Beyond QMIX: Advances in Multi-Agent Value Decomposition

Data Mining & Quality Analytics Open Seminar

2025.03.21

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

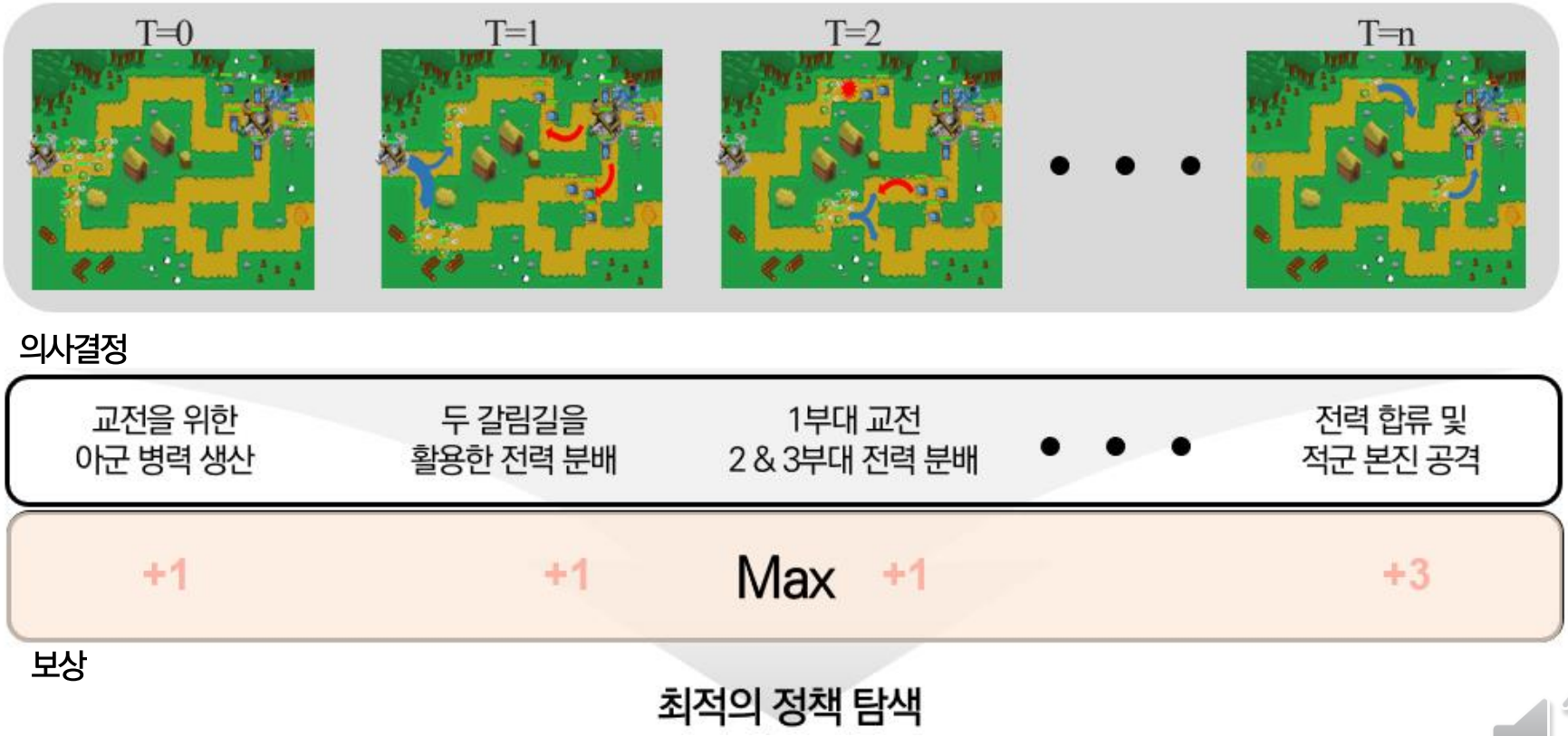
목차

- ❖ **Introduction**
- ❖ **Methods**
  - QPLEX (2021, ICLR)
  - TransMix (2022, AAAI)
  - ACORM (2024, ICLR)
- ❖ **Conclusion**

Introduction

강화학습 정의

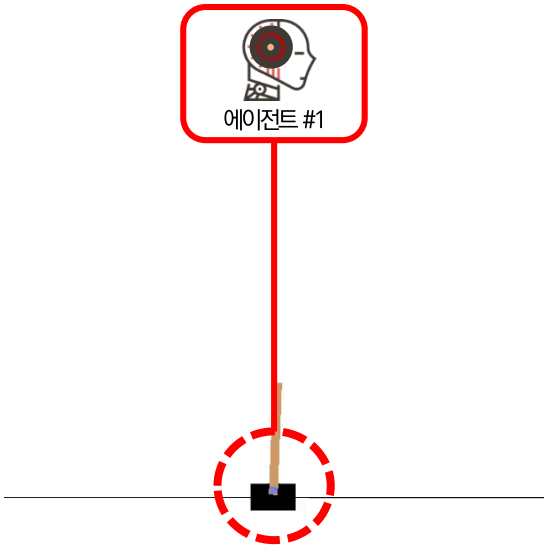
- ❖ 순차적인 의사결정 문제에서 에이전트가 환경으로부터 받는 누적 보상 값을 최대화하는 행동 정책을 학습하는 방법
- ❖ 정책: 에이전트가 특정 상태에서 어떤 행동을 선택할지 정해주는 함수



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



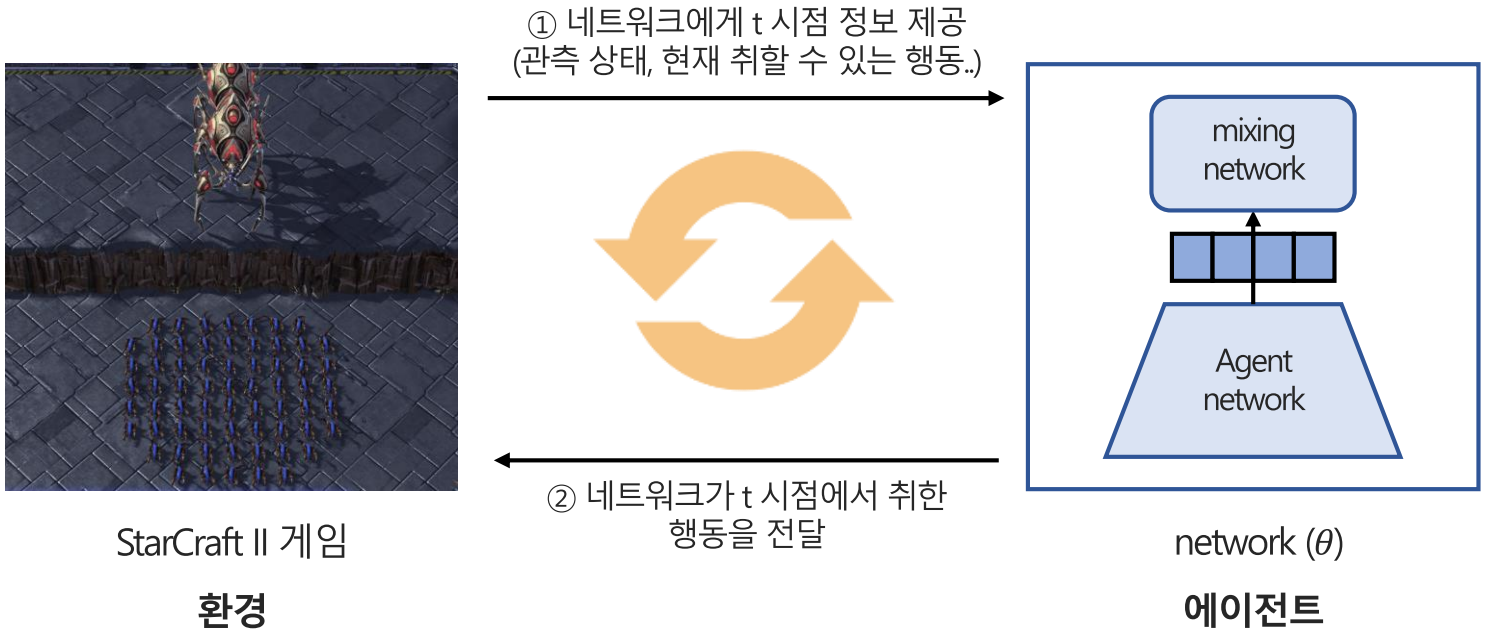
단일 에이전트 강화학습 (SARL)

다중 에이전트 강화학습 (MARL)

Introduction

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

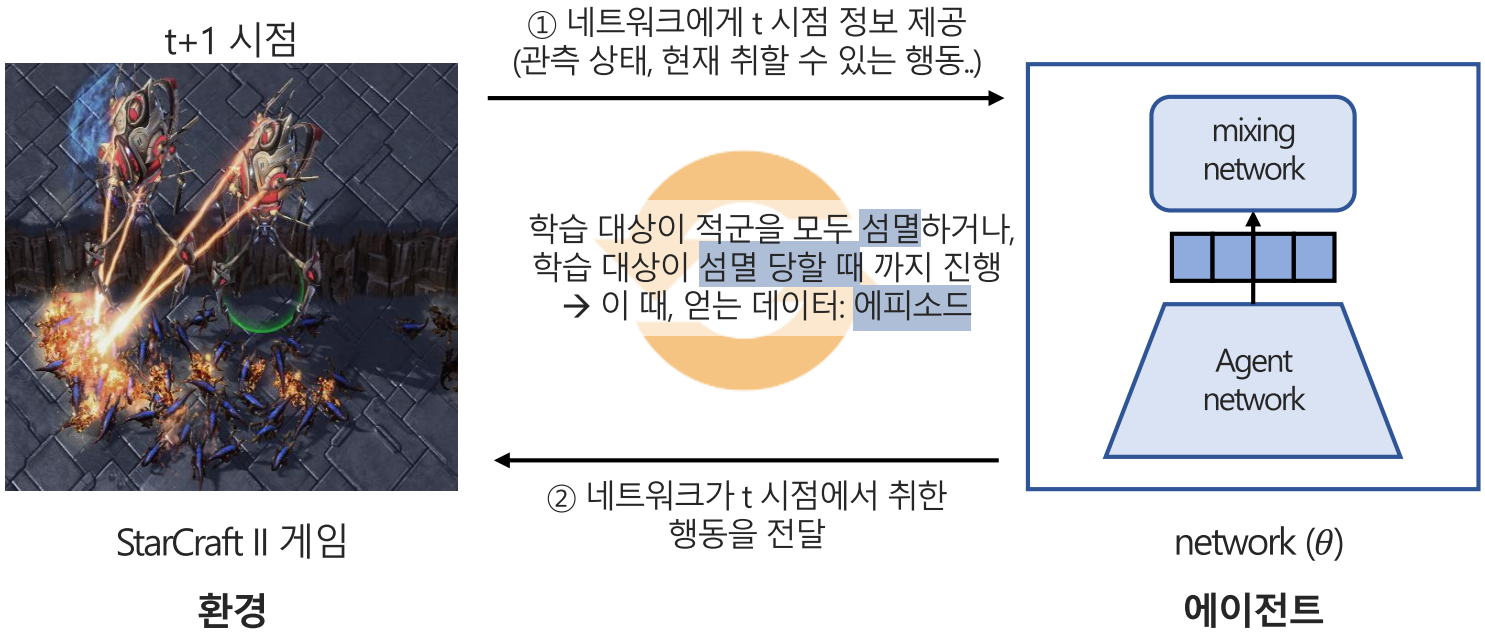


상호작용: 데이터를 수집하는 과정

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

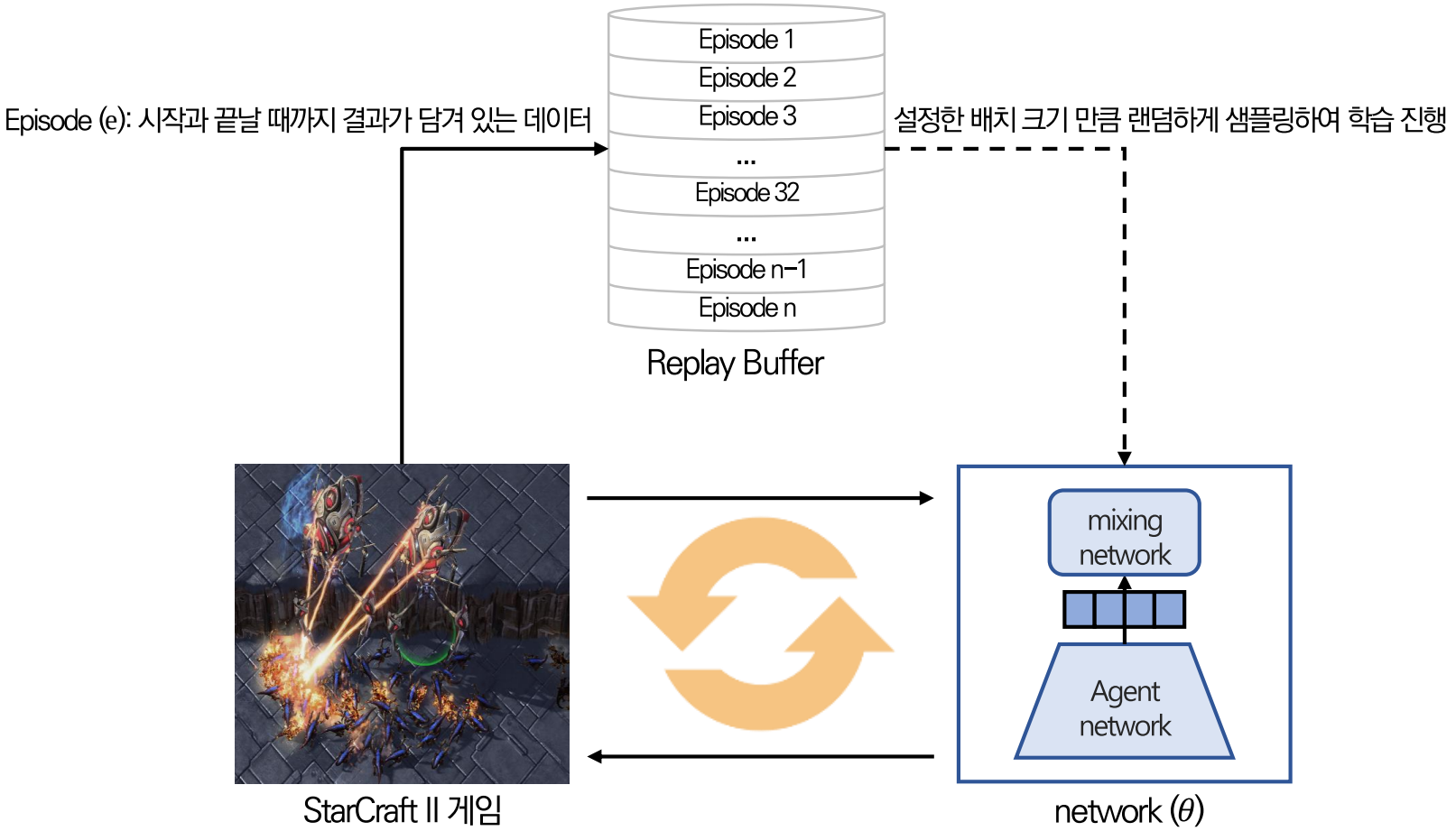


상호작용: 데이터를 수집하는 과정

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

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

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-

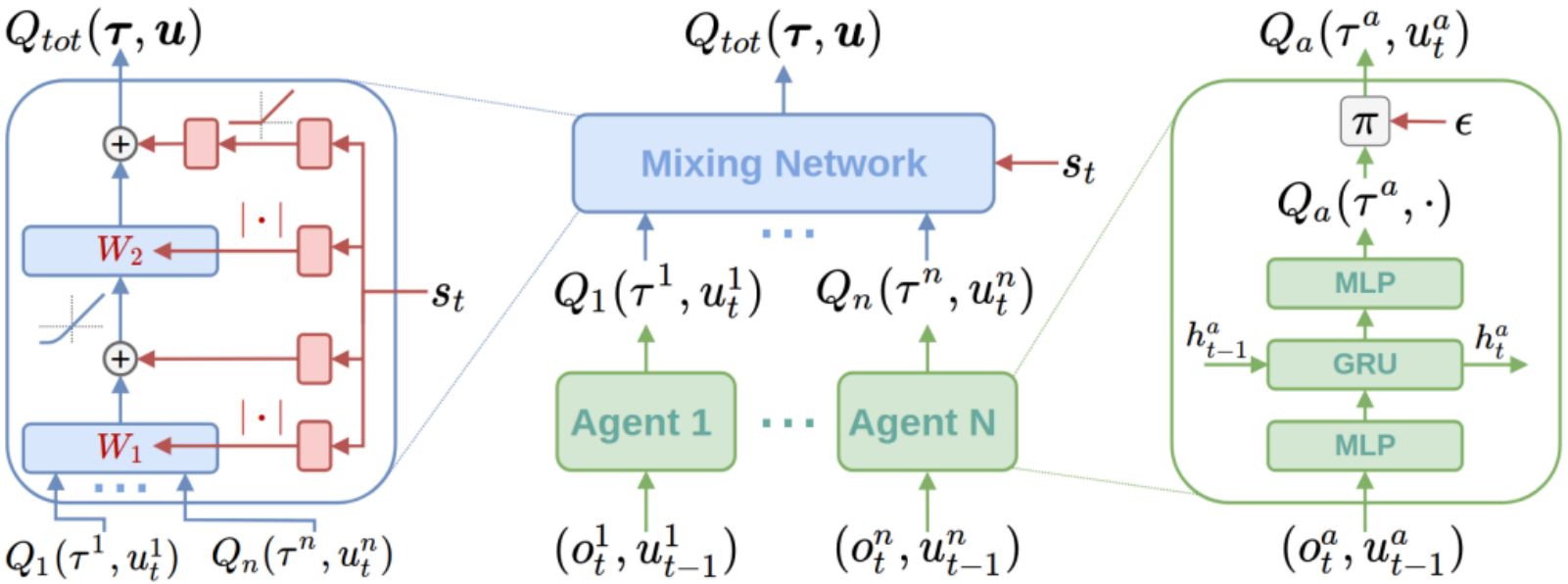
(a) 5 Marines map

(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

QPLEX: Duplex Dueling Multi-Agent Q-Learning (2021, ICLR)

- ❖ 기존 방법(VDN, QMIX)들은 표현력의 한계를 가지며, IGM 일관성을 유지하기 어려운 문제가 존재
- ❖ Credit assignmet를 적절히 분배하는 것이 어려움 (QMIX는 Q-value의 선형 결합으로 공동 행동 가치 함수를 구성했기 때문에, 특정 에이전트의 기여도를 세밀하게 조절하기 어려움)
- ❖ 위 한계점을 개선하기 위해 Advantage-based value factorization과 multi-head attention 사용한 QPLEX 제안

Published as a conference paper at ICLR 2021

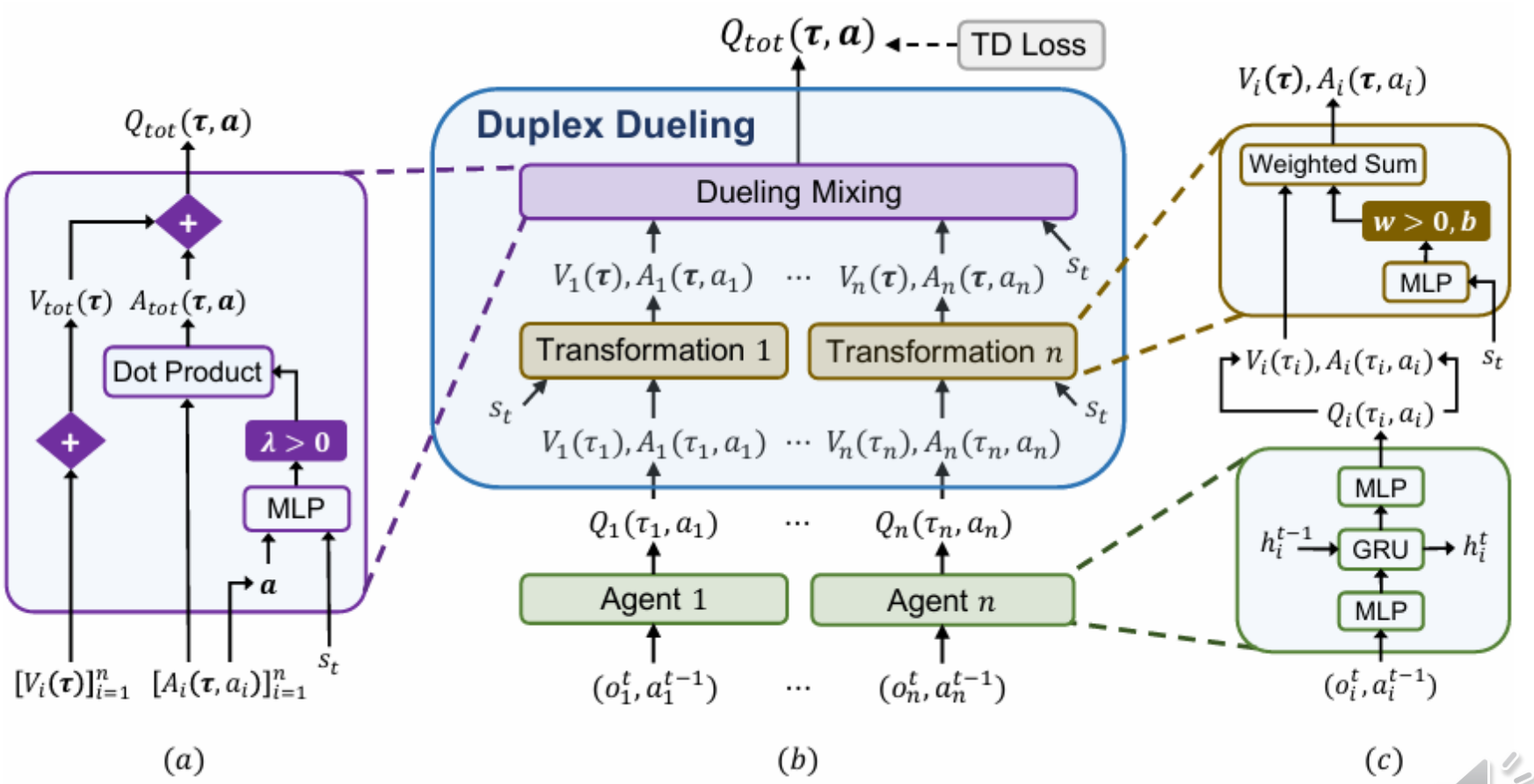
QPLEX: DUPLEX DUELING MULTI-AGENT Q-LEARNING

Jianhao Wang[*1], **Zhizhou Ren**[*1], **Terry Liu**[1], **Yang Yu**[2], **Chongjie Zhang**[1]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, China
[2]Polixir Technologies, China
{wjh19, rzz16, liudr18}@mails.tsinghua.edu.cn
yuy@nju.edu.cn
chongjie@tsinghua.edu.cn

Methods

QPLEX – 전체적인 구조

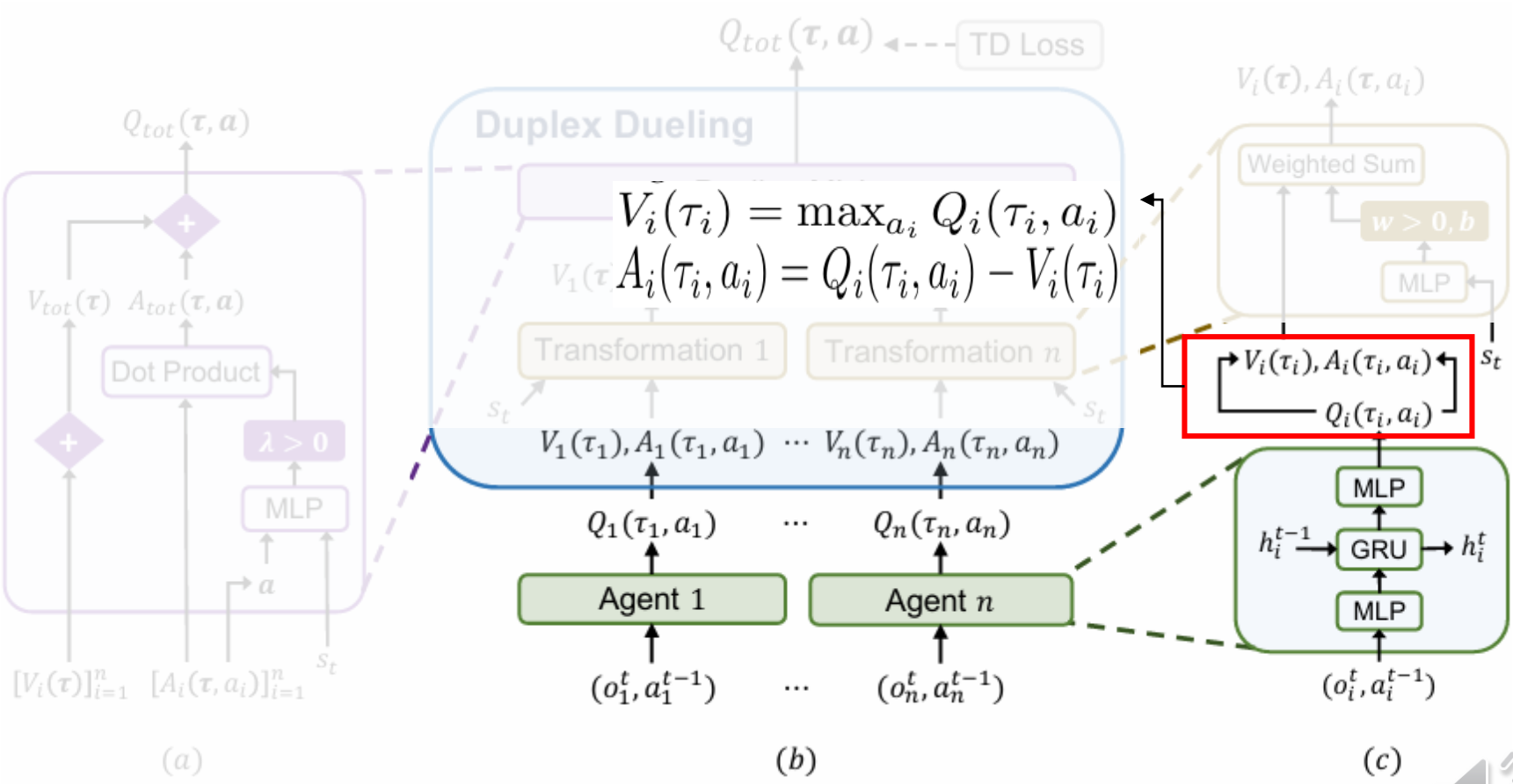
- ❖ Duplex: 공동 행동 가치 함수와 개별 가치 함수 모두 고려하는 이중 구조를 가지기에 붙여진 이름
- ❖ Dueling: advantage를 활용하여 행동 간 차이를 학습하는 방식이기에 붙여진 이름



Methods

QPLEX – 전체적인 구조

- ❖ 에이전트 네트워크는 QMIX에서 사용하고 있는 네트워크와 동일
- ❖ 에이전트 네트워크를 통해 나온 개별 가치 함수를 가치 함수와 이점 함수로 표현



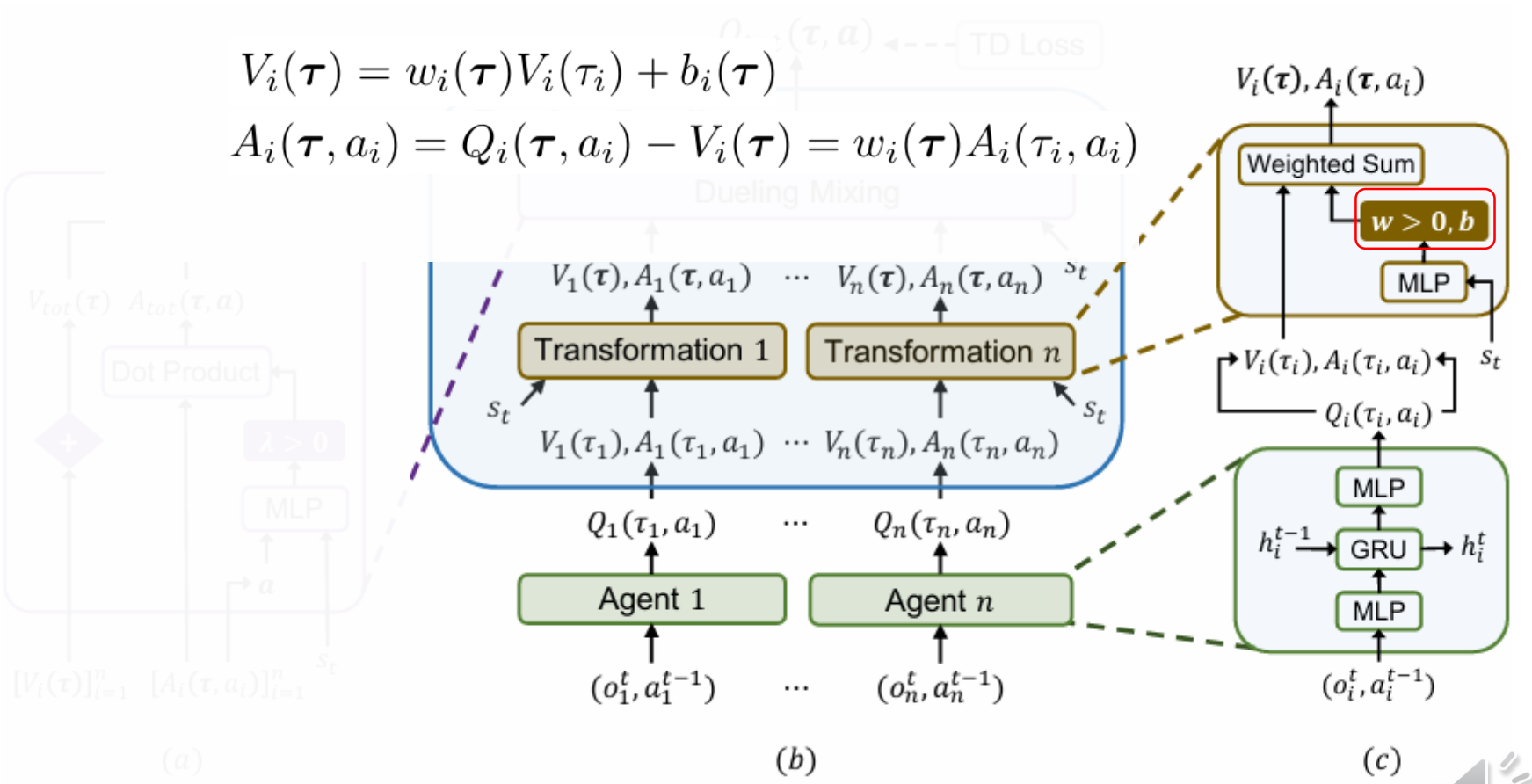
Methods

QPLEX – Transformation network module

- ❖ 현재 가치 함수와 이점 함수는 local-observation history에 기반하여 각 에이전트의 부분적인 정보 반영
- ❖ 전역 상태 정보를 추가하여 부분 관찰성 문제 해결 (w > 0 → 최적 행동 선택의 일관성 유지)

$$V_i(\boldsymbol{\tau}) = w_i(\boldsymbol{\tau})V_i(\tau_i) + b_i(\boldsymbol{\tau})$$

$$A_i(\boldsymbol{\tau}, a_i) = Q_i(\boldsymbol{\tau}, a_i) - V_i(\boldsymbol{\tau}) = w_i(\boldsymbol{\tau})A_i(\tau_i, a_i)$$



Methods

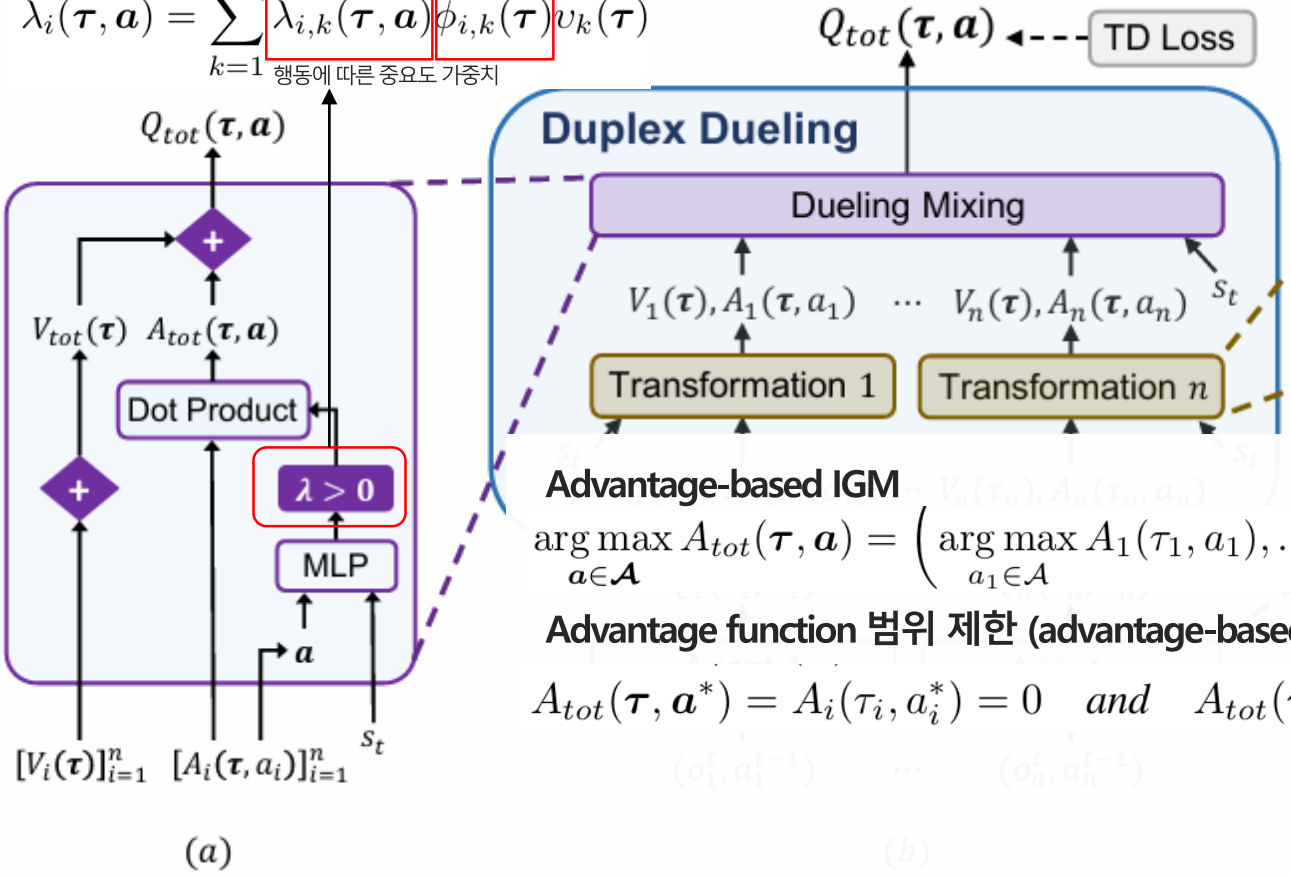
QPLEX – Dueling Mixing network module

- ❖ 이점 함수에만 제약을 걸어 IGM 원칙 유지 → advantage-based IGM
- ❖ 이점 함수에 제약을 가하면, 공동 행동 가치 함수에서 최적 행동이 항상 선택되도록 유도 가능

현재 환경에서 특정 에이전트가 얼마나 중요한지를 나타내는 스코어

$$\lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) = \sum_{k=1}^{K} \lambda_{i,k}(\boldsymbol{\tau}, \boldsymbol{a}) \phi_{i,k}(\boldsymbol{\tau}) \upsilon_k(\boldsymbol{\tau})$$

행동에 따른 중요도 가중치



Advantage-based IGM

$$\arg\max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = \left(\arg\max_{a_1 \in \mathcal{A}} A_1(\tau_1, a_1), \ldots, \arg\max_{a_n \in \mathcal{A}} A_n(\tau_n, a_n) \right)$$

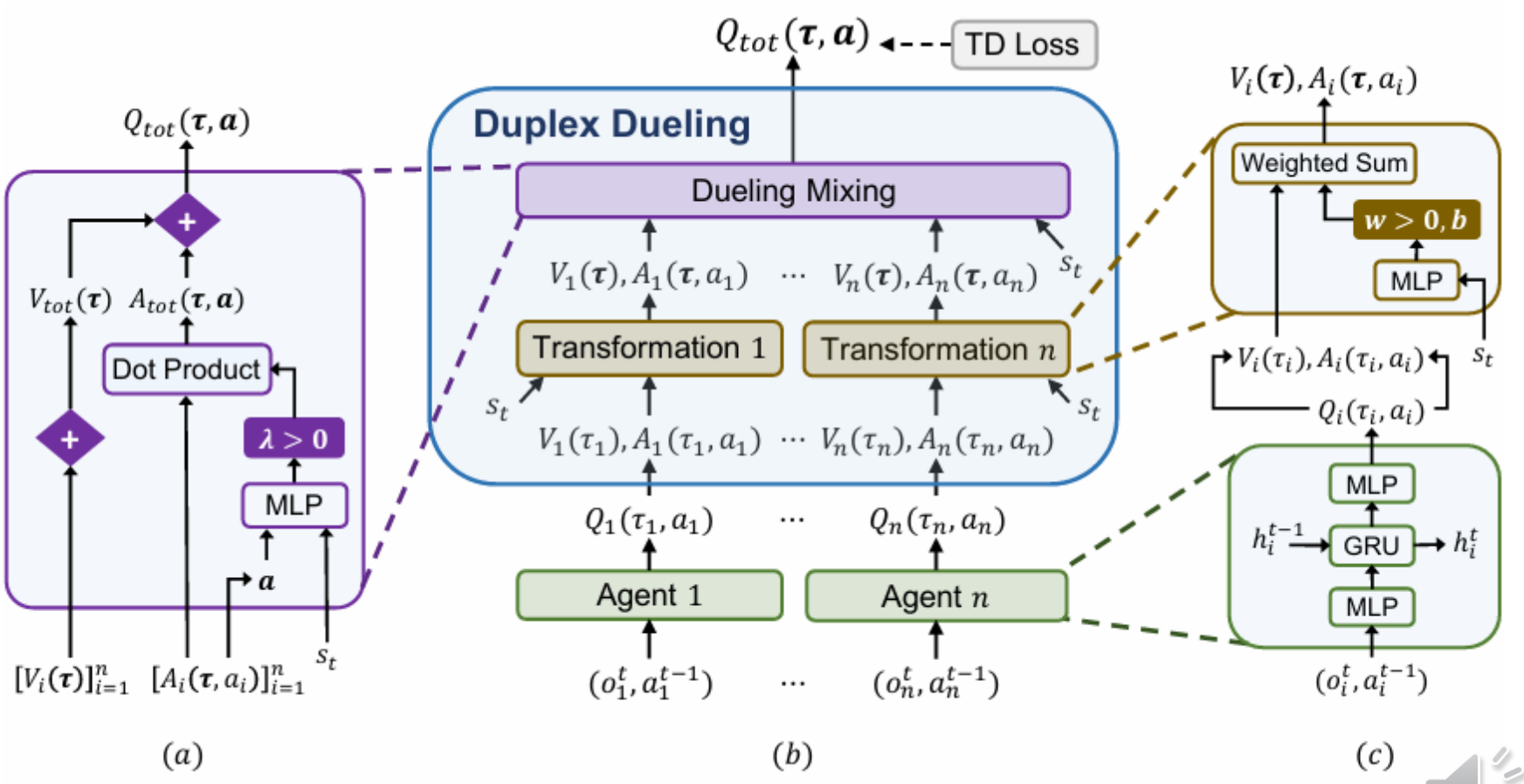
Advantage function 범위 제한 (advantage-based IGM을 보장)

$$A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}^*) = A_i(\tau_i, a_i^*) = 0 \quad and \quad A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) < 0, A_i(\tau_i, a_i) \leq 0$$

Methods

QPLEX – 최종 출력

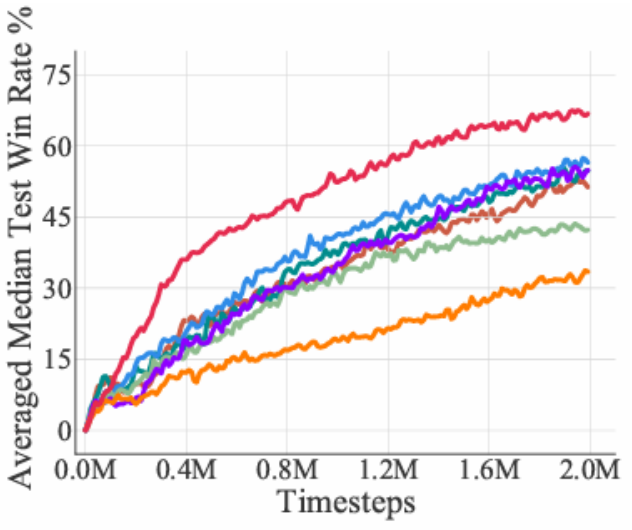
- ❖ 가치 함수와 이점 함수를 단순 합한 공동 가치 함수를 출력
- ❖ 공동 가치 함수를 사용한 TD Loss로 모델 업데이트



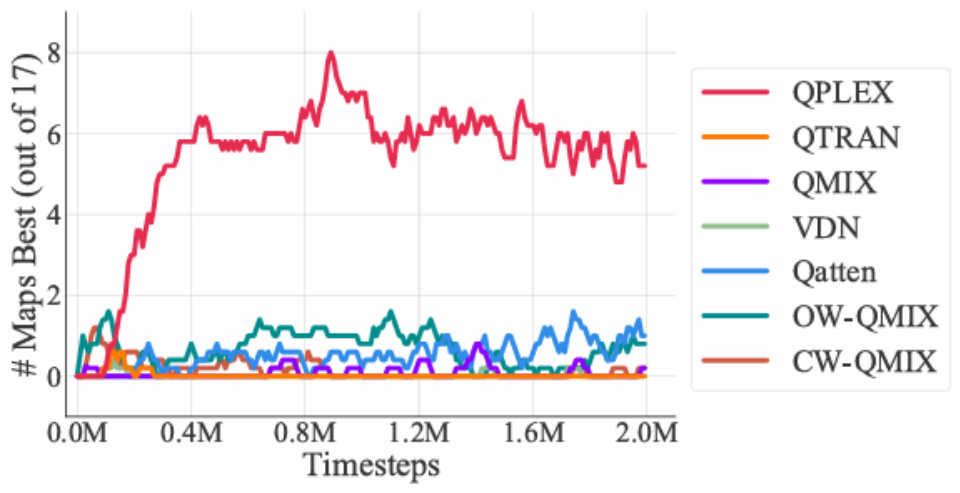
Methods

Experiment ①

- ❖ Figure (a): 17개의 SMAC 시나리오에 대한 평균 test win rate, QPLEX가 가장 우수함
- ❖ Figure (b): 17개의 시나리오 중에서 몇 개의 시나리오에서 최고의 성능을 보였는지 나타내는 그래프
  - ▪ 다른 알고리즘 대비, 시간이 지나면서 몇 개의 시나리오에서 최고 성능을 보이고 있는지를 보여줌



(a) Averaged test win rate

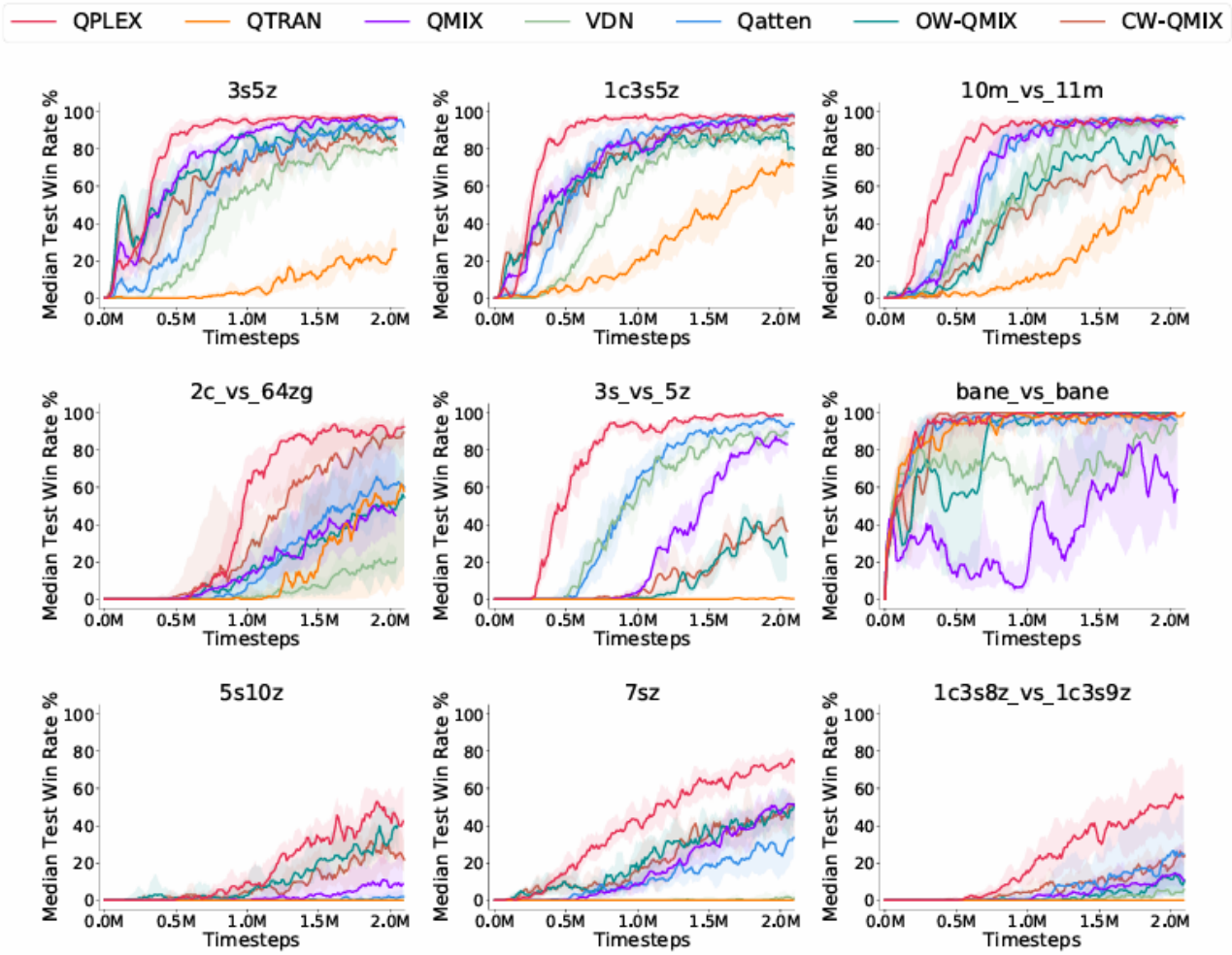


(b) # Maps best out of 17 scenarios

Methods

Experiment ②

❖ 모든 시나리오에서 QPLEX의 성능이 가장 우수함



Methods

Transformer-Based Value Function Decomposition for Cooperative Multi-Agent Reinforcement Learning in Starcraft (2022, AAAI)

- ❖ Credit assignment 문제와 공동 행동 가치 평가 문제를 해결해야함
- ❖ 기존 가치 함수 분해 방식(VDN, QMIX, QPLEX)의 한계 존재: 비선형적인 상호 작용 및 모든 형태의 복잡한 상호작용 학습하는데 어려움
- ❖ 기존 가치 함수 분해 방식의 한계를 개선하기 위해 트랜스포머 기반 Mixing network 제안 → TransMix
  - ① 비선형적인 가치 함수 분해가 가능 → TRANSFORMER의 어텐션 덕분에 비선형 조합 가능
  - ② 순열 불변성 확보
  - ③ 잡음이 있는 상태에서도 성능이 저하되지 않는 강인한 모델 제안

Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2022)

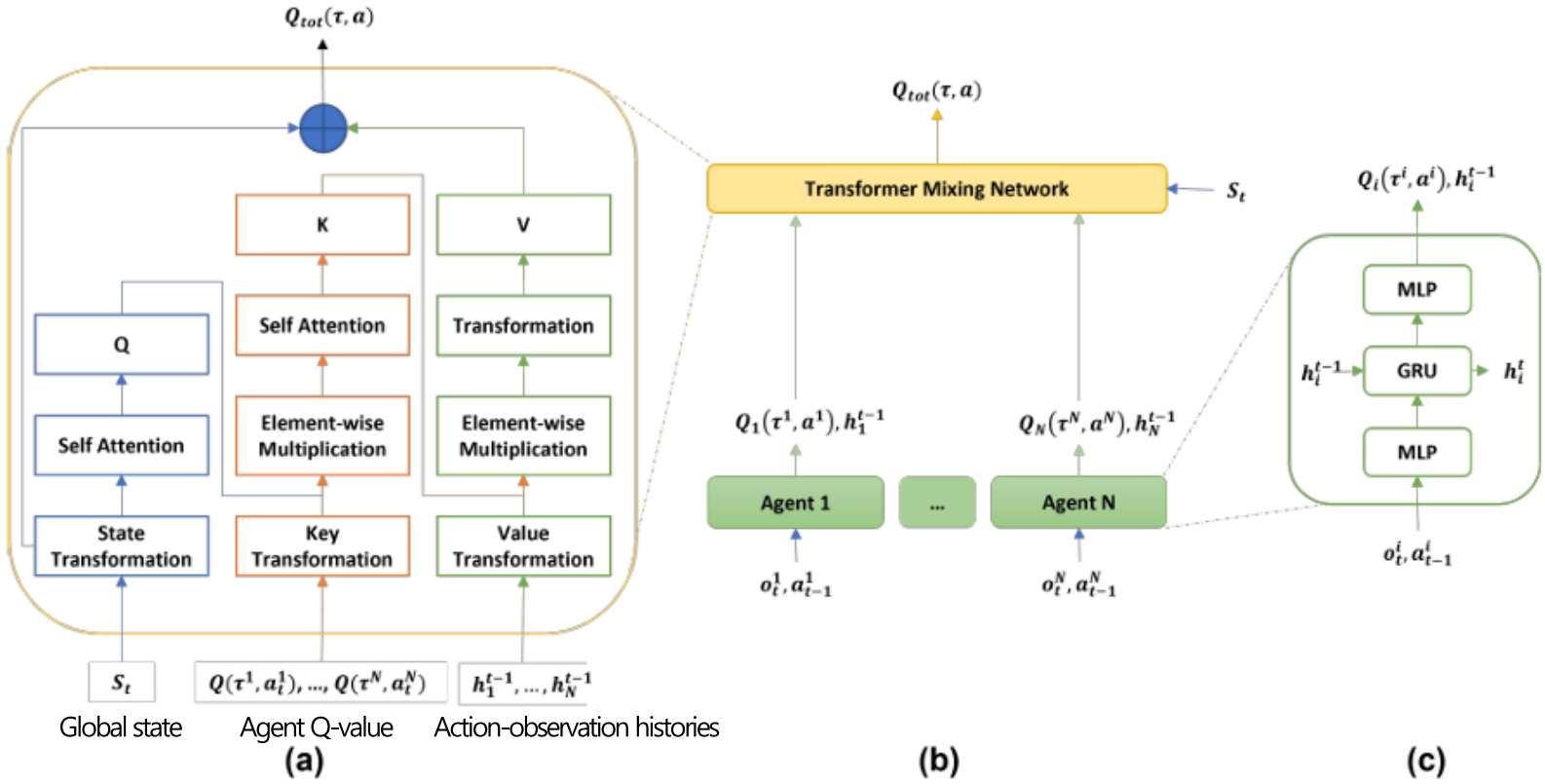
Transformer-Based Value Function Decomposition for Cooperative Multi-Agent Reinforcement Learning in StarCraft

Muhammad Junaid Khan, Syed Hammad Ahmed, Gita Sukthankar
Department of Computer Science
University of Central Florida
Orlando, FL US
{junaid_k, hammad.ahmed}@knights.ucf.edu, gitars@eecs.ucf.edu

Methods

TransMix – Overall Architecture

- ❖ QMIX의 전체적인 구조에서 Mixing network만을 Transformer 구조로 변환한 형태
- ❖ Agent network는 QMIX에서 사용한 GRU 기반 DRQN을 사용
- ❖ Mixing network는 Transformer 인코더를 사용



Methods

Experiment ①

- ❖ 각 시나리오는 2M 타입스텝 동안 학습하였으며, median win rate 결과
- ❖ 대부분의 시나리오에서 TransMix가 QMIX와 QPLEX보다 우수한 성능을 보임
- ❖ 5회 반복 실험 결과

Maps	Difficulty	TransMix	QMIX	QPLEX
2m_vs_1z	Easy	99. 7	95.3	**100**
3m	Easy	**100**	96.9	**100**
8m	Easy	**100**	97.7	**100**
2s3z	Easy	**100**	88.25	**100**
1c3s5z	Easy	**97.62**	93.37	97.2
MMM	Easy	**100**	95.3	96.9
bane_vs_bane	Easy	**100**	89.21	98.47
3s_vs_5z	Hard	95.91	88.7	**99.1**
3s5z	Hard	**96.68**	88.3	95.9
5m_vs_6m	Hard	**77.41**	69.2	73.4
8m_vs_9m	Hard	**96.88**	92.2	87.5
2c_vs_64zg	Hard	**92.62**	84.38	91.2
10m_vs_11m	Hard	**91.77**	89.2	90.12

Methods

Experiment ②

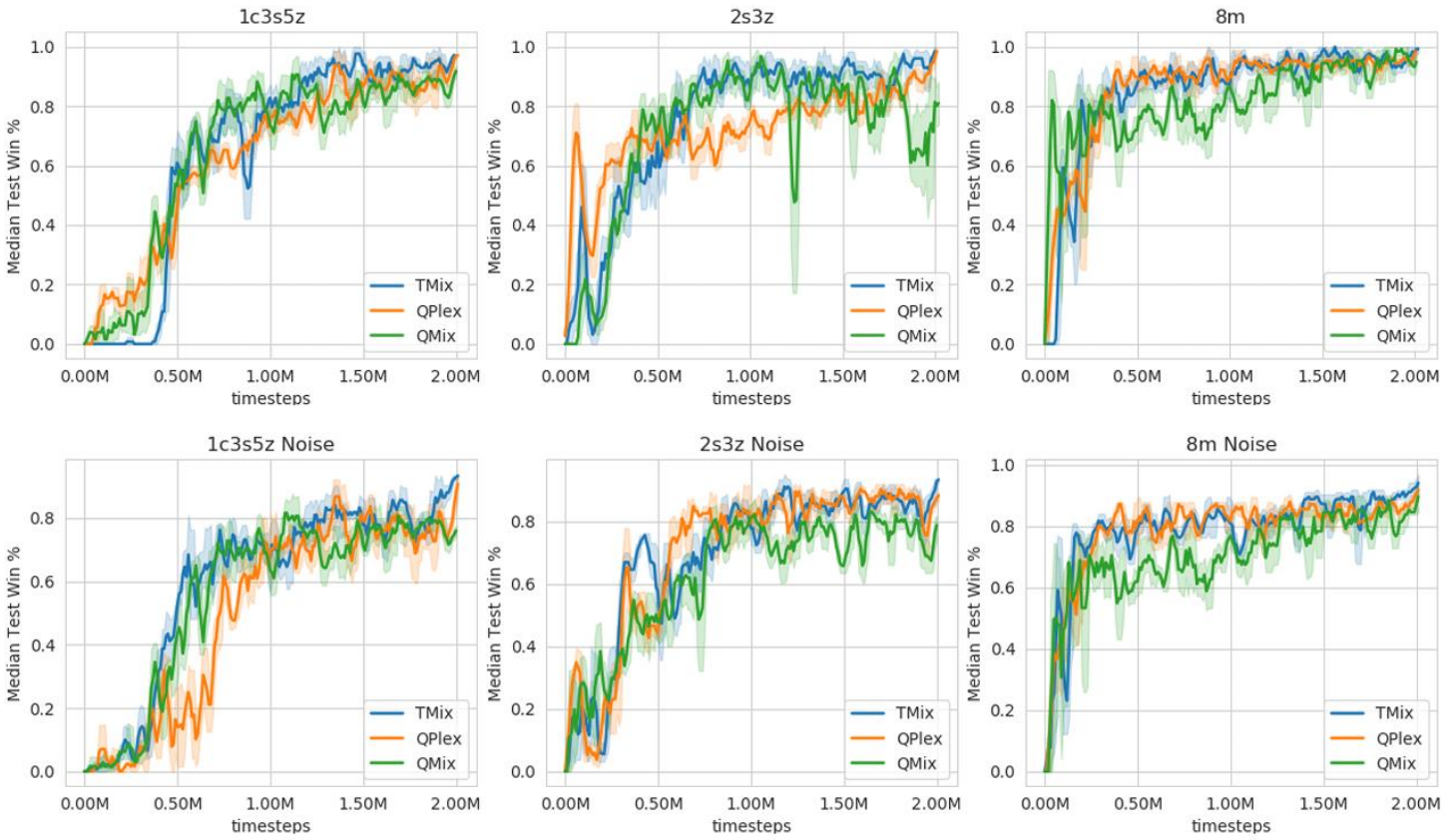
- ❖ Global state에 가우시안 노이즈를 추가하여 마치 전쟁에 안개가 뒤덮인 듯한 상태를 표현함 ($N(0, 0.05)$)
- ❖ Median win rate 결과이며, TransMix가 QMIX와 QPLEX보다 노이즈에 더 강건한 것을 확인함
- ❖ 맵의 복잡도가 높아질수록, 성능이 저하되는 것을 확인함

Maps	TransMix	QMIX	QPlex
3m	**99.82**	95.31	97.41
8m	**96.87**	93.75	93.75
2s3z	**93.75**	81.67	89.15
1c3s5z	**93.75**	78	83.88
3s5z	**84.18**	65.76	72.62

Methods

Experiment ③

- ❖ 첫 번째 행은 일반적인 상태에서 각 방법론의 성능 비교 결과
- ❖ 두 번째 행은 노이즈가 낀 상태에서 각 방법론의 성능 비교 결과
- ❖ TransMix가 QMIX와 QPLEX보다 노이즈에 더 강건하고 우수한 성능을 보인다고 함



Methods

Attention-Guided Contrastive Role Representations For Multi-Agent Reinforcement Learning (2024, ICLR)

- ❖ 유사한 행동을 학습하는 문제 → 기존 방법론은 정책을 공유하는 문제로 인해 에이전트가 유사한 행동을 학습하는 경향을 보임
- ❖ 효과적인 팀 구성의 필요성 → 역할 기반 학습 → 기존 방법은 역할의 동적 변화를 반영하기 어려움 (정적인 역할) → 환경의 변화에 따라 에이전트의 기여도를 올바르게 평가하기 어려움 (credit assignment problem)
- ❖ 대조 학습과 어텐션을 사용하여 기존 연구의 한계점을 개선한 ACORM 제안

Published as a conference paper at ICLR 2024

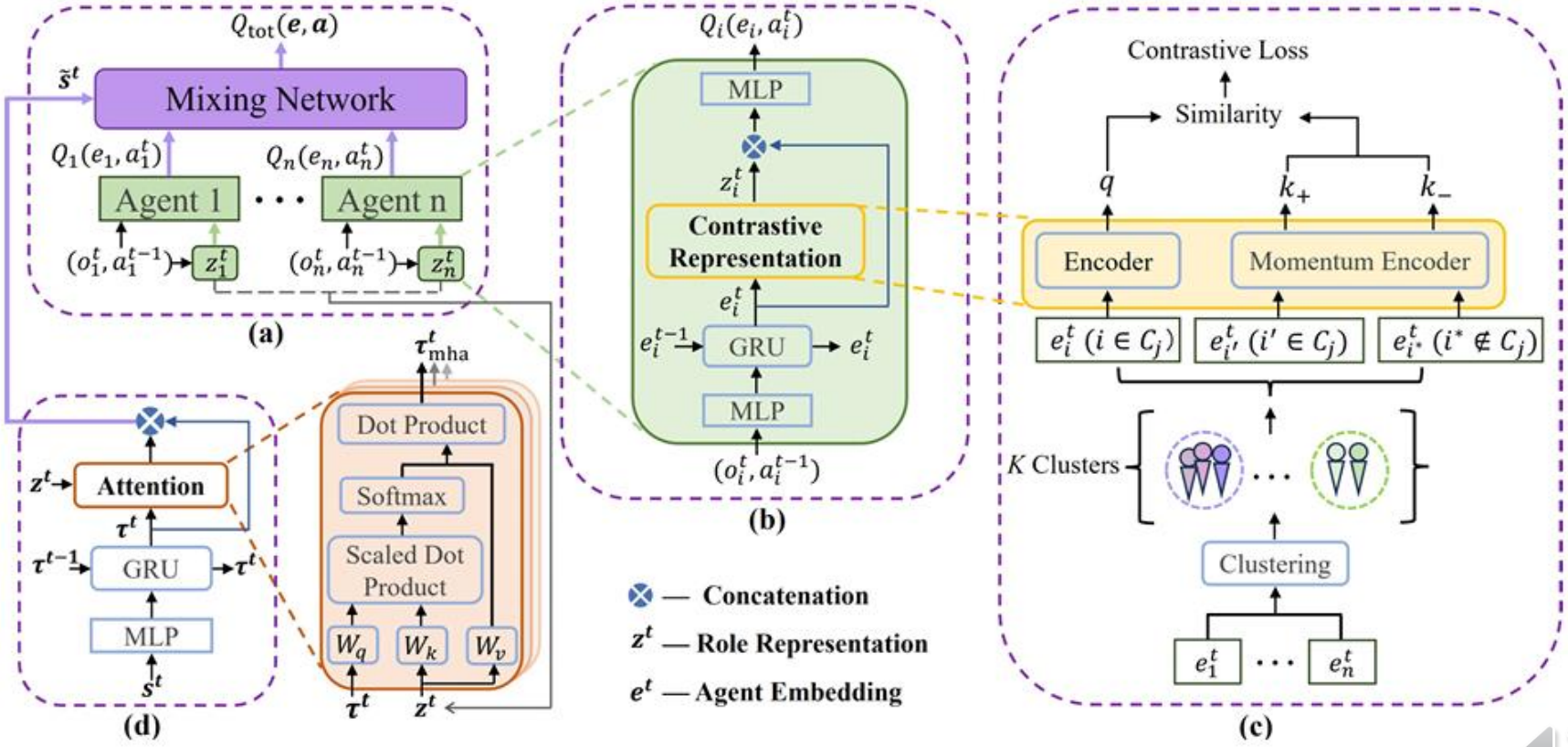
ATTENTION-GUIDED CONTRASTIVE ROLE REPRESENTATIONS FOR MULTI-AGENT REINFORCEMENT LEARNING

Zican Hu[1], **Zongzhang Zhang**[2], **Huaxiong Li**[1], **Chunlin Chen**[1], **Hongyu Ding**[1], **Zhi Wang**[1*]
[1] Department of Control Science and Intelligent Engineering, Nanjing University
[2] School of Artificial Intelligence, Nanjing University
{zicanhu,hongyuding}@smail.nju.edu.cn
{zzzhang,huaxiongli,clchen,zhiwang}@nju.edu.cn

Methods

ACORM – 전체적인 구조

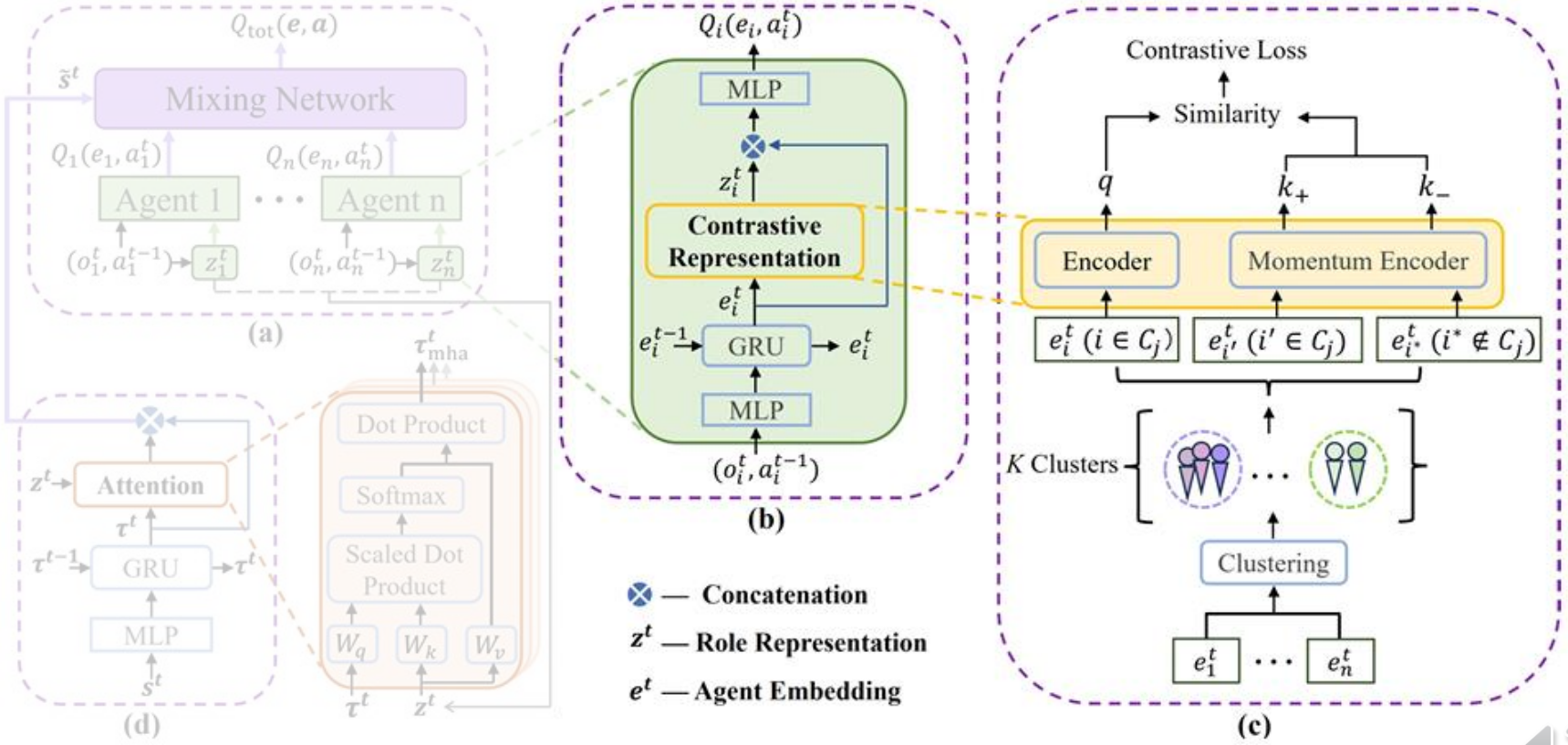
- ❖ 대조 학습 기반 역할 표현 학습 프레임워크 제안 → 유사한 행동을 하는 문제 해결
- ❖ 멀티 헤드 어텐션 사용 → 정적인 에이전트의 역할 문제를 해결하여, 역할 변화에 따른 신용 할당을 동적으로 최적화



Methods

ACORM – 에이전트 네트워크

- ❖ QMIX의 에이전트 네트워크에서 대조 학습을 통해 역할 표현을 학습하고 활용함
- ❖ 대조 학습을 통해, 유사한 행동 패턴을 가진 에이전트들이 보다 가까운 역할 표현을 가지도록, 반대로 다른 전략을 가진 에이전트들은 멀어지도록 학습



Methods

ACORM – 에이전트 네트워크 (대조 학습을 통한 역할 표현 학습)

- ❖ Mutual Information 최대화: 역할 표현(z)가 역할(M)을 최대한 반영하도록 학습하기 위함
- ❖ MI를 실제로 계산하기 어렵기 때문에 InfoNCE 손실 함수를 사용함

Mutual information 최대화

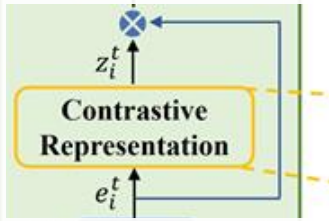
$$\max\ I(z;M) = \mathbb{E}_{z,M}\left[\log \frac{p(M|z)}{p(M)}\right]$$

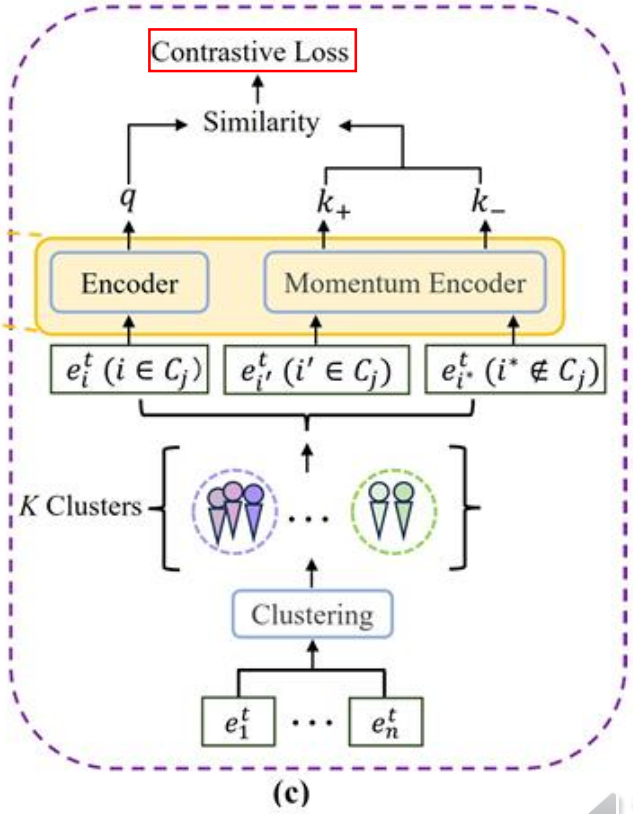
Contrastive Loss(→ InfoNCE)

$$\min_{\theta} \mathcal{L}_K = -\log \frac{\exp(q^\top W k_+)}{\exp(q^\top W k_+) + \exp(q^\top W k_-)} = -\log \frac{\sum_{i' \in C_j} \exp(z_i^\top W z_{i'})}{\sum_{i' \in C_j} \exp(z_i^\top W z_{i'}) + \sum_{i^* \notin C_j} \exp(z_i^\top W z_{i^*})}.$$

Momentum update

$$\theta_k \leftarrow \beta\theta_k + (1-\beta)\theta_q$$





(c)

— Concatenation

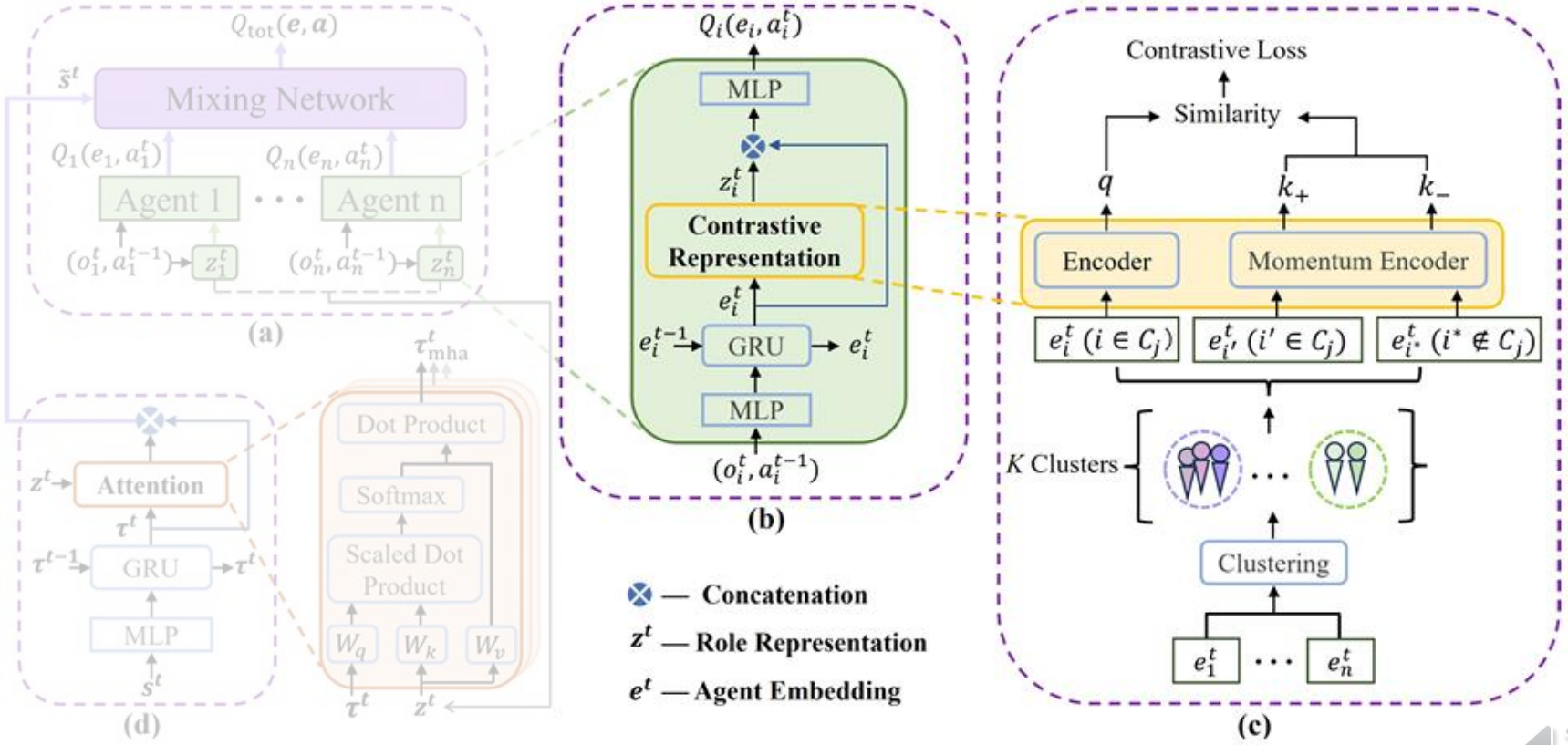
— Role Representation

— Agent Embedding

Methods

ACORM – 에이전트 네트워크

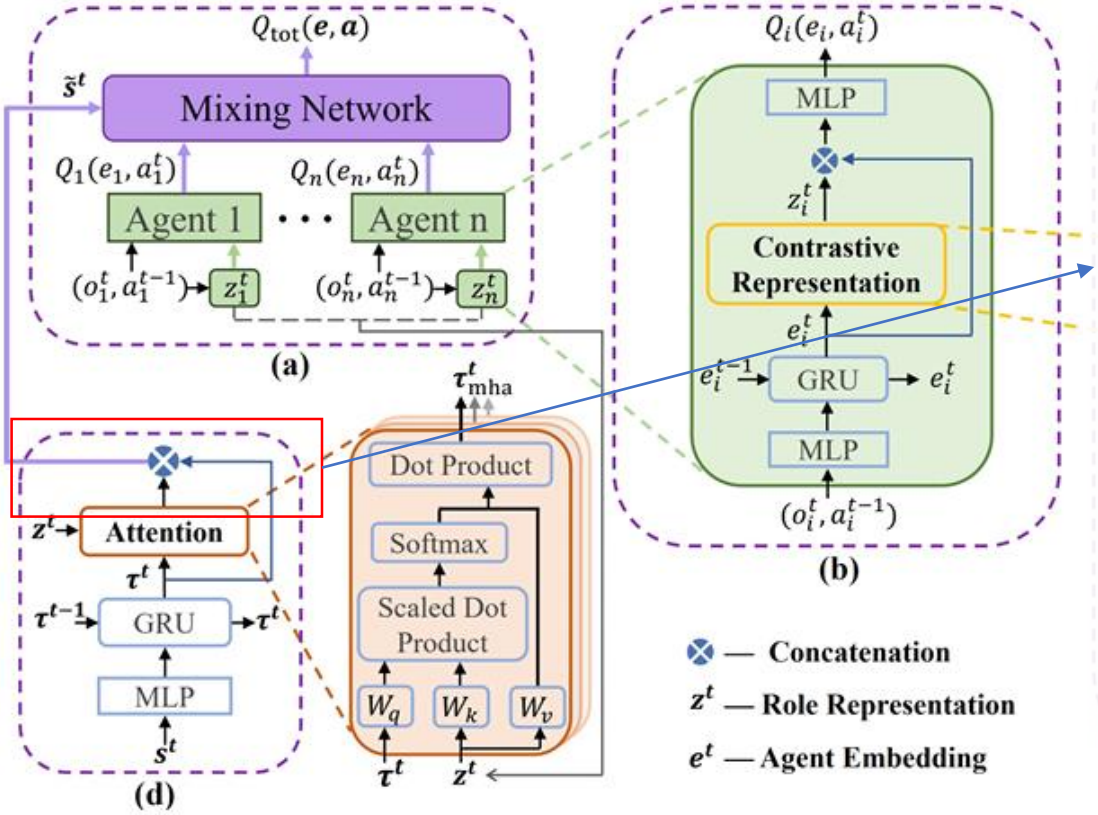
- 최종적으로, e_i^t를 Encoder에 입력해서 각 에이전트에 대한 역할 정보를 담은 임베딩 벡터(z_i^t) 산출
- 에이전트 행동 패턴과 관련된 정보를 담고 있는 e_i^t와 역할 정보를 담고 있는 z_i^t를 합친 정보를 활용하여 개별 가치 함수$Q_i(e_i, a_i^t)$ 산출



Methods

ACORM – 혼합 네트워크(멀티 헤드 어텐션 사용)

- ❖ 혼합 네트워크의 가중치는 전역 상태 정보에 따라 결정됨
- ❖ 멀티 헤드 어텐션을 사용하여 역할 정보가 반영된 새로운 전역 상태 정보 생성 → 역할 간 협력을 보다 효과적으로 수행하도록 유도



Attention을 통해 나온 값을 그대로 사용하지 않는 이유?

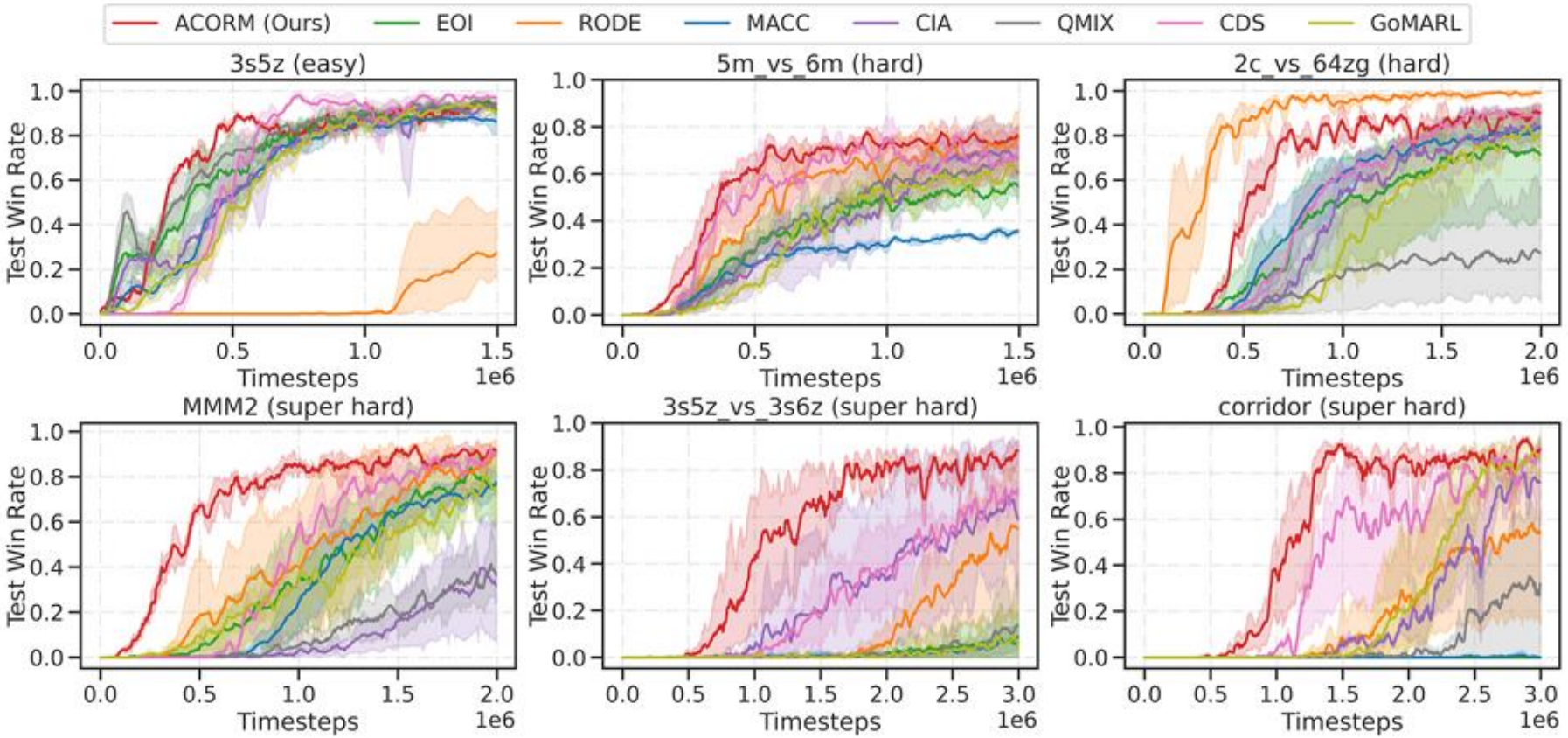
→ MHA 출력만 사용하게 되면, 원래 전역 상태에 포함된 정보를 잃을 수 있음

→ 따라서, 전역 상태의 원본 정보와 역할이 반영된 정보를 함께 제공하기 위해 concat을 한 새로운 전역 상태 생성

Methods

Experiment ①

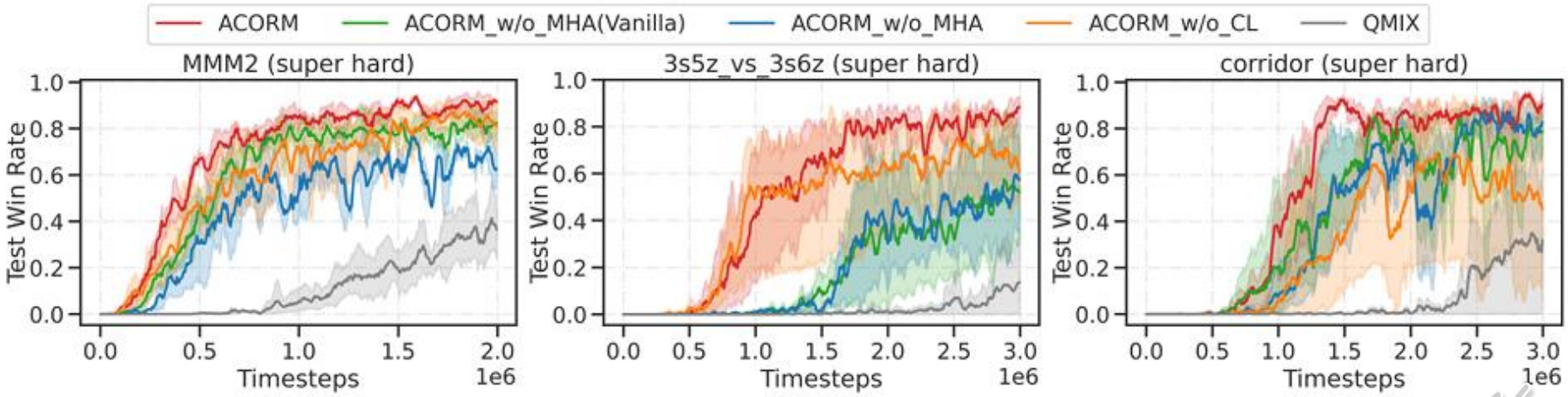
- ❖ 대부분의 시나리오에서 ACORM 성능이 우수함
- ❖ 특히, 매우 어려운 시나리오에서 ACORM 성능이 다른 방법론을 압도함



Methods

Experiment ② - Ablation study

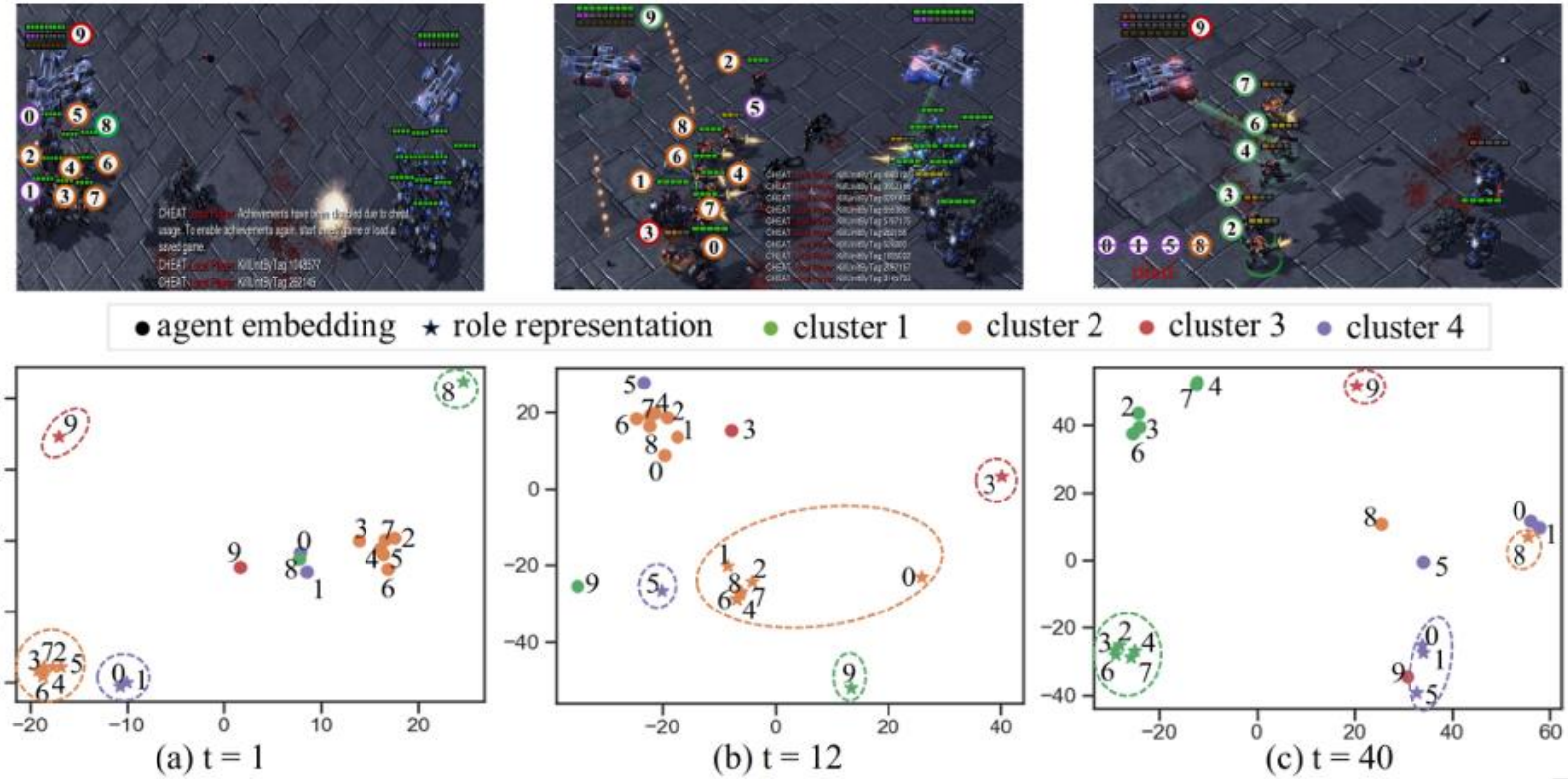
- ACORM_w/o_MHA(Vanilla): 새로운 상태 인코딩, 멀티 헤드 어텐션 제거한 ACORM
- ACORM_w/o_MHA: 멀티 헤드 어텐션을 제거한 ACORM
- ACORM_w/o_CL: 대조 학습을 사용하지 않는 ACORM
- ACORM은 모든 요소를 사용했을 때, 성능이 가장 우수함
- ACORM_w/o_CL과 ACORM_w/o_MHA의 성능 저하가 확인되었으며, 이는 두 요소가 ACORM의 성능에 중요한 역할을 한다는 것을 의미
- ACORM_w/o_MHA(Vannilla)의 성능이 ACORM_w/o_MHA와 유사하며, MHA의 핵심 기여는 상태 인코딩이 아니라 주어진 어텐션 모듈 자체에서 비롯됨



Methods

Experiment ③

- ❖ Agent embedding 관점: 현재 행동 패턴이 유사한 에이전트들은 가까운 위치에 존재하고, 시간이 지나면서 역할에 따라 구분됨
- ❖ Role representation 관점: 각 역할에 대한 대표적인 특징 벡터를 학습하여, 에이전트들이 자연스럽게 역할별 그룹으로 형성됨



Methods

Experiment ④

- 같은 군집안에 있는 역할은 유사한 어텐션 값을 보이며, 다른 군집은 상당히 다른 값을 가짐
- Head 2는 부상-구조 패턴을 보이고 있음을 보여줌
  - 가장 큰 가중치는 Medivac(9)와 체력이 낮은 유닛들에게 부여
- 대부분의 헤드에서, Marauder (0, 1)의 어텐션 가중치는 초반에 높으며, 시간이 지나면서 크게 감소
  - 이는 초반 공격에서 Marauder가 중요한 역할을 하다가 점차 marine에게 공격 임무가 넘겨지는 패턴을 보임
- 승리가 확실시 되는 시점인 t=36에서, 주요 전략이 마린 (2, 3, 5)을 사용한 최종 공격 수행이며, 체력이 낮은 마린 (4, 6, 7)은 보조 지원 역할 진행

(a) t = 4

(b) t = 10

(c) t = 36

Conclusion

- QMIX의 구조를 변형하여 가치 분해 기법의 한계를 극복하고, 표현력과 학습 안정을 개선하는 방법론에 초점을 맞춤
- QPLEX (2021, ICLR): advantage based IGM을 활용하여 IGM 원칙을 유지하면서도 표현력을 높이고, 학습 속도와 안정성을 개선한 가치 분해 기반 알고리즘
- TransMix (2022, AAAI): Transformer 기반의 가치 분해 기법을 적용하여, 시퀀스 의존성이 강한 환경에서 효과적인 성능을 입증
- ACORM (2024, ICLR): 다중 헤드 어텐션과 대조 학습을 활용하여 에이전트별 역할 표현을 학습하고 협력 전략을 최적화하는 알고리즘

Reference

1. Wang, J., Ren, Z., Liu, T., Yu, Y., & Zhang, C. (2020). Qplex: Duplex dueling multi-agent q-learning. arXiv preprint arXiv:2008.01062.
2. Khan, M. J., Ahmed, S. H., & Sukthankar, G. (2022, October). Transformer-based value function decomposition for cooperative multi-agent reinforcement learning in starcraft. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 18, No. 1, pp. 113-119).
3. Hu, Z., Zhang, Z., Li, H., Chen, C., Ding, H., & Wang, Z. (2023). Attention-guided contrastive role representations for multi-agent reinforcement learning. arXiv preprint arXiv:2312.04819.

Thank you