f >$ as the state space for communication emitter behavior's cognition network. Δv, $\Delta\varphi$, Δf represent velocity, direction(angle), and communication frequency change values sequence between sampling points respectively. The Agent will continuously select the action to be performed, analyze the corresponding state parameters and output cognition results. Definition of parameter in time t and cognition results are shown in TABLE 4 and TABLE 5. We assume that the working mode of the radio is abnormal when it exists communication frequency conversion. The cognition criterion of the frequency conversion in detail is shown in the demonstration of data sets below. And the definition of cognition results can be adjusted according to the model settings.

TABLE 4. Parameter definition (in time *t*) of network state space.

State	Raw physical parameters	Mathematical representation
Δv	Position coordinates x, y Sampling interval T	$v_t = \frac{\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}}{T}$ $\Delta v_t = \lvert v_{t+1} - v_t \rvert$
$\Delta\varphi$	Position coordinates x, y	$\varphi_t = \frac{\tan(\lvert y_t - y_{t-1} \rvert}{\lvert x_t - x_{t-1} \rvert)}$ $\Delta\varphi_t = \lvert \varphi_{t+1} - \varphi_t \rvert$
Δf	Communication frequency f	$\Delta f_t = \lvert f_{t+1} - f_t \rvert$

(1)The spatio-temporal data of data set 1 are obtained from publicly available flight trajectory data provided by Flightradar24. The simulation data set consists of 5 categories, each representing the motion trajectory of the same communication emitter. There are 140 groups of data with 150-300 sampling points in each group.

In data set 1, according to the characteristics of the simulation data set, when A3 occurs, Δv, $\Delta\varphi$, and Δf of the emitter is greater than 35, 150, and 500kHz respectively at same position/region. In the motion data of the same emitter, if the motion state of A3 only happens occasionally in a certain place, it will be judged as A2. At this time, sharp and large changes in state parameters usually occur. The rest are all determined as A1.

TABLE 5. Cognition results definition.

Cognition results	Definition
A1	Normal work (little change in velocity and direction, no frequency conversation)
A2	Probably abnormal work (existing changes in velocity and direction or frequency conversation, but cannot be determined as the abnormal)
A3	Abnormal work or be disturbed/struck (velocity and direction change greatly with frequency conversation)

(2)The spatio-temporal data of data set 2 comes from the Geolife project [30]–[32] of Microsoft Research Asia. In order to increase the diversity and complexity of motion states in the simulation data, nine groups of pedestrian trajectories from the Geolife Trajectory 1.3 dataset are adopted, with a total of 6,862 sampling points.

In data set 2, according to the characteristics of the simulation data set, when A3 occurs,Δv, $\Delta\varphi$, and Δf of the emitter is consecutively greater than 10, 100, and 500kHz respectively at a certain region. In the motion data of the same emitter, if the motion state of A3 only happens once in a certain place, it will be judged as A2. The rest are all determined as A1.

B. EXPERIMENTS RESULTS OF DDPG-BC

FIGURE 5 shows the cognition result of DDPG-BC, which is displayed by highlighting the experimental results (the red represents A3, and the blue represents A2). For all the simulation data graphs in this paper, coordinates of all positions have been expressed as longitude and latitude coordinates. To measure the performance of the algorithm, we use accuracy which can be calculated by

$$Accuracy = \frac{\sum_{i=1}^{3} P(i\,|i)}{\sum_{i=1}^{3}\sum_{j=1}^{3} P(i\,|j)} \tag{7}$$

$P(i\,|j)$ represents the number of samples when actual sample is i, while cognition result is j, $i, j = A1, A2, A3$.

By observing FIGURE 5 and TABLE 6, it can be found that DDPG-BC is able to realize cognition task of emitter behavior, and the accuracy is 90.434%. However, A2 and A3 cannot be well differentiated.

Cognition results of data set 2 with DDPG-BC are demonstrated in FIGURE 6. Combining FIGURE 6 and TABLE 7, we find that the cognition results of DDPG-BC roughly conform to the experimental setting, and the accuracy is 85.427%. The accuracy is reduced due to the complexity of

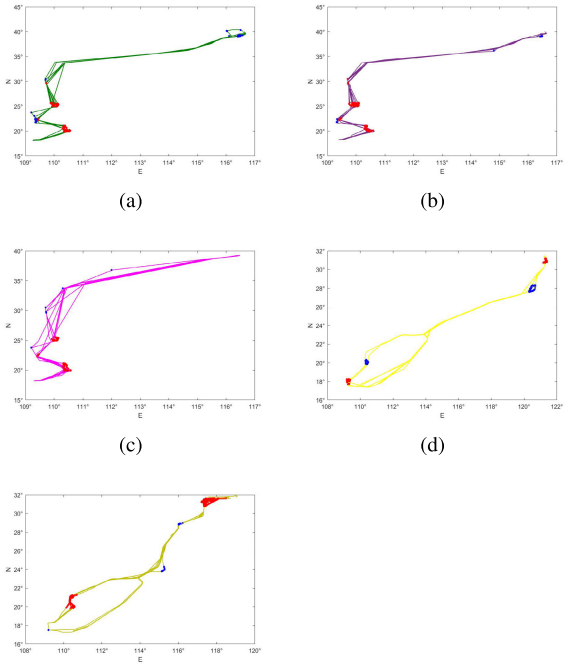


FIGURE 5. Cognition results of data set 1.

TABLE 6. Sample truths and cognition results.

		Sample truths			Total
		A1	A2	A3	
Cognition results	A1	25011	191	0	25202
	A2	427	2565	1813	4805
	A3	0	938	4275	5213
Total		25438	3694	6088	35220
Accuracy			90.434%		

TABLE 7. Sample truths and cognition results.

		Sample truths			Total
		A1	A2	A3	
Cognition results	A1	3218	11	0	3229
	A2	30	380	288	698
	A3	0	671	2264	2935
Total		3248	1062	2552	6862
Accuracy			85.427%		

the data, and it also cannot be able to distinguish A2 and A3 well.

C. EXPERIMENT RESULTS OF DDPG+A-BC

In data set 1, considering data characteristic and situation, we define the same region as exploration region (marked by box) for three groups to limit the scope of exploration, as shown in FIGURE 7. And then, A3 will be regarded as attention position where will be randomly selected as the initial state of the network. It should be noted that the addition of attention mechanism depends on whether the effect

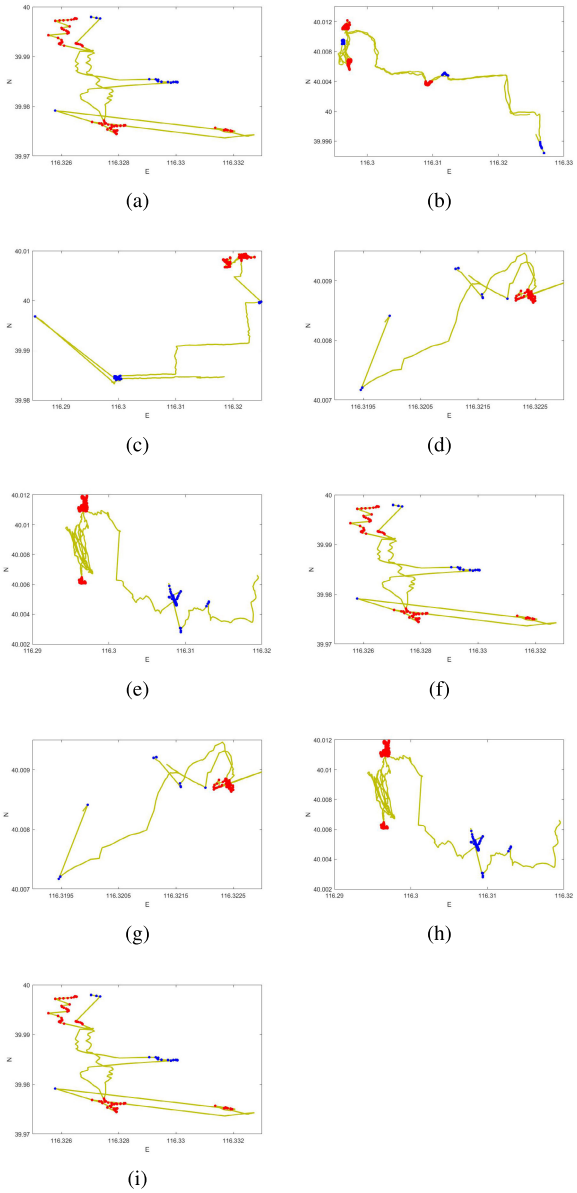


FIGURE 6. Cognition results of data set 2.

of attention is good or bad. For example, the cognition for A2 will be left out because of the limitation of the attention region in experiment.

$$Accuracy_{(\mathrm{A}_2,\mathrm{A}_3)} = \frac{\sum_{i=2}^{3} P(i\,|\,i)}{\sum_{i=2}^{3}\sum_{j=2}^{3} P(i\,|\,j\,)} \tag{8}$$

From TABLE 8, it is obvious that the cognition accuracy has been greatly improved compared with DDPG-BC, reaching 99.029%, and the discrimination effect of A2 and A3 has increased from 71.317% to 93.153% according to Eq.(8), indicating that the cognition process of using geographic

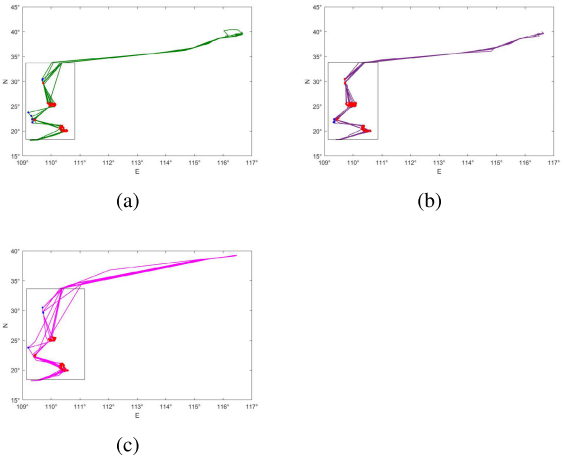


FIGURE 7. Cognition results of data set 1.

TABLE 8. Sample truths and cognition results.

		Sample truths			Total
		A1	A2	A3	
Cognition results	A1	16360	14	0	16374
	A2	52	383	69	504
	A3	0	40	1100	1140
Total		16412	437	1169	18018
Accuracy			99.029%		

information as attention for communication emitter behavior can effectively improve the cognition performance.

FIGURE 8 shows the final cognition results of data set 2 with DDPG+A-BC. Combining with TABLE 9, it can be found that the cognition accuracy increased by about 7% compared with DDPG-BC, reaching 92.495%, and the discrimination effect of A2 and A3 increased from 73.383% to 81.876%.

TABLE 9. Sample truths and cognition results.

		Sample truths			Total
		A1	A2	A3	
Cognition results	A1	1503	3	0	1506
	A2	12	175	59	246
	A3	0	132	861	993
Total		1515	310	920	2745
Accuracy			92.495%		

V. DISCUSSION

A. DISCUSSION FOR TRAINING EPISODE

Emitter behavior's cognition network is trained for 200 and 500 episodes, respectively. Each round explores for 200 steps. Losses, Qvalues, and TotalRewards in Figure 9 are used to observe the training results of the network, and the dotted line in TotalRewards represents the threshold to determine whether the network learns correctly.

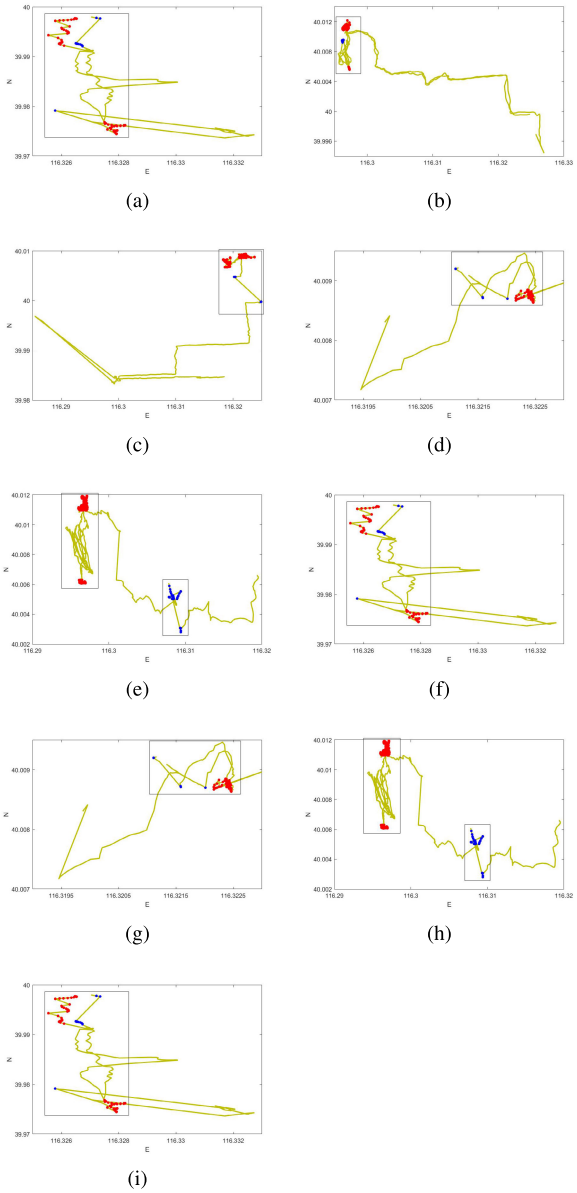


FIGURE 8. Cognition results of data set 2.

The training goal of DDPG network is to maximize target function (rewards) and minimize loss of value network. The loss of 200 episodes of training (FIGURE 9(a)) is approximate zero without convergence, and action value (Q value) is unstable. After 500 episodes (FIGURE 9(b)), losses and cumulative rewards converge, correct cognition is achieved after 235 episodes, and Q value tends to be stable.

B. DISCUSSION ON REWARD FUNCTION

Another problem of DDPG-BC is that the reward function may make the difference between cumulative rewards of A2 and A3 is too small after training in steps times, leading to Agent's inability to accurately distinguish two situations.

In this article, two reward functions are applied on cognition network to discuss the influence of reward function

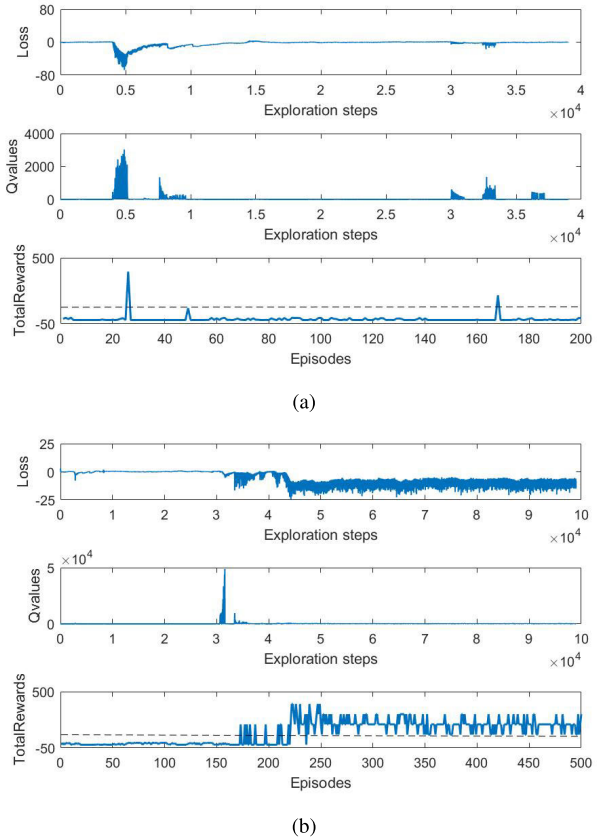


FIGURE 9. Training effect in 200 and 500 episodes.

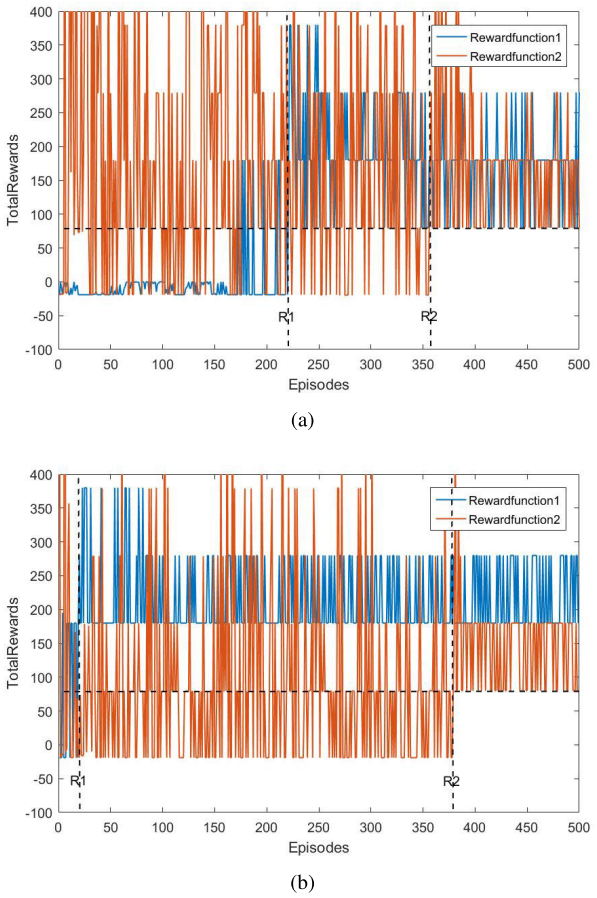


FIGURE 10. TotalRewards with different reward function.

during cognition process. We have referred to the reward function of 'MountainCarContinuous-v0', a continuous controlling environment from gym and of automated vehicle behavior decision making proposed in [8]. Eq.(9) and Eq.(10) are used to observe the influence of reward functions. r2 changes the reward for A3 in r1 from the fixed to the associated with state value. During the training, the reward will be continuously provided.

$$r1 = \begin{cases} -0.1 & for\ \text{A1} \\ 1 & for\ \text{A2} \\ 100 & for\ \text{A3} \end{cases} \tag{9}$$

$$r2 = \begin{cases} -0.1 & for\ \text{A1} \\ 1 & for\ \text{A2} \\ 100 * |\Delta v - \Delta\varphi - \Delta f| & for\ \text{A3} \end{cases} \tag{10}$$

FIGURE 10 shows the average cumulative reward of r1 and r2 at DDPG-BC and DDPG+A-BC, respectively. In any algorithm, r1 achieves correct cognition effect faster than r2, and attention fails to have a beneficial effect on cognition process with r2. In addition, it indicates that the reward with state values is not conducive to network learning and may even reduce the efficiency of the cognition process.

C. COMPARISON BETWEEN DDPG-BC AND DDPG+A-BC

Figure 11 shows the average cumulative reward after 500 episodes. DDPG+A-BC has been able to correctly cognize in the first 18 episodes and performs stably after the 19th episode, while DDPG-BC does not have such ability until after the 235 rounds. In addition, as shown in TABLE 10, the average training time of DDPG+A-BC is

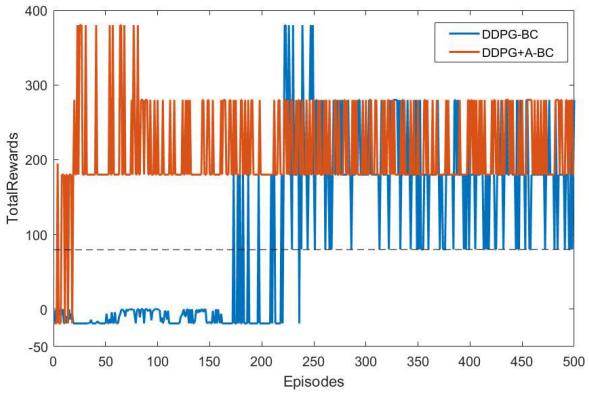


FIGURE 11. Comparison in TotalRewards.

TABLE 10. Training time.

	DDPG-BC	DDPG+A-BC
Average training time(s)	217.326	156.161

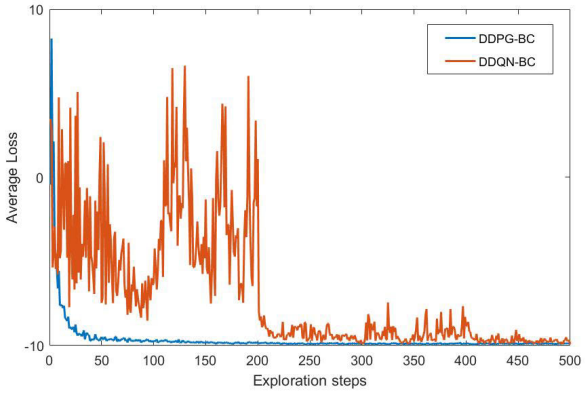


FIGURE 12. Average loss of DDPG and Double-DQN.

TABLE 11. Training time.

	DDPG	Double-DQN
Average training time(s)	89.608	177.183

about 60s shorter than that of DDPG, which proves that DDPG+A-BC can improve the efficiency of cognition.

D. COMPARISON WITH DQN

Due to the limitation of DQN algorithm in practical application, researchers proposed a Double-DQN algorithm [33], [34] to solve the overestimation problem of DQN. Double-DQN also has two Q network structures. By decoupling the selection of target Q value action and the calculation of target Q value, the network can avoid overestimation while approaching the optimal target as soon as possible.

In Section 3.2, we have mentioned that the value network of DDPG is based on Q network, and the experience replay of DQN algorithm is adopted to eliminate the correlation between samples. Therefore, it is possible to observe whether the algorithm in this paper has a better performance by comparing the behavior cognition results based on DDPG and Double-DQN. Since DQN can only deal with discrete actions, the selection of actions is limited to a certain range, divided equally into 11 actions (refer to 'Pendulum-v0' in gym) to be selected by the network.

FIGURE 12 and Table 11 respectively show the average loss and average training time of DDPG and Double-DQN network. It can be intuitively seen that the loss of DDPG converges faster, and DDPG has shorter average training

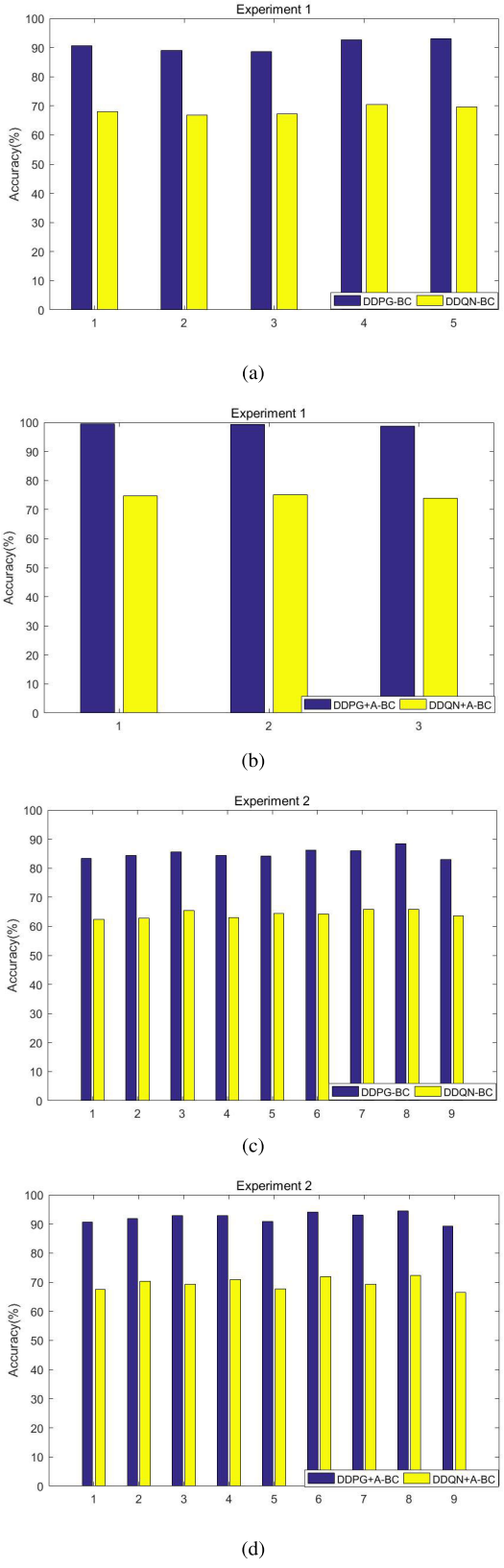


FIGURE 13. Accuracy of cognition results.

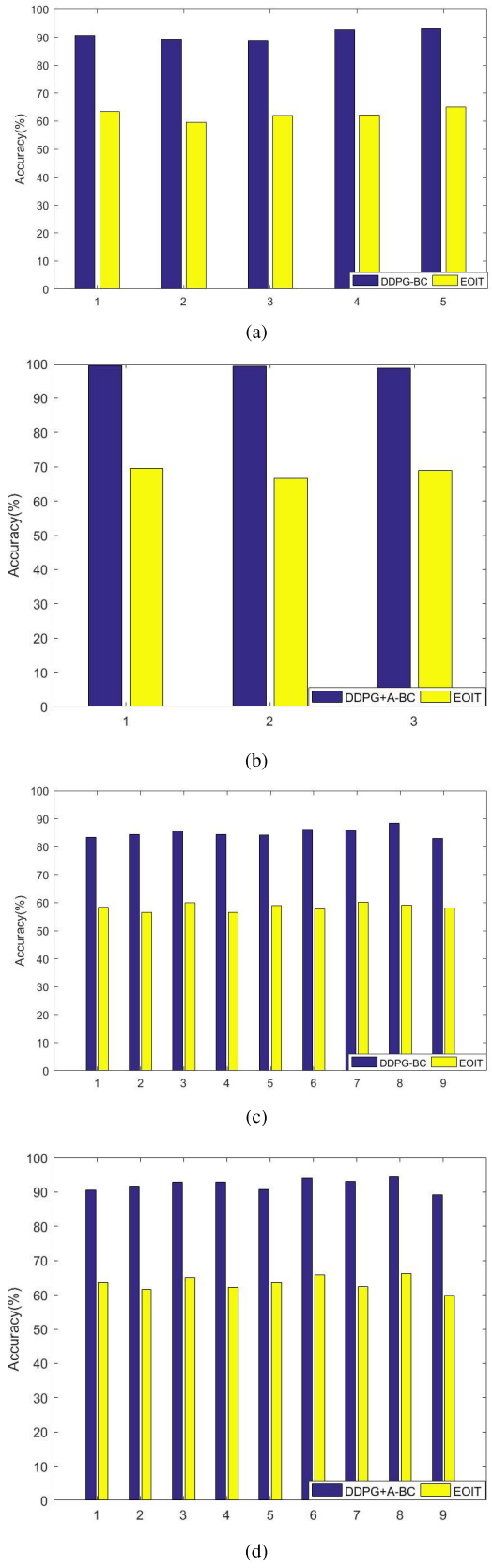


FIGURE 14. Accuracy of cognition results.

TABLE 12. Training time(s).

	DDPG+A-BC	EOIT
data set 1	156.161	177.183
data set 2	39.720	116.892

time. The training time required for DDPG is reduced due to the reduction of action space in the comparison experiment.

DDQN-BC and DDQN+A-BC represent the methods based on Double-DQN algorithm by imitating DDPG-BC and DDPG+A-BC algorithms. The accuracy of cognition results under four algorithms are compared on different experiment data sets respectively, as shown in FIGURE 13. (a), (c) show the comparison of the accuracy of cognition results of DDPG-BC and DDQN-BC on the data sets 1 and data set 2 respectively; (b), (d) compare that of DDPG+A-BC and DDQN+A-BC on two data sets with attention. After comparison and synthesis, it is found that: 1) the accuracy of cognition results based on DDQN is about 75% of that based on DDPG, and the cognition effect of DDPG-BC and DDPG+A-BC is better; 2) regardless of any of the algorithms, the introduction of attention mechanism can improve the cognition accuracy; 3) in general, the more complex the cognitive sample, the lower the cognitive accuracy, and vice versa.

E. COMPARISON WITH EOIT

EOIT is a conceptual framework and an experience-based approach [3]. According to EOIT's idea, the first 70% of each group of data set in data set 1 is taken as experience. For data set 2, 27 pedestrian trajectories from Geolife Trajectory 1.3 are reselected to constitute simulation experimental data as experience. Cognition results are obtained based on experience, and accuracy is used to measure performance.

FIGURE 14 demonstrates the cognition accuracy of EOIT and the algorithms in this paper. It can be found that the cognition accuracy of DDPG-BC and DDPG+A-BC is on average 31.27% higher than that of EOIT. Since the acquisition of experience requires consideration of all data in the specified range, EOIT will take longer time than DDPG+A-BC as shown in TABLE 12.

VI. CONCLUSION

In order to cognize the motion behavior of communication emitter, DDPG-BC and DDPG+A-BC are innovatively proposed in this paper. Firstly, considering the characteristics of emitter in multiple dimensions, large data and continuity and DRL's good learning ability and wide application in motion problems, we propose DDPG-BC based on DDPG for cognition tasks and set change values of velocity, direction, and communication frequency as state space. DDPG-BC will obtain specific cognition results directly and gain experience

from the interaction between network and environment. And then, we further propose a novel cognition algorithm named DDPG+A-BC with the introduction of attention mechanism. In addition to emitter's physical parameters, it uses geographic information (but not limited) to focus on attention positions in the process of DDPG network exploration, which can limit exploration scope and initial randomness of network to improve cognition efficiency.

The simulation results show that DDPG-BC can complete the cognition task on two different data sets with accuracy reaching 90.434% and 85.427% respectively. The addition of attention mechanism increased the cognition accuracy by 8.311% and 7.068%, leading to more precise cognition results and less cognition time. And compared with Double-DQN algorithm and existing cognition method EOIT, the algorithms proposed are all superior with less time and higher accuracy. In addition, the influence of training episode, reward function, and data complexity on cognition results are discussed respectively.

ACKNOWLEDGMENT

The authors would like to thank everyone for their comments and suggestions for this paper.

REFERENCES

[1] M. S. Zitouni, A. Sluzek, and H. Bhaskar, "Towards understanding socio-cognitive behaviors of crowds from visual surveillance data," *Multimedia Tools Appl.*, vol. 79, nos. 3–4, pp. 1781–1799, Jan. 2020.

[2] M. H. Goldberg, J. R. Marlon, S. A. Rosenthal, and A. Leiserowitz, "A meta-cognitive approach to predicting hurricane evacuation behavior," *Environ. Commun.*, vol. 14, no. 1, pp. 6–12, Jan. 2020.

[3] H. Zhang, F. Li, J. Wang, Y. Zhou, C. Sanin, and E. Szczerbicki, "Experience-based cognition for driving behavioral fingerprint extraction," *Cybern. Syst.*, vol. 51, no. 2, pp. 103–114, Feb. 2020.

[4] H. Hornischer, S. Herminghaus, and M. G. Mazza, "Structural transition in the collective behavior of cognitive agents," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, Dec. 2019.

[5] L. Pei, R. Guinness, R. Chen, J. Liu, H. Kuusniemi, Y. Chen, L. Chen, and J. Kaistinen, "Human behavior cognition using smartphone sensors," *Sensors*, vol. 13, no. 2, pp. 1402–1424, Jan. 2013.

[6] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.

[7] N. G. Nguyen, D. Phan, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, B. Purnama, M. K. Delimayanti, K. R. Mahmudah, M. Kubo, and K. Satou, "Applying deep learning models to mouse behavior recognition," *J. Biomed. Sci. Eng.*, vol. 12, no. 2, pp. 183–196, 2019.

[8] Y. Ye, X. Zhang, and J. Sun, "Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment," *Transp. Res. C, Emerg. Technol.*, vol. 107, pp. 155–170, Oct. 2019.

[9] D. Kim, G. Oh, Y. Seo, and Y. Kim, "Reinforcement learning-based optimal flat spin recovery for unmanned aerial vehicle," *J. Guid., Control, Dyn.*, vol. 40, no. 4, pp. 1076–1084, Apr. 2017.

[10] D. N. T. How, C. K. Loo, and K. S. M. Sahari, "Behavior recognition for humanoid robots using long short-term memory," *Int. J. Adv. Robot. Syst.*, vol. 13, no. 6, pp. 1–14, 2016.

[11] C. J. Cohen, G. Beach, B. Cavell, G. Foulk, C. J. Jacobus, J. Obermark, and G. Paul, "Behavior recognition system," U.S. Patent 7 036 094, Apr. 25, 2006.

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 65–72.

[13] Z. Wang, B. Guo, Z. Yu, and X. Zhou, "Wi-Fi CSI-based behavior recognition: From signals and actions to activities," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 109–115, May 2018.

[14] C. Wang, S. Chen, Y. Yang, F. Hu, F. Liu, and J. Wu, "Literature review on wireless sensing-Wi-Fi signal-based recognition of human activities," *Tsinghua Sci. Technol.*, vol. 23, no. 2, pp. 203–222, Apr. 2018.

[15] Y. Li, H. He, A. Khajepour, H. Wang, and J. Peng, "Energy management for a power-split hybrid electric bus via deep reinforcement learning with terrain information," *Appl. Energy*, vol. 255, Dec. 2019, Art. no. 113762.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013, *arXiv:1312.5602*. [Online]. Available: http://arxiv.org/abs/1312.5602

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[18] H. Shen, H. Hashimoto, A. Matsuda, Y. Taniguchi, D. Terada, and C. Guo, "Automatic collision avoidance of multiple ships based on deep Q-learning," *Appl. Ocean Res.*, vol. 86, pp. 268–288, May 2019.

[19] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*. [Online]. Available: http://arxiv.org/abs/1509.02971

[20] G. Matheron, N. Perrin, and O. Sigaud, "The problem with DDPG: Understanding failures in deterministic environments with sparse rewards," 2019, *arXiv:1911.11679*. [Online]. Available: http://arxiv.org/abs/1911.11679

[21] M. Hausknecht and P. Stone, "Deep reinforcement learning in parameterized action space," 2015, *arXiv:1511.04143*. [Online]. Available: http://arxiv.org/abs/1511.04143

[22] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," Tech. Rep., 2014.

[23] N. Heess, J. J. Hunt, T. P. Lillicrap, and D. Silver, "Memory-based control with recurrent neural networks," 2015, *arXiv:1512.04455*. [Online]. Available: http://arxiv.org/abs/1512.04455

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[26] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," 2017, *arXiv:1703.10106*. [Online]. Available: http://arxiv.org/abs/1703.10106

[27] M. Etchart, P. Ladosz, and D. Mulvaney, "Spatio-temporal attention deep recurrent q-network for pomdps," in *Proc. EPIA Conf. Artif. Intell.* Cham, Switzerland: Springer, 2019, pp. 98–105.

[28] I. Sorokin, A. Seleznev, M. Pavlov, A. Fedorov, and A. Ignateva, "Deep attention recurrent Q-network," 2015, *arXiv:1512.01693*. [Online]. Available: http://arxiv.org/abs/1512.01693

[29] M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia, "Learning visual question answering by bootstrapping hard attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–20.

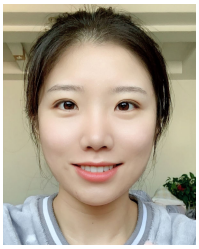
[30] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, 2009, pp. 791–800.

[31] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on GPS data," in *Proc. 10th Int. Conf. Ubiquitous Comput. (UbiComp)*, 2008, pp. 312–321.

[32] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, Jun. 2010.

[33] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.

[34] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, and H. Ishiguro, "Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2017, pp. 1639–1645.



YUFAN JI was born in China, in 1996. She received the B.S. degree in information engineering from the National University of Defense Technology, Hefei, China, where she is currently pursuing the master's degree in information and communication engineering. Her research interests include communication systems, deep learning, reinforcement learning, and data mining.

JIANG WANG was born in China, in 1975. He received the M.S. degree from the Electronic Engineering Institute of the PLA, Hefei, China. He is currently an Associate Professor and was engaged in communication equipment teaching over 15 years. His research interest includes technology on communication information processing.

WEILU WU was born in China, in 1973. She received the M.S. degree from the Electronic Engineering Institute of the PLA, Hefei, China. She is currently an Associate Professor and has engaged in communication equipment teaching for 15 years. Her research interest includes communication countermeasure.

LUNWEN WANG was born in 1966. He received the B.S. and M.S. degrees from the Electronic Engineering Institute of the PLA, Hefei, China, and the Ph.D. degree in communication and information systems from Anhui University, Hefei, in 2002. He is currently a Professor and a Ph.D. Supervisor with the National University of Defense Technology. His research interests include neural networks and data mining.

CHUANG PENG received the B.S. degree in information engineering from Dalian Maritime University, in 2016, and the M.S. degree in communication engineering from the National University of Defense Technology, in 2018, where he is currently pursuing the Ph.D. degree with the College of Communications Engineering. His research interests include data analytics, wireless communications, and cognitive radio networks.

HAO SHAO received the B.S. degree in communication engineering from the Nanjing University of Posts and Telecommunications, in 2017, and the M.S. degree in information and communication engineering from the National University of Defense Technology, China, in 2019, where he is currently pursuing the Ph.D. degree. His current research interests include link prediction, data mining, and machine learning.