Check for updates

Journal of Korea Robotics Society (2023) 18(2):225-232
https://doi.org/10.7746/jkros.2023.18.2.225 ISSN: 1975-6291 / eISSN: 2287-3961

다중 로봇 제조 물류 작업을 위한 안전성과 효율성 학습

Safety and Efficiency Learning for Multi-Robot Manufacturing Logistics Tasks

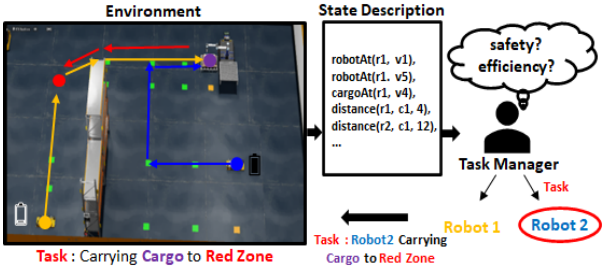
강 민 교[1] · 김 인 철†
Minkyo Kang[1], Incheol Kim†

Abstract: With the recent increase of multiple robots cooperating in smart manufacturing logistics environments, it has become very important how to predict the safety and efficiency of the individual tasks and dynamically assign them to the best one of available robots. In this paper, we propose a novel task policy learner based on deep relational reinforcement learning for predicting the safety and efficiency of tasks in a multi-robot manufacturing logistics environment. To reduce learning complexity, the proposed system divides the entire safety/efficiency prediction process into two distinct steps: the policy parameter estimation and the rule-based policy inference. It also makes full use of domain-specific knowledge for policy rule learning. Through experiments conducted with virtual dynamic manufacturing logistics environments using NVIDIA's Isaac simulator, we show the effectiveness and superiority of the proposed system.

Keywords: Smart Factory, Multi-Robot Logistics, Logic Rule Learning, Relational Reinforcement Learning, Domain Knowledge

1. 서 론

최근 들어 복수의 로봇들을 도입한 스마트 제조, 스마트 물류 환경들이 증가하면서, 동적인 작업 환경에 맞추어 어떤 작업 공정을 어떤 로봇에게 할당하는 것이 작업의 안전성과 효율성을 담보할 수 있을지 예측하는 일이 매우 중요해졌다. 예컨대, [Fig. 1]의 예시와 같이, 하나의 제조 물류 환경 내에 2대의 이동 로봇들(노랑, 파랑)과 운반해야 할 화물(보라)이 있을 때, 화물을 목표 위치(빨강)로 이동 적재하는 작업을 현재 상황에서는 어느 로봇에게 맡기는 것이 상대적으로 작업의 안전성과 효율성이 더 높을 것인지 예측할 필요가 있다. 이러한 예측 추론을 위해 시도해볼 다양한 지식-기반 접근 방법들[1,2]이 존재할 수 있으나, 본 논문에서는 영역-고유 지식(domain-specific knowledge)의 활용과 더불어 경험에 따른 학습(experience-based learning) 능력을 효과적으로 활용하기 위해 심층 관계형 강화 학습(Deep Relational Reinforcement Learning, DRRL) 기반의 논리 추론 규칙 학습법을 적용한다. 일반적으로 심층 관계형 강화 학습[3,4]은 작업 상태(state), 로봇 행동(action), 학습된 정책 규칙(policy rule)들을 모두 신경망 기반의 논리 서술자(logic predicate)들로 표현함으로써, 학습 과정뿐만 아니라 학습 결과물인 정책 규칙들의 해석 가능성(interpretability)과 일반성(generality)이 매우 높다는 장점이 있다. 하지만, 영역-고유 지식을 효과적으로 활용하지 못하고 시행-착오에 따른 보상 피드백에만 의존하여 학습을 수행할 경우, 학습 복잡도가



[Fig. 1] An example of multi-robot logistics task

Received : Feb. 7. 2023; Revised : Mar. 12. 2023; Accepted : Mar. 23. 2023
※ This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00096, Robot task planning for single and multiple robots connected to a cloud system)
1. Student, Computer Science, Kyonggi University, Suwon, Korea (alsry5786@kyonggi.ac.kr)
† Full Professor, Corresponding author: Computer Science, Kyonggi University, Suwon, Korea (kic@kyonggi.ac.kr)

Copyright ⓒKROS

높아서 제한된 시간 내에 실용적인 정책 규칙들을 학습해내기 어렵다는 한계성도 있다.

본 논문에서는 이와 같은 문제점을 극복하고 다중 로봇 물류 환경에서 효과적으로 작업의 안전성과 효율성을 예측하기 위해, dNL-RRL[4] 심층 관계형 강화 학습 엔진 기반의 작업 정책 학습기(Task Policy Learner, TPL)를 제안한다. 제안 시스템에서는 학습 복잡도를 낮추기 위해, 작업의 안전성과 효율성 예측 추론 단계를 정책 파라미터 추정 단계, 규칙 기반 정책 추론 단계 이렇게 두 단계로 나누고, 규칙 기반 정책 추론 단계에 적용할 추론 규칙 학습에 영역-고유 지식을 최대한 활용한다. 또한, 계층적 서술자 규칙 설계를 통해 더 효과적인 작업의 안전성과 효율성을 예측한다. 본 논문에서 제안한 작업 정책 학습기의 유효성과 효율성을 평가하기 위해, NVIDIA에서 제공하는 Isaac 시뮬레이터[5]를 활용하여 실제 스마트 제조 물류 환경과 유사한 실험 환경을 구축하고 다양한 검증 실험들을 진행하였다.

2. 관련 연구

복수 로봇들이 동시에 작업하는 동적 제조 물류 환경에서 어느 로봇에게 어떤 작업 공정을 할당하는 것이 안전성(safety)과 효율성(efficiency)을 보장할 수 있는가는 전체 작업 공정의 성공으로 이어지기에 매우 중요하다. 일반적으로 로봇 작업의 안전성은 작업 수행 시 로봇의 충돌 가능성, 적재 화물의 낙하 가능성, 배터리 잔여량 등을 고려하여 평가하며, 효율성은 작업 수행 시 소모되는 시간, 자원 등을 고려하여 평가한다. 이러한 안전성과 효율성에 대한 평가 예측을 위해 기존 연구들에서 다양한 방법들이 시도되었다. 대표적인 지식-기반 접근 방법[1,2]들은 안전성과 효율성 평가를 위한 규칙을 작업 수행 이전에 미리 정의하여 사용하였으며, 이러한 방법은 영역-고유 지식 활용 측면으로는 효과적이지만 로봇을 조작하는 제조 물류 환경과 같은 불확실성이 존재하는 환경에서는 강건하지 못하며, 여러 동적인 상황에서 발생하는 예외적 상황을 고려하기 어렵다는 한계를 갖는다.

특히, [2] 연구는 장애물이 많은 복잡한 작업 환경에서의 이동 작업(navigation task)의 성공을 담보하고자 정형화된 안전성 규칙을 작성하여 안전성이 담보된 상황에서 행동이 실행되도록 하였다. 특히, 안전성 논리 규칙(safety logic rule)은 환경 내의 장애물들과의 위치와 같은 공간적 맥락(spatial context) 정보와 과거 행동들의 시간적 맥락(temporal context) 정보를 활용하여 정형화하였다. 하지만, 정형화된 규칙을 통한 방법은 여전히 실시간으로 변하는 작업 환경에서 정확한 안전성을 추론하기 어려운 한계를 갖는다.

한편, 개별 및 군집 로봇들의 행동 정책, StarcraftII, Atrai 게임과 같은 온라인 게임 에이전트 학습에 활용되는 심층 강화학습(Deep Reinforcement Learning, DRL)[6,7]은 환경과의 시행착오적 상호작용을 통해 학습 데이터를 얻어 능동적으로 학습한다는 장점을 갖지만 다수의 인공 신경망으로 표현되는 행동 정책을 해석할 수 없다는 한계를 갖는다. 반면에, 일차 술어 논리 형태로 상태 및 행동을 표현하고 학습하는 관계형 강화학습(Relational Reinforcement Learning, RRL)은 높은 해석 가능성과 일반화 능력을 갖는다. 이러한 장점이 있는 관계형 강화학습은 최근 다양한 작업에 적용되어 연구되고 있다. NLRL (Neural Logic Reinforcement Learning)[3]은 대표적인 관계형 강화학습 연구로 학습 가능한 귀납적 논리 프로그래밍 엔진인 δILP[8]를 활용한 학습 가능한 순환 논리 머신(Differential Recurrent Logic Machine, DRLM)을 통하여 모든 구체화된 서술자(ground atom)들의 가치 평가 벡터(valuation vector)를 입력받아 N-단계의 전향 추론을 수행하여 가치 평가 값을 계산한다. 추론된 가치 평가 값은 학습자가 수행할 행동들의 확률 분포이며, 이 확률 분포를 이용하여 수행할 행동을 확률적으로 선택한다. 한편, NLRL은 상태-행동공간과 정책 공간을 모두 일차 술어 논리와 규칙으로 표현함으로써 설명 가능성을 보장하지만, 학습 이전에 미리 정의한 구문적 템플릿을 이용하여 조합적으로 상태-행동 공간(state-action space)과 정책 공간(policy space)을 초기화함에 따라 계산 복잡도가 높아지고 확장성이 크게 떨어지는 한계를 갖는다.

한편, dNL-RRL (differentiable Neural Logic-Relational Reinforcement Learning)[4]은 상태, 행동, 정책 모두를 논리 서술자로 표현하여 설명 가능성을 보장함과 동시에 논리 곱(logical conjunction), 논리 합(logical disjunction) 계층의 인공 신경망으로 이루어진 미분 가능한 뉴로-논리 귀납적 논리 프로그래밍 엔진(differentiable Neural-Logic Inductive Logic Programming, dNL-ILP)[9]을 채택하여 학습의 확장성을 NLRL과 같은 기존 관계형 강화학습 연구에 비해 크게 높였다. 그러나 dNL-RRL 또한 구문적 템플릿을 사용하여 인공 신경망 구조를 초기화함에 따라 초기 학습의 부담을 높이고 불필요한 자원을 소모하게 된다.

이 밖에도 관계형 강화학습은 다양한 형태로 연구되었다. 이들 중 GBFS-GNN[10], RDRL[11], SR-DRL[12], SYMNET[13] 등은 그래프 기반의 상태 표현(graphical state representation)을 이용하는 관계형 강화학습 연구로 그래프 표현을 통해 확장성(scalability)을 높였다. 또, 작업 환경 내의 관계 유추가 중요한 도메인으로부터 관계형 상태 표현을 모델 입력으로 받는 SRN[14], RFQL[15] 등도 있다. 또한, 한 쌍의 이미지들로부터 변화되는 과정을 논리 서술자로 표현된 행동 시퀀스로 생성하는 Gokhale[16] 등도 제안되었다.

3. 작업 정책 학습기

3.1 모델 개요

본 논문에서 제안하는 작업 정책 학습기의 구성은 [Fig. 2]와 같다. 작업 정책 학습기는 크게 정책 파라미터 추론기(Policy Parameter Reasoner, PPR)를 이용한 정책 파라미터 추론 단계와 심층 관계형 강화 학습기(Deep Relational Reinforcement Learner, DRRL)를 이용한 정책 추론 단계로 구성되며, 정책 추론 파라미터 추론 단계에서는 환경으로부터 로봇 및 화물들의 위치, 속도 등에 대한 관측값(observation, o_t)을 받아, 상태 인코더(state encoder)에 입력한다. 상태 인코더는 신호(signal) 형태의 관측값을 술어 논리(predicate logic) 형태의 상태(state, s_t)로 인코딩한다. 파라미터 등급 추정기(parameter grade estimator)는 영역-고유 지식을 바탕으로 미리 정의해놓은 다수의 평가 함수들을 이용하여 작업의 종합적인 안전성과 효율성 예측에 필요한 다양한 정책 파라미터(policy parameter, s_t^P)들의 등급을 추정하는 역할을 수행한다. 이와 같이 사전에 미리 정의해 둔 다면적 평가함수들을 이용한 정책 파라미터들의 등급 추정 방식은 작업 정책 학습기(TPL)의 전체 학습 부담을 줄여주고 동적으로 변화하는 작업 상태에 따라 실시간적으로 작업의 안전성과 효율성 예측을 가능하게 해주는 장점이 있다.

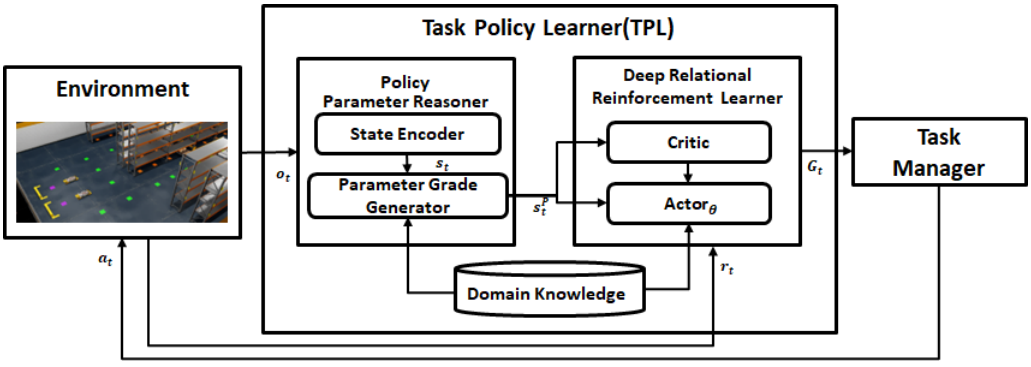
정책 추론 단계에서는 심층 관계형 강화 학습을 통해 습득한 작업 정책 규칙들을 이용하여, 각 로봇별 작업의 안전성과 효율성을 추론한다. 심층 관계형 강화 학습기는 행동가-비평가(actor-critic) 구조를 따르며, 먼저 비평가(critic)는 보상(reward)과 경험 데이터를 토대로 행동자(actor)의 정책 규칙들의 가치 Q를 평가해줌으로써, 행동자가 올바른 정책 규칙을 더 신속하게 학습하도록 돕는 역할을 수행한다. 행동가는 비평가의 가치 평가와 현재 입력 상태를 기반으로 작업 정책 규칙들을 학습하기도 하고, 현재 입력 상태와 학습된 규칙들을 토대로 전향 추론을 통해 작업의 안전성과 효율성을 예측하기도 한다.

작업의 안전성과 효율성을 예측하는 정책 규칙들은 논리 서술자들의 곱의 합 표준형(Disjunctive Normal Form, DNF)로 표현되며, 이러한 각 개별 규칙은 논리곱 계층들과 논리합 계층들로 구성된 하나의 인공 신경망 형태로 학습한다. 하지만, 무작위로 선택된 초기 신경망 가중치로부터 의미 있는 정책 규칙들을 학습하는 것은 많은 시간과 자원을 요구한다.

3.2 심층 관계형 강화 학습기

본 논문에서 제안하는 심층 관계형 강화 학습기는 대표적인 관계형 강화 학습 프레임 워크인 dNL-RRL을 기초로, 다양한 인간의 영역-고유 지식을 활용 방법들을 제안한다. 심층 관계형 강화 학습기는 [Fig. 3]과 같이, 행동자-비평가(actor-critic) 구조를 따른다. 비평가(critic)는 환경으로부터 얻은 경험 데이터들과 보상을 기반으로 행위자의 정책 규칙들의 가치 Q를 가치 누산기(value accumulator)를 통해 평가한다. 비평가의 평가 가치 Q는 행동가가 더 올바른 정책 규칙을 학습하도록 돕니다. 행동가(actor)는 비평가로부터 평가된 가치 평가 값 Q와 경험 데이터들을 기반으로 정책 규칙을 학습하고, 이 학습된 정책 규칙들을 기반으로 매 순간 입력되는 상태로부터 작업의 안전성과 효율성을 추론하는 역할을 수행한다.

한편, 본 논문에서 제안하는 작업 정책 학습기에 안전성과 효율성 규칙을 포함하는 행위자는 다시 상태 임베더(state embedder), 미분 가능한 뉴로-논리 귀납적 논리 프로그래밍 엔진(differentiable Neural-Logic Inductive Logic Programming, dNL-ILP) 그리고 정책 디코더(policy decoder)로 구성된다. 상태 임베더에서는 입력되는 상태인 정책 파라미터 s_t^P를 일차 술어논리(First-Order predicate Logic, FOL) 형태로 변경하고, 인공 신경망이 학습 가능하도록 다차원의 벡터로 변환해주는 역할을 담당한다. 미분 가능한 뉴로-논리 귀납적 논리 프로그래밍(dNL-ILP)은 평가자(critic)의 가치 평가와 입력 상태를 토대로 정책 규칙을 학습하며, 학습된 정책 규칙은 n 단계의 전향 추론(n-step forward

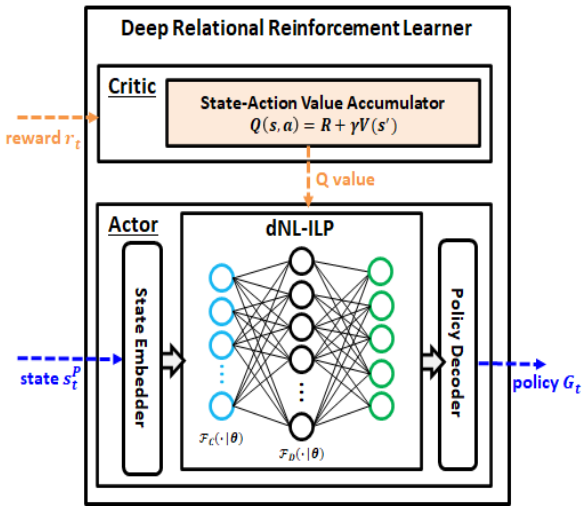


[Fig. 2] Architecture of task policy learner

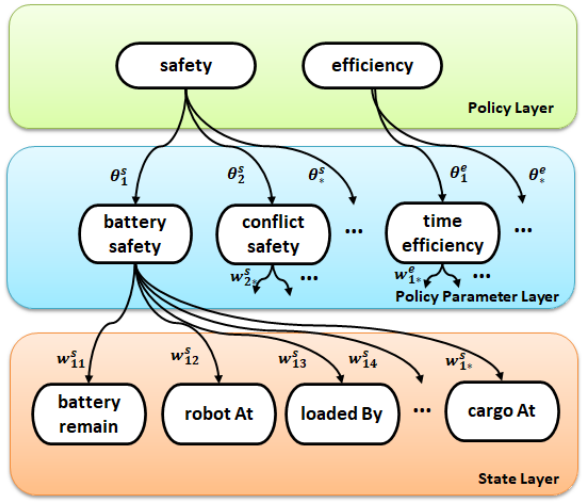
chaining)을 통해 안전성과 효율성 평가 값을 추론한다. 정책 디코더는 앞서 추론한 안전성과 효율성 평가 값을 기반으로 실제 작업 계획 및 수립에 도움을 주기 위한 등급 G_t을 계산하는 함수의 역할을 수행한다.

한편, 본 논문에서는 학습 부담을 줄이고 학습 효율성 향상을 위해 계층적 서술자 규칙 설계를 제안한다. [Fig. 4]는 본 논문에서 제안하는 서술자 규칙 계층 구조이다. 서술자 규칙들은 정책 계층(policy layer), 정책 파라미터 계층(policy parameter layer), 상태 계층(state layer)의 총 3계층으로 구성되며, 각 계층에 따라 서술자 규칙의 결론부(head) 및 조건부(body)가 결정된다. 예컨대, 정책 계층의 안전성 규칙의 조건부에는 정책 파라미터 계층의 배터리 안전성, 충돌 안전성 등이 포함되며, 다시 배터리 안전성 규칙의 조건부에는 현재 로봇의 위치, 배터리 잔량 등이 포함된다. 이를 통해 불필요한 학습 파라미터 생성을 방지하며, 더 신속하고 올바른 정책 규칙 학습이 가능하다.

또한, 본 논문에서는 불필요한 시간과 자원 소모를 줄이고 효율적인 정책 규칙 학습을 위해 영역-고유 지식을 활용한다. 영역-고유 지식의 활용 방법은 영역-고유 지식을 토대로 정책 규칙들 중 일부를 미리 심층 관계형 강화 학습에 제공함으로써, 신규로 학습해야 할 규칙의 수를 제한해주고 새로운 규칙들에 학습을 집중하도록 학습의 효율성을 높여주는 방법이다. 이와같은 과정을 통해 예측된 로봇별 작업의 안전성과 효율성 평가값 G_t은 작업 관리자에게 전달되어 해당 작업을 가용 로봇들에게 동적으로 할당하는데 효과적으로 활용된다.



[Fig. 3] Deep relational reinforcement leaner

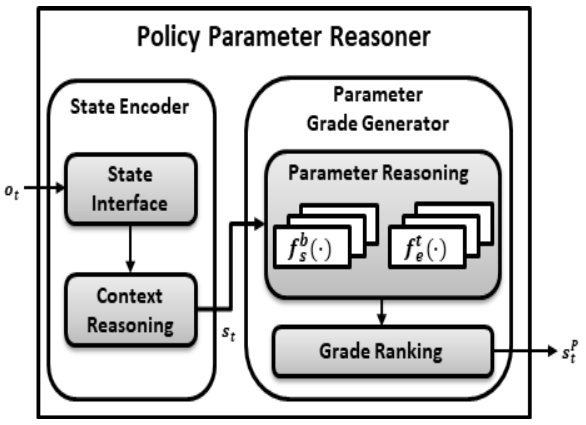


[Fig. 4] Hierarchical predicate rules

3.3 정책 파라미터 추론기

정책 파라미터 추론기는 본 논문에서 제안하는 작업 정책 학습기의 정책 파라미터 추론 단계를 담당한다. 정책 파라미터 추론 단계는 작업 환경으로부터 인식되는 관측값(observation) o_t으로부터 상태 인코더(state encoder), 파라미터 등급 생성기(parameter grade generator)를 거쳐 안전성과 효율성 규칙 학습을 위한 다양한 정책 파라미터(policy parameter) s_t^P를 생성한다. 한편, 심층 관계형 강화 학습기의 입력은 일차-술어 논리 형식으로 표현되며, 이로 인해 구문적 템플릿에 정의되는 서술자들의 개수에 따라 학습 파라미터가 크게 증가하는 문제가 있다.

본 논문에서는 이러한 한계를 극복하고자 함수기반 추론기를 채용한 정책 파라미터 추론기를 제안한다. [Fig. 5]는 본 논문에서 제안하는 정책 파라미터 추론기를 나타낸다. 먼저 상태 인코더는 다시 상태 인터페이스(state interface)와 파라미터 등급 생성기(parameter grade generator)로 나뉜다. 상태 인터페이스는 환경으로부터 입력되는 관측값(observation)을 전달받아 맥락 추론(context reasoning) 과정을 거쳐 기호(symbol)형



[Fig. 5] Policy parameter reasoner

태의 상태 서술자 s_t를 생성한다. 예컨대, 환경 내의 로봇에 대한 위치, 배터리 잔량 등과 같은 벡터로 표현되는 상태 관측값을 입력 받는다면 robotAt (R, L), batteryRemain (R, B)와 같은 상태 서술자를 생성한다. 이때, 서술자의 인자인 R은 로봇, L은 위치, B는 배터리 잔량이다. 생성된 상태 서술자 s_t 는 정책 파라미터 생성을 위해 파라미터 등급 생성기에 전달된다. 정책 파라미터 생성기는 각 상태 서술자들을 입력으로 삼아 추론 함수 f 들을 거쳐 파라미터 평가값을 계산한다. 이후 등급 랭킹(grade ranking) 과정을 거쳐 평가값으로부터 정책 파라미터 등급을 산정한다. 예를 들면, [Fig. 1]의 작업 환경 상태에서 안전성(safety) 규칙을 학습할 때, 고려할 정책 파라미터 중 하나인 배터리 측면의 안전성(battery safety) 등급을 추론하는 과정은 이동 적재 작업 수행 시 필요한 배터리 잔량에 대한 값인 batteryNeed (“amr01”, “transport (“cargo01”, “station04”)”, “57”)와 현재 로봇의 배터리 잔량에 대한 상태 서술자 batteryRemain (“amr01”, “20”)을 입력받아 추론 함수인 f_s^b를 거쳐 정책 파라미터 s_t^p 중 하나인 batterySafetyGrade-2 (“robot“robot1”, “transport (“cargo01”, “station04”)”)를 생성한다. 이때, -2는 안전성 등급을 나타낸다. 심층 관계 학습기에 이와 같은 정책 파라미터를 입력으로 학습 및 추론 과정을 거친다면 보다 효율적이고 안정적인 학습 및 추론을 할 수 있다.

4. 구현 및 실험

4.1 실험 환경 및 구현

본 논문에서 제안하는 작업 정책 학습기를 활용한 안전성, 효율성 규칙 학습 성능 검증을 위해 NVIDIA에서 제공하는 Isaac 시뮬레이터를 이용해 [Fig. 1]과 같은 제조 물류 실험 환경을 구축하였다. 환경 내에는 제조 물류 작업을 위한 모바일 이동 로봇 2대, 적재할 화물들과 이동 장애물들을 배치하였으며, 모바일 이동 로봇은 정해진 정점(vertex) 위를 직진(move) 혹은 회전(turn)하는 행위와 화물을 싣고(load) 내려놓는(unload) 행위를 환경 내에서 수행한다. 규칙 학습과 추론 검증에 이용된 목표 작업은 모바일 이동 로봇이 화물을 적재할 위치에 운송(transport)하는 작업이다. 특히 규칙 학습 시엔 무작위로 초기화된 환경 내의 위치, 배터리 잔량 등과 같은 모바일 이동 로봇의 상태들로부터 안전성, 효율성 규칙을 적용하여 작업 할당을 통해 실행한 작업을 평가하여 규칙을 학습한다. 심층 관계형 강화 학습기를 이용한 규칙 학습에는 Adam Optimizer 최적화 알고리즘(optimizer)을 사용하였으며, 학습률(learning rate)은 0.01, 비평가의 감가율(γ)은 0.9로 설정하였다. 규칙 학습 및 성능 비교 실험들은 Intel Xeon Silver 4210R CPU와 NVIDA RTX A5000 GPU 3개가 장착된 하드웨어 환경, Ubuntu 20.04, python 3.6, Tensorflow 1.15의 소프트웨어 환경에서 수행되었다.

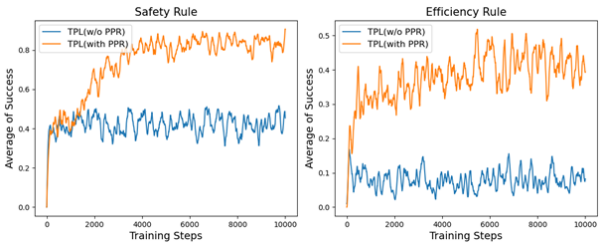
4.2 성능 비교 실험

학습된 안전성과 효율성 규칙 학습에 대한 첫 번째 비교 실험은 규칙 학습 시 정책 파라미터 추론기(Policy Parameter Reasoner, PPR)가 작업 정책 학습기(Task Policy Learner, TPL) 학습에 미치는 효과를 분석하기 위한 실험이다. 비교 실험에서는 정책 파라미터 추론기를 이용하여 추론된 배터리, 작업 시간 측면을 고려한 등급 서술자인 정책 파라미터 서술자를 사용하는 경우와 상태 서술자들로부터 직접 규칙을 학습하는 경우에 대해 비교하였다. 이 실험에서는 최근 100회의 규칙 추론 결과에 대한 평균 성공률(average of success rate)을 성능 척도로 사용하였다. 실험 결과는 [Fig. 6]과 같다.

[Fig. 6]에 학습 그래프에서 볼 수 있듯이, 안전성과 효율성 규칙 학습 모두 본 논문에서 제안한 작업 정책 학습기에 정책 파라미터 추론기를 사용한 것(with PPR)이 사용하지 않은 것(w/o PPR)보다 더 빠르게 수렴함과 동시에 더 높은 성공률을 보였다. 이러한 결과는 작업 정책 파라미터가 규칙을 학습 시 작업 정책 학습을 통해 알아내야 할 학습 파라미터를 줄여 더 안정적이고 효율적인 학습을 돕는 것을 확인할 수 있었다.

안전성과 효율성 규칙 학습에 대한 두 번째 비교 실험은 계층적 서술자 구조가 학습 성능 향상에 미치는 긍정적 효과를 확인하기 위한 실험이다. 이 실험에서는 [Fig. 4]와 같이 본 논문에서 제안한 계층적 구조로 서술자 규칙들을 표현할 때와 모든 서술자 규칙을 계층 구조 없이 하나의 단일 계층으로 표현할 때의 학습 성능을 서로 비교하였다. 이 실험에서도 앞선 실험과 마찬가지의 평균 성공률을 성능 측도로 사용하였다.

이 실험의 결과는 [Fig. 7]과 같다. 실험 결과를 살펴보면, 안전성과 효율성 규칙 학습에서 본 논문에서 제안하는 계층적 구조로 설계된 서술자 규칙들(with hierarchy)이 그렇지 않은 경우(w/o hierarchy)보다 더 높은 성공률을 보임을 확인할 수 있었다. 특히, 효율성 규칙보다 비교적 학습해야 할 파라미터 수가 많은 안전성 규칙의 경우에 더 큰 성능 차이를 보이며, 이를 통해 본 논문에서 제안하는 계층적 서술자 규칙 구조가 정



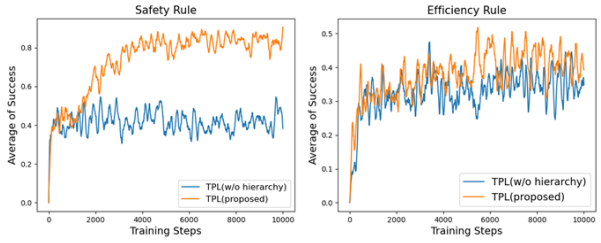
[Fig. 6] Performance evaluation of policy parameter reasoner

책 규칙들을 더 신속하고 올바르게 학습하는데 큰 도움을 줄 수 있음을 확인할 수 있었다.

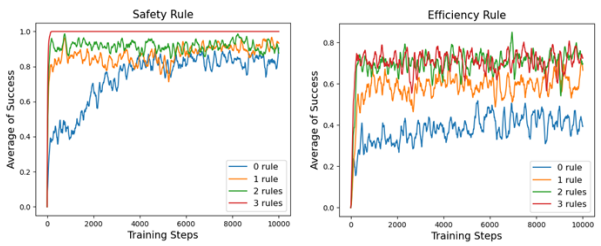
세 번째 실험은 전문가의 영역-고유 지식을 반영해 사전에 정의된 정책 규칙(predefined rule)들을 학습에 부분적으로 활용하는 것이 학습 효율에 미치는 긍정적 효과를 분석하기 위한 실험이다. 이 실험에서는 안전성과 효율성 등급을 판정하는 사전 정의된 정책 규칙들 중에 1가지 규칙(grade 1)을 활용하는 경우, 2가지 규칙(grade 1, 2)을 활용하는 경우, 3가지 규칙(grade 1, 2, 3)을 활용하는 경우, 사전 정의된 규칙을 전혀 활용하지 않는 경우 등 총 4가지 경우를 서로 비교하였다. 실험에서는 다른 실험들과 같이 총 100회의 규칙 추론 결과에 대한 평균 성공률을 성능 척도로 사용하였다.

[Fig. 8]은 이 실험의 결과를 그래프로 나타낸 것이다. [Fig. 8]의 그래프에서 확인할 수 있듯이, 영역-고유 지식을 바탕으로 사전에 정의된 규칙들을 활용한 경우들이 그렇지 않은 경우(0 rule)보다 더 빠른 학습 수렴 속도를 보여준 것을 확인할 수 있다. 또 사전 정의된 규칙들을 점점 더 많이 사용할수록 신규로 학습해야 할 규칙의 수는 감소하기 때문에 학습 수렴 시간이 훨씬 단축되었고, 신규 규칙 학습에만 계산을 더 집중할 수 있어 성공률이 더 높은 양질의 규칙들을 학습할 수 있었다.

다음 실험은 학습된 작업의 안전성, 효율성 정책 규칙들을 중심으로 본 논문에서 제안하는 작업 정책 학습기의 성능을 정성적으로 평가해보는 실험이다. 이 실험에서 작업의 안전성 예측을 위해서는 배터리 안전성(B_safety), 충돌 안전성(C_safety), 적재 안전성(S_safety) 등을, 작업의 효율성 예측을 위해서는 시간 효율성(T_efficiency), 자원 효율성(R_efficiency) 등을 정책 파라미터들로 삼아, 개별 평가 함수를 통해 각각 -2~2의 5가지 레벨로 파라미터 등급을 추론하였다. 최종적으로 심층 관계형 강화학습을 통해 학습된 안전성, 효율성 정책 규칙들은 [Table 1]과 같다.



[Fig. 7] Performance evaluation of hierarchical predicate rules



[Fig. 8] Performance evaluation of using domain-specific predefined rules

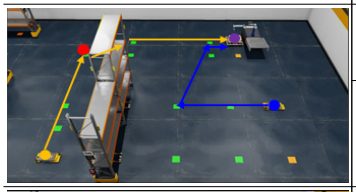
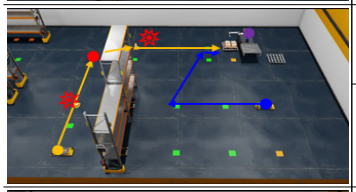
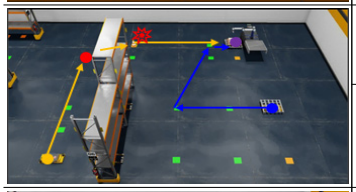
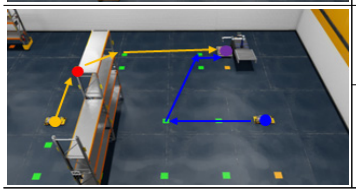
규칙 결론부(head)인 정책 서술자(policy predicate)도 각각 -2~2의 5가지 등급의 안전성, 효율성을 나타내며, 규칙 조건(condition)들은 다면적으로 평가한 정책 파라미터들의 평가값이다. [Table 1]의 학습된 관계형 정책 규칙들은 기존의 심층 신경망이나 강화학습으로 학습하는 벡터 형태의 규칙들에 비해 매우 직관적으로 의미 해석이 가능하며 적용 가능 범위도 더 넓다, 또 필요하다면, 학습된 규칙들을 설계자나 전문가가 쉽게 수정 보완하는 것도 가능하다.

다음은 시뮬레이션 작업 환경을 이용해 [Table 1]의 학습된 안전성과 효율성 정책 규칙들의 유효성을 검증해보았다. 이 실험에서는 동적으로 변화하는 작업 환경 상태마다 학습된 정책 규칙들을 적용하여 작업의 안전성과 효율성 등급을 예측하였고, [Table 2]는 해당의 실험 결과들을 나타낸다. Isaac 시뮬레이터로 구현한 작업 환경에는 총 2대의 작업 로봇(robot 1, robot 2)이 등장하며 동일한 작업에 대해 환경 상태 변화에 따라 해당 작업에 관한 각 로봇의 안전성과 효율성 등급을 예측해보았다. [Table 2]의 첫 번째 상태의 경우, 거리가 멀어 작업 소요 시간이 오래 걸리고 배터리 잔량도 적은 상태인 robot 2에 비해 동일 작업에 대한 robot 1의 안전성, 효율성 등급이 더 높게 추론되었다. 두 번째 상태의 경우도 robot 2의 배터리 잔량은 충분하지만, 이동 경로에 2번의 충돌 가능성(conflict)이 있기 때문에 해당 작업에 관해 robot 1보다 더 낮은 안전성 등급을 예측하였으며, 화물까지 이동해야 할 거리가 더 멀어 작업 시간 더 오래 걸릴 것으로 예상되는 robot 2의 효율성 등급이 robot 1보다 더 낮게 예측되었다. 세 번째 상태의 경우, robot 1은 이미 다른 화물을 싣고 있는 상황에서 해당 작업을 위해 추가 화물을 적재하면 안정성(stability)에 문제가 발생할 수 있어, 안전성 등급을 robot 2보다 낮게 예측하였다. 또한, robot 1은 자원(resource) 소모 측면에서도 robot 2에 비해 비-효율적이어서 더 낮은 효율성 등급을 예측하였다.

[Table 1] Learned safety and efficiency rules

Policy Predicates	Conditions
safety_2(R, T)	B_safety_2(R, T) ∧ C_safety_2(R, T) ∧ S_safety_1(R, T) ∧ ...
efficiency_2(R, T)	T_efficiency_2(R, T) ∧ R_efficiency_2(R, T) ∧ ...
safety_1(R, T)	B_safety_1(R, T) ∧ C_safety_0(R, T) ∧ S_safety_1(R, T) ∧ ...
...	...

[Table 2] Inference results based on learned rules

Environment		Policy Parameters					Inference Results	
		safety			efficiency		safety	efficiency
		battery	conflict	stability	time	resource		
	robot 1 (blue)	Grade 2	Grade 2	Grade 0	Grade 2	Grade 2	Grade 1	Grade 2
	robot 2 (yellow)	Grade -2	Grade 2	Grade 0	Grade -1	Grade -2	Grade 0	Grade -1
	robot 1 (blue)	Grade 2	Grade 2	Grade 0	Grade 2	Grade 2	Grade 1	Grade 2
	robot 2 (yellow)	Grade 1	Grade 0	Grade 0	Grade -1	Grade 1	Grade 0	Grade 0
	robot 1 (blue)	Grade -1	Grade 2	Grade -2	Grade 2	Grade -2	Grade 0	Grade 0
	robot 2 (yellow)	Grade 1	Grade 1	Grade 0	Grade -1	Grade 2	Grade 1	Grade 1
	robot 1 (blue)	Grade -1	Grade 2	Grade 0	Grade 2	Grade 1	Grade 0	Grade 1
	robot 2 (yellow)	Grade -1	Grade 2	Grade 0	Grade 2	Grade 1	Grade 0	Grade 1

마지막으로 매우 희소하게 발생하는 네 번째 상태의 경우는 robot1과 robot2에 관해 추론된 정책 파라미터들의 등급이 모두 같아서, 최종적으로 안전성 등급과 효율성 등급도 모두 동일하게 예측된 경우이다. 이와 같은 경우에는 어느 로봇에게 작업을 할당해야 될지 판정하기 어려워진다. 이러한 문제점을 해결하기 위해서는 현재보다 더 다양한 정책 파라미터들과 학습 규칙들이 추가적으로 개발되어야 할 필요가 있다고 판단된다.

5. 결 론

본 논문에서는 복수의 로봇들이 함께 협업하는 제조 물류 작업 환경에서 로봇의 상태에 따른 동적 작업 할당을 위해 로봇별 작업의 안전성 및 효율성을 예측하는 작업 정책 학습기를 제안하였다. 제안 시스템은 학습 결과물들의 높은 해석 가능성과 일반성을 가진 심층 관계형 강화학습 프레임워크인 dNL-RRL을 기초로, 로봇별 작업의 안전성과 효율성 등급을 예측하는 정책 규칙들을 학습한다. 또한, 작업 환경의 복잡성과 실시간성을 고려하여 전문가의 영역-고유 지식들을 정책 규칙 학습에 효과적으로 활용하는 방안들도 제시하였다. NVIDIA의 Isaac 로봇 시뮬레이터를 이용한 가상의 제조 물류 환경에서 진행한 다양한 실험들을 통해, 본 논문에서는 제안한 작업 정책 학습기의 유효성과 우수성을 확인하였다.

계획하고 있는 향후 연구로는 더 세밀하고 다면적인 작업의 안전성, 효율성 예측이 가능하도록 다양한 정책 파라미터들을 추가하여 현재의 정책 파라미터 추론기를 확장하는 연구와 실시간성을 만족하기 위해 작업 정책 학습기를 미리 사전 학습(pretraining)시키기는 연구, 개발된 작업 정책 학습기를 하드웨어 로봇들을 포함한 실제 제조 작업 환경에 적용하여 성능을 검증하고 기능을 고도화하는 연구 등이 있다.

References

[1] E. Tunstel, A. Howard, and H. Seraji, “Rule-based reasoning and neural network perception for safe off-road robot mobility,” *Expert Systems*, vol. 19, no. 4, pp. 191-200, 2002, DOI: 10.1111/1468-0394.00204.

[2] A. Kreutzmann, D. Wolter, F. Dylla, and J. H. Lee, “Towards Safe Navigation by Formalizing Navigation Rules,” *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, vol. 7, no. 2, pp. 161-168, Jun, 2013, DOI: 10.12716/1001.07.02.01.

[3] Z. Jiang and S. Luo, “Neural Logic Reinforcement Learning,” *Machine Learning*, 2019, DOI: 10.48550/arXiv.1904.10729.
[4] A. Payani and F. Fekri, “Incorporating Relational Background Knowledge into Reinforcement Learning via Differentiable Inductive Logic Programming,” *Machine Learning*, 2020, DOI: 10.48550/arXiv.2003.10386.
[5] M. Rojas, G. Hermosilla, D. Yunge, and G. Faris, “An Easy to Use Deep Reinforcement Learning Library for AI Mobile Robots in Isaac Sim,” *Applied Sciences*, vol. 12, no. 17, 2022, DOI: 10.3390/app12178429.
[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riediller, “Playing atari with deep reinforcement learning,” *Machine Learning*, 2013, DOI: 10.48550/arXiv.1312.5602.
[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, “Proximal Policy Optimization Algorithms,” *Machine Learning*, 2017, DOI: 10.48550/arXiv.1707.06347.
[8] R. Evans and E. Grefenstette, “Learning Explanatory Rules from Noisy Data,” *Journal of Artificial Intelligence Research*, vol. 61, Jan, 2018, DOI: 10.1613/jair.5714.
[9] A. Payani and F. Fekri, “Inductive Logic Programming via Differentiable Deep Neural Logic Networks,” *Artificial Intelligence*, 2019, DOI: 10.48550/arXiv.1906.03523.
[10] O. Rivlin, T. Hazan, and E. Karpas, “Generalized Planning With Deep Reinforcement Learning,” *Artificial Intelligence*, 2020, DOI: 10.48550/arXiv.2005.02305.
[11] V. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. Reichert, T. Lillicrap, E. Lockhart, M. Shanahan, V. Langston, R. Pascanu, M. Botvinick, O. Vinyals, and P. Battaglia, “Relational Deep Reinforcement Learning,” *Machine Learning*, 2018, DOI: 10.48550/arXiv.1806.01830.
[12] J. Janisch, T. Pevný and V. Lisý, “Symbolic Relational Deep Reinforcement Learning based on Graph Neural Networks,” *Machine Learning*, 2021, DOI: 10.48550/arXiv.2009.12462.
[13] S. Garg and A. Bajpai, “Symbolic Network: Generalized Neural Policies for Relational MDPs,” *the 37th International conference on machine learning*, 2020, [Online] https://proceedings.mlr.press/v119/garg20a.html.
[14] D. Adjodah, T. Klinger, and J. Joseph, “Symbolic Relation Networks for Reinforcement Learning,” *32nd Conference on Neural Information Processing Systems (NIPS 2018)*, Montréal, Canada, 2018, [Online] https://r2learning.github.io/assets/papers/CameraReadySubmission%203.pdf.
[15] S. Das, S. Natarajan, K. Roy, R. Parr, and K. Kersting, “Fitted Q-Learning for Relational Domains,” *Machine Learning*, 2020, DOI: 10.48550/arXiv.2006.05595.
[16] T. Gokhale, S. Sampat, Z. Fang, Y. Yang, and C. Baral, “Blocksworld Revisited: Learning and Reasoning to Generate Event-Sequences from Image Pairs,” *Computer Vision and Pattern Recognition*, 2019, DOI : 10.48550/arXiv.1905.12042.

강 민 교

2020 경기대학교 컴퓨터과학과(학사)
2020~현재 경기대학교 컴퓨터과학과(석사)

관심분야: 인공지능, 지능로봇, 지식 표현 및 추론

김 인 철

1985 서울대학교 수학과(학사)
1987 서울대학교 전산과학과(석사)
1995 서울대학교 전산과학과(이학박사)
1996~현재 경기대학교 컴퓨터공학부 교수

관심분야: 인공지능, 지능로봇, 지식 표현 및 추론