Journal of Korea Robotics Society (2022) 17(3):264-272
https://doi.org/10.7746/jkros.2022.17.3.264 ISSN: 1975-6291 / eISSN: 2287-3961

표정 피드백을 이용한 딥강화학습 기반 협력로봇 개발

Deep Reinforcement Learning-Based Cooperative Robot Using Facial Feedback

전 해 인[1] · 강 정 훈[2] · 강 보 영†
Haein Jeon[1], Jeonghun Kang[2], Bo-Yeong Kang†

Abstract: Human-robot cooperative tasks are increasingly required in our daily life with the development of robotics and artificial intelligence technology. Interactive reinforcement learning strategies suggest that robots learn task by receiving feedback from an experienced human trainer during a training process. However, most of the previous studies on Interactive reinforcement learning have required an extra feedback input device such as a mouse or keyboard in addition to robot itself, and the scenario where a robot can interactively learn a task with human have been also limited to virtual environment. To solve these limitations, this paper studies training strategies of robot that learn table balancing tasks interactively using deep reinforcement learning with human's facial expression feedback. In the proposed system, the robot learns a cooperative table balancing task using Deep Q-Network (DQN), which is a deep reinforcement learning technique, with human facial emotion expression feedback. As a result of the experiment, the proposed system achieved a high optimal policy convergence rate of up to 83.3% in training and successful assumption rate of up to 91.6% in testing, showing improved performance compared to the model without human facial expression feedback.

Keywords: Interactive Reinforcement Learning, DQN, Cooperative Robot, Emotion Estimation, AI, NAO Robot

1. 서 론

인공지능과 로봇 기술의 발달로 머신러닝(machine learning) 기술 탑재 로봇을 일상 속에서 접할 수 있게 되었다. 그 예로는 공항 및 역 안내로봇[1], 카페 서빙 로봇[2], 조립 공정 협동로봇[3] 등이 있다. 일상 속 로봇은 다양한 환경에서 사람의 업무를 대신 또는 함께 수행하며, 이에 따라 사람과 협력하여 임무를 수행하는 협력로봇에 대한 연구가 이루어져 왔다.

머신러닝 기법은 크게 지도학습, 비지도학습 그리고 강화학습(Reinforcement Learning)으로 분류할 수 있다. 그 중 로봇 학습에 성공적으로 적용되고 있는 기술 중 하나인 강화학습[4,5]은 학습 에이전트인 로봇이 주어진 상태에서 특정 동작을 취했을 때 받는 보상을 통해 가장 적절한 동작을 학습하는 방법이다. 강화학습 시스템에서 보상은 일반적으로 상태 별 에이전트 동작에 따라 주어지며, 사람-에이전트 간 실시간 상호작용을 통해 보상이 주어질 경우 이를 인터액티브 강화학습(Interactive reinforce ment learning)이라고 한다. 인터액티브 강화학습 기술인 리워드 쉐이핑(Reward shaping)[6]은 사람인 트레이너가 에이전트의 동작에 긍정적 또는 부정적 피드백을 제공하여 보상 함수를 수정하는 기술이다. 그러나 선행된 리워드 쉐이핑 연구에서는 사람이 직접 피드백 값을 별도의 입력 장치를 통해 입력하거나, 적은 수의 정해진 피드백을 사용하여 피드백의 종류가 한정적이다. 일상 속 로봇 증가에 맞추어, 로봇 활용도 향상을 위해서는 다양한 환경에서 자연스러운 로봇 훈련이 가능하도록 다양한 피드백을 통한 학습 시스템의 필요성이 제기된다.

Received : May. 27. 2022; Revised : Jun. 29. 2022; Accepted : Jul. 6. 2022
※ This research was supported by Kyungpook National University Research Fund, 2021
1. PhD Student, Department of Artificial Intelligence, Kyungpook National University, Daegu, Korea (haeinjeon.knu@gmail.com)
2. Master Student, Department of Artificial Intelligence, Kyungpook National University, Daegu, Korea (jhkang.knu@gmail.com)
† Professor, Corresponding author: Department of Robot and Smart System Engineering, Kyungpook National University, Daegu, Korea (kby09@knu.ac.kr)

따라서 본 논문에서는 로봇 활용도 향상을 위한 표정 피드백을 이용한 인터액티브 딥강화학습 시스템을 제안한다. 제안된 시스템에서 로봇은 Deep Q-Network (DQN)[7]을 사용하여 사람과 협동이 필요한 테이블 균형맞춤과제[8]를 수행하며, 사람 표정 피드백에 기반하여 과제 수행 정책을 학습한다. 작업 수행 방법을 알고 있는 트레이너는 로봇 동작에 대해 얼굴 표정을 사용하여 실시간으로 긍정 및 부정의 감성 표정 피드백을 제공하며, 로봇은 표정 추정 모듈을 사용해 사람 표정 피드백을 해석한 후 강화학습에 반영한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 선행 연구들을 통해 인터액티브 강화학습의 기존 연구의 흐름과 한계점을 알아보고, 3장에서는 제안한 표정 피드백을 이용한 인터액티브 딥강화학습 시스템을 설명한다. 그리고 4장에서는 제안된 시스템에 따라 테이블 균형맞춤 학습 실행 결과를 설명하며, 학습 설정에 따른 시스템 성능의 차이를 비교한다. 마지막으로 5장에서는 결론과 향후 연구 방향을 제시한다.

2. 선행 연구

강화학습에서 학습 속도 및 성능 향상 기술 중 하나는 사람이 트레이너로서 에이전트를 지도하는 것이며, 대표적인 예로는 모방을 통한 학습[9], 시연을 통한 학습[10,11], 피드백을 제공하는 학습[12]이 있다.

그 중 피드백 제공 학습을 집중하여 살펴보면 사람이 마우스, 리모콘 등의 장치를 이용한 시스템 연구[6-12], 이진(binary) 피드백을 사용한 인터액티브 강화학습 알고리즘 연구[13-15], 사람 표정 피드백과 같은 사람-에이전트 간 자연스러운 상호방식에 초점을 맞춘 연구[16,17]가 있다. 이러한 연구들의 공통점은 로봇 또는 컴퓨터가 사람과 상호작용하는 리워드 쉐이핑 기술을 적용하여 강화학습의 학습 시간을 감소시켜 수렴 속도를 개선하고, 에이전트의 목표 동작 학습 성능을 향상시켰다는 점이다.

먼저, 마우스나 리모콘 등을 통해 직접 피드백 값을 입력해 강화학습을 진행한 연구가 있다. Thomaz et al.[6]은 요리 시뮬레이션 로봇 학습을 위한 인터액티브 Q-러닝 강화학습 플랫폼에서, 사람이 가상 로봇 동작의 보상으로 마우스 스크롤을 이용해 -1 에서 +1 사이의 수를 피드백으로 제공하여 로봇의 학습 효율을 향상시킬 수 있음을 밝혔다. Ullerstam and Mizukawa[12]의 연구에서 AIBO로봇은 사람으로부터 2가지 리모컨 반응 보상을 통해 노래 부르기 작업 등을 학습하였다.

그러나 이러한 인터액티브 강화학습 선행연구에서는 사람 피드백 제공에 마우스, 리모콘과 같은 로봇 이외의 입력 하드웨어가 요구되며 이는 사람-로봇간 자연스러운 상호작용이 이루어진다고 보기 어렵다.

긍정, 부정의 이진(binary) 피드백을 사용한 인터액티브 강화학습 알고리즘 개발에 초점을 맞춘 연구에는 TAMER[13], Advise[14], REPaIR[15]가 있다. Knox and Stone이 제안한 인터액티브 강화학습 알고리즘인 TAMER[13]는 에이전트가 사람으로부터 동작에 대해 긍정 및 부정의 2가지 평가 신호를 키보드로 입력 받아 사람 피드백 함수를 모델링하며, 이에 따라 좋은 피드백(good feedback)을 최대화하는 Tetris 등의 게임 동작을 학습하였다. Griffith et al.이 제안한 Advise 알고리즘[14]에서 사람은 에이전트에게 긍정 또는 부정의 두 가지 피드백을 제공하여 에이전트의 행동 선택 확률, 즉 정책(policy)을 수정하며, 에이전트는 사람 피드백 함수를 학습한다. 그 결과 기존의 강화학습 알고리즘보다 Pac-Man 등의 게임 과제에 더 나은 성능을 보였다. Faulkner et al.은 인터액티브 강화학습의 사람 제공 피드백 오류 보완을 위해 피드백의 정확함을 추정하는 REPaIR 알고리즘[15]을 제안하였다. 가상 로봇은 시뮬레이션 환경에서 상자에 공 넣기 과제를 수행했으며, REPaIR 알고리즘이 피드백을 사용하지 않는 Q-러닝보다 나은 성능을 보임을 증명하였다.

그러나 이와 같이 인터액티브 강화학습에 피드백 학습 알고리즘 개발에 초점을 맞춘 접근법들은 버튼이나 레버를 사용하여 사람이 긍정 또는 부정의 이진 피드백만을 입력하였으며, 이를 통한 사람 피드백 함수의 학습 자체에도 시간이 소요된다. 반면 본 연구에서 제안한 시스템에서는 표정 피드백을 연속적 수치로 제공하여 자연스러운 사람 반응을 실시간 피드백으로 제공함으로써 실질적인 사람-로봇 상호작용 시스템 구현 시 활용도가 높다.

사람의 표정과 같은 자연스러운 상호방식에 초점을 맞추어 인터액티브 강화학습을 진행한 연구에는 Veeriah et al[16], Arakawa et al.[17]의 연구가 있다. Veeriah et al[16]은 SARSA 강화학습에 카메라를 통한 긍정, 부정, 중립의 표정 피드백과 버튼 입력을 사용하여 에이전트에 그립 선택 작업을 학습시켰으며, 가상환경에서의 실험을 통해 에이전트가 피드백을 통해 더 빠른 속도로 그립 선택 작업을 수행할 수 있음을 밝혔다. Arakawa et al.[17]은 심층 신경망 기반의 표정 분류를 통해 DQN-TAMER 에이전트를 훈련시켰으며 시뮬레이션 환경 속 에이전트는 긍정, 부정의 2가지 표정 피드백을 사용해 가상 경로계획 작업인 Maze와 Taxi를 학습하였다.

그러나 이와 같은 사람 표정 사용 접근법에서는 주로 가상환경의 에이전트가 피드백을 제공하는 사람과 함께 작업을 수행하기보다는, 가상의 게임과 같은 비연속적 이미지 데이터를 사용해 작업 학습에 집중하는 경향을 보인다. 반면 본 논문에서는 실시간으로 사람과 테이블 균형맞춤 작업을 수행하는 협력로봇을 대상으로 사람-로봇간 상호작용이 이루어지는 강화학습 시스템을 디자인하였다.

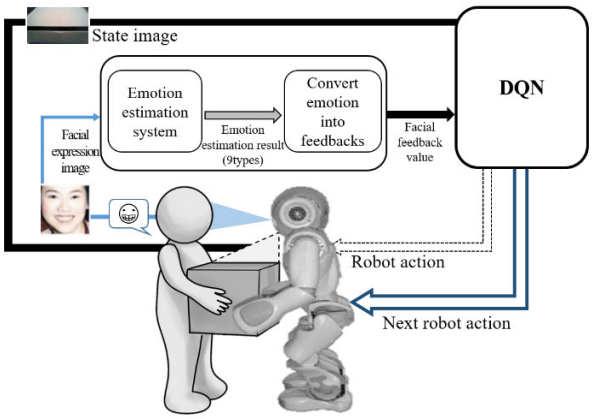
일련의 선행연구를 통해 피드백을 통한 인터액티브 강화학습은 학습 성능 향상에 도움을 주나, 별도의 피드백 입력 하드웨어를 요구하여 실시간 환경에서 사람과 자연스러운 상호작용이 이루어지기 어려웠다. 본 논문에서는 이러한 한계점 해소를 위해 협력 로봇이 사람 표정 피드백과 표정 추정 시스템을 사용하여 상호작용하는 학습을 하기 위한 인터액티브 강화학습 시스템을 제안하였다.

3. 제안한 표정 피드백을 이용한 인터액티브 딥강화학습 기반 협력로봇

본 절에서는 사람과 로봇 간에 성공적으로 테이블 균형을 맞추기 위해, 표정 피드백에 기반한 딥강화학습 프레임워크를 제안하며, 전체적인 작업도는 [Fig. 1]과 같다.

[Fig. 1]에서 먼저 로봇은 카메라로 테이블 상태 이미지를 촬영하여 DQN에 전달한다. 이후 로봇은 DQN 이미지 분석을 통해 예측한 균형맞춤 동작을 구동한다. 로봇은 동작 구동 후 사람으로부터 구동 동작에 대한 평가인 표정 피드백을 받는다. 표정 피드백은 로봇의 카메라를 통해 입력되며, 표정 피드백 인식 및 변환 모듈을 거쳐 수치값으로 변환된 후 DQN 알고리즘의 환경 보상에 통합된다. 위 과정의 반복을 통해 로봇은 환경 보상과 사람 표정 피드백의 합이 최대화되는 정책을 학습하며, 학습의 결과로 협력 테이블 균형맞춤 동작을 수행할 수 있게 된다.

본 연구에서 테이블 균형맞춤 작업을 학습할 로봇은 Softbank의 NAO 로봇 V6[18]이며, 사용된 테이블은 가로 31 cm, 세로 23 cm, 높이 6 cm의 직육면체 모양의 테이블이다. 또한 표정 피드백 입력에는 NAO 로봇에 탑재된 상부 카메라를, 학습에 사용할 테이블 상태 이미지는 하부 카메라를 사용하여 촬영하였다.



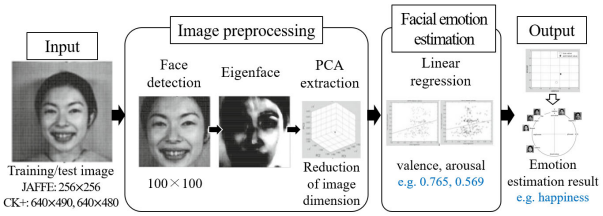
[Fig. 1] Interactive deep reinforcement learning model for table balancing based on human facial expression feedback

3.1 얼굴 표정 피드백 인식 및 변환 모듈

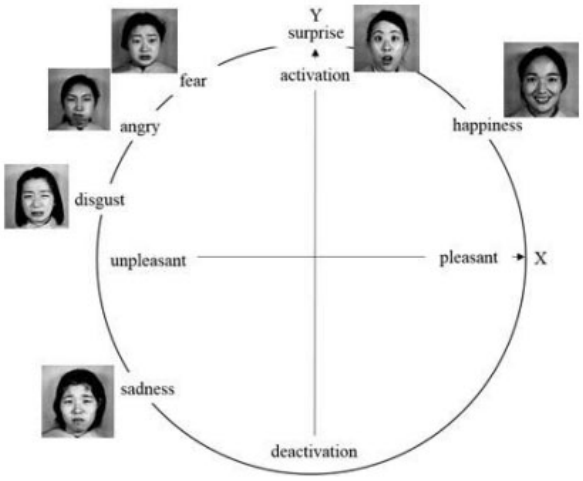
표정 피드백 인식 및 변환 모듈은 카메라를 통해 인식한 사람 감성 표정을 분석하며 수행한 로봇 동작이 긍정적, 중립적 혹은 부정적인지를 판단한다. 해당 모듈은 사람 표정 인식과 감성 표정 분석으로 구성된다. 감성 추정 시스템은 본 연구실에서 개발하여 발표되었으며[19], 사람 얼굴 표정 이미지를 입력 받아 해당 얼굴 이미지의 감성 추정 결과를 출력하는 컴포넌트이다.

개발된 사람 감성 표정 인식의 전체적인 작업도는 [Fig. 2]에 나타나 있다[19]. 사람 감성 표정 인식 모듈의 학습 데이터로는 JAFFE 데이터셋(Japanese female facial expression, JAFFE)[20]과 CK+ 데이터셋(Extended Cohn-Kanade, CK+)[21]을 사용하였으며, 사람 감성 표정의 표현에는 [Fig. 3]의 Russel의 2차원 감성 공간[22]을 기준으로 하여 접근하였다. Russel의 2차원 감성 공간에 따라 수평 축의 쾌/불쾌인 Valance 값과 수직 축의 각성/비각성인 Arousal 값의 조합으로 다양한 혼합 감성을 나타낼 수 있다. 사람 감성 표정은 Valance와 Arousal의 혼합 감성으로, Valance는 [Fig. 3] 가로축의 기쁨 정도를 나타내며 Arousal은 [Fig. 3] 세로축의 활성화 정도를 나타낸다.

먼저 데이터셋으로부터 사람 감성 표정 이미지를 입력 받아 얼굴 인식, 고유 얼굴(Eigenface)[23], 그리고 주성분분석



[Fig. 2] Workflow of emotion estimation module[19]



[Fig. 3] Russel's 2D emotion spaces[22]

(Principal Component Analysis, PCA)[24]의 전처리과정을 거친다. 이미지 전처리 과정 후에는 선형 회귀(Linear regression)알고리즘[25]을 통해 각 입력 이미지의 주성분에 대해 투영되는 값들을 학습하여 입력 이미지의 표정 추정 결과를 얻어낸다. 얻어낸 표정 추적 결과는[Fig. 3]의2차원 감성 공간에(Valance, Arousal) 값을 통해 나타내어 행복, 놀람, 기쁨, 슬픔, 화남, 불쾌함, 안정, 졸림, 중립의 9가지 감성 표정으로 분석된다. [Fig. 4]는 Valance와 Arousal의 조합으로 추정한 표정의 예시이다. [Fig. 4]에서 '행복'감성은 Valance, Arousal의 추정 값이 0.4383과 0.2272로 나타나 [Fig. 3]의 1사분면에 표현되고, '화남'감성은 -0.5101과 0.2846으로 [Fig. 3]의 2사분면에 표현된다. '불쾌'감성은 -0.7943과 -0.0222로 추정되어 2사분면과 3사분면 사이의 영역에 표현되어 얼굴 표정 감성이 추정된다.

이러한 학습 과정을 통해 각 얼굴 표정 피드백 이미지에 대한 Valance(쾌/불쾌)와 Arousal(각성/비각성)의 정도를 나타내는 값을 얻을 수 있다. 이 값들은 -1과 1사이의 실수이며 [Fig. 3]과 같이 Russel의 2차원 공간상에 배치되어 9가지의 감성 중 하나로 분석된다.

추정한 사람 감성 표정 추정 결과를 사용하여 [Table 1]과 같이 로봇에게 변환된 피드백을 제공한다. 로봇이 목표 작업인 테이블 균형맞춤 동작을 성공적으로 수행하였을 경우 놀람, 행복, 기쁨 표정으로 피드백을, 실패하였을 경우에는 중립, 졸림, 안정 표정을 통해 피드백을 제공하였다. 또한 균형맞춤 작업 모델에 정의되지 않은 동작[8]을 출력할 경우 화남, 불쾌, 슬픔 표정의 부정적 피드백을 제공하였다.

Happy (0.4383, 0.2272) | Angry (-0.5101, 0.2846) | Unpleasant (-0.7943, -0.0222)

[Fig. 4] Examples of facial expression feedback with converted feedback value (Valance, Arousal)

[Table 1] Emotion analysis and reward value for each expression

Emotion	Analysis	Reward
Happy	Positive feedback	+0.5
Surprise		
Pleasant		
Sad	Negative feedback	-0.5
Angry		
Unpleasant		
Calm	Neutral feedback	-0.3
Sleepy		
Neutral		

본 과정을 통해 사람 감성 표정 피드백을 로봇이 인식하여 상호작용 가능함을 확인하였다. 제안한 모듈은 테이블 균형맞춤 작업을 위한 로봇 강화학습에 적용되었으며 해당 과정에 대해서는 3.2절에서 더 자세히 설명한다.

3.2 테이블 균형맞춤을 위한 표정 피드백 도입 딥강화학습 프로세스

제안 시스템의 로봇은 표정 피드백 도입 DQN을 사용하여 테이블 상태 이미지 인식 및 테이블 균형맞춤 동작을 출력한다. DQN은 Q-러닝 기반 강화학습에 딥 합성곱 신경망을 결합한 기술이며, 입력 이미지와 동작이 주어졌을 때 상태-동작 가치 함수인 Q함수를 추정하는 알고리즘이다. DQN 시스템의 구현에는 Minh et al.이 제안한 DQN 2013[7]과 DQN 2015[26]의 두 가지 네트워크 구조가 고려되었다. 본 연구의 목표 작업인 테이블 균형맞춤 로봇에 대한 시스템 구현 및 학습 결과, 2013년 DQN 모델에서 더 높은 최적 정책 수렴 비율을 보이며 원활한 학습이 이루어짐에 따라 해당 구조를 채택하였다.

[Algorithm 1]은 표정 피드백 기반 인터액티브 DQN의 훈련 과정을 나타낸다. 본 훈련 과정은 DQN 훈련과정과 같으며 로

[Algorithm 1] Interactive Deep Q-Network Based on Facial Expression Feedback

Initialize action-value function with random weights θ
Initialize target action-value function $\hat{Q}$ with random weights $\theta^{-} = \theta$
for episodes = 1, 20000 **do**
  Initialize sequence
  for $t = 1,\ T$ **do**
    Get table state image $s_t = x_t$
    With probability ϵ select a random action a_t
    Otherwise select $a_t = \arg\max_{a \in A} Q_t(s_t, a_t)$
    Execute action a_t and observe reward r_t, image $x_t + 1$
    if Trainer provides facial expression feedback f_t on state s_t **then** Let $r_t \leftarrow r_t + f_t \times k$
    end if
    $$y_t \begin{cases} r_t & \text{if episode done at step } t+1 \\ r_t + \gamma \max_{a' \in A} \hat{Q}(s', a'; \theta^{-}) & \text{otherwise} \end{cases}$$
    Perform a gradient descent step on
    $$L(\theta) = \mathbb{E}[(y_t - Q(s_t, a_t; \theta_t))^2]$$
    with respect to the network parameters _
    Every 5 steps reset $\theta^{-} = \theta$
  end for
end for

봇 동작 후 표정 피드백 기반 인터액티브 과정이 추가되었다. 입력 상태 s는 테이블 이미지 x_t이며, x_t는 로봇 카메라로 촬영한 테이블 균형 상태를 나타내는 128×170 사이즈의 RGB 이미지이다.

환경은 훈련 데이터셋에서 테이블 상태 이미지를 선택한 후 DQN 에이전트인 로봇에게 전달한다. 로봇은 탐험을 위해 ϵ의 확률로 랜덤 동작을 선택하는 ϵ-greedy 정책에 따라 현재 동작을 결정한다. 랜덤 동작을 선택하지 않을 경우 에이전트는 상태-동작 가치 함수인 Q함수의 값을 최대화하는 동작을 선택한다. DQN이 예측하고자 하는 Q함수는 식 (1)과 같다.

$$Q_\pi(s,a) = \mathbb{E}_\pi \Sigma_{t=1}^{\infty} \gamma^t r_t \tag{1}$$

식 (1)에서 s는 상태, a는 로봇 동작, r은 현재 상태에서 동작을 수행하여 다음 상태로 갈 때 로봇이 받는 보상을 의미한다. Q함수는 상태 s에서 동작 a를 실행할 때 받는 누적 보상의 기댓값으로 표현되며, γ는 감가율로 미래 상태의 Q값의 영향력을 감소시킨다. 제안 시스템에서 테이블의 기울기 상태에 따라 정의된 사람 동작 상태는 총 5개로 올리기(s_{up}), 유지하기(s_0), 내리기(s_{down})로 나누어지고, 올리고 내린 정도에 따라 많이 올리기(s_{upup}), 많이 내리기($s_{downdown}$)로 구성되며 s의 아래 첨자는 사람의 동작을 나타낸다. 로봇은 무릎 관절 구동값을 조정하여 테이블 균형맞춤 동작 a를 구동한다. 테이블 이동 방향 및 정도에 따라 a_{upup}, a_{up}, a_0, a_{down}, $a_{downdown}$의 5가지 로봇 동작이 정의된다.

동작 구동 후 에이전트는 사람의 표정 피드백과 환경 보상을 받는다. 강화학습 시스템의 환경 보상은 [Table 2]와 같이 정의된다. 환경은 로봇이 목표 상태인 균형 유지 상태(s_0)에 도달할 경우 +0.5의 보상을, 목표가 아닌 상태에 도달하는 동작을 출력할 경우 -0.3의 보상을 제공한다. 또한 사람 동작 상태를 s_{upup}로 인식한 상태에서 a_{down}동작을 반환하는 것과 같이 균형맞춤 작업 모델 외 동작[8]을 출력할 경우 -0.5의 보상이 제공된다.

인터액티브 표정 피드백은 로봇의 동작에 대한 사람의 감성 표정 평가이다. 사람은 로봇의 동작에 따라 변화한 테이블 균형 상태를 확인한 후 로봇이 목표 상태에 도달하였을 경우에는 긍정적 표정 피드백을, 그렇지 않을 경우에는 부정적 표정 피드백을 제공한다. 제공된 표정 피드백은 3.1절에서 상술한 표정 피드백 인식 및 변환 모듈을 통해 수치값으로 변환된 후, 피드백 스케일을 조정하는 상수 k와 곱해져 강화학습 환경 보상에 합산된다. 사람이 표정 피드백을 제공할 때 로봇은 피드백과 환경 보상을 사용하며, 피드백이 없을 때는 환경 보상만을 사용한다.

[Table 2] DQN environmental reward model

Agent action	Reward
Reaching the target state	+0.5
Returning undefined action	-0.5
Reaching non-target states	-0.3

θ는 신경망의 파라미터를 의미한다. DQN에서는 y_t를 타겟으로 보고 y_t와 신경망에 의한 추정치인 $Q(s_t, a_t; \theta_t)$의 오차를 줄이는 방향으로 학습을 진행한다. 따라서 DQN 모델 업데이트는 평균제곱오차를 계산한 손실 함수 $L(\theta)$를 통해 매 에피소드마다 이루어진다. $L(\theta)$를 최소화하는 방향으로 θ를 반복적으로 업데이트하면 Q함수는 점점 최적의 상태-동작 가치함수에 가까워지며 에이전트는 최적의 행동을 학습한다.

이러한 과정을 통해 로봇은 테이블 균형맞춤을 위해 사람의 표정 피드백을 적용한 인터액티브 DQN을 학습할 수 있으며, DQN 프레임워크에서 표정 피드백을 활용하기 위해 본 연구에서는 표정 피드백 인식 및 변환 모듈을 통해 환경 보상과 표정 피드백으로부터 보상 함수를 구현하였다.

4. 실험 결과

본 절에서는 제안한 인터액티브 딥강화학습 모델의 성능 검증을 위한 표정 피드백 인식 모듈 평가, 인터액티브 DQN 모델 실험결과에 대해 설명한다.

4.1 표정 피드백 인식 모듈 테스트

먼저 제안한 표정 피드백 인식 및 변환 모듈의 성능을 측정하였다. 실험에 사용할 사람 표정 피드백 인식 테스트 데이터셋은 JAFFE와 CK+ 데이터셋을 혼합하여 구축하였다. 또한 Leave-one-out-cross-validation (LOOCV)[27] 방법에 따라 해당 데이터셋의 790개 이미지 중 789개 이미지를 훈련 데이터로, 나머지 1개 이미지를 테스트 데이터로 사용하여 학습을 진행한 후 식 (2)와 같이 평균 제곱근 오차(Root mean squared error, RMSE)를 계산하였다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}[y(x_i) - \hat{y}(x_i)]^2} \tag{2}$$

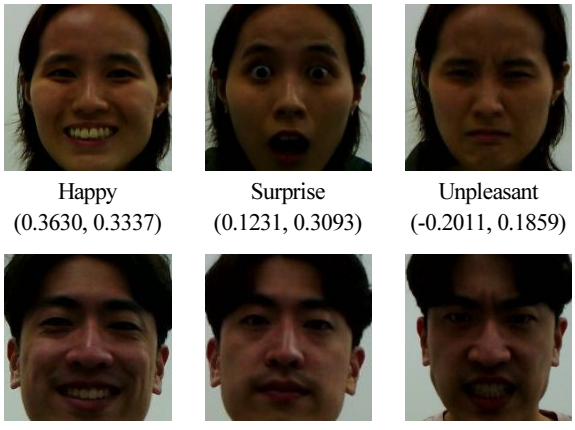
식 (2)에서 $y(x_i)$는 테스트 데이터셋의 정답 레이블에 의한 valance, arousal의 감성 표정 값을 나타내고 $\hat{y}(x_i)$는 제안한 모듈에 의해 추정된 감성 표정 값을 나타낸다. 제 제안한 시스

템이 추정한 감성 표정 값이 정답 값과 가까울수록 RMSE 값은 작아진다. 표정 피드백 인식 및 변환 모듈의 테스트 성능은 [Table 3]과 같이 나타난다. [Table 3]에서 RMSE_A는 얼굴 표정 피드백의 Arousal값 추정의 오차를, RMSE_V는 Valance 값의 추정 오차를 나타낸다. 제안한 표정 피드백 인식 시스템에서 RMSE_A는 0.2289, RMSE_V는 0.2609의 오차를 보였다. 이는 선형 회귀만 적용한 모델의 RMSE인 1.5598, 1.6918에 비해 현저히 낮은 오차를 보이는 우수한 성능을 보인다[19]. 이를 통해 제안한 표정 피드백 인식 모델에서 표정 추정 오차가 안정적으로 수렴하여 학습이 잘 이루어졌음을 알 수 있다.

또한 NAO V6 로봇에 제안한 표정 피드백 인식 및 변환 모듈을 탑재하여 실시간 표정 인식이 잘 이루어지는지 평가하였다. 이 경우에는 실시간으로 로봇에 탑재된 MT9M114 카메라를 통해 사람 표정 피드백 이미지를 촬영 및 모듈에 입력함으로써 표정 피드백 인식 및 변환이 이루어졌다. [Fig. 5]는 실제 환경에서 표정 피드백 인식 및 변환 모듈을 테스트한 결과인 (Valance, arousal)감성 추정 값과 해당 값들을 기반으로 추정한 사람 표정 피드백 감성 결과를 나타낸다. 이를 통해 제안한 표정 피드백 인식 및 변환 모듈을 로봇에 탑재하여 로봇이 사람 표정 피드백을 실시간 환경에서 올바르게 추정할 수 있음을 보였다.

[Table 3] The minimum RMSE of the proposed method

Method	Item	Value
Proposed method	RMSE_A	0.2289
	RMSE_V	0.2609
Linear regression	RMSE_A	1.5598
	RMSE_V	1.6918



[Fig. 5] Examples of facial expression feedback with converted feedback value (valance, arousal) by NAO V6 robot

4.2 제안한 인터액티브 딥강화학습 모델 학습 결과

본 절에서는 제안한 표정 피드백 사용 DQN 모델의 성능 테스트를 위한 실험을 진행한 후 결과를 분석하였다. 테스트에 사용한 피드백 모델은 연속적 제공 표정 피드백[28]으로, 해당 설정에서 사람은 학습 초기에 100회의 표정 피드백을 연속적으로 제공한다. 에이전트는 이 피드백을 강화학습 시스템의 보상 함수에 반영한다. 해당 피드백 제공 설정은 학습 초기에 집중적으로 사람 표정 피드백을 전달하여 초기 학습 방향을 잡는 것을 목표로 디자인되었다.

DQN 훈련을 위한 하이퍼 파라미터 설정은 [Table 4]에 나타나 있다. 표정 피드백 횟수를 제외한 모든 하이퍼 파라미터 설정은 제안한 인터액티브 DQN 모델과 베이스라인 DQN 모두에 동일하게 적용된다. 또한 표정 피드백 모델의 총 피드백 제공 횟수는 [Table 4]와 같이 20,000번 중 100번으로 동일하며, 이 외 에피소드는 [Table 2]의 환경 보상만을 사용하였다.

또한 학습과 테스트에 사용할 훈련 데이터셋과 테스트 데이터셋을 구축하였다. 각 상태별 500장의 이미지 데이터를 구성하였으며, 총 2500장의 테이블 데이터셋 이미지 중 2250장은 훈련 데이터셋, 250장은 테스트 데이터셋으로 분리하여 사용하였다.

먼저, 실험 과정에서 모델 설정별로 30회씩 실험을 진행하였으며 훈련이 끝난 후 최적 정책 수렴 비율을 계산하여 학습 성능을 평가하였다. 최적 정책 수렴 비율 계산식은 식 (3)과 같다.

$$\frac{(\text{모델의 최적 정책 수렴 횟수})}{(\text{전체 실험 횟수})} \times 100 \tag{3}$$

실험을 통해 피드백 제공 모델과 베이스라인인 표정 피드백 미사용 DQN의 학습 결과를 비교하였으며, 추가적으로 4가지의 모델 최적화 비교 실험을 진행하였다. 실험에 사용한 4가지 옵티마이저는 SGD[29], Adam[30], Adagrad[31], Adadelta[32]로 최근 ConvNet과 같은 이미지 데이터를 다루는 심층 신경망 학습 연구들에 활발히 적용되고 있으며, 향상된 모델 수렴 속도, 높은 분류 정확도와 같이 좋은 성능을 이끌어낸 최적화 알고리즘들이다[33,34].

[Table 4] Hyperparameters of DQN training

Hyperparameter	Value
Learning rate α	0.001
Discount factor γ	0.9
Epsilon ϵ	20
Number of episodes	20,000
Number of facial expression feedback	100

[Table 5] Training result with optimal policy convergence rate of baseline and interactive DQN with facial expression feedback in 4 different optimizer settings

Optimizer	Baseline	DQN with facial feedback
SGD	76.6%	83.3%
Adam	73.3%	76.6%
Adagrad	43.3%	70%
Adadelta	63.3%	73.3% ($k = 4$)

[Table 6] Test result with successful assumption rate in a single trial on the baseline and the proposed model in 4 different optimizer settings

Optimizer	Baseline	DQN with facial feedback
SGD	77.4%	91.6%
Adam	76.8%	90.4%
Adagrad	68.8%	76.6%
Adadelta	63.3%	87.8%

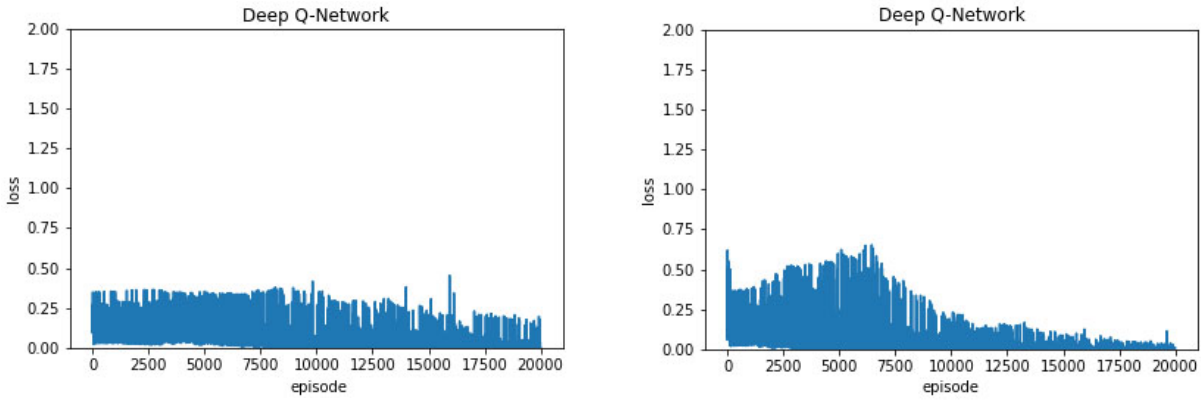
제안한 표정 피드백 사용 인터액티브 딥강화학습 모델과 베이스라인 DQN에 대해 30회의 실험을 진행한 후 학습 결과인 최적 정책 수렴율은 [Table 5]에 나타나 있다. 제안한 시스템인 표정 피드백 사용 DQN의 경우 4가지의 실험 설정 모두에서 30회의 실험 중 21회 이상(70%) 최적 정책으로 수렴하였으며, 베이스라인 DQN에 비해 최대 27% 향상된 최적 정책 수렴율을 보였다. 또한 전체적으로 베이스라인 DQN에 비해 제안한 인터액티브 딥강화학습 모델이 더 나은 성능을 보이고 있는데, [Fig. 6]을 통해 그 이유를 분석할 수 있다. [Fig. 6]는 베이스라인과 Adadelta 옵티마이저 설정을 사용한 인터액티브 DQN의 학습 손실 그래프이다. 학습 손실 그래프인 그림 [Fig. 6(a)], [Fig. 6(b)]를 살펴보면 (b)의 표정 피드백 사용 인터액티브 DQN 에서는 피드백이 제공되는 초반에는 상대적으로 손실이 높게 나타나나, 에피소드 횟수가 증가할수록 손실이 안정적으로 0에 수렴하고 있다. 이에 비해 표정 피드백 미사용 DQN인 베이스라인(a)에서는 손실이 0을 향해 감소하는 추세가 상대적으로 느리게 나타나 있다. 따라서 같은 에피소드 횟수를 사용해 학습을 수행하였을 때 표정 피드백 사용 모델에 비해 최적 정책으로 수렴하는 속도가 더 느리게 나타나며, 이는 제안 시스템에 비해 더 낮은 최적 정책 수렴율을 보이는 결과로 이어진다.

또한 구축한 테스트 데이터셋을 이용하여 제안한 표정 피드백 사용 DQN 모델과 베이스라인 DQN에 대해 250회 테스트를 실시하였으며, 전체 실험 횟수에 대하여 로봇이 한 번의 시도(trial)에 테이블 균형맞춤 작업에 성공한 비율을 [Table 6]에 비교하여 나타내었다. 제안한 표정 피드백 사용 모델은 모든 옵티마이저 설정에서 베이스라인 DQN 모델보다 더 높은 균형맞춤 성공율을 보이며 테이블 균형맞춤 동작을 더 잘 수행하였다. 특히 SGD, Adam 옵티마이저 설정에서는 90%가 넘는 성공률을 보였으며, 이는 기존 베이스라인 모델보다 13% 이상 높은 성공률로 성능의 향상이 이루어졌다.

이와 같은 실험 결과를 통해 테이블 균형맞춤 작업을 위한 DQN에 인터액티브 표정 피드백을 도입하여 대부분의 옵티마이저 셋팅에서 모델 성능을 개선할 수 있음을 확인하였다.

5. 결론 및 고찰

본 연구는 사람-로봇 협동 테이블 균형맞춤 작업을 위한 표정 피드백 기반 인터액티브 딥강화학습 모델을 제안하였다. 제안한 기술은 표정 추정 기술을 사용하여 감성 표정 피드백 인식 및 변환 모듈을 구현 및 DQN에 적용하였으며, 이를 통해 표정 피드백 제공이라는 가장 자연스러운 방식으로 로봇 학습



(a) Baseline (b) Interactive DQN

[Fig. 6] Loss graph of baseline and interactive DQN in Adadelta optimizer setting

이 이루어지는 시스템을 구현하였다. 학습의 결과로 로봇은 사람과 협력 테이블 균형맞춤 동작을 수행할 수 있게 된다. 실험 결과, 학습 초기에 연속적 표정 피드백을 제공하는 인터액티브 DQN 모델이 피드백 미사용 DQN보다 높은 최적 정책 수렴 비율을 달성하여 일관적으로 향상된 성능을 보였다.

후속 연구의 작업 방향은 다양한 로봇 센서를 사용하여 DQN에 음성+표정, 표정+제스처와 같은 멀티모달 피드백을 통합하는 것을 포함한다. 또한 다양한 학습 설정에서 인터액티브 딥강화학습의 모델의 학습 성능을 향상시키는 심층 모델 최적화 기술 개발에 집중할 수 있다. 또한 로봇이 학습 과정에서 표정 피드백 인식의 신뢰도에 따라 표정 피드백을 사용할 때와 폐기할 때를 학습하여 더 안정적인 학습을 하도록 확장 개발이 가능하다.

References

[1] M. Tonkin, J. Vitale, S. Herse, M.-A. Williams, W. Judge, and X. Wang, "Design Methodology for the UX of HRI: A Field Study of a Commercial Social Robot at an Airport," *2018 ACM/IEEE International Conference on Human-Robot Interaction*, Chicago IL, USA, pp. 407-415, 2018, DOI: 10.1145/3171221.3171270.

[2] T. Morita, N. Kashiwagi, A. Yorozu, H. Suzuki, and T. Yamaguchi, "Evaluation of a multi-robot cafe based on service quality dimensions," *The Review of Socionetwork Strategies,* vol. 14, no.1, pp. 55-76, 2020, DOI: 10.1007/s12626-019-00049-x.

[3] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "IkeaBot: An autonomous multi-robot coordinated furniture assembly system," *2013 IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, DOI: 10.1109/ICRA.2013.6630673.

[4] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research,* vol. 32, no. 11, 2013, DOI: 10.1177/0278364913495721.

[5] H. Nguyen and H. La, "Review of deep reinforcement learning for robot manipulation," *2019 Third IEEE International Conference on Robotic Computing (IRC),* Naples, Italy, pp. 590-595, 2019, DOI: 10.1109/IRC.2019.00120.

[6] A. L. Thomaz, G. Hoffman, and C. Breazeal, "Reinforcement learning with human teachers: Understanding how people want to teach robots," *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication,* Hatfield, UK, 2006, DOI: 10.1109/ROMAN.2006.314459.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013, DOI: 10.48550/arXiv.1312.5602.

[8] Y. Kim and B.-Y. Kang, "Cooperative robot for table balancing using q-learning," *The Journal of Korea Robotics Society,* vol. 15, no. 4, pp. 404-412, Dec., 2020, DOI: 10.7746/jkros.2020.15.4.404.

[9] B. Price and C. Boutilier, "Accelerating reinforcement learning through implicit imitation," *Journal of Artificial Intelligence Research*, vol. 19, pp. 569-629, 2003, DOI: 10.1613/jair.898.

[10] Z. Wang and M. E. Taylor, "Interactive Reinforcement Learning with Dynamic Reuse of Prior Knowledge from Human/Agent's Demonstration," *arXiv preprint arXiv:1805.04493*, 2018, DOI: 10.48550/arXiv.1805.04493.

[11] T. Brys, A. Harutyunyan, H. B. Suay, S. Chernova, M. E. Taylor, and A. Nowé, "Reinforcement learning from demonstration through shaping," *24th International Conference on Artificial Intelligence*, pp. 3352-3358, 2015, [Online], https://dl.acm.org/doi/abs/10.5555/2832581.2832716.

[12] M. Ullerstam and M. Mizukawa, "Teaching robots behavior patterns by using reinforcement learning: how to raise pet robots with a remote control," *SICE 2004 Annual Conference*, Sapporo, Japan, 2004, [Online], https://ieeexplore.ieee.org/document/1491384.

[13] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," *Fifth International Conference on Knowledge Capture*, pp. 9-16, 2009, DOI: 10.1145/1597735.1597738.

[14] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," *Advances in neural information processing systems 26 (NIPS 2013)*, 2013, [Online], https://proceedings.neurips.cc/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html.

[15] T. A. Kessler Faulkner, E. S. Short, and A. L. Thomaz, "Interactive reinforcement learning with inaccurate feedback," *2020 IEEE International Conference on Robotics and Automation (ICRA),* Paris, France, 2020, DOI: 10.1109/ICRA40945.2020.9197219.

[16] V. Veeriah, P. M. Pilarski, and R. S. Sutton, ''Face valuing: Training user interfaces with facial expressions and reinforcement learning,'' *arXiv:1606.02807*, 2016, [Online], http://arxiv.org/abs/1606.02807.

[17] R. Arakawa, S. Kobayashi, Y. Unno, Y. Tsuboi, and S. Maeda, Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback," *arXiv preprint arXiv:1810.11748*, 2018, [Online], https://arxiv.org/abs/1810.11748.

[18] NAO the humanoid and programmable robot | SoftBank Robotics, [Online], https://www.softbankrobotics.com/emea/en/nao, Access: Jun. 7, 2022.

[19] H.-S. Lee and B.-Y. Kang, "Continuous emotion estimation of facial expressions on JAFFE and CK+ datasets for human-robot interaction," *Intelligent Service Robotics*, vol. 13, no.1, 2020, DOI: 10.1007/s11370-019-00301-x.

[20] M. Lyons, M. Kamachi, M., and J. Gyoba, "The Japanese Female Facial Expression (JAFFE) Dataset," *Third International Conference on Automatic Face and Gesture Recognition*, Apr., 1998, DOI: 10.5281/zenodo.3451524.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, 2010, DOI: 10.1109/CVPRW.2010.5543262.

[22] J. A. Russel, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980, DOI: 10.1037/h0077714.

[23] P. Shakyawar, P. Choure, and U. Singh, "Eigenface method through facial expression recognition," *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2017, DOI: 10.1109/ICECA.2017.8212714.

[24] H. Zou and L. Xue, "A selective overview of sparse principal component analysis," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1311-1320, Aug., 2018, DOI: 10.1109/JPROC.2018.2846588.

[25] R. Ewing and K. Park, "Linear regression," *Basic Quantitative Research Methods for Urban Planners*. Routledge, pp. 220-269, 2020, [Online], https://books.google.co.kr/books?hl=ko&%20lr=&id=Gzz3DwAAQBAJ&oi=fnd&pg=PP1&ots=HJz-Tw6pgs&sig=oFD_mUSrG3iw3pFr8_uL9bc0STw&redi%20r_esc=y#v=onepage&q&f=false.

[26] 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 2015, DOI: 10.1038/nature14236.

[27] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition* vol. 48, no. 9, pp. 2839-2846, May, 2015, DOI: 10.1016/j.patcog.2015.03.009.

[28] H. Jeon, Y. Kim, and B.-Y. Kang, "Interactive Reinforcement Learning for Table Balancing Robot," *Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, 2021, DOI: 10.18653/v1/2021.splurobonlp-1.8.

[29] T. Schaul, I. Antonoglou, and D. Silver, "Unit Tests for Stochastic Optimization," *arXiv Preprint arXiv:1312.6055*, 2013, DOI: 10.48550/arXiv.1312.6055.

[30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412. 6980*, 2014, DOI: 10.48550/arXiv.1412.6980.

[31] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011, [Online], https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf.

[32] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *arXiv Preprint arXiv:1212.5701*, 2012, DOI: 10.48550/arXiv.1212.5701.

[33] E. M. Dogo, O. J. Afolabi, N. I. Nwulu, B. Twala, and C. O. Aigbavboa, "A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks," *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, 2018, DOI: 10.1109/CTEMS.2018.8769211.

[34] I. Kandel, M. Castelli, and A. Popovič, "Comparative study of first order optimizers for image classification using convolutional neural networks on histopathology images," *Journal of Imaging*, vol. 6, no. 9, 2020, DOI: 10.3390/jimaging6090092.

전 해 인

2020 경북대학교 영어교육과(학사)
2022 경북대학교 인공지능학과(석사)
2022~현재 경북대학교 인공지능학과(박사)

관심분야: 로보틱스, 인공지능, 딥러닝

강 정 훈

2020 경북대학교 정밀기계공학과(학사)
2020 경북대학교 컴퓨터학부(학사)
2021~현재 경북대학교 인공지능학과(석사)

관심분야: 인공지능, 강화학습, 딥러닝, 협력로봇

강 보 영

2004 경북대학교 컴퓨터공학과박사
2005 KAIST ICC 박사후 연구원
2006 서울대학교 치의학전문대학원 연구조교수
2018~현재 경북대학교 정교수

관심분야: 딥러닝, 강화학습, 인공지능, 협력로봇