Deep-Reinforcement-Learning-Based Energy Management Strategy for Supercapacitor Energy Storage Systems in Urban Rail Transit

Zhongping Yang, *Member, IEEE*, Feiqin Zhu, *Student Member, IEEE*, and Fei Lin, *Member, IEEE*

***Abstract*— The modeling complexity of the traction power system and variation of traffic conditions bring challenges for the optimization of energy management strategy for supercapacitor energy storage systems in urban rail transit. Therefore, in this paper a deep-reinforcement-learning-based energy management strategy is proposed: the energy management system is modeled as an intelligent agent, the reward function is formulated comprehensively considering the energy-saving and voltage-stabilizing effects of supercapacitor, a traction power system simulator is developed to emulate the environment, and the agent's behavior is improved in each headway through the deep Q-learning algorithm, and converges to the nearly-optimal policy. The proposed strategy is verified through simulation based on the Beijing Subway Batong Line. The study results show that it dynamically adjusts the voltage thresholds so as to better allocate the supercapacitor capacity along the time horizon. The energy-saving and voltage-stabilizing effects are significantly improved compared with the fixed-threshold strategy and genetic optimization, and demonstrating to be in close proximity to the optimal benchmark deduced from dynamic programming.**

***Index Terms*— Supercapacitor energy storage systems, energy management, reinforcement learning, urban railway.**

I. Introduction

IN RECENT years, wayside energy storage systems for urban rail systems have been widely investigated, with the aims of recycling the considerable regenerative energy and stabilizing the network voltage [1], [2]. Supercapacitors, which have an advantage of high power density, possess great potential for energy recovery [3], [4]. In general, the supercapacitor energy storage system (SCESS) plays the role of an energy buffer, which releases energy when the train is accelerating, and recycles the surplus energy when the train is braking, achieving a good balance of charge and discharge. Therefore, energy management strategy (EMS) has a significant impact on the energy recycling ability of SCESS.

Manuscript received March 26, 2018; revised January 31, 2019 and July 22, 2019; accepted December 17, 2019. Date of publication January 10, 2020; date of current version February 2, 2021. This work was supported by National Key R&D Program of China under Grant 2017YFB1201105 (2017-2020). The Associate Editor for this article was D. Chen. *(Corresponding author: Feiqin Zhu.)*

The authors are with the School of Electrical Engineering, University of Beijing Jiaotong, Beijing 100044, China (e-mail: zhpyang@bjtu.edu.cn; 15117405@bjtu.edu.cn; flin@bjtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2019.2963785

At present, different energy management strategies have been proposed in literatures, and these can be classified into three categories, i.e. the line-voltage-based strategy, the state of charge (SOC)-based strategy, and the optimal control strategy. The line-voltage-based strategy sets charge and discharge voltage thresholds for SCESS, and determines the working mode on the basis of network voltage [5], [6]. The working states of supercapacitor were managed by the state machine in [7], [8]. The relationship between the energy saving amount and threshold value was presented in [9], and the optimal thresholds were obtained from the gradient descent method. However, the energy recycling effect is affected by numerous aspects, such as the no-load network voltage and the trains' operation states, thus the formula cannot be determined explicitly. As for the SOC-based strategy, Petar J proposed a three-loop control structure, which implemented accurate tracking of SOC with the supercapacitor voltage controller [10]–[12]. The expected SOC was calculated with reference to the train's kinetic energy in [13]–[15]. Nevertheless, tracking of the expected SOC and matching with vehicle power demand could not be satisfied simultaneously. In addition, the charge energy could not be accurately predicted as multiple trains, substations and ESSs perform energy interaction through the supply network. For the optimal control strategy, a dynamic programming (DP)-based energy management strategy was proposed with the goal of minimizing the energy dissipation on the braking resistor in [16]. However, DP suffers from the curse of dimensionality and is unable to be realized in real-time. An EMS based on model predictive control (MPC) was evaluated in a hybrid system of tramway [17]. MPC implements on-line suboptimal control, however, the performance is influenced by the prediction accuracy and temporal horizon. Zhang proposed a real-time EMS based on equivalent consumption minimization strategy (ECMS) with velocity forecasting, and the equivalent factor (EF) was regulated by an adaptive law considering the predicted velocity [18]. However, the complexity of the traction power supply system in urban rail transit makes it difficult to solve the optimal control strategy of energy storage systems analytically. And the performance of the theoretical solution cannot be guaranteed due to the inevitable model error and system operation deviation. Besides, the distance between the train and SCESS, the train powers, the fluctuation

1558-0016 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

of substation no-load voltage, etc have significant influences on the energy saving effect of the SCESS. It's challenging to adjust the control parameters adaptively with respect to the change of system states.

In recent years, the rapid advancement of artificial intelligence inspires us to tackle the energy management problem with data-driven approaches. Compared with supervised learning and unsupervised learning, RL does not rely on a supervisor, it aims to optimize the sequential decision process through the mechanism of "trial and error". The intelligent agent takes actions, obtains feedbacks of reward signals and makes online improvement of its policy to maximize the accumulative rewards. Recently, The RL algorithms, combined with deep learning, have achieved significant progress, and show strong performance in different test environments, such as Atari video games [19]–[21] and chess playing [22], [23]. Deep learning models progressively extract higher level features from raw input. Therefore, when they are trained to approximate the value functions in RL, the curse of dimensionality with tabular approach is effectively avoided, and they enable the DRL algorithms to be a powerful tool to solve complex control tasks end to end. DRL has been also drawing increasingly attention in various application fields, such as traffic signal control [24] in traffic systems, optimal power management [25], [26], robotics and so on. In this paper, the deep Q learning algorithm, which is the first success and one of the most representative DRL algorithms, is adopted to implement real-time and nearly-optimal energy management of SCESS in the complex traction power system. It adjusts the control parameters dynamically with respect to the states of multiple trains, the substation and SCESS, performs online learning within each headway, so that it continuously improves its policy and achieves optimization of energy saving and voltage stabilization effect.

This paper is organized as follows: First, the working principle and basic control strategy of SCESS are presented. Then, the mathematical model of the power supply system is introduced, based on which the optimal control problem of the SCESS is formulated, comprehensively considering the voltage stabilization and energy saving purposes. Next in Section III, the reinforcement learning framework for SCESS is established, in which the states, actions, rewards of the learning agent are defined, and the deep Q-learning algorithm is presented. Finally, the DRL-based strategy is verified through simulation based on Beijing Metro Batong Line, and compared with the fixed-threshold strategy, the genetic algorithm and DP.

II. Problem Formulation

A. SCESS and Its Control

In the DC traction power supply system of urban rail transit, the trains are generally supplied by the unidirectional 24-pulse wave diode rectifiers in traction substations, as seen in Fig. 1. In order to effectively utilize the trains' regenerative braking energy, the supercapacitor-based energy storage system (SCESS), which consists of the DC/DC converter and supercapacitor modules, is installed in the traction substation, as shown in Fig. 1. The line-voltage-based control strategy is employed to manage the charge/discharge power of SCESS, as seen in Fig. 2. When there are trains braking near the SCESS, and the substation voltage rises higher than the charge threshold, the SCESS works in the charge state, the surplus braking energy is recovered by the supercapacitor. When the trains accelerate, and the substation voltage drops lower than the discharge threshold, the SCESS discharges to stabilize the network voltage at the discharge threshold.

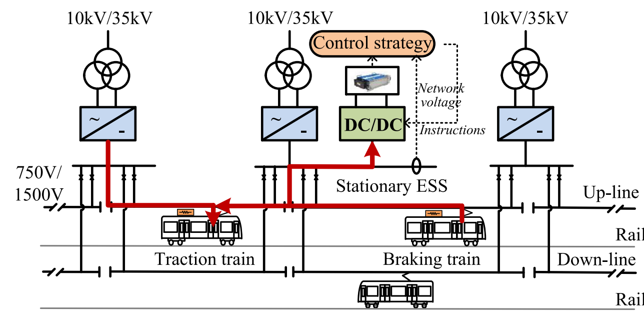

Fig. 1. Stationary supercapacitor energy storage system in urban railway.

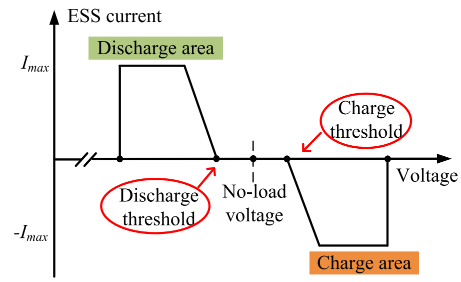

Fig. 2. The voltage-threshold-based control strategy.

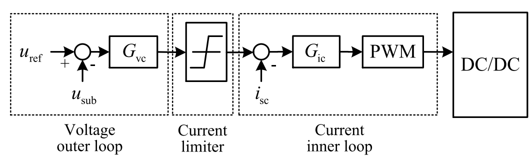

Fig. 3. The double closed loop control structure.

In order to track the voltage reference value and implement control of charge/discharge power, the double-closed loop control structure of the DC/DC converter is adopted, as seen in Fig. 3. The outer-loop proportional-integral(PI) controller G_{vc} regulates the network voltage, generating a reference current for the inner loop. The reference current is restricted by the current limiter so as to prevent overcurrent of supercapacitor. And the inner-loop controller G_{ic} regulates the supercapacitor current at the reference value by adjusting the duty cycles of the IGBT driving pulse. G_{vc} and G_{ic} are expressed as (1) and (2), respectively.

$$G_{\mathrm{vc}} = K_{\mathrm{p,v}} + \frac{K_{\mathrm{i,v}}}{s} \tag{1}$$

$$G_{\mathrm{ic}} = K_{\mathrm{p,i}} + \frac{K_{\mathrm{i,i}}}{s} \tag{2}$$

where $K_{\mathrm{p,v}}$, $K_{\mathrm{p,i}}$ are proportional coefficients of the two controllers, and $K_{\mathrm{i,v}}$, $K_{\mathrm{i,i}}$ are the integration coefficients.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

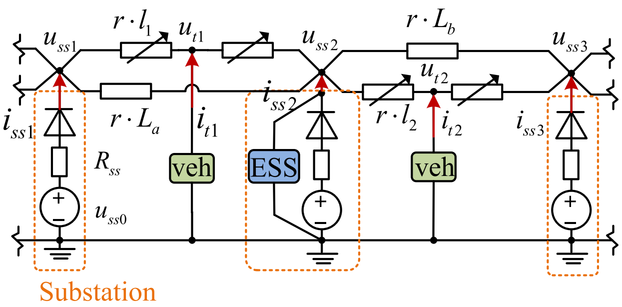

Fig. 4. The circuit model of the traction power supply system.

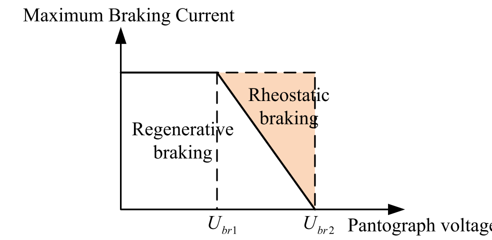

Fig. 5. The regenerative limiting curve of trains.

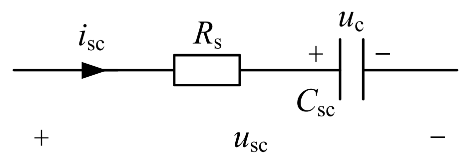

Fig. 6. The equivalent model of supercapacitor.

B. Modeling of Power Supply System

The equivalent circuit model of the traction power supply system is depicted in Fig. 4. u_{ss} and i_{ss} represent the voltages and currents of substations, and u_t and i_t denote the train voltages and currents, respectively. r is the unit resistance of the contact line, l_1, l_2, L_a and L_b are distances between nodes. They satisfy the nodal voltage equation, as in (3), shown at the bottom of this page. Due to the movement of trains, the topology of the network and line resistances between trains and substations is time-variant. In order to obtain the voltage-current relationships(VCR) of each element, their mathematic models are presented in this section.

1) Substations: Considering the output characteristic curve of the rectifier [27], the substation is modeled as a voltage source that is in series with a diode and an equivalent resistance, and its volt-ampere characteristic equation is shown in (4).

$$u_{ss} = u_{ss0} - R_{ss} \cdot i_{ss}, i_{ss} \geq 0 \tag{4}$$

where $u_{\rm ss0}$ and $R_{\rm ss}$ are the no-load voltage and equivalent resistance of the substation.

2) Trains: In the urban rail system, the trains have four operating states, namely traction, cruise, coasting and braking. When the train operates under the traction condition, it absorbs power from the network, and converts the electrical power to the mechanical power through the traction drive system. When the train is braking, the dynamic energy is converted to electric energy and released to the power network. Therefore, the train is modeled as a power source, and its current is derived from the train power and DC bus voltage.

$$i_{\rm t} = \frac{p_{\rm t}}{u_{\rm t}} \tag{5}$$

where $i_{\rm t}$ and $p_{\rm t}$ are the DC-side current and power of the traction converter, respectively, and $u_{\rm t}$ is the DC bus voltage.

It is important to note that the pantograph voltage will rise rapidly when the train brakes and there is no adjacent traction train. In order to prevent overvoltage, the braking resistor will start and consume part of the braking energy, as shown in Fig. 5.

3) Supercapacitor: The supercapacitor is modeled as a voltage source in series with the equivalent internal resistance [28], [29], as seen in Fig. 6. $C_{\rm sc}$ is the capacitive part of the supercapacitor model, and represents the charge-holding capability; $R_{\rm s}$ is the equivalent internal resistance. $u_{\rm sc}$ and $i_{\rm sc}$ respectively represent the terminal voltage and current of supercapacitor. Therefore, the relationship between the supercapacitor current and voltage can be obtained, as seen in (6) and (7) . And the state of charge (SOC) of supercapacitor, which represents the ratio of stored energy to its nominal energy, is expressed as (8):

$$u_{\rm c}(t) = \frac{1}{C_{\rm sc}} \int_0^t i_{\rm sc}(\tau) d\tau + U_{\rm c0} \tag{6}$$

$$u_{\rm sc}(t) = u_{\rm c}(t) + R_s \cdot i_{\rm sc}(t) \tag{7}$$

$$SOC = \frac{u_{\rm c}^2}{u_{\rm c,nom}^2} \tag{8}$$

where $U_{\rm c0}$ is the initial voltage of supercapacitor, and $u_{\rm c,nom}$ is the nominal voltage of supercapacitor.

$$\begin{bmatrix} \frac{1}{rl_1}+\frac{1}{rL_a} & -\frac{1}{rl_1} & -\frac{1}{rL_a} & 0 & 0 \\ -\frac{1}{rl_1} & \frac{1}{rl_1}+\frac{1}{r\left(L_a-l_1\right)} & -\frac{1}{r\left(L_a-l_1\right)} & 0 & 0 \\ -\frac{1}{rL_a} & -\frac{1}{r\left(L_a-l_1\right)} & \frac{1}{rL_a}+\frac{1}{r\left(L_a-l_1\right)}+\frac{1}{rl_2}+\frac{1}{rL_b} & -\frac{1}{rl_2} & -\frac{1}{rL_b} \\ 0 & 0 & -\frac{1}{rl_2} & \frac{1}{rl_2}+\frac{1}{r\left(L_b-l_2\right)} & -\frac{1}{r\left(L_b-l_2\right)} \\ 0 & 0 & -\frac{1}{rL_b} & -\frac{1}{r\left(L_b-l_2\right)} & \frac{1}{rL_b}+\frac{1}{r\left(L_b-l_2\right)} \end{bmatrix} \begin{bmatrix} u_{\rm ss1} \\ u_{\rm t1} \\ u_{\rm ss2} \\ u_{\rm t2} \\ u_{\rm ss3} \end{bmatrix} = \begin{bmatrix} i_{\rm ss1} \\ i_{\rm t1} \\ i_{\rm ss2} \\ i_{\rm t2} \\ i_{\rm ss3} \end{bmatrix} \tag{3}$$

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

C. Optimal Control Problem of SCESS

It's seen from Fig. 2 that the setting of the threshold voltages affects the charge/discharge power of the supercapacitor and substation voltage, which accordingly affects the voltage distribution of the traction network and energy flow among the energy storage system, substation and trains. Therefore, in order to improve the energy efficiency and reduce the voltage fluctuation of the traction network, the threshold voltages should be optimized and adjusted dynamically. The optimal energy management problem for the SCESS is formulated as finding the decision sequence that maximize the performance index J while satisfying the constrains, as seen in (9).

$$\max_{u(t)} J = \omega \cdot \upsilon\% + (1-\omega) \cdot e\%$$
$$s.t. \begin{cases} \text{KVL equations} \\ \text{VCR equations,} \\ |i_{sc}| \le I_{\lim} \\ SOC_{\min} \le SOC \le SOC_{\max} \end{cases} \tag{9}$$

1) Objective Function: The optimization objective J is expressed as the weighted sum of energy saving index $e\%$ and voltage stabilization index $\upsilon\%$, and ω is the weighting factor. $e\%$ and $\upsilon\%$ are defined as (10) and (11), respectively.

$$e\% = \left(1 - \frac{\sum_{j=1}^{n} \int_0^T u_{sub\_j}^{sc} \cdot i_{sub\_j}^{sc} dt}{\sum_{j=1}^{n} \int_0^T u_{sub\_j}^{nosc} \cdot i_{sub\_j}^{nosc} dt} \right) \times 100\% \tag{10}$$

where

$u_{sub\_j}^{sc}$, $u_{sub\_j}^{nosc}$: the voltage of substation j with and without SCESS;

$i_{sub\_j}^{sc}$, $i_{sub\_j}^{nosc}$: the current of substation j with and without SCESS;

n: the number of substations.

$$\upsilon\% = \left\{ 1 - \frac{\sum_{k=1}^{Nt} \left[\int_{\Delta_h T} \Delta_h u_{pk}^{sc} dt + \int_{\Delta_l T} \Delta_l u_{pk}^{sc} dt \right]}{\sum_{k=1}^{Nt} \left[\int_{\Delta_h T} \Delta_h u_{pk}^{nosc} dt + \int_{\Delta_l T} \Delta_l u_{pk}^{nosc} dt \right]} \right\} \times 100\% \tag{11}$$

where

$\Delta_h T$: the time interval when the train voltage is higher than u_h;

$\Delta_l T$: the time interval when the train voltage is lower than u_l;

$\Delta_h u_{pk}^{sc}$, $\Delta_h u_{pk}^{nosc}$: the difference between the k-th train voltage and u_h with and without SCESS.

$\Delta_l u_{pk}^{sc}$, $\Delta_h u_{pk}^{nosc}$: the difference between the k-th train voltage and u_l with and without SCESS.

N_t: the total number of trains in the line section.

Therefore, $e\%$ represents the percentage of energy reduction amount after installing SCESS. The voltage stabilization effect is evaluated by the integration of difference between the train voltage and the upper/lower value.

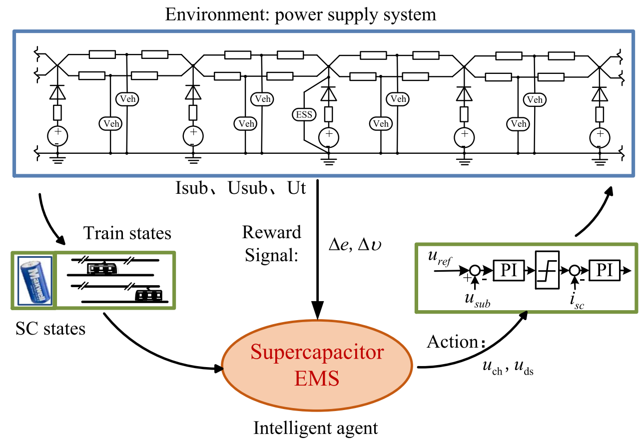

Fig. 7. Diagram of reinforcement learning framework for supercapacitor energy storage system.

2) Decision Variables: The decision variable $u(t)$ is the combination of charge and discharge voltage threshold at time t, as seen in (12).

$$u(t) = [u_{ch}(t), u_{ds}(t)] \tag{12}$$

3) Constraints: The constraints of the optimal control problem include the Kirchhoff circuit equations of the power supply system (as seen in (3)), the VCR equations of each elements(including (4), (5), (6), (7) and (8)), and the operating restrictions of the SCESS, i.e. the SOC and current of supercapacitor should be within the permissible range.

III. DRL-Based Energy Management Strategy

A. Reinforcement Learning for SCESS

The reinforcement learning problem is one in which an agent attempts to achieve the control objective through the interaction with the environment. The block diagram of the RL model for SCESS is shown in Fig. 7. The supercapacitor energy management system is modeled as an intelligent agent which performs learning and decision-making, while the traction power system is regarded as the environment. The agent observes the environment state and its own state, executes specific actions, which influences the environment, and causes the environment to generate a corresponding reward signal. Then the agent improves its policy with respect to the obtained reward, so as to maximize the accumulative rewards during the time horizon. The model elements, including the state s, the action a and the reward r are defined below.

1) State s: s consists of operating states of trains which are near the SCESS, SOC of the supercapacitor, the rectifier current and voltage of the substation where the SCESS is installed. In this paper, it's assumed that the train-to-ground communication is realizable, so the energy management system obtains the train information. s is denoted as:

$$s = [d_1, p_1, \cdots, d_N, p_N, soc, i_{ss}, u_{ss}] \tag{13}$$

where d_k and p_k represent the displacement and power of the k-th train, respectively.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

2) Action a: The action of SCESS is defined as (12). And the policy determines the behavior of the agent, which is a mapping from the state set S to the action set A: $\pi : S \rightarrow A$.

The action set A contains all possible actions. In this paper, we consider the range of voltage fluctuation in substations and discretize the charge/discharge thresholds to obtain A, as shown in (14).

$$A = \{[u_{ss0} - k_1 \cdot \Delta u, u_{ss0} + k_2 \cdot \Delta u]; k_1, k_2 \in [1, N]\} \quad (14)$$

where $u_{\mathrm{ss}0}$ is the no-load voltage, Δu is the unit change value of voltage threshold.

3) Reward r: The reward signal r is the feedback of the environment to agent's action, and the goal of agent learning is to obtain the maximum cumulative reward. In this paper, r is defined as the increment of energy-saving rate and voltage stabilization rate in each time step ΔT, as shown in (15), the weight coefficient ω is taken as 0.5.

$$r = -0.5 \cdot \Delta v\% - 0.5 \cdot \Delta e\% \quad (15)$$

B. Deep Q-Learning Algorithm

The Q learning algorithm (Watkins, 1989) is one of reinforcement learning algorithms that learns the optimal strategy through empirical sampling under the condition that the Markov decision process (MDP) is unknown. $Q(s, a)$ is defined as the expected cumulative reward when the agent performs action a at state s, and follows a certain policy in subsequent steps. It evaluates the benefit of the action taken in a particular state. The Q-learning algorithm belongs to the off-policy temporal difference (TD) strategy, which uses the TD target to estimate the action-value function. It attempts to find the optimal action-value Q^* by value iteration, and the updating formula of $Q(s, a)$ is formulated as (16).

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right] \quad (16)$$

In (16), $r_{t+1} + \gamma \cdot \max_a Q(s_{t+1}, a)$ is the estimated return, which is calculated without model knowledge of the MDP. α is the learning rate, and γ is the discounting factor. The larger γ means the agent tends to consider more about the future rewards.

However, the tabular Q-learning algorithm is unable to deal with continuous states, and suffers from curse of dimensionality when the discretized state-action space is large. In this paper, the state s is expressed as (13). If there are N_t trains near the SCESS, the train power and displacement are discretized with P_t and D_t feasible values, respectively, and the SOC of supercapacitor is discretized with M_s feasible values, the substation current and voltage are discretized with N_{us}, N_{is} elements, respectively, and there are N_a actions for each state, then the state-action table will have $E = N_a M_s N_{is} N_{us} (P_t D_t)^{N_t}$ elements. Take $N_t = 8$, $P_t = 10$, $D_t = 10$, $M_s = 100$, $N_{is} = 100$, $N_{us} = 10$, $N_a = 18$, then $E = 1.8 \cdot 10^{22}$, which causes curse of dimensionality.

Therefore, the value function approximation method is used to improve the efficiency and generalization ability. The action-value Q is approximated by an artificial neural network with parameters ω, i.e. $Q(s, a) \approx \hat{Q}(s, a; \omega)$. The Q-network is trained by gradient descent to reduce the mean squared error between the target-network and the Q-network, as seen in (17). N is the mini-batch size, and ω^- is the weight of the target-network. In order to break the correlation of the training data and improve stability, the experience tuples $\langle s, a, r, s' \rangle$ are stored in the replay buffer, and later sampled randomly for network training [20]. The DQN-based energy management algorithm is shown in Table. I.

TABLE I
THE DQN-BASED ENERGY MANAGEMENT ALGORITHM

Initialize replay buffer B, Q-network Q with random weights ω and initialize target network Q' with zero weights ω^-
Repeat (for each episode):
Initialize operation states s for the traction power simulator
Repeat (for each step of episode):
Choose a from s with ε -greedy policy, as seen in (18)
Execute action a in the traction power system simulator
The simulator solves the circuit equations, transforms the system state to the next state s', and the agent gets reward signal r
Store the transition $\langle s, a, r, s' \rangle$ in B
Sample a mini-batch of transitions from B
Update ω of Q-network with gradient descent on (17)
$\omega^- \leftarrow \omega$ for every n steps
Until s is terminal
Until stopping criterion satisfied

$$L(\omega) = \frac{1}{N} \sum_{k=1}^{N} \left[r_k + \gamma \max_{a_k'} Q\left(s_k', a_k'; \omega^-\right) - Q(s_k, a_k; \omega)\right]^2 \quad (17)$$

In order to realize the trade-off between exploration and exploitation, i.e. the agent takes full advantage of the obtained experience while continues to explore the unknown environment, the $\varepsilon-$greedy policy is used for action selection, as seen in (18). The agent chooses the greedy action with probability $1 - \varepsilon$, and a random action with probability ε.

$$\pi(a|s) = \begin{cases} 1 - \varepsilon, & \arg\max_{a \in A} Q(s, a) \\ \varepsilon, & otherwise \end{cases} \quad (18)$$

C. Training Implementation

The training process is implemented by the interaction of DQN-based strategy with the DC traction power environment, as seen in Fig. 10. We use the fully-connected neural network to approximate the Q-values, as seen in Fig. 8. The input of the Q-network is the normalized state vectors with 19 components. Followed by the input layer are three hidden layers, which respectively have 512, 256 and 128 neural units. The rectified linear unit (RELU) is used for activation of the hidden layers. The output layer is fully-connected with linear activation function, it outputs the action-value $Q(s, a_i)$ for each action a_i. Adam is used to update the weights of the Q-network. The DQN-based strategy is realized with TensorFlow on a personal computer with a 6-core CPU.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

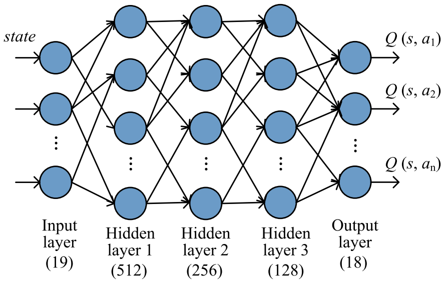

Fig. 8. The architecture of Q-network.

The traction power system simulator has been developed in a number of previous studies [7], [30]–[32]. In this work, the simulator is built in Matlab/Simulink. It is capable of emulating multiple-train operation under different headways in a two-side feeding power network. The energy dissipation of on-board resistors is also considered so the distribution of regenerative braking energy is well depicted. The simulator is developed in a modular manner, it consists of the train operation calculation module, energy storage module, and power network module. The power network module is further divided into the equivalent train model, contact line resistance model and substation model, as described in Section II-B. In order to deal with the time-variance of circuit topology and parameters caused by train movement, the fictitious train in each section between two substations is used in simulation. The powers of fictitious trains are deduced from the multi-train operation calculation, and the contact line resistance changes with the train displacement. The implementation process of the simulator is as follows: First, the displacement-time and power-time curves for a single train are calculated with the fixed-time traction calculation algorithm, and they are distributed to the fictitious trains in different sections given the headway. Also, the action generated by the DQN-based strategy is sent into the energy storage module. Then, the KVL equation at the current time instant is solved by means of DC-power calculation to obtain the currents and voltages of the substations, trains and on-board braking resistors. Finally, the SOC of the supercapacitor, the powers and displacements of the trains are updated, and the system transfers to the next time state.

IV. Simulation Studies

A. Simulation Conditions

In order to verify the DQN-based energy management strategy, a case study is presented here based on the Beijing Subway Batong Line. The route map of the Batong Line is shown in Fig. 9. It has a total length of 18.96 km, with 13 passenger stations. In this paper, the section between Sihui Station and Shuangqiao Station is chosen for simulation, and the SCESS is assumed to be installed in Gaobeidian Station. The simulation section includes five substations, and the station distances are shown in Table. II.

The parameters of the vehicle, power supply system and supercapacitor module are listed in Table. III. The train is composed of six B-type cars, and the total mass is 279.6 t under the standard load condition (AW2). The metro line is supplied by a DC 750V two-side feeding traction power system. The no-load voltage of substation fluctuates with the change of urban power load, and in this paper is set at 860 V for simulation. The SCESS is connected with 15 Maxwell 48-V/165-F modules in series and 14 in parallel. The rated power is 1.3MW, and the total stored energy is 8.4kWh.

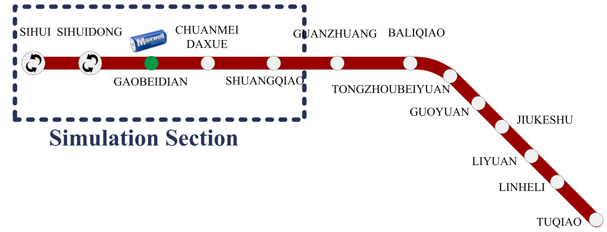

Fig. 9. The route map of beijing batong line.

TABLE II
Station Distances of the Simulation Section

Section number	Section	Distance(m)
1	Sihui-Sihuidong	1715
2	Sihuidong-Gaobeidian	1375
3	Gaobeidian-Chuanmeidaxue	2002
4	Chuanmeidaxue-Shuangqiao	1894

TABLE III
Simulation Parameters

Parameters	Values	Unit
train set	3M3T	-
weight(AW2)	279.6	t
maximum speed	80	km/h
motor rated power	180*12	kW
motor efficiency	92	%
regenerative limiting voltage	900/970	V
no-load voltage	860	V
equivalent resistance of substation	0.0161	Ω
pantograph resistance	0.015	Ω
contact line unit impedance	0.007	Ω/km
rail unit impedance	0.009	Ω/km
rated voltage of SC module	48	V
rated current of SC module	130	A
rated capacitance of SC module	165	F
equivalent series resistance of SC module	6.3	mΩ

B. Simulation Results

A multi-train operation cycle under the headway of 360s is used for simulation studies. The train power-time curves and displacement-time curves are shown in Fig. 13(a) and (b), respectively. The detailed training parameters of the DQN-based strategy are listed in Table. IV. The discounting factor γ reflects the consideration of future rewards. In this paper, it's set at 0.99, so the energy saving effect and voltage-stabilization effect in the long run will be taken

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

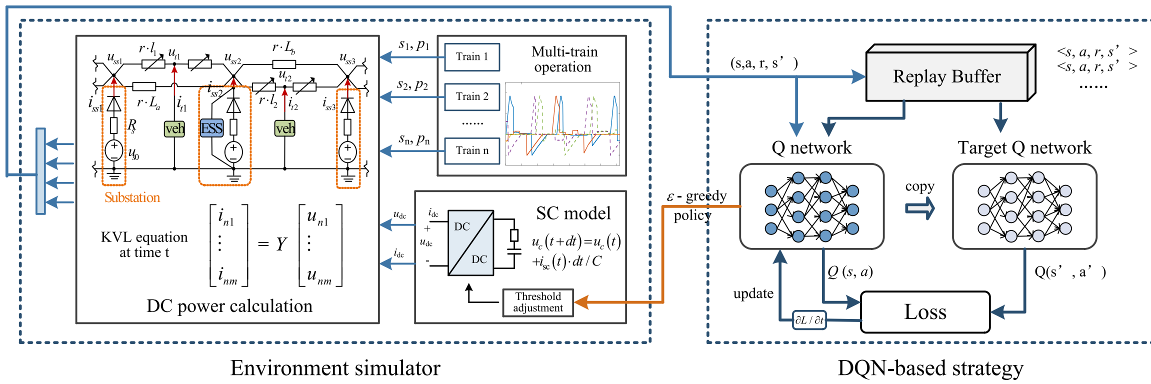

Fig. 10. The training structure of DQN-based strategy.

TABLE IV
TRAINING PARAMETERS OF DQN-BASED STRATEGY

Parameters	Descriptions	Values
α	learning rate	$1 \times e^{-3}$
N_b	batch size	32
N_r	replay buffer size	1000000
f_t	target update frequency	10000
γ	discounted factor	0.99
ε	probability of exploration	decay with time

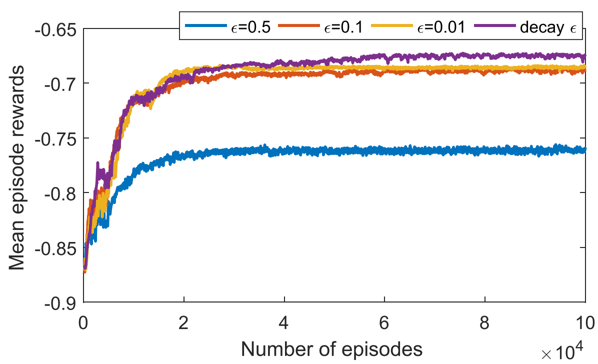

Fig. 11. The converging curve of the DQN-based strategy.

into account. In order to achieve the trade-off between the exploration and exploitation, in this paper, ε decays with time in a piecewise manner: ε reduces linearly from 1 to 0.1 after $1 \times e^6$ steps, and continues to decay until $1 \times e^7$ steps, then finally remains constant at 0.01. Fig. 11 shows the iteration curves of the DQN-based strategy with different exploration parameters. When ε is large ($\varepsilon = 0.5$), the agent interacts with the environment more stochastically, which is less likely to produce good training data, thus the final episode rewards is relatively low. When ε is small ($\varepsilon = 0.01$), the average rewards are much better than that when $\varepsilon = 0.5$. However, as the agent hardly explores, it's not easy to find better solutions. And when ε decays with time, the agent chooses good actions and generates more effective data to improve its policy while keeps the ability of exploring. Therefore, it achieves good trade-off between exploration and exploitation, and the final mean reward is better than that when ε is constant.

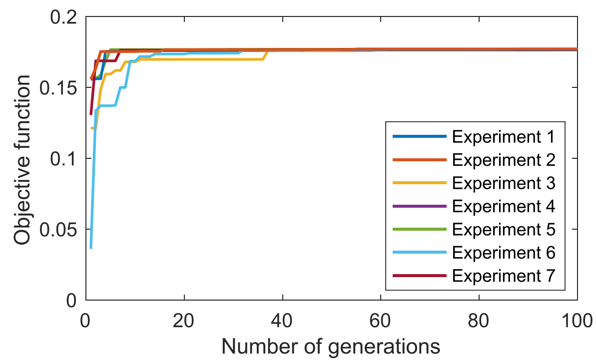

Fig. 12. The iteration curves of of GA.

In order to better evaluate the performance of the DQN-based strategy, the fixed-threshold strategy and the energy management strategies based on genetic algorithm (GA) and dynamic programming (DP) are also presented for comparison. DP solves the optimal control problem based on Bellman optimality principle, and achieves global optimization on the premise that the traffic condition is known in advance [33], [34]. However, it cannot be implemented online because of lacking the prior knowledge and high computation complexity. Therefore, in this work DP is adopted as an offline optimal benchmark for performance evaluation. For the fixed-threshold strategy, the supercapacitor is encouraged to charge/discharge as possible, thus the threshold voltages are set near the no-load voltage of substation, i.e. $u_{ch,fix} = 861V$, $u_{ds,fix} = 859V$. And for the GA-based strategy, the charge/discharge voltage thresholds are optimized with the genetic algorithm. GA is an evolutionary algorithm which is inspired by the natural selection mechanism, and the fitness function of the population is improved with the genetic operators, including selection, crossover and mutation. As the parameters of the algorithm have an important influence on the results, seven sets of experiments with different parameters are presented. The optimization parameters and results are listed in Table. V. The population size (denoted as N_p) is set at 25, 35 and 50, the crossover fraction (denoted as p_x) is chosen within range [0.4,0.99] and the mutation fraction (denoted as p_m) is chosen within range [0.01,0.1]. Fig. 12 shows the evolution curves of the objective function. It's seen

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

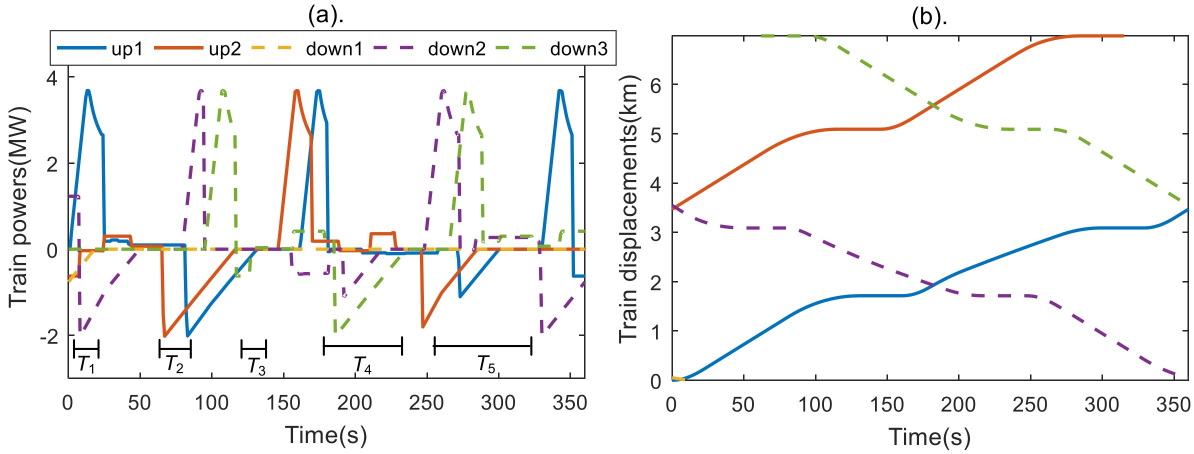

Fig. 13. The trains' operating curves under 360s headway. (a) Train powers. (b) Train displacements.

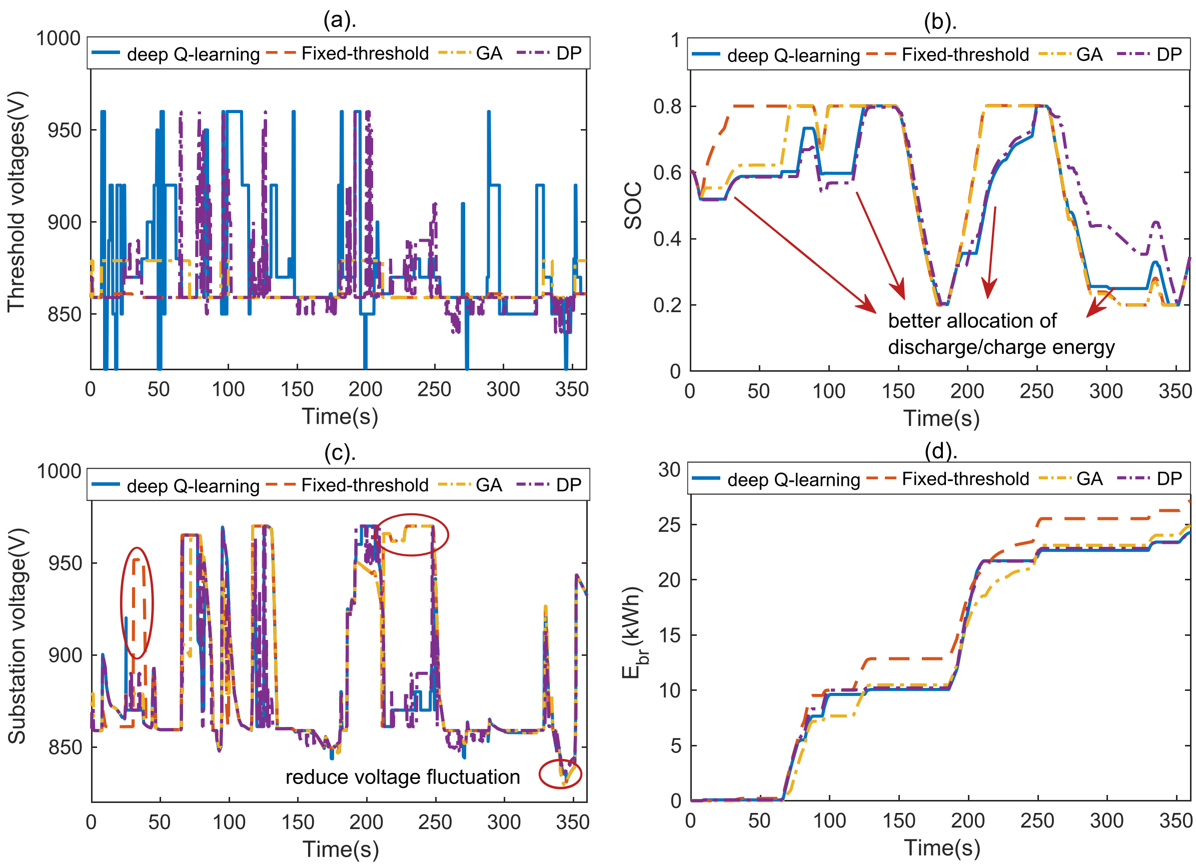

Fig. 14. The operating curves under different strategies. (a) Charge/discharge threshold voltages. (b) SOC of supercapacitor. (c) Substation voltage. (d) Energy consumptions on braking resistors.

that the under different parameters, the objective function is improved, and converges to nearly 0.177. When p_m is small, it's hard to keep population diversity, the algorithm is trapped in the local optimum (as seen in experiment 1); and when p_m is large, it tends to random search, which is not benefit for generation evolution, thus the convergence speed is much slower in experiment 3. Also, small values of p_x and N_p make it hard to produce new individuals, and take more iterations to converge. In experiment 2, the individuals with good fitness functions are reserved as possible and new individuals are also generated to maintain the population diversity, so it obtains the best objective value, and the results of it are used in strategy verification.

Fig. 14, 15 and 16 compares the simulation results under EMSs based on DQN, fixed-thresholds, GA and DP. Figs. 14 (a), (b), (c) and (d) respectively present the curves of threshold voltages, the supercapacitor SOC, the substation voltages and energy dissipated on braking resistors. Figs. 15 and 16 show the pantograph voltages and braking resistor currents of the trains. It can be observed from

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

TABLE V
THE PARAMETERS AND RESULTS OF GA

No.	N_p	p_x	p_m	Best values	Best solutions
1	35	0.5	0.01	0.1765	{ 858.9, 885.8}
2	35	0.5	0.05	0.1771	{ 858.9, 879.0}
3	35	0.5	0.1	0.1765	{ 857.8, 876.8}
4	35	0.4	0.05	0.1768	{ 857.7, 880.1}
5	35	0.7	0.05	0.1769	{ 857.9, 878.3}
6	25	0.5	0.05	0.1771	{ 857.6, 880.6}
7	50	0.5	0.05	0.1769	{ 858.8, 876.8}

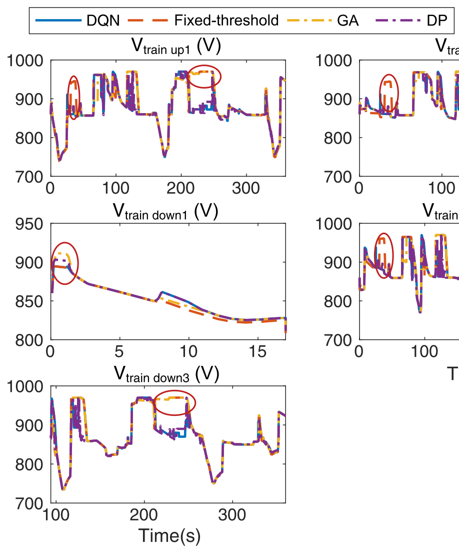

Fig. 15. The pantograph voltages of trains.

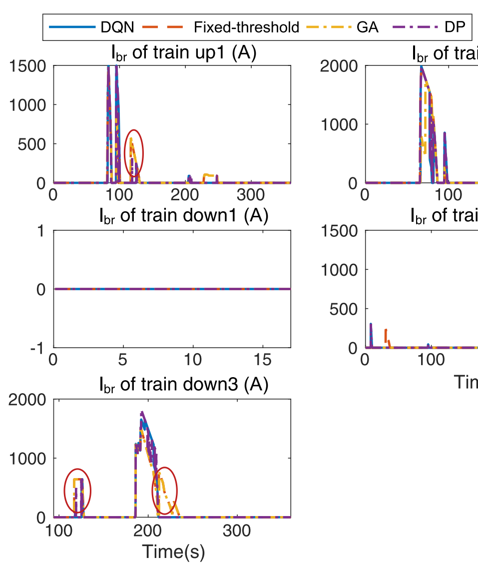

Fig. 16. The braking resistor currents of trains.

Fig. 14 that during time interval T_1, when the 1-st uplink train accelerates, and the 2-nd downlink train brakes (as seen in Fig. 13), the braking energy of down-link train 2 is absorbed by the SCESS under the fixed-threshold strategy, and the SOC of supercapacitor immediately rises to its upper limit.

TABLE VI
COMPARISON OF ENERGY STATISTICS UNDER DIFFERENT CONDITIONS

	without ESS	fixed-threshold	GA	deep Q-learning	DP
total traction energy(kWh)	137.1	137.1	137.1	137.1	137.1
total braking energy(kWh)	94.4	94.4	94.4	94.4	94.4
braking resistor loss(kWh)	38.8	27.1	24.92	24.30	24.25
line loss (kWh)	12.97	13.21	13.39	13.45	13.40
total output energy of substation(kWh)	94.5	80.89	78.83	78.28	78.20

Therefore, in the subsequent time interval T_2, when the 2-nd uplink train and the 1-st uplink train brake, the SCESS is not able to recover the braking energy, the pantograph voltages rise rapidly, and the braking resistors start. However, under the DQN-based strategy, the braking energy of the 2-nd downlink train is absorbed by the powering train 1 during interval T_1, and the supercapacitor leaves a certain capacity to recover regenerative energy during T_2, reducing the energy dissipation on the braking resistors and suppressing the voltage rise. Similarly, during time interval T_4 (when two trains brake simultaneously) and T_5 (when multiple trains traction successively), the supercapacitor rapidly charges/discharges to the upper/lower limit of SOC, and exits working under the fixed-threshold strategy and GA optimization. However, under the DQN-based strategy, the supercapacitor SOC is managed appropriately to improve the voltage stabilization effect. And it's seen from Fig. 14 (b) that the SOC trajectory of the DQN-based strategy is similar with that of DP. Major difference is observed during T_5. The discharge energy of SCESS under DP is further reduced to maintain SOC at a higher value, which enhances the power of SCESS. Therefore, the DQN-based strategy adjusts the voltage thresholds dynamically with respect to the states of trains, substation and SCESS, optimizes the energy distribution of the system. It achieves a better allocation of the supercapacitor capacity with time, hence the SCESS is utilized more effectively to recover the regenerative energy and reduce the network voltage fluctuation.

Table. VI shows the energy statistics under different conditions. Fig. 17 presents the distribution of the trains' braking energy. E_{br} is the energy loss on braking resistor, $E_{sc,ch}$ represents the energy recovered by SCESS, $E_{interaction}$ denotes the energy interaction between trains. E_{br} is significantly reduced after installing the SCESS. Under the fixed-threshold strategy, E_{br} is reduced to 27.1 kWh, and 14.05 kWh of braking energy is recovered by the SCESS. However, the energy interaction between trains also decreases slightly to 53.25 kWh. This signifies that the SCESS recovers the braking energy originally transmitted to the traction trains, which is inefficient as it increases the load burden of SCESS and may result in increase of line loss. However, under the DQN-based strategy, the SCESS absorbs the part of braking energy that cannot be utilized by the traction trains and increases the energy

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

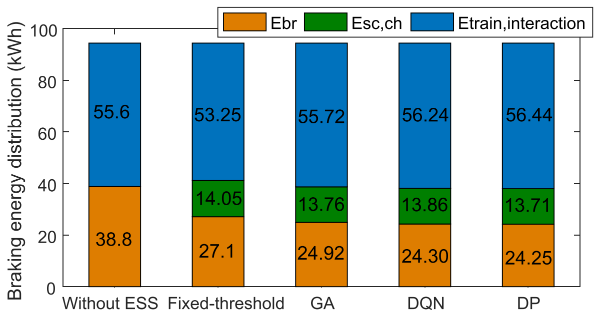

Fig. 17. Comparison of braking energy distribution.

TABLE VII
COMPARISON OF EVALUATION INDEX UNDER DIFFERENT CONTROL STRATEGIES

strategies	fixed-threshold	GA	deep Q-learning	DP
$e\%$	14.4	16.6	17.16	17.25
$v\%$	12.5	18.8	48.8	51.07
J	13.5	17.7	33.0	34.16

interaction between trains by improving the voltage distribution of the system. $E_{interaction}$ is enhanced to 56.24 kWh and E_{br} is then further reduced to 24.30 kWh. The total energy consumption of substations in one headway is 78.28 kWh under the DQN-based strategy, which is 2.61kWh lower than the fixed-threshold strategy, 0.55kWh lower than the GA-based strategy, and deviates little from the DP benchmark. Therefore, the DQN-based strategy effectively improves the energy efficiency of the traction power system, and presents significant energy saving potential for long-time operation.

Table. VII compares the evaluation indexes under the four strategies. It' seen that the control performance of the deep Q-learning algorithm is significantly improved. Compared with the fixed-threshold strategy, the energy saving rate $e\%$ is increased by 2.76% and the voltage stabilization rate $v\%$ is enhanced by 36.3%. In comparison with the results of GA, the energy saving performance is similar. However, $v\%$ is significantly improved. The objective value J reaches 33.0% under the DQN-based strategy, which achieves 96.6% of the DP benchmark, demonstrating that the DQN-based strategy improves its policy by interacting with the environment, and attains online nearly-optimal performance.

V. CONCLUSION

In this paper, a deep-reinforcement-learning-based energy management strategy for SCESS in urban rail transit is proposed, which implements real-time and nearly-optimal control by adjusting control parameters dynamically with respect to the states of trains, substations and SCESS. The proposed strategy is verified through simulation based on the Beijing Subway Batong Line. The study results show that, in comparison with the fixed-threshold strategy and genetic optimization, it has a better management of supercapacitor SOC, and the energy interaction among trains is improved, so the SCESS is utilized more effectively. Under the proposed strategy, the voltage stabilization rate reaches 48.8% and the energy saving rate reaches 17.16%, which deviate little from the optimal solution. In future works, the learning efficiency of the algorithm will be improved by making full use of the pre-known information, and it's also significant to improve the transfer ability of the algorithm, so it keeps good performance when transferring from simulation to real environment, and new conditions that has not encountered before.

REFERENCES

[1] T. Ratniyomchai, P. Tricoli, and S. Hillmansen, "Recent developments and applications of energy storage devices in electrified railways," *IET Elect. Syst. Transp.*, vol. 4, no. 1, pp. 9–20, Mar. 2014.

[2] A. González-Gil, R. Palacin, and P. Batty, "Sustainable urban rail systems: Strategies and technologies for optimal management of regenerative braking energy," *Energy Convers. Manage.*, vol. 75, pp. 374–388, Nov. 2013.

[3] I. Hadjipaschalis, A. Poullikkas, and V. Efthimiou, "Overview of current and future energy storage technologies for electric power applications," *Renew. Sustain. Energy Rev.*, vol. 13, nos. 6–7, pp. 1513–1522, Aug. 2009.

[4] S. Vazquez, S. M. Lukic, E. Galvan, L. G. Franquelo, and J. M. Carrasco, "Energy storage systems for transport and grid applications," *IEEE Trans. Ind. Electron.*, vol. 57, no. 12, pp. 3881–3895, Dec. 2010.

[5] H. Douglas, C. Roberts, S. Hillmansen, and F. Schmid, "An assessment of available measures to reduce traction energy use in railway networks," *Energy Convers. Manage.*, vol. 106, pp. 1149–1165, Dec. 2015.

[6] F. Ciccarelli, D. Iannuzzi, K. Kondo, and L. Fratelli, "Line–voltage control based on wayside energy storage systems for tramway networks," *IEEE Trans. Power Electron.*, vol. 31, no. 1, pp. 884–899, Jan. 2016.

[7] R. Barrero, X. Tackoen, and J. Van Mierlo, "Stationary or onboard energy storage systems for energy consumption reduction in a metro network," *Proc. Inst. Mech. Eng., F, J. Rail Rapid Transit*, vol. 224, no. 3, pp. 207–225, May 2010.

[8] R. Barrero, X. Tackoen, and J. Van Mierlo, "Improving energy efficiency in public transport: Stationary supercapacitor based Energy Storage Systems for a metro network," in *Proc. IEEE Vehicle Power Propuls. Conf.*, Sep. 2008, pp. 1–8.

[9] H. Lee, J. Song, H. Lee, C. Lee, G. Jang, and G. Kim, "Capacity optimization of the supercapacitor energy storages on dc railway system using a railway powerflow algorithm," *Int. J. Innov. Comput., Inf. Control*, vol. 7, no. 5, pp. 2739–2753, 2011.

[10] P. J. Grbovic, P. Delarue, P. Le Moigne, and P. Bartholomeus, "Modeling and control of the ultracapacitor–based regenerative controlled electric drives," *IEEE Trans. Ind. Electron.*, vol. 58, no. 8, pp. 3471–3484, Aug. 2011.

[11] P. J. Grbovic, P. Delarue, P. Le Moigne, and P. Bartholomeus, "The ultracapacitor–based controlled electric drives with braking and ride–through capability: Overview and analysis," *IEEE Trans. Ind. Electron.*, vol. 58, no. 3, pp. 925–936, Mar. 2011.

[12] P. J. Grbovic, P. Delarue, P. Le Moigne, and P. Bartholomeus, "The ultracapacitor–based regenerative controlled electric drives with power–smoothing capability," *IEEE Trans. Ind. Electron.*, vol. 59, no. 12, pp. 4511–4522, Dec. 2012.

[13] F. Ciccarelli and D. Iannuzzi, "A novel energy management control of wayside li–ion capacitors-based energy storage for urban mass transit systems," in *Proc. Int. Symp. Power Electron. Power Electron., Elect. Drives, Autom. Motion*, Jun. 2012, pp. 773–779.

[14] F. Ciccarelli, D. Iannuzzi, and P. Tricoli, "Control of metro-trains equipped with onboard supercapacitors for energy saving and reduction of power peak demand," *Transp. Res. C, Emerg. Technol.*, vol. 24, pp. 36–49, Oct. 2012.

[15] F. Ciccarelli, A. Del Pizzo, and D. Iannuzzi, "Improvement of energy efficiency in light railway vehicles based on power management control of wayside lithium–ion capacitor storage," *IEEE Trans. Power Electron.*, vol. 29, no. 1, pp. 275–286, Jan. 2014.

[16] E. Bilbao, P. Barrade, I. Etxeberria-Otadui, A. Rufer, S. Luri, and I. Gil, "Optimal energy management strategy of an improved elevator with energy storage capacity based on dynamic programming," *IEEE Trans. Ind. Appl.*, vol. 50, no. 2, pp. 1233–1244, Mar. 2014.

[17] J. P. Torreglosa, P. Garca, L. M. Fernndez, and F. Jurado, "Predictive control for the energy management of a fuel-cellbatterysupercapacitor tramway," *IEEE Trans. Ind. Informat.*, vol. 10, no. 1, pp. 276–285, Feb. 2014.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.

[18] F. Zhang, J. Xi, and R. Langari, "Real–time energy management strategy based on velocity forecasts using V2V and V2I communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 2, pp. 416–430, Feb. 2017.

[19] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," Dec. 2013, *arXiv:1312.5602*. [Online]. Available: https://arxiv.org/abs/1312.5602

[20] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[21] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[22] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[23] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[24] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi–agent deep reinforcement learning for large–scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[25] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[26] R. Liessner, C. Schroer, A. Dietermann, and B. Bäker, "Deep reinforcement learning for advanced energy management of hybrid electric vehicles," in *Proc. 10th Int. Conf. Agents Artif. Intell.*, 2018, pp. 61–72.

[27] P. Pozzobon, "Transient and steady-state short-circuit currents in rectifiers for DC traction supply," *IEEE Trans. Veh. Technol.*, vol. 47, no. 4, pp. 1390–1404, Nov. 1998.

[28] J. M. Miller, *Ultracapacitor Applications*. London, U.K.: Institution of Engineering and Technology, 2011.

[29] N. Devillers, S. Jemei, M.-C. Péra, D. Bienaimé, and F. Gustin, "Review of characterization methods for supercapacitor modelling," *J. Power Sour.*, vol. 246, pp. 596–608, Jan. 2014.

[30] F. Ciccarelli, D. Iannuzzi, and D. Lauria, "Supercapacitors-based energy storage for urban mass transit systems," in *Proc. 14th Eur. Conf. Power Electron. Appl.*, Aug./Sep. 2011, pp. 1–10.

[31] H.-S. Lee, H.-M. Lee, C.-M. Lee, G.-S. Jang, and G.-D. Kim, "Energy storage application strategy on dc electric railroad system using a novel railroad analysis algorithm," *J. Elect. Eng. Technol.*, vol. 5, no. 2, pp. 228–238, 2010.

[32] T. Ratniyomchai, S. Hillmansen, and P. Tricoli, "Energy loss minimisation by optimal design of stationary supercapacitors for light railways," in *Proc. Int. Conf. Clean Elect. Power (ICCEP)*, Jun. 2015, pp. 511–517.

[33] D. E. Kirk, *Optimal Control Theory: An Introduction*. Chelmsford, MA, USA: Courier Corporation, 2012.

[34] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic Programming And Optimal Control*, vol. 1, no. 2. Belmont, MA, USA: Athena Scientific, 1995.

Zhongping Yang (Member, IEEE) received the B.Eng. degree in electrical engineering from the Tokyo University of Mercantile Marine, Tokyo, Japan, in 1997, and the M.Eng. and Ph.D. degrees in electrical engineering from the University of Tokyo, Tokyo, in 1999 and 2002, respectively.

He is currently a Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include high-speed rail integration technology, traction and regenerative braking technology, and wireless power transfer of urban rail vehicles.

Prof. Yang received the Zhan Tianyou Award for Science and Technology in 2010, the Excellent Popular Science and Technology Book Award in 2011, and the Science and Technology Progress Award (second prize) of Ministry of Education in China in 2016, respectively.

Feiqin Zhu (Student Member, IEEE) received the B.S. degree in electrical engineering from Beijing Jiaotong University, Beijing, China, in 2015, where she is currently pursuing the Ph.D. degree with the School of Electrical Engineering.

Her research interests include modeling and optimization of energy storage systems, cooperative control, and energy management in urban railway.

Fei Lin (Member, IEEE) received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1997, the M.S. degree in electrical engineering from Shandong University, Jinan, China, in 2000, and the Ph.D. degree in electrical engineering from Tsinghua University, Beijing, China, in 2004.

He is currently a Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing. His research interests include traction converter and motor drives, energy management for railway systems, and digital control of power-electronic-based devices.

Authorized licensed use limited to: Beijing Jiaotong University. Downloaded on January 17,2022 at 02:50:33 UTC from IEEE Xplore. Restrictions apply.