Cooperative Multi-Agent Reinforcement Learning: Solving Credit Assignment with Value Decomposition

Data Mining & Quality Analytics Open Seminar

2024.08.30

김 정 인

발표자 소개

❖ 김정인 (Jung In Kim)

- Data Mining & Quality Analytics Lab
- Ph.D. Student (2021.09 ~)
- 지도 교수: 김성범 교수님

❖ 관심 연구 분야

- Deep Reinforcement Learning
- Self-supervised Learning

❖ Contact

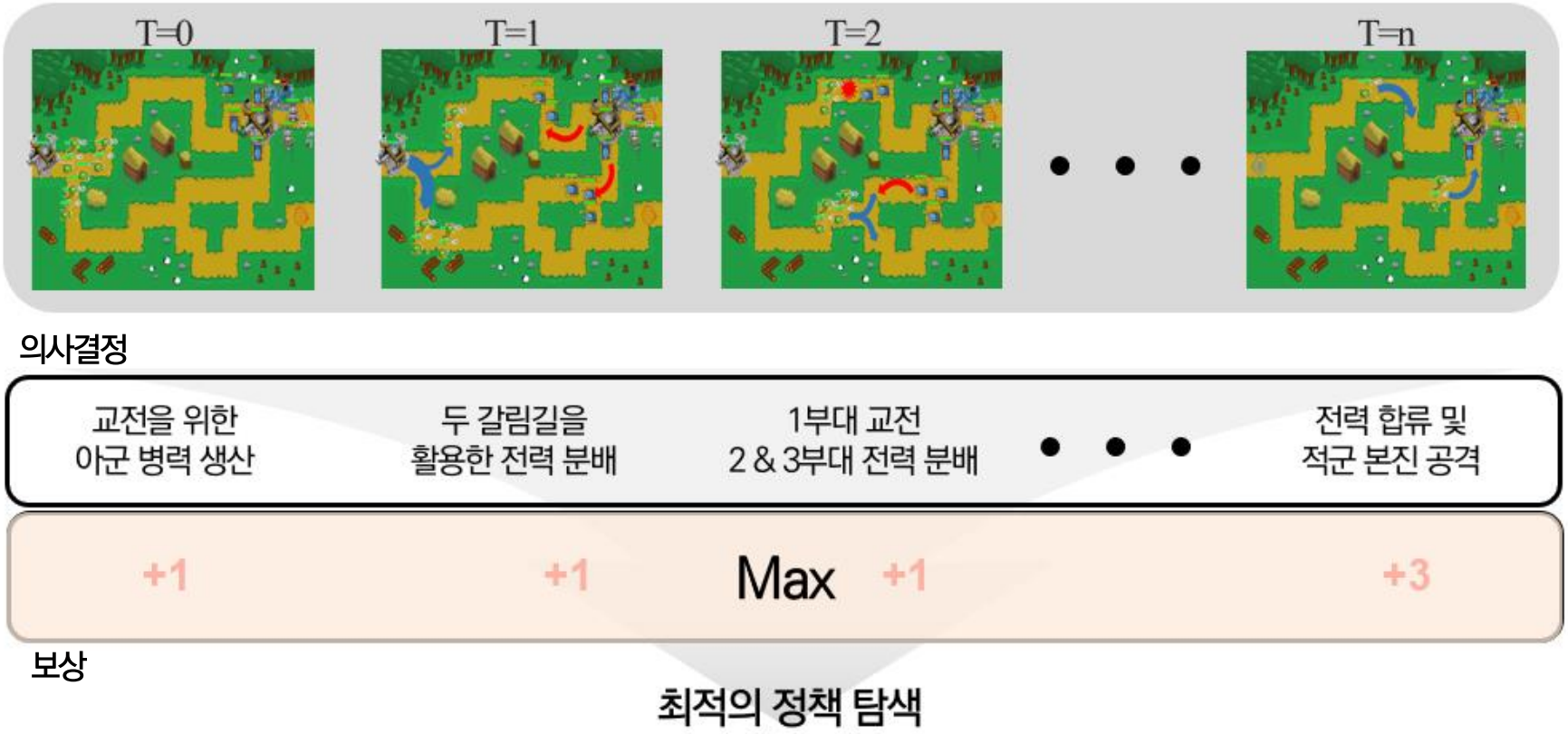
- Jungin_kim23@korea.ac.kr

목차

Introduction

강화학습 정의

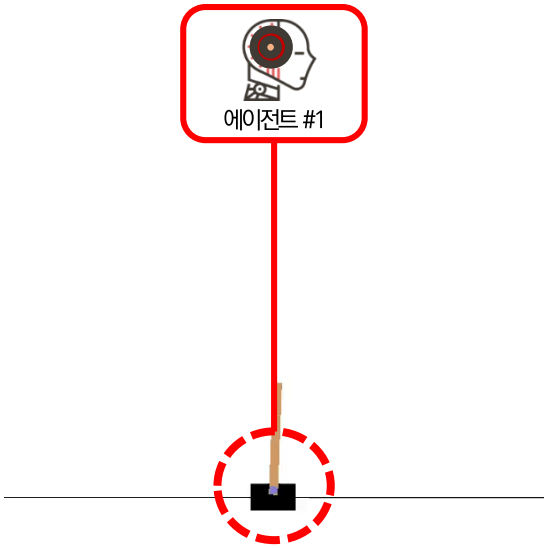
- ❖ 순차적인 의사결정 문제에서 에이전트가 환경으로부터 받는 누적 보상 값을 최대화하는 행동 정책을 학습하는 방법
- ❖ 정책: 에이전트가 특정 상태에서 어떤 행동을 선택할지 정해주는 함수



Introduction

단일 에이전트 강화학습 (SARL) vs 다중 에이전트 강화학습 (MARL)

- SARL: 하나의 에이전트가 환경과 상호작용하여 상황에 따른 최적의 행동 정책을 학습하는 방법론
- MARL: 두 개 이상의 에이전트가 서로 협력하고 환경과 상호작용하여 상황에 따른 최적의 행동 정책을 학습하는 방법론



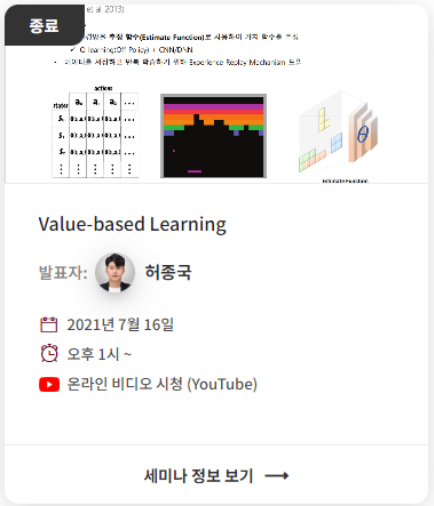
단일 에이전트 강화학습 (SARL)

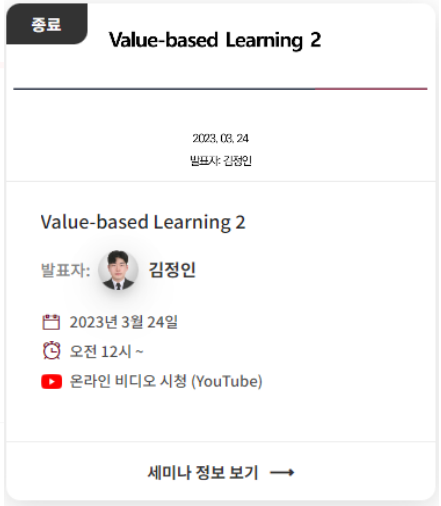
다중 에이전트 강화학습 (MARL)

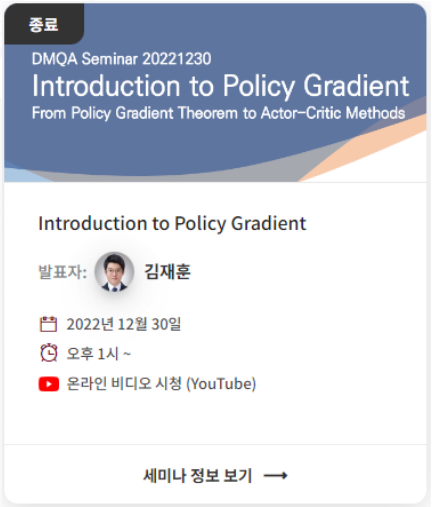
Introduction

단일 에이전트 강화학습 (SARL) vs 다중 에이전트 강화학습 (MARL)

- ❖ SARL: 하나의 에이전트가 환경과 상호작용하여 상황에 따른 최적의 행동 정책을 학습하는 방법론
- ❖ MARL: 두 개 이상의 에이전트가 서로 협력하고 환경과 상호작용하여 상황에 따른 최적의 행동 정책을 학습하는 방법론







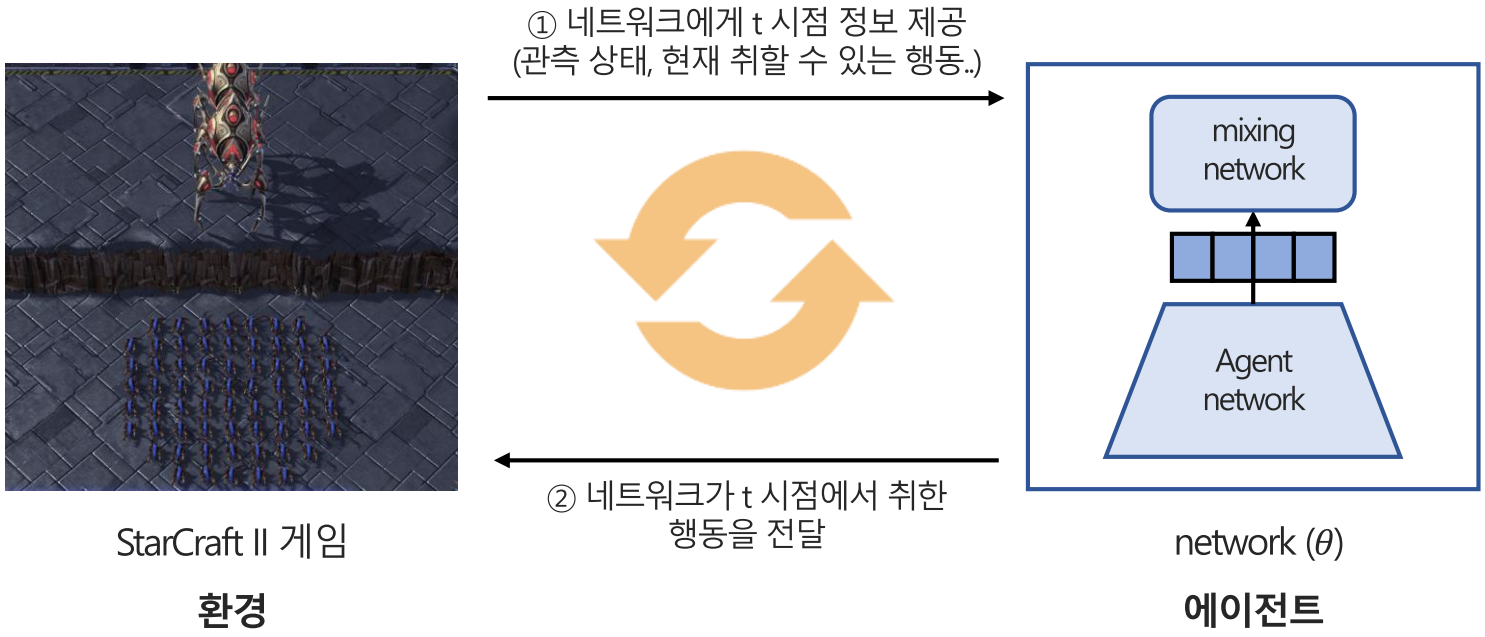
단일 에이전트 강화학습 (SARL)

다중 에이전트 강화학습 (MARL)

Introduction

다중 에이전트 강화학습 용어 설명 및 데이터 수집 과정

- ❖ 환경: 네트워크(에이전트)의 학습을 위한 데이터를 제공해주는 역할
- ❖ 에이전트: 학습 대상, 네트워크 또는 모델
- ❖ 관측 상태(observation): 게임 내에서 학습 대상이 처한 상황에 대한 주변 정보
- ❖ 행동: 학습 대상이 현 상황에서 취한 동작

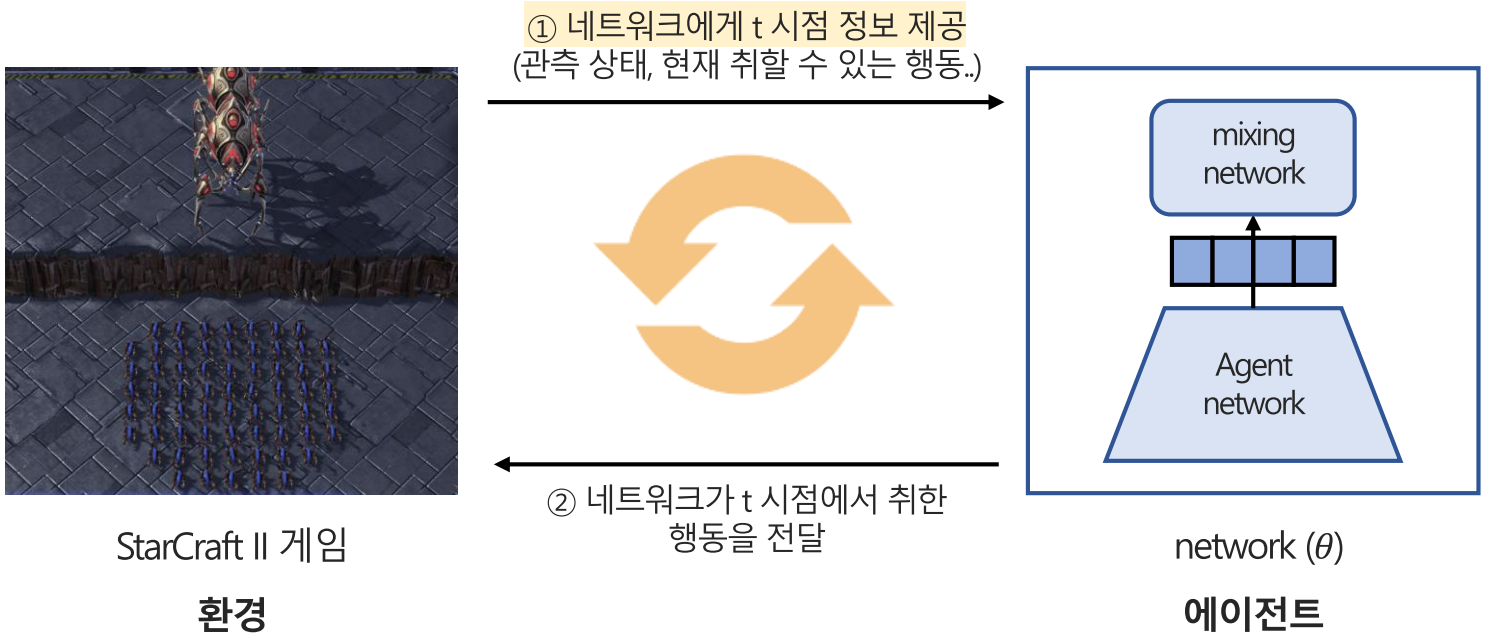


상호작용: 데이터를 수집하는 과정

Introduction

다중 에이전트 강화학습 용어 설명 및 데이터 수집 과정

- 환경: 네트워크(에이전트)의 학습을 위한 데이터를 제공해주는 역할
- 에이전트: 학습 대상, 네트워크 또는 모델
- 관측 상태(observation): 게임 내에서 학습 대상이 처한 상황에 대한 주변 정보
- 행동: 학습 대상이 현 상황에서 취한 동작

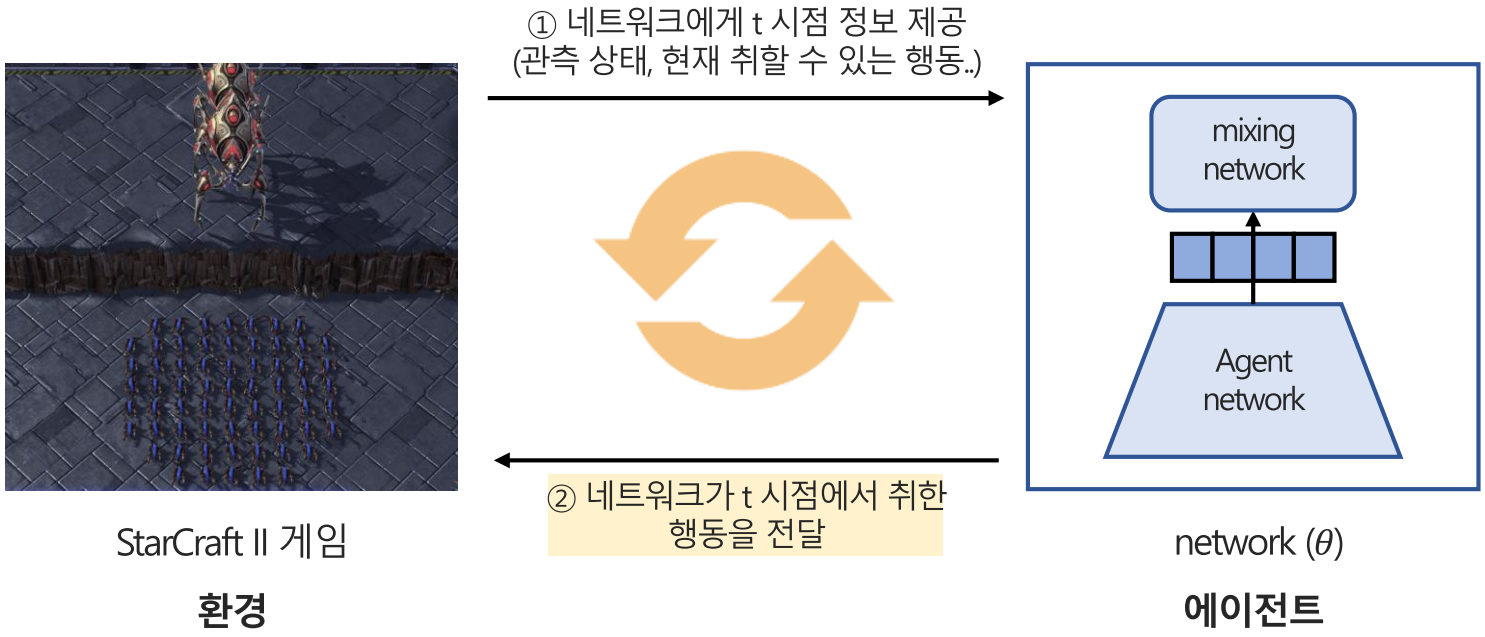


상호작용: 데이터를 수집하는 과정

Introduction

다중 에이전트 강화학습 용어 설명 및 데이터 수집 과정

- ❖ 환경: 네트워크(에이전트)의 학습을 위한 데이터를 제공해주는 역할
- ❖ 에이전트: 학습 대상, 네트워크 또는 모델
- ❖ 관측 상태(observation): 게임 내에서 학습 대상이 처한 상황에 대한 주변 정보
- ❖ 행동: 학습 대상이 현 상황에서 취한 동작

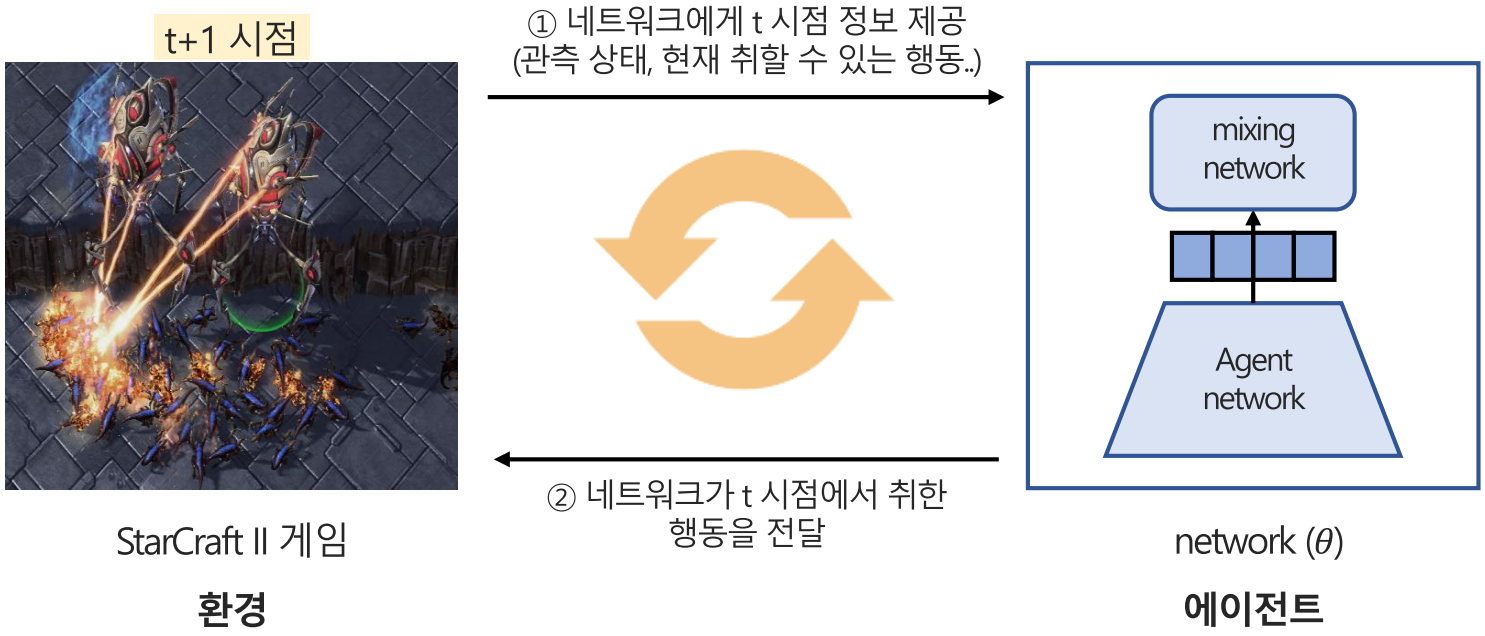


상호작용: 데이터를 수집하는 과정

Introduction

다중 에이전트 강화학습 용어 설명 및 데이터 수집 과정

- ❖ 환경: 네트워크(에이전트)의 학습을 위한 데이터를 제공해주는 역할
- ❖ 에이전트: 학습 대상, 네트워크 또는 모델
- ❖ 관측 상태(observation): 게임 내에서 학습 대상이 처한 상황에 대한 주변 정보
- ❖ 행동: 학습 대상이 현 상황에서 취한 동작

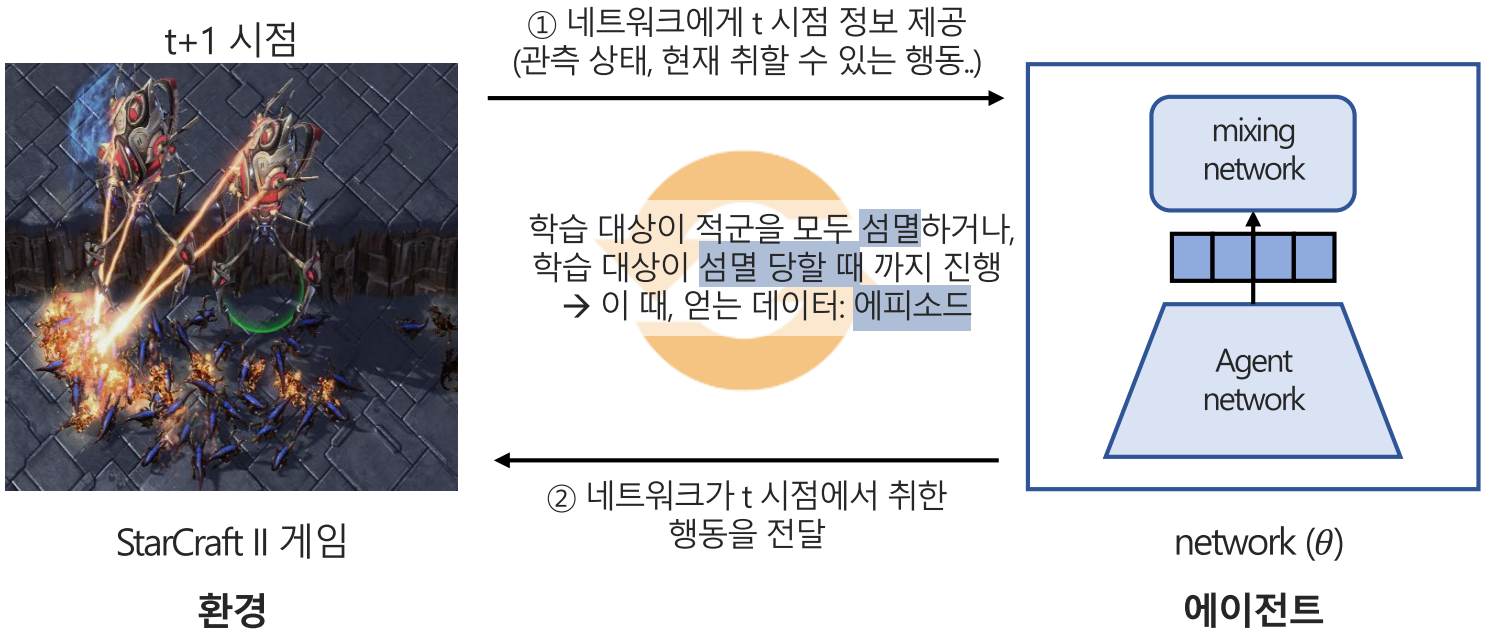


상호작용: 데이터를 수집하는 과정

Introduction

다중 에이전트 강화학습 용어 설명 및 데이터 수집 과정

- ❖ 환경: 네트워크(에이전트)의 학습을 위한 데이터를 제공해주는 역할
- ❖ 에이전트: 학습 대상, 네트워크 또는 모델
- ❖ 관측 상태(observation): 게임 내에서 학습 대상이 처한 상황에 대한 주변 정보
- ❖ 행동: 학습 대상이 현 상황에서 취한 동작

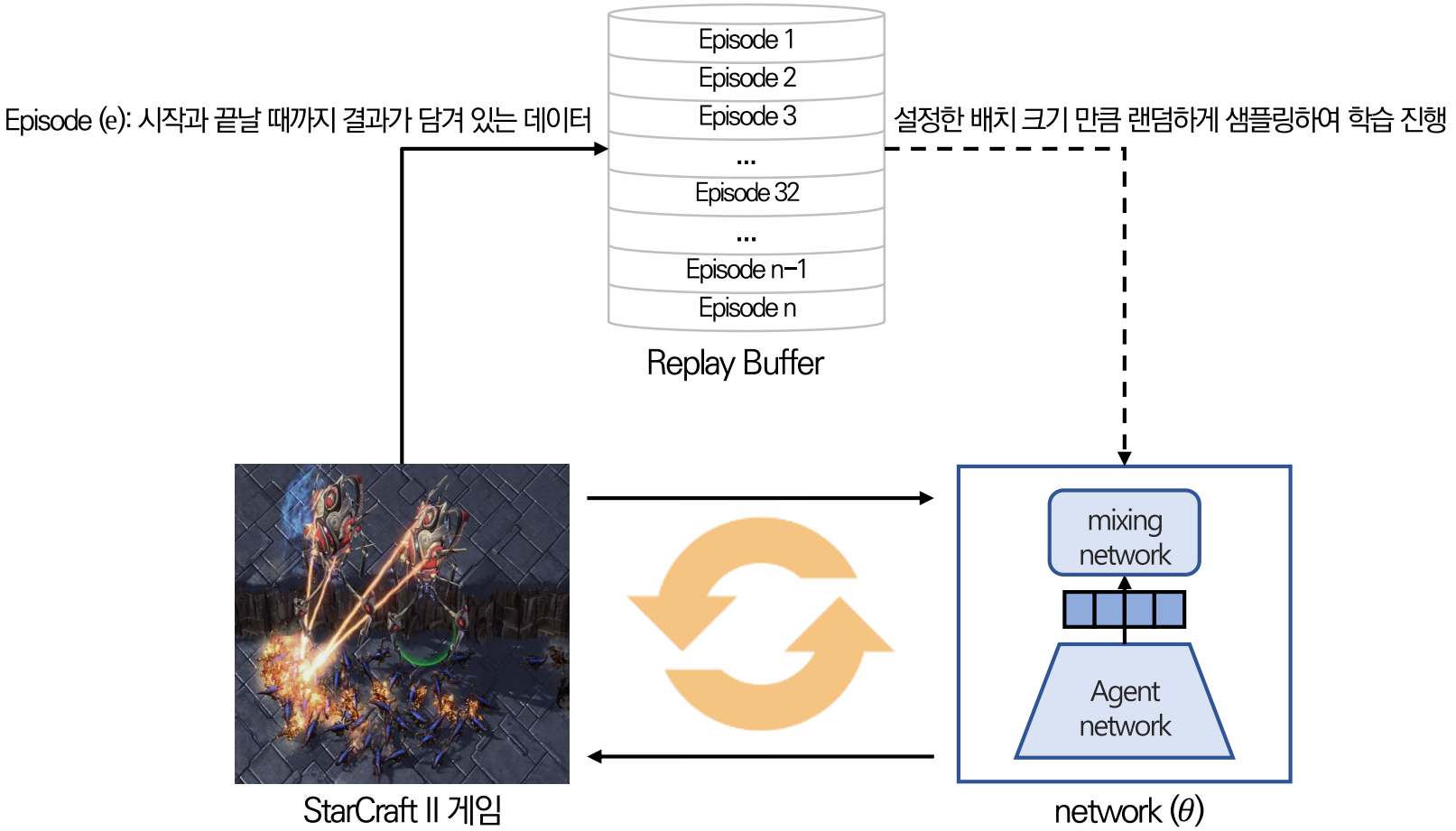


상호작용: 데이터를 수집하는 과정

Introduction

다중 에이전트 강화학습의 학습 과정

- ❖ 시작과 종료될 때까지의 데이터(에피소드)를 데이터 저장 공간(replay buffer)에 적재
- ❖ 설정한 배치 크기만큼 데이터가 적재되었을 때, 학습 진행



Introduction

데이터 설명

- ❖ 시작과 종료될 때까지의 데이터(에피소드)를 데이터 저장 공간(replay buffer)에 적재
- ❖ 설정한 배치 크기만큼 데이터가 적재되었을 때, 학습 진행

Episode={Global State, Observation, Actions, Avail_actions, Reward}

Global State

dimension

Timestep (시점)							
0	**1.0**	**1.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
1	0.78	0.69	0.0	0.0	0.0	0.0	0.0
2	1.0	1.0	…	…	0.69	0.0	0.0
…	…	…	…	…	…	…	…
n	0.0	0.58	…	…	0.89	0.0	0.0

Observation

dimension

Timestep (시점)

Agent 수

0	**1.0**	**1.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
1	0.78	0.69	0.0	0.0	0.0	0.0	0.0
2	1.0	1.0	…	…	0.69	0.0	0.0
…	…	…	…	…	…	…	…
x	0.0	0.58	…	…	0.89	0.0	0.0

Introduction

데이터 설명

- ❖ 시작과 종료될 때까지의 데이터(에피소드)를 데이터 저장 공간(replay buffer)에 적재
- ❖ 설정한 배치 크기만큼 데이터가 적재되었을 때, 학습 진행

Episode={Global State, Observation, Actions, Avail_actions, Reward}

Introduction

데이터 설명

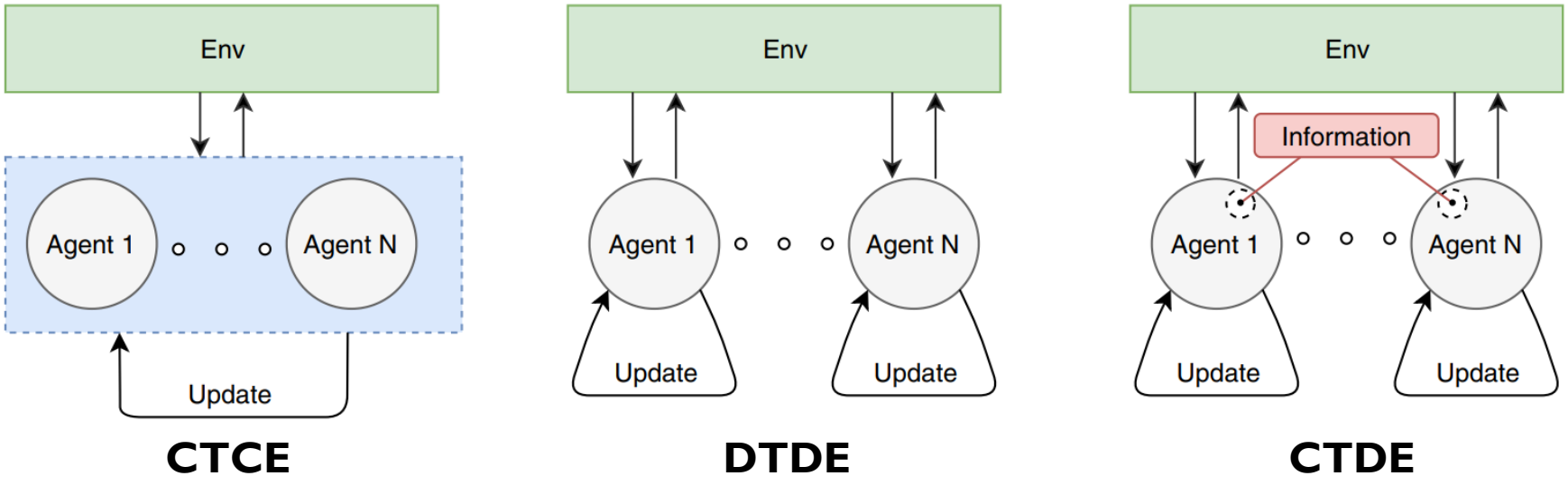
- ❖ 시작과 종료될 때까지의 데이터(에피소드)를 데이터 저장 공간(replay buffer)에 적재
- ❖ 설정한 배치 크기만큼 데이터가 적재되었을 때, 학습 진행

Episode={Global State, Observation, Actions, Avail_actions, Reward}

Introduction

훈련 방식

- ❖ **C**entralized **T**raining **C**entralized **E**xecution (CTCE): 모든 에이전트가 공동 정책을 유지하며, 훈련 중에 중앙 집중식으로 정보를 공유하고 이를 기반으로 행동을 결정하는 방식
- ❖ **D**ecentralized **T**raining **D**ecentralized **E**xecution (DTDE): 각 에이전트가 독립적으로 자신의 정책을 업데이트, 다른 에이전트와의 정보 공유 없이 개별적으로 행동을 결정하는 방식
- ❖ **C**entralized **T**raining **D**ecentralized **E**xecution (CTDE): 훈련 중에 에이전트들이 추가 정보를 교환하여 공동으로 학습하지만, 테스트 시에는 각 에이전트가 독립적으로 행동하는 방식



Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55(2), 895-943.

Introduction

다중 에이전트 강화학습 도메인에서 발생하는 여러 문제

Table 3 Overview of MADRL challenges and approaches proposed in recent literature

	Challenge	Approach	Literature
①	Non-stationarity	Experience replay	Foerster et al. (2017), Palmer et al. (2018), Tang et al. (2018), and Zheng et al. (2018a)
		Centralized training	Bono et al. (2019), Foerster et al. (2016), Foerster et al. (2018b), Iqbal and Sha (2019), Jorge et al. (2016), Lowe et al. (2017), Rashid et al. (2018), and Wei et al. (2018)
		Meta-learning	Al-Shedivat et al. (2018), and Rabinowitz et al. (2018)
②	Communication	Broadcasting	Foerster et al. (2016), Peng et al. (2017), and Sukhbaatar et al. (2016)
		Targeted	Das et al. (2019), Hoshen (2017), Jain et al. (2019), Jiang and Lu (2018), and Singh et al. (2019)
		Networked	Chu et al. (2020), Chu et al. (2020), Qu et al. (2020), Zhang et al. (2018), and Zhang et al. (2019)
		Extensions	Celikyilmaz et al. (2018), Jaques et al. (2018), Jaques et al. (2019), Kim et al. (2019), Li et al. (2019b), Singh et al. (2019), and Wang et al. (2020c)
③	Coordination	Independent learners	Foerster et al. (2018b), Lyu and Amato (2020), Omidshafiei et al. (2017), Palmer et al. (2018), Palmer et al. (2019), Sunehag et al. (2018), and Zheng et al. (2018a)
		Constructing models	Barde et al. (2019), Everett and Roberts (2018), Foerster et al. (2018a), Foerster et al. (2019), Grover et al. (2018), He et al. (2016), Hong et al. (2017), Hoshen (2017), Jaques et al. (2019), Le et al. (2017), Letcher et al. (2019), Raileanu et al. (2018), Tacchetti et al. (2019), Yang et al. (2018a), and Zheng et al. (2018b)
		Hierarchical methods	Ahilan and Dayan (2019), Cai et al. (2013), Han et al. (2019), Jaderberg et al. (2019), Kumar et al. (2017), Lee et al. (2020), Ma and Wu (2020), Tang et al. (2018), and Vezhnevets et al. (2019)
④	Credit assignment	Decomposition	Castellini et al. (2019), Chen et al. (2018), Nguyen et al. (2017b), Rashid et al. (2018), Son et al. (2019), Sunehag et al. (2018), Wang et al. (2020a), Wang et al. (2020c), Yang et al. (2018b)
		Marginalization	Foerster et al. (2018b), Nguyen et al. (2018), and Wu et al. (2018)
		Inverse RL	Barrett et al. (2017), Le et al. (2017), Lin et al. (2018), Song et al. (2018), and Yu et al. (2019)
⑤	Scalability	Knowledge Reuse	Baker et al. (2020), Da Silva et al. (2017), Da Silva and Costa (2017), Gupta et al. (2017), Hernandez-Leal et al. (2019), Jiang and Lu (2018), Long et al. (2020), Luketina et al. (2019), Narvekar et al. (2016), Omidshafiei et al. (2019), Peng et al. (2017), Sukhbaatar et al. (2016), Sukhbaatar et al. (2017), Sunehag et al. (2018), and Svetlik et al. (2017)
		Complexity reduction	Chen et al. (2018), Lin et al. (2018), Nguyen et al. (2017a), Nguyen et al. (2017b), and Yang et al. (2018b)
		Robustness	Baker et al. (2020), Bansal et al. (2018), Berner et al. (2019), Gleave et al. (2020), Heinrich and Silver (2016), Lanctot et al. (2017), Li et al. (2019a), Liu et al. (2020), Lowe et al. (2017), Pinto et al. (2017), Raghu et al. (2018), Silver et al. (2016), Silver et al. (2018), Spooner and Savani (2020), and Sukhbaatar et al. (2017)

Table 3 (continued)

	Challenge	Approach	Literature
⑥	Partial observability	Memory mechanism	Dibangoye and Buffet (2018), Foerster et al. (2018b), Foerster et al. (2019), Gupta et al. (2017), and Omidshafiei et al. (2017)

Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55(2), 895-943.

Introduction

Credit Assignment 문제

- ❖ 여러 에이전트가 협력하여 얻은 전체 보상에 대한 개별 에이전트의 기여도를 파악하기 어려운 문제
- ❖ 개별 기여도에 따라 각 에이전트가 선택한 행동에 대한 적절한 피드백 반영을 위해서 해결되어야 할 문제

https://news.mt.co.kr/mtview.php?no=2021083107224771572

Methods

VDN: Value-Decomposition Networks For Cooperative Multi-Agent Learning (2018, AAMAS)

❖ 협력적인 환경에서 ① fully centralized and decentralized approach로 학습할 때, ② **"lazy agent"**와 **"spurious rewards"** 문제가 발생 → ③ Value-Decomposition Networks (VDN)을 제안하여 문제 해결

❖ VDN의 특징

- Centralized Training Decentralized Execution (CTDE) 학습 방식 사용
- Addictive Factorization 사용 → **Credit Assignment 문제**를 간접적으로 해결

Main Track Extended Abstract AAMAS 2018, July 10-15, 2018, Stockholm, Sweden

Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward

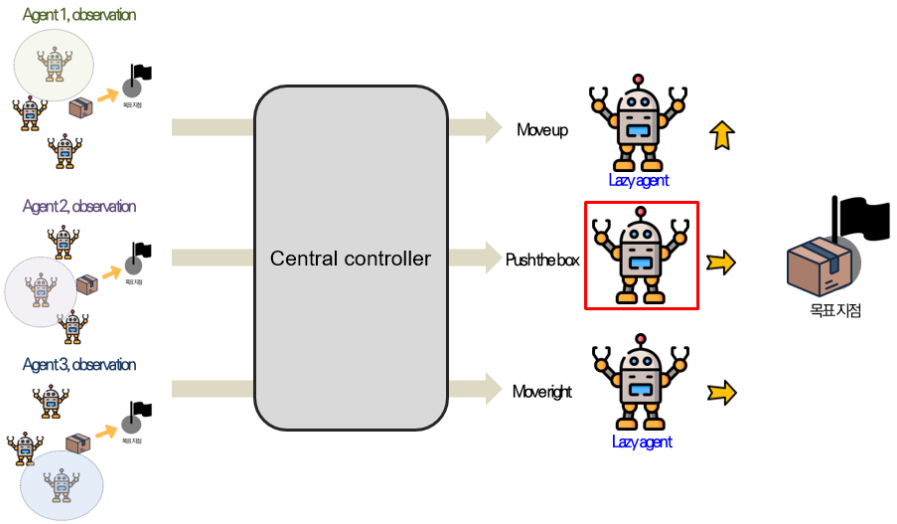
Extended Abstract

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, Thore Graepel
DeepMind, London, United Kingdom
sunehag@google.com

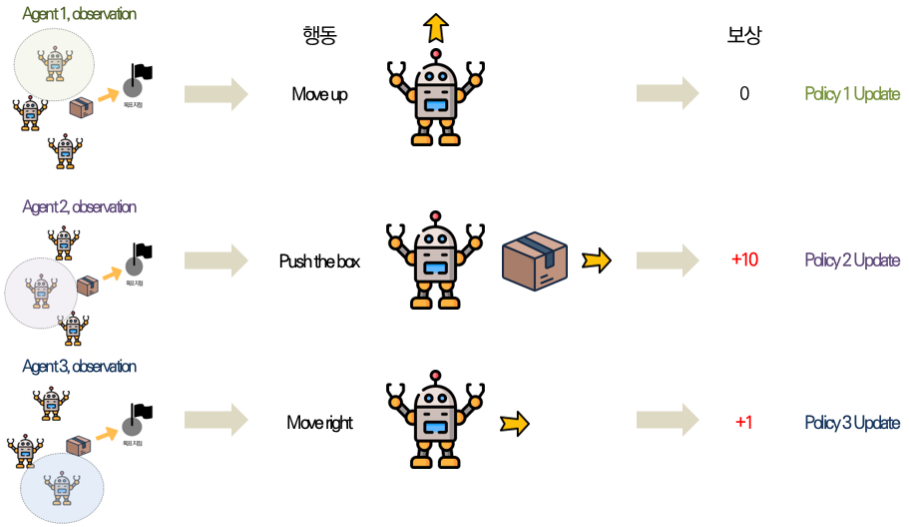
Methods

Fully Centralized and Decentralized Approach

- ❖ Fully Centralized Approach: 모든 에이전트의 행동과 정보를 중앙에서 한꺼번에 처리하고 결정 내리는 방식
- ❖ Fully Decentralized Approach: 각 에이전트가 자신의 로컬 정보만을 바탕으로 독립적으로 행동하고 결정을 내리는 방식



Fully Centralized Approach

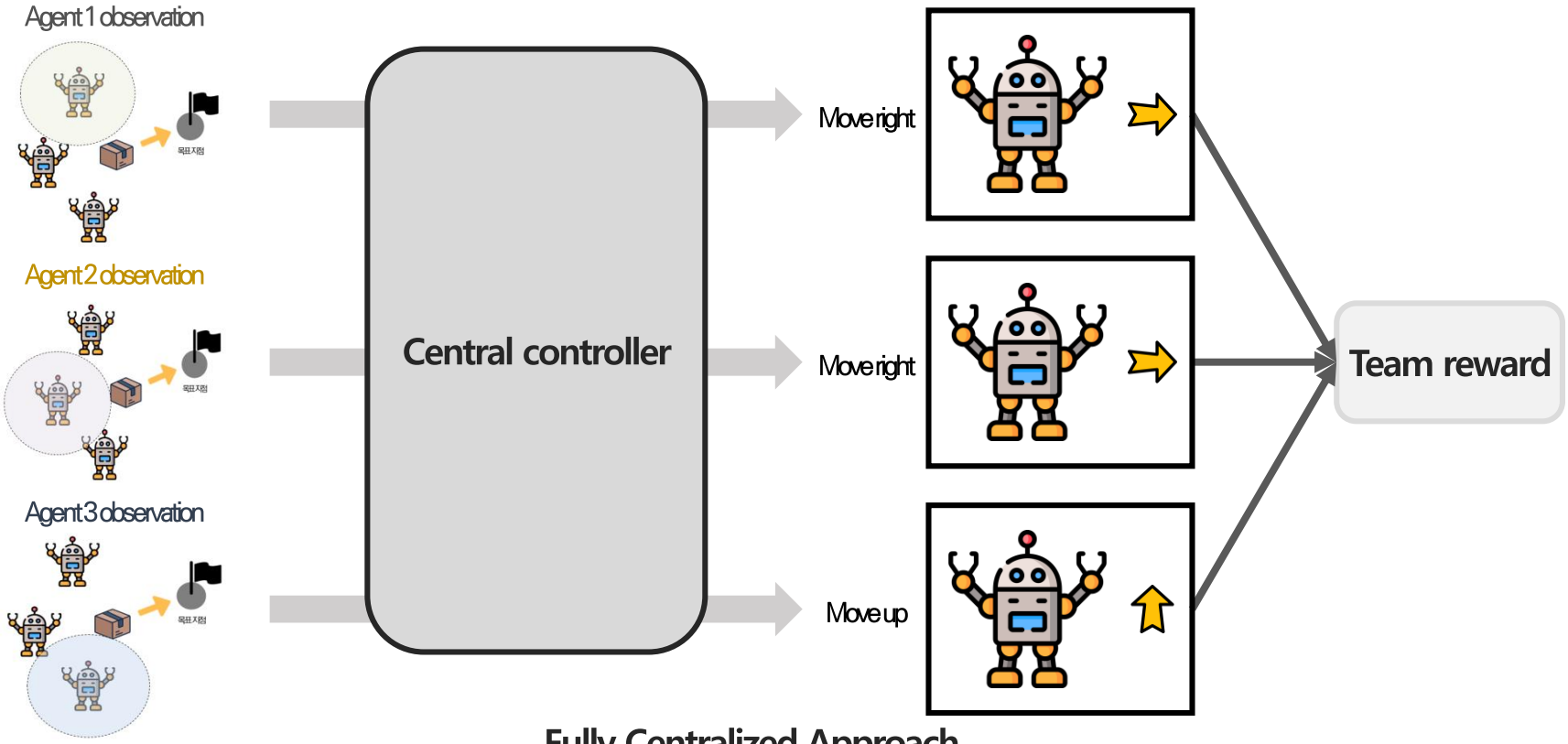


Fully Decentralized Approach

Methods

Fully Centralized and Decentralized Approach

- ❖ Fully Centralized Approach: 모든 에이전트의 행동과 정보를 중앙에서 한꺼번에 처리하고 결정 내리는 방식
- ❖ 중앙 컨트롤러가 모든 에이전트의 상태와 행동을 통합적으로 관리하여 팀 보상을 최대화하는 정책을 탐색하는 것이 목표

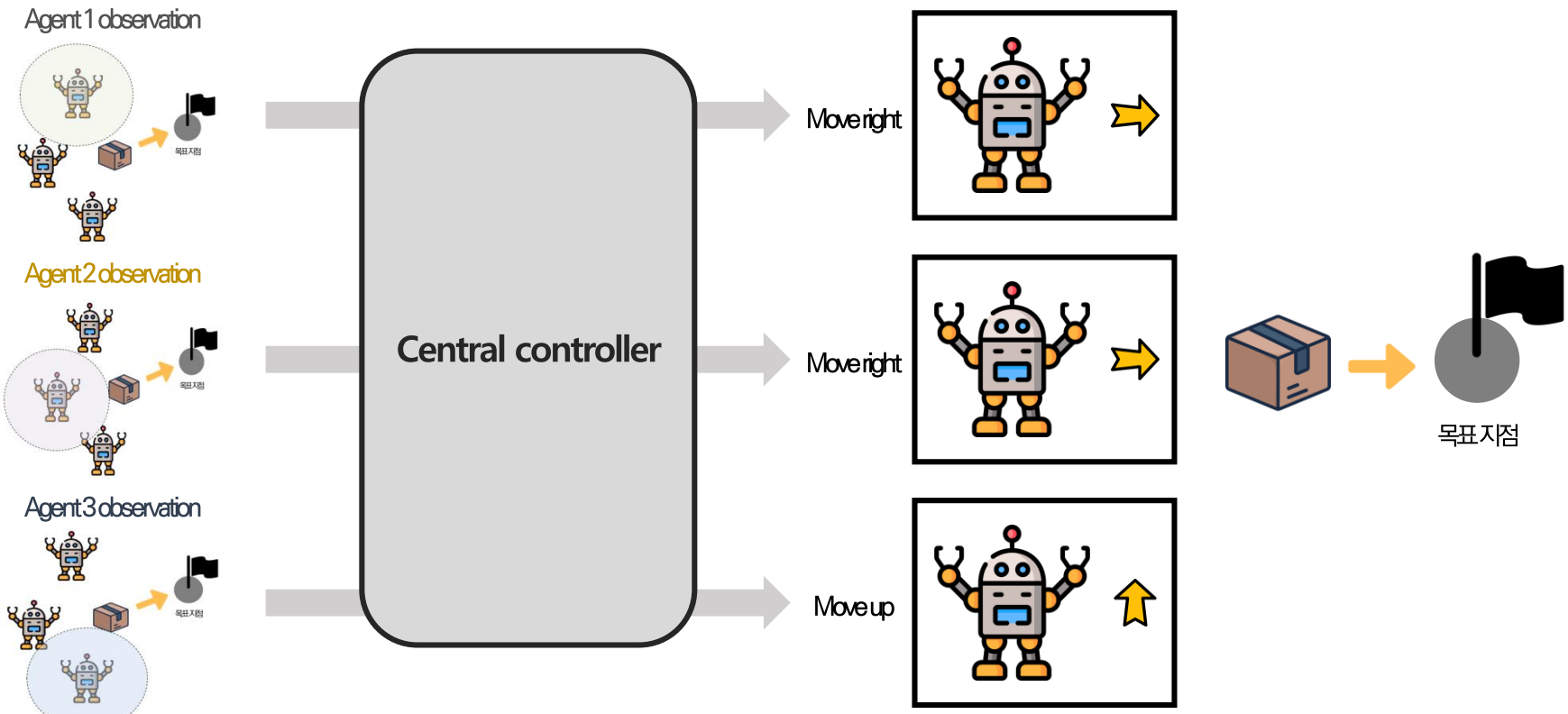


Fully Centralized Approach

Methods

Fully Centralized and Decentralized Approach

- ❖ Fully Centralized Approach: 모든 에이전트의 행동과 정보를 중앙에서 한꺼번에 처리하고 결정 내리는 방식
- ❖ 중앙 컨트롤러가 모든 에이전트의 상태와 행동을 통합적으로 관리하여 팀 보상을 최대화하는 정책을 탐색하는 것이 목표

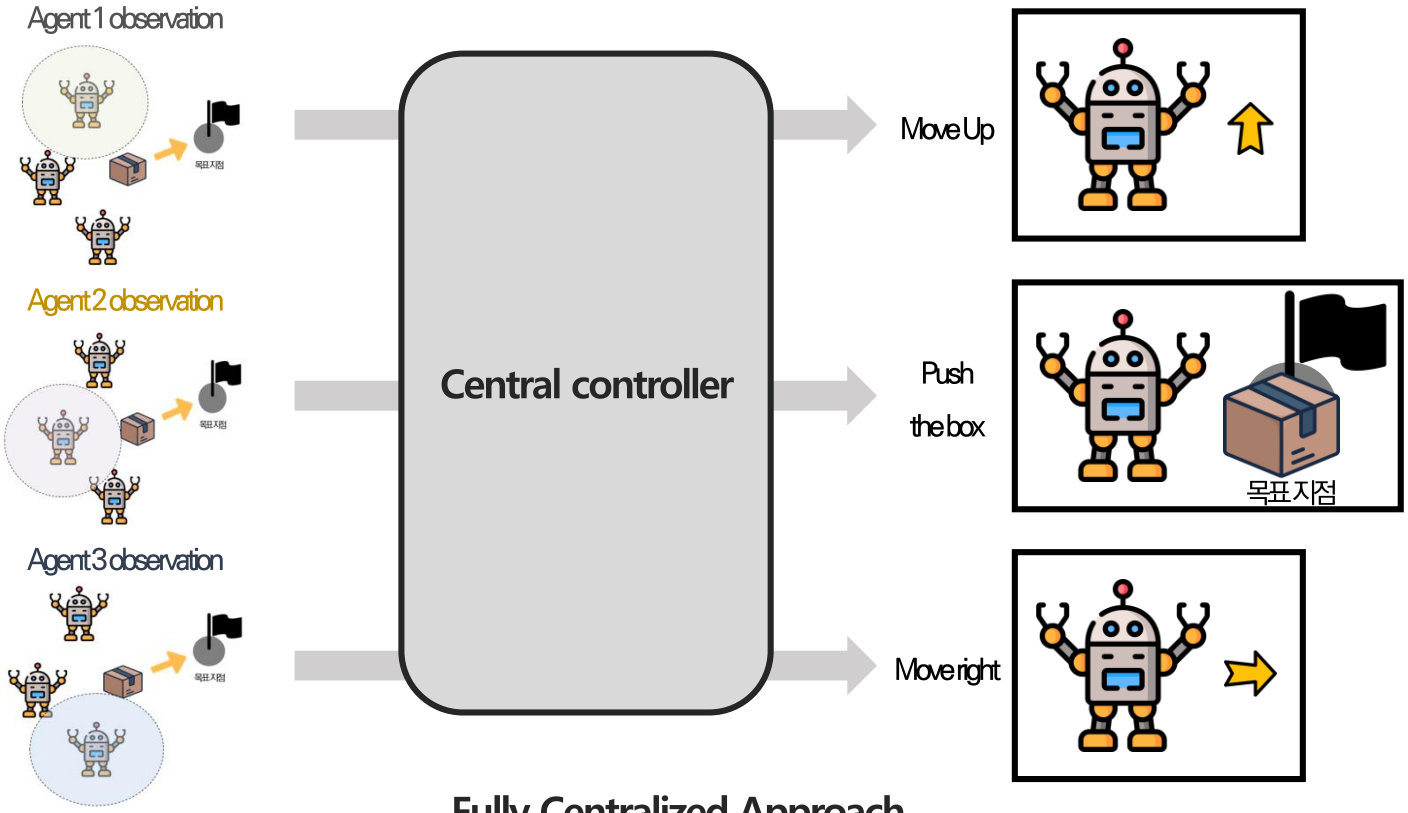


Fully Centralized Approach

Methods

Fully Centralized and Decentralized Approach

- ❖ Fully Centralized Approach: 모든 에이전트의 행동과 정보를 중앙에서 한꺼번에 처리하고 결정 내리는 방식
- ❖ 중앙 컨트롤러가 모든 에이전트의 상태와 행동을 통합적으로 관리하여 팀 보상을 최대화하는 정책을 탐색하는 것이 목표

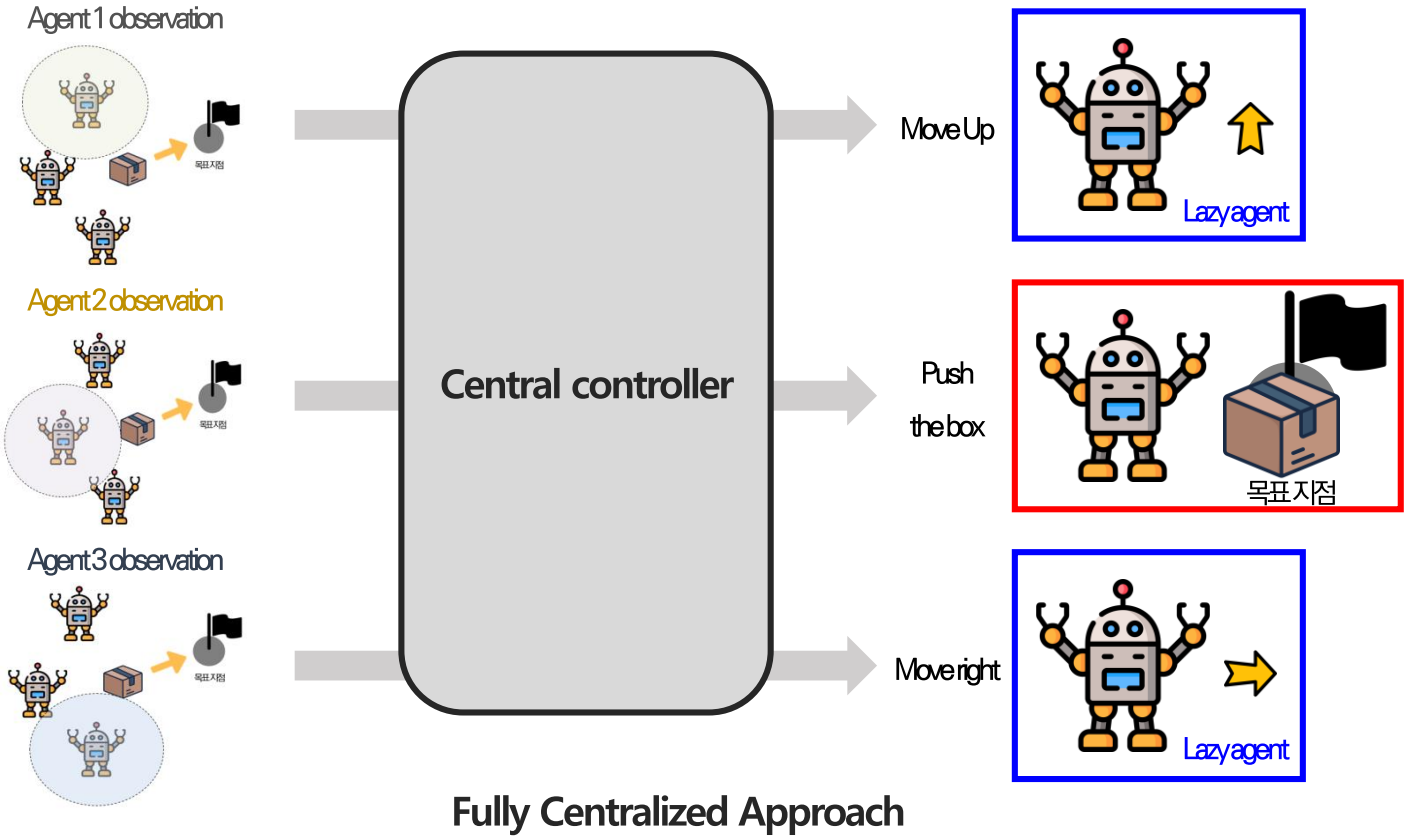


Fully Centralized Approach

Methods

Fully Centralized and Decentralized Approach

- ❖ Fully Centralized Approach: 모든 에이전트의 행동과 정보를 중앙에서 한꺼번에 처리하고 결정 내리는 방식
- ❖ 중앙 컨트롤러가 모든 에이전트의 상태와 행동을 통합적으로 관리하여 팀 보상을 최대화하는 정책을 탐색하는 것이 목표



Fully Centralized Approach

Methods

Fully Centralized and Decentralized Approach

❖ Fully Centralized Approach: 모든 에이전트의 행동과 정보를 중앙에서 한꺼번에 처리하고 결정 내리는 방식

❖ 중앙 컨트롤러가 모든 에이전트의 상태와 행동을 통합적으로 관리하여 팀 보상을 최대화하는 정책을 탐색하는 것이 목표

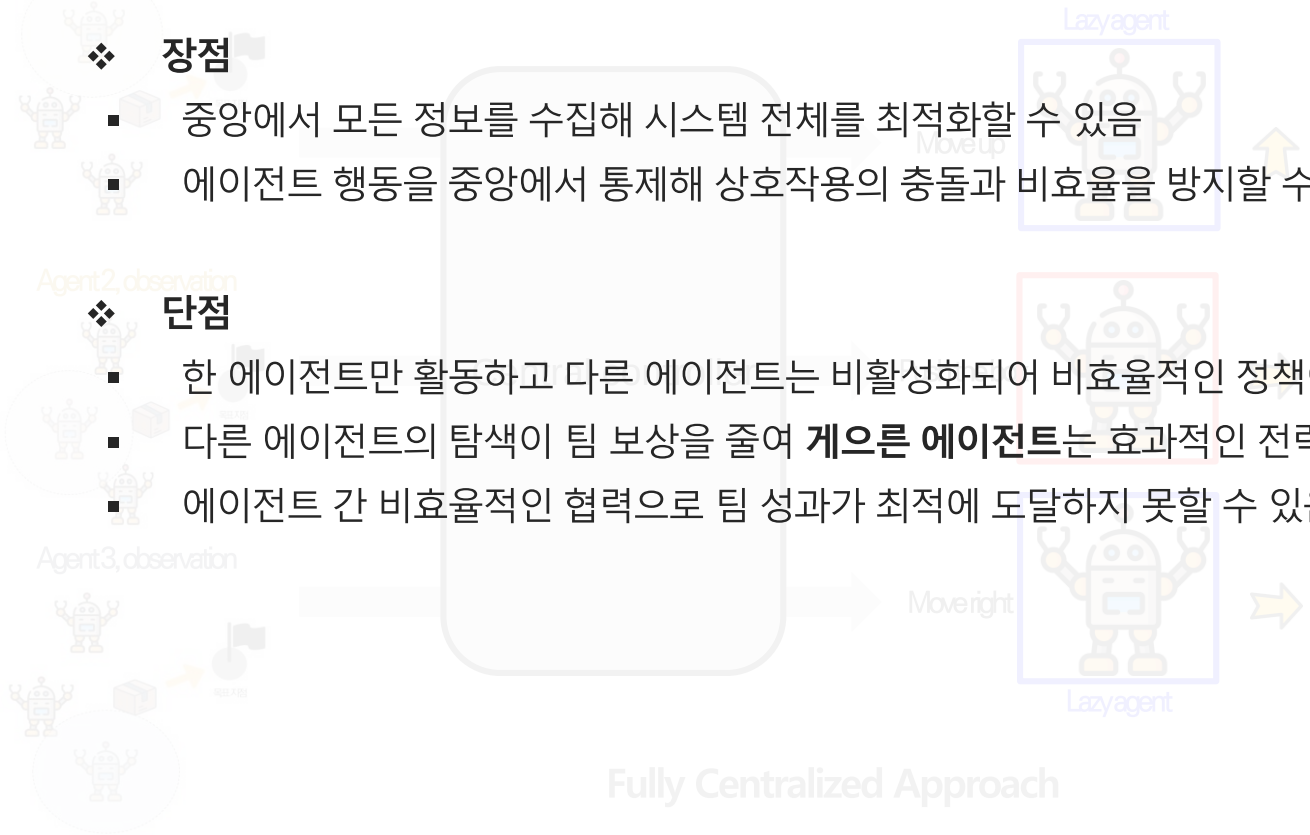
Fully Centralized Approach 장단점

❖ **장점**

- 중앙에서 모든 정보를 수집해 시스템 전체를 최적화할 수 있음
- 에이전트 행동을 중앙에서 통제해 상호작용의 충돌과 비효율을 방지할 수 있음

❖ **단점**

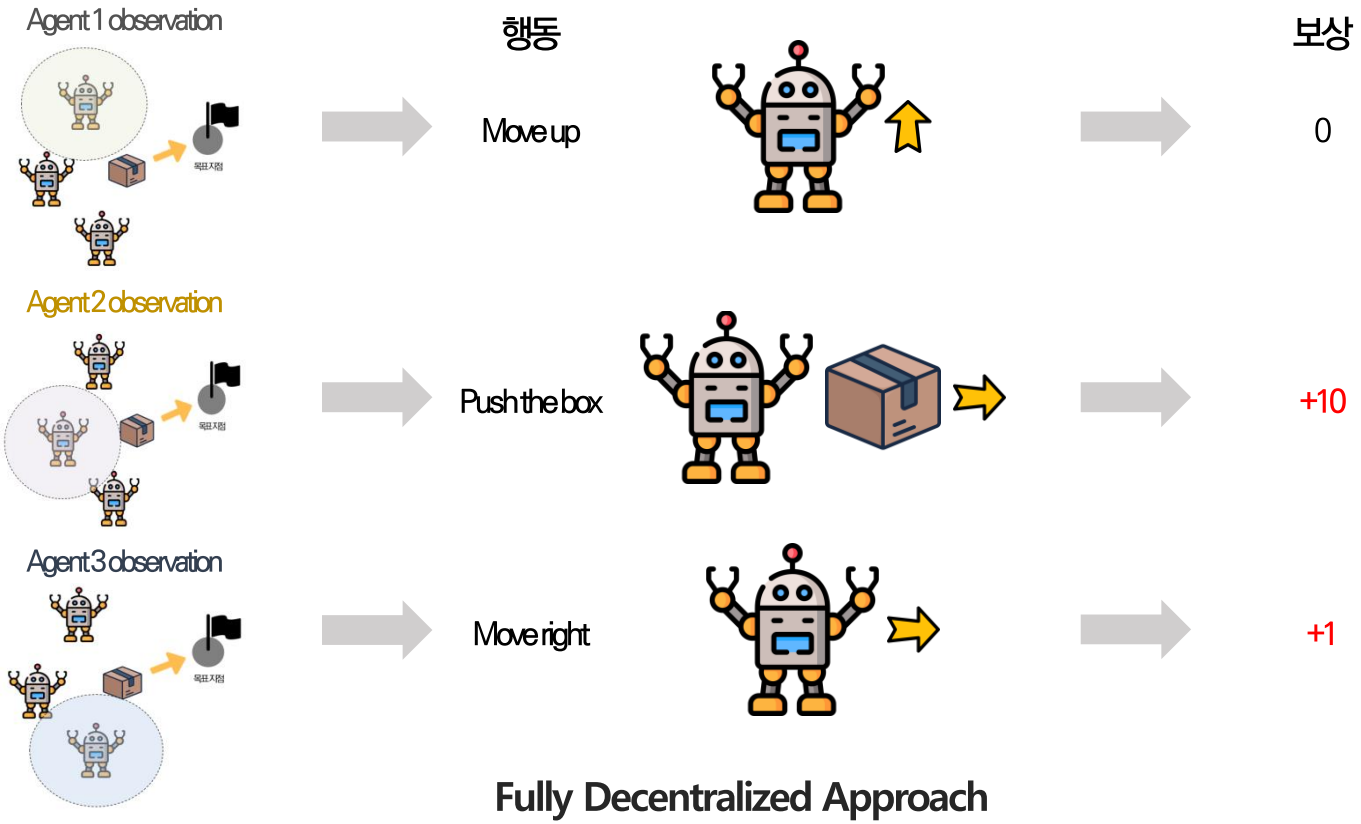
- 한 에이전트만 활동하고 다른 에이전트는 비활성화되어 비효율적인 정책이 학습될 수 있음
- 다른 에이전트의 탐색이 팀 보상을 줄여 **게으른 에이전트**는 효과적인 전략을 찾지 못함
- 에이전트 간 비효율적인 협력으로 팀 성과가 최적에 도달하지 못할 수 있음



Methods

Fully Centralized and Decentralized Approach

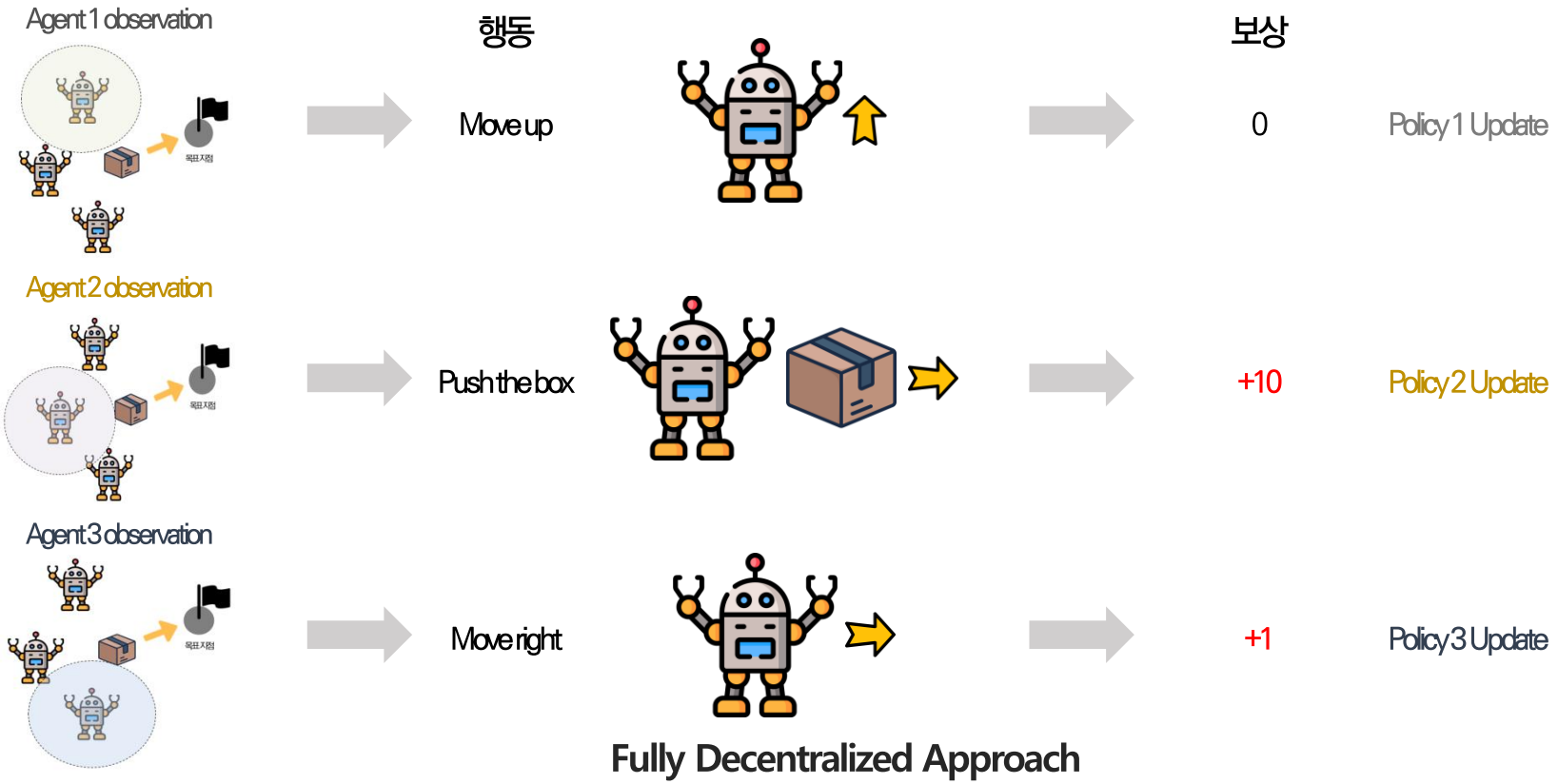
- ❖ Fully Decentralized Approach: 각 에이전트가 자신의 로컬 정보만을 바탕으로 독립적으로 행동하고 결정을 내리는 방식
- ❖ 다른 에이전트의 행동을 고려하거나 협력하지 않고 독립적으로 행동을 선택 및 실행하고 그에 따른 보상을 최대화하는 정책을 탐색하는 것이 목표



Methods

Fully Centralized and Decentralized Approach

- ❖ Fully Decentralized Approach: 각 에이전트가 자신의 로컬 정보만을 바탕으로 독립적으로 행동하고 결정을 내리는 방식
- ❖ 다른 에이전트의 행동을 고려하거나 협력하지 않고 독립적으로 행동을 선택 및 실행하고 그에 따른 보상을 최대화하는 정책을 탐색하는 것이 목표



Fully Decentralized Approach

Methods

Fully Centralized and Decentralized Approach

Fully Decentralized Approach 장단점

❖ **장점**

- 각 에이전트가 자신의 정보만으로 결정을 내려, 정보 처리 부담이 중앙에 집중되지 않음

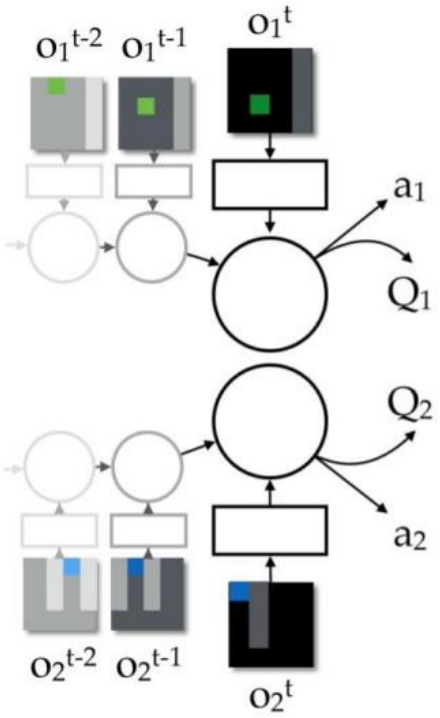
❖ **단점**

- 에이전트들이 독립적으로 학습하면서 동일 작업을 중복하거나 상충되는 행동을 할 수 있음
- 정보 공유가 없어 환경의 역학을 파악하는 것이 어려움
- 이로 인해, 보상이 자신의 행동 때문인지 팀원의 행동 때문인지 구분하기 어려운 **spurious reward** 문제가 발생할 수 있음

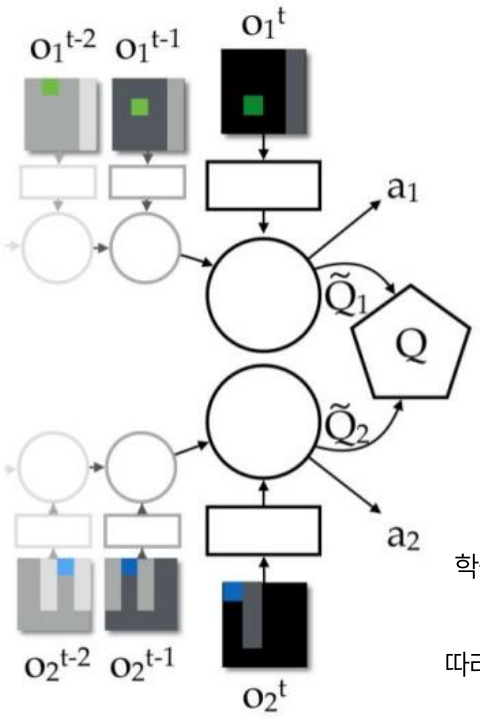
Methods

Value-Decomposition Network (VDN)

- ❖ Centralized, decentralized 접근법의 장점만을 사용하는 CTDE 학습 방식을 사용
- ❖ 즉, 학습할 때는 팀의 전체 보상을 활용하여 학습하고 실제 테스트 시에는 에이전트가 독립적으로 실행
- ❖ **Additive factorization**을 통해 팀 보상을 개별 에이전트의 가치 함수로 분해하여, 각 에이전트가 자신의 지역 정보만으로도 팀 전체 성과를 극대화할 수 있도록 학습



Independent agents



Value-Decomposition Network

에이전트 주변 정보 (observation)

에이전트 수

$$Q\left(\left(h^1, h^2, \ldots, h^d\right), \left(a^1, a^2, \ldots, a^d\right)\right) \approx \sum_{i=1}^{d} \tilde{Q}_i\left(h^i, a^i\right)$$

팀 전체 보상

개별 에이전트 보상

Q-Learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)$$

손실 함수

개별 가치 함수 ($\tilde{Q}_i(h^i, a^i)$)는 팀 전체 보상으로부터의 backpropagation에 의해 학습됨

학습이 진행되면서, 각 에이전트는 자신의 로컬 관찰에 기반한 최적의 행동 선택

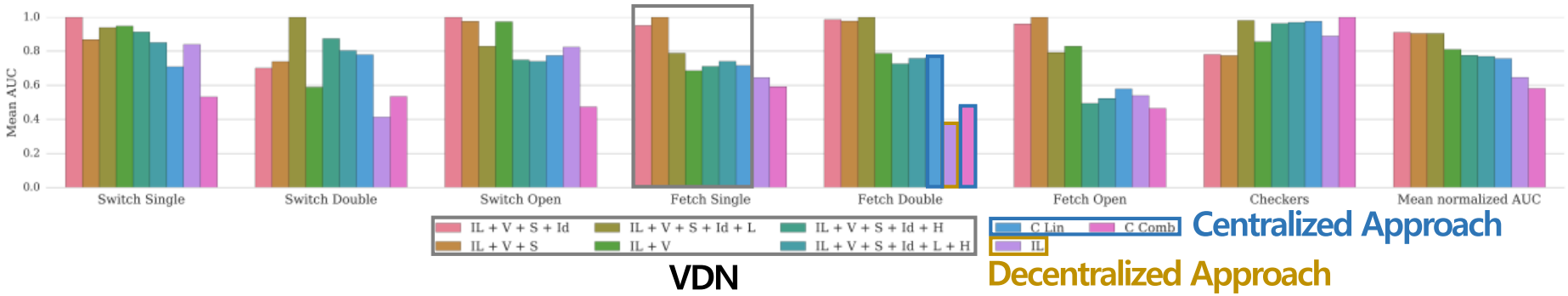
이 행동은 팀 보상을 극대화하기 위한 것
따라서 팀 보상에 더 많은 기여하는 행동을 한 에이전트는 더 높은 Q-값을 얻게 됨

간접적으로 기여도 반영 → **Credit Assignment 문제 해결**

Methods

Experiment ①

- ❖ 여러 환경에서 centralization, decentralization 접근법과 VDN의 성능 비교를 AUC를 통해 보여주고 있음
- ❖ VDN은 여러 component로 구성되어 있으며, 구성된 요소에 상관없이 centralization, decentralization 접근법 보다 더 우수한 성능을 보임



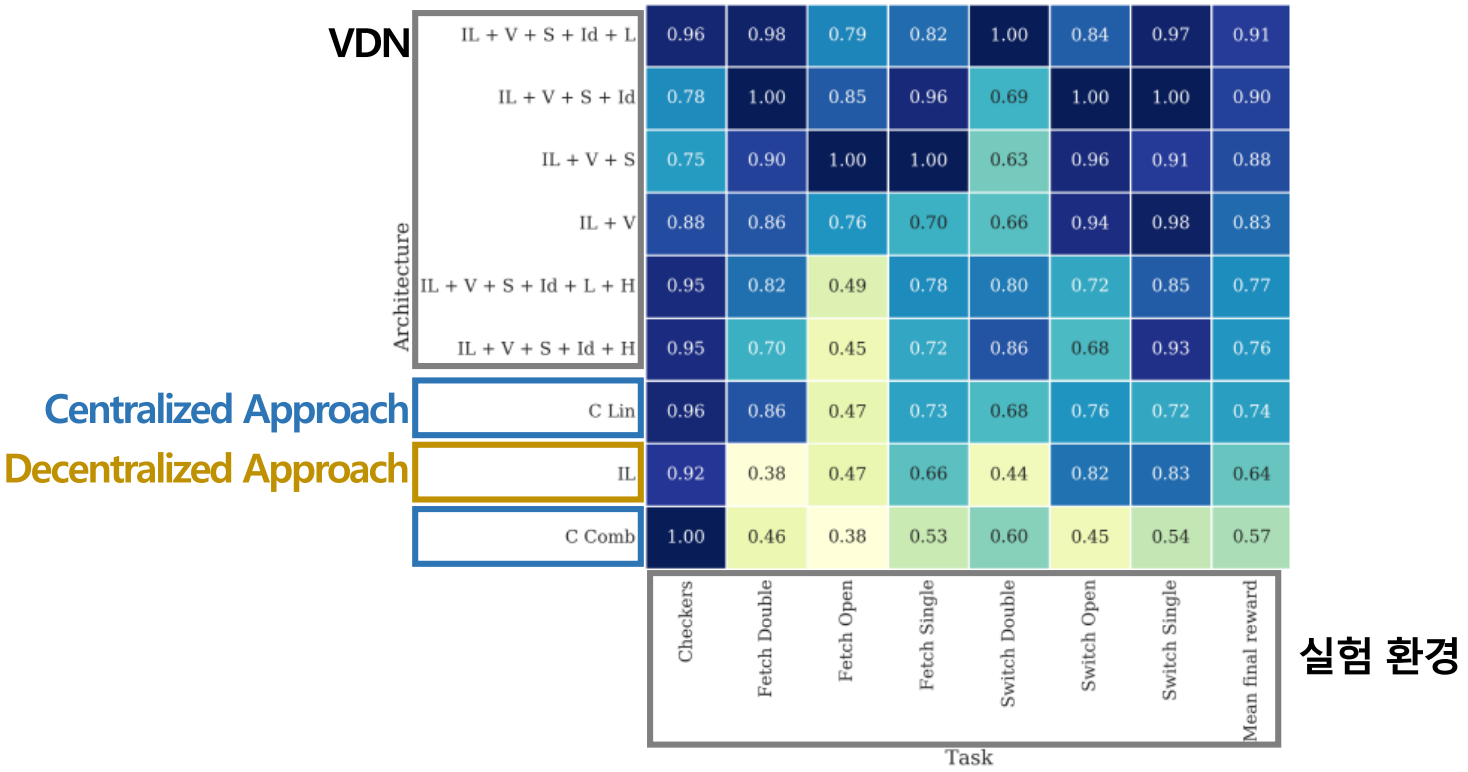
IL: **I**ndependent **L**earners – Decentralized Approach
V: **V**alue-Decomposition – 팀 보상을 개별 에이전트의 가치 함수로 분해하여 학습하는 방법
S: **S**hared Weights – 여러 에이전트가 동일한 네트워크 가중치를 공유하는 경우
Id: **I**dentity or Role Information – 각 에이전트가 고유한 역할 정보를 받는 경우
L: **L**ow-level Communication – 에이전트 간의 저수준 정보 공유
H: **H**igh-Level Communication – 에이전트 간의 고수준 정보 공유
C Lin: **C**entralized **L**inear – Centralized Approach + Linear Combination
C Comb: **C**ombinatorial **C**entralized – Centralized Approach + Complex Combination

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., ... & Graepel, T. (2017). Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296.

Methods

Experiment ②

- ❖ 여러 방법론이 여러 실험 환경에서 얻은 성능을 시각적으로 나타낸 히트맵
- ❖ 히트맵 내에 숫자는 해당 방법론이 각 환경에서 달성한 정규화된 최종 성능을 의미
- ❖ 색상은 성능의 정도를 나타내며, 어두운 파란색일수록 높은 성능, 밝은 녹색일수록 낮은 성능을 의미



Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., ... & Graepel, T. (2017). Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296.

Methods

QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning (2018, PMLR)

- ❖ 실제 현실에서 CTDE 학습 방식을 효율적으로 사용하는 것의 필요성을 언급
- ❖ VDN(2018, AAMAS)의 한계점
  - 팀의 전체 보상을 각 에이전트의 보상들의 단순 합으로 표현 → 선형 합산 방식 → 복잡성 크게 제한
  - 학습 중에 사용할 수 있는 추가적인 전체 상태 정보를 무시 → 복잡한 시나리오에서 성능 제한
  - 이로 인해, 학습할 때의 최적 정책이 실제 실행할 때의 최적 정책으로 이어지지 않을 가능성 존재
- ❖ 따라서, 비선형 조합 방식(+단조성 추가)과 추가 상태 정보를 사용하는 QMIX 제안

Decentralized Execution　　Centralized Training

QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning

Tabish Rashid [*1] Mikayel Samvelyan [*2] Christian Schroeder de Witt [1]
Gregory Farquhar [1] Jakob Foerster [1] Shimon Whiteson [1]

Abstract

In many real-world settings, a team of agents must coordinate their behaviour while acting in a decentralised way. At the same time, it is often possible to train the agents in a centralised fashion in a simulated or laboratory setting, where global state information is available and communi-

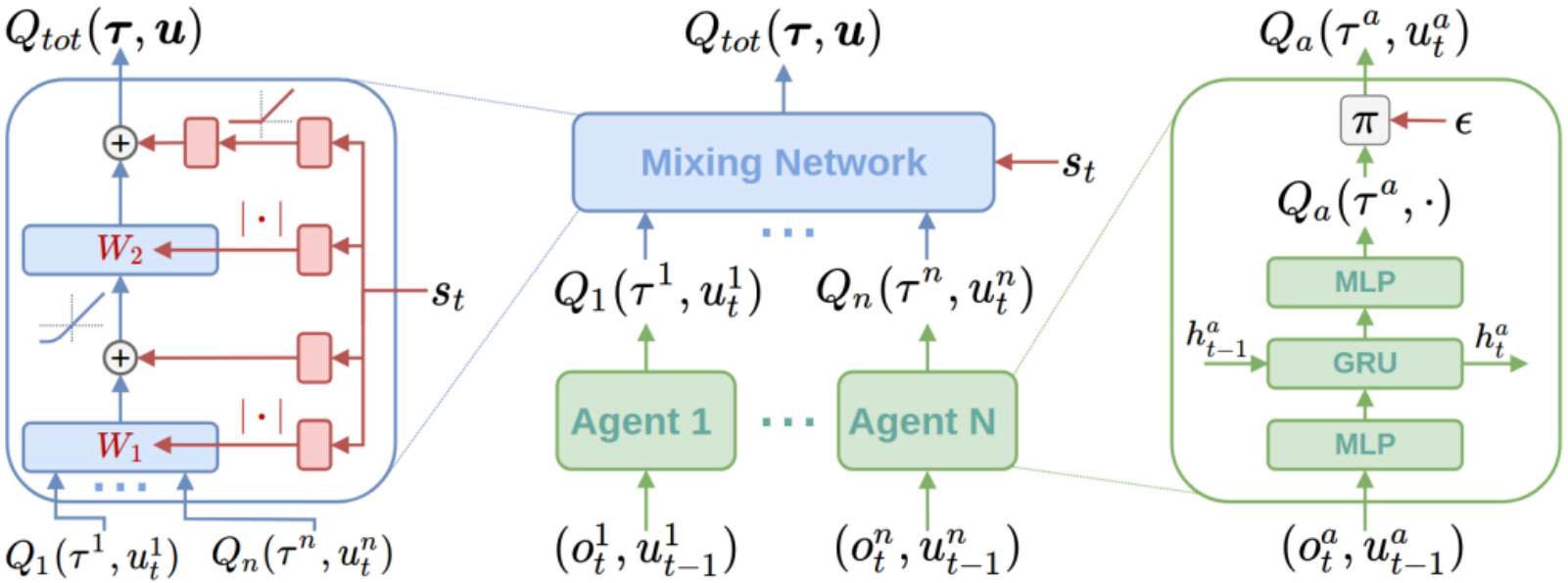
(a) 5 Marines map

(b) 2 Stalkers & 3 Zealots map

Methods

QMIX – Overall Architecture

- ❖ QMIX는 **에이전트 네트워크**와 **mixing 네트워크**로 구성되어 있음
- ❖ 에이전트 네트워크는 각 에이전트의 가치 함수를 산출하는 역할
- ❖ Mixing 네트워크
  - 개별 에이전트의 가치 함수를 비선형적으로 결합하여 전체 공동 행동 가치 함수 Q_{tot} 생성
  - 추가 상태 정보를 활용한 단조성 제약을 적용하여 CT와 DE간의 일관성을 보장



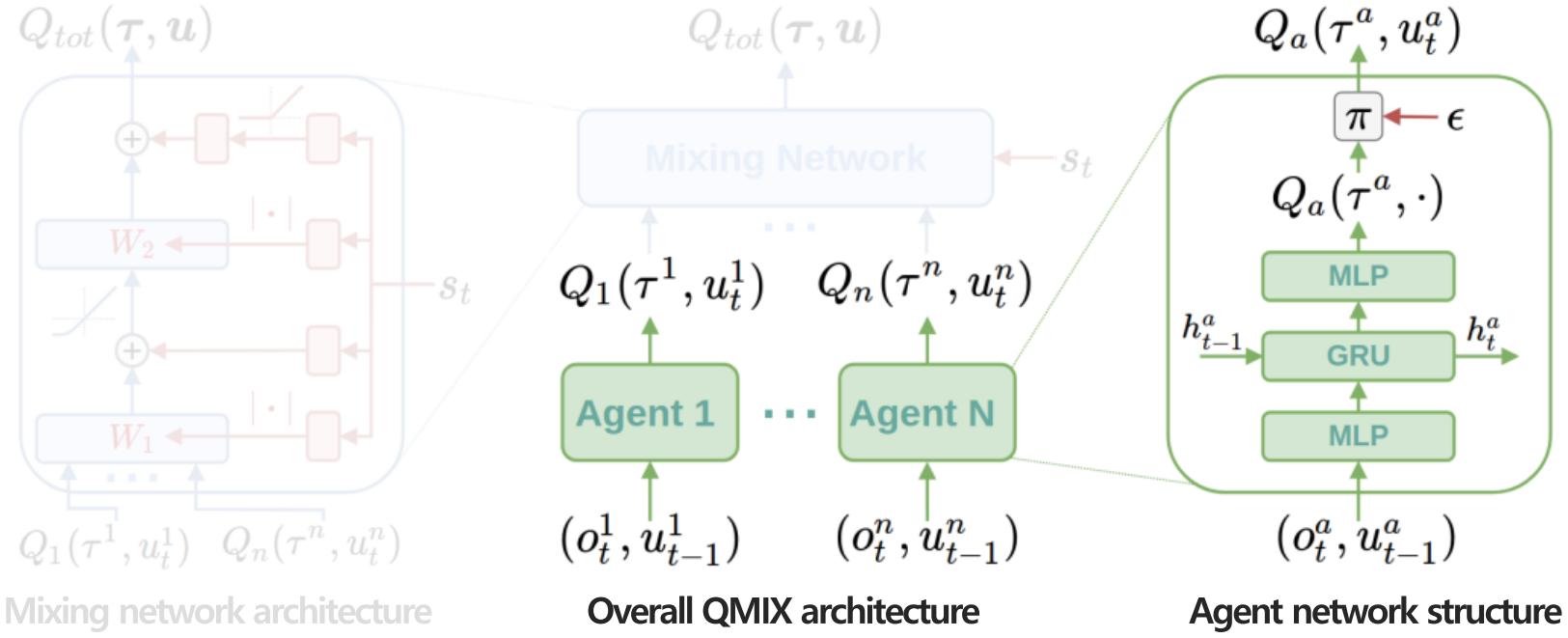
Mixing network architecture | Overall QMIX architecture | Agent network structure

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

QMIX – Agent network

- QMIX는 에이전트 네트워크와 mixing 네트워크로 구성되어 있음
- **에이전트 네트워크**는 각 에이전트의 가치 함수를 산출하는 역할
- Mixing 네트워크
  - 개별 에이전트의 가치 함수를 비선형적으로 결합하여 전체 공동 행동 가치 함수 Q_{tot} 생성
  - 추가 상태 정보를 활용한 단조성 제약을 적용하여 CT와 DE간의 일관성을 보장

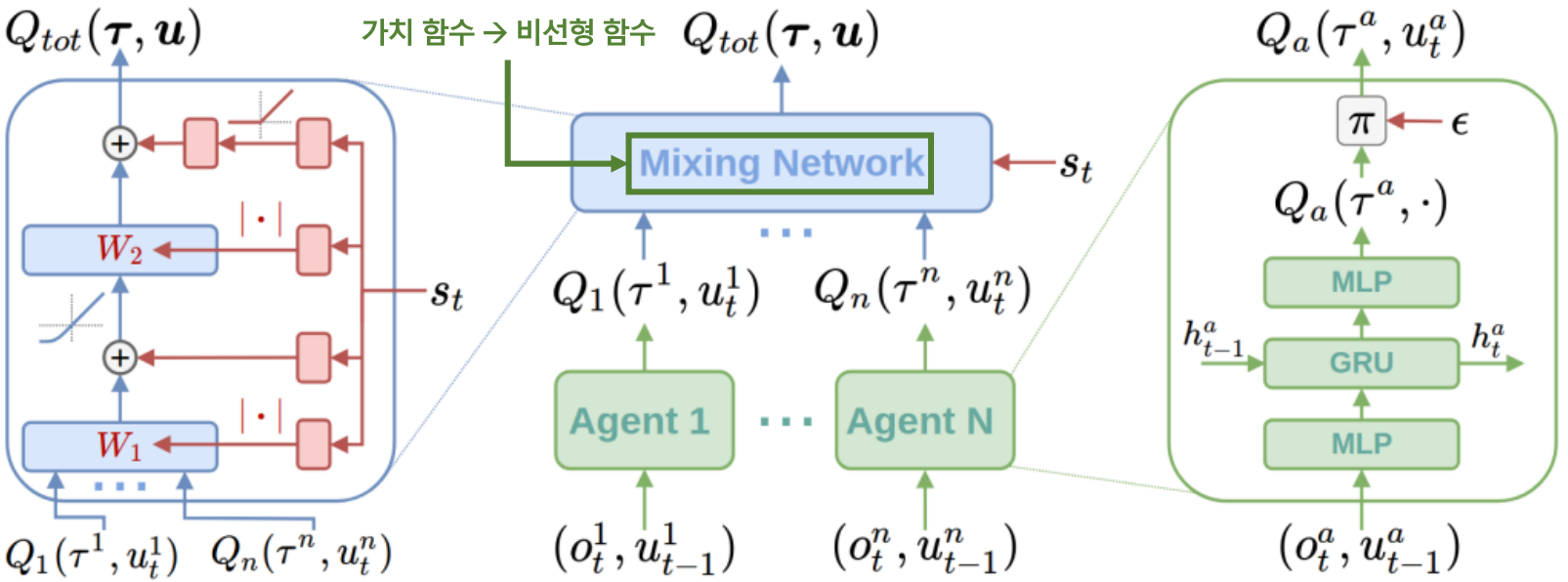


Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

QMIX – Mixing network

- ❖ QMIX는 에이전트 네트워크와 mixing 네트워크로 구성되어 있음
- ❖ 에이전트 네트워크는 각 에이전트의 가치 함수를 산출하는 역할
- ❖ **Mixing 네트워크**
  - 개별 에이전트의 가치 함수를 비선형적으로 결합하여 전체 공동 행동 가치 함수 Q_{tot} 생성
  - 추가 상태 정보를 활용한 단조성 제약을 적용하여 CT와 DE간의 일관성을 보장



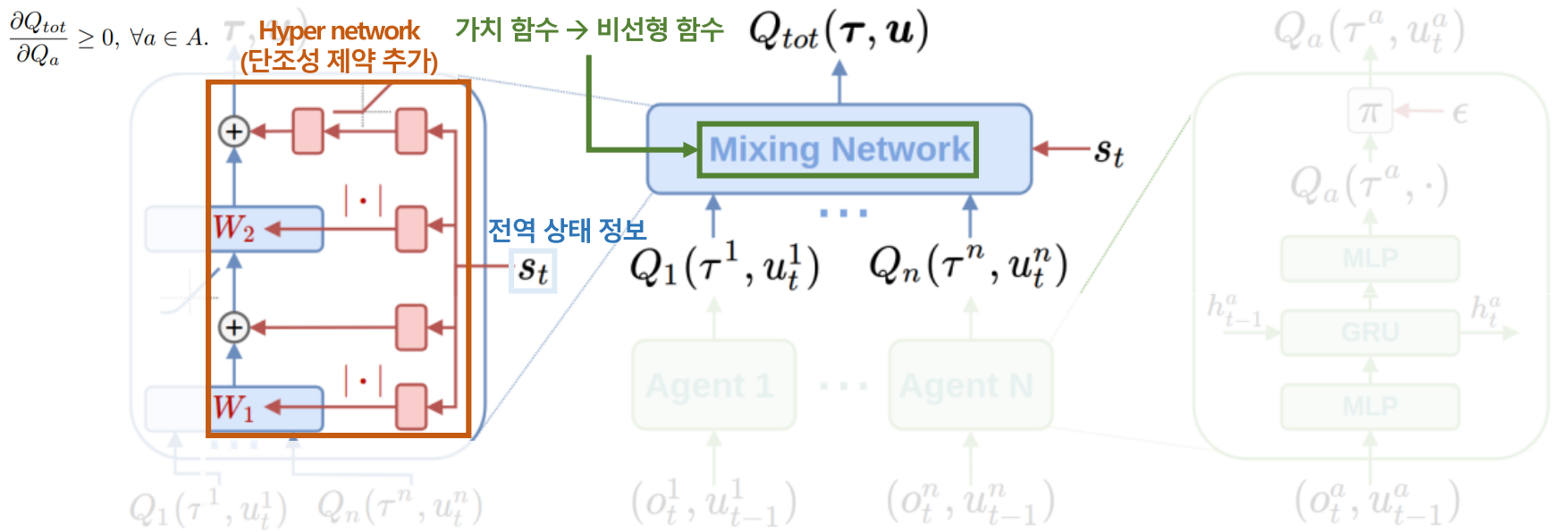
Mixing network architecture | Overall QMIX architecture | Agent network structure

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

QMIX – Mixing network

- ❖ QMIX는 에이전트 네트워크와 mixing 네트워크로 구성되어 있음
- ❖ 에이전트 네트워크는 각 에이전트의 가치 함수를 산출하는 역할
- ❖ **Mixing 네트워크**
  - 개별 에이전트의 가치 함수를 비선형적으로 결합하여 전체 공동 행동 가치 함수 Q_{tot} 생성
  - 추가 상태 정보를 활용한 단조성 제약을 적용하여 CT와 DE간의 일관성을 보장



Mixing network architecture | Overall QMIX architecture | Agent network structure

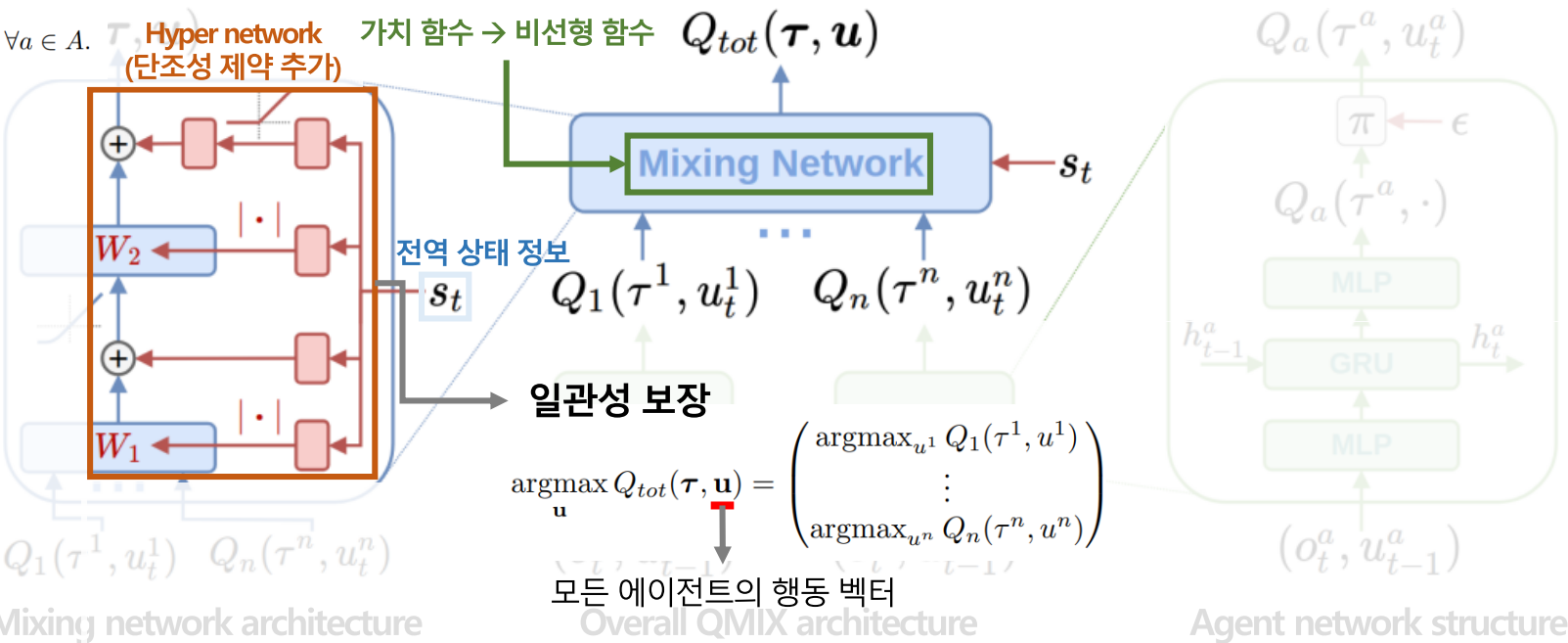
Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

QMIX – Mixing network

- ❖ QMIX는 에이전트 네트워크와 mixing 네트워크로 구성되어 있음
- ❖ 에이전트 네트워크는 각 에이전트의 가치 함수를 산출하는 역할
- ❖ **Mixing 네트워크**
  - 개별 에이전트의 가치 함수를 비선형적으로 결합하여 전체 공동 행동 가치 함수 Q_{tot} 생성
  - 추가 상태 정보를 활용한 단조성 제약을 적용하여 CT와 DE간의 일관성을 보장

$\frac{\partial Q_{tot}}{\partial Q_a} \geq 0, \forall a \in A.$



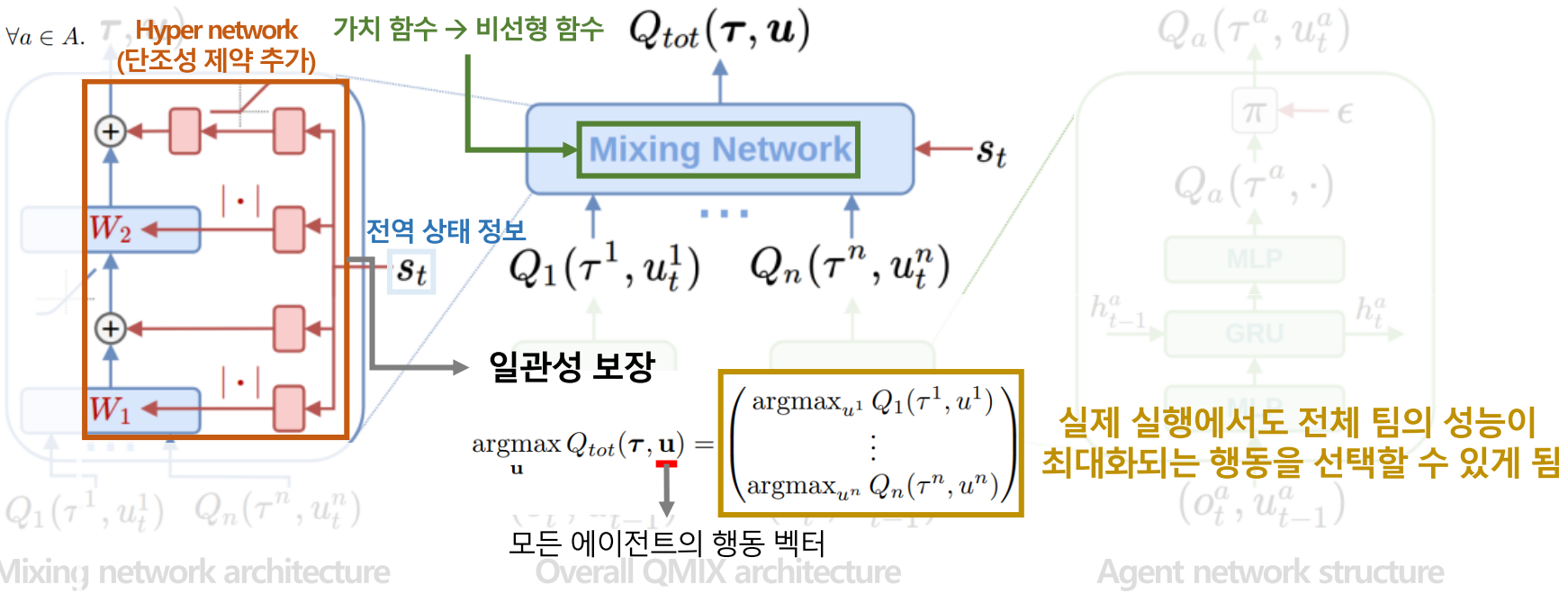
Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

QMIX – Mixing network

- ❖ QMIX는 에이전트 네트워크와 mixing 네트워크로 구성되어 있음
- ❖ 에이전트 네트워크는 각 에이전트의 가치 함수를 산출하는 역할
- ❖ **Mixing 네트워크**
  - 개별 에이전트의 가치 함수를 비선형적으로 결합하여 전체 공동 행동 가치 함수 Q_{tot} 생성
  - 추가 상태 정보를 활용한 단조성 제약을 적용하여 CT와 DE간의 일관성을 보장

$\frac{\partial Q_{tot}}{\partial Q_a} \geq 0, \forall a \in A.$

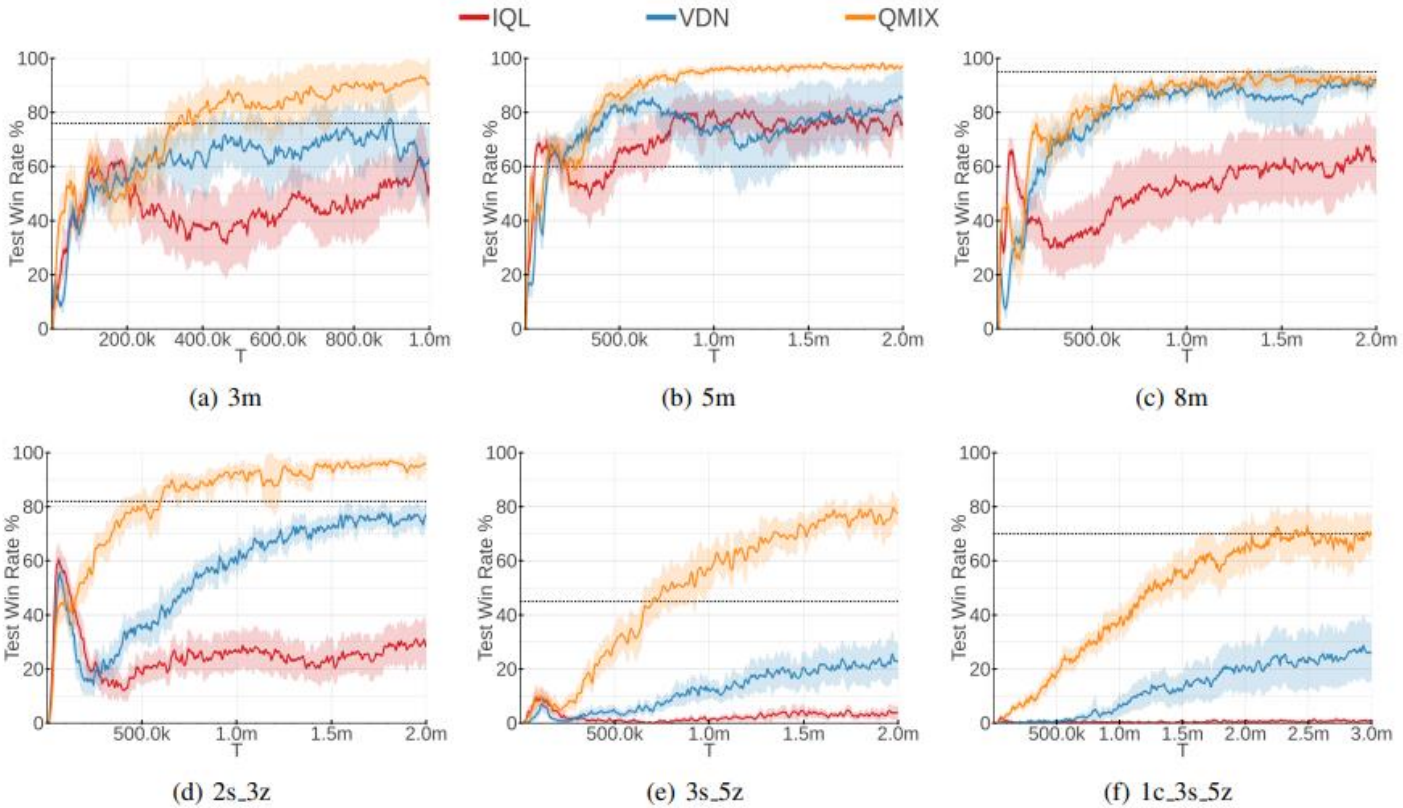


Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

Experiment ①

- ❖ StarCraft Ⅱ 환경에서 실험을 돌렸으며, 각 에이전트는 시나리오에서 move[direction], attack[enemy id], stop and noop의 discrete action space를 가짐 → direction: 4방향, 적 유닛이 shooting range 안에 있다면 사격
- ❖ **IQL**: 다중 에이전트 문제를 각각의 에이전트들이 독립적으로 해결하는 단일 에이전트 문제로 전환
  - **에이전트들의 정책 변화로** 인해 발생하는 **비정상성** 문제 때문에 불안정하고 저조한 성능을 보임

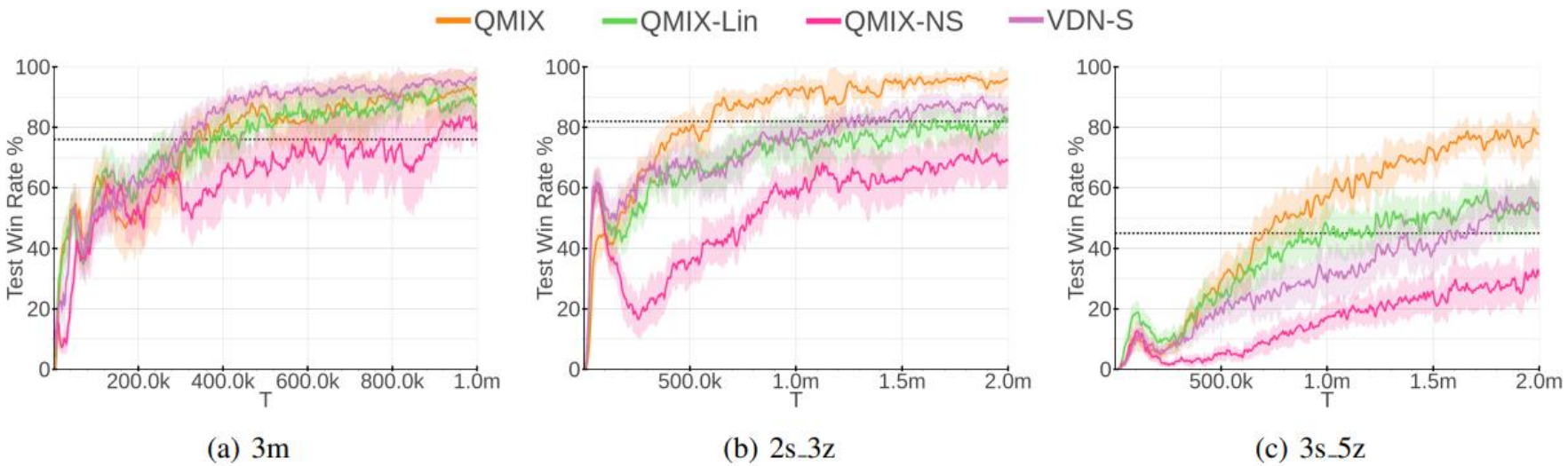


Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

Experiment ②

- ❖ QMIX의 변형과 수정된 VDN간의 성능 비교를 통해 **비선형 조합 방식, 전역 상태 정보 활용의 중요성**을 확인
- ❖ Figure (a): **동질적인 에이전트 유형**이 있는 맵에서 비선형 조합 방식이 꼭 필요하지 않음을 보여주고 있음
- ❖ Figure (b) & (c): **이질적인 에이전트**가 있는 맵에서 전역 상태 정보와 비선형 조합 방식의 사용은 좋은 성능을 달성하는데 꼭 필요하다는 것을 확인
- ❖ **전역 상태 정보를 활용하는 VDN-S**와 **선형 조합 방식을 사용하는 QMIX-Lin**간 성능 비교를 통해 전역 상태 정보를 온전히 활용하기 위해서는 비선형 조합 방식이 꼭 필요함을 확인할 수 있음



Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.

Methods

QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning (2019, ICLR)

- 다중 에이전트 강화학습 문제에서 CTDE 학습 방식이 각광 받고 있으며, 본 연구에서도 사용한다고 언급
- 기존 연구인 VDN(2018, AAMAS)과 QMIX(2018, PMLR)의 한계점
  - 가산성(additivity)과 단조성(monotonicity)이라는 구조적 제약으로 인해, 해결할 수 있는 문제의 범위가 제한적이라는 한계점을 지님
  - 제한적인 문제 상황: 에이전트 간의 상호작용이 비선형적이거나 비협력적 행동을 고려하지 못하는 상황
- 따라서, 구조적 제약에서 자유로운 새로운 가치 분해 방법을 적용한 QTRAN 제안

QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement learning

Kyunghwan Son [1] **Daewoo Kim** [1] **Wan Ju Kang** [1] **David Hostallero** [1] **Yung Yi** [1]

Abstract

We explore value-based solutions for multi-agent reinforcement learning (MARL) tasks in the centralized training with decentralized execution (CTDE) regime popularized recently. However, VDN and QMIX are representative examples that use the idea of factorization of the joint action-neural networks, which can represent action-value functions and policy functions in reinforcement learning problems as a high-capacity function approximator. However, more complex tasks such as robot swarm control and autonomous driving, often modeled as cooperative multi-agent learning problems, still remain unconquered due to their high scales and operational constraints such as distributed execution.

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

VDN과 QMIX의 구조적 제약으로 인해 해결할 수 없는 문제

- ❖ 제한적인 문제 상황: 에이전트 간의 상호작용이 비선형적이거나 비협력적 행동을 고려하지 못하는 상황
- ❖ 간단한 매트릭스 게임을 통해 VDN과 QMIX의 한계점을 명시
- ❖ One step pay off matrix game
  - 매트릭스 게임: 다중 에이전트 강화학습 또는 게임 이론에서 사용되는 간단한 게임 구조로, 각 에이전트가 특정 행동을 선택하고 그 행동 조합에 따른 보상이 미리 정의된 행렬 형태의 보상 구조에 의해 결정되는 게임
  - one step은 한 에피소드의 최대 스텝이 1임을 의미
  - 총 2개의 에이전트가 존재하며, 각 에이전트가 선택할 수 있는 행동의 수는 3개
  - VDN과 QMIX는 $\epsilon = 1$로 설정하여 20,000 timestep동안 학습을 진행한 후 얻은 결과

u_1 \ u_2	A	B	C
A	**8**	-12	-12
B	-12	0	0
C	-12	0	0

Payoff of matrix game
: 각 행동 조합에 대한
사전에 정의된 공동 보상

Q_1 \ Q_2	-3.14(A)	**-2.29**(B)	-2.41(C)
-2.29(A)	-5.42	-4.57	-4.70
-1.22(B)	-4.35	-3.51	-3.63
-0.73(C)	-3.87	**-3.02**	-3.14

VDN

Q_1 \ Q_2	-0.92(A)	0.00(B)	**0.01**(C)
-1.02(A)	-8.08	-8.08	-8.08
0.11(B)	-8.08	0.01	**0.03**
0.10(C)	-8.08	0.01	0.02

QMIX

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

VDN과 QMIX의 구조적 제약으로 인해 해결할 수 없는 문제

- ❖ 제한적인 문제 상황: 에이전트 간의 상호작용이 비선형적이거나 비협력적 행동을 고려하지 못하는 상황
- ❖ 간단한 매트릭스 게임을 통해 VDN과 QMIX의 한계점을 명시
  - ▪ 기존 연구는 구조적 제약으로 인해, 에이전트 간의 비선형적인 상호작용 또는 비협력적인 행동을 충분히 반영하지 못하기 때문에 팀 전체의 최적 성과를 놓치는 경우가 발생

u_1 \ u_2	A	B	C
A	**8**	-12	-12
B	-12	0	0
C	-12	0	0

Payoff of matrix game
: 각 행동 조합에 대한
사전에 정의된 공동 보상

Q_1 \ Q_2	-3.14(A)	**-2.29**(B)	-2.41(C)
-2.29(A)	-5.42	-4.57	-4.70
-1.22(B)	-4.35	-3.51	-3.63
-0.73(C)	-3.87	**-3.02**	-3.14

VDN

Q_1 \ Q_2	-0.92(A)	0.00(B)	**0.01**(C)
-1.02(A)	-8.08	-8.08	-8.08
0.11(B)	-8.08	0.01	**0.03**
0.10(C)	-8.08	0.01	0.02

QMIX

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

VDN과 QMIX의 구조적 제약으로 인해 해결할 수 없는 문제

- 제한적인 문제 상황: 에이전트 간의 상호작용이 비선형적이거나 비협력적 행동을 고려하지 못하는 상황
- 간단한 매트릭스 게임을 통해 VDN과 QMIX의 한계점을 명시
  - 기존 연구는 구조적 제약으로 인해, 에이전트 간의 비선형적인 상호작용 또는 비협력적인 행동을 충분히 반영하지 못하기 때문에 팀 전체의 최적 성과를 놓치는 경우가 발생
- 반면에, QTRAN은 최적의 행동 조합을 탐색하여 사전에 정의된 공동 보상과 동일한 결과를 얻음

u_1 \ u_2	A	B	C
A	**8**	-12	-12
B	-12	0	0
C	-12	0	0

Payoff of matrix game
: 각 행동 조합에 대한
사전에 정의된 공동 보상

Q_1 \ Q_2	-3.14(A)	**-2.29**(B)	-2.41(C)
-2.29(A)	-5.42	-4.57	-4.70
-1.22(B)	-4.35	-3.51	-3.63
-0.73(C)	-3.87	**-3.02**	-3.14

VDN

Q_1 \ Q_2	-0.92(A)	0.00(B)	**0.01**(C)
-1.02(A)	-8.08	-8.08	-8.08
0.11(B)	-8.08	0.01	**0.03**
0.10(C)	-8.08	0.01	0.02

QMIX

Q_1 \ Q_2	**4.16**(A)	2.29(B)	2.29(C)
3.84(A)	**8.00**	6.13	6.12
-2.06(B)	2.10	0.23	0.23
-2.25(C)	1.92	0.04	0.04

QTRAN(Q'_{jt})

u_1 \ u_2	A	B	C
A	**8.00**	-12.02	-12.02
B	-12.00	0.00	0.00
C	-12.00	0.00	-0.01

QTRAN(Q_{jt})

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN – 새로운 가치 분해 방법

- ❖ Credit assignment 문제를 해결하기 위한 가치 분해 방법은 Individual-Global-Max (IGM) 조건을 만족해야함
  - IGM 조건: 공동 행동 가치 함수의 최적 액션이 개별 행동 가치 함수들의 최적 행동과 동일해야함을 의미
  - IGM 조건을 만족할 때, "분해 가능하다 " 라고 말할 수 있음
- ❖ 기존 연구인 VDN과 QMIX는 각각 가산성과 단조성 제약을 통해 IGM 조건을 충족시킬 수 있는 충분 조건을 만족함

VDN **(Additivity)** $Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^{N} Q_i(\tau_i, u_i),$

QMIX **(Monotonicity)** $\frac{\partial Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u})}{\partial Q_i(\tau_i, u_i)} \geq 0, \quad \forall i \in \mathcal{N}.$

$\Longrightarrow$

IGM

$$\arg\max_{\boldsymbol{u}} Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \begin{pmatrix} \arg\max_{u_1} Q_1(\tau_1, u_1) \\ \vdots \\ \arg\max_{u_N} Q_N(\tau_n, u_N) \end{pmatrix}$$

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN – 새로운 가치 분해 방법

- ❖ Credit assignment 문제를 해결하기 위한 가치 분해 방법은 Individual-Global-Max (IGM) 조건을 만족해야함
  - IGM 조건: 공동 행동 가치 함수의 최적 액션이 개별 행동 가치 함수들의 최적 행동과 동일해야함을 의미
  - IGM 조건을 만족할 때, "분해 가능하다 " 라고 말할 수 있음
- ❖ QTRAN은 상태 보정 함수와 변환된 공동 행동 가치 함수를 도입함으로써, IGM 조건을 만족
  - 상태 보정 함수: 학습 과정에서 발생할 수 있는 오류나 불확실성을 수정하고, 더 정확한 가치를 계산하도록 도움
  - Affine transformation: 특정한 경우뿐만 아니라 다양한 상황에서도 일관되게 최적화를 수행하는데 도움

$$\sum_{i=1}^{N} Q_i(\tau_i, u_i) - Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) + V_{\text{jt}}(\boldsymbol{\tau}) = \begin{cases} 0 & \boldsymbol{u} = \bar{\boldsymbol{u}}, \quad (4a) \\ \geq 0 & \boldsymbol{u} \neq \bar{\boldsymbol{u}}, \quad (4b) \end{cases}$$

where

$$V_{\text{jt}}(\boldsymbol{\tau}) = \max_{\boldsymbol{u}} Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - \sum_{i=1}^{N} Q_i(\tau_i, \bar{u}_i).$$

$\Longrightarrow$

IGM

$$\arg\max_{\boldsymbol{u}} Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \begin{pmatrix} \arg\max_{u_1} Q_1(\tau_1, u_1) \\ \vdots \\ \arg\max_{u_N} Q_N(\tau_n, u_N) \end{pmatrix}$$

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN – 새로운 가치 분해 방법

- Credit assignment 문제를 해결하기 위한 가치 분해 방법은 Individual-Global-Max (IGM) 조건을 만족해야함
  - IGM 조건: 공동 행동 가치 함수의 최적 액션이 개별 행동 가치 함수들의 최적 행동과 동일해야함을 의미
  - IGM 조건을 만족할 때, "분해 가능하다 " 라고 말할 수 있음
- QTRAN은 상태 보정 함수와 변환된 공동 행동 가치 함수를 도입함으로써, IGM 조건을 만족
  - 상태 보정 함수: 학습 과정에서 발생할 수 있는 오류나 불확실성을 수정하고, 더 정확한 가치를 계산하도록 도움
  - Affine transformation: 특정한 경우뿐만 아니라 다양한 상황에서도 일관되게 최적화를 수행하는데 도움

$$\sum_{i=1}^{N} Q_i(\tau_i, u_i) - Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) + V_{\text{jt}}(\boldsymbol{\tau}) = \begin{cases} 0 & \boldsymbol{u} = \bar{\boldsymbol{u}}, \quad (4a) \\ \geq 0 & \boldsymbol{u} \neq \bar{\boldsymbol{u}}, \quad (4b) \end{cases}$$

where

$$V_{\text{jt}}(\boldsymbol{\tau}) = \max_{\boldsymbol{u}} Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - \sum_{i=1}^{N} Q_i(\tau_i, \bar{u}_i).$$

$$\begin{aligned} Q_{\text{jt}}(\boldsymbol{\tau}, \bar{\boldsymbol{u}}) &= \sum_{i=1}^{N} Q_i(\tau_i, \bar{u}_i) + V_{\text{jt}}(\boldsymbol{\tau}) \quad \text{(From (4a))} \\ &\geq \sum_{i=1}^{N} Q_i(\tau_i, u_i) + V_{\text{jt}}(\boldsymbol{\tau}) \\ &\geq Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) \quad \text{(From (4b))}. \end{aligned}$$

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN – 새로운 가치 분해 방법

- ❖ Credit assignment 문제를 해결하기 위한 가치 분해 방법은 Individual-Global-Max (IGM) 조건을 만족해야함
  - IGM 조건: 공동 행동 가치 함수의 최적 액션이 개별 행동 가치 함수들의 최적 행동과 동일해야함을 의미
  - IGM 조건을 만족할 때, "분해 가능하다 " 라고 말할 수 있음
- ❖ QTRAN은 상태 보정 함수와 변환된 공동 행동 가치 함수를 도입함으로써, IGM 조건을 만족
  - 상태 보정 함수: 학습 과정에서 발생할 수 있는 오류나 불확실성을 수정하고, 더 정확한 가치를 계산하도록 도움
  - Affine transformation: 특정한 경우뿐만 아니라 다양한 상황에서도 일관되게 최적화를 수행하는데 도움

$$\sum_{i=1}^{N} Q_i(\tau_i, u_i) - Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) + V_{\text{jt}}(\boldsymbol{\tau}) = \begin{cases} 0 & \boldsymbol{u} = \bar{\boldsymbol{u}}, \quad (4a) \\ \geq 0 & \boldsymbol{u} \neq \bar{\boldsymbol{u}}, \quad (4b) \end{cases}$$

where

$$V_{\text{jt}}(\boldsymbol{\tau}) = \max_{\boldsymbol{u}} Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - \sum_{i=1}^{N} Q_i(\tau_i, \bar{u}_i).$$

$$Q'_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) := \sum_{i=1}^{N} Q_i(\tau_i, u_i).$$

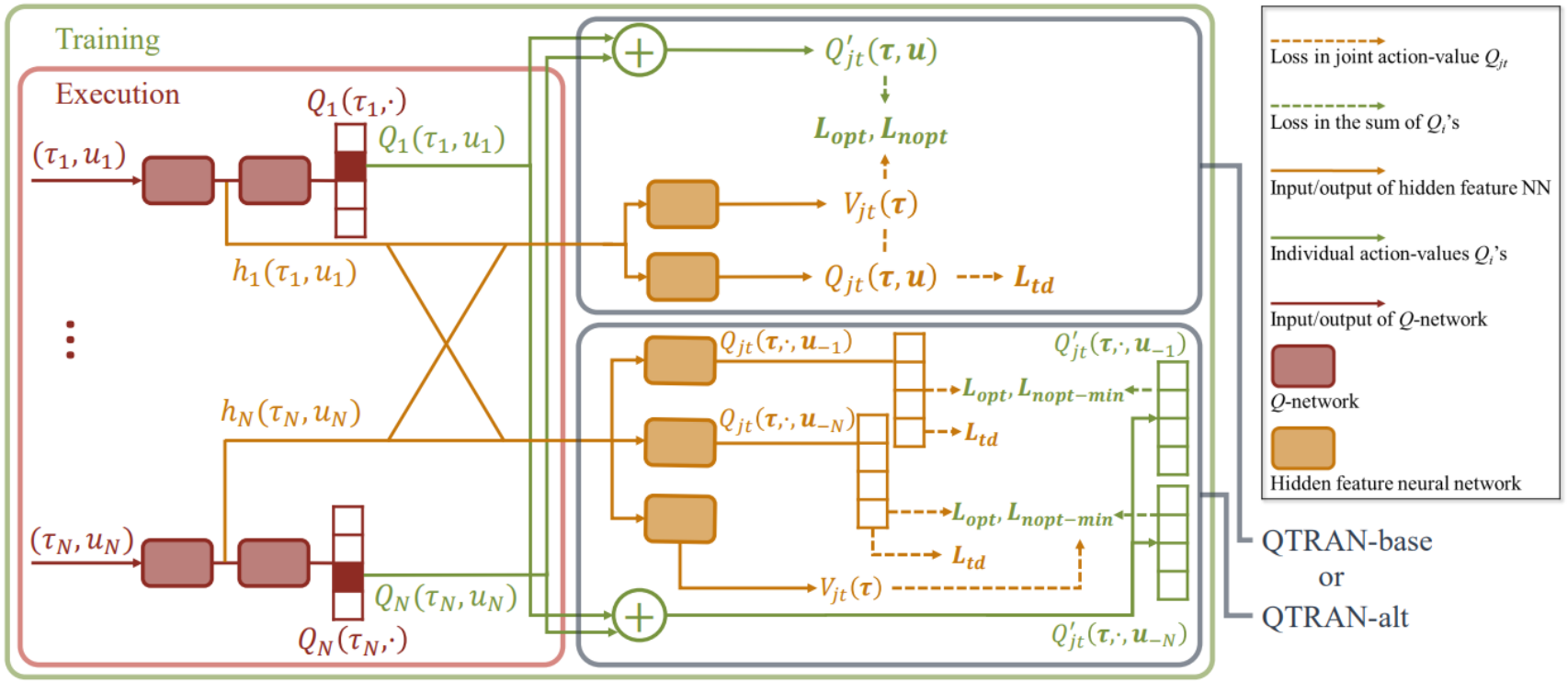
Affine transformation

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN – Overall Architecture

- ❖ QTRAN은 ①Individual action-value network, ②Joint action-value network, ③State-value network로 구성됨
- ❖ Individual action-value network: 각 에이전트의 히스토리를 입력으로 받아 개별 에이전트의 Q_i를 출력
- ❖ Joint action-value network: 각 에이전트의 히든 피처를 입력받아 Q_{jt} 출력
- ❖ State-value network: 각 에이전트의 상태 히든 피처를 입력 받아 V_{jt} 출력

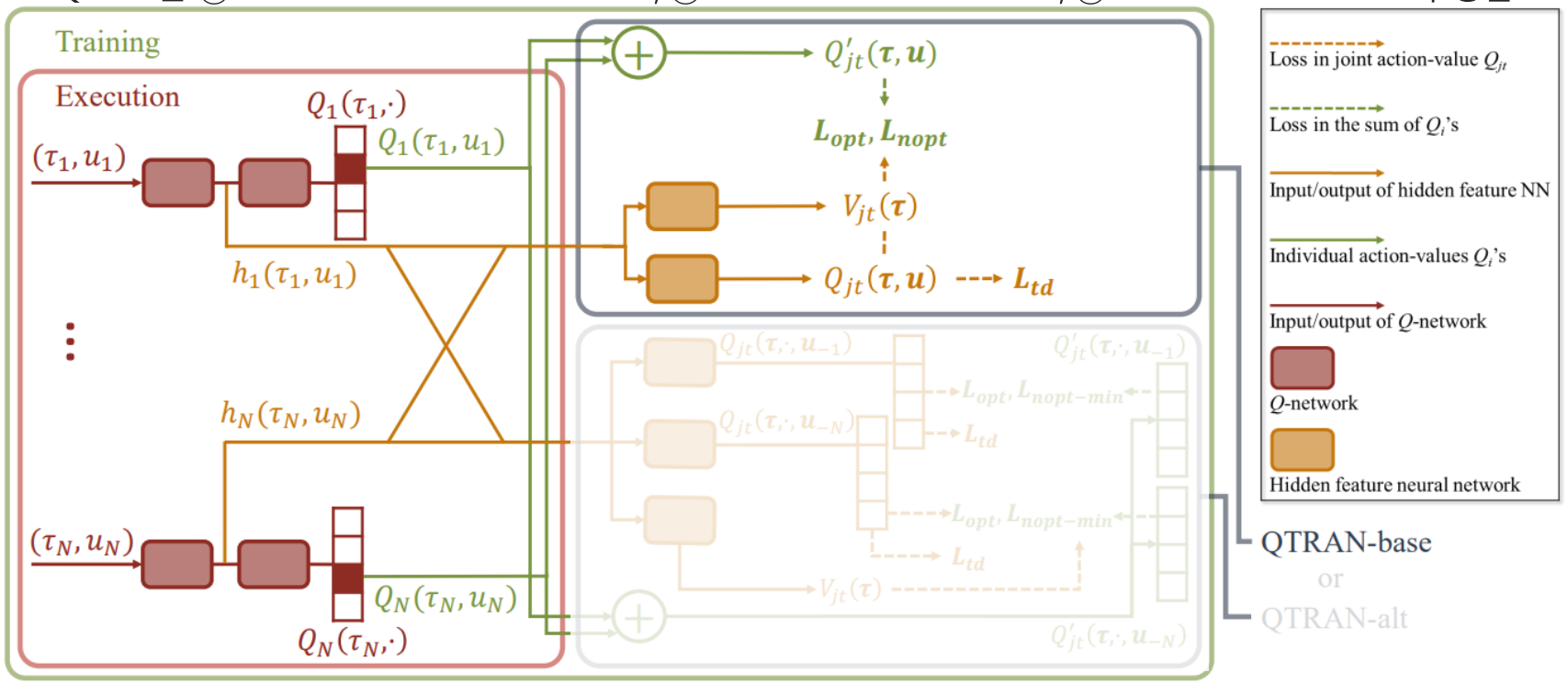


Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN-base

❖ QTRAN은 ①Individual action-value network, ②Joint action-value network, ③State-value network로 구성됨



$$L(\boldsymbol{\tau}, \boldsymbol{u}, r, \boldsymbol{\tau}'; \boldsymbol{\theta}) = L_{\text{td}} + \lambda_{\text{opt}} L_{\text{opt}} + \lambda_{\text{nopt}} L_{\text{nopt}}$$

$$L_{\text{td}}(;\boldsymbol{\theta}) = \left(Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - y^{\text{dqn}}(r, \boldsymbol{\tau}'; \boldsymbol{\theta}^-)\right)^2, \text{ Q-Learning loss}$$

$$L_{\text{opt}}(;\boldsymbol{\theta}) = \left(Q'_{\text{jt}}(\boldsymbol{\tau}, \bar{\boldsymbol{u}}) - \hat{Q}_{\text{jt}}(\boldsymbol{\tau}, \bar{\boldsymbol{u}}) + V_{\text{jt}}(\boldsymbol{\tau})\right)^2, \text{ 조건 (4a)를 만족시키도록 하는 손실 함수}$$

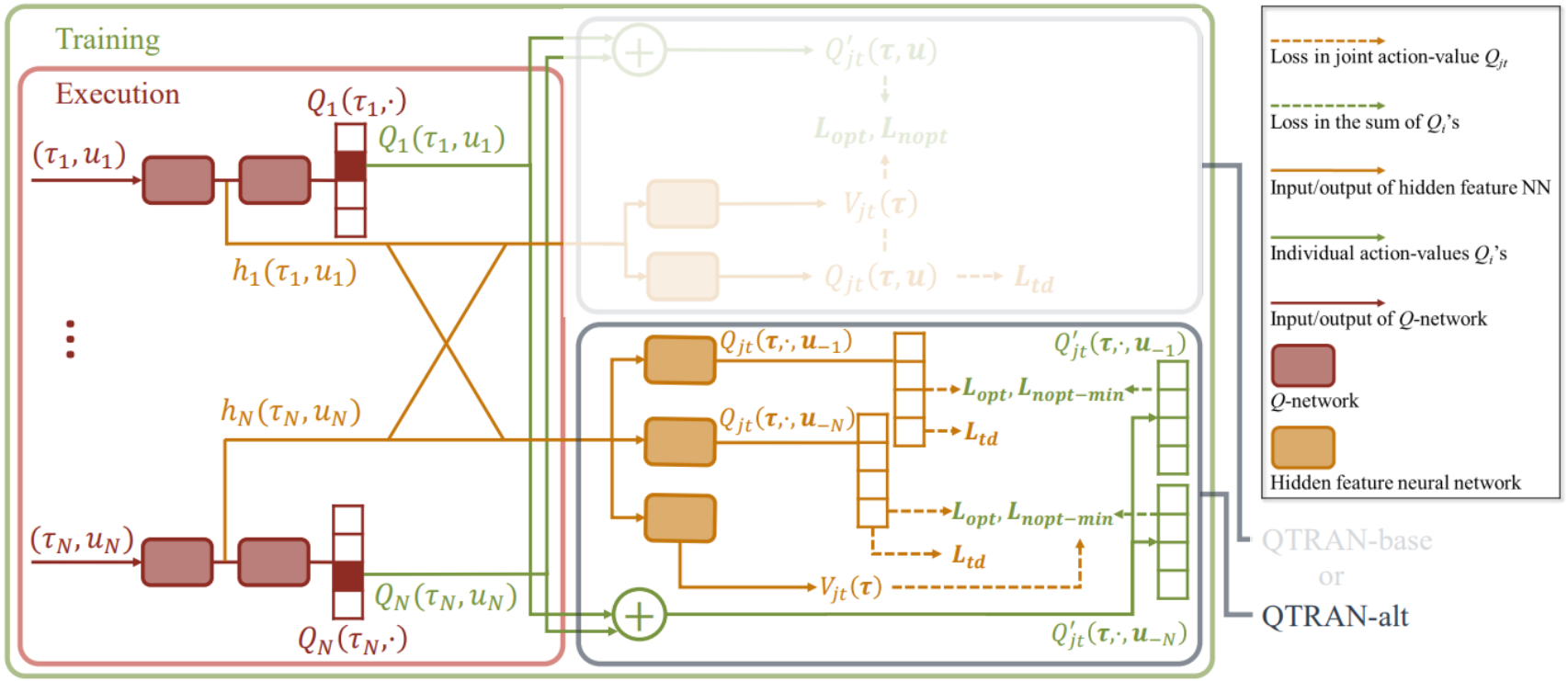
$$L_{\text{nopt}}(;\boldsymbol{\theta}) = \left(\min\left[Q'_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - \hat{Q}_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) + V_{\text{jt}}(\boldsymbol{\tau}), 0\right]\right)^2 \text{ 조건 (4b)를 만족시키도록 하는 손실 함수}$$

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

QTRAN-alt

❖ QTRAN-alt는 비최적 행동(조건 4b)에 대해 더 강력한 제약 조건을 부과하여 학습의 안정성과 수렴 속도를 개선한 프레임워크



$$\sum_{i=1}^{N} Q_i(\tau_i,u_i) - Q_{\mathrm{jt}}(\boldsymbol{\tau},\boldsymbol{u}) + V_{\mathrm{jt}}(\boldsymbol{\tau}) = \begin{cases} 0 & \boldsymbol{u}=\bar{\boldsymbol{u}}, \quad (4a) \\ \geq 0 & \boldsymbol{u}\neq\bar{\boldsymbol{u}}, \quad (4b) \end{cases} \Rightarrow \min_{u_i\in\mathcal{U}} \left[Q'_{\mathrm{jt}}(\boldsymbol{\tau},u_i,\boldsymbol{u}_{-i}) - Q_{\mathrm{jt}}(\boldsymbol{\tau},u_i,\boldsymbol{u}_{-i}) + V_{\mathrm{jt}}(\boldsymbol{\tau}) \right] = 0, \quad \forall i=1,\ldots,N,$$

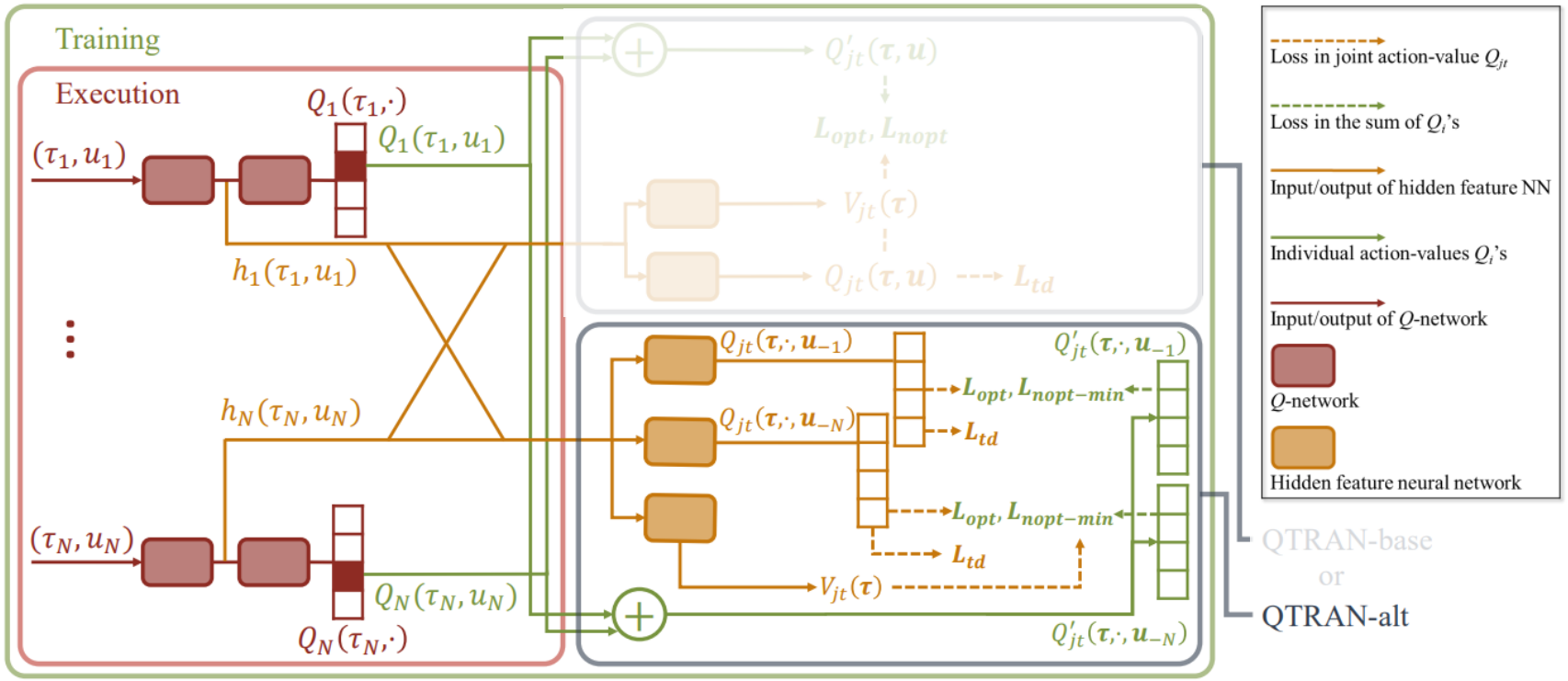
Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

$$L(\boldsymbol{\tau}, \boldsymbol{u}, r, \boldsymbol{\tau}'; \boldsymbol{\theta}) = L_{\text{td}} + \lambda_{\text{opt}} L_{\text{opt}} + \lambda_{\text{nopt}} L_{\text{nopt}}$$

QTRAN-alt

❖ QTRAN-alt는 비최적 행동(조건 4b)에 대해 더 강력한 제약 조건을 부과하여 학습의 안정성과 수렴 속도를 개선한 프레임워크



$$L_{\text{td}}(;\boldsymbol{\theta}) = \left(Q_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - y^{\text{dqn}}(r, \boldsymbol{\tau}'; \boldsymbol{\theta}^-)\right)^2,$$

$$L_{\text{opt}}(;\boldsymbol{\theta}) = \left(Q'_{\text{jt}}(\boldsymbol{\tau}, \bar{\boldsymbol{u}}) - \hat{Q}_{\text{jt}}(\boldsymbol{\tau}, \bar{\boldsymbol{u}}) + V_{\text{jt}}(\boldsymbol{\tau})\right)^2,$$

$$L_{\text{nopt}}(;\boldsymbol{\theta}) = \left(\min\left[Q'_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) - \hat{Q}_{\text{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) + V_{\text{jt}}(\boldsymbol{\tau}), 0\right]\right)^2 \Longrightarrow L_{\text{nopt-min}}(\boldsymbol{\tau}, \boldsymbol{u}, r, \boldsymbol{\tau}'; \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\min_{u_i \in \mathcal{U}} D(\boldsymbol{\tau}, u_i, \boldsymbol{u}_{-i}))^2$$

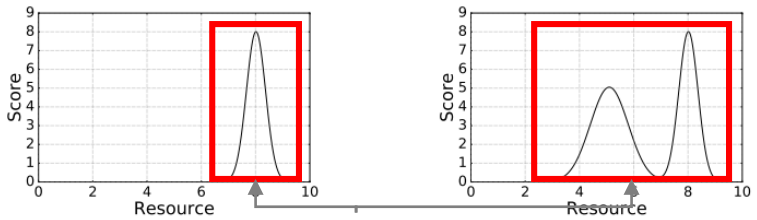
Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

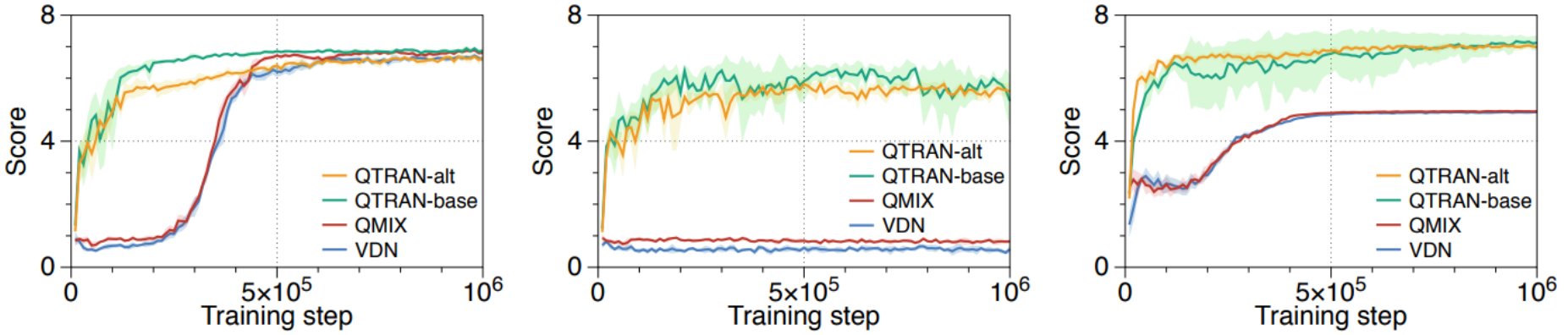
Experiment ①

- ❖ Gaussian Squeeze (GS)와 Multi-domain Gaussian Squeeze (MGS)는 자원 할당 문제를 다룬 환경
- ❖ GS는 자원 사용에서 단일 최적 지점을 찾는 환경이라면, MGS는 여러 최적 지점이 존재하는 더 복잡한 환경
- ❖ 그림 3(a): QTRAN의 수렴 속도가 눈에 띄게 차이가 나며, 비단조적 특성을 더 잘 처리하는 것을 확인할 수 있음
- ❖ 그림 3(b): epsilon-decay가 없는 상황에서도 QTRAN은 일관되게 성능을 유지
- ❖ 그림 3(c): 여러 개의 최적점이 존재하는 복잡한 환경에서도 각 에이전트가 효과적으로 자원을 배분하여 높은 보상을 유지

자원 사용에 따른 보상은 Gaussian 분포를 따름



(a) Gaussian Squeeze **비단조적 구조를 띔** (b) Multi-domain Gaussian Squeeze



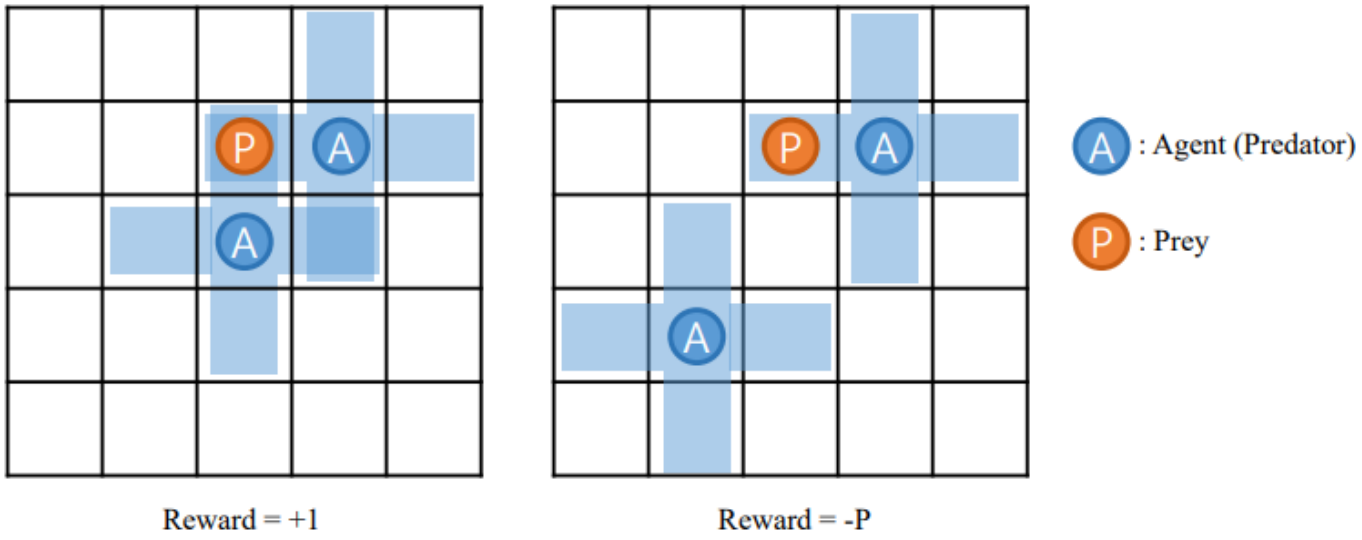
(a) Gaussian Squeeze (b) Gaussian Squeeze without epsilon decay (c) Multi-domain Gaussian Squeeze $K = 2$

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

Experiment ②

- ❖ Modified predator-prey 환경은 predator와 prey가 존재하며, predator는 협력하여 도망가는 prey를 잡아야 함
- ❖ Predator는 매 timestep마다 왼쪽, 오른쪽, 위, 아래, 멈추기 행동 중 하나 선택
- ❖ Prey는 매 timestep마다 5가지 행동 중 한 가지 행동을 무작위로 선택
- ❖ 보상 구조
  - 긍정적 보상: 두 명 이상의 포식자가 피식자를 동시에 잡을 때 팀 전체의 보상을 받음
  - 부정적 보상: 한 명의 포식자가 단독으로 피식자를 잡으면, 그 행동에 대한 페널티가 주어짐

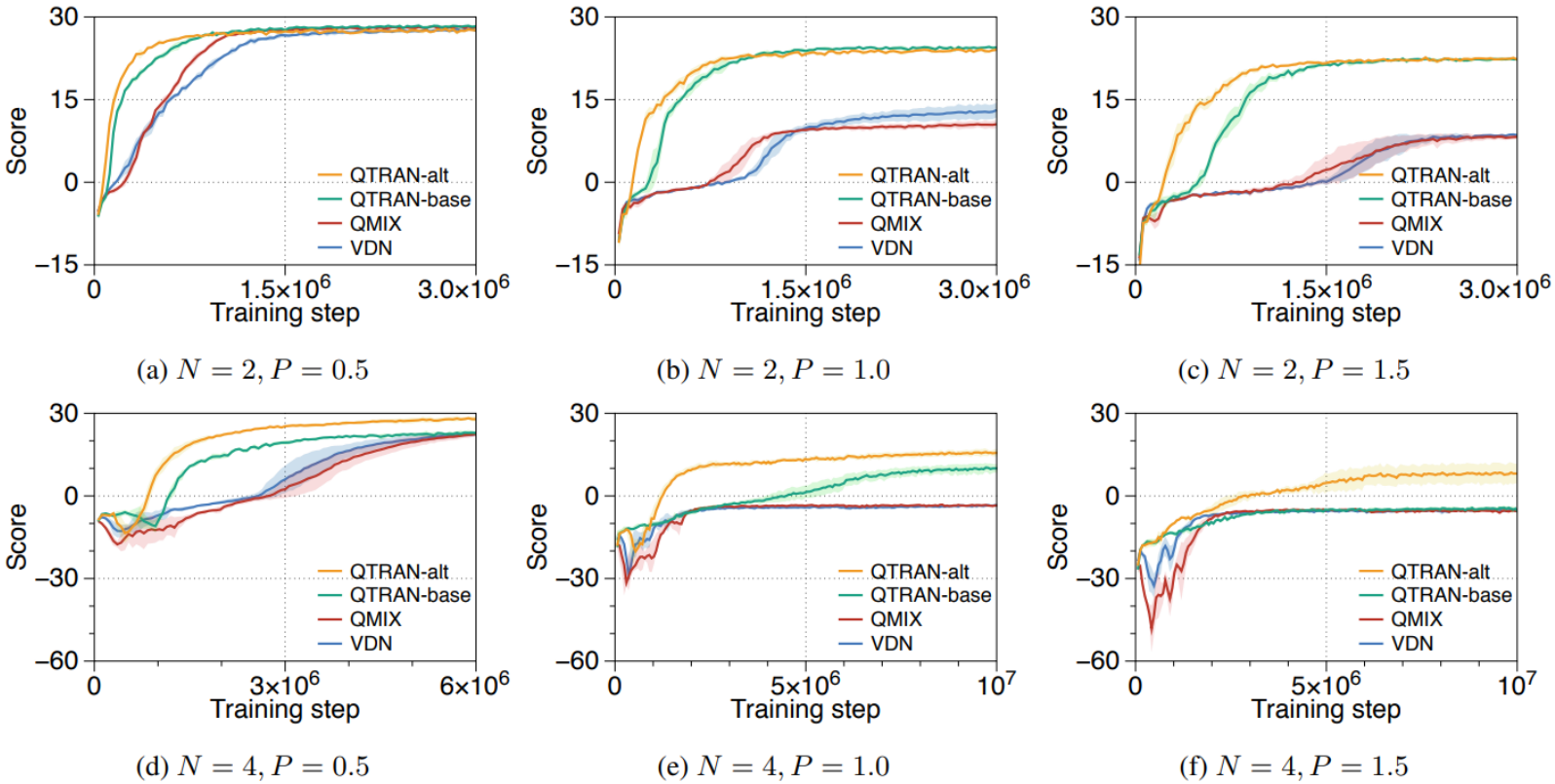


Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Methods

Experiment ②

- ❖ 해당 환경에서 predator의 수(N)를 2(prey 수: 1) 또는 4(prey 수:2)로 설정하여 실험 진행
- ❖ 페널티 값(P)에 변화를 주어, 이에 따른 각 방법론의 성능을 확인하고자 함
- ❖ Predator의 수 및 페널티 값 변화에 상관없이 QTRAN이 가장 우수한 성능을 보임



(a) $N = 2, P = 0.5$ (b) $N = 2, P = 1.0$ (c) $N = 2, P = 1.5$

(d) $N = 4, P = 0.5$ (e) $N = 4, P = 1.0$ (f) $N = 4, P = 1.5$

Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.

Conclusion

- 다중 에이전트 강화학습 도메인의 다양한 문제 중 Credit Assignment 문제에 초점을 맞춤
- 이를 해결하기 위한, 다양한 접근법 중 value decomposition 접근법을 사용한 방법론 3가지 설명
  - VDN: 가산성 제약 조건을 통해, credit assignment 문제를 간접적으로 해결
  - QMIX: 단조성 제약 조건을 통해, 해당 문제를 해결
  - QTRAN: 기존 연구의 구조적 제약으로 인해, 제한적인 상황을 해결하지 못하는 문제를 개선하는 새로운 가치 분해 방법을 제안함

Reference

1. Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., ... & Graepel, T. (2018, July). Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (pp. 2085-2087).
2. Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018, July). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In International Conference on Machine Learning (pp. 4295-4304). PMLR.
3. Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In International conference on machine learning (pp. 5887-5896). PMLR.
4. Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55(2), 895-943.

Thank you