energies

MDPI

Article

A Novel Deep Reinforcement Learning-Based Current Control Method for Direct Matrix Converters

Yao Li, Lin Qiu *, Xing Liu, Jien Ma, Jian Zhang and Youtong Fang

College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China
* Correspondence: qiu_lin@zju.edu.cn

Abstract: This paper presents the first approach to a current control problem for the direct matrix converter (DMC), which makes use of the deep reinforcement learning algorithm. The main objective of this paper is to solve the real-time capability issues of traditional control schemes (e.g., finite-set model predictive control) while maintaining feasible control performance. Firstly, a deep Q-network (DQN) algorithm is utilized to train an agent, which learns the optimal control policy through interaction with the DMC system without any plant-specific knowledge. Next, the trained agent is used to make computationally efficient online control decisions since the optimization process has been carried out in the training phase in advance. The novelty of this paper lies in presenting the first proof of concept by means of controlling the load phase currents of the DMC via the DQN algorithm to deal with the excessive computational burden. Finally, simulation and experimental results are given to demonstrate the effectiveness and feasibility of the proposed methodology for DMCs.

Keywords: matrix converter; current control; deep reinforcement learning; deep Q-network

Citation: Li, Y.; Qiu, L.; Liu, X.; Ma, J.; Zhang, J.; Fang, Y. A Novel Deep Reinforcement Learning-Based Current Control Method for Direct Matrix Converters. *Energies* **2023**, *16*, 2146. https://doi.org/10.3390/en16052146

Academic Editor: Fernando Sánchez Lasheras

Received: 7 February 2023
Revised: 17 February 2023
Accepted: 21 February 2023
Published: 22 February 2023

1. Introduction

The direct matrix converter (DMC) is a promising topology due to its numerous advantages, such as sinusoidal input and output currents, controllable input power factor, and compact design without a DC-link capacitor [1–3]. These prominent features make the DMC an alternative to the traditional back-to-back converter in various industrial applications where size and lifetime are critical issues.

In the past few decades, numerous modulation and control methods for DMCs have been introduced in the literature, among which the space vector modulation (SVM) has gained the most popularity for its inherent capability to track both the reference output voltage vector and input current vector simultaneously [4,5]. However, with the rapid development of digital processors and power devices, the SVM is now being challenged by the model predictive control (MPC) due to its simpler theoretical complexity, easier implementation, and better dynamic response [6–8]. The MPC method involves solving a finite-horizon optimization problem at each time step by predicting future system behavior, optimizing a cost function and applying only the first control input of the sequence. Although the MPC has been considered an emerging alternative to the traditional SVM, the computational burden of solving the optimization problem, the accurate modeling of system dynamics and constraints, and the selection of appropriate cost functions are well-known obstacles to its real-world applications [9–16]. The first major challenge is the computational complexity associated with solving the optimization problem due to the fast switching frequencies and complex dynamics of the system. This can result in high processing times and control delays. To address this challenge, various approaches have been proposed, such as reduced-order models and optimization algorithm improvements [9–11]. Another critical challenge is the precise modeling of the system dynamics and constraints, which are typically complex and nonlinear. Researchers have proposed adaptive and robust methods that can account for uncertainties and modeling errors in

real time [12–14]. Additionally, the selection of an appropriate cost function for the MPC controller is crucial, as it affects the control performance, energy efficiency, and system stability. Recent research has focused on developing new cost functions that can balance these competing objectives more effectively [15,16]. Addressing these challenges is crucial for the continued development and application of MPC in power electronics, and ongoing research is focused on developing new and improved methods to overcome these issues.

Recently, the fast growth of artificial intelligence technology has changed the traditional control strategy of the past few decades [17–19]. Reinforcement learning (RL) is a subfield of machine learning concerned with how an agent can learn to take actions that maximize a cumulative reward signal in an uncertain environment. RL is a powerful approach for building intelligent systems that can learn from experience and make decisions based on complex and dynamic inputs. In recent years, there has been growing interest in RL as a result of its success in a wide range of domains, from playing complex games such as Go and chess to controlling complex robotic systems. RL has also shown promise in addressing real-world problems, such as optimizing energy consumption and navigating autonomous vehicles [20–23]. In contrast to the MPC, RL agents try to find the optimal control policy during the training process before their real-world implementation, which makes it possible to avoid the computationally costly online optimization in each sampling period. Furthermore, the RL control method can be trained in field applications to take parameter variations and parasitic effects into account. As a result, RL has become an active research area with many ongoing studies exploring new algorithms, applications, and theoretical foundations.

Motivated by the aforementioned shortcomings of the MPC method and the superiority of the RL method, the potential of utilizing RL methods in power electronics is being explored [24–27]. Deep Q-Network (DQN) is a type of RL algorithm that uses a neural network to approximate the Q-function, which estimates the expected return for taking a particular action in a given state. DQN has shown promising results in various application scenarios with continuous states and discrete actions [28,29]. Although the DQN algorithm has emerged as a promising method for controlling power electronics systems, it faces significant challenges that need to be addressed. One of the primary challenges of using DQN in power electronics is the issue of high-dimensional state and action spaces. This can make it difficult to train the neural network effectively and can result in slow convergence and poor performance. Another challenge is the stability of DQN during training. DQN can suffer from issues such as overfitting, instability, and divergence, which can result in poor performance or even catastrophic failure of the controller. Addressing this challenge requires developing methods for stabilizing DQN during training, such as target network updating, experience replay, and parameter initialization. Due to the aforementioned challenges, no attempt has been made to incorporate the DQN algorithm with the DMC system.

In view of the above observations, this paper is concerned with a novel approach to the current control problem for the DMC, which makes use of the DQN algorithm. Specifically, an agent is trained without any plant-specific knowledge to find the optimal control policy by direct interaction with the system. Thus, the online optimization process is carried out in advance. The main merit of this proposal is that the computational burden problems can be alleviated by deploying the proposed solution. Furthermore, the proposal can be easily expanded to different power converters with finite switching states. Finally, the performance evaluation of the proposed methodology for DMCs in comparison to the state-of-the-art finite control set model predictive current control approach is given to confirm the effectiveness and feasibility of the proposal.

We contribute two main points to the relevant literature. (1) To the best of the authors' knowledge, this is the first time the DQN algorithm is incorporated with the current control method of the DMC. (2) Another important contribution of this paper is that the heavy computational burden can be reduced dramatically by the utilization of the trained agent

so as to carry out the optimization problem in the training phase in advance, which allows for low-cost processors.

2. Proposed DQN-Based Current Control Method for DMC

The common topology of a three-phase DMC is shown in Figure 1, which consists of nine bi-directional switches to connect the input voltage source to the output load. An input filter (L_i, R_i, C_i) is installed to eliminate high-frequency harmonics of the input current and reduce the input voltage distortion supplied to the DMC. The DMC performs AC/AC power conversion in a single stage, while the indirect matrix converters (IMC) achieve this in two stages, namely, rectification and inversion stages. The implementation of DMC requires 18 reverse-blocking IGBTs while the IMC consists of 12 reverse-blocking IGBTS and 6 reverse-conduction IGBTs. In comparison, the virtual DC-link stage of IMC makes it easier to construct with fewer switches, such as the sparse matrix converter, which is beyond the scope of this paper.

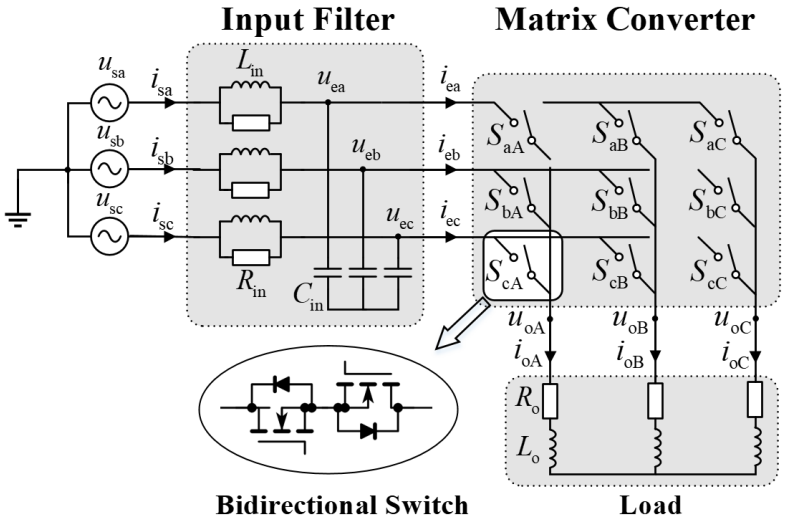

Figure 1. Common topology of the three-phase DMC.

According to Figure 1, the instantaneous relationship between the input and output quantities can be described as

$$\begin{bmatrix} u_{oA} \\ u_{oB} \\ u_{oC} \end{bmatrix} = \begin{bmatrix} S_{aA} & S_{bA} & S_{cA} \\ S_{aB} & S_{bB} & S_{cB} \\ S_{aC} & S_{bC} & S_{cC} \end{bmatrix} \begin{bmatrix} u_{ea} \\ u_{eb} \\ u_{ec} \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} i_{ea} \\ i_{eb} \\ i_{ec} \end{bmatrix} = \begin{bmatrix} S_{aA} & S_{aB} & S_{aC} \\ S_{bA} & S_{bB} & S_{bC} \\ S_{cA} & S_{cB} & S_{cC} \end{bmatrix} \begin{bmatrix} i_{oA} \\ i_{oB} \\ i_{oC} \end{bmatrix} \tag{2}$$

where u_{oA}, u_{oB}, u_{oC} and u_{ea}, u_{eb}, u_{ec} are the output and input phase voltages of the DMC, i_{oA}, i_{oB}, i_{oC} and i_{ea}, i_{eb}, i_{ec} are the output and input currents of the DMC, respectively, $S_{xy} = 1$ with $x \in (a, b, c)$ and $y \in (A, B, C)$ means the switch is on while $S_{xy} = 0$ means the switch is off.

For safe operation, the input phases should not be short-circuited, and the load should not be open-circuited. Thus, the switching constraints of the DMC can be expressed as

$$\begin{cases} S_{aA} + S_{bA} + S_{cA} = 1 \\ S_{aB} + S_{bB} + S_{cB} = 1 \\ S_{aC} + S_{bC} + S_{cC} = 1 \end{cases}. \tag{3}$$

Therefore, there are 27 valid switching states for the DMC.

The basic RL setting consists of an agent and environment. At each time step k, the agent observes the current state $\boldsymbol{O_k}$ of the environment, and an action A_k is taken according to the policy π. Based on $\boldsymbol{O_k}$ and A_k, the environment is updated to $\boldsymbol{O_{k+1}}$, and a reward R_k is produced, both of which are received by the agent. The observation–action–reward cycle continues until the training process is complete. The goal of the agent is to use RL algorithms to learn the best policy as it interacts with the environment so that given any state, it will always take the most optimal action that produces the most reward in the long run [30]. The action-value function $Q^\pi(\boldsymbol{O_k}, A_k)$ is introduced to evaluate the expected cumulative discounted reward as [28]

$$Q^\pi(\boldsymbol{O_k}, A_k) = \mathbb{E}\left\{\sum_{i=k}^{\infty} \gamma^{i-k} R_i \mid \boldsymbol{O} = \boldsymbol{O_k}, A = A_k\right\} \tag{4}$$

$$= \mathbb{E}\{R_k + \gamma Q^\pi(\boldsymbol{O}_{k+1}, A_{k+1}) \mid \boldsymbol{O} = \boldsymbol{O_k}, A = A_k\} \tag{5}$$

where $\gamma \in [0, 1)$ is the discount factor allowing the control task to be adjusted from short-sighted to far-sighted, and $\mathbb{E}\{\cdot\}$ denotes the expected value.

In the DMC, the observation consists of the measured input phase voltage ($u_{e\alpha}$, $u_{e\beta}$), output load current ($i_{o\alpha}$, $i_{o\beta}$), and the errors between the measured and the reference load current ($\Delta i_{o\alpha}$, $\Delta i_{o\beta}$), which looks as follows:

$$\boldsymbol{O} = \left[u_{e\alpha}, u_{e\beta}, i_{o\alpha}, i_{o\beta}, \Delta i_{o\alpha}, \Delta i_{o\beta}\right]. \tag{6}$$

According to the constraints in Equation (3), when only one zero switching state is included, the action space A contains 25 options, which can be defined as

$$A = \{S_0, S_1, S_2, \ldots, S_{24}\}. \tag{7}$$

To improve the policy of the agent with trial and error, an appropriate reward function should be designed. In this paper, the DMC should operate with the load current accurately following the reference value. Thus, the reward function is defined as

$$R = -(\Delta i_{o\alpha}^2 + \Delta i_{o\beta}^2). \tag{8}$$

The reference value of the load current is given as

$$\boldsymbol{i_o^*} = [I_{om}^* \cos\phi_o \quad I_{om}^* \cos(\phi_o - 2\pi/3) \quad I_{om}^* \cos(\phi_o + 2\pi/3)] \tag{9}$$

where ϕ_o is the expected angle of the load current and I_{om}^* is the amplitude of the expected load current.

According to Equations (4) and (5), the expectable return is represented by the action-value function based on the state–action pair at each time step. To maximize the expected cumulative reward over time, a new policy π' better than π can be found as

$$\pi'(\boldsymbol{O_k}) = \arg\max_A Q^\pi(\boldsymbol{O_k}, A) \tag{10}$$

Thus, one major challenge in the DQN algorithm is to derive an accurate mapping from state–action pairs to values. With the help of the neural network, the $Q^\pi(\boldsymbol{O_k}, A_k)$ can be estimated by a universal function approximator $Q_{\boldsymbol{\theta}}^\pi(\boldsymbol{O_k}, A_k)$ with weights and biases (critic parameters) represented by $\boldsymbol{\theta}$. The network has four layers: an input layer, two hidden layers, and an output layer. The hidden layers are fully connected, and the ReLU function is adopted as the activation function.

To train the network, state transition experiences $\mathcal{E}_i = \{\boldsymbol{O}_i, A_i, R_i, \boldsymbol{O}_{i+1}\}$ are stored in the experience buffer, from which a random mini-batch $\mathcal{M}$ of M experiences is sampled to update $\boldsymbol{\theta}$ by reformulating the Bellman equation in Equation (5) as a minimization problem of the loss L_Q:

$$\min_{\boldsymbol{\theta}} L_Q$$
$$\text{s.t. } L_Q = \frac{1}{M} \sum_{\mathcal{E}_\mathrm{i} \in \mathcal{M}} \left(Q^{\pi}_{\boldsymbol{\theta}}(\boldsymbol{O_i}, A_\mathrm{i}) - \left(R_\mathrm{i} + \gamma \max_{A} Q^{\pi}_{\boldsymbol{\theta}_t}(\boldsymbol{O_{i+1}}, A) \right) \right)^2 \tag{11}$$

where $Q^{\pi}_{\boldsymbol{\theta}_t}(\boldsymbol{O_i}, A_\mathrm{i})$ is the target critic, which improves the stability of the bootstrapping methods. The parameters $\boldsymbol{\theta_t}$ of the target network are updated periodically:

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}, \quad \text{after every } N_\mathrm{T} \text{ steps.} \tag{12}$$

At last, the tradeoff between exploration and exploitation is performed to avoid the learning algorithm converging into a suboptimal policy. Therefore, the ϵ-greedy policy is introduced as

$$A_\mathrm{k} = \begin{cases} \arg\max_{A} Q^{\pi}_{\boldsymbol{\theta}}(\boldsymbol{O_k}, A), & \text{with probability } 1-\epsilon \\ \text{a random element from } A, & \text{with probability } \epsilon \end{cases} \tag{13}$$

where ϵ updates at the end of each training step:

$$\epsilon = \epsilon \cdot (1 - \epsilon_{\text{decay}}). \tag{14}$$

Note that ϵ is set to zero when the training process has been completed. The schematic of the overall control structure with a learning routine is presented in Figure 2, and the learning pseudocode is given in Algorithm 1.

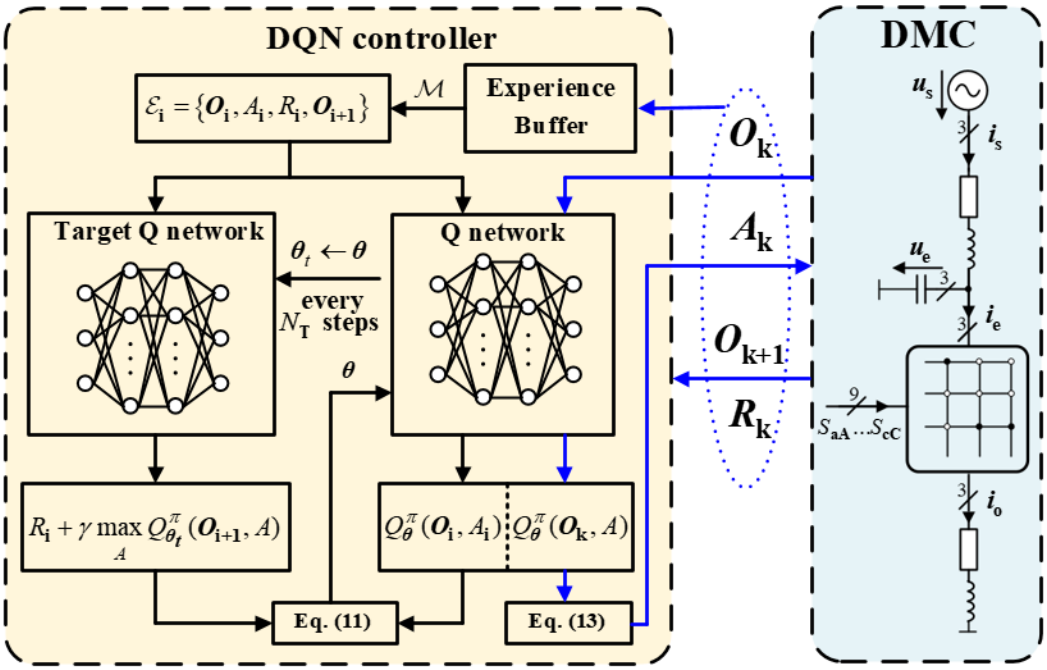

Figure 2. Schematic depiction of the DQN learning routine.

Algorithm 1 DQN pseudocode

Initialize the critic $Q_{\boldsymbol{\theta}}^{\pi}(\boldsymbol{O}, A)$ with random parameter values $\boldsymbol{\theta}$.
Initialize the target critic $Q_{\boldsymbol{\theta_t}}^{\pi}(\boldsymbol{O}, A)$ with parameters: $\boldsymbol{\theta_t} = \boldsymbol{\theta}$.
for episode=1 to max-episode **do**:
  Observe the initial state $\boldsymbol{O}_0$.
  for step=1 to max-step **do**:
    1. For the current observation $\boldsymbol{O}_{\mathrm{k}}$, select the action A_{k} based on Equations (13) and (14).
    2. Execute action A_{k}. Observe the next observation $\boldsymbol{O}_{\mathrm{k}+1}$ and reward R_{k}.
    3. Store ($\boldsymbol{O}_{\mathrm{k}}$, A_{k}, R_{k}, $\boldsymbol{O}_{\mathrm{k}+1}$) in the experience buffer.
    4. Sample a random mini-batch of experiences ($\boldsymbol{O}_{\mathrm{i}}$, A_{i}, R_{i}, $\boldsymbol{O}_{\mathrm{i}+1}$) from the experience buffer.
    5. Update the critic parameters using Equation (11).
    6. Update the target critic parameters using Equation (12).
    7. Reset the environment and break if $\boldsymbol{O}_{\mathrm{k}+1}$ is the terminal state.
  end for
end for

3. MPC Method for DMC

First, the input filter model is established for the prediction of input voltages and currents. In this paper, the LC filter with a damping resistor is adopted, as shown in Figure 3.

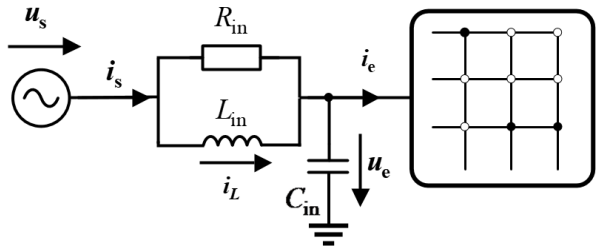

Figure 3. Circuit of the input filter.

The continuous system model of the input filter in Figure 3 can be described by the following equations:

$$\begin{aligned}\begin{bmatrix} \frac{du_e}{dt} \\ \frac{di_L}{dt} \end{bmatrix} &= A\begin{bmatrix} u_e \\ i_L \end{bmatrix} + B\begin{bmatrix} u_s \\ i_e \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{R_{in}C_{in}} & \frac{1}{C_{in}} \\ -\frac{1}{L_{in}} & 0 \end{bmatrix}\begin{bmatrix} u_e \\ i_L \end{bmatrix} + \begin{bmatrix} \frac{1}{R_{in}C_{in}} & -\frac{1}{C_{in}} \\ \frac{1}{L_{in}} & 0 \end{bmatrix}\begin{bmatrix} u_s \\ i_e \end{bmatrix}\end{aligned} \tag{15}$$

where L_{in}, C_{in}, and R_{in} are the filter inductance, the filter capacitance, and the filter damping resistance, respectively.

A discrete state space model can be derived when a forward Euler approximation is applied to a continuous-time system described in the state space form of Equation (15). Considering a sampling period T_s, the discrete-time input filter model can be described as

$$\begin{aligned}\begin{bmatrix} u_e(k+1) \\ i_L(k+1) \end{bmatrix} &= G\begin{bmatrix} u_e(k) \\ i_L(k) \end{bmatrix} + H\begin{bmatrix} u_s(k) \\ i_e(k) \end{bmatrix} \\ &= \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}\begin{bmatrix} u_e(k) \\ i_L(k) \end{bmatrix} + \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}\begin{bmatrix} u_s(k) \\ i_e(k) \end{bmatrix}\end{aligned} \tag{16}$$

where $G = e^{AT_s}$, $H = A^{-1}(G - I)B$. Using Equation (16), the value of u_e and i_L in the next sampling instant can be predicted.

The model of the resistance–inductance load is given by

$$\frac{d\boldsymbol{i}_o}{dt} = \frac{1}{L_o}(\boldsymbol{u}_o - R_o\boldsymbol{i}_o) \tag{17}$$

where L_o and R_o are the inductance and resistance of the load.

Similarly, using the forward Euler approximation, the equation for the load current prediction can be derived as

$$\boldsymbol{i}_o(k+1) = \left(1 - \frac{R_o}{L_o}T_s\right) \cdot \boldsymbol{i}_o(k) + \frac{T_s}{L_o}\boldsymbol{u}_o(k) \tag{18}$$

For 27 different switching states of the DMC, the corresponding load voltage vector $\boldsymbol{u}_o(k)$ and input current vector $\boldsymbol{i}_e(k)$ are calculated to predict the value of $\boldsymbol{i}_o(k+1)$, $\boldsymbol{i}_L(k+1)$, and $\boldsymbol{u}_e(k+1)$ in the next sampling interval. The source current $\boldsymbol{i}_s(k+1)$ is calculated by

$$\boldsymbol{i}_s(k+1) = \frac{\boldsymbol{u}_s(k+1) - \boldsymbol{u}_e(k+1)}{R_{in}} + \boldsymbol{i}_L(k+1). \tag{19}$$

The current control objectives of the DMC are twofold: to regulate the grid-side current $\boldsymbol{i}_s$ for unit power factor operation and to adjust the output current $\boldsymbol{i}_o$ for symmetrical and sinusoidal three-phase load current. The reference values for $\boldsymbol{i}_o$ are the same as Equation (9), and $\boldsymbol{i}_s$ and are defined as follows:

$$i_s^* = \begin{bmatrix} I_{sm}^* \cos\varphi_{in} & I_{sm}^* \cos(\varphi_{in} - 2\pi/3) & I_{sm}^* \cos(\varphi_{in} + 2\pi/3) \end{bmatrix}^T \tag{20}$$

where φ_{in} and I_{sm}^* are the expected phase angle and amplitude of the source current.

The errors of the predicted source current i_s^P and load current i_o^P in static two-phase coordinates can be expressed as

$$\Delta i_s = \begin{pmatrix} \Delta i_{s\alpha} & \Delta i_{s\beta} \end{pmatrix}^T = T_{abcto\alpha\beta}\left(i_s^P - i_s^*\right) \tag{21}$$

$$\Delta i_o = \begin{pmatrix} \Delta i_{o\alpha} & \Delta i_{o\beta} \end{pmatrix}^T = T_{abcto\alpha\beta}\left(i_o^P - i_o^*\right) \tag{22}$$

where

$$T_{\text{abcto}\alpha\beta} = \frac{2}{3}\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} \tag{23}$$

The cost function is designed to penalize differences from the reference value:

$$\begin{cases} g = \lambda * g_1 + g_2 \\ g_1 = \Delta i_{s\alpha}^2 + \Delta i_{s\beta}^2 \\ g_2 = \Delta i_{o\alpha}^2 + \Delta i_{o\beta}^2 \end{cases} \tag{24}$$

where λ is the weighting factor for the source current control. In this paper, the DQN method is trained to focus on the output current. Thus, for a fair comparison, λ is set to 0. In practice, $\lambda = 1$ provides a fairly good load current in comparison to $\lambda = 0$ due to the fact that $\boldsymbol{u}_e$ is controlled to be more sinusoidal.

In each sampling period, all 27 possible switching states are used to calculate the cost function, and the switching state corresponding to the minimum value of the cost function is applied to the DMC in the next sampling time.

In practical applications, due to the delay of the digital controller, the switching state selected at a certain moment can only be applied to the converter in the next moment, and the switching state applied at that moment may not be the optimal one for the next moment, which may result in significant errors. In order to make the selected optimal switching state act on the converter at a reasonable time, a two-step prediction strategy is

usually adopted. The specific implementation process is as follows: based on the sampled value of the current system state $x(k)$, predict the value of the controlled variable $x(k+1)$ in the next moment, and then further traverse all switching states based on this prediction to obtain the predicted value of the controlled variable $x(k+2)$ in moment $k+2$, which means the optimal switch is selected and applied to the system at moment $k+1$.

4. Results

To verify the effectiveness and feasibility of the proposed DQN-based current control method, a 3 × 3 DMC model is established, and the training of the DQN is handled with the use of the Reinforcement Learning Toolbox. Further, the experimental prototype (see Figure 4) has been built. The high-speed insulated gate bipolar transistor module (FF300R12KE4_E), which consists of two common-emitter-IGBTs, is used in the prototype. The controller includes a Digital Signal Processor (TMS320F28377) and Field Programmable Gate Array (10M50DAF484). The three-step commutation is implemented. The detailed model parameters are listed in Table 1, and the training parameters used in the DQN method are listed in Table 2.

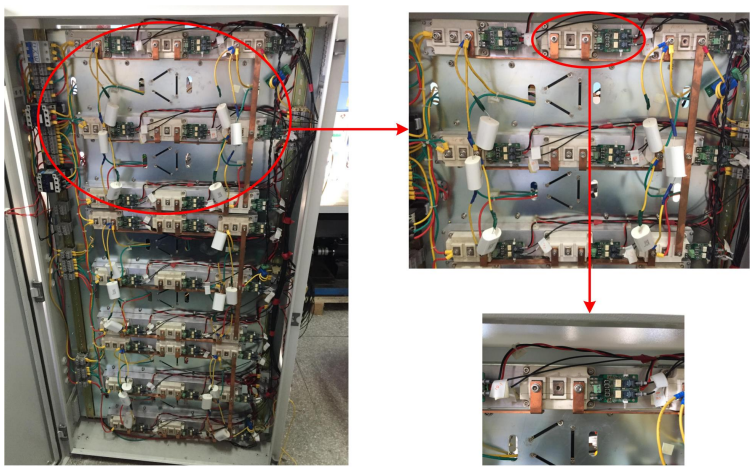

Figure 4. Experimental prototype.

Table 1. Circuit parameters of the DMC.

Parameters	Value
Source phase voltage (U_s)	50 V
Source voltage frequency (f_{in})	50 Hz
Sampling period (T_s)	200 µs
Input filter inductance (L_{in})	2 mH
Input filter capacitance (C_{in})	20 µF
Input filter resistance (R_{in})	20 Ω
Load frequency (f_o)	70 Hz
Load resistance (R_o)	10 Ω
Load inductance (L_o)	10 mH
Load current reference (I^*_{om})	3 A

Table 2. Training parameters of the DQN method.

Parameters	Value
Discount factor (λ)	0.85
Hidden network layer number (l)	2
Hidden layer 1 neuron number (n_1)	6
Hidden layer 2 neuron number (n_2)	8
Target network update frequency (N_T)	20
Mini-batch size (M)	256
Replay buffer size (D)	1×10^5
Maximum training steps (S)	1200
Maximum episode length (K)	2000

Figure 5 shows the output performance of the three-phase DMC with the MPC and proposed DQN methods. The input voltage of the DMC is set to 50 V, and a 3 A load current reference is imposed on the load. As is depicted in Figure 5, sinusoidal load currents are generated, which means the reference can be accurately tracked. From the perspective of waveform qualities, the proposed DQN method achieves a similar output performance to the MPC method.

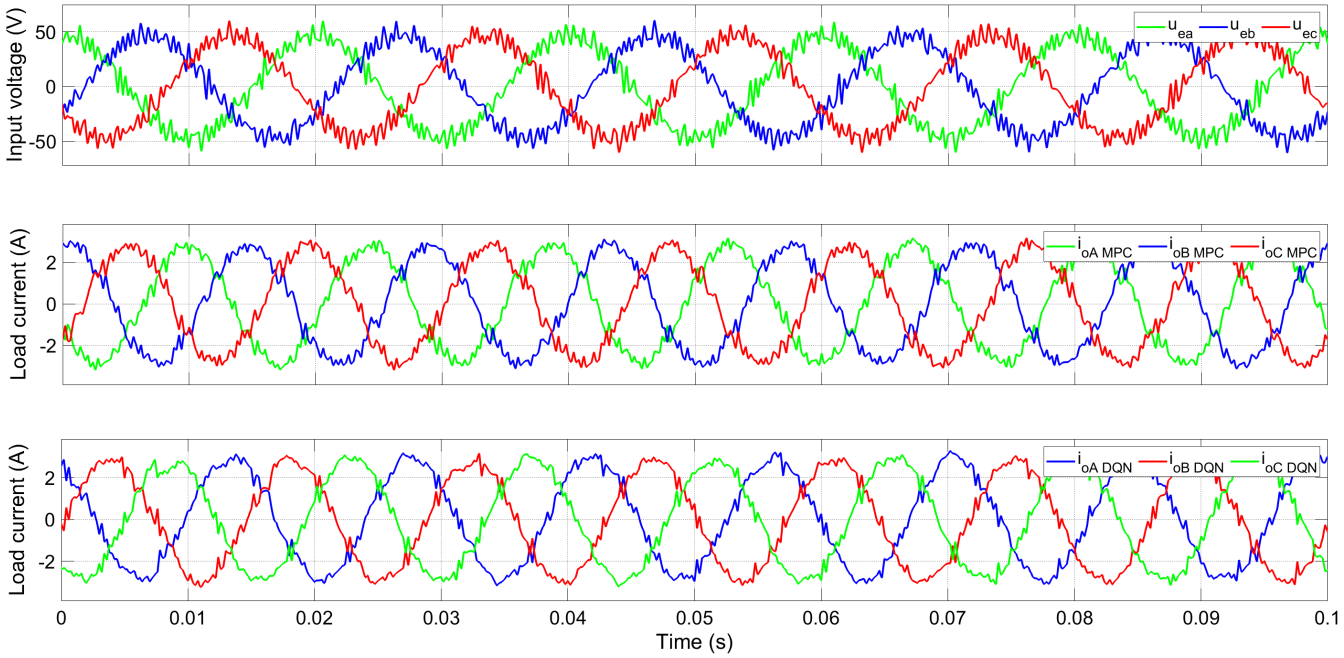

Figure 5. Comparison of the load current with the MPC and proposed method.

$$\text{MAE} = \frac{1}{N}\sum_{k=1}^{N}|I(k) - I^*(k)|$$
$$\text{MSE} = \frac{1}{N}\sum_{k=1}^{N}(I(k) - I^*(k))^2. \quad (25)$$

To present the comparison of the two aforementioned control schemes clearly, some measurements (defined in Equation (15)) in the steady state are listed in Table 3. The values of the total harmonic distortion (THD) show that MPC achieves slightly better performance, but it has higher mean absolute errors (MAE) and mean square errors (MSE) due to the fact that MPC does not ensure a zero error in the steady state. Based on the results, it can be indicated that the proposed DQN method has almost the same performance as the MPC method.

Table 3. System measurements of i_o.

Control Method	Measurement	Value
MPC	THD (i_o)	8.44%
	MAE (i_o)	0.398
	MSE (i_o)	0.202
DQN	THD (i_o)	8.73%
	MAE (i_o)	0.1536
	MSE (i_o)	0.0396

The goal of the proposed DQN method is to train an agent that learns the best policy as it interacts with the environment so that, given any state, it will always take the most optimal action that produces the most reward in the long run. As for MPC, the best switching state is selected by solving an optimization problem at each time step. The objective is to minimize a cost function that captures the desired behavior and any penalties for violating constraints. In this paper, the agent is trained to learn the policy that is similar to MPC.

However, the proposed DQN method is not identical to the MPC method. First, the policy used in the proposed method is pre-trained, which alleviates the time-consuming traversal process in the MPC method. Second, in the training process, a discount factor is used to compute the expected reward, which not only helps the agent to learn more quickly but also ensures the future reward. In this sense, the DQN method is more like a multi-step MPC. Third, the RL-based method has the potential to take parameter variations, parasitic effects, and commutation processes into account through online training.

After the training process, the parameters of the learned agent policy are obtained by using the function "getLearnableParameters". The weight w and bias b for each fully-connected layer are derived. The input variable x consists of the sampled input phase voltage, output load current, and the errors between the measured and the reference load current. A Relu function is adopted as the activation function for the output $y = wx + b$ of each hidden layer, which sets the negative value of y to zero. At last, the action corresponding to the output layer neuron with the maximum value is selected as the optimal switching state in this sampling interval.

Finally, experimental tests are conducted to verify the effectiveness of the proposed method. As shown in Figure 6, lower THD values are achieved by the MPC method. However, in the proposed DQN method, the agent is trained in a Simulink environment, which fails to consider the influence of the three-step commutation process of the DMC. Further, the training process of the agent might be improved, and the neural network can be optimized, which is beyond the scope of this paper. Although a deteriorated output current waveform is generated by the DQN method, the enumerating process in the MPC method is excluded for the reason that the agent is trained before its application. In this sense, the calculation time in each sampling period is significantly reduced, which means that the output performance of the proposed method can be improved by a higher sampling frequency.

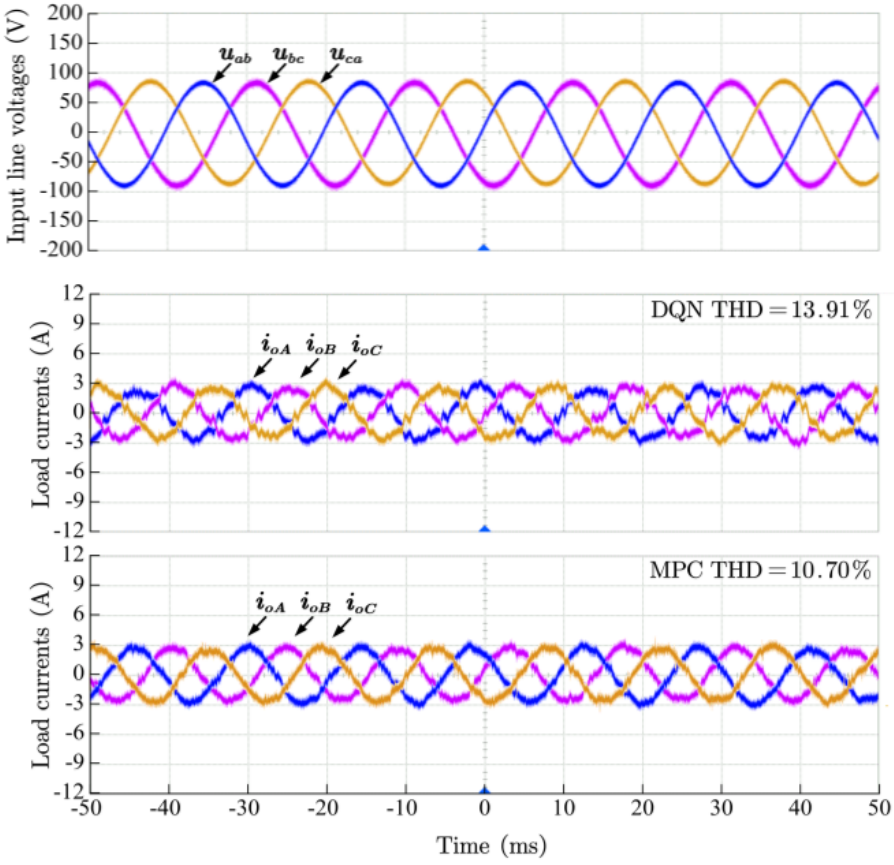

Figure 6. Experimental results of the load current with the MPC and proposed method.

5. Conclusions

In this paper, a novel DQN-based current control methodology for DMC systems was presented. By incorporating the DQN algorithm with the conventional current control method, we considered a fundamentally different solution to long-standing research problems with the use of an RL method. In addition, performance evaluations were provided to demonstrate the effectiveness and feasibility of the proposed methodology for DMCs. In the simulation, we showed that the proposed methodology can reduce the computational burden in comparison to the MPC method while maintaining feasible control performance. First, the time-consuming traversal process is replaced by an offline trained agent, making the proposed method available for a higher sampling frequency. Second, the agent is trained to ensure a zero error in the steady state, which achieves a smaller value of MAE and MSE in comparison to the MPC method. Third, the policy learned by the proposed method selects the optimal switching states in a similar manner to the MPC method. Therefore, the proposed DQN method achieves a similar output performance as the MPC method. However, in experiments, the proposed method fails to achieve a lower THD for the following reasons. First, the agent is trained in the Simulink environment, which neglects the commutation process and parasitic effects of the DMC. Second, the neural network can be improved by adding more hidden layers and neurons so as to fit the nonlinearity mapping from the high-dimensional input to the output. Finally, possible interesting directions for future research could be controlling multiple objectives such as common-mode voltage reduction and efficiency improvement and training an agent online.

Author Contributions: Conceptualization, Y.L., L.Q. and X.L.; Data curation, Y.L.; Formal analysis, Y.L.; Investigation, Y.L.; Methodology, Y.L. and X.L.; Project administration, L.Q. and Y.F.; Resources, Y.L.; Software, Y.L.; Supervision, J.M. and Y.F.; Validation, Y.L. and L.Q.; Visualization, Y.L. and J.Z.; Writing—original draft, Y.L. and X.L.; Writing—review and editing, J.M. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Program of China (2018YFB1201804), the National Natural Science Foundation of China under Grant (No. 52293424, 51827810, 51977192).

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Empringham, L.; Kolar, J.W.; Rodríguez, J.; Wheeler, P. W.; Clare, J.C. Technological issues and industrial application of matrix converters: A review. *IEEE Trans. Ind. Electron.* **2013**, *60*, 4260–4271. [CrossRef]
2. Gili, L.C.; Dias, J.C.; Lazzarin, T.B. Review, Challenges and Potential of AC/AC Matrix Converters CMC, MMMC, and M3C. *Energies* **2022**, *15*, 9421. [CrossRef]
3. Maidana, P.; Medina, C.; Rodas, J.; Maqueda, E.; Gregor, R.; Wheeler, P. Sliding-Mode Current Control with Exponential Reaching Law for a Three-Phase Induction Machine Fed by a Direct Matrix Converter. *Energies* **2022**, *15*, 8379. [CrossRef]
4. Casadei, D.; Grandi, G.; Serra, G.; Tani, A. Space vector control of matrix converters with unity input power factor and sinusoidal input/output waveforms. In Proceedings of the 1993 Fifth European Conference on Power Electronics and Applications, Brighton, UK, 13–16 September 1993.
5. Rodríguez, J.; Rivera, M.; Kolar, J.W.; Wheeler, P.W. A review of control and modulation methods for matrix converters. *IEEE Trans. Ind. Electron.* **2012**, *59*, 58–70. [CrossRef]
6. Rivera, M.; Wilson, A.; Rojas, C.A.; Rodriguez, J.; Espinoza, J.R.; Wheeler, P.W.; Empringham, L. A comparative assessment of model predictive current control and space vector modulation in a direct matrix converter. *IEEE Trans. Ind. Electron.* **2012**, *60*, 578–588. [CrossRef]
7. Liu, X.; Qiu, L.; Wu, W.; Ma, J.; Fang, Y.; Peng, Z.; Wang, D. Predictor-based neural network finite set predictive control for modular multilevel converter. *IEEE Trans. Ind. Electron.* **2021**, *68*, 11621–11627. [CrossRef]
8. Toledo, S.; Caballero, D.; Maqueda, E.; Cáceres, J.J.; Rivera, M.; Gregor, R.; Wheeler, P. Predictive Control Applied to Matrix Converters: A Systematic Literature Review. *Energies* **2022**, *15*, 7801. [CrossRef]
9. Mousavi, M.S.; Davari, S.A.; Nekoukar, V.; Garcia, C.; Rodriguez, J. Computationally Efficient Model-Free Predictive Control of Zero-Sequence Current in Dual Inverter Fed Induction Motor. *IEEE J. Emerg. Sel. Top. Power Electron.* **2022**. [CrossRef]
10. Mao, J.; Li, H.; Yang, L.; Zhang, H.; Liu, L.; Wang, X.; Tao, J. Non-Cascaded Model-Free Predictive Speed Control of SMPMSM Drive System. *IEEE Trans. Energy Convers.* **2022**, *37*, 153–162. [CrossRef]
11. Liu, X.; Qiu, L.; Wu, W.; Ma, J.; Fang, Y.; Peng, Z.; Wang, D. Event-Triggered Neural-Predictor-Based FCS-MPC for MMC. *IEEE Trans. Ind. Electron.* **2022**, *69*, 6433–6440. [CrossRef]
12. Liu, X.; Qiu, L.; Rodriguez, J.; Wu, W.; Ma, J.; Peng, Z.; Wang, D.; Fang, Y. Neural Predictor-Based Dynamic Surface Predictive Control for Power Converters. *IEEE Trans. Ind. Electron.* **2023**, *70*, 1057–1065. [CrossRef]
13. Wu, W.; Qiu, L.; Liu, X.; Ma, J.; Zhang, J.; Chen, M.; Fang, Y. Model-Free Sequential Predictive Control for MMC with Variable Candidate Set. *IEEE J. Emerg. Sel. Top. Power Electron.* **2021**. [CrossRef]
14. Xu, W.; Qu, S.; Zhang, C. Fast terminal sliding mode current control with adaptive extended state disturbance observer for PMSM system. *IEEE J. Emerg. Sel. Top. Power Electron.* **2023**, *11*, 418–431. [CrossRef]
15. Vazquez, S.; Rodriguez, J.; Rivera, M.; Franquelo, L.G.; Norambuena, M. Model Predictive Control for Power Converters and Drives: Advances and Trends. *IEEE Trans. Ind. Electron.* **2017**, *64*, 935–947. [CrossRef]
16. Dragičević, T.; Novak, M. Weighting Factor Design in Model Predictive Control of Power Electronic Converters: An Artificial Neural Network Approach. *IEEE Trans. Ind. Electron.* **2019**, *66*, 8870–8880. [CrossRef]
17. Li, D.; Ge, S.S.; Lee, T.H. Fixed-Time-Synchronized Consensus Control of Multiagent Systems. *IEEE Trans. Control Netw.* **2021**, *8*, 89–98. [CrossRef]
18. Li, Y.; Che, P.; Liu, C.; Wu, D.; Du, Y. Cross-scene pavement distress detection by a novel transfer learning framework. *Comput.-Aided Civ. Infrastruct. Eng.* **2021**, *36*, 1398–1415. [CrossRef]
19. Li, J.; Deng, Y.; Sun, W.; Li, W.; Li, R.; Li, Q.; Liu, Z. Resource Orchestration of Cloud-Edge–Based Smart Grid Fault Detection. *ACM Trans. Sens. Netw.* **2022**, *18*, 1–26. [CrossRef]
20. Chen, C.; Modares, H.; Xie, K.; Lewis, F.L.; Wan, Y.; Xie, S. Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics. *IEEE Trans. Automat. Contr.* **2019**, *64*, 4423–4438. [CrossRef]
21. Duan, J.; Yi, Z.; Shi, D.; Lin, C.; Lu, X.; Wang, Z. Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC–DC microgrids. *IEEE Trans. Ind. Informat.* **2019**, *15*, 5355–5364. [CrossRef]

22. Sun, L.; You, F. Machine learning and data-driven techniques for the control of smart power generation systems: An uncertainty handling perspective. *Engineering* **2021**, *7*, 1239–1247. [CrossRef]
23. Wang, N.; Gao, Y.; Zhao, H.; Ahn, C.K. Reinforcement learning-based optimal tracking control of an unknown unmanned surface vehicle. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3034–3045. [CrossRef] [PubMed]
24. Wei, C.; Zhang, Z.; Qiao, W.; Qu, L. An adaptive network-based reinforcement learning method for MPPT control of PMSG wind energy conversion systems. *IEEE Trans. Power Electron.* **2016**, *31*, 7837–7848. [CrossRef]
25. Zhao, S.; Blaabjerg, F.; Wang, H. An overview of artificial intelligence applications for power electronics. *IEEE Trans. Power Electron.* **2021**, *36*, 4633–4658. [CrossRef]
26. Tang, Y.; Hu, W.; Cao, D.; Hou, N.; Li, Y.; Chen, Z.; Blaabjerg, F. Artificial intelligence-aided minimum reactive power control for the DAB converter based on harmonic analysis method. *IEEE Trans. Power Electron.* **2021**, *36*, 9704–9710. [CrossRef]
27. Rodríguez, J.; Garcia, C.; Mora, A.; Flores-Bahamonde, F.; Acuna, P.; Novak, M.; Zhang, Y.; Tarisciotti, L.; Davari, S.A.; Zhang, Z.; et al. Latest advances of model predictive control in electrical drives—Part I: Basic concepts and advanced strategies. *IEEE Trans. Power Electron.* **2022**, *37*, 3927–3942. [CrossRef]
28. Schenke, M.; Wallscheid, O. A deep Q-learning direct torque controller for permanent magnet synchronous motors. *IEEE Open J. Ind. Electron. Soc.* **2021**, *2*, 388–400. [CrossRef]
29. Chen, Y.; Bai, J.; Kang, Y. A non-isolated single-inductor multi-port DC-DC topology deduction method based on reinforcement learning. *IEEE J. Emerg. Sel. Top. Power Electron.* **2022**. [CrossRef]
30. Schenke, M.; Kirchgassner, W.; Wallscheid, O. Controller design for electrical drives by deep reinforcement learning: A proof of concept. *IEEE Trans. Ind. Inform.* **2021**, *16*, 4650–4658. [CrossRef]