Korean Chem. Eng. Res., 63(1), 50-58 (2025)
https://doi.org/10.9713/kcer.2025.63.1.50
PISSN 0304-128X, EISSN 2233-9558

강화학습을 이용한 천연가스 액화 공정 최적화에 관한 연구

이지은* · 박경태*,**,†

*숙명여자대학교 화공생명공학부
04310 서울특별시 용산구 청파로 47길 100
**주식회사 엔엑스엔시스템즈
06588 서울특별시 서초구 방배로 26길 32
(2024년 8월 23일 접수, 2024년 12월 11일 수정본 접수, 2024년 12월 11일 채택)

A Study on Optimization of Natural Gas Liquefaction Process Using Reinforcement Learning

Jieun Lee* and Kyungtae Park*,,†**

**Department of Chemical & Biological Engineering, Sookmyung Women's University, Seoul, 04310, Korea*
***Research and Development Center, NXN Systems Co., Ltd., Seoul, 06588, Korea*
(Received 23 August 2024; Received in revised from 11 December 2024; Accepted 11 December 2024)

요 약

본 연구에서는 강화학습 방법론 중 Deep Q-Network(DQN)와 Advantage Actor-Critic(A2C)알고리즘을 이용하여 천연가스 액화공정 중 단일혼합냉매 공정을 최적화하고 각 알고리즘에 따른 에너지 소모량 결과를 유전 알고리즘(Genetic algorithm)을 통한 최적화 결과와 비교 분석하였다. 그 결과 DQN 최적화 결과가 A2C보다 낮은 에너지 소모량을 보였으며 학습 시간은 A2C 알고리즘이 짧은 것을 확인하였다. 그러나 유전 알고리즘과 비교분석 결과 유전 알고리즘의 최적화 결과가 가장 좋았으며, 강화학습을 통한 공정의 최적화를 위해 연속적인 변수를 다루는 행동 지정에 대한 연구가 필요함을 제시하였다.

Abstract – In this study, Deep Q-Network and Advantage Actor-Critic algorithms among reinforcement learning methodologies were used to optimize the single-mixed refrigerant process for a natural gas liquefaction. And optimization results using these algorithms were compared with the results of genetic algorithm (GA). The results showed that the optimization results using the DQN algorithm had lower energy consumption than A2C, and the learning time was shorter for the A2C algorithm. However, the comparison analysis with the genetic algorithm (GA) showed that the GA had the best performance, suggesting that research on specifying actions that deal with continuous variables is necessary for optimizing the process through reinforcement learning.

Key words: Liquefied natural gas, Single mixed refrigerant process, Reinforcement Learning, Optimization, Deep Q-Network, Advantage actor-critic algorithm

1. 서 론

경제 성장, 인구 증가에 따른 에너지 수요의 증가 및 에너지안보에 대한 관심의 증가로 전세계적으로 천연가스에 대한 수요가 증가하고 있으며, 2050년까지 이 추세가 이어질 것으로 전망하고 있다[1]. 또한, 천연가스는 화석연료 중 이산화탄소, 질소산화물 및 황 산화물의 배출이 가장 적다는 장점도 있어 신재생에너지 기반의 새로운 에너지 패러다임으로 전환하기 전까지 주요 에너지원으로서 역할을 충분히 수행할 수 있을 것으로 예측된다[2].

천연가스는 이송 거리에 따라 파이프라인을 통한 이송과 액화천연가스(Liquefied natural gas)를 통한 이송이 있다. 보통 장거리 이송의 경우 액화천연가스를 통한 이송이 더 경제적이라고 알려져 있다[3]. 액화천연가스는 가스전이나 유전에서 채취한 천연가스를 전처리 한 후 액화천연가스 액화 플랜트에서 -162 ℃의 온도로 냉각하여 액화시킨 것이다. 이후 액화천연가스는 액화천연가스 운반선을 통해 최종 목적지까지 운반된다.

천연가스를 액화천연가스로 만들기 위해서는 액화천연가스의 낮은 끓는점(-162 ℃)으로 인해 많은 에너지가 투입된다. 따라서, 액화천연가스 액화 공정의 효율을 높이기 위해 다양한 공정이 개발되었으며, 대표적으로 Propane pre-cooled mixed refrigerant process (C3MR), Dual mixed refrigerant process (DMR), Single mixed refrigerant

† To whom correspondence should be addressed.
E-mail: ktpark@sm.ac.kr

process (SMR), Dual nitrogen expansion process 등이 있다. 이러한 공정들은 공정의 복잡성과 다양한 결정변수(Decision variable)들로 인해 최적화 조건을 찾기가 매우 어려우며, 이에 따라 다양한 최적화 방법론이 꾸준히 연구되어 왔다[2-5].

구글 딥마인드의 알파고는 강화학습(Reinforcement Learning, RL)을 이용한 인공지능(Artificial Intelligence) 바둑 프로그램으로, 실제 대국을 진행하여 뛰어난 성능을 증명한 사례가 있다[6]. Trial-and-error를 통해 경험으로부터 학습하는 강화학습은 복잡한 계산을 요하는 환경에서 최적의 경로를 찾는 데에 적합하며, 로보틱스, 자율주행 등 다양한 산업에서의 활용을 위한 연구가 진행되고 있다. 화학 공장에서는 주로 공장 제어 분야에서 강화학습 모델이 연구되어 왔으며, 2023년 실제 증류탑 제어에 자율제어 인공지능 모델을 도입하여 안정성을 검증한 결과가 발표되었다[7]. 해당 모델은 날씨 등 외부 요인의 변화에 유연하게 대응할 뿐만 아니라 공정 자체의 변화가 생겨도 동일한 모델을 사용할 수 있으며, 이는 제어하고자 하는 공정 시스템마다 동적 모델이 필요한 모델 예측 제어 방법을 효과적으로 보완하는 대안책으로서 이점을 가진다.

최근에는 강화학습을 통한 공정 최적화 연구가 시도되고 있다. 공정 시스템 공학에서 최적의 공정 설계를 위해 주로 쓰이는 superstructure 방법은 유전 알고리즘을 포함하여 다양하게 발전했지만, 복잡한 초기 모델링과 더불어 계산 과정에서 드는 시간과 비용 측면에서 낮은 효율을 보인다. 이러한 한계에 대한 대안으로서 강화학습은 공정설계 방법론에서 최적화기반의 superstructure-free 방법에 속하며, 강화학습 알고리즘은 고차원 문제의 계산과 모델의 유연한 적용이 가능해 superstructure 방법의 한계를 극복하는 효과적인 최적화 방법론으로서 공정 산업에서 다양한 활용이 가능할 것으로 예상된다. Gao와 Schweidtmann은 이러한 강화학습을 이용한 공정설계 연구에 대한 다양한 사례를 조사하였다[8]. Kim 등은 SMR공정에 강화학습의 한 종류인 Deep Q-Network (DQN)을 적용하여 최적화를 진행하였고, 적절한 행동(action)과 보상함수의 설정을 통해 비선형성이 강한 최적화 문제에 DQN을 성공적으로 적용할 수 있음을 보였다[9]. Seidenberg 등은 메탄개질반응(Steam methane reforming)에 대해 Hierarchical reinforcement learning approach를 적용하였다[10]. 그 결과 데이터뿐만 아니라 전문지식도 정형화하여 공정 설계에 적용할 수 있는 가능성을 보였다. Stops 등은 hierarchical, hybrid actor-critic 에이전트(agent)를 이용하여 공정합성이 가능함을 보였다[11]. 이 연구에서는 공정도를 구성하기 위해 graphical neural network (GNN)을 이용하였고 구성된 공정도의 최적화를 위해 multi-layer perceptron (MLP)를 이용하였다. Chen과 Wang은 이산화탄소 포집 공정의 최적 운전점을 actor-critic 알고리즘을 통해 찾고자 하였다[12]. 먼저, 공정 모사 소프트웨어를 이용하여 강화학습 에이전트를 학습시킨 후 파일롯 설비에 강화학습 에이전트를 적용하여 최적 운전점으로 운전가능한지를 확인하였다. 그 결과 공정모사만으로 학습한 에이전트는 실제 공정을 운전하는데 한계가 있다는 사실을 밝혀내었다.

지금까지 살펴보았듯이, 기존의 연구는 강화학습 방법론 중 주로 한 가지 방법론을 이용하여 연구를 진행해왔다. 하지만, 강화학습 방법론에는 정책 기반, 가치 기반 또는 이 들을 결합한 방법 등 다양한 방법이 있으며, 공정설계에 어떤 방법론이 더 좋은지에 대한 연구는 제대로 이루어지지 않고 있다. 따라서, 이번 연구에서는 오프폴리시(Off-policy) 강화학습 방법 중 하나인 DQN과 정책 및 가치를 둘 다 사용하는 A2C 알고리즘을 이용하여 천연가스 액화 공정 중 SMR 공정의 최적화를 수행하고 이 두 알고리즘을 비교 분석하였다. 또한, 베이스라인 설정을 위해 대표적인 최적화 방법인 유전 알고리즘을 통한 최적화도 수행하여 강화학습을 통한 최적화 방법과 비교하였다.

2. 연구 방법론

2-1. 강화 학습(Reinforcement Learning)

강화 학습은 기계 학습의 한 분류로 주어진 데이터를 바탕으로 학습하는 지도학습(Supervised Learning) 및 비지도학습(Unsupervised Learning)과는 달리 에이전트가 환경으로부터 수집한 경험을 토대로 학습한다[13]. Fig. 1에서 나타내었듯이 에이전트는 시행착오를 통해 학습하며, 주어진 특정 상황에서 어떤 행동을 하는 것이 더 좋은 보상을 얻을 수 있는지 스스로 발견한다는 점에서 인간의 학습 방법과 매우 비슷한 점이 특징이다.

따라서, 강화학습은 순차적인 행동의 결정이 필요한 문제에 대한 좋은 해결책이 될 수 있으며, 순차적 행동결정 문제를 수학적으로 정의하기 위해 Markov Decision Process (MDP)가 사용된다[9]. MDP의 구성요소는 상태, 행동, 보상함수, 상태 변환 확률 및 할인율이 있으며, 각 구성요소는 다음과 같이 정의할 수 있다.

- 상태(state) s, s∈S이며 S는 에이전트가 관찰가능한 상태의 집합, S_t는 시간 t에서의 가능한 상태의 집합
- 행동(action) a, a∈A이며, A는 에이전트가 취할 수 있는 행동의 집합, A_t는 시간 t에서 에이전트가 취할 수 있는 행동의 집합
- 보상함수 r(s, a), 시간 t에서 $S_t = s$이고 $A_t = a$일 때 받을 보상에 대한 기대값 $r(s, a) = E[R_t|S_t = s, A_t = a]$, R_t은 t시점에 취한 행동에 의해 받게 되는 보상
- 전이 확률 P, 상태 s에서 행동 a를 했을 때 다음 상태 s'에 도달할 확률
- 감쇠인자 γ, 나중에 받는 보상에 대한 할인율, $\gamma \in [0, 1]$.

먼 미래에 받는 보상은 현재에 받는 보상에 비해 가치가 떨어지므로

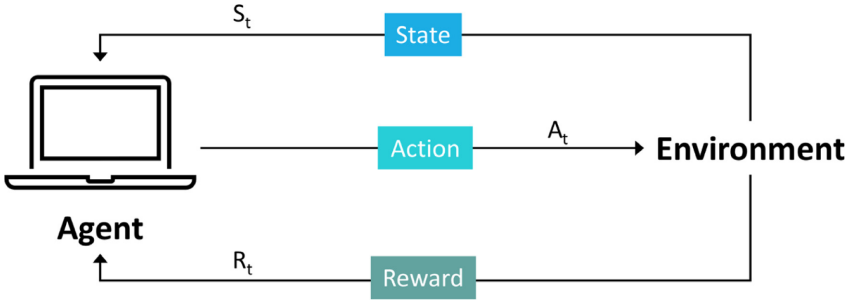

Fig. 1. Conceptual diagram of reinforcement learning.

이를 반영하기 위해 보상을 계산할 때, 할인률 γ를 고려한다. 또한, 에이전트가 실제로 환경을 경험하면서 받게 되는 누적 보상의 합을 반환값(Return)이라고 하며, 반환값 G는 다음과 같이 표현할 수 있다[13].

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} R_k \tag{1}$$

에이전트는 강화학습을 통해서 수식 1로 정의되는 반환값을 최대화할 수 있는 최적 정책을 학습하여야 하며, 정책은 다음과 같이 나타낼 수 있다.

$$\pi(a|s) = P[A_t = a|S_t = s] \tag{2}$$

수식 2에서 나타내었듯이 정책이란 상태 s에서 행동 a를 취할 확률이다. 결국 최적의 정책을 학습하기 위해서는 에이전트가 선택하는 행동에 대한 가치를 알아야 하며, 이는 다음과 같다.

$$Q_\pi(s,a) = R + \gamma \sum_{s' \in S} P^a_{SS'} \sum_{a' \in A} \pi(a'|s') Q_\pi(s',a') \tag{3}$$

여기서 $Q_\pi(s, a)$는 정책 π에 따라 상태 s에서 행동 a를 선택했을 때, 선택한 행동의 가치를 나타내는 함수이며, 행동가치함수(action-value function) 또는 Q함수(Q function)라고 불린다. R은 상태 s에서 행동 a를 선택했을 때의 보상, $P^a_{SS'}$ 는 상태 s에서 행동 a를 취했을 때 s' 상태가 되는 전이확률이다.

최적의 정책을 학습하기 위한 최적의 Q함수는 벨만 최적 방정식(Bellman optimality equation)에 의해 표현할 수 있다.

$$Q_*(s,a) = R + \gamma \sum_{s' \in S} P^a_{SS'} (\max_a, Q_*(s',a')) \tag{4}$$

결국 최적 정책을 위한 최적의 Q함수는 다음 상태(s')에서 선택가능한 행동 중 가능 높은 값을 가지는 Q함수에 감쇠인자를 곱하여 현재 상태의 보상에 더한 것과 같다. 따라서, 최적 정책 $\pi(s, a)$는 현재 상태에서 Q함수를 최대화할 수 있는 행동을 선택하면 되므로, 다음과 같이 표현할 수 있다.

$$\pi_*(s,a) = \text{argma}\ x_a Q_*(s,a) \tag{5}$$

2-2. Deep Q-Network (DQN) and Advantage Actor-Critic (A2C) 알고리즘

DQN알고리즘은 구글 딥마인드(DeepMind)에 의해 처음 소개되었다[14]. DQN은 대표적인 오프폴리시(Off-Policy) 강화학습 방법인 큐러닝의 단점-예를 들어 큐러닝의 경우 모든 상태-행동 쌍에 대한 Q함수를 학습해야 함, 작은 규모의 문제에서는 가능하나 대규모 문제에서는 현실적으로 불가능함-을 극복하기 위해 Q함수를 인공신경망을 통해 근사하는 것이 특징이다. 이를 위해 DQN은 경험 리플레이(Experience Replay)와 타깃 네트워크(Target Network)를 도입하여 학습의 안정성 및 수렴성을 높였다[9].

우선 전통적인 큐러닝에서 전이확률 P = 1이라고 가정하면, Q함수의 업데이트는 다음 식에 의해 이루어진다.

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R + \gamma(\max_{a'} Q(s',a') - Q(s,a))) \tag{6}$$

여기서 α는 학습율이며(learning late) $\max_{a'} Q(s',a')$ 는 다음 상태 s'에서 선택 가능한 행동 중 가장 높은 Q함수의 값이다.

한편, DQN에서는 타깃 네트워크를 사용하여 다음 상태 s'에서의 최대 Q값을 계산하고 이를 사용하여 식 7과 같이 목표 타깃(y)을 설정한다. 여기서 θ^-는 타깃 네트워크의 가중치로 학습의 안정성을 위해 일정주기 마다 업데이트를 한다.

$$y = R + \gamma(\max_{a'} Q(s',a';\theta^-)) \tag{7}$$

따라서, DQN의 손실함수는 타깃 네트워크가 계산한 목표타깃과 현재 네트워크의 예측값의 평균 제곱 오차(MSE)를 사용하여 다음과 같이 나타낸다.

$$L(\theta) = (y - Q(s,a;\theta))^2 \tag{8}$$

연구에 사용된 DQN 알고리즘을 Table 1에 나타내었다.

Table 1. DQN Algorithm used in this study

DQN Algorithm:
Initialize the Q-network with random weights θ
Initialize the target network with weights $\theta^- = \theta$
Initialize the replay memory D with capacity N
for episode e = 0, E-1 **do**
  Initialize the environment (process simulation)
  Observe the initial state s_0 from the environment
    for time step t = 0, $T-1$ **do**
      With probability ε, select a random action a_t; otherwise select $a_t = argmax_a Q(s_t, a; \theta)$
      using the current Q-network
      Execute the selected action a_t and get reward r_t and the next state s_{t+1} from the enviro nment
      Store set of a tuple (s_t, a_t, R_t, s_{t+1}) in the replay memory D
if $e \geq e_{start}$ **then**
  if $\varepsilon > \varepsilon_{min}$ **then**
    Decay ε with decay rate r_d
  Randomly sample a mini-batch of n tuples (s_j, a_j, R_j, s_{j+1}) from the replay memory D
  Compute the target Q-value $y_j = R_j + \gamma(\max_{a'} Q(s_{j+1}, a'; \theta^-))$
  Update the current Q-network by minimizing the loss function $L(\theta)$:

$$L(\theta) = \frac{1}{N}(y_j - Q(s_j, a_j; \theta))^2$$

end for
Update the target Q-network: $\theta^- \leftarrow \theta$
end for

따라서, DQN의 손실함수는 타깃 네트워크가 계산한 목표타깃과 현재 네트워크의 예측값의 평균 제곱 오차(MSE)를 사용하여 다음과 같이 나타낸다.

A2C알고리즘은 액터(actor)가 정책에 따라 현재 상태에서 어떤 행동을 취할 것인지 선택하고 크리틱(critic)은 선택된 행동의 가치를 평가하는 알고리즘이다. 또한 어드밴티지(advantage)함수를 도입하여 액터가 선택한 행동에 대한 보상과 크리틱이 평가한 행동의 가치의 차이를 이용해 액터의 정책을 업데이트하며, 이 때, 어드밴티지 함수는 다음과 같이 정의된다.

$$A(s,a) = R + \gamma V(s') - V(s) \tag{9}$$

여기서, V는 크리틱의 가치신경망을 이용하여 계산한 가치함수가 된다. 따라서, A2C 알고리즘의 가중치는 다음 식을 이용하여 업데이트 한다.

$$\theta' \sim \theta + \alpha[\nabla_\theta \log \pi_\theta(a|s) A(s,a)] \tag{10}$$

여기서, θ는 액터의 정책신경망의 가중치이다, 원래는 어드밴티지 함수 자리에 실제 반환값이 들어가야하지만, 크리틱의 가치신경망을 통해 근사한 가치 함수를 통해 계산되는 어드밴티지 함수가 들어가므로 θ'를 근사값의 형태로 표현하였다.

한편, 크리틱의 가치신경망은 MSE를 이용하여 손실이 최소화되는 방향으로 가중치를 업데이트하며 손실함수는 다음과 같이 정의된다.

$$L(\theta_v) = (R + \gamma V(s') - V(s'))^2 \tag{11}$$

Table 2. A2C Algorithm used in this study

A2C Algorithm:
Initialize the actor (policy network) with random weights θ
Initialize the critic (value network) with random weights θ_v
for episode $e = 0, E\text{-}1$ **do**
  Initialize the environment (process simulation)
  Observe the initial state s_0 from the environment
    for time step $t = 0, T-1$ **do**
    Select a_t using the actor (current policy network, $\pi(a_t|s_t)$)
    Execute the selected action a_t and get reward r_t and the next state s_{t+1} from the enviro nment
    Compute the advantage $A(s_t, a_t) = R_t + \gamma V(s_{t+1}) - V(s_t)$
    Update the policy network (actor) parameters using the policy gradient:

$$\nabla_\theta \log \pi_\theta(a_t|s_t) A(s_t, a_t)$$

    Update the value network (critic) parameters by minimizing loss function:

$$L(\theta_v) = (R_t + \gamma V(s_{t+1}) - V(s_t))^2$$

  end for
end for

여기서, θ_v는 크리틱의 가중치이다. 연구에 사용된 A2C 알고리즘을 Table 2에 나타내었다.

2-3. Problem statement

본 연구에 사용된 천연가스 액화 공정을 다음 Fig. 2에 나타내었다. Fig. 2에서 보듯이 단일혼합냉매(Single Mixed Refrigerant. SMR) 공정은 가장 간단한 형태의 천연가스 액화공정으로 Air cooler를 포함한 혼합냉매 압축기(MR Compressor), 콜드박스(Main Cryogenic Heat Exchanger, MCHE), JT 밸브, 그리고 기액분리기(Separator)로 구성되어 있다. 일반적으로 혼합냉매는 질소, 메탄, 에탄, 프로판 그리고 부탄 등으로 구성되어 있으며, 혼합냉매의 조성에 따라 공정의 성능이 많이 달라진다는 특징이 있다.

단일혼합냉매 공정에서 천연가스는 콜드박스를 통과하며 완전히 액화되고 이후 JT-NG를 거치면서 팽창하여 더욱 온도가 내려간다. 이 때, 비응축성가스들(Boil-off gas, BOG)은 기액분리기에서 LNG와 분리된다. 한편, 혼합냉매는 혼합냉매압축기를 통해 적정 압력으로 가압된 후, Air-cooler와 콜드박스를 거치면서 냉각된다. 이후 JT-MR을 거치면서 팽창하여 더욱 온도가 내려간다. JT-MR을 통과한 혼합냉매는 해당 공정에서 가장 온도가 낮은 상태가 되며, 콜드박스에서 천연가스 액화와 혼합냉매를 냉각시키기 위한 냉열을 제공하여 기화한 후 다시 혼합냉매 압축기로 공급된다.

Table 3. Major assumptions [4,5,15]

Parameters	Unit	Values
Feed NG Compositions	-	Table 4
Feed NG flowrate	kg/h	626.4
Feed NG temperature	℃	20
Feed NG inlet pressure	bar	60
LNG production pressure	bar	1.05
LNG production temperature	℃	Saturated
Ambient temperature	℃	20
Pressure drop in heat exchangers	bar	0.2
Compressor/Expander Efficiency	-	0.8
Thermodynamic Package	-	Peng-Robinson

Table 4. Natural gas compositions[4]

Component	Mole fraction
Nitrogen	0.00185
Methane	0.92345
Ethane	0.0428
Propane	0.0221
n-Butane	0.0049
i-Butane	0.0049

공정모사에는 상용 공정모사기인 Aspen Plus V14를 사용하였으며, 공정모사에 사용된 주요 가정은 다음 Table 3에 정리하였다. 또한, 천연가스의 입구 조성은 Table 4에 정리하였다.

또한, 최적화를 위한 결정 변수들은 주어진 공정의 자유도 분석을 통해 선정되었으며, 변수들의 초기값과 최대, 최소 경계는 다음 Table 5에 정리하였다.

2-4. 최적화 방법론

강화학습을 위한 상태는 Table 5에 주어진 변수(총 10개)의 현재 상태로 정의하였다. 또한, 행동의 경우에는 주어진 변수를 일정한 크기로 증가 또는 감소하는 것으로 정의하였고, 자세한 정의는 Table 6에 정리하였다.

먼저, a_1~a_6는 해당 성분의 질량 유량을 10 kg/h씩 증가, a_7, a_8은 해당 압력을 0.5 bar씩 증가, a_9, a_{10}은 온도를 0.5 ℃ 증가하는 것으로 정의하였다. 또한, a_{11}~a_{16}은 해당 성분의 질량 유량을 10 kg/h씩 감소, a_{17}, a_{18}은 해당 압력을 0.5 bar씩 감소, a_{19}, a_{20}은 온도를 0.5 ℃ 감소하는 것으로 정의하였다. 따라서, 행동은 총 20개로 정의하였다.

또한, 강화학습을 통한 천연가스 액화 공정의 최적화를 위해서는

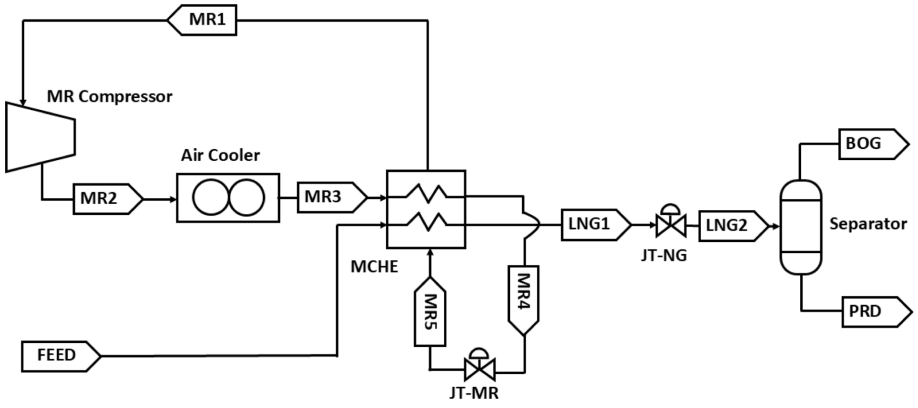

Fig. 2. Process flow diagram of the single mixed refrigerant process.

Table 5. Initial values, upper and lower bound of decision variables

Variables	Initial Values	Lower bound	Upper bound
Nitrogen flow rate (kg/h)	300	150	500
Methane flow rate (kg/h)	400	200	700
Ethane flow rate (kg/h)	1500	700	1900
Propane flow rate (kg/h)	100	0	200
n-Butane flow rate (kg/h)	2200	1000	3000
i-Butane flow rate (kg/h)	0	0	200
Compressor discharge pressure (bar)	20	10	30
JT-MR outlet pressure (bar)	3.8	2	10
MR 4 stream temperature (℃)	-160	-162	-150
LNG 1 stream temperature (℃)	-160	-162	-150

Table 6. Definitions of actions a_1~a_{20}

Decision variable	action	value	action	value
Nitrogen flow rate (kg/h)	a_1	+ 10 kg/hr	a_{11}	- 10 kg/hr
Methane flow rate (kg/h)	a_2	+ 10 kg/hr	a_{12}	- 10 kg/hr
Ethane flow rate (kg/h)	a_3	+ 10 kg/hr	a_{13}	- 10 kg/hr
Propane flow rate (kg/h)	a_4	+ 10 kg/hr	a_{14}	- 10 kg/hr
n-Butane flow rate (kg/h)	a_5	+ 10 kg/hr	a_{15}	- 10 kg/hr
i-Butane flow rate (kg/h)	a_6	+ 10 kg/hr	a_{16}	- 10 kg/hr
Compressor discharge pressure (bar)	a_7	+ 0.5 bar	a_{17}	- 0.5 bar
JT-MR outlet pressure (bar)	a_8	+ 0.5 bar	a_{18}	- 0.5 bar
MR 4 stream temperature (℃)	a_9	+ 0.5 ℃	a_{19}	- 0.5 ℃
LNG 1 stream temperature (℃)	a_{10}	+ 0.5 ℃	a_{20}	- 0.5 ℃

보상함수의 설정이 중요하다. 보상함수는 크게 제약사항을 잘 지켰을 때의 보상 R_{const}와 Specific Energy Consumption(이하, SEC)를 낮은 값으로 갈 때의 보상 R_{SEC}로 나누어 구성하였다. 먼저 R_{const}의 구성을 위해 다음과 같은 제약사항을 고려하였다.

1) 압축기의 입구 흐름의 기상 분율은 1이다.

2) 모든 열교환기의 최소 접근 온도(minimum temperature approach, MTA)는 2K이다.

이 제약사항을 고려한 보상함수 R_{const}는 다음과 같이 설정하였다.

$$R_{const,comp} = 3^{vf\text{-}vf_{min}},\ \text{if } vf_{min} > vf$$

$$R_{const,HX} = \begin{cases} 3^{MTA-MTA_{min}}, & \text{if } MTA > MTA_{min} \\ 3^{MTA-MTA_{min}}, & \text{otherwise} \end{cases} \quad (12)$$

$$R_{const} = \min(R_{const,comp},\ R_{const,HX})$$

여기서 $R_{const,\,comp}$는 압축기 입구에서 기상 분율을 고려한 보상함수이다. vf_{min}은 압축기 입구에서의 기상 분율 제약값이며 본 연구에서는 1로 설정하였다. vf는 압축기 입구에서의 기상 분율이다. $R_{const,HX}$는 열교환기의 MTA를 고려한 보상함수이며, MTA_{min}은 MTA의 제약값으로 본 연구에서는 2로 설정하였다. 결과적으로 R_{const}는 $R_{const,comp}$와 $R_{const,HX}$ 중 최소값으로 설정하였는데 제약사항을 모두 만족할 경우 R_{const}는 1이되고 그렇지 않을 경우는 수식 12에 의해 1보다 작은 값을 가지게 된다.

R_{SEC}는 다음 수식에 의해 정의되는 SEC를 최소화하는 경우 제공하는 보상함수이다.

$$SEC\left(\frac{kWh}{kg\ LNG}\right) = \frac{\text{Total Sum of Energy Consumption (kW)}}{\text{LNG Production Rate (kg/h)}} \quad (13)$$

이를 고려하여 다음과 같이 R_{SEC}를 구성하였다.

$$R_{SEC} = 1 - M(SEC - SEC_{max}) \quad (14)$$

여기서 SEC_{max}는 SEC의 최대값으로 이 연구에서는 1 kWh/kg LNG로 설정하였다. M은 계수로 적절한 값을 설정해야 하며, 이 연구에서는 0.5로 설정하였다. 따라서 R_{SEC}는 수식 13에 의해 SEC가 SEC_{max}보다 작을 경우는 R_{SEC}는 1보다 큰 값을 가지게 되며, SEC가 SEC_{max}보다 클 경우에는 1보다 작은 값을 가지게 된다.

따라서, 보상함수 R은 최종적으로 다음과 같이 표현할 수 있다.

$$R = \begin{cases} R_{SEC} + R_{const}, & \text{if lower bound} \le a_n \le \text{upper bound},\ 1 \le n \le 20 \\ -1, & \text{otherwise} \end{cases} \quad (15)$$

여기서, a_n은 행동으로 에이전트가 취하는 행동이 Table 5에 주어진 최소 및 최대값 경계를 충족할 경우 보상은 R_{SEC}와 R_{const}의 합이 되고 그렇지 않을 경우는 -1이 된다.

연구에 사용된 모든 코드는 파이썬을 통해 작성되었으며, 파이썬과 Aspen V14는 COM 인터페이스를 통해 연결하였다. 사용된 DQN과 A2C알고리즘의 자세한 정보는 Table 7에 정리하였다. 이때 각 알고리즘의 하이퍼파라미터는 공정의 복잡도를 반영하여 선택되었다. 앞서 정의한 결정 변수와 행동에 따라 두 신경망의 입력층과 출력층의 차원을 지정했고, 그 외의 하이퍼파라미터도 학습 과정에서 충분한 학습과 활용이 이루어지도록 적절한 값을 설정하였다. 또한, 최적화의 베이스라인을 설정하기 위해 유전알고리즘을 통한 최적화도 진행하였다. 사용된 GA에 대한 자세한 정보는 다음 Table 8에 정리하였다.

Table 7. Details of DQN and A2C algorithms

Item	DQN	A2C	
Discount factor (γ)	0.99	0.99	
Learning rate (α)	0.001	0.001	
Episodes	500	500	
Decay rate(r_d)	0.999	-	
ε_{min}	0.05	-	
Batch size	256	-	
Network Details		Actor	Critic
Input layer	10 × 64	10 × 64	10 × 64
2nd layer	64 × 64	64 × 64	64 × 64
3rd layer	64 × 32	64 × 32	64 × 32
Output layer	32 × 20	32 × 20	32 × 1
Activation function	ReLU	ReLU	
Optimizer	Adam	Adam	

Table 8. Details of genetic algorithms used in this study

Item	Values
Used package	PyGAD [16]
Objective function	Minimize SEC
Constraints	MTA ≥ 2 and vf ≥ 1
Decision variables	See Table 5
Generations	150
Population	50
Elite parent	10
Crossover type	Single point
Mutation type	Random
Mutation percentage	20%

3. 결과 및 고찰

모든 최적화는 AMD Ryzen Threadripper 3990X(2.92 GHz)의 cpu와 NVIDIA Titan RTX의 환경에서 진행되었다. 유전알고리즘은 CPU에서 DQN과 A2C 알고리즘은 GPU에서 진행되었다. 본 연구의 최적화 결과를 Table 9 및 Table 10에 정리하였다.

Table 9에서 보듯이 DQN 및 A2C 알고리즘 모두 베이스라인인 GA 알고리즘에 비해 SEC에서 더 좋은 결과를 보이진 못했다. 일반적인 SMR 공정의 에너지 효율은 MCHE의 MTA 값의 영향을 가장 크게 받는데, GA를 통한 최적화는 결정 변수를 연속적인 공간에서 다뤘기 때문에 MTA의 세밀한 조절이 가능했다. 하지만 DQN이나 A2C 알고리즘은 결정 변수를 이산적으로 다루기 때문에(Table 6참고) 최적화된 MCHE의 MTA 값이 제약값인 2 ℃와 각각 2%, 6%의 오차율을 보였으며, 결과적으로 최적화된 공정의 SEC가 GA보다 높게 나타난 것으로 보인다. 최적화된 결정 변수값은 Table 10에 정리

Table 9. Optimization results

Item	GA	DQN	A2C
SEC (kWh/kg LNG)	0.290	0.306	0.305
Power Consumption (kW)	172.1	183.4	178.2
LNG production rate (kg/h)	593.3	598.6	585.1
Reward	-	2.35	2.35
Vapor fraction @ compressor inlet	1	1	1
MTA @ MCHE (℃)	2.00	2.04	2.12
Execution time (min)	216.4	395.7	225.0

Table 10. Optimized decision variables

Decision Variables	GA	DQN	A2C
Nitrogen flow rate (kg/h)	421.3	320.0	360.0
Methane flow rate (kg/h)	583.7	480.0	450.0
Ethane flow rate (kg/h)	1617.3	1480.0	1450.0
Propane flow rate (kg/h)	2.9	120.0	10.0
n-Butane flow rate (kg/h)	2780.4	2240.0	2300.0
i-Butane flow rate (kg/h)	9.3	60.0	40.0
Compressor discharge pressure (bar)	17.0	15.5	19.5
JT-MR outlet pressure (bar)	4.8	3.8	4.3
MR 4 stream temperature (℃)	-156.7	-157.5	-158.0
LNG 1 stream temperature (℃)	-154.6	-156.0	-152.5

해두었다.

실행 시간의 경우에는 DQN이 가장 길고 유전알고리즘이 가장 짧았으며, A2C 알고리즘은 그 중간에 위치했다. 유전알고리즘의 경우 세대 당 인구수의 값을 50, 총 150세대를 지정하여 총 7500번의 연산이 이루어졌으며 DQN과 A2C는 각각 평균 step 수가 87.7, 61.1이며 500 에피소드의 학습을 진행했기 때문에 각각 43850, 30550번의 연산이 진행되었으며 이로 인해 DQN의 실행시간이 가장 길고, 유전알고리즘의 실행시간이 가장 짧았다. 다만, DQN이나 A2C의 경우 GPU 상에서 알고리즘을 실행했지만 대부분의 계산 시간은 Aspen Plus에 의존적인 것으로 보여, 실행 시간상 GA에 비해 이점은 없어 보였다.

Fig. 3에는 DQN 및 A2C의 학습곡선을 나타내었다. DQN의 경우 경험리플레이(본 연구에서는 에피소드 100까지)를 진행하는 동안 한 에피소드당 진행되는 step의 수의 편차가 크게 나타났다. 그러나 에피소드가 진행될수록 한 에피소드당 진행되는 step의 수가 매우 안정적으로 유지되었으며, 49번째 에피소드에서 최대 보상값인 2.35가 나왔다. A2C의 경우에는 초반부터 매우 작은 값의 손실을 보였지만 학습 내내 step 수의 편차가 크게 나타났으며, 403번째 에피소드에서 최대 보상값인 2.35가 나왔다.

더 나아가 DQN과 A2C 알고리즘의 하이퍼파라미터 중 학습률과 은닉층의 차원 및 보상함수의 M값을 조정하여 parametric study를 진행하였다. 강화학습 알고리즘의 하이퍼파라미터는 학습 결과를 결정하는 중요한 요소로, 본 연구에서는 학습률의 경우 0.0005와 0.005로, 은닉층의 차원은 32와 128로 조정하여 그 결과를 비교하였다. 또한 보상함수의 설정에서 0.5로 지정한 M의 값을 0.1과 1.0으로 바꾸어 최적화를 진행하여 SEC의 보상함수에 주어지는 가중치에 따른 최적화 결과를 분석하였다. 각 경우에 대해 최적화된 SEC 값과 최대 보상, 최대 보상에 도달한 에피소드 및 MTA는 Table 11에 정리하였다.

DQN 알고리즘의 학습률을 base 값인 0.001에서 0.0005와 0.005로 바꾸어 최적화를 진행한 결과, 학습률이 0.0005일 때 최적화된 SEC 값은 0.300 kWh/kg으로 base case의 SEC 값인 0.306 kWh/kg보다 낮은 값을 보였다. 반면 학습률을 0.005로 두고 최적화를 진행한 결과 손실함수가 발산하여 학습이 제대로 이루어지지 않았다.

또한 DQN 신경망의 은닉층 차원을 32와 128로 조정하여 최적화를 진행한 결과, 은닉층의 차원이 32일 때 모든 경우 중 가장 낮은 SEC 값인 0.294 kWh/kg을 가졌으며 차원이 128일 때는 학습 base case와 비슷한 보상과 SEC 값을 보였다.

본 연구에서 보상함수의 M값은 0.1일 경우 전체 보상함수의 구성에

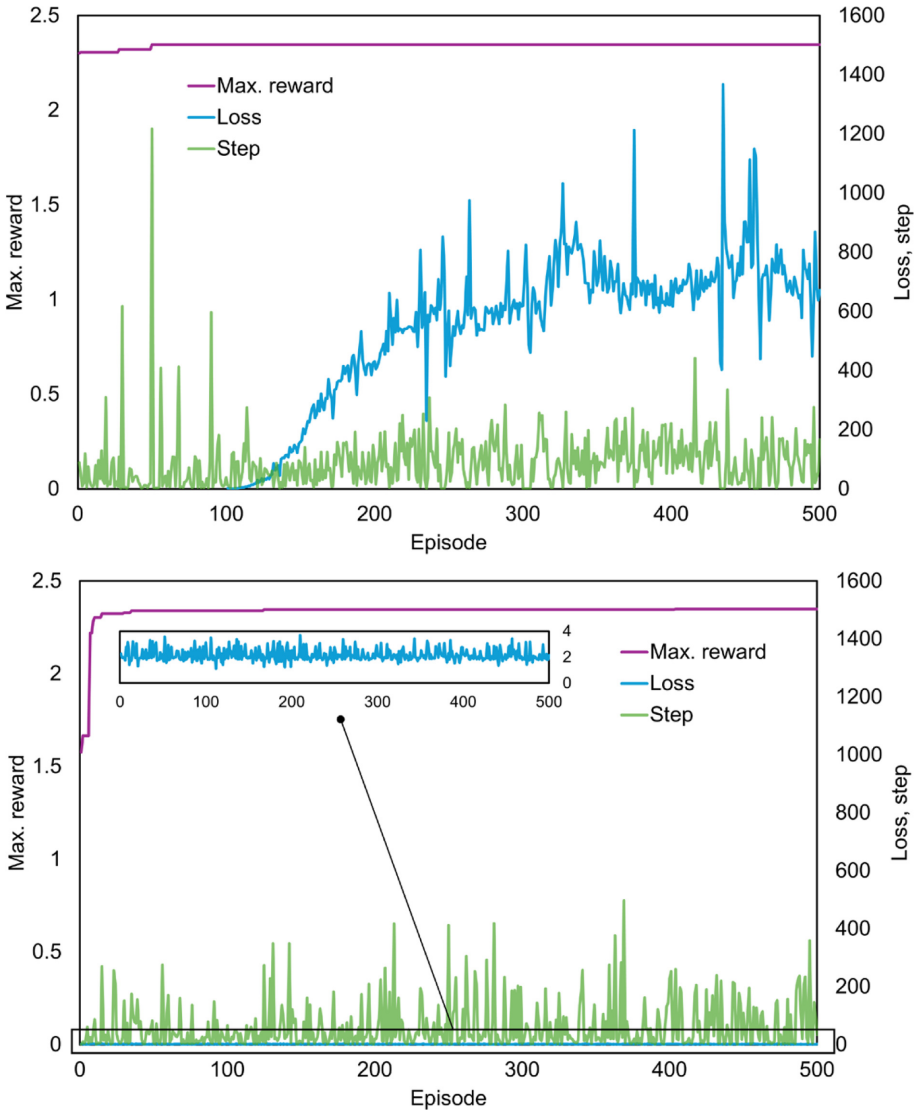

Fig. 3. DQN의 학습곡선(상) 및 A2C의 학습곡선(하).

Table 11. Optimized results of parametric study

DQN		SEC (kWh/kg)	Maximum reward	Episode	MTA (℃)
Base case	(0.001, 64)	0.306	2.35	49	2.04
Learning rate	0.0005	0.300	2.35	224	2.03
	0.005	-	-	-	-
Dimension of hidden layer	32	0.294	2.35	209	2.01
	128	0.306	2.35	49	2.04
M value in reward function	0.1	0.306	2.07	168	2.19
	1.0	0.309	2.69	454	2.07
Best case	(0.0005, 32)	0.293	2.35	157	2.10
A2C		SEC (kWh/kg)	Maximum reward	Episode	MTA (℃)
Base case	(0.001, 64)	0.305	2.35	403	2.12
Learning rate	0.0005	0.316	2.34	186	2.04
	0.005	0.310	2.34	416	2.02
Dimension of hidden layer	32	0.334	2.33	475	2.28
	128	0.324	2.33	187	2.02
M value in reward function	0.1	0.348	2.07	204	2.22
	1.0	0.346	2.65	149	2.02

SEC 값이 가지는 영향을 비교적 낮게 가지도록 하며 1.0으로 둘 경우 그 영향을 크게 가지도록 함을 의미한다. 따라서 유사한 SEC 값을 가지더라도 M의 값에 따라 보상함수의 값이 크게 변화하며, Table 11에서 볼 수 있듯이 M이 0.1일 경우 보상은 2.07, 1.0일 경우 가장 큰 값인 2.69를 가지는 것으로 확인되었다. 하지만 각 경우에서 최적화된 SEC 값은 결과적으로 큰 차이를 보이지 않았고, base case와 M이 0.1일 때 가장 낮은 0.306 kWh/kg의 SEC 값을 가졌다. 최적화된 공정에서 제약사항 중 하나인 MTA의 값을 확인한 결과, 0.1일 때는 보상함수의 설정으로 인해 SEC값이 낮은 편임에도 불구하고 MTA 값이 2.19 ℃로 비교적 높은 값을 보였고, 결과적으로 낮은 보상을 가졌다. 그와 반대로 M을 1.0으로 두었을 때는 세 경우 중 가장 높은 SEC 값을 가졌지만 MTA의 최적화된 값이 제약사항 값으로 설정한 2 ℃와 큰 차이를 가지지 않았고, 보상함수의 값은 최대로 나타났다. 앞선 결과로부터 보상함수의 설정이 제약사항과 목표함수를 만족하며 최적의 결과를 얻기 위한 중요한 요소가 된다는 점을 알 수 있었다.

DQN 알고리즘의 경우 parametric study 과정에서 더 나은 성능을 보이는 하이퍼파라미터를 찾았기 때문에 추가적으로 학습율은 0.0005, 은닉층의 차원은 32로 두고 학습을 진행하였다. 이때 보상함수의 M 값은 그대로 0.5로 두었고, 그 결과 SEC 값이 0.293 kWh/kg 으로 가장 낮은 값을 가졌다. 강화학습에서 학습율을 낮은 값으로 설정할수록 학습 속도는 느려지지만 학습이 안정적으로 진행될 수 있다. 또한 은닉층의 차원 수가 낮을수록 신경망의 복잡도가 줄어들고 학습 속도가 빨라진다. 따라서 두 하이퍼파라미터의 적절한 조합을 통해 base case보다 나은 최적화 성능을 가지도록 조정이 가능함을 확인하였다.

A2C 알고리즘의 parametric study 중 학습율을 0.0005와 0.005로 조정한 결과 0.0005의 경우 비교적 이른 186번 째 에피소드에서, base case와 0.005의 경우 각각 403, 416번째 에피소드에서 최대 보상을 얻었다. 이는 학습률을 낮은 값으로 두고 최적화를 진행할 때 모델의 업데이트가 느리게 진행되면서 좁은 영역에서의 탐색이 이루어진 결과로 보인다. 최적화 결과는 Table 11에 나타나 있듯이 base case의 SEC 값이 0.305 kWh/kg로 가장 낮게 나타났다. 학습률이 0.005일 때 최적화된 MTA 값은 2.02 ℃로 제약사항 값인 2 ℃에 가까운 값을 보였으나, COMP의 입구 증기분율이 0.998로 제약사항인 1보다 낮은 값을 가졌으며, 이와 더불어 SEC 값은 0.310 kWh/kg으로 base case보다 높은 값을 가지는 결과를 보였다. A2C 알고리즘의 정책신경망과 가치신경망의 은닉층 차원을 32와 128로 바꿔 parametric study를 진행한 결과, Table 11에서 볼 수 있듯이 두 경우 각각 0.334, 0.324 kWh/kg으로 높은 SEC 값을 보였다. 특히 32차원으로 두고 학습을 진행한 경우 최적화된 MTA 값 또한 2.28 ℃로 제약사항 값과 비교적 큰 차이를 가졌다.

마지막으로 A2C 알고리즘의 M값을 0.1과 1.0으로 조정하여 학습을 진행한 결과, DQN과 마찬가지로 각각 0.348, 0.346 kWh/kg의 높은 SEC 값을 보였다. 또한 M이 0.1일 때, DQN과 마찬가지로 최적화된 MTA값이 2.22 ℃로 다른 조건에 비해 제약사항과 큰 차이를 보였다.

결과적으로 A2C 알고리즘은 base case의 조건대로 학습률은 0.001, 은닉층의 차원은 64, M은 0.5로 설정했을 때 가장 나은 최적화 성능을 보이는 것을 확인하였다. 두 알고리즘을 비교할 경우, DQN의 best case가 0.293 kWh/kg으로 가장 낮은 SEC 값을 가졌으며 앞선 parametric study로부터 적절한 하이퍼파라미터의 조합으로 최적화 성능을 크게 개선할 수 있음을 알 수 있었다.

결과적으로 아직 DQN이나 A2C 알고리즘이 기존의 최적화 알고리즘인 유전알고리즘보다 더 좋은 성능을 내지는 못했다. 다만, 이는 연속적인 도메인의 결정 변수들을 비연속적인 행동으로 결정하는 데서 오는 한계점으로 생각된다.

4. 결 론

본 연구에서는 강화학습 방법론 중 대표적인 DQN과 A2C 알고리즘을 통해 천연가스 액화 공정 중 단일혼합냉매 액화공정을 최적화하였다. 또한, 베이스라인 설정을 위하여 기존 최적화 알고리즘인 유전알고리즘을 통해 최적화하여 결과를 비교했으며 두 강화학습 알고리즘의 하이퍼파라미터 중 학습률과 은닉층의 차원 및 보상함수에 대한 parametric study를 진행하였다.

두 가지 강화학습 알고리즘을 비교해보면, 최적화 측면에서는 DQN이 A2C보다 더 뛰어났고 에피소드 당 step이 안정적으로 유지되는 측면이 있었다. 또한, 에피소드당 평균 step 수가 A2C에 비해 높아 한 에피소드에서 좀 더 많은 영역을 탐색할 수 있는 것으로 보인다. 다만, A2C의 경우에는 학습 시간이 DQN에 비해 짧았으며, 손실도 안정적으로 감소하였다. 결론적으로 DQN과 A2C 중 DQN이 더 나은 성능을 보였으며, 이는 경험 리플레이의 장점이 잘 발현된 것으로 보인다. Parametric study의 경우에는 학습률과 은닉층의 차원, 보상함수에서의 M값 모두 결과에 주요한 영향을 미쳤다. 다만, DQN 및 A2C 각각의 알고리즘에 맞는 하이퍼파라미터 조합이 모두 달랐으며, 일관적인 경향이 발견되지는 않았다.

하지만 DQN이나 A2C 모두 유전알고리즘에 비해 최적화 측면에서 좋은 성능을 얻지는 못하였다. 그 이유는 DQN이나 A2C 알고리즘 모두 연속적인 결정 변수들을 비연속적인 행동으로 조정함으로 인해 제약사항 중 하나인 MTA 값을 세밀하게 조절하지 못했기 때문으로 보인다. 따라서, 강화학습 알고리즘에서 행동을 통해 결정 변수들을 연속적으로 결정할 수 있는 방법에 대한 추가적인 연구가 필요한 것으로 보인다.

앞서 제시한 행동 영역에 대한 개선점 외에도 하이퍼파라미터 튜닝, 보상함수 설정 등 강화학습 알고리즘의 성능을 높일 수 있는 경로가 아직 다양하게 존재한다. 본 연구는 기존의 방법론과 비교대조를 통해 강화학습 알고리즘의 공정 최적화 방법론으로서 발전 방향을 제시했다는 점에서 의의가 있으며, 이후 실제 화학 공장에서의 강화학습 활용에 기여할 수 있을 것이다.

감 사

이 연구는 산업통상자원부 및 산업기술평가관리원의 사용온도(영하 40도에서 영상 65도씨)에서 사용압력 70MPa급의 수소공급용 고내구성 호스 어셈블리 국산화 개발(4/4) 연구비 지원에 의해 수행된 연구입니다(과제번호: 20017444).

Reference

1. Looney, Energy Outlook 2020 Edition. BP, London, UK, 2020.
2. He, T., Chong, Z. R., Zheng, J., Ju, Y. and Linga, P., "LNG Cold Energy Utilization: Prospects and Challenges," *Energy*, **170**, 557-568(2019).

3. He, T. and Ju, Y., "Optimal Synthesis of Expansion Liquefaction Cycle for Distributed-Scale LNG (Liquefied Natural Gas) Plant," *Energy*, **88**, 268-280(2015).
4. Aspelund, A., Gundersen, T., Myklebust, J., Nowak, M. P. and Tomasgard, A., "An Optimization-Simulation Model for a Simple LNG Process," *Computers & Chemical Engineering*, **34**(10), 1606-1617(2010).
5. Silver, D., Huang, A., Maddison., C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, L., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchnrenner, N., Sutskever, I., Lilicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, **529**, 484-489(2016).
6. https://www.yokogawa.com/news/press-releases/2022/2022-03-22/, accessed: 26. Oct. 24.
7. Lee, S. H., Lim, D.-H. and Park, K., "Optimization and Economic Analysis for Small-Scale Movable LNG Liquefaction Process with Leakage Considerations," *Appl. Sci.*, **10**(15), 5391(2020).
8. Gao, Q. and Schweidtmann, A. M., "Deep Reinforcement Learning for Process Design: Review and Perspective," *Current Opinion in Chemical Engineering*, **44**, 101012(2024).
9. Kim, S., Jang, M.-G. and Kim, J.-K., "Process Design and Optimization of Single Mixed-Refrigerant Processes with the Application of Deep Reinforcement Learning," *Applied Thermal Engineering*, **223**, 120038(2023).
10. Seidenberg, J. R., Khan, A. A. and Lapkin, A. A., "Boosting Autonomous Process Design and Intensification with Formalized Domain Knowledge," *Computers and Chemical Engineering*, **169**, 108097(2023).
11. Stops, L., Leenhouts, R., Gao, Q. and Schweidtmann, A. M., "Flowsheet Generation through Hierarchical Reinforcement Learning and Graph Neural Networks," *AlChe J*, **69**(1), 17938(2023).
12. Chen, J. and Wang, F., "Cost Reduction of CO_2 Capture Processes Using Reinforcement Learning Based Iterative Design: A Pilot-Scale Absorption–Stripping System," *Separation and Purification Technology*, **122**, 149-158(2014).
13. Roh, E. J., Lee, H., Park, S., Kim, J., Kim, K. and Kim, S., "Directional Autonomous Torpedo Maneuver Control Using Reinforcement Learning," *The Journal of Korean Institute of Communications and Information Sciences*, **49**(5), 752-761(2024).
14. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D., "Human-Level Control through Deep Reinforcement Learning," *Nature*, **518**(7540), 529-533(2015).
15. Bao, J., Lin, Y., Zhang, R., Zhang, N. and He, G., "Effects of Stage Number of Condensing Process on the Power Generation Systems for LNG Cold Energy Recovery," *Applied Thermal Engineering*, **126**, 566-582(2017).
16. https://pygad.readthedocs.io/en/latest/, accessed: 26. Oct. 24.

Authors

Jieun Lee: Researcher, Department of Chemical & Biological Engineering, Sookmyung Women's University, Seoul, 04310, Korea; jieundk22@sm.ac.kr

Kyungtae Park: Associate, Department of Chemical & Biological Engineering, Sookmyung Women's University, Seoul, 04310, Korea; ktpark@sm.ac.kr