2021 26th International Computer Conference, Computer Society of Iran (CSICC) | 978-1-6654-1241-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/CSICC52343.2021.9420561

Optimal ATM Cash Replenishment Planning in a Smart City using Deep Q-Network

1st Mohammadhossein Kiyaei
Finance Dept., Faculty of Management & Accounting
Farabi Campus, University of Tehran
Qom, Iran
mh.kiaee@alumni.ut.ac.ir

2nd Farkhondeh Kiaee
Dept. of Electrical & Computer Engineering, Faculty of Shariaty
Technical and Vocational University (TVU)
Tehran, Iran
kiaei@shariaty.ac.ir

***Abstract*—ATMs are no longer just machines, these connected devices are smart, intelligent things in the Internet of Things (IoT). Access to cash for many in society is remaining essential during the current COVID-19 lock-down around the globe. A cash inventory management system is necessary to decide whether ATM should be replenished on each day of the week. In this paper, we study the real-time cash replenishment planning problem under outflow uncertainty where the fee of the security companies grows if the replenishment ends up falling on a weekends/holidays. Our model is based by the Double Deep Q-Network (DQN) algorithm which combines popular Q-learning with a deep neural network. The proposed method is used to control replenishment operation in order to minimize replenishment cost where the cash demand changes dynamically at each day. Experiment results show that our proposed method can work effectively on the real outflow time-series and it is able to reduce the ATM operational cost compared with the other state-of-the-art cash demand prediction schemes.**

***Index Terms*—cash replenishment planning, deep learning, ATM, reinforcement learning, double Q-network.**

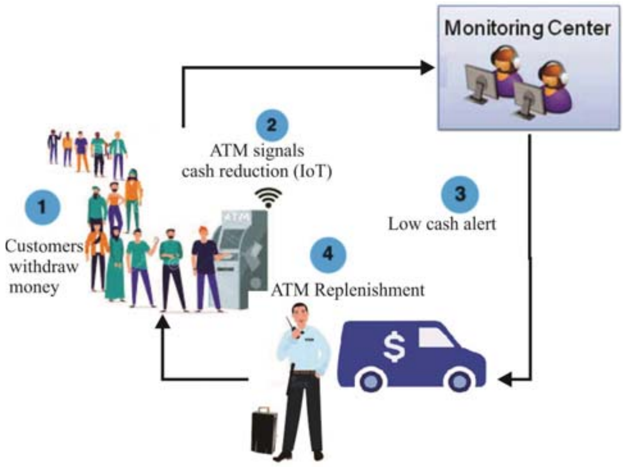

Fig. 1. Overview of the CRP system.

I. Introduction

A smart city requires the banking industry to find a completely new way of thinking about cash-points, branch offices, and the services it offers. IoT and Artificial Intelligence (AI) allows a bank to track its ATM network to help predict outages due to cash shortages. The ATM is playing an even more critical role in ensuring that consumers have access to cash and wider banking services while branches have reduced hours or closures, or customers want to avoid face-to-face or in branch interactions completely. The modern banking industry that emerges from the COVID-19 crisis embraces Artificial Intelligence (AI) and the Internet of Things (IoT) that will entirely replace humans by taking management decisions.

Cash replenishment planning (CRP) system helps formulate a cost efficient operating plan that includes the fee for secure ATM replenishment service. The overview of the system is shown in Fig. 1. The ATMs using the IoT technology are linked with the monitoring center, which enables automatically analysing the cash remaining in the ATM and interacting with the security company necessary for replenishment plan execution.

CRP problem for ATMs has been extensively investigated by researchers in the economic and financial management community. We may classify the developed approaches according to their underlying assumptions into two main groups. The first group assume the future ATM cash withdrawal information, is known in advance and thus are not real-time. These methods seek for an optimization method towards optimum planning of replenishment, namely particle swarm optimization [1], Heuristic Optimization [2], and decision trees [3]. Several optimal CRP solutions based on dynamic programming (DP) algorithm are proposed in literature [4]–[6]

The second group is more challenging due to the lack of cash demand information. The models in this group are mainly based on the learning systems and use the historical data to predict the future required information (prediction based CRP). In particular, an AI model namely, feed-forward neural networks (NNs) [7], [8], recursive NNs [9], [10], or combination of clustering and NNs [11], [12] is trained to make real-time decisions.

In this paper, an AI decision-making model based on reinforcement learning (RL) is introduced. In the proposed model, an agent explores the unknown CRP environment and learns the value associated to each action taken in a different set of states. The instructed action-value function defines a

978-1-6654-1241-4/21/$31.00 ©2021 IEEE

decision criterion which helps to take the optimum actions in an immediate (real-time) manner [13].

In particular, in this paper, the deep Q-learning as the most established realization of RL approaches, is applied to the CRP system. Our goal is to build a smart system that determines when to perform replenishment, 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 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 effectively reduces the correlations of the action-values with the target [17].

In this paper a Double DQN+IoT method is proposed which benefits from both IoT and AI technologies. The method is tested on the real *NN5* dataset that contains daily ATM cash withdrawal amounts over 2 years across the UK. the proposed method can efficiently model cash demand uncertainty and take into account replenishment cost variability in order to reduce ATM operational cost. The proposed Double-DQN+IoT system is compared with the system based solely on the IoT, shallow Q-network, and prediction based NN methods under diverse testing conditions. The comparisons show that the proposed method can make reliable cost savings under real ATM outflow dataset.

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

II. Problem Formulation

The aim of CRP system is to reduce replenishment cost through efficient planning that benefits both the financial institutions and customers by reducing the ATM downtime without increasing ATM operational costs.

Let $0 \leq b_t \leq 1$ be the percentage of the remaining ATM cash that is available at time point t and l_t denote the elapsed downtime after ATM runs out of money. If the ATM runs out of cash, customers are dissatisfied due to bad service. The maximum allowable ATM downtime due to lack of cash is then restricted to be D_{max} days. It is assumed that the contract of financial institution with the security company is on a per replenishment basis without a fixed fee for all week days. The cost of replenishment runs on business days is 10$. However, if replenishment day falls on a weekend/holiday, the security company performs replenishment by the fee of 30$.

The state space of the cash replenishment planning problem comprises of the ATM remaining cash space, the replenishment cost space and the elapsed ATM downtime space. The state of the system at time t is then defined as $s_t = [b_t, p_t, l_t]$ The action in the cash replenishment planning can be interpreted as choosing one operation from the action space $A = \{\text{replenishment, no-action}\}$. However, due to the constraints in the 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 the maximum allowable ATM downtime due to lack of cash D_{max} . Taking "no action" is then not allowed when $l_t > D_{max}$.

Reward function is a key ingredient of the reinforcement learning systems. The reward is defined as the ATM operational cost savings for the financial institutes. The corresponding rewards for the replenishment action is then negative of the money paid by the financial institute to the security company. However, the corresponding rewards for the no (replenishment) action is positive, as the financial institute saves money by reducing ATM operational cost.

III. Double-DQN CRP System

Reinforcement learning (RL) is a general framework to deal with sequential decision tasks. Fig. 2 shows the schematic of the CRP system using RL with Double-DQN method. 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. The RL method suggests that, given sufficient pairs of (s_t, a_t, r_t), the optimal policy Q^* is to maximize 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 Bellman equation property formulated as:

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

For systems with 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 consists of several layers of nodes, we obtain the deep Q-learning architecture. Deep learning has been known to have the ability to learn hierarchical patterns, and the patterns learned by the upper layers tend to be abstract and invariant against disturbance. The Q-network is trained to minimize 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. In addition, y_i is the prediction of $Q(s,a)$ given by the Bellman Eq. 2.

The deep loss functions are typically minimized using stochastic gradient descend (SGD) algorithm. The gradient of Eq. 3 with respect to θ is given by

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

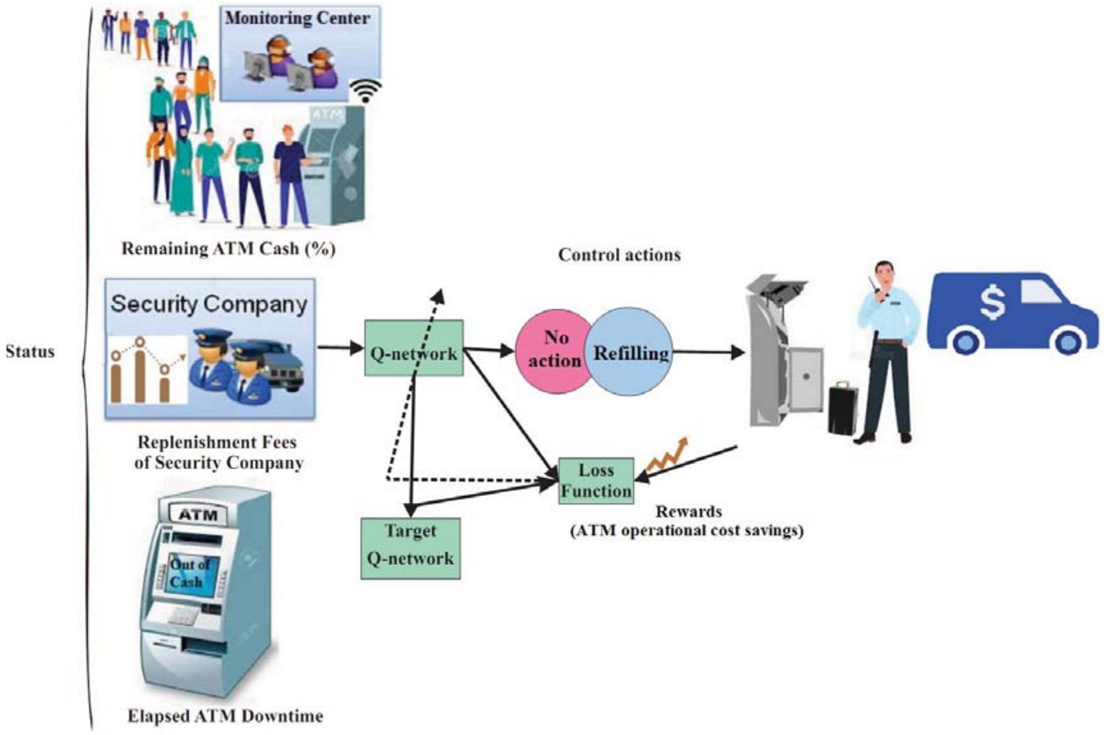

Fig. 2. Overview of the proposed Double-DQN CRP system.

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 system with neural networks due to using the same Q-network in calculating the target value y_i in (3). Our solution is similar to the target network used in Fig. 2 for Q-learning. The authors in [16] show that stable targets y_i are required in order to train the system, consistently. A copy of the Q-network (represented by $\tilde{Q}$) is then created and 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 then approaches to a case of robust supervised learning problem. Although, the delays in updating the target values may slow learning, in practice the stability of learning is greatly outweighed. Note that the decision of the replenishment is made based on the target network $\tilde{Q}$, rather than the present network Q.

IV. Performance Evaluation

In the results reported in this section, our proposed CRP system is evaluated using the actual cash demand *NN5* dataset [12] for two ATMs with different daily outflow characteristics illustrated in Fig. 3. The ATM-1 dataset consists of historical daily outflow from 01 Nov 2003, to 16 Nov 2005. The dataset corresponding to ATM-2 includes daily outflow form 01 Jan 2002 to 17 Dec 2003. In order to improve the customer convenience, the maximum allowable ATM downtime due to lack of cash, D_{max}, is selected to be 2 days. If the replenishment ends up falling on a weekends i.e. Saturday and Sunday, the fee of the security companies is considered to be three times greater than business days (e.g. 30$ for weekends vs 10$ for business days is considered here for the replenishment cost computations).

In the evaluation, we consider two different scenarios:

- CRP method based on Internet of Things (IoT): In this method, the ATM remaining cash (%) is signalled to the monitoring center by the smart ATM. Therefore, replenishment is done if the elapsed ATM downtime is more than D_{max} days.
- CRP method using both IoT and proposed Double-DQN method (IoT+DQN): In this method, using the proposed DQN method, the optimal action is taken based on the learned Q-value function. In other words, depending on the current state of the system, the real-time decision to whether replenish the ATM's cash or not, is made automatically. The signalled ATM remaining cash, the elapsed ATM downtime, and the replenishment fee determine the state of the system.

The average monthly replenishment cost of the two different scenarios on the *NN5* dataset are provided in Fig. 3. As the DQN algorithm learns from experience and adapts to different cash demands and variable costs, the monthly cost of the proposed algorithm begins to decrease. From the results, it is observed that after the initial training, which takes about

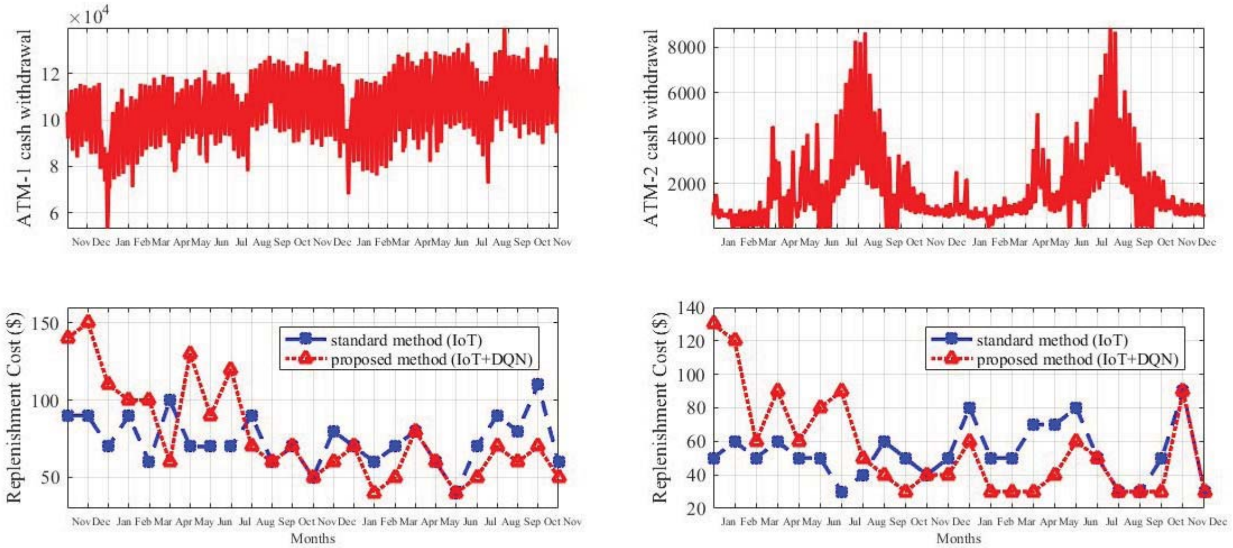

Fig. 3. Top: Daily cash flow of the two ATMs (NN5 dataset). Bottom: Performance of the proposed IoT+DQN system vs the system based solely on the IoT.

12 months, the proposed IoT+DQN system makes more cost savings than the system based solely on the IoT.

Figure 4 shows, the actual actions taken in the IoT method (Scenario 1) and the IoT+DQN method (Scenario 2) on different days of the week corresponding to ATM-1 data during June and July 2005. As indicated by the arrow in certain places in the figure, in the second scenario the replenishment operation is not postponed to the 6th and 7th days of the week, which successfully reduces the ATM operational cost.

We compare the proposed Double-DQN CRP system with the performance of 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 cash demand of the next day is going to surpass the ATM remaining cash which corresponds to taking replenishment action. Two models based on shallow networks i.e. SQN and SNN is considered where a network with two layers is combined with clustering and feature selection [18]. A small representation of data is first achieved 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. A sigmoid transfer function is used in the hidden layer of the network and the output layer is linear, trained with the Levenberg-Marquardt algorithm.

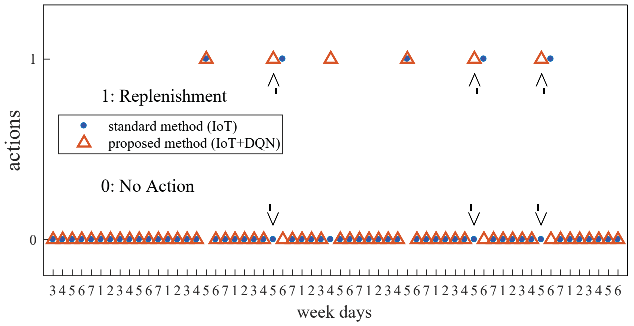

Fig. 4. Review of the actions taken by the systems during June & July.

The python-based DL package tensorflow is used to implement the deep CRP system structures. Tensorflow provides the benchmark implementations of convolution, pooling and fully-connected layers for public usages. The proposed DQN is composed of five layers: 1) an input layer (28 dimension input composed of the 26 ATM remaining cash of 26 days in the past along with their corresponding replenishment cost and the elapsed ATM downtime); 2) a convolutional layer with 32 convolutional kernels (the length of each kernel is considered to be 6); 3) a max pooling layer; 4) a fully connected layer; and 5) a soft-max layer with two outputs. The RNN contains an input layer, a dense layer (32 hidden neurons), a recurrent layer, and a soft-max layer for classification. The LSTM shares similar configuration to RNN except for replacing the recurrent layer with the LSTM module.

The training strategy of the deep networks involves iterative updating of the weights in an online manner. In practice, the first 200 time points are used to set up the network weights. At each day, a new training example (s_t, a_t, r_t, s_{t+1}) is added to a defined buffer B (with a finite capacity) consisting of recent CRP 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 used to control the CRP system from 201 to 250. In the next iteration, the sliding window of the training data is moved 50 ticks forward covering a new training set from 50 to 250. The parameters in the network are then iteratively updated with the recently released data. This online strategy allows the model to get aware of the latest CRP system condition and update its parameters accordingly.

The average monthly replenishment costs of the proposed algorithm for two ATMs are shown in Table I (Note that the first 12 months time points of the input time series (before convergence) employed for system initialization and is not used for performance calculation).

TABLE I
Results of the average monthly replenishment cost ($) for the two ATMs.

	ATM-1	ATM-2
IoT+DQN	45.0	61.9
(based solely on) IoT	56.1	70.6
IoT+SQN	49.7	63.7
IoT+CDNN	51.4	67.1
IoT+LSTM	49.1	64.6
IoT+RNN	54.1	68.3
IoT+SNN	55.5	69.2

The results in Table I shows that for both ATMs, the lowest replenishment costs are made by Double-DQN CRP system. This is due to its novel structure which allows simultaneous environment sensing and optimum action learning for CRP system.

When considering the results of CDNN, RNN, LSTM and SNN, the pitfalls of prediction-based NN methods become immediately apparent. By investigating the total profit values in Table I, only the LSTM achieves comparable cost with the other RL-based systems. This is because prediction-based systems only consider the cash demand data to make decisions. The Double-DQN learns both cash demand condition and the action-value function $Q(s;a)$ in a joint framework.

V. Conclusion

ATMs are used by the majority of the costumers to withdraw cash. In this work we presented a CRP system based on the deep Q-network (DQN) structure. The system is composed of two main components: a deep learning component that learns the cash demand dynamic 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 stable targets during temporal difference calculations, a separate target network is attached to the system thereby forming the final Double-DQN structure. Experimental results show that the proposed method outperforms the other state-of-the-art deep CRP systems. The results on real cash demand timeseries demonstrate the effectiveness of the learning system in joint system dynamic acquisition and optimal action learning.

References

[1] Yaoguo Li, Huiye Sun, Chen Zhang, and Guohui Li, "Sites selection of atms based on particle swarm optimization," in *2009 International Conference on Information Technology and Computer Science*. IEEE, 2009, vol. 2, pp. 526–530.

[2] Valeria Platonova, Elena Gubar, and Saku Kukkonen, "Heuristic optimization for multi-depot vehicle routing problem in atm network model," in *Advances in Dynamic Games*, pp. 201–228. Springer, 2020.

[3] Mithat Zeydan and Sümeyra Kayserili, "A rule-based decision support approach for site selection of automated teller machines (atms)," *Intelligent Decision Technologies*, vol. 13, no. 2, pp. 161–175, 2019.

[4] Fazilet Ozer, Ismail Hakki Toroslu, Pinar Karagoz, and Ferhat Yucel, "Dynamic programming solution to atm cash replenishment optimization problem," in *International Conference on Intelligent Computing & Optimization*. Springer, 2018, pp. 428–437.

[5] Şeyma Batı and Didem Gözüpek, "Joint optimization of cash management and routing for new-generation automated teller machine networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.

[6] Canser Bilir and Adil Döşeyen, "Optimization of atm and branch cash operations using an integrated cash requirement forecasting and cash optimization model," *Business & Management Studies: An International Journal*, vol. 6, no. 1, pp. 237–255, 2018.

[7] Soodabeh Poorzaker Arabani and Hosein Ebrahimpour Komleh, "The improvement of forecasting atms cash demand of iran banking network using convolutional neural network," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 3733–3743, 2019.

[8] Sarveswararao Vangala and Ravi Vadlamani, "Atm cash demand forecasting in an indian bank with chaos and deep learning," *arXiv preprint arXiv:2008.10365*, 2020.

[9] S Poorzaker Arabani and H Ebrahimpour Komleh, "The optimization of forecasting atms cash demand of iran banking network using lstm deep recursive neural network," *Journal of Operational Research In Its Applications (Applied Mathematics)-Lahijan Azad University*, vol. 16, no. 3, pp. 69–88, 2019.

[10] Hossein Abbasimehr, Mostafa Shabani, and Mohsen Yousefi, "An optimized model using lstm network for demand forecasting," *Computers & Industrial Engineering*, p. 106435, 2020.

[11] Pankaj Kumar Jadwal, Sonal Jain, Umesh Gupta, and Prashant Khanna, "K-means clustering with neural networks for atm cash repository prediction," in *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, 2017, pp. 588–596.

[12] Kamini Venkatesh, Vadlamani Ravi, Anita Prinzie, and Dirk Van den Poel, "Cash demand forecasting in atms by clustering and neural networks," *European Journal of Operational Research*, vol. 232, no. 2, pp. 383–392, 2014.

[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] Farkhondeh Kiaee, "Integration of electric vehicles in smart grid using deep reinforcement learning," in *2020 11th International Conference on Information and Knowledge Technology (IKT)*. IEEE, 2020, pp. 40–44.

[18] 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.