KIPS Trans. Softw. and Data Eng.
Vol.11, No.3 pp.141~148
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.3.141

Effective Utilization of Domain Knowledge for Relational Reinforcement Learning

MinKyo Kang† · InCheol Kim††

ABSTRACT

Recently, reinforcement learning combined with deep neural network technology has achieved remarkable success in various fields such as board games such as Go and chess, computer games such as Atari and StartCraft, and robot object manipulation tasks. However, such deep reinforcement learning describes states, actions, and policies in vector representation. Therefore, the existing deep reinforcement learning has some limitations in generality and interpretability of the learned policy, and it is difficult to effectively incorporate domain knowledge into policy learning. On the other hand, dNL-RRL, a new relational reinforcement learning framework proposed to solve these problems, uses a kind of vector representation for sensor input data and lower-level motion control as in the existing deep reinforcement learning. However, for states, actions, and learned policies, It uses a relational representation with logic predicates and rules. In this paper, we present dNL-RRL-based policy learning for transportation mobile robots in a manufacturing environment. In particular, this study proposes a effective method to utilize the prior domain knowledge of human experts to improve the efficiency of relational reinforcement learning. Through various experiments, we demonstrate the performance improvement of the relational reinforcement learning by using domain knowledge as proposed in this paper.

Keywords : Relational Reinforcement Learning, Domain Knowledge, Policy, Logic Predicate, Generality, Interpretability

관계형 강화 학습을 위한 도메인 지식의 효과적인 활용

강 민 교† · 김 인 철††

요 약

최근 들어 강화 학습은 심층 신경망 기술과 결합되어 바둑, 체스와 같은 보드 게임, Atari, StartCraft와 같은 컴퓨터 게임, 로봇 물체 조작 작업 등과 같은 다양한 분야에서 매우 놀라운 성공을 거두었다. 하지만 이러한 심층 강화 학습은 행동, 상태, 정책 등을 모두 벡터 형태로 표현한다. 따라서 기존의 심층 강화 학습은 학습된 정책의 해석 가능성과 일반성에 제한이 있고, 도메인 지식을 학습에 효과적으로 활용하기도 어렵다는 한계성이 있다. 이러한 한계점들을 해결하기 위해 제안된 새로운 관계형 강화 학습 프레임워크인 dNL-RRL은 센서 입력 데이터와 행동 실행 제어는 기존의 심층 강화 학습과 마찬가지로 벡터 표현을 이용하지만, 행동, 상태, 그리고 학습된 정책은 모두 논리 서술자와 규칙들로 나타내는 관계형 표현을 이용한다. 본 논문에서는 dNL-RRL 관계형 강화 학습 프레임워크를 이용하여 제조 환경 내에서 운송용 모바일 로봇을 위한 행동 정책 학습을 수행하는 효과적인 방법을 제시한다. 특히 본 연구에서는 관계형 강화 학습의 효율성을 높이기 위해, 인간 전문가의 사전 도메인 지식을 활용하는 방안들을 제안한다. 여러 가지 실험들을 통해, 본 논문에서 제안하는 도메인 지식을 활용한 관계형 강화 학습 프레임워크의 성능 개선 효과를 입증한다.

키워드 : 관계형 강화 학습, 도메인 지식, 행동 정책, 논리 서술자, 일반성, 해석 가능성

1. 서 론

최근 들어 신경망(neural network)을 이용하는 심층 강화 학습(deep reinforcement learning) 기술은 Starcraft, Atari 등과 같은 PC 게임 분야뿐만 아니라, 지능형 서비스 로봇, 자율 주행 자동차 분야에 이르기까지 폭넓게 활용되어 오고 있다. 하지만 기존의 심층 강화 학습은 정책(policy), 상태(state), 행동(action) 등을 모두 벡터 형태로 표현하는 강화 학습으로서, 학습된 정책의 해석 가능성(interpretability)과 일반성(generality)에 제한이 있고 도메인 지식(domain knowledge)을 학습에 효과적으로 활용하기도 어렵다는 한계성을 갖는다.

이러한 문제점들을 극복하고자, 정책, 상태, 행동 등을 논리 서술자(logic predicate) 기반의 논리식(logic expression)으로 표현하는 NLM(Neural Logic Machine)[1], RDRL(Relational Deep Reinforcement Learning)[2], NLRL(Neural Logic Reinforcement Learning)[3], dNL-RRL(differentiable

※ 정보통신기획평가원 / 정보통신방송 기술개발사업 / 클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발 / 2020-0-00096
※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회에서 “효율적인 관계형 강화학습을 위한 사전 영역 지식의 활용”의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 종신회원 : 경기대학교 컴퓨터과학과 교수
Manuscript Received : July 14, 2021
Accepted : August 29, 2021
* Corresponding Author : InCheol Kim(kic@kyonggi.ac.kr)

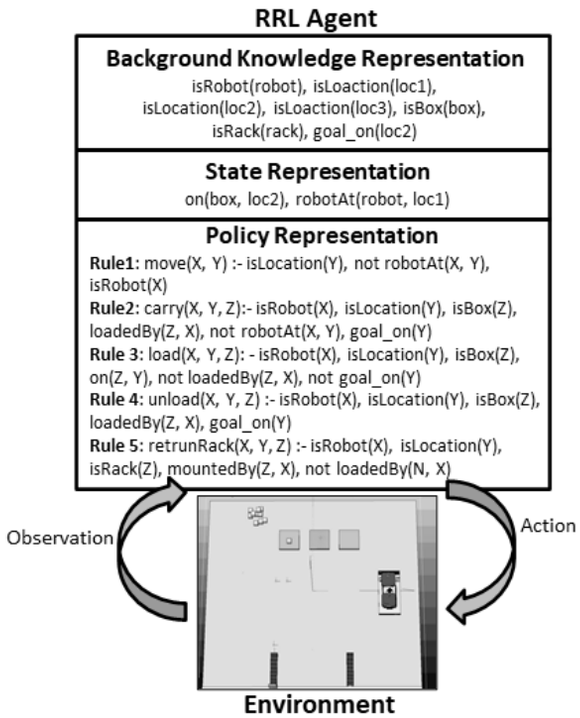


Fig. 1. Example of Relational Reinforcement Learning

Neural Logic-Relational Reinforcement Learning)[4] 등과 같은 여러 관계형 강화 학습(relational reinforcement learning) 방법들이 제안되었다.

이들 중 특히 관계형 강화 학습 프레임워크인 dNL-RRL은 미분 가능한 귀납적 논리 프로그래밍(differentiable Inductive Logic Programming, dNL-ILP)[5] 엔진을 채용함으로써, 관계형 강화 학습 프레임워크 전체에 대해 종단간 학습(end-to-end training)이 가능하고, 비-기호 형태인 센서 데이터과 입력 영상, 그리고 출력 제어 신호도 처리 가능하다는 장점이 있다. 이를 통해 다양한 응용 분야들에서 활용 가능성이 높다.

Fig. 1은 관계형 강화 학습 프레임워크인 dNL-RRL 에서 표현할 수 있는 상태(state), 배경 지식(background knowledge), 학습되는 행동 정책(action policy) 의 예시를 나타낸다. Fig. 1은 본 논문에서 다루는 응용 분야인 물류 공장 환경에서 작동하는 운송용 모바일 로봇의 행동 정책을 학습하는 예를 보여준다. Fig. 1에서 볼 수 있듯이, dNL-RRL과 같은 관계형 강화 학습 프레임워크에서는 변화하는 상태와 배경 지식이 모두 논리 서술자(logic predicate)들의 집합으로 표현되고, 학습되는 행동 정책은 각 행동에 관한 논리 서술자 규칙(logic rule)들로 표현된다. 예컨대, 행동 move(X, Y)에 관한 정책은 isRobot(X), not robotAt(X, Y), isLocation(Y)와 같이 변수를 포함하는 논리 서술자들을 규칙 조건부로 갖도록 표현된다. 그러므로 이와 같은 논리 서술자 규칙 형태의 정책들은 매우 높은 해석 가능성과 일반성을 갖는다.

본 논문에서는 대표적인 관계형 강화 학습 프레임워크인 dNL-RRL을 기초로 물류 공장 환경 내 로봇의 제어를 위한 행동 정책 학습을 하였으며, 학습의 효율성을 높이기 위해 인간 전문가(human expert)의 사전 도메인 지식(prior domain knowledge)을 활용하는 다양한 방안들을 제안한다. 관계형 강화 학습 프레임워크인 dNL-RRL 에서는 여러가지 방법으로 인간의 도메인 지식을 활용할 수 있다.

본 논문에서는 (1) 학습하여야 하는 각 행동 규칙(action rule)에 반드시 배제되어야 하는 상태 조건들(excluded conditions)과 반드시 포함되어야 하는 상태 조건들(included conditions)을 미리 정의해주는 방식, (2) 학습하여야 하는 행동 규칙들의 일부를 미리 정의해주는 방식(predefined rules) 등의 2 가지 사전 도메인 지식 활용 방법들을 제안한다. CopelliaSim 로봇 시뮬레이터로 구현한 가상의 물류 공장 환경과 운송용 모바일 로봇을 이용한 다양한 실험들을 통해, 본 논문에서 제안하는 도메인 지식 기반의 관계형 강화 학습 방법의 학습 효율성 개선 효과 및 정책의 해석 가능성(interpretability), 일반성(generality) 등을 분석해본다.

2. 관련 연구

관계형 강화 학습(Relational Reinforcement Learning, RRL)은 일차 술어 논리(First-Order predicate Logic, FOL) 형태의 상태 및 행동 표현으로 인해 좋은 일반화 능력과 해석 가능성을 갖는다. 이러한 관계형 강화 학습의 대표적인 연구 중 NLRL(Neural Logic Reinforcement Learning)[3]은 학습 가능한 순환 논리 머신(Differential Recurrent Logic Machine, DRLM)을 통하여 모든 구체환된 서술자들의 가치평가 벡터를 입력받고 n-단계 추론을 수행하여 구체화된 행동 서술자들에 대한 가치평가 값을 계산한다. 이를 통해 얻은 가치 평가 값은 구체화된 행동들의 확률 분포이며, 이 확률 분포를 이용하여 확률적으로 행동을 선택한다. NLRL은 상태-행동공간과 정책 공간을 모두 논리 서술자와 규칙으로 표현함으로써 완전한 설명 가능성을 보장하지만, 조합 가능한 모든 상태-행동공간과 정책 공간의 생성으로 인해 계산 복잡도가 높아지고 확장성이 크게 떨어지는 한계성이 있다.

반면에, dNL-RRL[4] 관계형 강화 학습 프레임워크는 상태, 행동, 정책을 모두 논리 서술자로 표현하여 완전한 설명 가능성을 보장한다. 또 dNL-RRL은 미분 가능한 귀납적 논리 프로그래밍(differentiable Inductive Logic Programming)[5]을 통하여 논리 곱, 논리 합 계층의 인공 신경망을 통하여 학습하여 확장성을 높였다. dNL-RRL 관계형 강화 학습 프레임워크는 3장에서 더 자세히 다룬다.

이 밖에도 최근에는 다양한 형태의 관계형 강화 학습 방법들이 제안되었다. 이들 중 그래프 기반의 상태 표현(graphical state representation)을 이용하는 관계형 강화 학습 모델들로는 SR-DRL[6], SYMNET[7], GBFS-GNN[8], RDRL[2] 등이 있다. 또, 관계 유추가 필요한 도메인으로부터 관계형 상태 표현(relational state representation)을 입력받는

RFQL[9], SRN[10] 등도 있다. 또한, 한 쌍의 이미지들로부터 논리 서술자로 표현된 행동 시퀀스(event sequence)를 구하는 Gokhale[11]와 관계형 인공 신경망을 모델 학습에 이용하는 A3SL[12], Fair-A3SL[13], Skinner[14] 등도 제안되었다.

3. 사전 도메인 지식 기반 관계형 강화 학습

3.1 관계형 강화 학습 프레임워크

본 논문에서는 대표적인 관계형 강화 학습 중 하나인 dNL-RRL 프레임워크를 기초로, 사전 도메인 지식 활용 방법들을 제안한다. dNL-RRL 관계형 강화 학습 프레임워크는 Fig. 2와 같이, 행동자-비평가(actor-critic) 구조를 따른다. 행동자(actor)는 환경에서의 경험 데이터(experience data)들을 토대로 행동 정책을 구성하는 논리 규칙들을 학습(policy learning)하고, 이 학습된 행동 규칙들에 따라 정책을 추론함(policy inference)으로써 매 순간 학습 에이전트가 실행해야 할 행동을 결정(action decision)하는 역할을 수행한다. 반면에, 비평자(critic)는 경험 데이터와 보상(reward)을 토대로 행동 가치 함수 Q를 학습하고, 이를 토대로 행동자가 정한 행동들을 평가함으로써, 행동자가 신속하게 올바른 정책 갱신에 도움을 주는 역할을 수행한다.

한편, 행동 결정과 정책 학습을 담당하는 행동자(actor)는 다시 상태 인코더(state encoder), 미분 가능한 뉴로-로직 귀납적 논리 프로그래밍 엔진(differentiable Neural-Logic Inductive Logic Programming, dNL-ILP), 행동 디코더(action decoder)들로 구성된다. 상태 인코더는 환경(environment)으로부터 관측 데이터(observation)를 입력받아, 해당 관측 상태(state)에서 만족되는 논리 서술자들(logic predicates)을 생성함으로써 상태를 하나의 일차-술어논리(First-Order predicate Logic, FOL) 형태로 표현하는 역할을 담당한다. 미분 가능한 뉴로-로직 귀납적 논리 프로그래밍 엔진(dNL-ILP)은 논리 서술자들 기반의 상태 표현과 이전에 실행한 행동에 관한 평가자(critic)의 가치 평가를 토대로 행동 규칙(action rule)들을 학습하기도 하고, 앞서 학습된 행동 규칙들을 토대로 n 단계의 전향 추론(n-step forward chaining)을 통해 학습 에이전트가 실행할 행동을 결정하기도 한다.

행동 디코더는 실행이 결정된 서술자 형태의 행동을 실제 환경 안에서 에이전트가 실행할 수 있도록 하는 제어 함수의 역할을 수행한다. 한편, 미분 가능한 뉴로-로직 귀납적 논리 프로그래밍 엔진(dNL-ILP)에서는 모든 일차-술어논리식을 곱의 합 표준형(Disjunctive Normal Form, DNF)으로 변환하여 표현한다. 따라서 하나의 행동 결론부(action head)와 다양한 상태 조건들(state conditions)로 구성된 각각의 행동 규칙도 모두 곱의 합 표준형(DNF)으로 표현한다. 그리고 이러한 임의의 곱의 합 표준형(DNF) 논리식을 논리합(logical disjunction, F_D)계층들과 논리곱(logical conjunction,

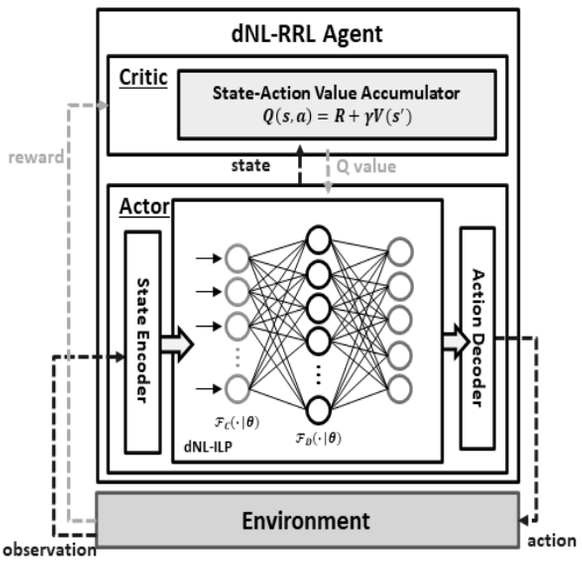


Fig. 2. Architecture of dNL-RRL, a Novel Relational Reinforcement Learning Framework

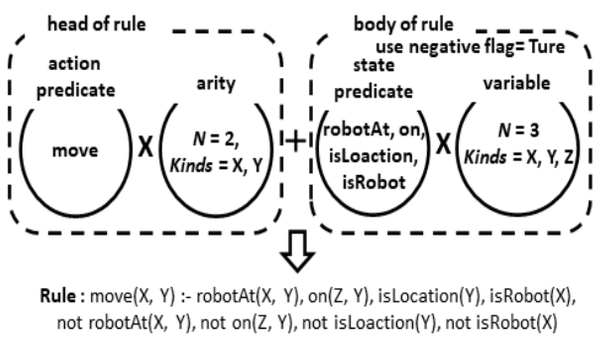


Fig. 3. Example of Action Rule Template

F_C)계층들로 구성된 인공 신경망으로 학습한다. 각 논리합, 논리곱 계층은 완전 연결 계층(fully connected layer)으로 구성되어 있으며, 각 계층의 입력 및 출력 크기는 사전 정의된 템플릿에 의하여 결정된다. 따라서 행동 규칙들을 논리합 계층들과 논리곱 계층들로 구성된 전체 신경망으로 학습하기 위해서, 첫째로 학습할 행동 규칙의 구성 및 특징을 나타내는 행동 규칙 템플릿(policy rule template)을 필요로 한다.

Fig. 3은 행동 규칙 move(X, Y)에 관한 행동 규칙 생성 템플릿을 보여준다. 이 템플릿에 의하면 각 행동 규칙은 크게 행동 결론부(head)와 상태 조건부(body)로 구성되며, 행동 규칙의 결론부는 다시 move(X, Y) 행동 서술자(action predicate)와 2개의 인자(arity)들로 구성되고 인자의 종류(kind)는 X와 Y이다. 또한, 규칙의 조건부는 robot_At, is_Location, is_Robot, on 등의 상태 서술자(state predicate)들이 사용될 수 있으며, 종류가 X, Y, Z인 3개의 변수(variable)들을 포함할 수 있다. 따라서 이러한 규칙 생성 템플릿에 따라, Fig. 3의 하단에 표시된 행동 move(X, Y)에 관한 행동 규칙을 학습시킬 수 있다.

3.2 행동 규칙 조건 제약

이전 절에서 다룬 내용대로, 관계형 강화 학습 프레임워크인 dNL-RRL 에서는 행동 규칙을 학습하기 위하여 사전에 정의해둔 규칙 생성 템플릿에 따라 가능한 모든 행동 규칙들을 생성한 후, 행동가-비평가(actor-critic) 기반의 심층 강화 학습(deep reinforcement learning)을 통해 행동 규칙을 표현하는 신경망의 가중치(weight)들을 학습한다. 대부분의 경우 이러한 템플릿에 의해 생성 가능한 후보 행동 규칙들(candidate action rules)의 수는 후속 학습 과정에 부담을 줄 정도로 매우 많이 생성된다. 그중에는 실제 환경에서는 동시에 성립하기 어려운 상태 조건식들이 포함된 규칙들이나 의미가 없는 규칙들도 다수 포함되어 있다. 학습이 충분히 이루어진 후에는 성립 불가능한 상태 조건식들의 조합이 해소될 수도 있으나, 해소되기 위해서는 수많은 계산 자원(computational resources)이 소모될 수 있다. 사전 정의된 규칙 템플릿을 사용하는 것 이외에, 해당 특정 도메인에 관한 인간 전문가의 사전 지식(prior knowledge)을 활용해 의미가 있는 후보 행동 규칙들만 생성될 수 있도록 미리 제한할 수 있다면, 이러한 문제들을 해결하고 학습 효율성을 크게 높일 수 있을 것이다.

관계형 강화 학습 프레임워크인 dNL-RRL에서 후보 행동 규칙들을 생성할 때 도메인 지식을 활용할 수 있는 제안 방법의 하나는 행동 규칙의 조건 제약(condition constraint)들을 이용하는 것이다. 행동 규칙 조건 제약에는 두 가지 종류가 있다. 첫째는 조건 포함 제약(condition inclusion constraint)으로서, 행동 규칙의 조건부(body)에 항상 특정 상태 조건식(state condition)들을 행동 규칙의 조건부에 포함하도록 하는 제약이다. 둘째는 조건 배제 제약(condition exclusion constraint)으로서, 행동 규칙의 조건부에 항상 특정 상태 조건식들을 행동 규칙 조건부에 포함되지 않도록 배제하는 제약이다.

Fig. 4는 행동 move(X, Y)를 위한 행동 규칙 학습에 조건 배제 제약과 조건 포함 제약을 적용하는 예를 보여주고 있다. Fig. 4의 상단에는 Fig. 3과 같이 규칙 생성 템플릿에 따라 생성된 행동 move(X, Y)의 초기 행동 규칙(initial action rule)을 나타내며, 이 행동 규칙에 조건 포함 제약인 not robotAt(X, Y)와 조건 배제 제약들인 not isRobot(X), not isLocation(Y), robotAt(X, Y) 등이 적용되어 Fig. 4의 하단과 같이 제약된 행동 규칙이 생성된다.

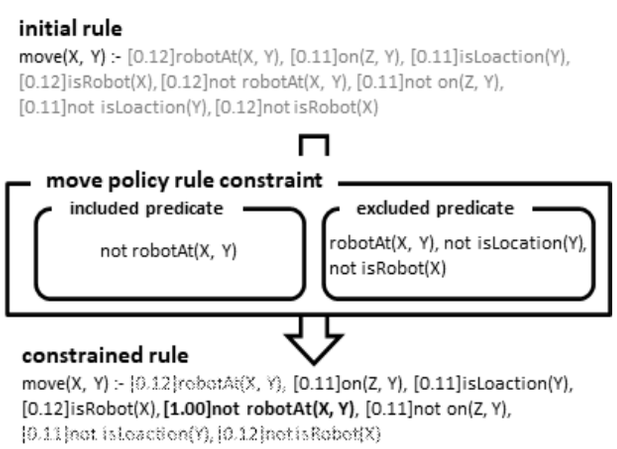


Fig. 4. Example of Rule Condition Constraint

3.3 사전 정의된 행동 규칙의 활용

인간 전문가의 사전 도메인 지식을 활용해 관계형 강화 학습 프레임워크인 dNL-RRL의 학습 효율성을 높일 수 있는 또 다른 방법은 이미 알고 있는 행동 규칙들의 일부를 행동 정책 학습이 본격적으로 시작되기 이전에 미리 제공하는 방법이다. 예를 들어, Fig. 1과 같은 물류 공장 내 운송용 모바일 로봇의 행동 정책은 carry, move, load, returnRack, unload의 총 5가지 행동 각각에 관한 규칙들로 구성된다. 따라서 이 로봇 에이전트의 경우 완전한 행동 정책을 습득하려면 이 5가지 행동 규칙들을 모두 학습해야 한다. 하지만 이 행동 규칙들 중 알고 있는 행동 규칙들이 존재한다면, 이러한 규칙까지 처음부터 학습할 필요가 없이 다른 행동들에 관한 행동 규칙 학습이 시작되기 전에 사전 행동 정책을 구성하는 행동 규칙 집합에 포함시켜줄 수 있을 것이다. 예를 들어, 위의 5가지 학습해야 할 행동들 중 사전 도메인 지식을 토대로 행동 load(X, Y, Z)에 관한 행동 규칙을 load(X, Y, Z):-isBox(Z), isRobot(X), isLocation(Y), on(Z, Y), robotAt(X, Y) 와 같이 사전에 정의해줄 수 있다면, 학습은 해당 행동 규칙을 제외한 나머지 4가지 행동들에 관한 행동 규칙들에만 집중할 수 있을 것이다. 이처럼 사전에 정의된 행동 규칙(predefined action rule)들을 사용한다면 꼭 필요한 행동 규칙들에만 학습을 집중함으로써 행동 정책을 학습하는데 소요되는 자원과 시간을 크게 절약할 수 있으며, 보다 실효성이 높은 행동 정책을 학습하기에도 도움이 될 것이다.

사전에 정의된 행동 규칙들은 학습을 통해 얻게되는 나머지 행동 규칙들과 함께 통합적으로 하나의 행동 정책(policy)을 형성하여, 입력 상태로부터 대처할 수 있는 에이전트의 행동을 결정하는데 사용된다. Fig. 5는 이와 같이 사전에 정의된 행동 규칙들과 학습된 행동 규칙들이 뉴로-로직 귀납적 논리 프로그래밍 엔진(dNL-ILP) 내부에서 행동 결정에 이용되는 과정을 나타낸다.

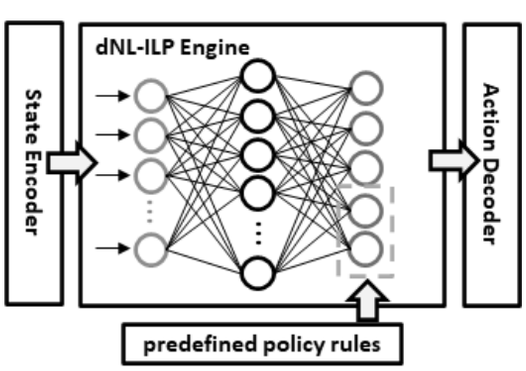


Fig. 5. Inference with Predefined Action Rules

4. 구현 및 실험

4.1 실험 환경 및 모델 학습

본 논문에서 제안하는 관계형 강화 학습의 사전 도메인 지식 활용 방법의 학습 효과 평가를 위해, CopelliaSim 로봇 시뮬레이터를 이용해 Fig. 6과 같은 물류 공장 내 운송용 모바일 로봇을 포함한 실험 환경을 구축하였다. 모바일 로봇의 전체적인 인식과 행동 제어는 로봇 지능 구조를 구성하는 AM(Action Manager), PM(Perception Manager), TM (Task Manager), CM(Context Manager) 모듈들 간의 실시간 메시지 기반의 상호 통신 작용에 의해 구현된다.

이 모듈들 중 PM은 로봇 환경으로부터 인식 정보(observation)를 취득하여 CM에게 전달하고, CM은 인식 정보들을 바탕으로 모듈 내부 추론을 통하여 상태 서술자들로 구성된 논리적 상태 서술자를 생성하여 TM에게 전달한다. 이 상태 표현을 토대로 TM은 관계형 강화 학습 프레임워크가 학습한 행동 정책에 의해 실행할 행동을 결정하여 AM에게 전달한다. AM은 해당 로봇의 행동을 로봇 환경 내에서 실제로 실행하는 역할을 수행한다. 특히 본 논문에서 다루는 관계형 강화 학습 프레임워크인 dNL-RRL은 로봇의 실시간 행동 결정을 담당하는 TM 내부에서 이용한다.

정책 학습과 테스트에 이용된 로봇 작업은 로봇이 물류를 정해진 목표 지점까지 옮기는 운송하는 작업이다. 초기 상태는 로봇이 충전 스테이션 위치에 있고 물류 박스는 선반 위에 올려진 상태로서 robotAt(robot, loc3), on(box, loc1)와 같이 표현되고, 목표 상태는 물류 박스는 컨베이어 벨트로 옮겨지고 로봇은 다시 충전 스테이션 위치로 복귀하는 상태로 robotAt(robot, loc3), on(box, loc2)와 같은 술어 논리식으로 표현된다. 관계형 강화 학습 프레임워크인 dNL-RRL의 학습을 위해 사용한 최적화 알고리즘(optimizer)은 Adam Optimizer를 사용했으며. 학습률(learning rate)은 0.025, 비평가의 감가율(γ)은 0.85로 설정하였다. 관계형 강화 학습 프레임워크의 학습과 성능 실험들은 Tensorflow 1.15, Python 3.6, Ubuntu 18.04 의 소프트웨어 환경, Geforce RTX 1080ti GPU, 128GB RAM 하드웨어 환경에서 실행하였다.

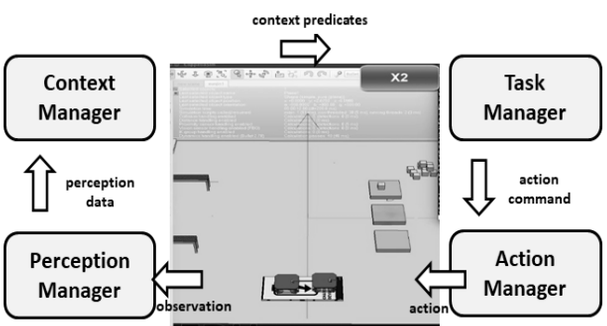


Fig. 6. Experimental Environment with Mobile Robots

4.2 성능 평가 실험

첫 번째 실험은 행동 규칙 조건 제약으로 표현되는 사전 도메인 지식이 학습 성능에 미치는 효과를 분석하기 위한 실험이다. 이 실험에서는 (1) 행동 규칙 조건 제약이 없는 경우(no constraint), (2) 행동 조건 배제 제약이 적용된 경우(excluded constraint) (unload와 load의 행동 규칙 조건부에 isRack과 mountedBy 행동 조건들을 배제), (3) 조건 포함 제약이 적용된 경우(included constraint) (retrunRack, unload, load의 행동 규칙 조건부에는 robotAt(X, Y) 조건을, carry와 move의 행동 규칙 조건부에는 not robotAt(X, Y) 조건을 반드시 포함) 의 총 3가지 서로 다른 경우를 비교 실험하였다. 이 실험에서는 총 100 episode의 평균 작업 성공률(success rate)을 성능 척도로 사용하였다. 실험 결과는 Fig. 7과 같다.

Fig. 7에서 볼 수 있듯이, 행동 규칙 조건 제약이 적용된 경우들이 적용하지 않은 경우에 비해 빠른 성능 향상을 보였다. 특히 본 실험에서는 조건 배제 제약을 적용한 경우에 비해 조건 포함 제약을 적용한 경우가 더 빠른 성능 향상을 보였다. 본 실험 도메인에서는 템플릿에 따라 생성되는 행동 규칙들에 불필요하거나 의미 없는 조건들이 많이 포함되지 않아서 해당 결과가 나타난 것으로 판단된다. Fig. 7의 실험 결과를 토대로, 본 논문에서 제안한 바와 같이 행동 규칙 조건 제약으로써 주어지는 사전 도메인 지식이 관계형 강화 학습의 학습 효율성과 작업 성공률을 높이는데 큰 효과가 있음을 확인할 수 있었다.

두 번째 실험은 사전 정의된 행동 규칙들로 표현되는 사전 도메인 지식이 학습 성능에 미치는 효과를 분석하기 위한 실험이다. 해당 실험에서는 (1) 사전 정의된 행동 규칙을 사용하지 않는 경우(0 rule), (2) 사전 정의된 carry 행동 규칙을 사용하는 경우(1 rule), (3) 사전 정의된 load, carry 행동 규칙들을 사용하는 경우(2 rules), (4) 사전 정의된 load, unload, carry 행동 규칙들을 사용하는 경우(3 rules)들을 서로 비교 실험하였다. 총 100 episode의 평균 작업 성공률을 성능 척도로 사용하였으며, 실험 결과는 Fig. 8과 같다.

Fig. 8에서 볼 수 있듯이, 사전 정의된 규칙들을 많이 활용할수록, 더 빠른 성능 향상을 보였다. 반면에, 사전 정의된 규

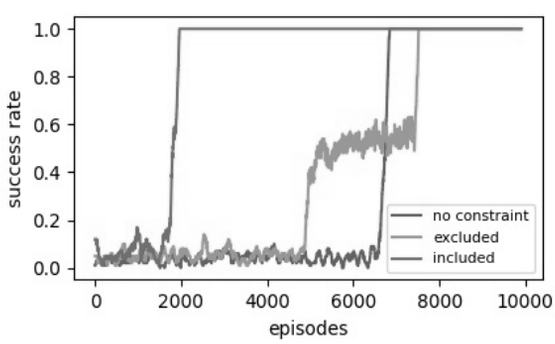


Fig. 7. Performance Evaluation of Rule Condition Constraints

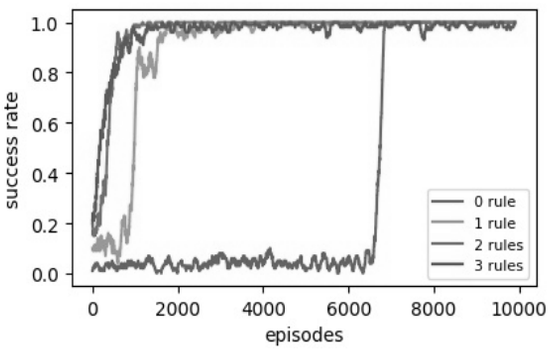


Fig. 8. Performance Evaluation of Predefined Rules

칙들을 활용하지 않은 경우가 동일한 성공률에 도달하기 위해서는 가장 긴 학습 시간을 요구했다. 이와 같은 결과를 토대로, 본 논문에서 제안한 바와 같이 사전 정의된 행동 규칙 형태로 도메인 지식을 활용하는 것이 관계형 강화 학습의 학습 효율성과 작업 성공률 향상에 도움을 줄 수 있음을 확인할 수 있었다.

세 번째 실험은 관계형 강화 학습 프레임워크 dNL-RRL을 통해 학습되는 행동 정책의 일반성(generality)과 해석 가능성(interpretability)을 확인하기 위한 실험이다. 관계형 강화 학습 프레임워크가 물류 환경에서 운송용 모바일 로봇을 위해 학습한 정책은 Table 1과 같다. 학습된 행동 규칙들의 조건부는 논리 서술자들로 표현되어 의미 해석이 용이하며, Fig. 1에 설명된 인간 전문가에 의해 사전 정의된 행동 규칙들과 높은 일치성을 보였다. 다음은 관계형 강화 학습 프레임워크인 dNL-RRL을 통해 학습되는 행동 정책의 일반성을 검증해보기 위한 비교 실험을 수행하였다. 이 실험에서는 행동 정책 학습을 위한 작업들의 에이전트 초기 상태들과는 다른 테스트 작업들의 에이전트 초기 상태들을 가정하고, 이와 같이 학습용 작업(training task)들과는 서로 다른 테스트 작업(test task)들에 대한 행동 정책의 작업 성공률을 비교해보았다. 학습용 작업들의 초기 상태(initial state)는 Table 2와 같이 로봇의 초기 위치의 X 좌표는 [2.85, 3.25] 범위 안에서, Y 좌표는 [1.85, 2.25] 범위 안에서 각각 임의의 실수 값으로 선택하여 결정하였다. 그리고 Table 1과 같은 행동 정책들이 학습이 완료된 후에는 행동 정책의 성능 테스트를 위해 학습용 작업들의 로봇 초기 위치와는 다른 범위의 로봇 초기 위치들을 임의로 선택하여 테스트 작업군들(test set 1, 2, 3, 4)을 설정하였다.

Table 2의 실험 결과를 살펴보면, 학습용 작업들의 초기 위치와 동일한 범위의 작업들(train set)에 대해서는 학습된 행동 정책은 100%의 작업 성공률을 보였다. 그러나 테스트 작업군들의 초기 위치의 범위가 학습용 작업들의 초기 위치 범위와의 차이가 커질수록, 작업 성공률이 점차 낮아지는 결과를 보였다. 하지만 test set 1과 test set 2와 같이 학습용 작업들과는 비교적 차이가 있는 테스트 작업군들의 경우에도 각각 0.53, 0.27 정도의 작업 성공률을 보여주었다. 이러한

Table 1. Learned Policy: A Set of Action Rules

Head of Rule	Body of Rule
move (X, Y)	not robotAt (X, Y), not goal_on (Y), isLocation (Y), not loadedBy (Z, X)
	not on (Z, Y), isLocation (Y), not loadedBy (Z, X), not robotAt (X, Y), not goal_on (Y)
load (X, Y, Z)	isLocation (Y), not goal_on (Y), not loadedBy (Z, X), robotAt (X, Y), on (Z, Y)
unload (X, Y, Z)	not on (Z, Y), loadedBy (Z, X), isLocation (Y), goal_on (Y)
carry (X, Y, Z)	not robotat (X, Y), loadedBy (Z, X), isLocation (Y), goal_on (Y)
returnRack (X, Y, Z)	mountedBy (Z, X), robotAt (X. Y), not goal_on (Y), not loadedBy (N, X)

Table 2. Analysis of Policy Generality

Initial State			Success Rate
Robot Location	X-axis	Y-axis	
train set	[2.85, 3.25]	[1.85, 2.25]	1.00
test set 1	[1.55, 4.55]	[0.55, 3.55]	0.53
test set 2	[1.05, 5.05]	[0.05, 4.05]	0.27
test set 3	[0.55, 5.55]	[-0.55, 4.55]	0.17
test set 4	[4.05, 8.05]	[3.05, 7.05]	0.00

실험 결과를 토대로, 관계형 강화 학습 프레임워크를 통해 학습되는 행동 정책이 다른 작업에 대한 높은 일반성을 확인할 수 있었다.

위의 실험들외에 추가적으로, 본 논문에서 제안한 방법들을 통해 학습된 행동 정책이 실제 환경에서 실행되었을 때, 어떤 결과를 보여주는지 정성적으로 분석해보았다. Fig. 9는 작업의 초기 상태에서 시작하여 목표 상태에 이르기까지 학습된 행동 정책에 의해 실시간으로 결정된 운송용 모바일 로봇의 행동 실행 순서를 나타낸다. 정책 학습 후 실행 단계에 놓인 로봇은 각 행동 실행 단계마다 그때의 환경 상태와 행동 정책에 따라 실행할 행동을 선택한다. 이때 각 환경 상태는 벡터 형태의 센서 데이터에 상태 인코더를 적용하여 일차 술어 논리 형태로 표현된다. 또한, 관계형 강화학습 프레임워크에 의해 학습된 행동 정책 역시 논리 서술자 기반의 규칙들로 표현된다. 따라서 센서와 모터 제어의 노이즈(noise)와 불확실성(uncertainty)으로 인한 환경 내의 실제 로봇 위치의 오차를 일정량 극복할 수 있다. 예컨대, Fig. 9의 환경의 초기 상태에 로봇의 좌표는 (3.2, 2.1), 박스의 좌표는 (1.3, 5.3)로 추론기를 통하여 논리 서술자로 만들게 되면 robotAt (robot, loc13), on(box, loc1)로 표현된다. 이 상태와 배경 지식을 통하여 행동 정책으로부터 행위를 선택하면 행동 규칙 :move(X,Y):-isLocation(Y), not robotAt(X, Y), not

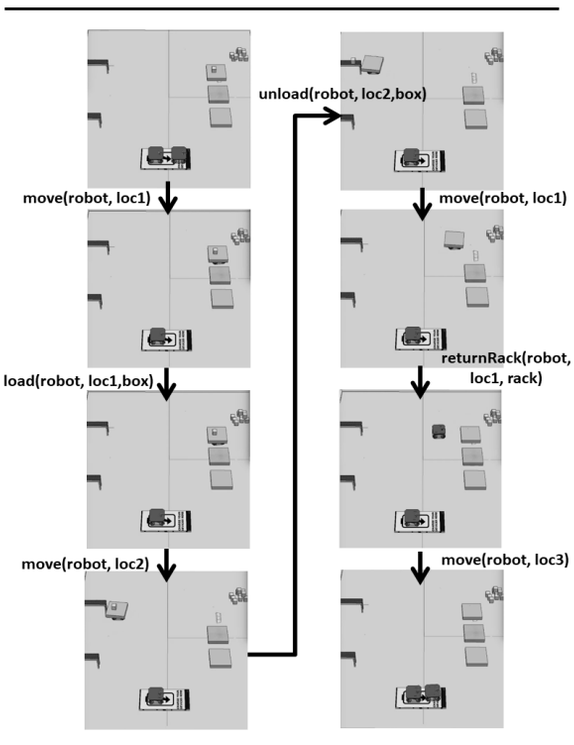


Fig. 9. Qualitative Analysis of Learned Policy during Execution

loadedBy(Z, X), not goal_on(Y)에 의해 실제 행위 move (robot, loc1)이 실행된다. 실행이 종료되면 로봇의 위치가 (3.1, 1.9)에서 (0.9, 5.1)로 옮겨진다. 해당 위치는 상태 인코더 추론기를 통하여 추론하면 robotAt(robot, loc1)가 된다. 이처럼, 벡터로 표현된 실제 위치에는 불확실성으로부터 오차가 생기지만, 이를 추론하여 논리 서술자로 만들게 되면 행동 정책으로부터 올바른 행위를 선택하여 실행할 수 있다. 이러한 실행 과정을 반복함으로써, 학습된 행동 정책은 작업 목표 상태에 성공적으로 도달할 수 있도록 로봇 행동 결정과 실행을 효율적으로 유도하였다. 이를 통하여 본 논문에서 제안한 도메인 지식을 활용한 관계형 강화 학습의 결과로서 얻어지는 행동 정책의 정당성(correctness)와 높은 일반성(high generality)을 다시 한번 확인할 수 있었다.

5. 결 론

본 논문에서는 대표적인 관계형 강화 학습 프레임워크인 dNL-RRL을 기초로 제조 공장 내 운송용 모바일 로봇의 제어를 위한 행동 정책 학습을 수행하였으며, 학습의 효율성 향상을 위해 인간 전문가의 사전 도메인 지식을 활용하는 방안들을 제안하였다. 그리고 다양한 정량적, 정성적 실험들을 통해, 제안한 도메인 지식 활용 방법의 긍정적 효과와 학습된 행동 정책의 해석 가능성, 일반성 등을 확인하였다.

본 연구에서는 학습하고자 하는 행동 정책을 나타내는 일부의 행동 규칙이나 규칙 조건을 인간 전문가가 미리 학습 이전에 제공하는 방식으로 사전 도메인 지식을 활용하는 방법을 제안하였다. 향후 연구에서는 관계형 강화 학습의 효율성 개선을 위해, 제안한 방식 이외에 환경 상태 제약 조건 명세(environmental state constraint specification)나 사전 학습된 신경망 모듈(pre-trained neural module)들을 이용하는 등 보다 더 다양한 방식으로 사전 도메인 지식을 활용할 수 있는 방법들을 연구해볼 계획이다. 또한, 본 연구에서는 단일 에이전트 관계형 강화 학습에서 도메인 지식 활용 문제를 다루었지만, 향후에는 멀티 에이전트 관계형 강화 학습(multiagent relational reinforment learning)에 적용가능하도록 제안한 방법들을 확장하는 연구를 진행해볼 계획이다.

References

[1] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, and D. Zhou, "Neural logic machines," *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[2] V. Zambaldi, et al., "Relational deep reinforcement learning," *arXiv preprint arXiv:1806.01830*, 2018.

[3] Z. Jiang and S. Luo, "Neural logic reinforcement learning," *Proceedings of International Conference on Machine Learning (ICML)*, 2019.

[4] A. Payani and F. Fekri, "Incorporating relational background knowledge into reinforcement learning via differentiable inductive logic programming," *arXiv preprint arXiv:2003.10386*, 2020.

[5] A. Payani and F. Fekri, "Inductive logic programming via differentiable deep neural logic networks," *arXiv preprint arXiv:1906.03523*, 2019.

[6] J. Janisch, T. Pevný, and V. Lisý, "Symbolic relational deep reinforcement learning based on graph neural networks," *arXiv preprint arXiv:2009.12462*, 2020.

[7] S. Garg and A. Bajpai, "Symbolic network: Generalized neural policies for relational MDPs," *Proceedings of International Conference on Machine Learning (ICML)*, pp.3397-3407, Nov. 2020.

[8] O. Rivlin, T. Hazan, and E. Karpas, "Generalized planning with deep reinforcement learning," *arXiv preprint arXiv: 2005.02305*, 2020.

[9] S. Das, S. Natarajan, K. Roy, R. Parr, and K. Kersting, "Fitted Q-learning for relational domains," *arXiv preprint arXiv: 2006.05595*, 2020.

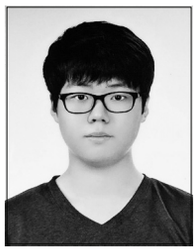
[10] D. Adjodah, T. Klinger, and J. Joseph, "Symbolic relation networks for reinforcement learning," *Proceedings of the Workshop on Relational Representation Learning in Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[11] T. Gokhale, S. Sampat, Z. Fang, Y. Yang, and C. Baral, "Blocksworld revisited: Learning and reasoning to generate event-sequences from image pairs," *arXiv preprint arXiv: 1905.12042*, 2019.

[12] Y. Zhang and A. Ramesh, "Learning interpretable relational structures of hinge-loss markov random fields," *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pp.6050-6056, Aug. 2019.

[13] Y. Zhang and A. Ramesh, "Learning fairness-aware relational structures," *Proceedings of European Conference on Artificial Intelligence (ECAI)*, 2020.

[14] M. A. Skinner, L. Raman, N, Shah, A. Farhat, and S. Natarajan, "A preliminary approach for learning relational policies for the management of critically ill children," *arXiv preprint arXiv:2001.04432*, 2020.



강 민 교

https://orcid.org/0000-0002-1266-0511
e-mail : alsry5786@kyonggi.ac.kr
2020년 경기대 컴퓨터공학부(학사)
2020년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 기계학습, 로봇지능

김 인 철

https://orcid.org/0000-0002-5754-133X
e-mail : kic@kyonggi.ac.kr
1985년 서울대학교 수학과(이학사)
1987년 서울대학교 전산과학과(이학석사)
1995년 서울대학교 전산과학과(이학박사)
1996년 ~ 현 재 경기대학교 컴퓨터공학부 교수
관심분야 : 인공지능, 기계학습, 로봇지능