Model Free Optimal Control of Two Whole Buildings using Deep Q-networks

Ki Uhn Ahn¹, Jae Min Kim¹, Youngsub Kim¹, Cheol Soo Park¹, Kwang Woo Kim²
¹Department of Architecture and Architectural Engineering, College of Engineering, Seoul National University, Seoul, South Korea
²Institute for Passive Zero Energy Building, Seoul, South Korea

Abstract

This study introduces an application of deep Q-networks (DQN) for whole building energy optimization. The DQN was coupled to an office building and a retail building: an EnergyPlus simulation model provided by the U.S. Department of Energy (DOE). The optimal controls can successfully minimize total building energy use with appropriate indoor CO₂ concentrations maintained (less than 1,000 ppm). It was shown that the DQN can find balanced control actions between different energy consumers in the building such as chillers, pumps, and air handling units (AHUs), leading to a whole building global optimization.

Introduction

Many buildings have been operated by the operators’ experience and knowledge. This rule-based control has been applied to determination of the HVAC start and stop time, set-point temperature and air/water flow rates, etc. (Afram et al., 2017). In contrast to the rule-based control, Model Predictive Control (MPC) has been highlighted since it can predict future states of the system and find optimal control variables that can minimize a cost function over the prediction horizon (Afram and Janabi-Sharifi, 2014; Afram et al., 2017). However, the performance of the MPC relies on the quality of the simulation model and the data availability for the simulation model (Liu and Henze, 2006; Afram and Janabi-Sharifi, 2014).

The reinforcement learning (RL), one type of artificial intelligence techniques, mimics the learning process of human beings. The RL learns what to do and how to map the states to actions so as to achieve a goal (Sutton and Barto, 1998). In contrast to the MPC, the RL is able to improve a policy for control actions based on the results of the previous actions, without resorting to the simulation model (Kazmi and D’Oca, 2016). In the area of building energy, a few studies (Kazmi and D’Oca, 2016; Yang et al., 2015; Cheng et al., 2016; Chen et al., 2018) have used the traditional RL methods (e.g. Markov Decision Process, Temporal Difference learning, Q-learning) for optimal control of building systems. However, an infinite number of state-action pairs can exist in building, because the outdoor environment, thermal characteristics of spaces, and HVAC systems are continuously time-varying. The aforementioned traditional RL methods are limited to local control of an individual system, owing to the demanding computation time for updating the policy with respect to the infinite state-action pairs.

Recently, Mnih et al. (2015) developed a novel artificial agent, termed a deep Q-network (DQN), which combines deep neural networks and Q-learning. Mnih et al. (2015) showed that the DQN is able to approximate policies directly from the high-dimensional states and actions, and the DQN was capable of learning and subsequent decision-making at the human level or higher. Therefore, in contrast to the MPC, the DQN would be able to control the system without the model. In addition, the number of state-action pairs that the DQN can handle would be much greater than traditional RL methods.

This study introduces the application of the DQN for ‘global’ optimization of a whole building’s energy use for two whole buildings: one is an office building located in Incheon and the other a retail building located in Ulsan, South Korea. The goal of this study is to minimize total building energy use with the average of CO₂ concentration maintained not exceeding 1,000ppm. The two buildings are generated using EnergyPlus simulation models developed by U.S. Department of Energy. The HVAC systems of two buildings include four air-handling units, two electric chillers, a cooling tower, and two pumps. This study investigated whether the DQN could successfully achieve the model-free control for the whole building energy systems.

Methods

Reinforcement learning

The RL differs from the supervised learning (e.g. regression or classification using ANN, support vector machine, Gaussian process, etc.) and unsupervised learning (e.g. clustering or dimensionality reduction using K-means, principle component analysis, etc.). One of the most significant differences between the RL and the others is that no representation of input/output pairs exist in the RL because the RL learns behavior through trial-and-error interactions with a dynamic environment (Kaelbling et al., 1996). Figure 1 presents a general framework of the RL problem. The learner and decision maker is called the agent, and the entity it interacts with, comprising everything outside the agent, is called the environment (Sutton and Barto, 1998). The agent selects an action a_t based on the information represented as state

https://doi.org/10.26868/25222708.2019.210385
Two value functions exist in RL: the state value function $V(s)$ and the action value function $Q(s,a)$. These functions represent the expected long-term reward under optimal policies. The state value function $V(s)$ is defined as the total expected discounted reward starting from state $s$ and following the policy $\pi$:

$$V(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

where $\gamma$ is the discount factor, $r_t$ is the reward at time $t$, and $\pi$ is the policy function. The action value function $Q(s,a)$ represents the expected return starting from state $s$, taking action $a$, and following the policy $\pi$:

$$Q(s,a) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r_t \right]$$

The action value function can be used to determine the optimal action at each state, as follows:

$$a^* = \arg \max_a Q(s,a)$$

where $a^*$ is the optimal action at state $s$.

**Deep Q Network (DQN)**

If all the possible state space could not be itemized into a table, the approximation technique could be used for the Q-table (Yang et al., 2015). Thus, a function approximator such as deep neural networks was introduced to represent the action-value function as follows:

$$Q(s,a; \theta) = Q(s,a)$$

where $\theta$ is the parameter set of the neural network. The network is trained using experience replay and target networks to stabilize the training process.

**Figure 1: Diagram of reinforcement learning problem.**
DQN uses an experience replay (Mnih et al., 2015), which stores the agents’ experience \( e_t = (s_t, a_t, r_t, s_{t+1}) \) in a dataset \( D_t = \{ e_1, e_2, \ldots, e_t \} \), called as a replay memory, at each time-step \( t \). The experience samples are randomly extracted from the replay memory to update the Q-networks, and it allows greater data efficiency and breaks the correlations between experience samples to reduce the variance of the updates (Mnih et al., 2015). The second feature of the DQN is the use of two Q-networks, a Q-network \( Q(s, a; \theta) \) and a target Q-network \( \hat{Q}(s, a; \theta^-) \). The former is used to determine optimal action, and the latter is used to generate a target value to update \( \theta \) of \( Q(s, a; \theta) \). At each iteration, \( Q(s, a; \theta) \) is updated to minimize the mean-square error with \( \hat{Q}(s, a; \theta^-) \) by optimizing a loss function (Nair et al., 2015).

In the DQN, exploration evaluates possible actions, while exploitation uses knowledge of prior experiences. The trade-off between exploration and exploitation must be carefully considered. One simplest way of balancing the exploration and the exploitation is the \( \epsilon \)-greedy method, where the agent chooses the action with the greatest estimated value with a probability of \( 1 - \epsilon \) (exploitation), and choose random actions with a probability of \( \epsilon \) (exploration) (Yang et al., 2015).

### Implementation of DQN

#### Target buildings

The DQN was applied to the office building located in Incheon, South Korea and the retail building located in Ulsan, South Korea. The two buildings were generated using the office building of EnergyPlus reference model (RefBldgLargeOfficeNew2004_Chicago.idf) developed by U.S. Department of Energy. The geometry and configuration of HVAC systems of the each building are identical to those of the reference model. Each building has the 46,320 m\(^2\) of total floor area with 12 stories above ground and a basement (B1F). The building’s above ground floor has five zones including four exterior zones and an interior zone, and the basement has a single zone. 10 times multiplier (DOE, 2018), a function of EnergyPlus was applied to the 6th floor, which is a typical floor. Therefore, there are 16 zones in each building (1 zone in the basement, 5 zones in 1F, 5 zones in 6F, 10 multiplier) + 5 zones in 12F).

In this study, 14 days of cooling season (July 1 to July 14) was analysed. The HVAC system includes four air handling units (AHU#1, AHU#2, AHU#3, AHU#4) with variable air volume units, two electric chillers (Chiller#1, Chiller#2), a cooling tower, two pumps for a chilled water loop (Pump#1) and a condenser loop (Pump#2). The air handling units deliver the amount of cold supplied by the chillers to each zone, and introduce outdoor air (OA) though OA dampers so that an appropriate CO\(_2\) level is maintained. Two chillers are arranged in parallel, and if the cooling load exceeds the capacity of the Chiller#1, then Chiller#2 start to operate. Pump#1 and Pump#2 are variable and constant, respectively. The HVAC systems are sized to meet the cooling loads in two locations. Table 1 shows the internal loads of two buildings. The cooling set-point temperatures of the two buildings were set at 26 °C. The operating hours are from 05:00 to 18:00 and 05:00 to 21:00 in the office building and retail building, respectively. In order to reflect a stochastic nature in occupants’ presence to the building model, white noise was intentionally added to the occupants’ schedule (20% of variation).

<table>
<thead>
<tr>
<th>Floors</th>
<th>Occupant (m(^2)/people)</th>
<th>Lighting (W/m(^2))</th>
<th>Equipment (W/m(^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basement</td>
<td>37.16 / 37.16</td>
<td>10.76 / 10.76</td>
<td>10.76 / 10.76</td>
</tr>
<tr>
<td>1F</td>
<td>18.58 / 11.61</td>
<td>10.76 / 18.29</td>
<td>10.76 / 5.38</td>
</tr>
<tr>
<td>6F</td>
<td>18.58 / 11.61</td>
<td>10.76 / 18.29</td>
<td>10.76 / 5.38</td>
</tr>
<tr>
<td>12F</td>
<td>18.58 / 18.58</td>
<td>10.76 / 10.76</td>
<td>10.76 / 10.76</td>
</tr>
</tbody>
</table>

#### DQN formulation and co-simulation

Table 2 shows 54 states and 6 actions used for the DQN. At each time-step, the number of possible actions is 972 (3x3x3x4x3, Table 2), and the DQN determines the optimal actions among 972 cases under the time-varying 54 states. The number of states and action is infinite, because the states vary continuously.

<table>
<thead>
<tr>
<th>Index</th>
<th>State</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>OA temperature</td>
<td>°C</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>OA humidity</td>
<td>%</td>
</tr>
<tr>
<td>( s_3 )</td>
<td>Direct solar radiation</td>
<td>W/m(^2)</td>
</tr>
<tr>
<td>( s_4 )</td>
<td>Diffuse solar radiation</td>
<td>W/m(^2)</td>
</tr>
<tr>
<td>( s_5 )</td>
<td>Indoor air temperatures of sixteen zones</td>
<td>°C</td>
</tr>
<tr>
<td>( s_{21} )</td>
<td>CO(_2) concentrations of sixteen zones</td>
<td>Ppm</td>
</tr>
<tr>
<td>( s_{37} )</td>
<td>Energy use of four AHUs</td>
<td>kWh</td>
</tr>
<tr>
<td>( s_{41} )</td>
<td>OA damper opening rate of four AHUs</td>
<td>%</td>
</tr>
<tr>
<td>( s_{45} )</td>
<td>Energy use of two chillers</td>
<td>kWh</td>
</tr>
<tr>
<td>( s_{47} )</td>
<td>CHWS temperature of two chillers</td>
<td>°C</td>
</tr>
<tr>
<td>( s_{49} )</td>
<td>CHWS temperature at main pipe</td>
<td>°C</td>
</tr>
<tr>
<td>( s_{50} )</td>
<td>Energy use of a cooling tower</td>
<td>kWh</td>
</tr>
<tr>
<td>( s_{51} )</td>
<td>WCS temperature of a cooling tower</td>
<td>°C</td>
</tr>
<tr>
<td>( s_{52} )</td>
<td>Energy use of two pumps</td>
<td>kWh</td>
</tr>
<tr>
<td>( s_{54} )</td>
<td>Number of time-step</td>
<td>-</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>Settings for the OA damper opening rate of four AHUs (100, 90, 80)</td>
<td>%</td>
</tr>
<tr>
<td>( a_5 )</td>
<td>Set-points for the CHWS temperature of two chillers (6, 7, 8, 9)</td>
<td>°C</td>
</tr>
<tr>
<td>( a_6 )</td>
<td>Set-point for the WCS temperature of a cooling tower (29, 30, 31)</td>
<td>°C</td>
</tr>
</tbody>
</table>

* CHWS: chilled water supply, CWS: cooling water supply

The objective of the DQN control in this study was to minimize energy use of the whole building, with indoor CO\(_2\) concentration not exceeding 1,000 ppm. The DQN explores the optimal policy based on the experiences of states, actions, and rewards. The reward was set as the total energy use of the building \( E_{tot}(F) \) at each time, and

---

Proceedings of the 16th IBPSA Conference Rome, Italy, Sept. 2-4, 2019
a weight of 1.5 was introduced to penalize any action inducing indoor CO₂ concentration greater than 1,000 ppm (Equation (8)). The authors developed the DQN codes to find actions that can minimize the reward (energy use) at every time-step.

\[ \begin{align*}
    r(t) &= \begin{cases} 
    E_{tot}(t), & \text{if } \frac{1}{16} \sum_{i=1}^{16} CO_{2}^i(t) \leq 1000 \\
    E_{tot}(t) \times 1.5, & \text{if } \frac{1}{16} \sum_{i=1}^{16} CO_{2}^i(t) > 1000
    \end{cases} 
\] (8)

where \( t \) is the time-step, \( CO_{2}^i \) is the CO₂ concentration of \( i \)th zone, and \( i \) is the index of 16 zones of each building. The replay memory of DQN deletes the oldest experience when the new experience is received, and the authors set the replay memory to store the experiences of 960 time-steps. The Q-network was designed with two hidden layers with 30 neurons that were determined by the trial and error method (Kumar et al., 2013). The Rectified Linear Unit (ReLU) function was used for the activation function, and \( \theta \) of Q-network \( Q(s,a;\theta) \) was updated with the Adam optimization method, which is appropriate for a large number of layers and neurons such as deep neural network (Kingma and Ba, 2014). An EnergyPlus simulation runs from July 1 to July 11 were regarded as one single episode, and 1,000 episodes were iterated to explore the optimal policy of the DQN. The \( \theta^- \) of the target Q-network \( \tilde{Q}(s,a;\theta^-) \) was replaced with \( \theta \) of the Q-network \( Q(s,a;\theta) \) at the end of every two episodes. The initial value for the \( \epsilon \)-greedy method was set to 1.0 (\( \epsilon(0) = 1.0 \)), and the \( \epsilon \) was changed by Equation (9) so as to decrease the explorations when the exploitations are sufficient.

\[ \epsilon((n-1) \times 96 + t) = 1.0 \times (0.9999)^{(n-1) \times 96+1} \] (9)

where \( t \) is the time-step, and \( n \) is the index of episodes.

An on-line co-simulation was implemented to couple with the target buildings of the EnergyPlus models and the DQN algorithm designed in python programming language. Using the external interface in the EnergyPlus, the authors defined the variables to be delivered (states, \( s_1-s_{54} \)) and the variables to be received (actions, \( a_1-a_6 \)). In python, the BCVTB Socket module was defined to receive the states from the EnergyPlus and deliver the values for the actions from the Q-network. At every time-step of each episode, the Q-network was updated and the optimal action was determined by the \( \epsilon \)-greedy method.

**DQN results**

For both buildings, in order to investigate the performance of DQN, the actions for the baseline operation were defined as follows: 90% of OA damper opening rate (\( a_1 \), \( a_2 \), \( a_3 \), \( a_4 \)), 6 °C for the chilled water (\( a_5 \)), and 29 °C for the cooling water (\( a_6 \)). The policy of DQN was improved for 11 days (July 1-July 11) (1,000 times iteration), and the policy was validated for 3 days (July 12-July 14).

**Case study#1: office building**

Figure 2 shows the return, the sum of rewards (Equation (8)) at every time-step during each episode (from July 1 to July 11). The return pattern is non-stationary, since the 1.5 of weight was multiplied to \( E_{tot} \) (Equation (8)) when the average of CO₂ concentration was over 1,000 ppm. Please note that the occupants’ presence of both buildings were randomly changed, so that the cooling load and CO₂ concentration of each zone were changed at every time-step of each episode. The greatest and least returns were 54,827 kWh at the 5th episode and 50,286 kWh at the 900th episode, respectively, resulting in an 8.3% difference between the greatest and least. The decreasing pattern in Fig. 6 shows the DQN’s self-learning capability.

For understanding of the policy improvement, the DQN performances at the 10th (Episode#10), the 100th (Episode#100) and the 900th (Episode#900) episodes were compared. Figure 3 shows the results based on the DQN policy for the chilled water and cooling water for the validation period (3 days, July 12-July 14). At Episode#900, the chilled water supply temperature was set to 9 °C during the operation time (05:00-18:00) (Figure 3 (a)), and the cooling water set-point temperature was set at 30 °C at the morning hours (about 05:00-10:00) and maintained 29 °C from about 10 a.m. when the cooling load began to increase (Figure 3 (b)).

Figure 4 and Table 3 shows the difference between the greatest and least. The decreasing pattern in Fig. 6 shows the DQN’s self-learning capability.

It is noteworthy that the DQN reduced the energy use of Chiller#1, which was the greatest energy consumer in the building. At Episode #900, 33.4% of Chiller#1’s energy use was reduced as compared to that of the baseline operation (Chiller#1 in Baseline: 36,540 kWh vs. Chiller #1 in DQN at Episode#900: 24,320 kWh) (Figure 4 (a)). Even though the energy use of Chiller#2 at Episode#900 increased by 0.6% compared to the that of baseline operation (Baseline: 19,712 kWh vs. Episode#900: 19,832 kWh), the energy use of Chiller#2 at Episode#900 was reduced by iterating the episodes. It is interesting that Pump#1 at DQN Episode#900 consumes far more energy.
than Pump#1 in the baseline operation (2,972 kWh) (Table 4), because the chilled water flow rate has to be increased as the set-point chilled water temperature increases (Figure 3(a)). The reason that DQN increased the set-point chilled water temperature from 6 to 9 °C (Figure 3(a)) is to reduce Chiller#1’s energy consumption, which far outweighs Pump#1’s energy (Table 3). However, the energy use of the chilled water pump only accounts for 4.0% of the total energy use of the building (chilled water pump: 2,972 kWh vs. total energy use: 73,936 kWh) (Table 3). The difference in the AHUs’ energy between the baseline operation and DQN is almost negligible (Table 3). The energy use of Pump #2 was not changed, because it operates on a constant flow rate.

Figure 3: DQN policy improvement for CHWS and CWS of the office building.

Figure 4: Energy use of the office building by DQN.

Table 3: Energy use of the office by DQN (unit:kWh).

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Episode #10</th>
<th>Episode #100</th>
<th>Episode #900</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHU#1</td>
<td>2,972</td>
<td>3,316</td>
<td>4,012</td>
<td>4,012</td>
</tr>
<tr>
<td>Chiller#1</td>
<td>36,540</td>
<td>36,832</td>
<td>38,488</td>
<td>38,488</td>
</tr>
<tr>
<td>Chiller#2</td>
<td>6,464</td>
<td>6,464</td>
<td>6,464</td>
<td>6,464</td>
</tr>
<tr>
<td>Total energy</td>
<td>73,936</td>
<td>70,540</td>
<td>63,284</td>
<td>62,312</td>
</tr>
<tr>
<td>Saving rate (%)</td>
<td>-</td>
<td>4.6</td>
<td>14.4</td>
<td>15.7</td>
</tr>
</tbody>
</table>

Figure 5 shows the results on the control of the OA damper opening rate. At Episode#900, the OA damper opening rates of AHUs#1 and #3 were maintained at a lowest level most of the time, while the others were kept at the highest level after about 10:00 when the number of occupants increased. Figure 6 and Table 4 show the ventilation load due to the OA intake, which were
calculated as the output of “zone mechanical ventilation cooling load increase energy” (DOE, 2018) in the EnergyPlus. The ventilation load gradually decreased as the experiences were accumulated (Table 4). For example, the DQN policy at Episode#900 could reduce the ventilation load by 14.9% compared to the baseline operation (Baseline: 47,584 kWh vs. Episode#900: 40,476 kWh). In addition, the DQN policy can maintain the average of CO$_2$ concentration level below 1,000 ppm at Episode#900 (Figure 7). The OA dampers were controlled so that the ventilation load was minimized (Figure 6, Table 4) while the averages of CO$_2$ concentration were kept under 1,000 ppm (Figure 7).

**Case study#2: retail building**

Figure 8 shows the return at each episode (from July 1 to July 12) for the retail building located in Ulsan. During the 1,000 episodes, the maximum return was 64,531 at the 8th episode, and 55,081, the minimum return was recorded at the 708th episode. The return at Episode#1000 is 55,407, which is the 11th lowest. The authors compare the DQN performance at Episode#1000 with the baseline operation, rather than the DQN policy improvement (e.g., Episode#10 vs. Episode#100 vs. Episode#900) as presented in the previous section.

Figure 9 shows the results of DQN actions at Episode#1000 for the chilled water and cooling water set-point temperatures. Similar to the office building, the DQN policy improved by itself to keep the chilled water temperature at a highest level of 9 °C (Figure 9 (a)). The cooling water set-point temperature was intermittent between 29 °C and 30 °C (Figure 9 (b)).

Figure 10 and Table 5 show the comparison between the DQN at Episode#1000 and baseline operation. In the baseline operation, the energy use of Chiller#1 (37,032 kWh) accounts for 57.2% of the total energy use of the retail building (64,688 kWh). During the baseline operation, Chiller#2 was turned on intermittently (Figure 10 (b)), because most of the heat was removed from the Chiller#1 (Figure 10 (a)). The DQN could increase the chilled water temperature up to 9 °C (Figure 9 (a), Figure 10 (a)) so that the Chiller#2 can operate continuously for a higher efficiency than the baseline. (Figure 10 (b)). Although the energy uses of Chiller#2 (Figure 10 (b)) and Pump#1 (Figure 10 (d)) were increased compared to the baseline operation, the total energy use by DQN is 7.0% less than that of the baseline operation (Baseline: 64,688 kWh vs. DQN: 60,156 kWh) (Figure 10 (e)).
In contrast to the office building (Figure 5), the OA damper opening rate of AHU#3 in the retail building was changed intermittently (Figure 11) when the average of CO$_2$ concentration was closed to 1,000 ppm (Figure 12). The average CO$_2$ concentrations could be kept below 1000 at all times (Figure 12), and the ventilation load of the building could be reduced by 9.2% (Baseline: 21,984 kWh vs. DQN: 19,956 kWh) (Figure 13, Table 6).

Ministry of Trade, Industry & Energy (MOTIE) of the Technology Evaluation and Planning (KETEP) and the This work was supported by the Korea Institute of Energy.

Acknowledgement

This study presents the DQN implementation for the whole building energy optimization. When the DQN was applied to the reference office and retail building, it could achieve 15.7% and 7.0% energy savings, respectively, with an indoor CO2 concentration not exceeding 1,000 ppm, as compared to the baseline operation. In addition, it was found that the DQN can balance energy consumptions of different mechanical systems in the building by itself. For example, in both cases of the office and retail buildings, the DQN wanted to increase the energy uses of Chiller#2 and Pump#1 for greater energy savings by Chiller#1, resulting in the net energy saving of the whole buildings.

The prominent feature of the DQN is that it requires neither any simulation model (system level or building level), nor any prediction horizon. What the DQN requires is only prior experiences with regard to actions, states, and rewards. In other words, the DQN can achieve the whole building optimization even without any deliberate simulation model.

Conclusion

This study presents the DQN implementation for the whole building energy optimization. When the DQN was applied to the reference office and retail building, it could achieve 15.7% and 7.0% energy savings, respectively, with an indoor CO2 concentration not exceeding 1,000 ppm, as compared to the baseline operation. In addition, it was found that the DQN can balance energy consumptions of different mechanical systems in the building by itself. For example, in both cases of the office and retail buildings, the DQN wanted to increase the energy uses of Chiller#2 and Pump#1 for greater energy savings by Chiller#1, resulting in the net energy saving of the whole buildings.

The prominent feature of the DQN is that it requires neither any simulation model (system level or building level), nor any prediction horizon. What the DQN requires is only prior experiences with regard to actions, states, and rewards. In other words, the DQN can achieve the whole building optimization even without any deliberate simulation model.

Acknowledgement

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20182010106460).

References


