ISSN 1975-8359 [Print] / ISSN 2287-4364 [Online]
The Transactions of the Korean Institute of Electrical Engineers, vol. 72, no. 11, pp. 1586~1592, 2023
https://doi.org/10.5370/KIEE.2023.72.11.1586

Reinforcement Learning-based Energy Storage System Control for Optimal Virtual Power Plant Operation

가상발전소 최적 운영을 위한 강화학습 기반 에너지 저장장치 제어

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

Abstract

In this paper, we design a framework of the energy storage system (ESS) controller in virtual power plant (VPP) that maximize the profit. We consider the VPP that includes photovoltaics, wind turbines and demand along with ESSs and describe the environment based on Markov decision process (MDP). To find the best policy for ESS charging and discharging control, we implement a deep Q-network (DQN) method that trains a neural network which estimates Q-function values for each possible discrete actions. In the numerical test utilizing real-world data of Namgwangju Station, ERCOT and US government, we train the DQN and demonstrate that the proposed algorithm converges. Through the test with the trained policy, we showcase that the policy functions effectively in the scenario with uncertainty from renewable generations and load, as it responds adaptively to electricity prices.

Key Words

Deep Q-Network, Markov Decision Process, Energy Storage System, Reinforcement Learning, Virtual Power Plant

1. 서 론

전 세계적인 기후변화의 위기를 극복하기 위해 에너지 시스템의 전환이 강조되면서 재생에너지 발전 확대가 가속화되고 있다 [1]. 또한 정부는 탄소 중립을 실현하기 위한 뉴딜 정책을 추진하고 있으며, 이 정책의 일환으로 신재생 에너지의 전력 네트워크 안정성을 확보하고 에너지 전환을 촉진하기 위해 가상 발전소(Virtual Power Plant, VPP) 플랫폼 기술을 적용하고 있다 [2].

가상발전소는 풍력발전, 태양광발전 및 기타 신재생 에너지원과 에너지 저장장치를 통합한 시스템으로, 전력 생산과 수요 관리를 조율함으로써 에너지 효율성을 극대화하는 것이 필요하다 [3]. 이러한 시스템의 핵심 부분 중 하나가 에너지 저장장치의 최적 제어이고 에너지 저장장치를 효율적으로 제어함으로써 전력 네트워크의 안정성을 유지하면서 에너지 비용을 최소화할 수 있다. 에너지 저장장치는 생산된 전기 에너지를 보관하고, 필요한 때에 공급할 수 있는 시스템으로, 전력 가격이 저렴한 시기에 충전하여, 가격이 높아질 때 에너지를 사용하거나 필요한 시설에 전력을 공급함으로써 에너지 비용을 절감할 수 있다[4].

그러나 가상 발전소와 에너지 저장장치를 결합하여 최적으로 운영하기 위해서는 다양한 요소를 고려해야 한다. 시간, 온도, 습도, 에너지 수요, 전력 가격 등 다양한 불확실성 요소가 에너지 운영을 결정하는 데 영향을 미치고 이러한 다양한 변수와 불확실성을 고려하여 최적의 운영 정책을 찾아야 한다 [5]. 이러한 문제는 확률분포를 사용하지 않고 모델 정보 없이도 불확실성 데이터를 활용하여 최적 운영 정책을 찾을 수 있는 강화학습(Reinforcement learning; RL)을 활용하여 해결할 수 있다.

강화학습은 현재 상태를 기반으로 선택 가능한 행동 중 총 평균 보상을 최대화하는 최적 정책(Policy)을 찾는 방법으로 [6], 이를 위해 매개변수화된 함수를 활용하여 정책을 직접 찾는 대신, 해당 정책에 대응하는 매개변수를 조절하는 방법이 효율적이다 [7]. 이 매개변수화된 함수는 주로 두 가지 방법으로 구현된다. 첫 번째는 Q 함수 또는 가치함수(Value function)를 매개변수화하는 방법이며, 두 번째는 정책을 직접 매개변수화하는 방법이다. 예를 들어, Deep Q-Network (DQN) 방법은 Q 값을 근사화하기 위해 인공신경망을 구축하고, 이를 통

† Corresponding Author: Electrification System Research Department, Korea Railroad Research Institute, Korea
E-mail:jypark@krri.re.kr
https://orcid.org/0000-0001-8821-3251
* RaonFriends Co., Ltd., Korea
https://orcid.org/0000-0002-1069-4750 https://orcid.org/0009-0008-4378-3750
https://orcid.org/0000-0001-6164-9178
** Electrification System Research Department, Korea Railroad Research Institute, Korea
https://orcid.org/0000-0003-4546-2041
Received: Sep. 06, 2023 **Revised:** Oct. 04, 2023 **Accepted**: Oct. 07, 2023

해 Q 함수를 매개변수화한다 [8]. 이때 Q 함수는 인공신경망의 가중치 행렬로 표현되며, 각 가중치 값을 조정하여 최적 정책을 찾아낸다 [9].

반면, 정책 경사법(Policy Gradient Method)은 정책을 조건부 확률로 표현하고, 해당 확률 분포의 매개변수를 최적화함으로써 최적 정책을 찾는 방법이다 [10]. 예를 들어, 선형 정규분포 정책(Linearized Gaussian Policy)을 가정한 경우, 확률 분포를 매개변수화하여 상태(state)에 따른 평균과 분산을 설정하고, 이를 최적화하여 최적 정책을 구하게 된다 [11].

본 연구에서는 풍력발전, 태양광발전, 에너지 저장장치와 수요로 이루어진 가상발전소를 모델링하고 운영 이득을 최대화하는 에너지 저장장치의 충·방전 제어 정책을 구하였다. 이를 위하여 2장에서는 가상발전소의 모델링을, 3장에서는 강화학습을 적용하기 위해 마르코브 의사결정 과정을 구현하였다. 4장에서는 강화학습 방법 중 Deep Q-Network (DQN) 방법을 적용하여 에너지 저장장치의 충·방전 제어의 최적 정책을 구하였다. 5장에서는 한국남동발전(주)와 미국 정부 및 ERCOT의 실제 데이터를 기반으로 사례연구를 진행하였으며, 학습 진행 여부와 학습으로 구한 정책의 성능을 분석하였다. 마지막으로 6장에서는 본 연구의 결론을 서술하였다.

2. 가상발전소 모델링

본 논문에서는 그림 1과 같이 풍력발전, 태양광발전, 에너지 저장장치와 수요로 이루어진 가상발전소를 고려하였다. 먼저 최대 출력을 각각 $\overline{W}$, $\overline{V}$인 풍력발전과 태양광 발전에 대하여, 시간 t에 각각 $w_t \in [0, \overline{W}]$와 $v_t \in [0, \overline{V}]$의 전력을 생산한다고 정의하였다. 가상발전소에 포함된 수요는 시간 t에 d_t의 전력을 소비한다. 에너지 저장장치의 최소, 최대 충전상태(State of charges; SoC)는 $\underline{E}$, $\overline{E}$로 나타낼 수 있으며, 따라서 시간 t의 충전상태는 $e_t \in [\underline{E}, \overline{E}]$로 정의한다. 시간 t에서 에너지 저장장치는 가상발전소 운영자의 제어에 따라 $b_t \in [\underline{B}, \overline{B}]$의 전력을 충전 또는 방전한다. 즉, $b_t > 0$인 경우 전력을 충전, $b_t < 0$인 경우 전력을 방전한다. 이때 b_t와 최소, 최대 충전상태 $\underline{E}$, $\overline{E}$를 고려하여 시간 $t+1$의 충전상태 e_{t+1}은 식 (1)와 같다 [12].

$$e_{t+1} = \max\left\{\min\left\{e_t + b_t, \overline{E}\right\}, \underline{E}\right\} \tag{1}$$

각 신재생에너지의 발전, 수요의 전력 소비 및 에너지 저장장치의 제어를 고려하였을 때 남거나 추가로 필요한 전력은 그림 1의 실시간 전력시장을 통해 구매하거나 판매한다고 가정한다. 따라서, 시간 t에서 가상발전소의 이득 r_t는 식 (2)과 같다.

$$r_t = \delta_t(w_t + v_t - b_t - d_t) \tag{2}$$

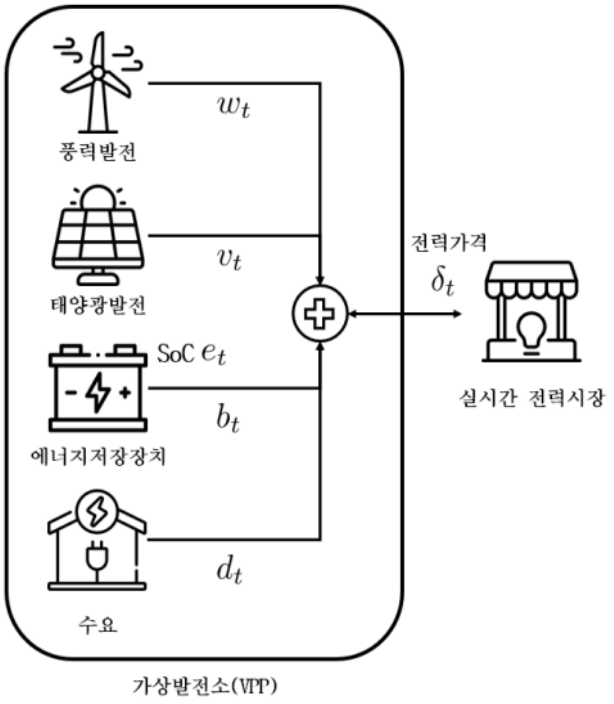


그림 1 가상발전소 모델링
Fig. 1 Modeling of a Virtual Power Plant

3. 마르코브 결정과정 모델링

본 장에서는 강화학습을 적용하기 위해서 2장에서 모델링한 가상발전소 모델링을 마르코브 결정과정(Markov Decision Process; MDP)로 나타낸다. 마르코브 결정과정은 행동(Action), 전이함수(Transition kernel), 보상(Reward), 감가율(Discount factor)로 구성된다 [13]. 가상발전소 모델링을 토대로 각각을 정의하면 아래와 같다.

- 상태(State)는 가상발전소의 구성요소와 실시간 전력시장의 상태를 모두 포함한다. 즉, 시간 t에 대하여 상태 $s_t = [w_t, v_t, e_t, d_t, \delta_t]$의 다섯 가지 요소로 정의할 수 있다.

- 행동(Action)은 가상발전소의 구성요소 중 제어 가능한 에너지 저장장치의 충·방전 전력 b_t로 정의할 수 있다. 이때 행동을 continuous action으로 표현할 경우 b_t의 값을 그대로 이용할 수 있으며, discrete action으로 표현할 경우 행동집합(action space)를 이산값으로 나타냄으로써 정의할 수 있다. 예를 들어 5개의 행동이 가능하다고 할 때, 행동집합 $A = [0,1,2,3,4]$로 정의되며, 각각의 행동 $a_t \in A$에 따른 b_t의 값은 아래와 같이 나타낼 수 있다.

$$b_t = \begin{cases} \underline{B}, & a_t = 0 \\ 0.5\underline{B}, & a_t = 1 \\ 0, & a_t = 2 \\ 0.5\overline{B}, & a_t = 4 \\ \overline{B}, & a_t = 5 \end{cases} \quad (3)$$

- 전이함수(Transition kernel)은 상태 s_t의 각 구성요소에 대하여 다르게 표현될 수 있다. 제어가 불가능한 w_t, v_t, d_t, δ_t의 경우 마르코브 성질을 따른다고 가정하고 아래와 같이 나타낼 수 있다 [14]. 예를 들어, w_t의 전이함수는 식 (4)와 같으며, v_t, d_t, δ_t의 경우도 같은 식으로 나타낼 수 있다. 반면 제어 가능한 e_t의 전이함수는 식 (1)의 정의를 그대로 사용할 수 있다.

$$\Pr(w_{t+1}|\{w_\tau\}_{\tau=1}^{t}) = \Pr(w_{t+1}|w_t) \quad (4)$$

- 보상(Reward)는 식 (2)의 표현을 그대로 사용할 수 있다. 즉, 실시간 전력시장의 가격이 높을수록, 가상발전소에 사용하고 남은 전력이 많을수록 보상은 커진다.

- 감가율(Discount factor) $\gamma \in (0,1]$는 현재의 보상과 미래의 보상 간의 비를 의미한다. 즉, γ의 값이 작을수록 현재의 보상을 미래의 보상보다 더 가치 있게 여김을 의미한다. 본 논문에서는 $\gamma = 1$을 사용하였으며, 이는 현재의 보상과 미래의 보상이 같은 가치를 가짐을 의미한다.

4. Deep Q-Network 기반 강화학습

4장에서는 3장에서 정의한 마르코브 결정과정을 토대로 최적 에너지 저장장치의 충·방전 제어 정책 π를 구하기 위해 Deep Q-Network(DQN) 기반 강화학습을 적용한다 [8]. 먼저 3장에서 정의한 보상함수 및 감가율을 토대로 에너지 저장장치의 최적 운영 정책 π는 다음의 최적화 문제로 나타낼 수 있다.

$$\max_\pi E_\pi[\sum_{t=1}^{T}\gamma^t r_t] \quad (5)$$

즉, 최적 정책 π는 시간 T 까지의 감가율을 고려한 시간대별 보상의 평균값을 최대화하는 정책으로 정의할 수 있다. 위를 만족하는 최적 정책 π를 찾기 위해서 본 논문에서는 DQN 방식을 적용하였다. DQN 방식은 상태 s_t에서 행동 a_t를 적용하였을 때의 가치를 나타내는 Q함수 $Q(s_t,a_t)$의 값을 추정하는 인공신경망을 학습한다. 이를 수식으로 나타내면 식 (6)와 같다.

$$Q(s_t, a_t) = E_\pi[\sum_{\tau=t}^{T}\gamma^{(\tau-t)} r_\tau(s_\tau, a_\tau)|s_t, a_t] \quad (6)$$

즉 Q값은 정책 π를 따라서 행동 a_t를 선택할 때, 시간 t부터 T까지의 기대 보상의 평균값을 의미한다.

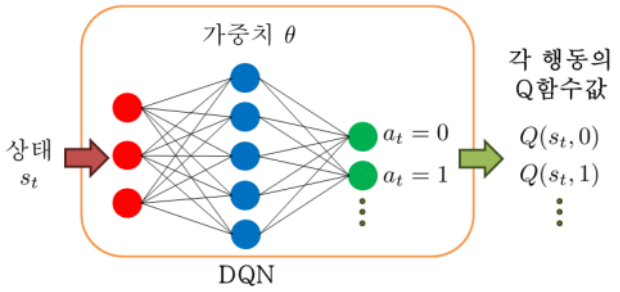


그림 2 DQN 기반 강화학습의 인공신경망
Fig. 2 Artificial Neural Network for DQN-based Reinforcement Learning

그림 2는 DQN 기반 강화학습에 활용하는 인공신경망을 나타내고 있다, 인공신경망의 입력값은 s_t이며, 따라서 입력 노드의 총 개수는 s_t의 구성요소 개수와 같다. 출력값은 각 행동에 따른 Q값이며, 따라서 출력 노드의 총 개수는 가능한 a_t의 개수이다. 따라서 최적 가중치 θ를 찾음으로써 Q값을 정확하게 추정할 수 있는 경우, 최적 행동은 아래와 같이 같은 s_t에 대하여 가장 큰 $Q(s_t,a_t)$값을 가지는 a_t를 최적 행동으로 선택할 수 있다.

$$\pi_\theta : a_t = \arg\max_{a'} Q(s_t, a') \quad (7)$$

여기서 π_θ는 가중치 θ로 나타내어지는 파라미터화된 정책을 의미한다. 즉, 식(5)는 최적 정책 π대신 θ를 찾는 식(7)으로 바뀌게 된다.

$$\max_\theta E_{\pi_\theta}[\sum_{t=1}^{T}\gamma^t r_t] \quad (8)$$

한편 Q함수는 식 (9)의 벨만 방정식(Bellman equation)을 만족하므로, 아래 수식을 만족하는지 확인함으로써 Q함수 값을 정확하게 추정하는지 확인할 수 있다 [15].

$$\begin{aligned} &Q^*(s_t, a_t) \\ &= r_t(s_t, a_t) + \gamma E_{s_{t+1}}[\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})|s_t, a_t] \end{aligned} \quad (9)$$

즉, 식 (10)과 같이 식 (9)의 좌변과 우변의 차이를 나타내는 손실함수 $\mathcal{L}(\theta)$를 최소화하는 최적 파라미터 θ값을 찾음으

로써 Q함수 값을 정확하게 추정하는 Q-network를 구성할 수 있다.

$$\mathcal{L}(\theta) = E_{\{s_t, a_t, s_{t+1}\}}[(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta) - Q(s_t, a_t; \theta))^2] \quad (10)$$

이때 좌변과 우변에 같은 가중치 θ를 적용하는 경우, 업데이트 시 좌변과 우변이 함께 변화하게 되므로 손실함수가 0으로 수렴하지 않을 수 있다. 이러한 수렴의 불안정성을 해결하기 위해 Fixed Q-target 방식을 적용하였다 [17]. Fixed Q-target 방식은 파라미터 θ를 좌변과 우변에 같이 사용하는 대신, 서로 다른 파라미터 θ, θ'을 식 (11)과 같이 각각 Train network와 Target network에 적용한다.

$$\mathcal{L}(\theta) = E_{\{s_t, a_t, s_{t+1}\}}[(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta') - Q(s_t, a_t; \theta))^2] \quad (11)$$

이때 Target network의 파라미터 θ'는 θ^-값으로 고정하고, N_O번의 업데이트에 1번씩 θ^-를 업데이트함으로써 Target 값의 파라미터를 고정한다. 반면 Train network의 파라미터 θ는 반복학습 시 매번 업데이트를 수행함으로써 손실함수 값을 0으로 수렴시킨다. Target의 파라미터가 고정되어있기 때문에, Fixed Q-target 방식은 수렴의 불안정성을 해소할 수 있다.

한편 식 (11)에서 볼 수 있듯이 손실함수 $\mathcal{L}(\theta)$의 최소값은 0이므로, 최적 파라미터 θ는 식 (11)을 최소화하는 값이다. 이를 구하기 위해 DQN 방식에서는 경사하강법을 적용한다. 따라서 손실함수를 파라미터 θ에 대하여 편미분한 기울기(gradient)는 아래와 같다.

$$\nabla \mathcal{L}(\theta) = E_{\{s_t, a_t, s_{t+1}\}}[-2(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) - Q(s_t, a_t; \theta)) \nabla_\theta Q(s_t, a_t; \theta)] \quad (12)$$

마지막으로 모든 $\{s_t, a_t, s_{t+1}\}$ 조합에 대하여 기댓값을 구하는 것은 어렵기 때문에 현재 정책을 기준으로 샘플링을 통해 경로(trajectory)를 구성하고, 이에 대한 평균값을 계산하여 기울기의 근사값을 구한다. 즉, 총 $|\psi|$개의 샘플링에 대하여, 근사화된 손실함수의 기울기는 식 (13)와 같다.

$$\nabla \widehat{\mathcal{L}}(\theta) = (-2/|\psi|) \sum_{t \in \psi} [(r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta^-) - Q(s_t, a_t; \theta)) \nabla_\theta Q(s_t, a_t; \theta)] \quad (13)$$

결과적으로 식 (13)의 기울기를 활용해 파라미터 θ는 식 (14)의 경사하강법을 통해 업데이트 된다.

$$\theta \leftarrow \theta - \eta \widehat{\nabla} \mathcal{L}(\theta) \quad (14)$$

이때 알고리즘 초기에는 충분히 서로 다른 정책을 탐색(exploration)하기 위해 ϵ-greedy 방법을 적용하였다 [6]. ϵ-greedy 방법은 ϵ의 확률로 무작위 행동을 선택하고, $(1-\epsilon)$의 확률로 식 (7)의 최적 행동을 선택한다. 이때 ϵ값은 반복학습이 진행됨에 따라 감소하게 함으로써 알고리즘 초기에는 무작위 행동에 따른 Q값을 확인하고, 이후 ϵ값이 충분히 감소하게 되면 최적 행동을 선택하게 된다. 추가로 학습 중 발생하는 샘플을 효과적으로 활용하기 위해 샘플 (s_t, a_t, s_{t+1}, r_t)를 메모리 Φ에 저장하고, 이 중 무작위로 $|\psi|$ 크기의 mini-batch를 구성하여 식 (13)을 계산한다 [16]. 위에서 설명한 DQN 알고리즘은 그림 3에서 확인할 수 있다.

Algorithm 1: DQN 기반 가상발전소의 에너지 저장장치 최적 정책 알고리즘

하이퍼파라미터: 감가율 $\gamma = 1$, 학습률 $\eta > 0$, ϵ-greedy 계수 $\kappa \in (0, 1)$, mini-batch 크기 $|\phi|$, 타겟 매개변수 업데이트 간격 N_0, DQN 최대 반복학습 수 N
입력값: 탐색 시간 T
초기값: ϵ-greedy 확률 $\epsilon \in (0, 1]$, Replay 메모리 $\Phi = \emptyset$, 반복수 초기값 $n = 0$, 매개변수 초기값 θ', 타겟 매개변수 초기값 $\theta^- = \theta'$
while $n \leq N$ **do**
  for t=0, $\cdots$,T **do**
    ϵ의 확률로 무작위 행동 a_t를 선택; 그 외의 경우 $a_t^* = \arg\max_{a_t} Q(s_t, a_t; \theta')$
    s_t, a_t를 토대로 식 (1), (4)를 통해 다음 상태 s_{t+1}을 계산.
    보상 r_t를 식 (2)를 통해 계산.
    (s_t, a_t, s_{t+1}, r_t)를 메모리 Φ에 저장.
    Φ에서 $|\phi|$ 크기의 mini-batch ϕ를 무작위로 구성
    식 (13)을 통해 손실함수 기울기의 근사값을 계산.
    매개변수 업데이트: $\theta' \leftarrow \theta' - \eta \nabla \mathcal{L}(\theta')$
    if t/N_o가 정수인 경우 **then**
      타겟 매개변수 업데이트 $\theta^- \leftarrow \theta'$.
    end
    ϵ 업데이트: $\epsilon = \kappa\epsilon$.
  end
  반복수 업데이트: $n \leftarrow n + 1$.
end

그림 3 DQN 기반 가상발전소의 에너지 저장장치 최적 정책 알고리즘
Fig. 3 Algorithm for Optimizing Energy Storage Device Policies in a DQN-based Virtual Power Plant

5. 사례연구

위에서 소개한 DQN 기반 강화학습의 효과를 입증하기 위해서, 실제 데이터를 기반으로 사례 연구를 진행하였다. 먼저, 학습에 필요한 풍력 발전 및 태양광 발전 데이터는 한국남동발전(주)의 2022년 시간대별 발전량을 기반으로 하여 각각 1.5MW 및 1MW로 스케일링하였다 [17]. 시간대별 수요 및 실시간 전력시장 가격 데이터는 미국 정부의 상업 지역 및 주거 지역 전력 사용량에 대한 공개 데이터 및 미국 ERCOT 데이터를 활용하였으며, 수요 데이터는 0.5MW로 스케일링했다 [18-19]. 에

너지 저장장치의 용량은 1MW이며, 최소 State of Charge (SoC)는 0.1MW로 설정되었다. 완전한 충전에 걸리는 시간은 1.5시간으로 설정되었고, 이에 따라 충전 및 방전의 최대 전력량은 각각 0.6MWh로 설정되었다. 에너지 저장장치의 제어는 식 (3)과 같이 총 5가지 제어 방식 중에서 선택되며, 이에 따라 0.6MWh 충전, 0.3MWh 충전, 충·방전 없음, 0.3MWh 방전, 0.6MWh 방전의 다섯 가지 제어 방식이 가능하도록 구성하였다. 마지막으로, 전력시장은 15분 간격으로 운영되며, 따라서 하루는 총 96개의 시간 단위로 분할되었다. 사용한 DQN 모델의 학습 파라미터는 표 1에 제시한 것과 동일하며, 매 학습 단계마다 무작위로 업데이트된 3일치 데이터를 15분 간격으로 활용하여 학습을 진행했다. 이때 은닉층 및 노드 개수는 5개의 입력값과 5개의 출력값을 기준으로 논문 [20]를 참고하여 정하였으며, 그 결과 그림 3에서 제시한 알고리즘을 토대로 Python에 기초하여 Tensorflow, Keras를 활용하여 DQN을 구성하였다 [21].

표 1 DQN 파라미터
Table 1 DQN parameters

하이퍼파라미터	값
은닉층 노드 개수	[32, 16]
Mini-batch 크기	64
활성화함수	Softmax
학습률	0.01
반복수	2000
초기 ϵ	0.3
ϵ-greedy 계수	0.997

먼저 표 1을 토대로 구성한 DQN을 활용하여 알고리즘 1을 토대로 진행한 학습 결과는 아래와 같다. 먼저 그림 4은 DQN의 학습 과정 중 총 보상의 변화를 나타내고 있다. 그래프에

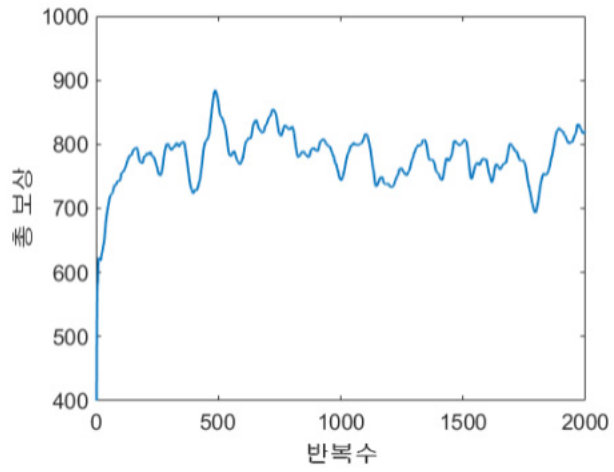


그림 4 DQN 학습 과정 중 총보상 값의 변화
Fig. 4 Variation in Total Reward Value during DQN Training Process

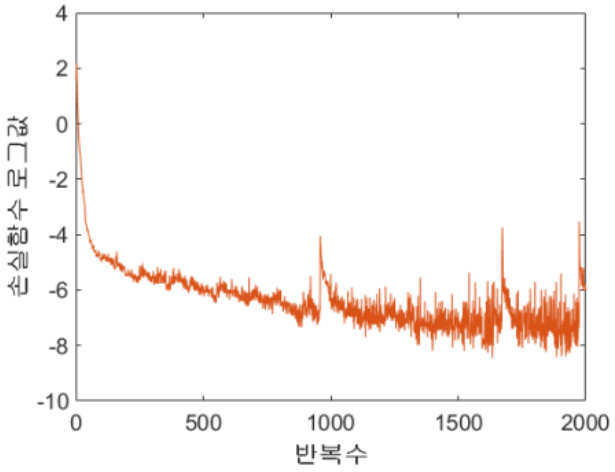


그림 5 DQN 학습 과정 중 손실함수 로그값의 변화
Fig. 5 Variation in Logarithmic Loss Function Values during DQN Training Process

서 확인할 수 있듯이, 무작위로 3일을 선택하기 때문에 총 보상 값이 변하는 것을 계속 변하지만 전체적으로 총보상 값이 학습을 진행함에 따라 증가하는 것을 확인할 수 있다. 이는 그림 5의 손실함수 로그값에서도 확인할 수 있다. 손실함수 값이 0에 수렴하는 것은 DQN의 출력값이 실제 Q함수 정확하게 추정한다는 것을 의미하며, 따라서 손실함수 로그값이 음수가 될수록 0에 더 가까운 것을 의미한다. 그림 4에서 확인할 수 있듯이 손실함수 로그값은 크게 감소하여 -6에 근접하는 것을 확인할 수 있으며, 이를 통해 손실함수 값이 0에 충분히 가까운 것을 알 수 있다.

이어서 학습한 DQN을 토대로 무작위로 선택한 3일간의 데이터에 해당 정책을 적용하여 에너지 저장장치의 운영에 대한 테스트를 총 5회 진행하였다. 그림 6는 테스트에 사용된 시간대별 풍력발전, 태양광발전, 수요 및 실시간 전력가격 데이터 중 하나를 나타내고 있다. 각 값은 표준화된 값으로 표현되어 [0,1] 사이의 값으로 표현되어 있다. 그림 6의 데이터를 토대로 학습된 DQN을 활용하여 정책을 적용한 에너지 저장장치 운영 결과는 그림 7와 같다. 그림 7에서 볼 수 있듯이, 에너지 저장장치는 3일 동안 매일 충·방전을 실시하여 전력가격의 차액을 통한 이득을 내는 것을 확인할 수 있다. 이때 에너지 저장장치의 충·방전은 전력 가격에 가장 큰 영향을 받는 것을 확인할 수 있다. 즉, 에너지 저장장치를 운영할 때 전력 가격이 낮을 때 충전을 한 뒤 전력가격이 높은 시간대에 방전함으로써 차액거래를 통해 이득을 최대화하게 된다. 이는 총 5회 진행한 테스트의 총 보상값과 에너지저장장치의 차액거래를 통한 보상값을 나타내고 있는 그림 8을 통해서 확인할 수 있다. 즉, 표 8에 나타내어지는 두 보상값이 모두 양수임을 확인할 수 있으며, 이는 에너지 저장장치의 차액 거래가 효과적으로 운영됨에 따라 양의 총 보상값을 가지게 되는 것을 나타낸다.

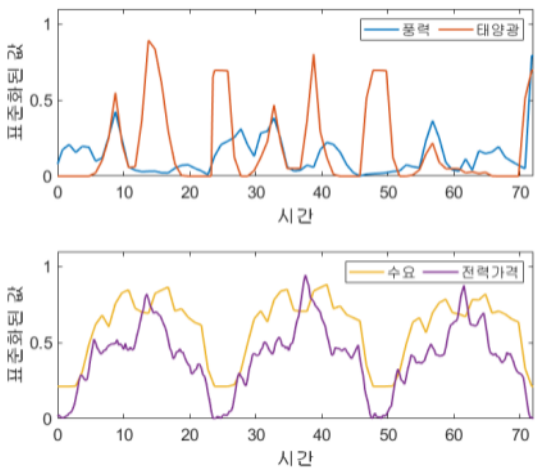


그림 6 시간대별 풍력발전, 태양광발전, 수요 및 실시간 전력가격 데이터
Fig. 6 Time-series Data of Wind Power Generation, Solar Power Generation, Demand, and Real-time Electricity Prices

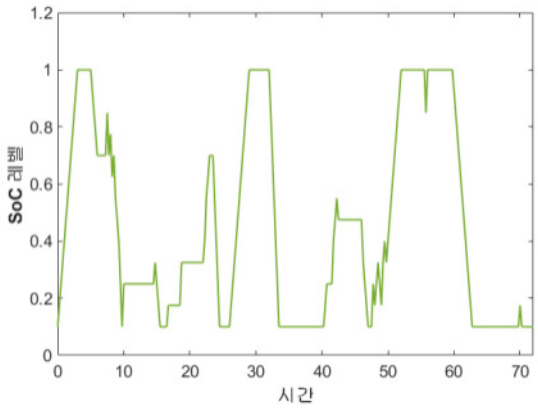


그림 7 에너지 저장장치의 SoC 레벨 변화
Fig. 7 State of Charge (SoC) Level Changes in Energy Storage Devices

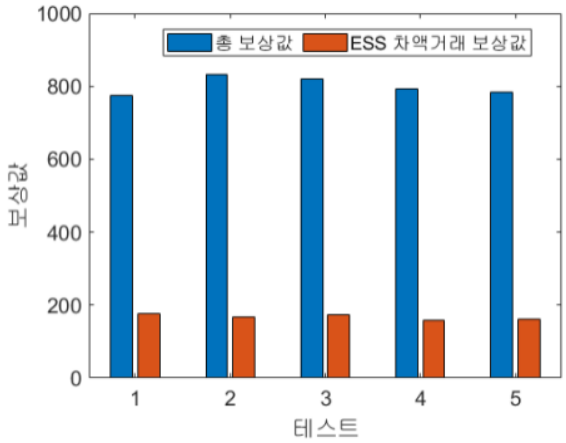


그림 8 총 보상값과 에너지 저장장치의 차액거래 보상값
Fig. 8 Total Reward Value and Differential Trading Reward Value of Energy Storage Devices

6. 결 론

본 논문에서는 풍력발전, 태양광발전, 수요와 에너지 저장장치로 구성된 가상발전기를 모델링하고, 에너지 저장장치의 충·방전 제어를 위한 최적 제어 정책을 강화학습을 통해 구하였다. 이를 위해 먼저 마르코브 결정과정으로 문제를 설정하고 에너지 저장장치의 파라미터를 토대로 discrete action으로 충·방전을 표현하였다. 최적 제어 정책을 구하기 위해 Q값을 추정하고 Q값을 최대로 하는 행동을 선택하는 정책을 구현하는 DQN 방식을 적용하였으며, 추정하는 Q값과 실제 Q값의 차이를 나타내는 손실함수를 최소화하도록 DQN의 가중치를 경사하강법을 이용하여 업데이트하였다. 사례연구에서는 한국남동발전(주)와 미국 정부 및 ERCOT의 실제 데이터를 활용하여 DQN의 학습을 진행하였으며, 손실함수가 0으로 수렴하는 것을 확인함으로써 Q값을 정확하게 추정하는 DQN을 구현하였다. 이어서 학습된 DQN 정책을 적용한 테스트에서는 에너지 저장장치가 실시간 전력가격에 반응하여 차액거래를 통해 이득을 내는 것을 확인하였다. 후속 연구로는 본 논문에서 개발한 에너지 저장장치의 최적 제어 정책을 기존의 운영 방법과 비교함으로써 본 논문에서 개발한 방법의 유효성을 입증하는 것과 온도, 풍속 등 날씨 데이터를 포함하여 보다 실용적인 제어 정책을 구성하는 것을 제안한다.

Acknowledgements

This research was supported by a grant from the R&D program (Development of smart energy management and performance evaluation technology for railway stations based on virtualization, PK2303E1) of the Korea Railroad Research Institute, Republic of Korea.

References

[1] J. Lee, C. Jung and S. Son, “P2H Technical Potential Analysis of Korea for Renewable Energy Acceptance,” The Proceedings of the Korean Institute of Electrical Engineers, pp. 187-188, 2022.

[2] Chung, K. H., Park, M. G., Cho, S. B., and Cho, K. S., “Analysis on the Prerequisites to Deploying Virtual Power Plant (VPP) in Smart Grid Environment,” The Proceedings of the Korean Institute of Electrical Engineers, pp. 493-494, 2014.

[3] J, Ryu and J, Kim. “Case Study of Overseas Virtual Power Plant Operation to activate VPP in Korea Electricity Market,” The Proceedings of the Korean Energy Society, pp. 133-135, 2021.

[4] G. An, “The importance and role of Energy Storage Systems,” In The Proceedings of the Korean Institute of Illuminating and Electrical Installation Engineers, vol. 26, no. 2, pp. 13-17, Mar. 2012.

[5] J. Lee, S. Lee and J. Kim, “Optimal operation of virtual power plants using machine learning-based new and renewable energy prediction,” pp. 81, 2020.

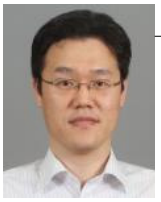
[6] R. S. Sutton and A. G. Barto, “Reinforcement Learning: An Introduction, 2nd ed,” The MIT Press, 2018.

[7] 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.
[8] M. Roderick, J. MacGlashan and S. Tellex, “Implementing the Deep Q-Network,” arXiv, 2017.
[9] D. Silver, A. Huang,, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, G., J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, and S. Dieleman, “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484-489, 2016.
[10] J. Peters and J. A. Bagnell, “Policy Gradient Methods. In: Sammut, C., Webb, G. (eds),” Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA., 2016.
[11] J. Peters and S. Schaal. “Natural actor-critic,” Neurocomputing, vol 71. no. 7-9, pp. 1180-1190, 2008.
[12] K. Kwon and H. Zhu, “Reinforcement Learning-Based Optimal Battery Control Under Cycle-Based Degradation Cost,” IEEE Transactions on Smart Grid, vol. 13, no. 6, pp. 4909-4917, 2022.
[13] A. Pieter and A. Y. Ng., “Apprenticeship learning via inverse reinforcement learning,” In Proceedings of the twenty-first international conference on Machine learning, pp. 1, 2004.
[14] A. A. Markov, “An Example of Statistical Investigation of the Text Eugene Onegin Concerining the Connection of Samples in Chains,” Science in Context, vol. 19, no. 4, pp. 591-600, 2006.
[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, “Playing Atari with Deep Reinforcement Learning,” arXiv preprint arXiv, 2013.
[16] L. Lin, “Self-improving reactive agents based on reinforcement learning, planning and teaching,” Machine Learning, vol. 8, no. 3, pp. 293-321, 1992.
[17] Ministry of Public Administration and Security [Online], Available: https://www.data.go.kr/data/15043275/fileData.do.
[18] Data.gov [Online], Available: https://catalog.data.gov/dataset/?tags=energy-consumption.
[19] ERCOT Market Price [Online], Available: http://www.ercot.com/mktinfo/prices
[20] Kyung-bin Kwon, Su-Min Hong, Jae-Haeng Heo, Hosung Jung, and Jong-young Park. “Development of Reinforcement Learning-based Energy Management Agent for HVAC Facilities and ESS,” The transactions of The Korean Institute of Electrical Engineers, 71(10), 1434-1440, 2022.
[21] Keras. [Online] Available: https://github.com/fchollet/keras

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.



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.

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.

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.