JESK J Ergon Soc Korea 2022; 41(2): 89-99
https://dx.doi.org/10.5143/JESK.2022.41.2.89
https://jesk.or.kr eISSN:2093-8462

Categorization of Safety Culture using a Reinforcement Learning Algorithm

Dhong Ha Lee

Department of Industrial Engineering, The University of Suwon, Suwon, 18323

강화학습알고리즘을 이용한 안전문화평가

이 동 하

수원대학교 산업공학과

Corresponding Author
Dhong Ha Lee
Department of Industrial Engineering,
The University of Suwon, Suwon, 18323
Email : dhonghalee@gmail.com

Received : February 03, 2022
Revised : February 13, 2022
Accepted : March 25, 2022

Objective: The aim of this study is to investigate whether a reinforcement learning (RL) algorithm is effective to improve the accuracy of the safety culture categorization which the author's previous study tried to solve with a convolutional neural net classifier.

Background: The RL algorithm using the optimal Bellman equation is nowadays popularly applied to many markov decision process situations.

Method: An asynchronous advantage actor critic (A3C) neural net was applied to learn the safety culture survey data collected from a nuclear power industry for safety culture level categorization.

Results: The RL algorithm applied to the randomly selected validation data resulted in 97.63% accuracy compared to 96.2% which showed in the previous study.

Conclusion: The safety culture level categorization using the A3C neural net is more stable and accurate than the previous convolution neural net classifier.

Application: The safety culture level classifier with the RL neural net learning massive survey data might be useful in place of expert interviewers for safety culture evaluation.

Keywords: Artificial intelligence, Convolution neural network, Reinforcement learning, Asynchronous advantage actor critic, A3C, Safety culture, Survey data

1. Introduction

빅데이터를 이용한 인간행동예측분야(Lee, 2016)에서 인공지능의 활약은 최근에 괄목할만한 성과를 보여주고 있다. 예를 들어 조직의 안전문화를 평가하는데 합성곱신경망(convolution neural network)을 학습시킨 분류기(classifier)를 이용하여 실무에 적용할 만한 수준의 정밀도를 달성할 수 있었다(Lee, 2019).

Lee (2019) 연구는 조직의 안전문화를 묻는 총 48개의 문항으로 설문지를 구성하였다.

Copyright@2022 by Ergonomics Society of Korea. All right reserved.

설문의 내용은 안전기법적용, 안전규칙/절차적용, 실수보고태도, 안전에 대한 헌신, 스트레스관리, 사고원인규명태도, 교육 및 훈련태도, 의사소통의 효율성 및 안전관리시스템에 대한 태도 등 9개 분야에 관한 것이다.

응답형식으로 안전문화설문에 대한 동의 정도를 5점 리커트 척도로 표현하였다. 문항 중에는 안전문화에 관한 긍정적 설문과 부정적 설문이 혼재되어 있는데 이 중 부정적 설문은 역방향척도를 사용하여 전 문항에서 5점이 안전문화에 대한 강한 긍정, 1점이 강한 부정을 표현하도록 하였다. 응답데이터는 신경망에 입력할 때 0~1사이의 값을 가지도록 정규화하였다.

안전문화설문응답을 인공신경망클래스분류기(neural net classifier)가 학습할 때 레이블데이터로서 안전문화수준의 높고 낮음을 분류하는 클래스분류데이터가 필요하다. Lee (2019) 연구에서는 안전문화수준을 "안전" 및 "개선요망"의 2 클래스(class)로 구분하였다. 분류기준은 전체응답평균이다. 각 응답자의 응답평균이 분류기준 이상이면 안전문화수준을 "안전" 클래스로, 그 미만이면 "개선요망" 클래스로 레이블링하였다.

인공신경망을 학습시키는 방법 중 강화학습(reinforcement learning; RL)은 시행착오행동을 통해서 받는 보상 중 좋은 보상을 받은 행동을 선별하고 선별된 좋은 행동을 학습하여 학습효율을 강화하는 방법이다. 최근 알파고(Wang et al., 2016)에 활용되어 주목을 받고 있다. 강화학습은 게임, 로보틱스, 주식예측 등 다양한 분야로 활용도를 확장하고 있다.

본 연구는 Lee (2019)의 합성곱신경망모델을 강화학습으로 확장하여 안전문화 분류정밀도를 더 향상시킬 수 있는지 확인하고자 한다.

2. Assumptions of Reinforcement Learning based on Markov Decision Process

강화학습은 학습시뮬레이터 상의 의사결정자인 에이전트(agent)가 현재의 상태(state)를 인식하여, 선택 가능한 액션(action)들 중 보상(reward)을 최대화하는 액션을 선택하도록 신경망을 학습시킨다. 그런데 여기서의 보상은 매 스텝마다 같은 유형의 보상이 아니고 다가오는 미래시점에서의 보상까지 포함한다. 보상의 시간적(temporal dynamics) 개념이 추가되어 일반 클래스분류기신경망으로 안전문화를 평가하는 것과 주요한 차이점을 보일 수 있다.

강화학습알고리즘은 마코프결정과정(Markov Decision Process; MDP)에 기반을 두고 개발되었다. 따라서 어떤 문제를 강화학습으로 해결하기 위해서는 문제의 모델링과정에서 다음 사항을 만족하는 것으로 가정해야 한다(Juliani, 2016).

1. 어떤 시점에 에이전트는 환경 중 상태집합 S의 원소에 해당하는 어떤 상태 s에 존재한다.

2. 에이전트는 환경으로부터 해당 상태 s에 관한 정보를 관찰하고 액션집합 A의 원소에 해당하는 어떤 액션 a를 취할 수 있다.

3. 액션을 취한 후 다음시점에 에이전트는 확률적으로 새로운 상태 s'으로 전이(transfer)하게 된다.

4. 새로운 상태 s'으로 전이한 후 에이전트는 상태전이에 대응하는 보상 r을 받는다.

5. 기존상태 s에서 새로운 상태 s'으로 전이하는 확률($P_a(s, s')$)은 에이전트의 액션에 영향을 받는다.

6. 다음상태 s'은 현재상태 s와 에이전트의 액션 a에만 영향을 받으며 현재 이전 시점의 모든 상태와는 확률적으로 독립적이라고 가정한다. 이 가정을 만족하는 전이를 마코프속성(Markov property)을 만족하는 상태전이라고 한다.

7. 미래에 받을 보상은 할인률(γ)을 사용하여 현재가로 보정한다. 할인률은 일반적으로 1보다 작으므로 동일한 가치가 미래에는 현재보다 적은 가치를 가지게 된다.

8. 위의 1-4의 과정은 정해진 조건에 따라 과정의 반복을 종결하는 상태정보 bool 변수(주로 done이라고 칭함)가 True가 될 때까지

계속된다. 이 MDP 시뮬레이션에서 done이 True가 될 때까지의 반복과정을 에피소드(episode)라고 한다.

위의 가정을 요약하면 전형적인 MDP는 주어진 환경에서 상태정보 s와 보상 r을 제공하고 상태와 보상은 에이전트의 액션결정(a)에 영향을 준다. 상태정보를 고려한 에이전트의 액션은 다시 환경의 다음시점의 상태 s'에 영향을 줄 수 있다(Figure 1).

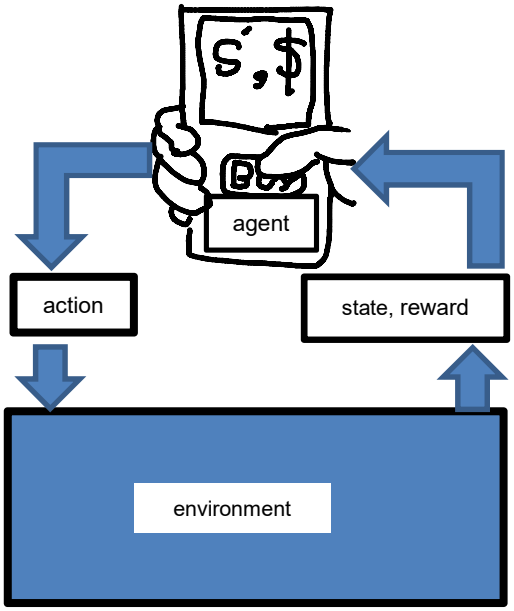

Figure 1. Markov decision process model

3. Modelling of Safety Culture Classification Problem

안전문화수준분류를 강화학습으로 해결하기 위해 MDP의 가정을 만족하도록 다음과 같이 문제를 구성하였다.

3.1 State

에이전트가 환경으로부터 매 시점마다 관찰하는 상태는 48개 문항에 대한 한 응답자의 응답데이터이다. 각 문항에 대한 가능한 응답이 5개이므로 상태집합은 $5^{48}\approx3.6*10^{33}$개의 원소를 가진다.

3.2 Action

안전문화수준범주의 수는 '안전'과 '개선요망'의 2개이므로 상태정보를 관찰한 에이전트가 취할 수 있는 액션은 2가지 판정 중 하나이다. 따라서 액션집합은 2개의 원소를 가진다.

3.3 Reward

에이전트가 바른 판정은 한 경우에는 1, 바르지 않은 경우에 0의 보상을 준다.

3.4 State transfer

안전문화응답데이터는 훈련데이터셋으로부터 에이전트에게 무작위로 복원추출한 표본(sampling with replacement)으로 제공되므로 현재상태와 다음상태는 거의 상호 독립적이다. 이는 전형적 MDP의 가정과 다르지만 상호독립을 의존도가 매우 낮은 상호의존의 한 형태로 볼 수 있어서 강화학습을 적용하는 데는 문제가 없었다.

3.5 Transfer probability

상태전이확률을 구하기 위해서는 상태전이에 관한 많은 관찰경험이 있어야 한다. 수많이 반복되는 MDP 시뮬레이션을 통해(s, a, r, s') 경험치를 축적하고 이 경험치 데이터집단으로부터 상태전이확률을 유추하였다.

3.6 Markov property

안전문화응답데이터가 매번 훈련데이터셋으로부터 무작위로 복원추출하여 제공되므로 에이전트의 안전문화판정액션은 다음상태 발생(generation)과 거의 상호독립적이다. 이 또한 전형적 MDP의 가정과 다르지만 상호독립을 의존도가 매우 낮은 상호의존의 한 형태로 볼 수 있으므로 강화학습을 적용하는 데는 문제가 없었다.

3.7 Discount rate

할인률 0.99를 사용하여 같은 가치의 현재의 보상이 미래의 보상보다 크도록 하였다.

3.8 Policy

강화학습은 장기간에 걸친 의사결정을 통해 주어진 환경에서 확률적으로 가장 많은 보상이 주어지는 액션을 선택할 수 있게 하는 상태-액션간의 매핑함수(mapping function)를 구하는 것이 목표다. 상태 s에서 의사결정자가 취할 행동 a를 지정하는 매핑함수를 정책(policy; π)이라고 한다. MDP에서 최적의 정책은 그 정책에 따라 액션을 취했을 때 장기간에 걸쳐 최선의 보상을 보장해야 한다. 최선의 보상은 벨만방정식(Bellman equation)의 최적 Q 함수로 표현된다(Lee et al., 2017).

$$\mathrm{Q(s,a)^* = r + \gamma\,[max_{\pi}\{Q(s',a')\}]} \quad (1)$$

식(1)에 의하면 현재시점에 주어진 상태와 액션에 대해 기대되는 최선의 보상은 현재시점의 즉각적인 보상에 다음시점에서 취할 수 있는 미래액션에서 기대할 수 있는 최선의 보상을 더한 값과 같다.

정책 중 가장 단순한 형태는 액션집합에서 무작위로 액션을 선택하는 방법이다. 정책탐색을 하는 시뮬레이션의 초기에 사용하였다.

그리디정책(greedy policy)은 Q 값이 최대가 되는 행동의 인덱스를 argmax 함수로 선택하는 방법이다(식(2)).

$$\pi(s) = argmax_a Q(s,a) \quad (2)$$

강화학습에서 가장 인기 있는 정책은 Q 값에서 각 행동의 확률분포를 softmax 함수로 도출하여 얻은 액션확률값분포 $\pi(a|s)$에 따라 액션을 선택하는 방법이다.

확률분포에 따라 액션을 선택하면 신경망파라미터의 변동에 따른 액션의 변동이 argmax 함수로 선택하는 방법보다 부드럽게 이어져 신경망이 안정적으로 학습되게 하는 장점이 있다. 또한 하나의 정해진 방식으로 선택하는 것보다 다양한 상황을 경험하게 함으로써 학습을 윤택하게 하여 최적정책을 보다 수월하게 찾을 수 있다. 확률분포에 따라 액션을 선택할 경우 최적화방법으로 정책그레디

언트(policy gradient)법을 적용할 수 있다. 정책그레디언트(∇J)는 식(3)과 같이 정의된다(Sutton et al., 2000).

$$\nabla J \approx E[Q(s,a)\nabla \log \pi(a|s)] \quad (3)$$

정책그레디언트는 주어진 상태에서의 액션가치 $Q(s,a)$와 선택한 액션의 로그확률미분 $\nabla \log \pi(a|s)$에 비례한다. 정책그레디언트는 누적보상을 최대화하기 위해 신경망파라미터가 움직여야 할 정책방향을 가리킨다. 따라서 최적화 과정을 진행함에 따라 좋은 보상이 주어지는 액션을 선택할 확률이 커지고 나쁜 보상이 주어지는 액션을 선택할 확률은 작아지게 된다. 본 연구에서는 액션을 선택하는 정책으로 정책그레디언트를 사용하였다.

4. Implementation of Reinforcement Learning Environment

환경구현에 사용한 딥러닝프레임워크는 페이스북의 AI 팀이 개발한 파이토치(pytorch)이다. 인공지능 코딩이 대부분 파이썬으로 이루어지고 있는데 파이토치는 파이썬언어에 기반을 두고 있어서 구글이 개발한 프레임워크 텐서플로에 비해 코딩이 직관적이고 수월한 장점이 있다.

4.1 Simulation environment

안전문화수준분류문제를 강화학습으로 처리하기 위해서는 에이전트에게 현재상태정보를 전달하고 에이전트의 액션에 대응하는 보상과 다음상태정보를 반복적으로 전달하는 시뮬레이션 환경이 필요하다. 본 연구는 open ai의 gym (Lapan, 2020)을 오버라이딩하여 이 환경을 SafetyCultureEnv 클래스로 구현하였다(Table 1).

Table 1. SafetyCultureEnv class

Method	Function
Constructor ()	Define observation_space Define action_space Get a new observation (s)
Reset ()	Get a new observation (s)
Step (action)	Observe current state (s) Process the action (a) to judge whether current state is 'safe' or 'need to improve' If judgement is correct, reward=1, done=false Else reward=0, done=true Get a next state observation (s') Return (s', r, done, {})
Get_observation ()	From train dataset choose a random sample Get a new observation from the sample

4.2 A3C neural net

에이전트가 사용할 정책, 즉 상태-액션간의 매핑함수로서 A3C (asynchronous Advantage Actor Critic) 신경망을 이용하였다. A3C 신경망은 계산시 컴퓨터의 다중프로세스(multi process)를 사용한다. 또한 정책을 개선하는 액터(actor)와 액터의 정책을 상태가치로 평가하

여 액터에게 피드백하는 크리틱(critic)으로 구성한다. A3C의 구조는 DQN (deep Q network)과 듀엘링 DQN (deuling DQN)의 구조를 기초로 이용하였다.

DQN은 상태-액션매핑함수 Q를 구하기 위해 다계층합성곱신경망을 사용한다. 상태전이확률을 추정하는 방법으로 상태전이경험치(s, a, r, s')의 최신자료를 버퍼에 저장하고 이 버퍼에서 학습에 필요한 자료를 꺼내어 사용한다. 경우에 따라서 제2신경망을 생성하여 학습을 안정적으로 진행하도록 돕기도 하지만 본 연구에서는 제2신경망 생성 대신 하나의 신경망을 생성하고 다중프로세스를 사용하여 각 프로세스에서 신경망의 파라미터를 공유하도록 하였다.

듀엘링 DQN은 DQN의 합성곱연산결과를 상태가치함수 V(s)와 주어진 상태에서의 액션가치를 나타내는 어드밴티지(advantage)함수 A(s, a)로 분리하여 표현한 신경망이다. 어드밴티지함수는 상태본연의 가치에 외에 액션의 기여에 의해 추가로 받은 보상을 의미한다. 듀엘링 DQN의 상태-액션매핑함수 $\mathrm{Q(s,a)}$를 식으로 표현하면 식(4)와 같다. 식(5)는 어드밴티지함수의 표현식이다. 계산 관점에서 보면 듀엘링 DQN은 DQN에 비해 학습속도를 증진시키는 효과가 있다.

$$\mathrm{Q(s,a)} = \mathrm{V}(s) + \mathrm{A(s,a)} \quad (4)$$

$$A(s,a) = Q(s,a) - V(s) \quad (5)$$

DQN이나 듀엘링 DQN은 컴퓨터의 단일 프로세스를 사용하여 연산한다. 따라서 단일 에이전트가 모든 처리를 담당하기 때문에 계산속도에서 제한이 있다.

A3C는 다중프로세스에 각각의 에이전트를 생성하여 각 에이전트가 서로 다른 경험데이터버퍼를 가지고 독자적(asynchronous)으로 신경망을 훈련시킨다. 시스템 전체로 보면 단일프로세스를 사용하는 것에 비해 다양한 경험소스로부터 신경망을 학습시키는 효과를 가진다. 학습속도도 단일프로세스를 사용하는 것보다 훨씬 빠르다.

A3C의 구성요소에서 액터는 정책그레디언트를 사용하여 정책을 향상시키고 정책의 평가는 크리틱이 담당한다. 크리틱이 정책평가에 사용하는 기준은 식(5)와 같은 어드밴티지함수의 한 형태인데 본 연구와 같이 상태집합원소 수가 많은 경우에는 Q 함수 대신 할인된 다음상태 s_{t+1}의 가치 $r_{t+1} + \gamma V(s_{t+1})$를 근사치로 사용하여 전체적으로 식(6)의 δ 함수로 표현된다.

$$\delta = \{r_{t+1} + \gamma V(s_{t+1})\} - V(s_t) \quad (6)$$

식(6)에서 $V(s_t)$는 베이스라인(baseline)으로 사용되었으며 어드밴티지함수의 변동이 커지는 것을 완화하는 역할을 한다.

신경망파라미터의 기울기를 구하기 위하여 역전파(back propagation) 시 사용하는 크리틱의 손실함수 J_{critic}은 어드밴티지의 평균제곱오차(mean squared error; MSE)함수이다(식(7)).

$$J_{critic} = MSE(\delta) \quad (7)$$

두 개의 액터손실함수 중 J_{actor1}는 에이전트가 실제로 수행한 액션로그확률을 어드밴티지와 곱한 값의 평균을 사용하였다. 액터의 크로스엔트로피(cross entropy) 손실함수라고 하는 이 식은 식(3)에서 Q 대신 어드밴티지 δ로 근사시킨 형태이다(식(8)).

$$J_{actor1} = -\mathrm{mean}[\delta \log \pi(\mathrm{a|s})] \quad (8)$$

에이전트가 액션을 선택할 때 특정액션에 대해 미리 편견(prejudice)을 가지지 못하도록 하는 장치가 필요하다. 액터의 또 다른 손실함수(J_{actor2})로서 액션확률의 마이너스엔트로피(minus entropy)를 추가하면 이 함수를 최소화하는 과정에서 액션확률이 어느 한쪽으로 치우치려는 경향을 막을 수 있다(식(9)).

$$J_{actor2} = -mean[-\Sigma_a \pi(a) log \pi(a)] \quad (9)$$

A3C 신경망을 갱신하기 위한 전체 손실함수 J는 크리틱과 액터 손실함수의 합(식(10))을 사용하였다.

$$J = J_{critic} + J_{actor1} + J_{actor2} \quad (10)$$

예비실험을 통해 정책확률(π)과 상태가치 $V(s)$를 구하기 위한 신경망으로서 전계층을 완전연결층(fully connected layer) 신경망으로 시도해 보았으나 입력데이터 처리에 합성곱층(Figure 2)을 사용하는 것이 모델의 예측정밀도를 더 높이는 것으로 확인되었다. 이 결과를 기초로 A3C 전체신경망은 듀엘링신경망 구조를 활용하여 Table 2와 같이 구성하였다.

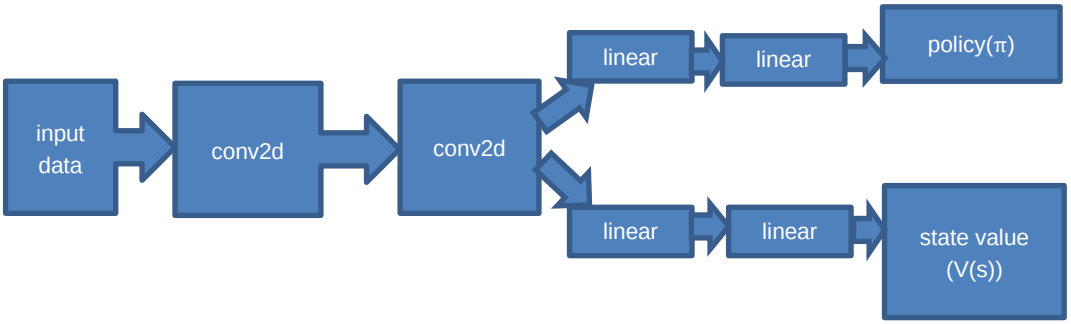

Figure 2. A3C net for policy and state value

Table 2. A3C neural net model arguments

Layer	Sublayer	Model arguments
Convolution	conv2d	in_channels=1 out_channels=8 kernel_size=4, stride=1 activation=relu
	conv2d	in_channels=8 out_channels=16 kernel_size=2, stride=1 activation=relu
Policy (π)	Linear	in_features=convolution output size out_features=512 activation=relu
	Linear	in_features=512 out_features=number of actions activation=softmax

Table 2. A3C neural net model arguments (Continued)

Layer	Sublayer	Model arguments
State value (V(s))	Linear	in_features=convolution output size out_features=512 activation=relu
	Linear	in_features=512 out_features=1

5. Training Safety Culture Data with A3C Net

5.1 Train dataset

Lee (2019)의 연구에서 사용되었던 1,045개의 안전문화설문응답 중 836개를 훈련용데이터셋으로, 209개를 검증용데이터셋으로 무작위로 분리하였다. 검증용데이터셋은 신경망의 훈련과정 중 예측정밀도를 추정하기 위해 사용하였다. 신경망에 입력되는 각 응답데이터는 각 데이터셋에서 무작위로 복원추출되어 공급되었다. 응답데이터 48개 특성(feature)은 8*8 행렬로 재형상(reshape) 처리하고 특성치가 들어가지 않은 빈 셀(cell)은 제로패딩(zero padding) 처리하였다. Table 2의 합성곱층에 입력하기 위한 입력데이터의 형상은 (-1, 1, 8, 8)로 주었다.

5.2 Procedure

강화학습은 아래와 같은 절차로 진행하였다:

1. MDP를 시뮬레이션할 수 있는 환경객체를 생성한다.

2. A3C 신경망 객체를 생성한다.

3. 신경망파라미터를 갱신할 옵티마이저(optimizer) 객체를 생성한다(갱신 방식=Adam, 학습률=1e-3).

4. 신경망의 분산학습을 진행할 차일드프로세스(child process) 객체를 생성하고 차일드프로세스가 수행할 업무를 지정한다. 본 연구에서는 4개의 차일드프로세스를 이용하였다. 각 차일드프로세스가 수행하는 업무는 4.1-4.10과 같다:

4.1 프로세스당 8개의 환경객체를 생성하여 프로세스당 8개의 MDP 시뮬레이션을 독립적으로 진행하였다. 이로써 학습에 필요한 경험치(s, a, r, s')를 다양한 독립적 소스로부터 구할 수 있었다.
4.2 경험치소스(experience source) 객체를 생성한다. 경험치소스객체는 MDP 시뮬레이션을 수행하여 경험치(s, a, r, s')를 하나씩 제공하는 제너레이터(generator) 역할을 한다. 본 연구에서 경험치소스객체는 강화학습도구 라이브러리 중 하나인 PyTorch Agent Net (PTAN)이 제공하는 것을 사용하였다(Lapan, 2020).
4.3 학습진행과정에서 신경망을 모니터링하기 위해 성능지표를 텐서보드(tensorboard)에 기록할 SummaryWriter 객체를 생성한다. SummaryWriter 객체는 매 반복마다 신경망학습성능지표를 기록하였다.
4.4 매 스텝마다 훈련데이터배치(batch) 저장소에 경험치를 누적하고 한 에피소드의 끝에 그 에피소드에서 받은 보상의 총합을 기록한다.
4.5 최근 100 에피소드의 평균보상총합이 100 이상이면 프로세스 작업을 종결하고 메인 프로세스로 돌아가도록 하였다. 보상총합이 100이라는 것은 에이전트가 한 에피소드 중 안전문화에 대한 판단을 연속 100회 올바른 판정을 한다는 의미이다.

4.6 훈련데이터배치에 64개의 경험치가 수집되면 이로부터 현상태배치, 실제 수행한 액션배치 및 할인된 다음상태가치배치를 재구성하고 Table 2의 신경망에 입력한다.
4.7 신경망의 출력으로부터 정책확률 π와 상태가치 V(s)를 구한다.
4.8 식(10)으로 표현되는 전체손실함수를 계산한다.
4.9 전체손실함수에 대한 역전파로 신경망의 기울기를 구한다. 기울기팽창(gradient explosion)을 방지하기 위한 클리핑(clipping) 처리를 한다(Lapan, 2020).
4.10 각 차일드프로세스에서 갱신한 신경망의 기울기를 메인프로세스로 전달한다.

5. 메인프로세스에서는 무한반복으로 다음 사항을 수행한다:

5.1 각 차일드프로세스에서 갱신한 파라미터를 가져와서 파라미터버퍼에 저장한다.
5.2 최근 100에피소드의 평균보상총합이 100을 넘으면 무한반복을 벗어나 학습을 종료한다.
5.3 각 프로세서에서 가져온 신경망기울기를 계산효율을 위해 두 개씩 묶어서 기울기클리핑한 후 옵티마이저를 호출하여 파라미터를 갱신한다.
5.4 파라미터버퍼를 지우고 5.1의 반복과정으로 돌아간다.

6. Results and Conclusions

런타임 환경으로 구글코랩프로(google colab pro)를 이용하였다. 구글코랩프로는 학습 가속을 위한 gpu 사용을 수월하게 제공하는 장점이 있다. 본 연구에서 4.833 mega (M) 프레임의 차일드프로세스 훈련진행과정 중 신경망의 학습속도는 초당 1200 프레임 내외의 속도로 진행되었다(Figure 3).

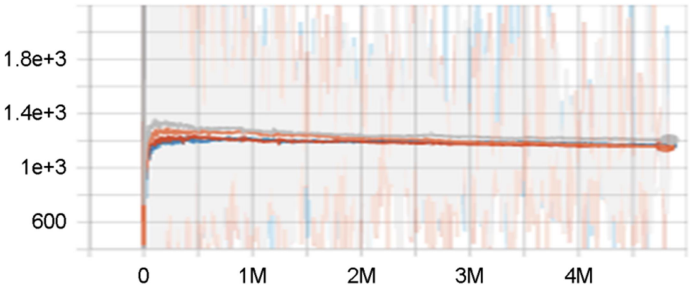

Figure 3. Neural net learning speed (fps)

에이전트가 실제로 수행한 액션확률과 이것의 이상적인 값이라 할 수 있는 어드밴티지값과의 차이를 나타내는 크로스엔트로피값은 하향추세인데 목표하는 방향으로 진행되었다(Figure 4).

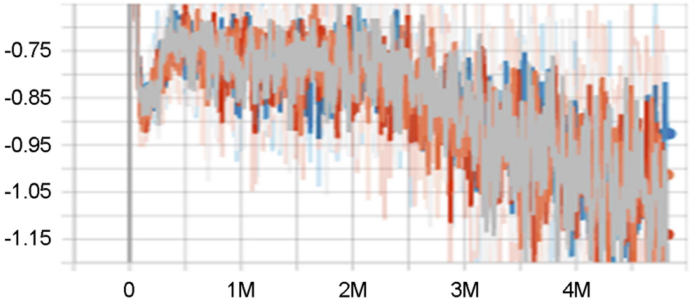

Figure 4. Trend of cross entropy

액터의 마이너스엔트로피는 액터가 학습초기부터 어떤 액션에 대한 편향을 갖지 못하게 손실함수로 처리하여 억제하지만 학습이 진행됨에 따라 최적정책을 찾아가므로 결과적으로 증가하는 추세로 관찰되었다(Figure 5).

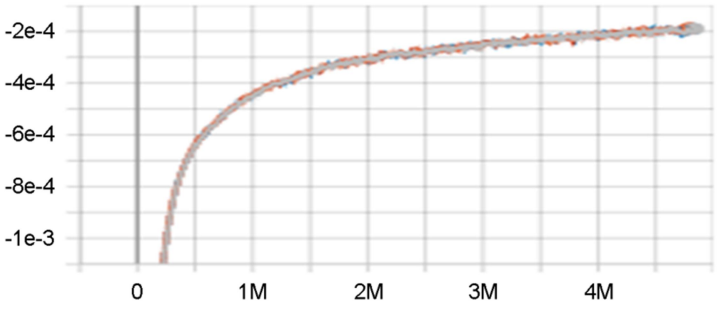

Figure 5. Trend of actor entropy

가장 최근의 100 에피소드에서 받은 보상평균은 목표지점인 100을 향하여 학습과정 중 내내 안정적으로 점근하였다(Figure 6).

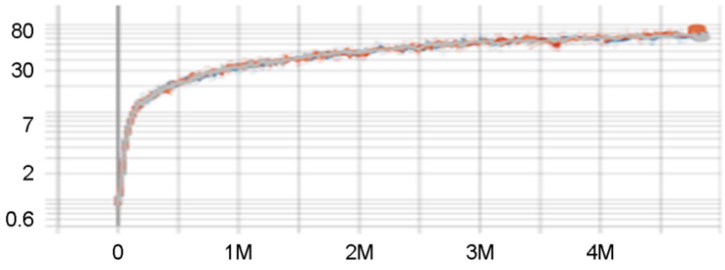

Figure 6. Trend of mean 100 episode reward

훈련용데이터셋에 대해 독립적으로 구성된 검증용데이터셋을 사용하여 메인프로세스가 신경망파라미터를 갱신하는 300k 스텝 중 매 10k 스텝마다 신경망의 예측정밀도를 측정하였다(Figure 7). 정밀도는 초기 70k 스텝 만에 97.33%에 급격히 도달하고 이후에는 최고치근방에서 완만하게 안정된 추세를 보였다. 측정치 중 최고치는 190k 스텝에서 98.67%이고 훈련종료시점의 측정치는 97.67%였다.

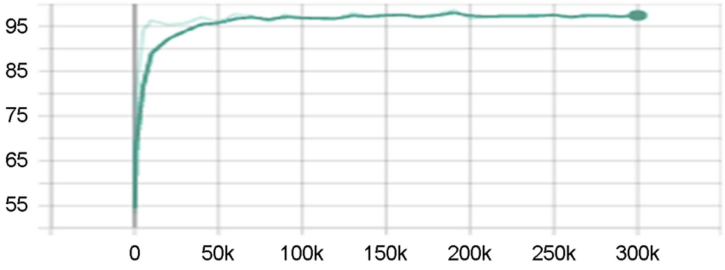

Figure 7. Trend of accuracy for validation dataset

이상의 시험결과로 Lee (2019)가 합성곱신경망을 클래스분류기로 사용하여 얻은 정밀도 96.2% 보다 본 연구의 강화학습으로 훈련한 신경망이 정밀도를 향상시킬 수 있음을 확인하였다.

본 연구에서는 시간과 비용의 제약으로 신경망훈련에 필요한 하이퍼파라미터(hyperparameter)-학습률 및 신경망구성 요소-를 최적화

하지 못하였다. 학습률의 경우 1e-3, 1e-4에 대해서만 시행하였다. 실험결과 양쪽의 성능지표는 유사했지만 후자의 경우 학습시간이 13시간 이상으로 전자(2시간 34분)에 비해 학습속도가 더딤을 확인하였다. 이외의 다른 하이퍼파라미터가 최적화되면 더 좋은 성능을 볼 수 있을 것으로 기대한다.

본 연구의 한계점은 강화학습에 사용한 훈련데이터셋의 규모가 작다는 점이다. 일부 원자력발전소 종사자로부터 수집된 1,045건의 설문응답자료는 원자력발전소 전체의 안전문화 수준을 일반화하기에는 턱없이 부족하다는 판단이다. 이를 해결할 수 있는 방법으로는 Lee (2019)의 연구에서 제안된 안전문화 설문평가방법이 주기적으로 전 원자력발전소로 확대실시 되어야 할 것이다. 일단 대규모의 설문자료가 축적되었다고 가정하면 본 연구에서 제안한 강화학습 알고리즘이 높은 정밀도를 보장하는 평가도구로 사용될 수 있을 것이다. 본 연구에서 제안한 강화학습 알고리즘이 안전문화수준평가 외에도 유사한 분류문제에서 충분한 자료를 바탕으로 유용한 범주분류기(category classifier)로 활용되기를 기대한다.

References

Juliani, A., Simple Reinforcement Learning with Tensorflow Part 0: Q-Learning with Tables and Neural Networks, https://medium.com/emergent-future/simple-reinforcement-learning-with-tensorflow-part-0-q-learning-with-tables-and-neural-networks-d195264329d0 (retrieved April 26, 2016).

Lapan, M., Deep Reinforcement Learning Hands-On: Apply modern RL methods to practical problems of chatbots, robotics, discrete optimization, web automation, and more, 2nd Edition, Packt Publishing, 2020.

Lee, D.H., Classification of safety culture level using a convolution neural network, *Journal of the Ergonomics Society of Korea*, 38(6), 435-443, 2019. doi:10.5143/JESK.2019.38.6.435

Lee, W.W., Yang, H.R., Kim, K.W., Lee, Y.M. and Lee, Y.R., Reinforcement Learning with Python and Keras, Wikibooks, Paju, 2017.

Lee, Y.H., Safety culture, a new challenge to human factors engineering for 21st century, *Journal of the Ergonomics Society of Korea*, 35(6), 473-492, 2016. doi:10.5143/JESK.2016.35.6.473

Sutton, R.S., McAllester, D.A., Singh, S.P. and Mansour, Y., Policy Gradient Methods for Reinforcement Learning with Function Approximation. In H. Larochelle (Ed), *Advances in Neural Information Processing Systems*, 1057-1063, 2000.

Wang, F.Y., Zhang, J.J., Zheng, X., Wang, X., Yuan, Y., Dai, X., Zhang, J. and Yang, L., Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond, *IEEE/CAA Journal of Automatica Sinica*, 3(2), 113-120, 2016.

Author listings

Dhong Ha Lee: dhonghalee@gmail.com
Highest degree: PhD, Department of Industrial Engineering, KAIST
Position title: Professor, Department of Industrial Engineering, Suwon University
Areas of interest: Application of artificial intelligence