Integration of Electric Vehicles in Smart Grid using Deep Reinforcement Learning

Farkhondeh Kiaee
Department of Electrical and Computer Engineering, Faculty of Shariaty, Tehran Branch
Technical and Vocational University (TVU)
Tehran, Iran
kiaei@shariaty.ac.ir

***Abstract*—The vehicle-to-grid (V2G) technology provides an opportunity to generate revenue by selling electricity back to the grid at peak times when electricity is more expensive. Instead of sharing a contaminated pump handle at a gas station during the current covid-19 pandemic, plugging in the electric vehicle (EV) at home makes feel much safer. A V2G control algorithm is necessary to decide whether the electric vehicle (EV) should be charged or discharged in each hour. In this paper, we study the real-time V2G control problem under price uncertainty where the electricity price is determined dynamically every hour. Our model is inspired by the Deep Q-learning (DQN) algorithm which combines popular Q-learning with a deep neural network. The proposed Double-DQN model is an update of the DQN which maintains two distinct networks to select or evaluate an action. The Double-DQN algorithm is used to control charge/discharge operation in the hourly available electricity price in order to maximize the profit for the EV owner during the whole parking time. Experiment results show that our proposed method can work effectively in the real electricity market and it is able to increase the profit significantly compared with the other state-of-the-art EV charging schemes.**

***Index Terms*—vehicle-to-grid, deep Learning, electric vehicles, reinforcement learning, double Q-network.**

I. Introduction

Efficient integration of electric vehicles (EVs) in the distribution network is necessary for the implementation of smart city. Vehicle-to-grid (V2G) technology allows the EV to either provide power to the grid or take power from the grid in a bidirectional manner (Fig. 1). V2G operations increase during COVID-19 outbreak and EV owners are the real winners in the crisis. Charging the car cheaply at home feels a whole lot safer than sharing a contaminated gas station pump. The transport system that emerges from the COVID-19 crisis embraces V2G technology to make the smartest choices possible.

Smart control of V2G systems has been extensively investigated by researchers in the power grid community. We may classify the developed approaches according to their underlying assumptions into two main groups. The first group assume the future V2G information including driving, environment, pricing, and demand time series is known in advance and thus are not real-time. These methods seek for an optimization method towards optimum scheduling of charge/discharge

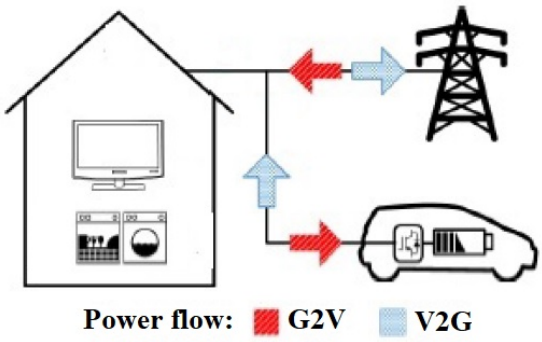

Fig. 1: Bidirectional V2G system connection.

actions, namely binary particle swarm optimization [1], fuzzy-logic method [2], Monte Carlo simulations [3], simulated annealing [4]. Several optimal control solutions based on dynamic programming (DP) algorithm are proposed in smart grid context [5]–[7]

The second group are more challenging due to the lack of the future information. The models in this group are mainly based on the learning systems. A subset of this group uses the historical data to predict the future required information (prediction based V2G systems). In particular, an AI model namely, support vector machine (SVM) [8], neural networks (NN) [9], or combination of SVM and NN [10] is trained to make real time decisions. In the other subset of methods, an agent explores the unknown V2G environment and learns the value associated to each action taken in different set of states (Reinforcement Learning (RL) integrated V2G systems). The instructed action-value function defines a decision criterion which helps to take the optimum actions in an immediate (real-time) manner [11], [12]. In this paper, we apply the deep Q-learning, the most established realization of RL approaches, to control V2G system. Our goal is to build a smart system that determines when to charge, discharge or be neutral (no charge/discharge), based on the current and historical system information.

For systems with continuous observation, the application of the neural networks to Q-learning, termed a Q-Network [13] is developed. Deep learning is the term given to neural networks with many layers, and has been shown to be effective

978-0-7381-4309-5/20/$31.00 ©2020 IEEE

in learning high level features from large input spaces [14], [15]. The training instability of Q-learning is addressed by the introduction of Double variant of Deep Q-Networks [16]. The Double DQN method reduces the correlations of the action-values with the target .

In this paper a Double-DQN V2G control system is proposed. The method is tested on the real *Nord pool* electricity price for decision making (charge/discharge). The Double-DQN V2G system is compared with DQN, shallow Q-network, and prediction based NN methods under diverse testing conditions. The comparisons show that the proposed method is much robust to different V2G uncertainty conditions and could make reliable profits under real electricity price markets.

The outline of this paper is as follows. Section II presents a general formulation of the V2G control problem. Section III describes the deep Q-learning approach and presents the implementation details of the Double-DQN V2G system. The experimental results and performance comparison with other methods are presented in Section IV. Finally, V concludes the paper.

II. V2G Control Problem Formulation

The V2G control problem is considered for a single EV arriving at time-step t_0. It is assumed that once the EV is parked, its departure time t_d and the expected state-of-charge (SoC) at departure, SoC_d, are notified. The hourly market pricing for purchasing and selling electricity (in Eur/kWh) is indicated by p_t. Let $0 \leq SoC_t \leq 1$ be the percentage of the battery power capacity that is available at time point t and l_t denote the time left for departure. The state space of the V2G control problem comprises of the pricing space, the SoC space, and the remaining time space. The state of the system at time t is then defined as $s_t = [p_t, b_t, l_t]$.

The action in the V2G control problem can be interpreted as choosing one control operation from the action space $A = \{\text{charge, discharge, neutral (no charge/discharge)}\}$. However, due to the constraints in the V2G control problem, not all the actions can be performed at a given state. The set of all possible actions given the state of the system is limited by two types of constraints in the V2G control problem.

The first constraint is that the EV must be charged to the expected SOC at departure. It must be considered that EV rate of charge/discharge is limited. Let C denote the charging rate which is the percentage of the battery energy capacity that can be charged per hour. We assume the absolute discharging rate is the same as the charging rate denoted by $-C$. The discharging action is then not allowed when $l_t \leq \lceil \frac{SoC_d - SoC_t}{C} \rceil$.

The V2G control requires a policy that accounts for charging characteristics of the EV battery in order to protect the EV battery from damaging. Then the charge and discharge actions are forbidden when the SoC is approaching the maximum (e.g., 95%) and minimum (e.g., 5%), respectively.

Reward function is a key ingredient of the reinforcement learning systems. The reward is defined as the financial revenue for the EV owner. As the charging/discharging operations are performed at the rate C, the energy flow is equal to EC, where E is the battery capacity of the EV in kWh. The corresponding rewards for the charging action is then negative i.e. $-p_t EC$ which means the money is paid by the owner. However, the corresponding rewards for the discharging action is positive ($p_t EC$), as the owner gains money during discharge actions by selling the battery energy to the grid.

III. Double Deep Q-Network V2G control system

Reinforcement learning (RL) is a general framework to deal with sequential decision tasks. Fig. 2 shows the schematic of the V2G control system using reinforcement learning with DQN method and its Double-DQN extension. At each time step t, RL observes the status s_t of the environment, takes an action a_t, and receives some reward r_t from the environment. With sufficient pairs of (s_t, a_t, r_t), RL can learn an optimal decision policy Q^* that maximizes the long-term accumulated reward.

$$Q^*(s,a) = max_\pi E_\pi\{R_t | s_t = s, a_t = a\} \tag{1}$$

The Q-function holds a nice property formulated as the Bellman equation:

$$Q^*(s_t, a_t) = r + \gamma max_a Q^*(s_{t+1}, a) \tag{2}$$

In the case of continuous state s, a neural network is often used to approximate the value $Q(s,a)$. This network is often referred as a Q-network [13]. If the Q-network involves multiple layers, we obtain the deep Q-learning architecture. It has been well-known that deep learning is capable of learning hierarchical patterns, and the patterns learned by the top-layers tend to be abstract and invariant against disturbance. The Q-network can be trained by minimizing the Q prediction error, i.e., the difference between the left-hand and right-hand side of Eq. 2. The loss function is then formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & L(\theta) = \sum_{i \in V} (y_i - Q(s_i, a_i \mid \theta))^2, \\ & y_i = r_i + \gamma max_a \tilde{Q}(s_{i+1}, a \mid \theta), \end{aligned} \tag{3}$$

where i and θ denote the training iteration and the parameters of the Q-network, respectively. The training examples are in the form of (s_i, a_i, r_i, s_{i+1}), and B denotes the buffer containing the recent training examples. Additionally, y_i is the prediction of $Q(s,a)$ given by the Bellman equation 2.

This loss function can be minimized by the stochastic gradient descend (SGD) algorithm. The gradient with respect to θ is given by

$$\nabla_\theta L = \sum_{i \in V} (y_i - Q(s_i, a_i \mid \theta)) \nabla_\theta Q(s_i, a_i \mid \theta) \tag{4}$$

where $\nabla_\theta Q(s_i, a_i \mid \theta)$ can be easily computed by the back-propagate (BP) algorithm.

We avoid the divergence of direct implementation of the trading system with neural networks due to using the same Q-network in calculating the target value y_i in (3). Our solution

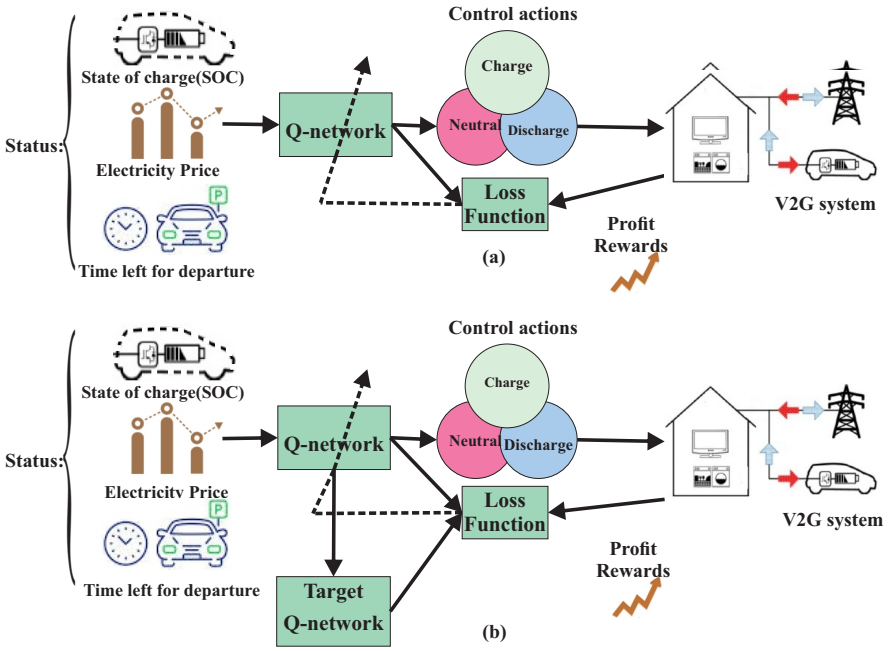

Fig. 2: Schematic of the RL-based V2G control system. (a) DQN , (b) Double-DQN.

Algorithm 1 Training process of the Double-DQN V2G control system
function Double-DQN-V2G$\{s_i = [p_i, b_i, l_i],\ i = 1, \ldots, t_d\}$

Initialize replay buffer B to a defined capacity.
Initialize action-value Q-function with random weights θ
Initialize target action-value $\tilde{Q}$-function with weights $\tilde{\theta} \leftarrow \theta$
for episode 1,M **do**
    Receive initial electricity price sequence p_1 and form initial state $s_1 = [p_1, b_1, l_1]$
    for t steps **do**
        With probability ϵ select a random action, otherwise select $a_t = \arg max_a Q(s_t, a; \theta)$
        Execute a_t and observe reward r and new price sequence p_{t+1} (next state: $s_{t+1} = [p_{t+1}, b_{t+1}, l_{t+1}]$)
        Store transition $\{s_t, a_t, r_t, s_{t+1}\}$ in buffer
        Sample random batch of transitions from B
        Perform a gradient descent step on (4) with respect to the network parameters θ
        Update the target network using (5).
    end for
end for
Return $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$

is similar to the target network used in Fig. 2 (b) for Q-learning. The authors in [16] show that it is required to have a target $\tilde{Q}$ to have stable targets y_i in order to train the system, consistently. A copy of the Q-network is created and then used for calculating the target values. The weights of the target $\tilde{Q}$ network (indicated by $\tilde{\theta}$ are softly updated by interpolating with the latest θ, as follows:

$$\tilde{\theta} = \tau\theta + (1 - \tau)\tilde{\theta}, \tag{5}$$

where τ is the interpolation factor. The relatively unstable problem of learning the action-value function is then moved closer to a case of robust supervised learning problem. Although, the delay in the update of target values may slow learning, in practice the stability of learning is greatly outweighed. Note that the decision of the trading is made based on the target network $\tilde{Q}$, rather than the present network Q. An overview of the Double-DQN method for V2G system control is outlined in Algorithm. 1.

IV. Experimental Result

In the results reported in this section, our proposed real-time V2G control algorithm is evaluated using the actual electricity spot price data of Oslo zone from *Nord pool* power market illustrated in Fig. 3. The dataset consists of historical hourly

electricity prices (Eur/MWh) from January 1, to December 31 corresponding to the year 2019. For simplicity, we consider the case where the proposed V2G control algorithm is run for a single EV on different days with exactly the same conditions. We consider the vehicle type, Nissan Leaf 2016, with battery energy capacity of $E = 24kWh$. The charging rate C = 0.1. We assume that the EV arrives at 8:00 in the morning every day and departs at 16:00 in the afternoon with expected departure SoC of 70%. The range of arrival SoC is selected to be uniformly distributed in $[0.2, 0.8]$.

We compare the proposed Double-DQN V2G system with the performance of standard Deep Q-network (DQN) and Shallow Q-network (SQN). Moreover, the performance of proposed method is compared with the prediction-based NNs namely, Shallow Neural Network (SNN), Convolutional Deep Neural Network (CDNN), RNN and long short-term memory (LSTM).

The goal of the prediction-based NN is to predict whether the electricity price of the next hour is going higher, lower, or suffering no change which corresponds to taking discharge, charge or neutral actions, respectively. In models based on shallow networks i.e. SQN and SNN, a network with two layers is combined with clustering and feature selection [17]. A small representation of data using clustering. A sequential stepwise process, called Backward Greedy Selection, is then used to remove variables (features) that are irrelevant to the neural network performance. The hidden layer of the network is using a sigmoid transfer function and the output layer is linear, trained with the Levenberg-Marquardt algorithm.

The python-based DL package tensorflow is used to implement the deep V2G control structures. Tensorflow provides the benchmark implementations of convolution, pooling and fully-connected layers for public usages. The DQN is composed of five layers: 1) an input layer (52 dimension input composed of the 50 delta electricity price $p_t - p_{t-1}$ of 50 hours in the past along with their corresponding SoC and the time left for departure); 2) a convolutional layer with 64 convolutional kernels (each kernel is of length 12); 3) a max pooling layer; 4) a fully connected dense layer; and 5) a soft-max layer with three outputs. The RNN contains an input layer, a dense layer (128 hidden neurons), a recurrent layer, and a soft-max layer for classification. The LSTM shares the same configuration as RNN except for replacing the recurrent layer with the LSTM module.

The training strategy of the deep networks is composed of an iterative update of the weights in an online manner. In practice, the first 1500 time points are used to set up the network weights. At each parking hour, a new training example (s_t, a_t, r_t, s_{t+1}) is added to a defined buffer B (with a finite capacity) that consists of recent V2G system history. The examples in the buffer are used as a mini-batch to train the Q-network following Eq. 3. The trained system is then exploited to control the V2G system from 1501 to 2000. In the next iteration, the sliding window of the training data is moved 500 ticks forward covering a new training set from 500 to 2000 (Note that the first 500 time points of the input time

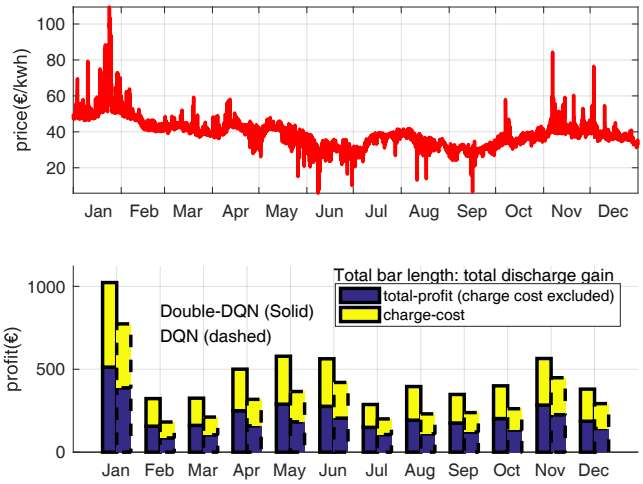

Fig. 3: (Top panel) Hourly prices for the Nordpool electricity market (Bottom panel) Proposed Double DQN vs DQN performance

series employed for system initialization will not be used for test phase). The parameters in the network are then iteratively updated with the recently released data. This online updating strategy allows the model to get aware of the latest V2G system condition and revise its parameters accordingly.

The performances of the proposed Double-DQN system on the electricity price data from January to December 2019 are provided in Fig. 3. The total profit (blue part of the bar) consists of the total gain obtained using energy discharged (total bar length) excluding the cost of the energy charged (yellow part of the bar) during each month. From the results, it is observed that compared with the DQN system (bars with dashed line-edges), the Double-DQN system (bars with solid line-edges) makes more profits. It is because of the increased stability of the Double-DQN due to attaching the target network to the structure of the standard DQN. In the Double-DQN structure, an iterative update adjusts the action-values (Q) towards target values that are only periodically updated. The network is then given more time to consider many recent actions instead of updating all the time, thereby reducing correlations with the target and resulting in a more robust model.

The performance results of the proposed algorithm when the departure SOC varies from 50% to 80% and the battery charging rate varies from 0.1 to 0.2 are shown in Table I. As we can see from the Table, for all V2G control methods, the average monthly profit increases as the charging rate increases while the departure SoC stays the same. This is because increasing charging rate gives rise to decreasing required charging hours which means that more hours during the parking time can be utilized to exchange power in order to increase the revenue. If we compare the result of different departure SOC at the same charging rate, it can be found that increasing the departure SOC can result in an decreasing monthly profit. This is due to

TABLE I: Results of the average monthly profit under different charging rates C and expected departure SOCs.

		Double DQN	DQN	SQN	CDNN	LSTM	RNN	SNN
C=0.1	SoC=0.5	521	438	302	283	431	119	91
	SoC=0.6	513	425	294	275	427	95	88
	SoC=0.7	502	411	288	264	410	79	72
	SoC=0.8	488	396	276	251	400	66	59
C=0.2	SoC=0.5	527	441	311	295	438	121	93
	SoC=0.6	519	432	303	281	429	101	90
	SoC=0.7	511	417	295	275	421	83	81
	SoC=0.8	505	408	282	262	409	74	69

the fact that more hours are required for charging in order to meet the increasing departure SOC, which costs more money.

The results in Table I shows that in all charging rate and departure SoC setting, the highest profits are made by Double-DQN V2G system. This is due to its novel structure which allows simultaneous environment sensing and optimum action learning for V2G system control.

When taking the results of CDNN, RNN, LSTM and SNN into considerations, the pitfalls of prediction-based NN methods become apparent. By examining the total profit values in Table I, only the LSTM could make comparable profits while other deep RL-based systems. This is because prediction-based systems only consider the electricity price market to make decisions. The Double-DQN learns both price condition and the action-value function $Q(s; a)$ in a joint framework. Moreover, electricity price signal is not like other stationary or structured sequential signals, such as music, that exhibit periodic and repeated patterns. The conventional RNN configurations recursively remember the historical price information cannot show an acceptable performance. Instead, the LSTM only takes the current and the recent rice history into consideration which helps the system to take a relatively better actions.

V. Conclusion

Emerging smart grid systems will allow for more control over the charging of EVs. In this work we presented a V2G control system based on the deep Q-network (DQN) structure. The system is composed of two major components: a deep learning component that learns the system status, and a Q-learning component that learns the action-value function. However, the two components are integrated as one, in the real implementation of the system. In order to obtain consistent targets during temporal difference calculations, a separate target network is considered in the system thereby forming the final Double-DQN structure. Experimental results show that the proposed method outperforms the other state-of-the-art deep V2G systems. The results on real electricity price demonstrate the effectiveness of the learning system in simultaneous system condition summarization and optimal action learning.

References

[1] Vishu Gupta, Srikanth Reddy, Lokesh Panwar, Rajesh Kumar, and BK Panigrahi, "Optimal v2g and g2v operation of electric vehicles using binary hybrid particle swarm optimization and gravitational search algorithm," in *2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2017, pp. 157–160.

[2] Felice De Luca, Vito Calderaro, and Vincenzo Galdi, "A fuzzy logic-based control algorithm for the recharge/v2g of a nine-phase integrated on-board battery charger," *Electronics*, vol. 9, no. 6, pp. 946, 2020.

[3] Claes Sandels, Ulrik Franke, Niklas Ingvar, Lars Nordstrom, and Roberth Hamren, "Vehicle to gridmonte carlo simulations for optimal aggregator strategies," in *Power System Technology (POWERCON), 2010 International Conference on*. IEEE, 2010, pp. 1–8.

[4] Tiago Sousa, Tiago Soares, Hugo Morais, Rui Castro, and Zita Vale, "Simulated annealing to handle energy and ancillary services joint management considering electric vehicles," *Electric Power Systems Research*, vol. 136, pp. 383–397, 2016.

[5] Yu Wu, Alexandre Ravey, Daniela Chrenko, and Abdellatif Miraoui, "Demand side energy management of ev charging stations by approximate dynamic programming," *Energy Conversion and Management*, vol. 196, pp. 878–890, 2019.

[6] Zhong-kai Feng, Wen-jing Niu, Chun-tian Cheng, and Xin-yu Wu, "Optimization of large-scale hydropower system peak operation with hybrid dynamic programming and domain knowledge," *Journal of Cleaner Production*, vol. 171, pp. 390–402, 2018.

[7] Ahmed Abdulaal, Mehmet H Cintuglu, Shihab Asfour, and Osama A Mohammed, "Solving the multivariant ev routing problem incorporating v2g and g2v options," *IEEE Transactions on Transportation Electrification*, vol. 3, no. 1, pp. 238–248, 2016.

[8] Yu-Wei Chung, Behnam Khaki, Chicheng Chu, and Rajit Gadh, "Electric vehicle user behavior prediction using hybrid kernel density estimator," in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*. IEEE, 2018, pp. 1–6.

[9] Qiyun Dang, Di Wu, and Benoit Boulet, "Ev charging management with ann-based electricity price forecasting," in *2020 IEEE Transportation Electrification Conference & Expo (ITEC)*. IEEE, 2020, pp. 626–630.

[10] Yu-Wei Chung, Behnam Khaki, Tianyi Li, Chicheng Chu, and Rajit Gadh, "Ensemble machine learning-based algorithm for electric vehicle user behavior prediction," *Applied Energy*, vol. 254, pp. 113732, 2019.

[11] Sunyong Kim and Hyuk Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, no. 8, pp. 2010, 2018.

[12] Qiyun Dang, Di Wu, and Benoit Boulet, "A q-learning based charging scheduling scheme for electric vehicles," in *2019 IEEE Transportation Electrification Conference and Expo (ITEC)*. IEEE, 2019, pp. 1–5.

[13] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al., "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[14] Farkhondeh Kiaee, Christian Gagné, and Mahdieh Abbasi, "Alternating direction method of multipliers for sparse convolutional neural networks," *arXiv preprint arXiv:1611.01590*, 2016.

[15] Farkhondeh Kiaee, Hamed Fahimi, and Hossein Rabbani, "Intra-retinal layer segmentation of optical coherence tomography using 3d fully convolutional networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2795–2799.

[16] Hado Van Hasselt, Arthur Guez, and David Silver, "Deep reinforcement learning with double q-learning," *arXiv preprint arXiv:1509.06461*, 2015.

[17] Karol Lina López, Christian Gagné, and Marc-André Gardner, "Demand-side management using deep learning for smart charging of electric vehicles," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2683–2691, 2018.