ISSN 1975-8359 [Print] / ISSN 2287-4364 [Online]
The Transactions of the Korean Institute of Electrical Engineers, vol. 70, no. 10, pp. 1594~1600, 2021
https://doi.org/10.5370/KIEE.2021.70.10.1594

Reinforcement Learning-based HVAC Control Agent for Optimal Control of Particulate Matter in Railway Stations

역사 내 미세먼지 농도 조절을 위한 강화학습 기반의 공조설비 제어 에이전트 구축

Kyung-bin Kwon · Sumin Hong · Jae-Haeng Heo · Hosung Jung · Jong-young Park
권경빈* · 홍수민* · 허재행* · 정호성** · 박종영**†

Abstract

This study developed a reinforcement learning-based energy management agent that controls the concentration of fine dust by controlling the power consumption of energy facilities such as air conditioners and blowers in stations. To apply reinforcement learning, the problem was first defined based on the Markov decision-making process, and a model was developed to predict the concentration of fine dust in history using data correlated with fine dust. Based on the linear compensation function created based on this, the Deep Q-Network (DQN) method was applied to obtain the optimal policy based on the artificial neural network. In the case study, it was confirmed that convergence to the optimal policy was achieved through the learning process, and it was confirmed that the learned agent lowers the fine dust concentration by increasing the power consumption of the air conditioner when the fine dust concentration in the station rises above a certain level.

Key Words

Key Words : Deep Q-Network, Energy management, Markov decision process, Particulate matter, Reinforcement learning

1. 서 론

최근 국민들의 건강에 대한 관심과 함께 미세먼지와 초미세먼지에 대한 관심이 높아지고 있다. 이러한 측면에서 국민들이 자주 이용하는 지하철 및 철도의 미세먼지 관리 또한 중요한 이슈로 떠오르고 있다 [1]. 특히 지하철 및 철도 역사의 경우 환기가 어려운 구조 특성상 미세먼지와 초미세먼지에 취약하기 때문에, 역사 내 미세먼지를 줄이기 위한 기술의 적용 및 관련 연구가 요구된다. 이에 발맞춰 도시철도 및 지하 역사의 미세먼지를 모니터링하거나 줄이는 방법에 대한 연구가 활발하게 진행되고 있다. 예로 미세먼지의 농도를 실시간으로 측정 가능한 제어시스템을 구축하여 미세먼지의 농도를 확인하는 연구 [2]-[3], 미세먼지의 확산 및 분포에 관한 연구 [4]-[5], 미세먼지 저감을 위한 공기조화기, 객실 유입 차단장치 등에 대한 연구 [6]-[8], 미세먼지 저감을 위한 추가적인 개선 방안에 대한 연구 [9] 등이 다양하게 진행되었다.

미세먼지 절감을 위해선 도시철도 역사 내에 설치된 공조기와 송풍기를 동작시켜 공기질을 향상시키는 방법이 있다[10]. 공조기는 외부에서 실내로 공급되는 공기를 사용 목적에 적합하도록 조절하는 것으로, 외부의 공기를 실내로 유입할 때 불순물을 제거하기 위한 필터를 거침으로써 내부의 미세먼지 농도를 줄일 수 있다[11]. 송풍기는 역사 내의 미세먼지를 밖으로 나오게 하고, 필터를 통해 통과한 외부 공기를 공급함으로써 내부 공기질을 향상시킬 수 있다[12]. 그렇지만 공조기와 송풍기의 설치 위치와 전력사용량에 따라 미세먼지 농도 제어의 효과가 다르기 때문에 미세먼지 농도 변화에 대한 불확실성이 존재한다.

이러한 불확실성에 대하여 역사 내 미세먼지 농도를 충분히 관리함과 동시에 공조기와 송풍기의 전력사용량을 효율적으로 제어하기 위한 방법으로 강화학습(Reinforcement Learning)을 적용할 수 있다. 강화학습이란 머신러닝(Machine learning)의 범주 안에 있는 학습 방법 중 하나로, 에이전트가 역동적인 환경에서 반복적인 시행착오 상호작용을 통해 작업수행 방법을 학습하는 것이다 [13]. 불확실성이 포함된 최적화 문제를 해결하기 위해서 기존에는 불확실성을 확률분포로 표현한 후 몬테 카를로 방법(Monte Calro algorithm)을 활용하여 최적해를 구하였다 [14]. 이러한 방법은 확률분포 모델을 구성해야 하는 Model-based 방법으로, 데이터에서부터 불확실성의 확률

† Corresponding Author: Electrification System Research Department, Korea Railroad Research Institute, Korea
E-mail:jypark@krri.re.kr
http://www.krri.re.kr
* RaonFriends Co., Ltd., Korea
** Electrification System Research Department, Korea Railroad Research Institute, Korea
Received : Aug. 30, 2021 **Revised :** Sep. 17, 2021 **Accepted :** Sep. 27, 2021

분포를 구하기 어렵거나 하나의 확률분포로 표현하여 특성을 반영하기 어려운 경우에는 최적의 해를 구하는 데 어려움이 있다. 반면 강화학습은 대표적인 Model-free 방법으로, 마르코브 의사결정 과정(Markov Decision Process)에 기초하여 불확실한 환경을 학습하기 위해 반복적으로 샘플링을 진행하고, 이를 토대로 최적 행동을 정하는 최적 정책(optimal policy)을 구하는 과정을 거친다 [15].

이러한 측면에서 불확실성을 가지는 외부 미세먼지 농도나 역사 내부의 습도 등의 환경이 주어졌을 때 강화학습을 적용하여 역사 내 미세먼지 농도를 관리하는 송풍기, 공조기의 최적 제어정책을 구할 수 있다. 본 논문은 이를 위해 강화학습의 다양한 알고리즘 중 Deep Q-Network (DQN) 방식을 활용하여 송풍기 및 공조기의 전력사용량을 제어하는 에너지 관리 에이전트를 구축하였다. 이를 위하여 2장은 강화학습 적용을 위하여 마르코브 의사결정 과정에 기초하여 시스템 모델링을 구성하였다. 3장에서는 실제 데이터에 기초하여 역사 내 미세먼지 농도와 상관관계가 있는 데이터를 선택하고, 이를 토대로 선형 보상함수를 개발하였다. 4장에서는 3장에서 개발한 보상함수를 토대로 DQN 알고리즘을 적용하여 최적 정책을 구성하였다. 5장에서는 사례연구를 통해 강화학습을 적용한 결과를 분석하였으며, 6장에서는 본 연구의 결론을 서술하였다.

2. 시스템 모델링

송풍기, 공조기 등 에너지 설비 최적 운영를 통한 역사 내 미세먼지 농도 조절에 대한 시스템은 다음과 같이 정의할 수 있다. 먼저 시간 t에 대하여 역사 내 미세먼지 농도는 지름이 2.5㎛보다 작은 미세먼지(PM2.5) 농도 $i_t^{(1)}$(㎍/㎥), 지름이 10㎛보다 작은 미세먼지(PM10) 농도 $i_t^{(2)}$(㎍/㎥)로 나타낼 수 있다. 다음으로 시간 t에 대하여 K개의 송풍기와 L개의 공조기는 각각 $v_t^{(1)},\cdots,v_t^{(K)}$ 와 $w_t^{(1)},\cdots,w_t^{(L)}$의 전력을 사용하며, 이로 인해 미세먼지 농도가 조절된다. 이때 시간 t의 전력 가격이 p_t로 주어질 경우, 총 전력비용 c_t는 (1)과 같다.

$$c_t = p_t\left(\sum_{k=1}^{K} v_t^{(k)} + \sum_{l=1}^{L} w_t^{(l)}\right) \tag{1}$$

이후 송풍기와 공조기의 제어로 인해 시간 $t+1$에서의 미세먼지 농도는 각각 $i_{t+1}^{(1)}$, $i_{t+1}^{(2)}$로 변하게 된다.

위에서 정의한 시스템 모델링에 강화학습을 적용하기 위해선 송풍기 및 공조기의 제어를 통한 미세먼지 조정에 대한 일련의 과정을 마르코브 의사결정 과정(Markov Decision Process; MDP) 기반의 확률적 모델로 표현해야 한다 [13]. Markov Decision Process는 (S, A, P, R, γ)의 다섯 가지 요소로 구성된다. 먼저 현재 상태(state) S는 현재 시간 t, 역사 내 미세먼지 농도 $i_t^{(1)}$, $i_t^{(2)}$ 및 변화에 영향을 주는 외부 환경 요소를 포함한다. 외부 환경 요소에는 실외 미세먼지 농도, 습도, 온도 등이 있으며, 이 중 역사 내 미세먼지 농도의 변화와 상관관계가 있는 요소가 포함되게 된다. A는 현재 취하는 행동(action)을 의미하며, 위 모델링에서는 시간 t의 행동 a_t는 아래와 같이 정의할 수 있다.

$$a_t = \left\{v_t^{(1)},\cdots,v_t^{(K)},w_t^{(1)},\cdots,w_t^{(L)}\right\} \tag{2}$$

다음으로 전이함수(transition function) P는 현재 상태 s_t에서 a_t라는 행동을 취하였을 때 s_{t+1}으로 이동하는 확률로 정의할 수 있다. 이때 마르코브 의사결정 과정에 기반한 모델링에서 s_t에서 s_{t+1}으로 상태가 변하는 전이확률은 '과거의 모든 상태 중 바로 이전 상태인 s_t에 의해서만 결정된다.'는 마르코브 성질(Markov property)를 따른다. 이러한 성질은 식 (3)와 같이 나타낼 수 있다.

$$\Pr\left(s_{t+1}|\{s_\tau\}^{t_{\tau=1}}\right) = \Pr\left(s_{t+1}|s_t\right) \tag{3}$$

보상(reward) R은 상태 s_t에서 a_t라는 행동을 취하였을 때 얻게 되는 보상을 의미한다. 시간 t에서 얻게 되는 보상 r_t는 s_t와 a_t에 대한 함수 $r_t(s_t, a_t)$로 표현할 수 있다. 역사 내 미세먼지 농도와 상관관계가 있는 상태 요소 및 송풍기, 공조기 제어에 따른 역사 내 미세먼지 농도 변화량과 이에 따른 전력사용량 간의 관계를 모두 포괄하는 보상함수에 개발은 3장에서 자세히 설명한다.

마지막으로 감가율(discount factor) $\tau \in (0,1]$는 현재 얻는 보상과 미래에 얻을 수 있는 보상 간의 중요도를 조절하는 변수이며, 작은 값을 가질수록 현재 얻는 보상을 미래에 얻을 수 있는 보상보다 더 가치 있게 여김을 의미한다. 강화학습에서 고려하는 시간이 유한한 경우, $\tau = 1$로 설정할 수 있다.

3. 보상함수 개발

강화학습은 현재 상태(state)에서 행동(action)을 반복적으로 취하고 그로 인해 발생하는 보상(reward)을 살펴봄으로써 보상의 총합이 최대가 되는 최적 정책 (optimal policy)를 만드는 학습과정을 의미한다. 본 연구에서는 행동인 역사 내 송풍기 및 공조기의 운영에 따른 미세먼지 농도 변화를 확인하기 위해 서로 다른 여러 상태에 대하여 실제 송풍기 및 공조기를 오랜 기간에 걸쳐 운전하면서 미세먼지 농도 변화를 관찰하기 어려우므로, 기존에 가지고 있는 외부, 내부의 온도, 습도, 미세먼지 농도 및 송풍기, 공조기의 전력사용량 데이터를 토대로 송풍기, 공조기의 전력사용량에 따른 역사 내 미세먼지 농

도를 수학적으로 모델링하여 선형 보상함수를 개발하고 이를 강화학습에 적용하였다.

3.1 데이터 분석을 통한 상태 변수 결정

먼저 역사 내 미세먼지 농도와 상관관계가 있는 데이터를 찾기 위해 기존의 데이터와 역사 내 미세먼지 농도와의 관계를 그림 1, 그림2와 같이 산점도로 나타내었다.

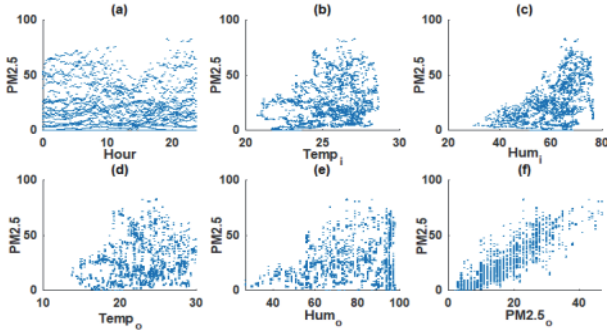


그림 1 (a) 시간 (b) 역사 내부온도 (c) 역사 내부습도 (d) 외부온도 (e) 외부습도 (f) 외부 PM2.5 농도 와 역사 내 PM2.5 농도 간의 산점도

Fig. 1 (a) time (b) indoor temp. (c) indoor humidity (d) outdoor temp. (e) outdoor humidity (f) scatter plot of PM2.5 between indoor and outdoor

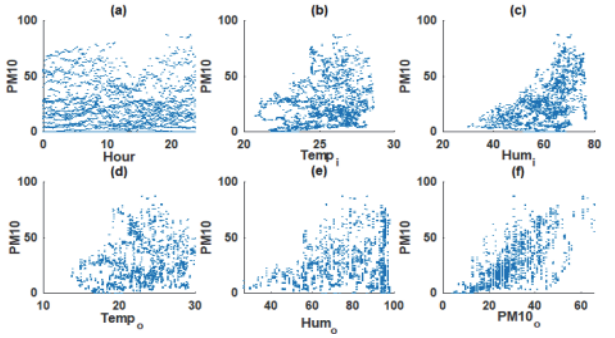


그림 2 (a) 시간 (b) 역사 내부온도 (c) 역사 내부습도 (d) 외부온도 (e) 외부습도 (f) 외부 PM10 농도 와 역사 내 PM10 농도 간의 산점도

Fig. 2 (a) time (b) indoor temp. (c) indoor humidity (d) outdoor temp. (e) outdoor humidity (f) scatter plot of PM10 between indoor and outdoor

그림 1과 그림 2에서 확인할 수 있듯이 역사 내 PM2.5 및 PM10 모두 외부 미세먼지 농도와 가장 강한 상관관계를 가지며, 추가적으로 역사의 내부 습도와 상관관계를 가짐을 확인할 수 있다. 이에 따라 현재 상태(state)에 역사 내 PM2.5, PM10 농도, 외부의 PM2.5, PM10 농도와 역사의 내부 습도를 포함하여 식 (4)와 같이 구성할 수 있다.

$$s_t = \left\{ i_t^{(1)}, i_t^{(2)}, o_t^{(1)}, o_t^{(2)}, h_t \right\} \tag{4}$$

3.2 선형 보상함수 모델 개발

다음으로 식 (4)의 상태에서 송풍기와 공조기의 제어를 하였을 때 발생하는 역사 내 미세먼지 변화에 대한 선형 예측모델로 개발하기 위하여 다음과 같은 최적화 문제를 구성하였다. 시간 t에 대하여 K개의 송풍기와 L개의 공조기를 제어하였을 때 발생하는 역사 내 미세먼지 $i_t^{(1)}$, $i_t^{(2)}$는 각각 계수 $\lambda_1 = [\lambda_{1,}^o \lambda_{1,}^h \lambda_{1,1}^v, \cdots, \lambda_{1,K}^v, \lambda_{1,1}^w, \cdots, \lambda_{1,L}^w]$, $\lambda_2 = [\lambda_{2,}^o \lambda_{2,}^h \lambda_{2,1}^v, \cdots, \lambda_{2,K}^v, \lambda_{2,1}^w, \cdots, \lambda_{2,L}^w]$ 및 상수 집합 $b = [b_1, b_2]$을 적용하여 식 (5), (6)과 같은 선형모델로 나타낼 수 있다.

$$i_{t+1}^{(1)} = (\lambda_1^o o_t^{(1)} + \lambda_1^h h_t) + (\sum_{k=1}^{K} \lambda_{1,k}^v v_t^{(k)} + \sum_{l=1}^{L} \lambda_{1,l}^w w_t^{(l)}) + b_1 \tag{5}$$

$$i_{t+1}^{(2)} = (\lambda_2^o o_t^{(2)} + \lambda_2^h h_t) + (\sum_{k=1}^{K} \lambda_{2,k}^v v_t^{(k)} + \sum_{l=1}^{L} \lambda_{2,l}^w w_t^{(l)}) + b_2 \tag{6}$$

따라서 실제로 측정한 역사 내 PM2.5, PM10의 값을 각각 $j_t^{(1)}$, $j_t^{(2)}$라고 할 때, 최적 예측모델을 구성하는 계수 λ_1, λ_2와 상수 b_1, b_2는 각각 다음의 최적화 문제를 풀어서 구성할 수 있다.

$$\min_{\lambda_1, b_1} \sum_{t=1}^{T} (j_{t+1}^{(1)} - i_{t+1}^{(1)})^2 \tag{7}$$

$$\min_{\lambda_2, b_2} \sum_{t=1}^{T} (j_{t+1}^{(2)} - i_{t+1}^{(2)})^2 \tag{8}$$

위 비선형 최적화 문제를 풀어 찾은 최적해를 각각 $\lambda_1^*, \lambda_2^*, b_1^*, b_2^*$라 할 때, 이를 활용하여 최종적인 선형 보상함수는 송풍기와 공조기를 제어하여 발생한 미세먼지 농도 감소로 인한 보상과 식 (1)에서 계산한 송풍기와 공조기를 제어할 때 발생하는 총 전력비용 c_t 의 차로 나타낼 수 있다. 즉, 최종적인 보상함수는 미세먼지 농도 감소로 인한 보상과 총 전력비용 간의 비를 ρ라고 할 때 식 (9)와 같이 나타날 수 있다.

$$r_t(s_t, a_t) = \rho\left\{ (i_{t+1}^{(1)} - i_t^{(1)}) + (i_{t+1}^{(2)} - i_t^{(2)}) \right\} - c_t \tag{9}$$

이때 ρ의 값이 커질수록 전력비용에 비해 미세먼지 감소로 인한 보상을 더 크게 평가한다는 의미를 가진다.

4. Deep Q-Network를 활용한 강화학습

2장과 3장에서 정의한 마르코브 의사결정 과정에 기초하여, 최적 송풍기 및 공조기 제어에 관한 문제는 식 (10)과 같은 최적화 문제로 표현할 수 있다.

$$\max_{\mu} E_{\pi_\mu}[\sum_{t=1}^{T} \gamma^t r_t] \tag{10}$$

최적화 문제의 목적은 시간 $t=1$부터 $t=T$까지 각 시간대별 보상 r_t의 총합의 평균을 최대화하는 정책 π_μ를 찾는 것으로, 이때 최적 정책 $\pi_\mu(\cdot) = \pi(\cdot;\mu)$는 최적 파라미터 μ을

찾음으로써 구할 수 있다. 식 (10)의 최적화 문제를 풀기 위해서 Policy Gradient Method 등의 강화학습 알고리즘을 적용할 수 있으며, 본 연구에서는 비선형 인공신경망에 기초한 deep Q-Network (DQN) 방법을 활용하여 최적화 문제를 풀었다. DQN 방법에서 파라미터 μ는 인공신경망의 가중치를 의미하며, 인공신경망은 학습을 통해 식 (11)의 Q-function의 근사값을 구하게 된다.

$$Q(s_t, a_t) = E_{\pi_\mu}[\sum_{\tau=t}^{T}\gamma^{(\tau-t)} r_\tau(s_\tau, a_\tau)|s_t, a_t] \tag{11}$$

식 (11)에서 표현한 바와 같이 Q-function은 상태 s_t에서 행동 a_t을 수행하였을 때 발생하는 총 보상의 기댓값을 의미하며, 이에 관하여 최적 Q-function는 벨만 방정식(Bellman equation)에 기초하여 식 (12)의 관계를 만족한다 [15].

$$\begin{aligned} &Q^*(s_t, a_t) \\ &= r_t(s_t, a_t) + \gamma E_{s_{t+1}}[\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})|s_t, a_t] \end{aligned} \tag{12}$$

최적 Q-function을 만들기 위하여, DQN 방법에서는 가중치 μ로 표현된 인공신경망에서 대하여 벨만 방정식의 양변의 차가 최소화되도록 파라미터 μ를 학습시킨다. 즉, 식 (12)에서 우변의 값은 고정된 파라미터 μ'에 기초한 target network로 계산하며, 좌변의 Q값이 벨만 방정식의 양변의 차를 최소화할 수 있도록 파라미터 μ를 조정함으로써 최적 파라미터를 구한다. 이와 관련하여 벨만 방정식의 양변의 차를 나타내는 손실함수(loss function) $\mathcal{L}(\mu)$는 식 (13)으로 계산할 수 있다.

$$\begin{aligned} \mathcal{L}(\mu) = E_{\{s_t, a_t, s_{t+1}\}}[(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \mu') \\ - Q(s_t, a_t; \mu))^2] \end{aligned} \tag{13}$$

식 (13)을 최소화하기 위해 손실함수의 기울기(gradient)를 식 (14)와 같이 계산하고 경사하강법(gradient descent method)을 적용하여 최적 파라미터 μ를 구할 수 있다.

$$\begin{aligned} \nabla\mathcal{L}(\mu) = E_{\{s_t, a_t, s_{t+1}\}}[-2(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \mu') \\ - Q(s_t, a_t; \mu))\nabla_\mu Q(s_t, a_t; \mu)] \end{aligned} \tag{14}$$

강화학습에서는 식 (14)의 기댓값을 정확하게 구하는 대신 샘플링을 통해 경로(trajectory)를 구성하고 이를 토대로 표본평균을 구하여 근사값을 구한다. 이때 충분한 탐색(exploration)을 보장하기 위하여 ϵ-greedy 방법을 적용하여 상태 s_t 가 주어졌을 때 $(1-\epsilon)$의 확률로 최적 행동 $a_t^* = \arg\max_{a_t} Q(s_t, a_t; \mu)$를 선택하고, 그 외의 경우 무작위로 행동 a_t를 결정할 수 있다[13]. ϵ값은 반복(iteration)이 진행됨에 따라 일정한 비율로 감소하게 된다.

이에 더하여 효과적으로 DQN 방법을 적용하여 안정적으로 최적 정책을 구하기 위해서, Experience replay 방법과 Fixed target network 방법을 추가적으로 적용할 수 있다. 먼저 Experience replay 방법은 이전에 샘플링 결과를 메모리 $\Phi = \{(s_t, a_t, s_{t+1}, r_t)\}$에 저장하고, 손실함수를 계산할 때 메모리 Φ에 저장한 샘플을 무작위로 선택하여 mini-batch ψ를 만들어 식 (14)를 계산하게 된다[16]. 이는 μ를 업데이트할 때마다 새로운 샘플을 만들 필요 없이 이전의 샘플을 활용하여 효율적으로 손실함수의 기울기를 구할 수 있게 한다. 다음으로 Fixed target network는 벨만 방정식의 우변을 계산할 때 적용하는 파라미터 μ'의 업데이트를 반복학습 시 매번 적용하는 대신 μ^-값으로 고정하고 N_O번의 업데이트에 1번씩 μ^-를 업데이트하여 최적화 과정에서 발생하는 불안정성 문제를 해결할 수 있다 [17]. 위에서 언급한 두 방법을 적용하여, 식 (14)의 근사값을 아래와 같이 구할 수 있다.

$$\begin{aligned} \nabla\widehat{\mathcal{L}}(\mu) = (-2/|\psi|)\sum_{t\in\psi}[(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \mu^-) \\ - Q(s_t, a_t; \mu))\nabla_\mu Q(s_t, a_t; \mu)] \end{aligned} \tag{15}$$

이를 토대로 DQN 방법을 활용한 최적 미세먼지 농도 조절을 위한 송풍기 및 공조기 제어는 그림 3과 같다.

5. 사례연구

4장에서 제시한 DQN 기반의 송풍기, 공조기 최적 제어 알고리즘의 효과를 입증하기 위하여 광주 남광주역의 데이터를 토대로 사례연구를 진행하였다. 3장의 보상함수 개발 단계에서는 GAMS를 활용하여 비선형 최적화 문제의 해를 구하였으며, 4장의 DQN 방법을 적용한 강화학습은 Python과 Keras 패키지를 이용하여 진행하였다 [18].

먼저 3장에서 설명한 바와 같이 현재 상태를 식 (4)와 같이 정의하였으며, 역사 내 송풍기 3개와 공조기 2개에 대하여 식 (7), (8)의 목적함수를 최소화하는 계수 λ_1, λ_2와 상수 b_1, b_2를 구한 결과는 표 1과 같다.

표 1 보상함수의 계수와 상수값

Table 1 Coefficients and constant values of the compensation function

λ_1^o	λ_1^h	$\lambda_{1,1}^v$	$\lambda_{1,2}^v$
1.830	0.488	-0.00047	-0.00009
$\lambda_{1,3}^v$	$\lambda_{1,1}^w$	$\lambda_{1,2}^w$	b_1
0	0	0	-35.64
λ_2^o	λ_2^h	$\lambda_{2,1}^v$	$\lambda_{2,2}^v$
1.218	1.005	-0.00039	-0.00019
$\lambda_{2,3}^v$	$\lambda_{2,1}^w$	$\lambda_{2,2}^w$	b_2
0	0	0	-71.77

Algorithm 1: DQN 기반 에너지 설비 최적 운영 알고리즘

하이퍼파라미터: 감가율 $\gamma = 1$, 학습률 $\eta > 0$, ϵ-greedy 계수 $\kappa \in (0,1)$, mini-batch 크기 $|\phi|$, 타겟 파라미터 업데이트 간격 N_o, 최대 반복학습 수 N.
입력값: 탐색 시간 T.
초기값: ϵ-greedy 확률 $\epsilon \in (0,1)$, replay 메모리 $\Phi = \emptyset$, 반복 수 초기값 $n = 0$, 파라미터 초기값 μ', Q-function 초기값 $Q(s,a;\mu^-)$, target network 파라미터 초기값 $\mu^- = \mu'$
while $n \leq N$ **do**
  for t=0, $\cdots$,T **do**
    ϵ 확률로 무작위 행동 a_t을 선택; 그 외의 경우 $a_t^* = \arg\max_{a_t} Q(s_t, a_t; \mu')$.
    a_t를 적용하여 식 (5), (6)을 활용하여 다음 상태 s_{t+1}를 계산.
    보상 r_t를 식 (9)를 토대로 계산.
    (s_t, a_t, s_{t+1}, r_t)를 replay 메모리 Φ에 저장.
    Φ에서 $|\phi|$ 크기의 mini-batch ϕ를 무작위로 구성.
    식 (15)를 토대로 손실함수 기울기의 근사값을 계산.
    파라미터 업데이트: $\mu' \leftarrow \mu' - \eta \nabla \hat{\mathcal{L}}(\mu')$.
    if t/N_o 가 정수인 경우 **then**
      target network 파라미터 업데이트: $\mu^- \leftarrow \mu'$.
    end
    ϵ 업데이트: $\epsilon = \kappa\epsilon$.
  end
  반복 수 업데이트: $n \leftarrow n + 1$.
end

그림 3 DQN 기반 에너지 설비 최적 운영 알고리즘

표 1에서 확인할 수 있듯이, 외부 미세먼지 농도 증가 및 내부 습도 증가는 내부 미세먼지 농도 증가와 상관관계가 있으며, 1번, 2번 공조기가 내부 미세먼지 농도를 낮추는 효과가 있음을 의미한다. 결과를 토대로 만들어지는 PM2.5, PM10에 대한 보상함수는 각각 식 (16), (17)과 같다.

$$i_{t+1}^{(1)} = 1.830 o_t^{(1)} + 0.488 h_t - 0.00047 v_k^{(1)} - 0.00009 v_k^{(2)} - 35.64 \quad (16)$$

$$i_{t+1}^{(2)} = 1.218 o_t^{(2)} + 1.005 h_t - 0.00039 v_k^{(1)} - 0.00019 v_k^{(2)} - 71.77 \quad (17)$$

이때 1번 공조기의 최대 전력사용량은 2000W, 2번 공조기의 최대 전력사용량은 1200W로 측정됨에 따라 각각 400W 단위로 운전을 제어하여 1번 공조기는 총 6가지, 2번 공조기는 총 4가지의 행동, 총 24가지의 행동을 선택할 수 있도록 구성하였다.

그림 3에서 제시한 알고리즘을 토대로 Python을 기초로 Tensorflow, Keras를 활용하여 DQN 알고리즘을 적용하였다. Q-function의 근사값을 계산하는 인공신경망의 파라미터는 표 2와 같이 구성하였으며, 학습 시 15분 단위로 업데이트 된 한 달 간의 데이터를 활용하여 학습을 진행하였다. (T=2880)

표 2 인공신경망 파라미터 세팅
Table 2 Parameters of artificial neural network

파라미터	값
은닉층 개수	2
노드 개수	[128, 32]
활성화함수	ReLU
학습률	0.001
최적화툴	Adam
배치 사이즈	64
총 반복 수	3500

표 2의 인공신경망을 이용하여 진행한 학습과정은 다음과 같다. 그림 4와 그림 5는 각각 학습과정 중 보상함수와 손실함수 값의 변화를 나타낸다. 두 그림에서 확인할 수 있듯이, 목적함수인 손실함수 값이 최소가 되도록 파라미터를 업데이트함에 따라 0에 가깝게 수렴함을 확인할 수 있으며, 동시에 보상함수의 값은 증가하여 수렴함을 알 수 있다. 이는 DQN 방법을 활용한 강화학습을 통해 최적 정책으로 수렴함을 의미한다.

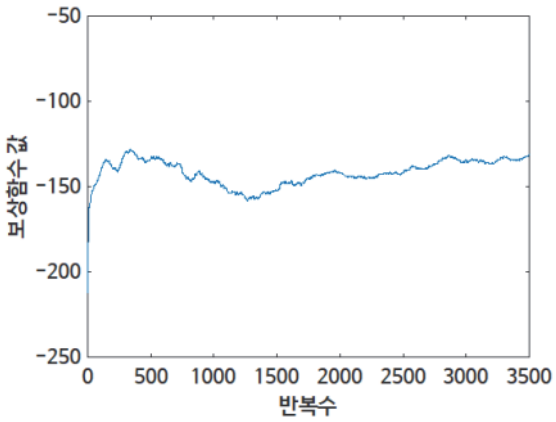


그림 4 학습 과정 중 보상함수 값의 변화
Fig. 4 Compensation function during deep learning training

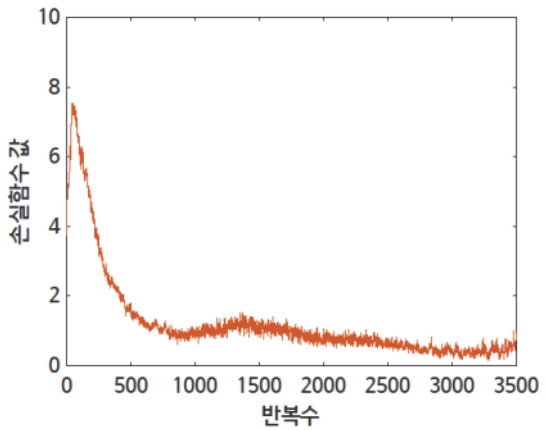


그림 5 학습과정 중 손실함수 값의 변화
Fig. 5 Loss function during deep learning training

이어서 학습된 인공신경망을 토대로 일주일 간(월요일~일요일)의 테스트를 진행한 결과 1번, 2번 공조기의 전력사용량은

그림 6과 같으며, 이를 각 요일별 평균 전력 사용량을 계산하여 정리한 결과는 표 3과 같다. 이어서 공조기의 제어에 따른 역사 내 PM2.5, PM10의 농도는 그림 7과 같음을 확인하였다. 그림 7에서 확인할 수 있듯이 미세먼지 농도가 증가하는 경우(월, 화요일과 토, 일요일) 에이전트는 공조기 1, 공조기 2의 전력사용량을 증가시켜 미세먼지 농도가 감소하는 것을 확인할 수 있다. 이는 외부 미세먼지 농도와 상관관계가 있는 역사 내 미세먼지 농도가 증가할수록 보상함수가 감소하며, 이때 에이전트는 공조기의 전력사용량을 증가시켜 미세먼지 농도를 제어하기 때문이다.

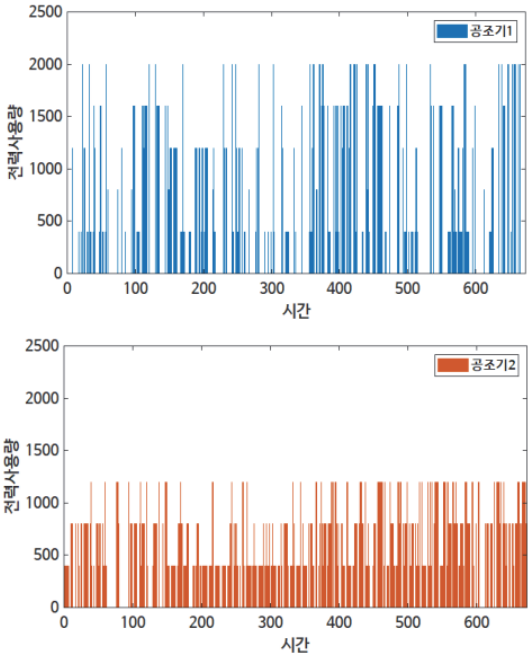


그림 6 공조기의 전력사용량
Fig. 6 Power consumption of HVAC system

표 3 공조기 각각의 전력사용량과 총 전력사용량
Table 3 Power consumption of each HVAC device

시간	월	화	수	목	금	토	일
공조기1	538	725	450	479	658	550	721
공조기2	542	550	475	600	508	788	725
총 전력량	1080	1275	925	1079	1166	1338	1446

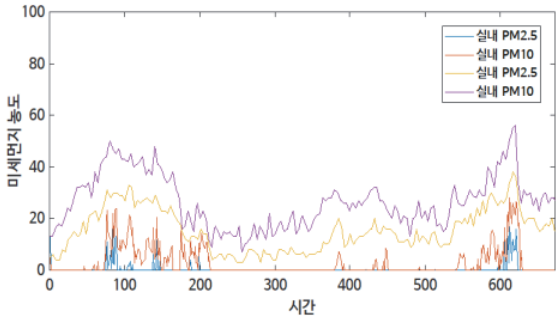


그림 7 공조기 제어에 따른 역사 내 미세먼지 농도 변화
Fig. 7 Particulate matter according to the control of HVAC system

6. 결 론

본 논문에서는 역사 내 미세먼지를 제어하기 위한 강화학습 기반의 에너지 관리 에이전트를 개발하였다. 이와 관련하여 먼저 공조기, 송풍기를 제어하여 역사 내 미세먼지를 관리하는 문제를 마르코브 의사결정 모델로 구성하였다. 이때 공조기, 송풍기의 제어에 따른 역사 내 미세먼지의 변화를 예측하기 위해 외부 미세먼지 농도, 내부 습도와의 상관관계에 기초하여 선형 보상함수 모델을 개발하여 모델에 포함시켰다. 위 모델에 기초하여 DQN 기반의 강화학습 에이전트를 학습시켜 역사 내 미세먼지 농도 감소의 가치와 전력사용량 간의 관계에 따라 공조기, 송풍기의 전력사용량을 제어하는 에이전트를 개발하였다. 사례연구에서는 남광주역의 데이터를 활용하여 선형 보상함수를 개발하고 DQN 방법의 인공신경망을 학습시켰으며, 이를 통해 공조기와 송풍기의 전력사용량을 제어하여 역사 내 미세먼지를 관리하는 에이전트를 개발하였다. 에이전트는 미세먼지가 증가하게 되면 공조기의 전력사용량을 증가시켜 미세먼지를 감소시켰으며, 이를 통해 강화학습에 기반한 에너지 관리 에이전트가 공조기의 전력사용량 및 역사 내 미세먼지 농도를 동시에 관리할 수 있음을 확인할 수 있었다.

Acknowledgements

This research was supported by a grant from R&D Program of the Korea Railroad Research Institute, Republic of Korea

References

[1] B. Jung, “Measurement and Management of Find Dust in Railroad and Station,” National Technology Proposal Insight, pp.1-29, vol. 2, Feb. 2021.

[2] J. Kim, K. Lee, and J. Bae, “Construction of real-time Measurement and Device of reducing fine Dust in Urban Railway,” 2020 Summer Conference of the Korea Society for Railway, pp.101-102, Jul, 2020.

[3] Y. Roh, W. Park, C. Lee, Y. Kim, D. Park, and S. Kim, “A Study of PM levels in Subway Passenger Cabins in Seoul Metropolitan are,” Journal of Korean Society of Occupational and Environmental Hygiene, vol. 17, no. 1, pp.13-20, Mar, 2007.

[4] M. Kim, K. Han, H. Kim, H. Gong, C. Kim, and J. Jeong, “A study on diffusion and distribution of PM10 in metropolitan subway tunnel,” Journal of The Korea Society For Urban Railway, vol. 4, no. 4, pp.577-582, Dec. 2016.

[5] J. Park, J. Park, and S, , “Estimation of Diffusion Direction and Velocity of PM10 in a Subway Station (For Gaehwasan Station of Subway Line 5 in Seoul,” Journal of Korean Society of Transportation, vol. 28, no. 5, Oct. 2010.

[6] J. Kang, C. Shin, S. Bae, S. Kwon, S. Kim, and S. Han, “Pre-study for the improvement of air filtration performance

in the air handling unit of subway station," 2008 Conference of Society of Air -Conditioning and Refrigerating Engineers of Korea, pp.541-545, 2008.

[7] Y. Lee, S. Park, S. Kwan, T. Kim, and D. Park, "A Study on the Building Energy Efficiency Rating Changes by Enhanced Thermal Insulation Performance of Building Envelope Standards in Apartment Houses," 2018 Summer Conference of Society of Air-Conditioning and Refrigerating Engineers of Korea, pp.935-936, Jun, 2018.

[8] S. Kwon, and S. Kim, "Intelligent AI-based Find Dust Management System," Railway Journal, pp.58-64, vol. 22, no. 1, Feb. 2019.

[9] J. Kim, K. Lee, J. Park, and M. Kim, "A Study on Management Measures for Aerosol Dust Reduction of the Urban Railway in DJET," 2019 Winter Conference of the Korean Society for Railway, pp.134-135, Nov. 2019.

[10] H. Lim, T. Yin, and Y. Kwon, "A Study on the Optimization of the Particulate Matter Reduction Device in Underground Subway Station," 2019 Spring Conference of the Korean Institute of Industrial Engineers, pp. 3786-3786, Apr. 2019.

[11] Y. Lee, Y. Woo, J. Koo, E. Jo, H. Park, and T. Lim, "A Study on vane and slot shape of air conditioner"s duct for improvement indoor cooling air temperature distribution," 2016 Spring Conference of the Korean Society for Railway, pp.646-651, May. 2016.

[12] Y. Cho, B. Roh, In. Yang, J. Lee, K. Jeong, Y. Lee, J. Kim, D. Park, and S. Kwon, "Study on the Effects of Fan on Particulate Matters Movement in Railroad Tunnel," 2019 Winter Conference of the Korean Society for Railway, pp.83-84, Nov. 2019.

[13] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. The MIT Press, 2018.

[14] R. Y. Rubinstein, D. P. Kroese, "Simulation and the Monte Carlo Method, 3rd ed. Wiley, 2016

[15] B. Recht, "A tour of reinforcement learning: The view from continuous control," Annual Review of Control, Robotics, and Autonomous Systems, vol. 2, no. 1, pp. 253-279, 2019.

[16] L.-J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," Machine Learning, vol. 8, no. 3, pp. 293-321, May 1992.

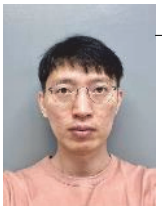
[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529-533, Feb 2015.

[18] Keras. https://github.com/fchollet/keras. Accessed: 2021-08-27.

저자소개

권경빈(Kyung-bin Kwon)

He received a B.S. and M.S. degree in Electrical and computer engineering from Seoul National University, Republic of Korea, in 2012 and 2014, respectively. He is currently pursuing a Ph.D. degree from The University of Texas at Austin from 2019. He is currently on an internship in R&D department of Raon Friends, Anyang, South Korea.



홍수민(Sumin Hong)

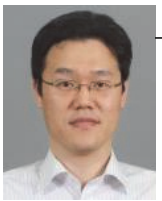
He received a B.S degree in Naval Architecture and Ocean Engineering from Seoul National University, Republic of Korea, in 2008. Currently, He is a team leader at RaonFriends Co., Ltd., Korea from 2019. He recent research interests include the Power system, Urban railroad and AI.

허재행(Jae-Haeng Heo)

He was born in Korea in 1978. He received his Ph.D. degree in Electrical Engineering from Seoul National University, Korea. Currently, he works at the RaonFriends Co. that is a consulting company for the power system and power system economics. His research field of interest includes power system reliability, equipment maintenance and urban railroad.

정호성(Hosung Jung)

He received a B.S and M.S. degree in Electrical engineering from Sungkyunkwan University, Republic of Korea, in 1995 and 1998, respectively. He received a Ph.D. degree from the Electrical Electronic and Computer Engineering from Sungkyunkwan University in 2002. He is currently a chief Researcher with the Smart Electrical & Signaling Division, Korea Railroad Research Institute, Uiwang, South Korea.



박종영(Jong-young Park)

Jong-young Park received the B.S., M.S., and Ph.D. degrees from Seoul National University, Seoul, Korea, in 1999, 2001, and 2007, respectively. He was a Senior Researcher at LS Electric Co., Ltd., Korea from 2009 to 2013. Currently, he is a Senior Researcher at Korea Railroad Research Institute (KRRI) since 2013. His recent research interests include the optimal operation of power systems in railway with the smart grid technology.