Addressing the Sight Range Dilemma in Multi-Agent Reinforcement Learning

Data Mining & Quality Analytics Open Seminar

2025.10.10

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
- Anomaly Detection

❖ Contact

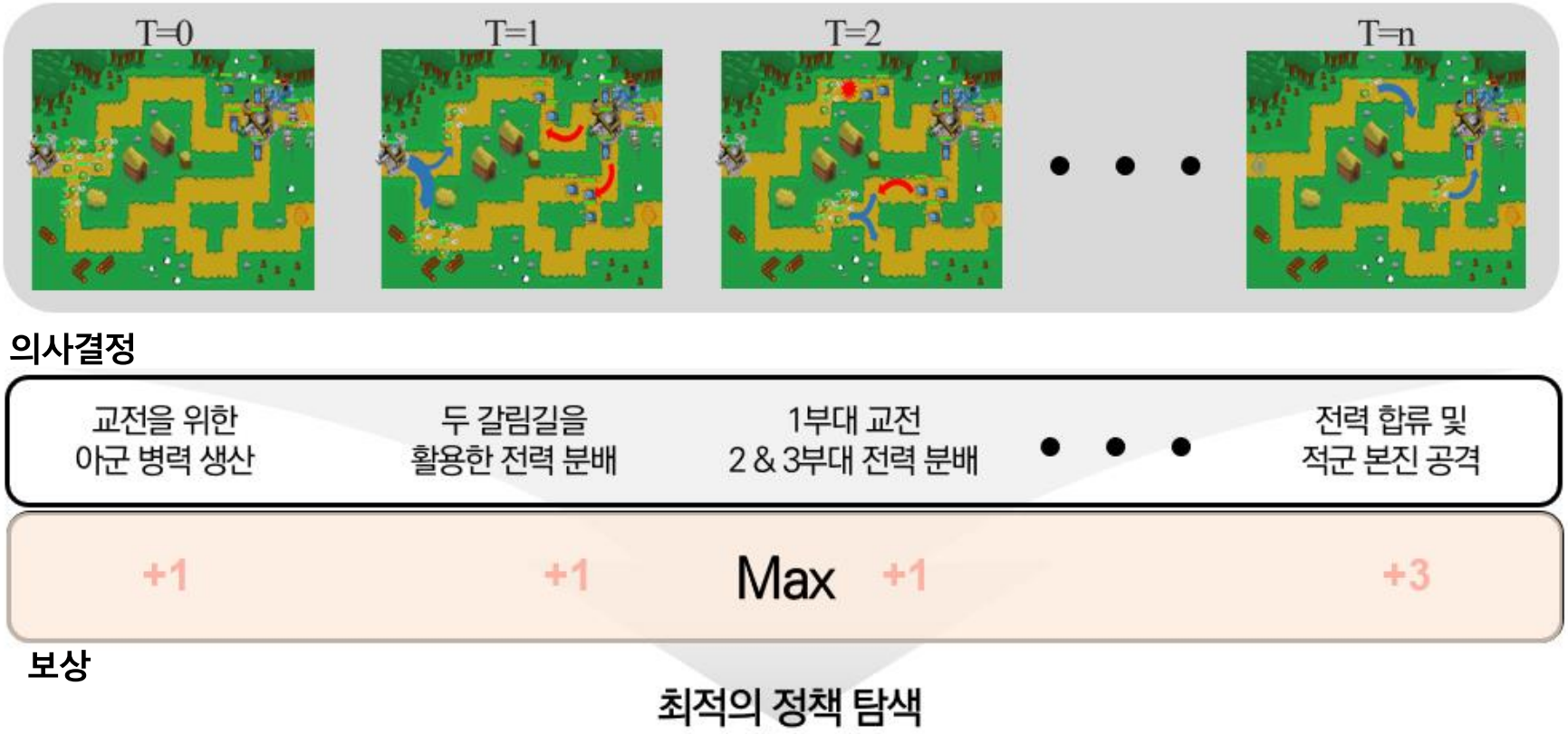
- Jungin_kim23@korea.ac.kr

목차

Introduction

강화학습 정의

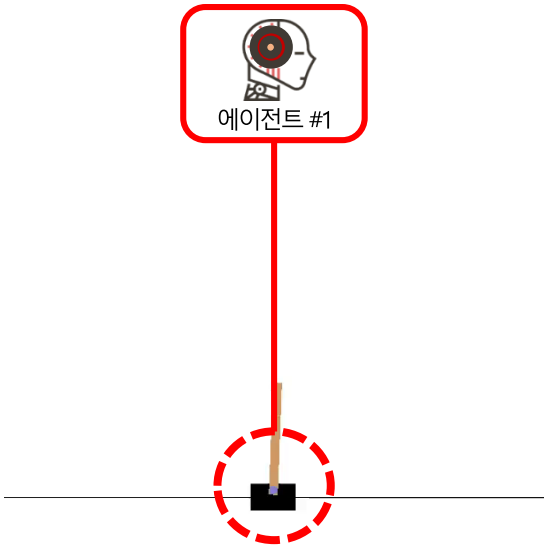
- ❖ 순차적인 의사결정 문제에서 에이전트가 환경으로부터 받는 누적 보상 값을 최대화하는 행동 정책을 학습하는 방법
- ❖ 정책: 에이전트가 특정 상태에서 어떤 행동을 선택할지 정해주는 함수



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)

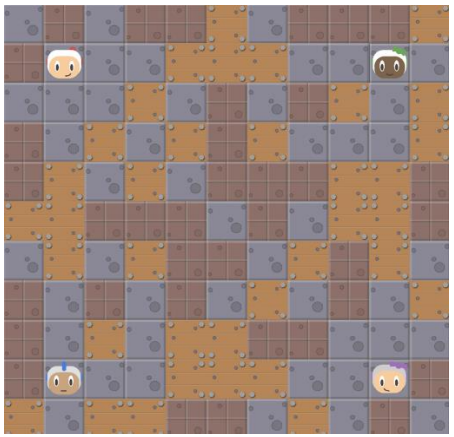
다중 에이전트 강화학습 (MARL)

Introduction

Cooperative vs Competitive setting

- ❖ 보상 구조에 따라서, 대표적으로 2가지 설정으로 분류 가능
- ❖ Cooperative setting: 팀 보상 공유 → 협력 필수
- ❖ Competitive setting: 보상 제로섬 → 경쟁 필수

Cooperative game benchmark (SMAC)



Cooperative game benchmark (Pommerman)

Cooperative vs Competitive (넓은 의미)

- ❖ Cooperative setting: 에이전트들이 협력하도록 유도되는 상황을 "cooperative" setting이라고 정의
- ❖ Competitive setting: 에이전트들이 상대보다 더 잘하도록 장려하는 경우를 "competitive" setting이라고 정의

Introduction

Cooperative setting = Decentralized Partially Observable Markov Decision Process (Dec-POMDP)

- ❖ Cooperative setting은 여러 에이전트가 하나의 팀으로 묶여 있고, 팀 성과가 보상으로 주어짐
- ❖ Dec-POMDP: 여러 에이전트가 partial observation만 가지고 협력해야 하는 문제 구조
- ❖ Cooperative setting이 Dec-POMDP가 정의하는 구조와 일치

Dec-POMDP 구성요소

$$M = \langle N, S, A, P, O, Z, R, \gamma \rangle$$

N: 에이전트 집합
S: 전역 상태 (global state, 직접 관측 불가)
A: joint action (모든 에이전트의 행동 조합)
P: 상태 전이 확률 함수
O: 관측 집합(각 에이전트는 부분 관측만 가능)
Z: 관측 함수 (상태 → 에이전트별 observation 분배)
R: 공유 보상 함수
γ: discount factor

Introduction

Dec-POMDP → Partial Observability

- ❖ Dec-POMDP 구조에서 각 에이전트는 전역 상태의 일부분만 관측 가능함 → 자신의 주변 정보
- ❖ 부분 관측만 가능하기 때문에, 발생하는 문제가 여러 존재 (ex: coordination problem, credit assignment problem)

Coordination Problem

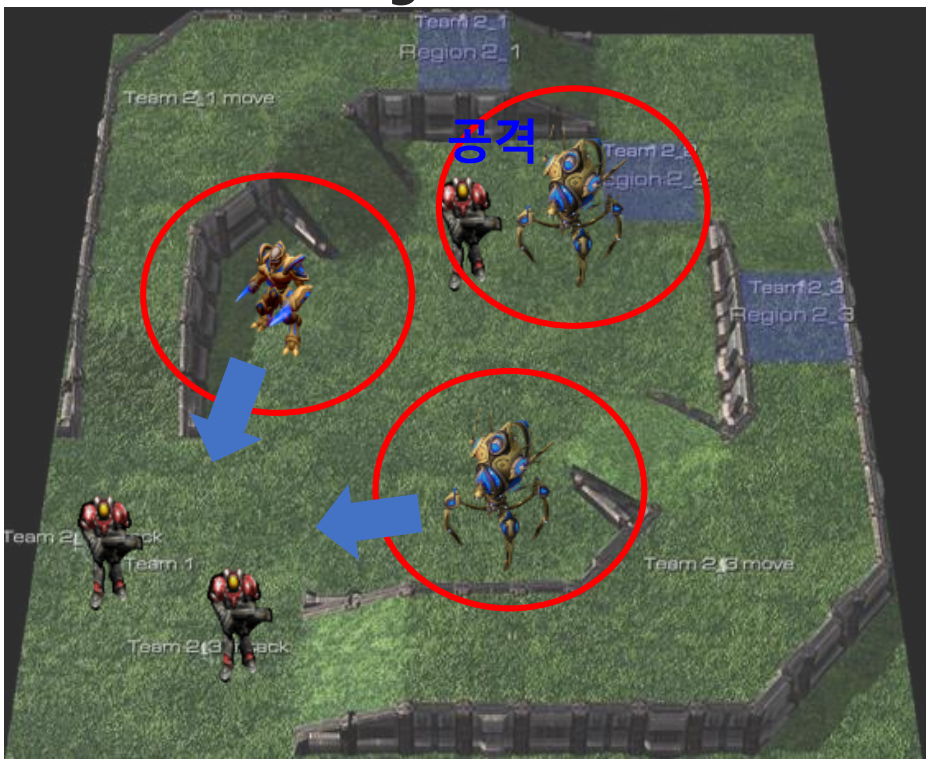
내 시야에 안 보이는데 어떻게 협력하지?

Introduction

Dec-POMDP → Partial Observability

- ❖ Dec-POMDP 구조에서 각 에이전트는 전역 상태의 일부분만 관측 가능함 → 자신의 주변 정보
- ❖ 부분 관측만 가능하기 때문에, 발생하는 문제가 여러 존재 (ex: coordination problem, credit assignment problem)

Credit Assignment Problem



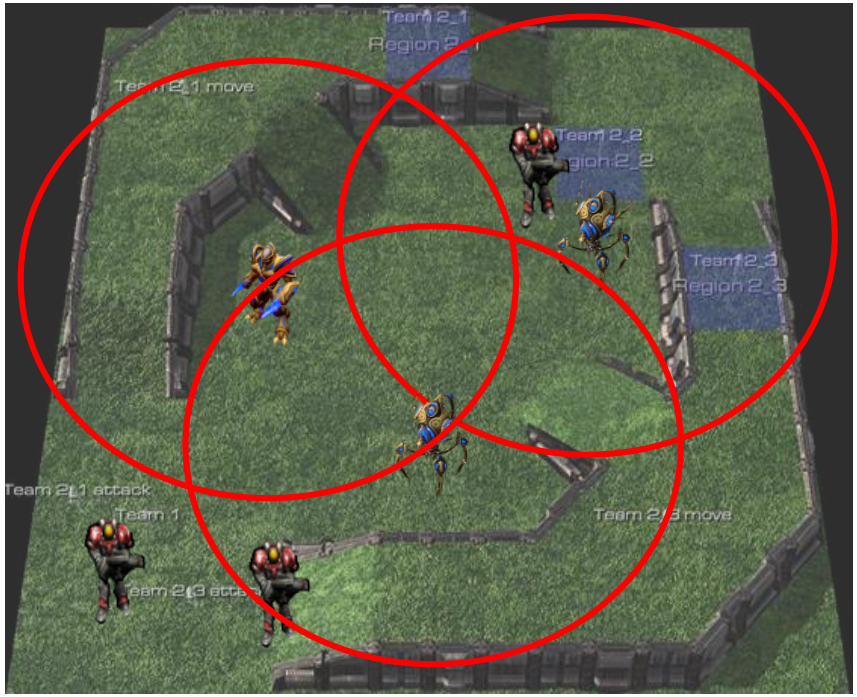
내 행동이 얼마나 기여했지??

Introduction

Dec-POMDP → Partial Observability → Sight Range Dilemma

- ❖ Sight Range Dilemma: 시야 범위 설정에 따라, 협력 학습이 오히려 방해 받을 수 있는 현상
  - 시야가 너무 좁을 때, 필요한 정보를 보지 못함 → 협력에 필요한 단서 부족
  - 시야가 너무 넓을 때, 너무 많은 불필요한 정보 포함 → 학습이 비효율적이고 정책이 혼란스러워짐
- ❖ Trade-off: 좁으면 정보 부족, 넓으면 정보 과잉 → 둘 다 협력 성능을 저하시킴

시야가 너무 좁을 때



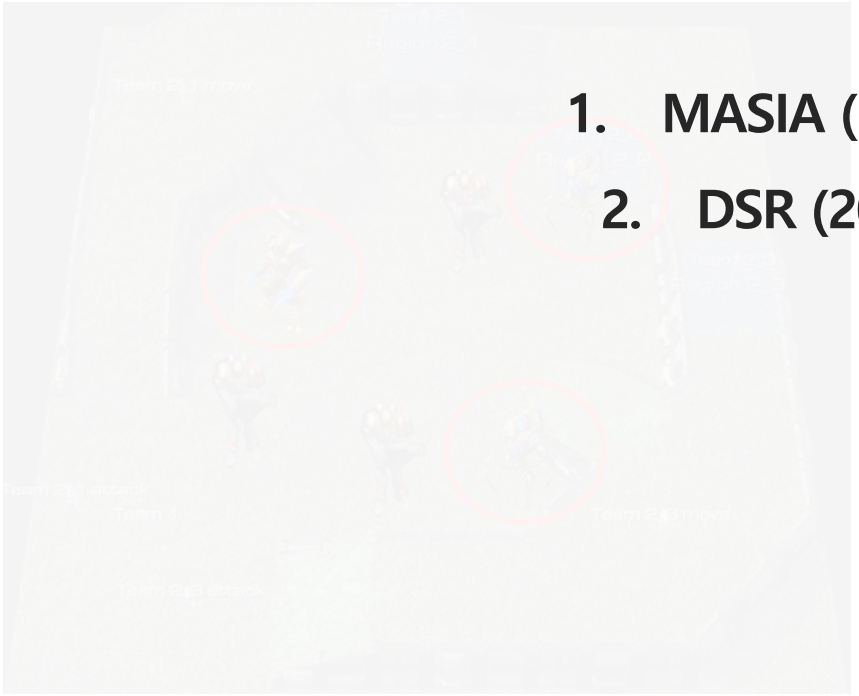
시야가 너무 넓을 때

Introduction

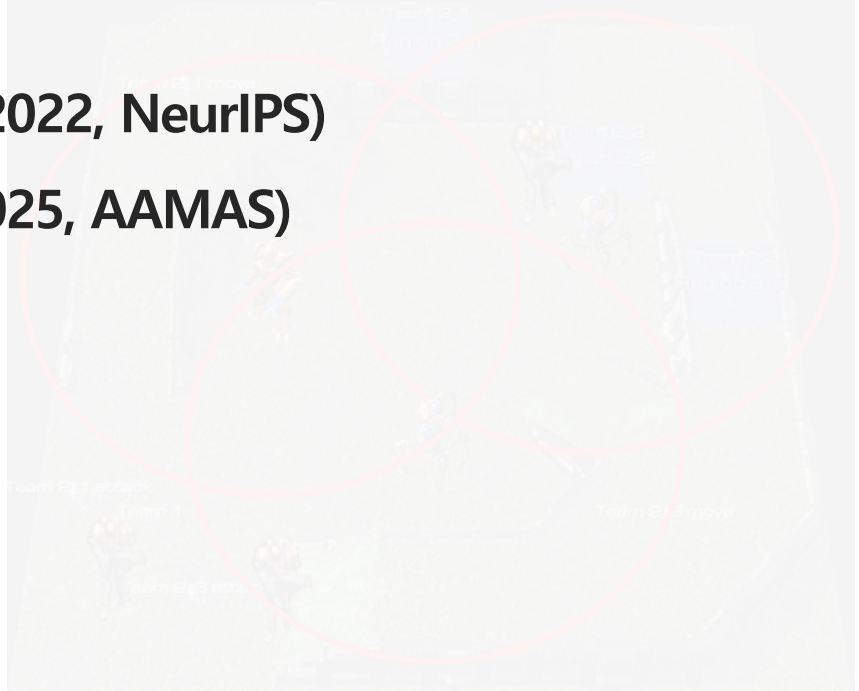
Dec-POMDP → Partial Observability → Sight Range Dilemma

Sight Range Dilemma를 다룬 2가지 연구 소개

1. MASIA (2022, NeurIPS)
2. DSR (2025, AAMAS)



시야가 너무 좁을 때



시야가 너무 넓을 때

Methods

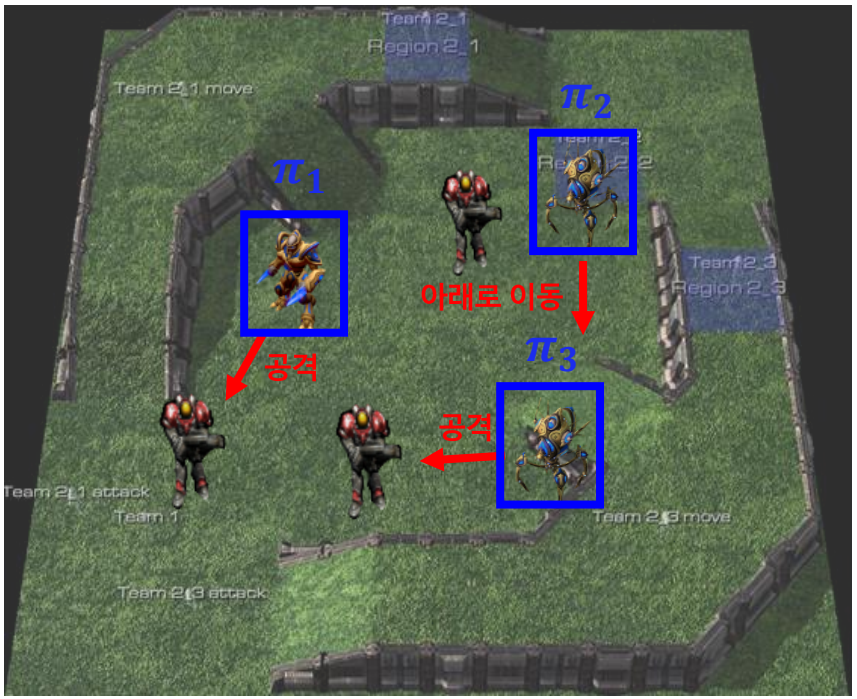
Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ 연구 배경

1. Cooperative MARL에서 **부분 관측성**과 다수의 에이전트가 동시에 학습하는 **비정상성 문제**가 존재
   → 협력 성능을 크게 저하시킴

동시에 행동 정책 학습 → 에이전트마다 전이 확률과 보상이 달라짐

부분 관측성 문제



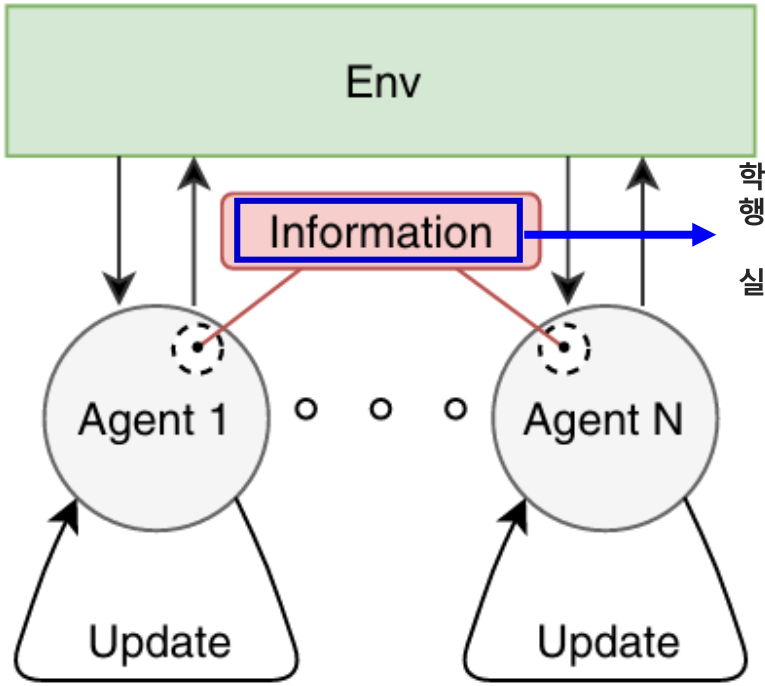
비정상성 문제

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ 연구 배경

1. Cooperative MARL에서 부분 관측성과 다수의 에이전트가 동시에 학습하는 비정상성 문제가 존재
   → 협력 성능을 크게 저하시킴
2. 이를 보완하기 위해, **CTDE 패러다임**이 널리 사용
   → 실행 시 다른 에이전트 상태나 행동에 대한 불확실성으로 인해 여전히 협력 실패가 발생



학습: 중앙집중식 정보(모든 에이전트의 관측, 행동, 전역 상태 등)를 활용

실행: 각 에이전트 본인의 관측값에 기반해서 행동

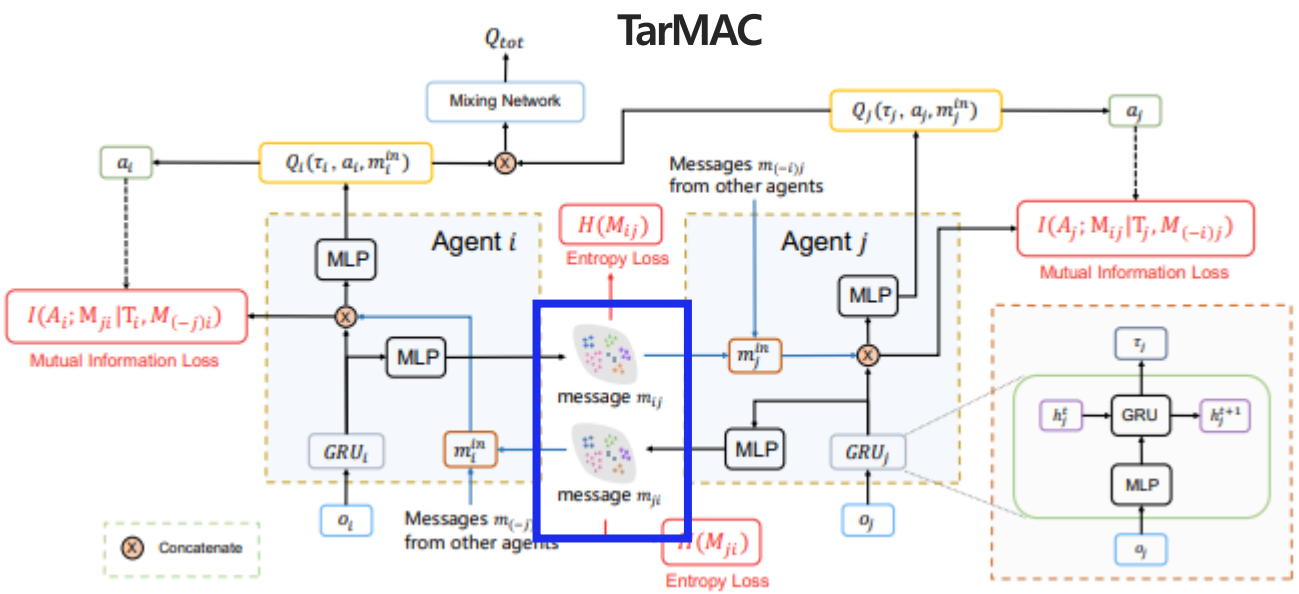
Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55(2), 895-943.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ 연구 배경

1. Cooperative MARL에서 부분 관측성과 다수의 에이전트가 동시에 학습하는 비정상성 문제가 존재
   → 협력 성능을 크게 저하시킴
2. 이를 보완하기 위해, CTDE 패러다임이 널리 사용
   → 실행 시 다른 에이전트 상태나 행동에 대한 불확실성으로 인해 여전히 협력 실패가 발생
3. 따라서, 에이전트 간의 **통신**이 핵심 대안으로 떠오름
4. 하지만, 여러 메시지를 **집약하는 과정을 명시적으로 다루지 않아 비효율성**이 발생함 (메시지 중복, 무관한 정보)



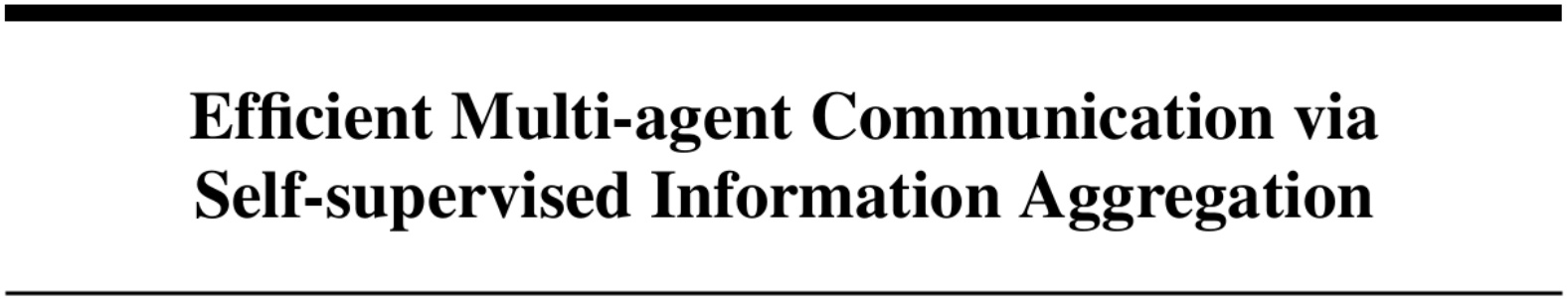
Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly de composable value functions via communication minimization. In International Conference on Learning Representations, 2020.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **문제 정의**

- Cooperative MARL에서 여러 에이전트 메시지를 어떻게 효율적으로 집계하고, 개별 에이전트의 의사 결정에 유용한 형태로 추출할 것인가?
- 단순히 raw message를 모두 policy 입력으로 넣는 방식은 불필요한 정보가 과다하게 포함되어 학습을 방해
- **따라서, 메시지 집약(aggregation) → 중요 정보 추출 (extraction) 과정이 필요함**



Efficient Multi-agent Communication via Self-supervised Information Aggregation

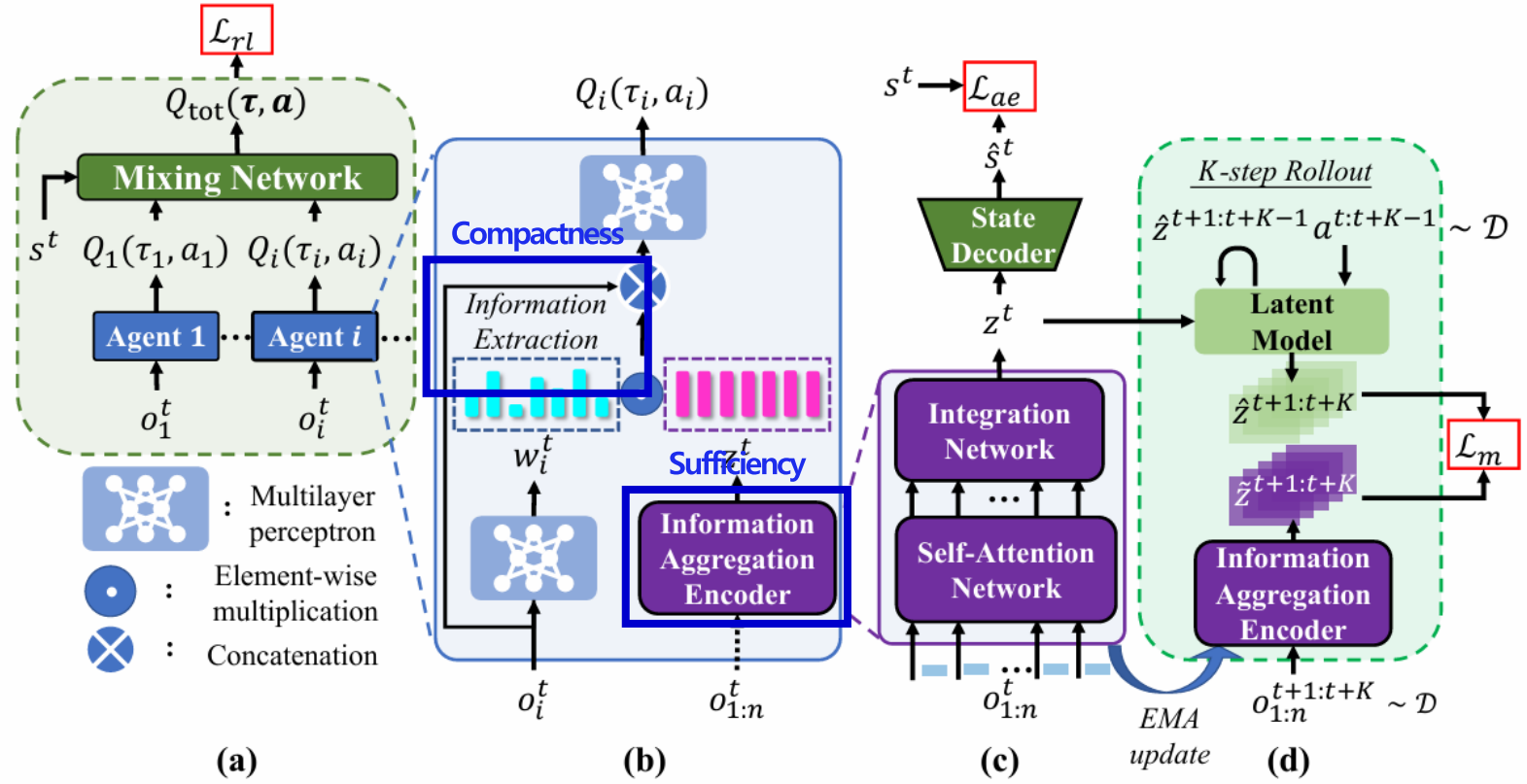
Cong Guan[1]*, Feng Chen[1]*, Lei Yuan[1,2], Chenghe Wang[1],
Hao Yin[1], Zongzhang Zhang[1], Yang Yu[1,2]†
[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2] Polixir Technologies
{guanc, chenf, yuanl, wangch, yinh}@lamda.nju.edu.cn
{zzzhang, yuy}@nju.edu.cn

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ MASIA

- 효율적인 통신 메커니즘 디자인을 위해 **"Sufficiency"**와 **"Compactness"** 고려
- Sufficiency → 많은 양의 정보 (메시지 집약) / Compactness → 밀도 높은 정보 (중요 정보 추출)
- Information Aggregation Encoder (IAE) for **Sufficiency** / focusing network for **Compactness**



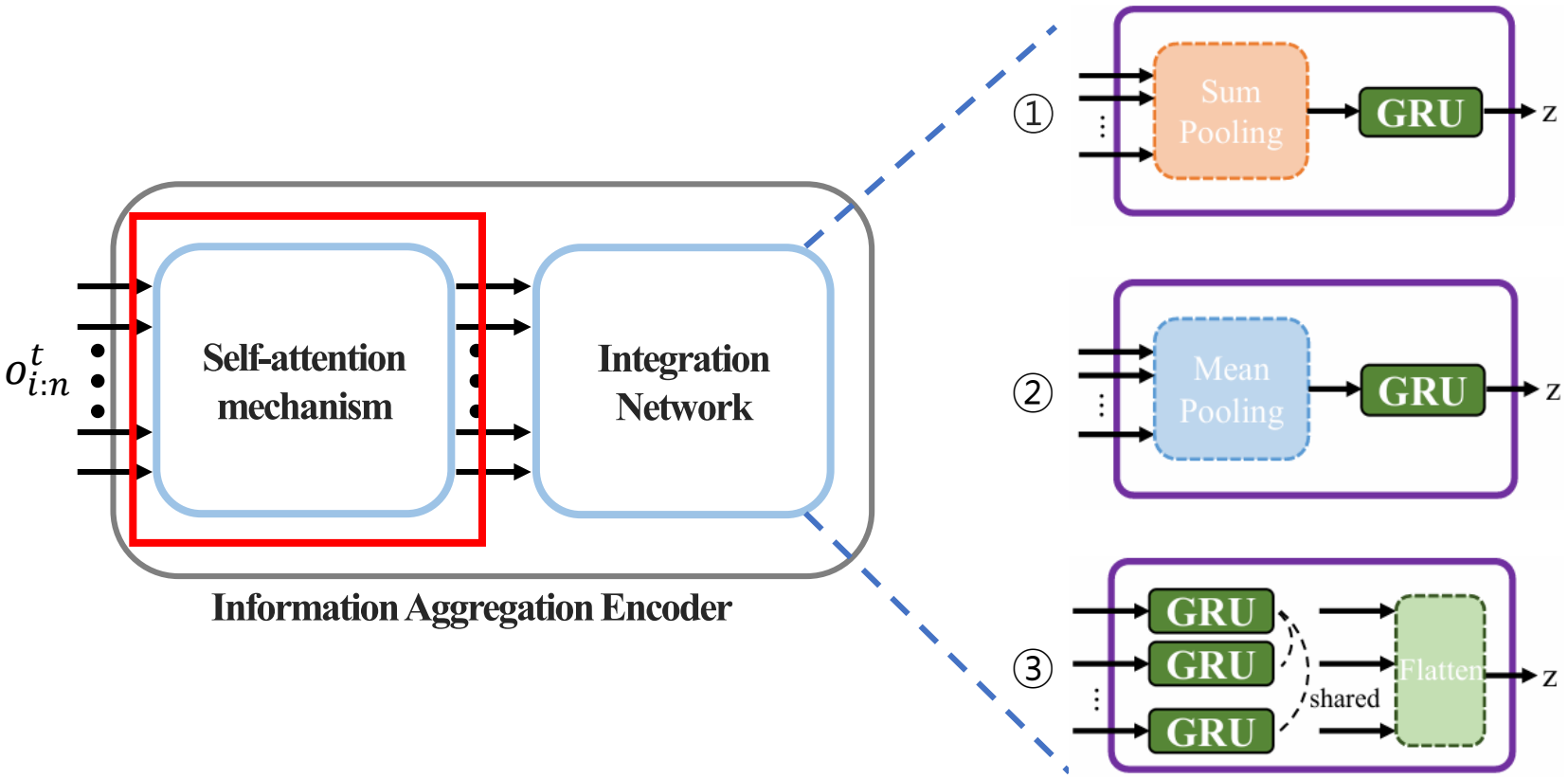
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **MASIA – IAE (for sufficiency)**

- 다중 에이전트 시스템에서의 통신은 메시지의 순서보다 어떤 메시지가 중요한지를 아는 것이 중요
- 따라서, IAE에서 self-attention mechanism이 사용됨 → **자신에게 어떤 에이전트의 정보가 중요한지 확인**



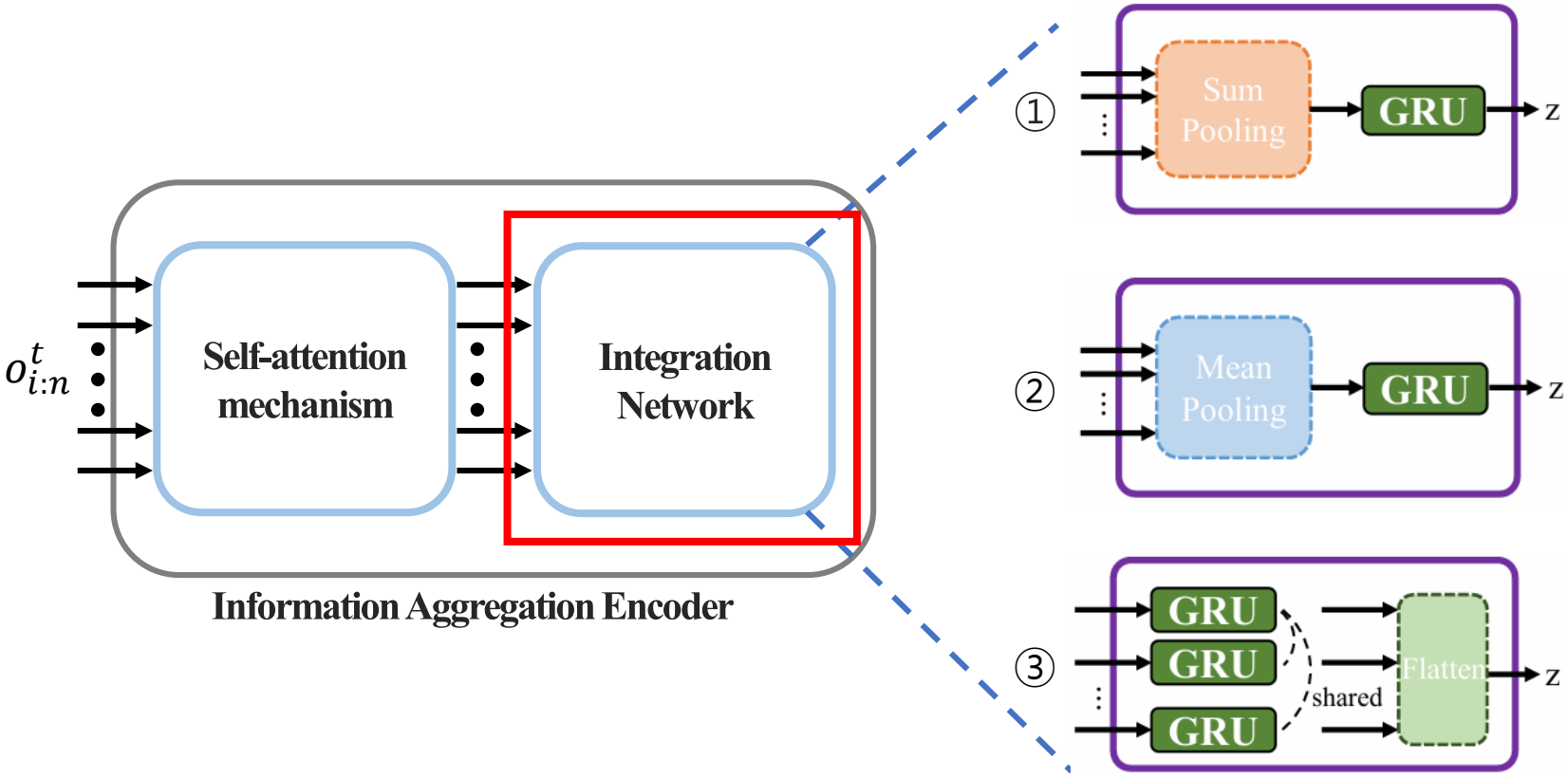
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **MASIA – IAE (for sufficiency)**

- 세 가지 다른 형태(①, ②, ③)의 integration network를 소개함
- 이 중 ③ 버전이 가장 좋은 성능을 보였으며, 본 논문의 모든 실험에서 ③ 구현을 기반으로 함



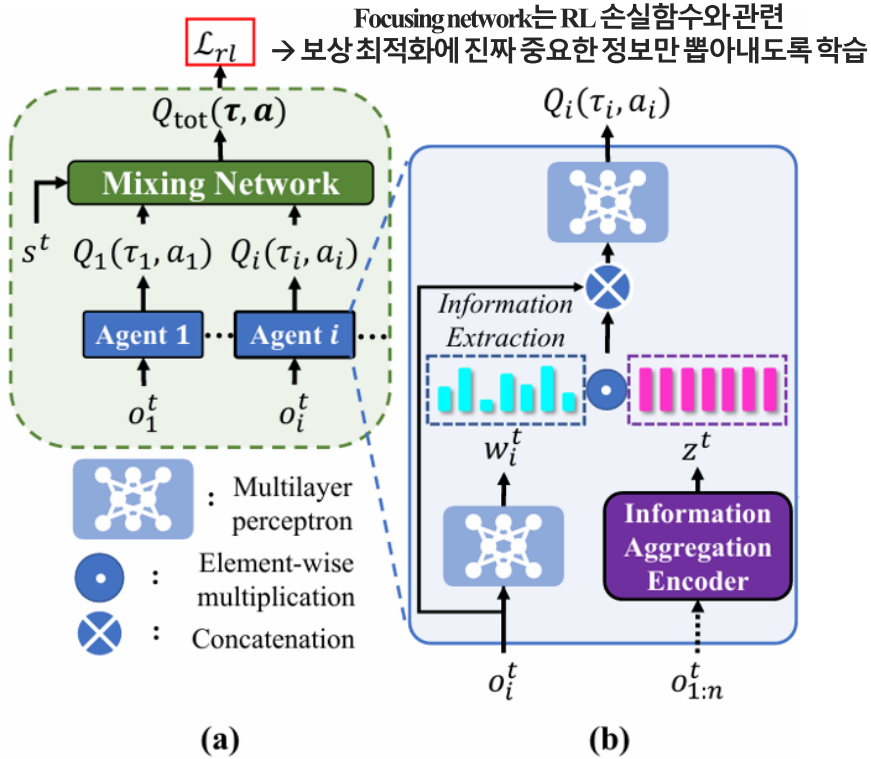
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ MASIA – Focusing network (for compactness)

- 각 에이전트에 대한 통합된 정보에서 중요 정보에 따른 가중치를 주기 위해 **focusing network** 사용
- Focusing network는 MLP+sigmoid로 구성되어 있음 → w_i^t가 모두 0~1 사이에 위치함
- w_i^t 내 특정 차원에 대해서 높은 가중치를 부여한다면, 에이전트는 그 특정 부분에 더 민감하게 반응
- 반면에, 특정 차원에 대해서 거의 0에 가까운 가중치가 출력되면, 그 차원의 정보는 걸러짐



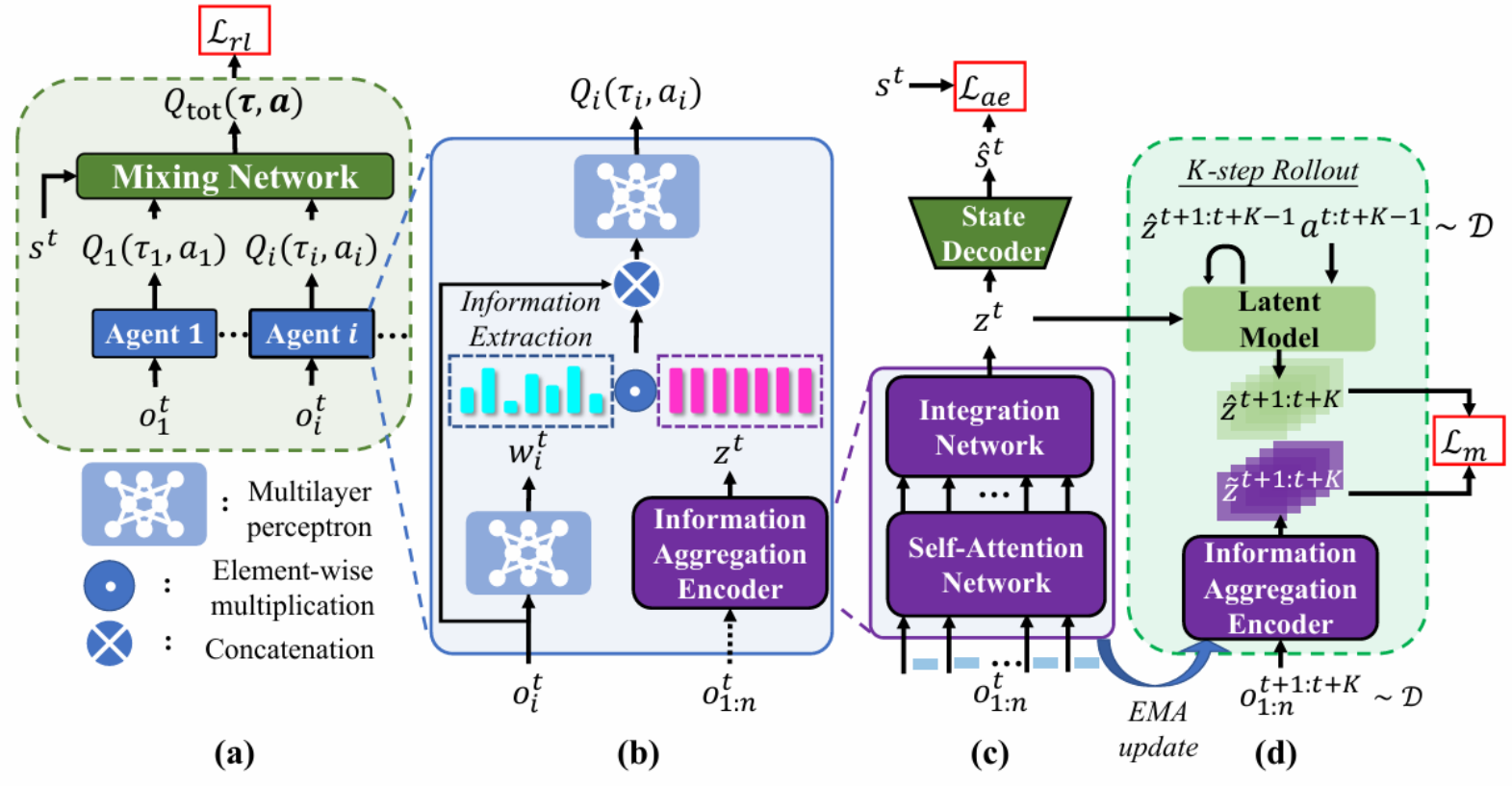
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **MASIA – how to optimize aggregation representation**

- **정보의 집계**와 **중요 정보만을 추출**한 통합 표현을 잘 만들기 위해 **두 가지 추가 보조 손실 함수** 사용
- L_{ae}, L_m: 재구성과 다단계 예측을 위한 손실 함수이며, 이는 각각 **충분성**과 **간결성**을 갖추도록 제약 두는데 사용



Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

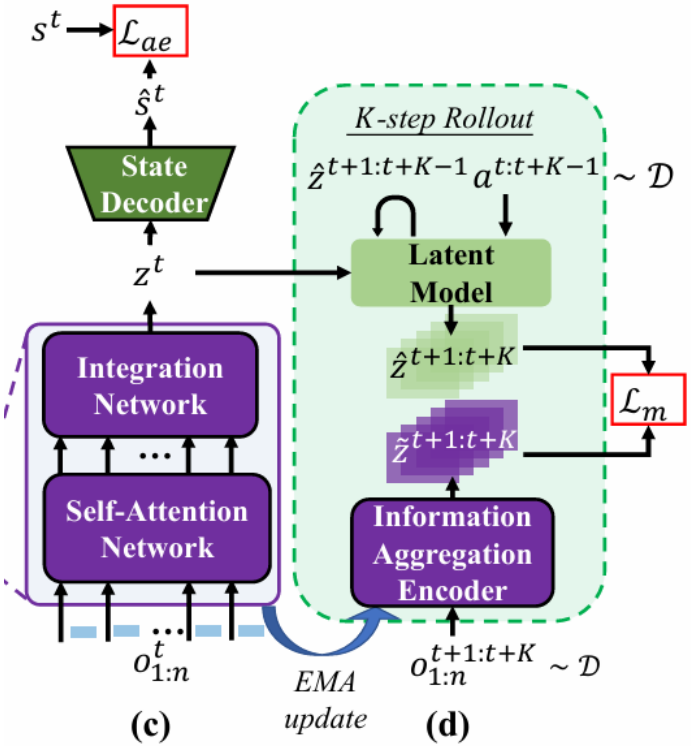
❖ **MASIA – how to optimize aggregation representation**

- Reconstruction task: 집약 표현으로부터 전역 상태를 재구성하는 것이 목표
- 재구성 손실 함수를 통해, 인코더가 전역 상태를 추론하는 데 도움이 되는 관측 특징을 추출하도록 유도
- 이로 인해, z_t 안에는 부분 관측들을 잘 통합한 충분한 특징 집합이 될 수 있음 → **Sufficient representation**

Reconstruction Loss

$$\mathcal{L}_{ae}(\theta, \eta) = \mathbb{E}_{\boldsymbol{o}^t, s^t} \| g_\eta(z^t) - s^t \|_2^2, \qquad z^t = f_\theta(\boldsymbol{o}^t)$$

State Decoder　　IAE



Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ MASIA – how to optimize aggregation representation

- Multi-step prediction task: t 시점의 집계된 표현(z_t)과 공동 행동(a_t)을 기반으로 t+1시점의 집약 표현(z_{t+1})을 예측
- 이 손실 함수를 통해, 집계 표현(z_t)이 의사결정에 중요한 정보와 강하게 연결되도록 유도 → **compactness**
- 학습 과정 안정화를 위해, k-step rollout과정에서 사용된 IAE는 지수 이동 평균으로 업데이트됨 → like in SPR

Reconstruction Loss

$$\mathcal{L}_{ae}(\theta,\eta)=\mathbb{E}_{\boldsymbol{o}^t,s^t}\|g_\eta(z^t)-s^t\|_2^2, \qquad z^t=f_\theta(\boldsymbol{o}^t)$$

State Decoder　　IAE

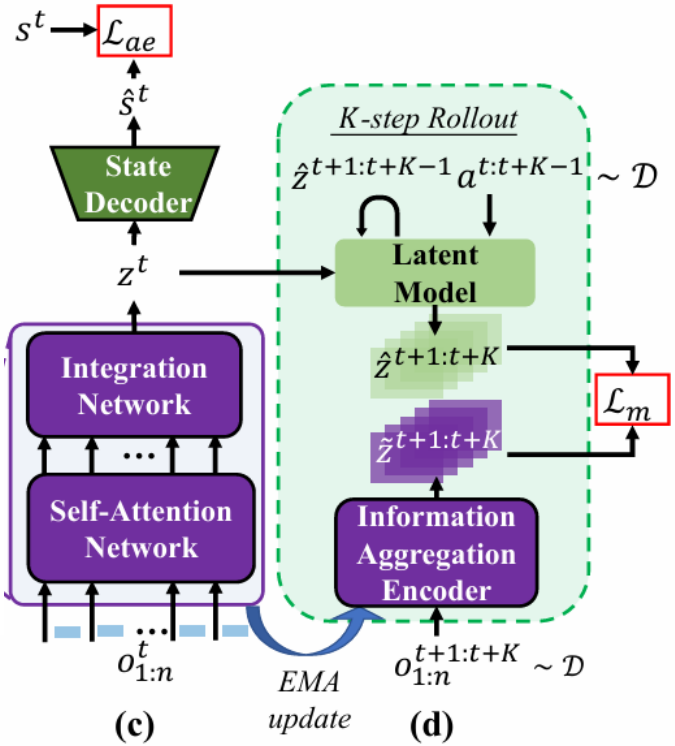
Multi-step prediction loss

$$\mathcal{L}_m(\theta,\psi)=\mathbb{E}_{\boldsymbol{o}^t,s^t,\boldsymbol{a}^t,\ldots,\boldsymbol{a}^{t+K-1},\boldsymbol{o}^{t+K},s^{t+K}}\left[\sum_{k=1}^{K}\|\hat{z}^{t+k}-\tilde{z}^{t+k}\|_2^2\right],$$

Transition Model

$$\hat{z}^{t+1}=h_\psi(\tilde{z}^t,\boldsymbol{a}^t),$$
$$\hat{z}^{t+k}=h_\psi(\hat{z}^{t+k-1},\boldsymbol{a}^{t+k-1}),\quad k=2,\ldots,K,$$
$$\tilde{z}^{t+k}=f_\theta(\boldsymbol{o}^{t+k}),\quad k=0,\ldots,K.$$

IAE



Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **MASIA – how to optimize aggregation representation**

Total Loss Function

$$\mathcal{L}_{rl}(\theta,\phi)+\lambda_1\mathcal{L}_{ae}(\theta,\eta)+\lambda_2\mathcal{L}_{m}(\theta,\psi)$$

$$\mathcal{L}_{rl}(\theta,\phi)=\mathbb{E}_{(\boldsymbol{\tau},\boldsymbol{a},r,\boldsymbol{\tau}')\in\mathcal{D}}\left[\left(r+\gamma\max_{\boldsymbol{a}'}Q_{\text{tot}}\left(\boldsymbol{\tau}',\boldsymbol{a}';\theta^-,\phi^-\right)-Q_{\text{tot}}(\boldsymbol{\tau},\boldsymbol{a};\theta,\phi)\right)^2\right]$$

$$\mathcal{L}_{m}(\theta,\psi)=\mathbb{E}_{\boldsymbol{o}^t,s^t,\boldsymbol{a}^t,\ldots,\boldsymbol{a}^{t+K-1},\boldsymbol{o}^{t+K},s^{t+K}}\left[\sum_{k=1}^{K}\|\hat{z}^{t+k}-\tilde{z}^{t+k}\|_2^2\right]$$

$$\mathcal{L}_{ae}(\theta,\eta)=\mathbb{E}_{\boldsymbol{o}^t,s^t}\|g_\eta(z^t)-s^t\|_2^2,\qquad z^t=f_\theta(\boldsymbol{o}^t)$$

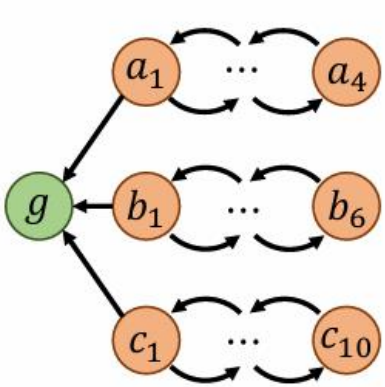
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

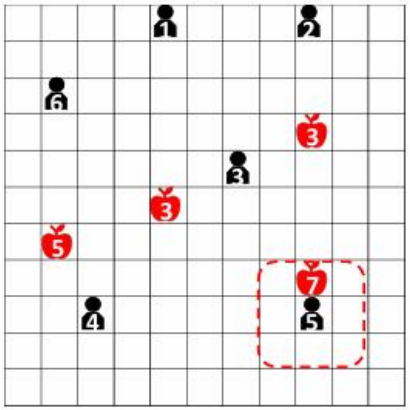
Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

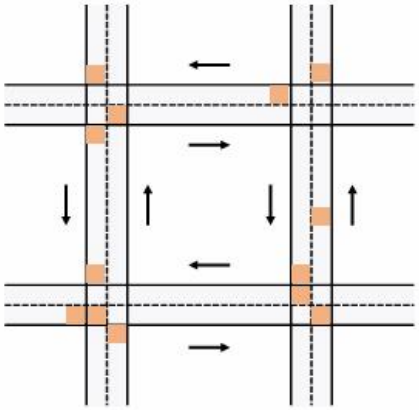
❖ 실험에 사용한 환경

- Hallway: 여러 에이전트가 서로 다른 위치에서 무작위로 초기화되며, 동시에 목표 지점 g에 도달해야 하는 환경
- Level Based Foraging (LBF): 에이전트들이 동시에 음식을 수집하기 위해 협력해야 하는 환경
- Traffic Junction: 통신 능력을 평가하기 위한 환경으로, 다수의 차량이 하나 이상의 교차로를 포함한 양방향 도로를 따라 미리 정의된 경로를 이동하는 환경
- SMAC: 적의 위치를 파악하고 통신을 통해 적군을 섬멸해야하는 combat mini scenario 게임 환경



(a) Hallway


(b) LBF


(c) Traffic Junction (TJ)

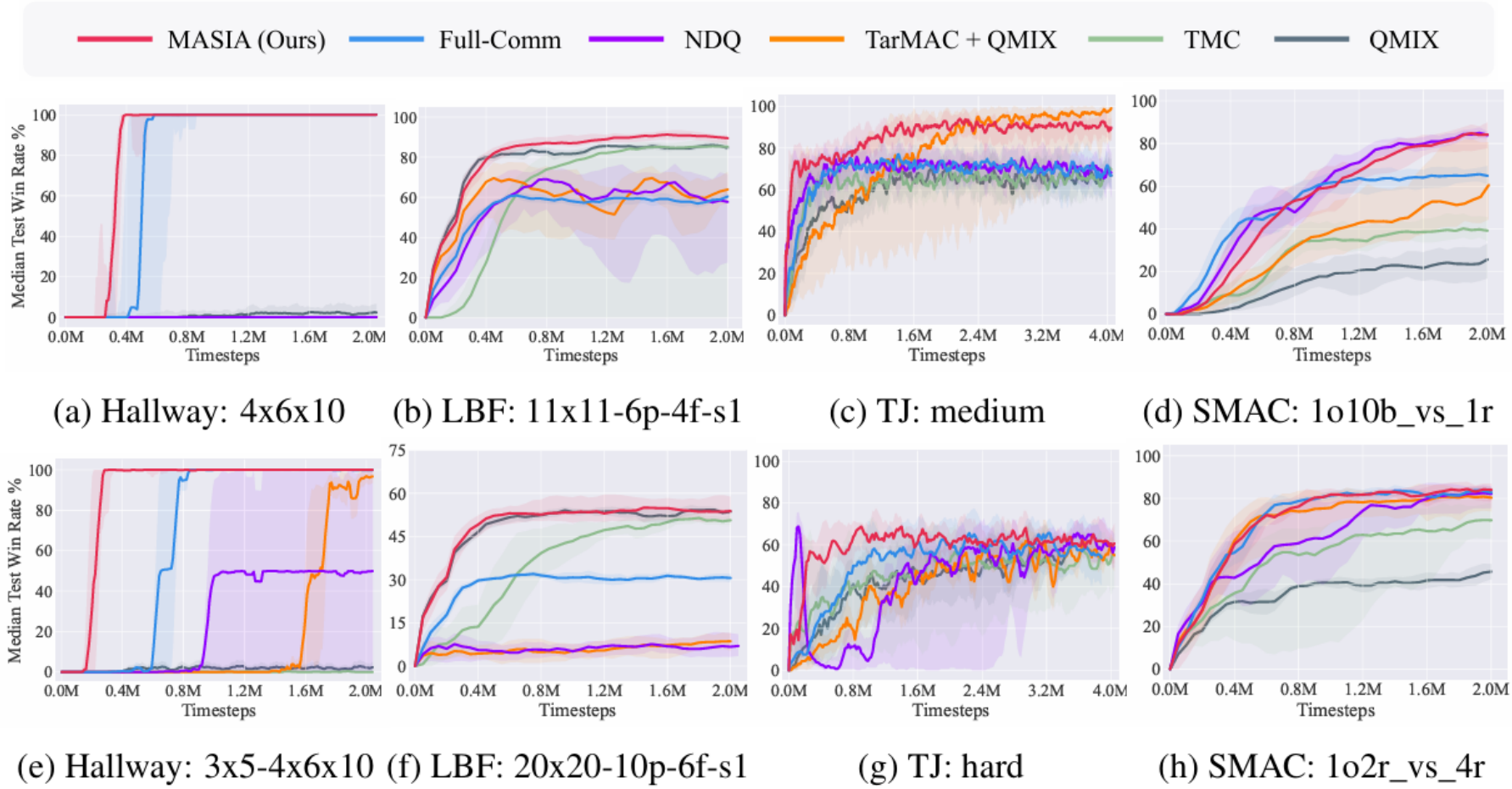

(d) SMAC

Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **Experiment ① - MASIA가 다양한 시나리오에서 여러 기준선과 비교했을 때, 어떻게 성능을 보이는가?**

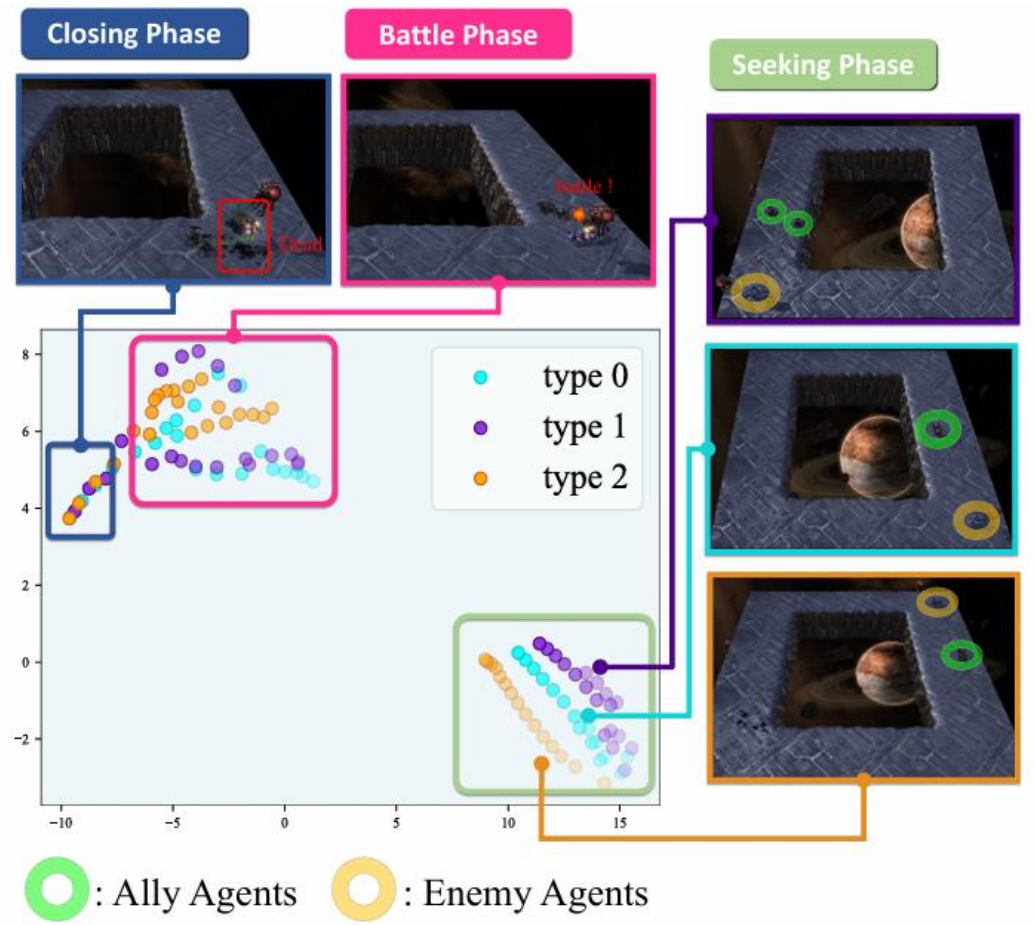


Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **Experiment ② - IAE는 어떤 지식을 학습하는가?**



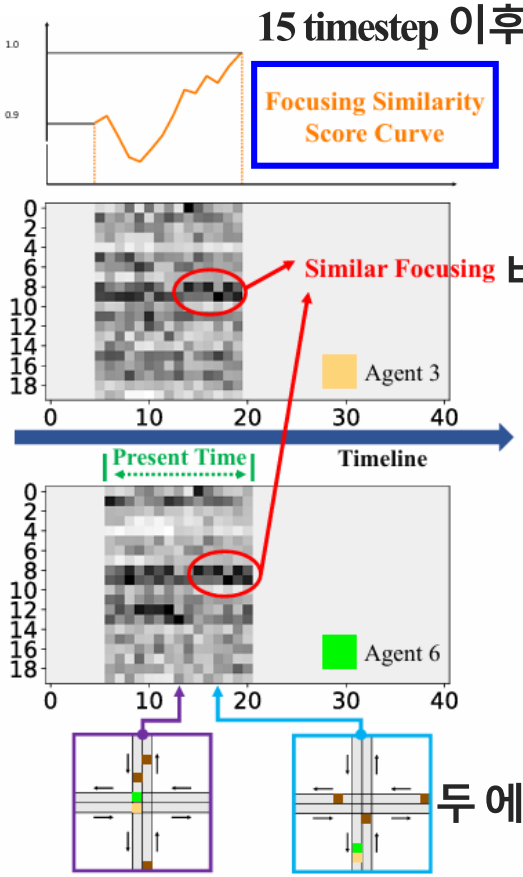
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ Experiment ③ - focusing network는 학습된 임베딩 공간으로부터 개별 에이전트에게 가장 관련성이 높은 정보를 추출할 수 있는가? (Traffic Junction 환경 사용)

15 timestep 이후 점점 유사도가 올라감



비슷한 특징 차원에 가중치를 많이 줌

두 에이전트가 동일한 경로를 향해 가고 있음

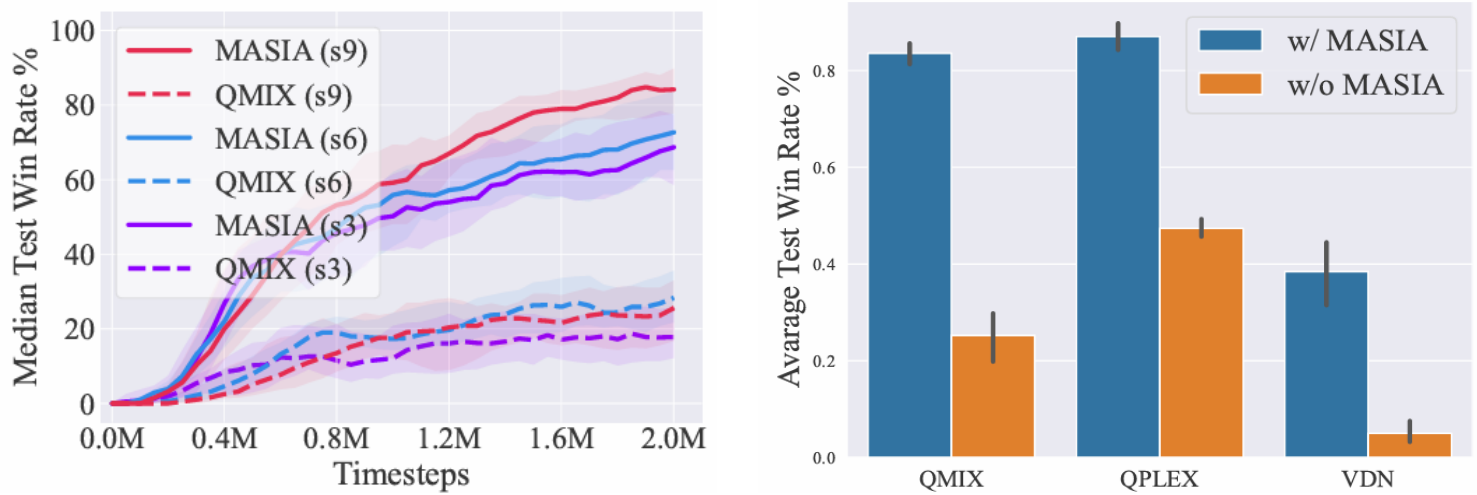
Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

Efficient Multi-agent Communication via Self-supervised Information Aggregation (2022, NeurIPS)

❖ **Experiment ④ - MASIA는 다양한 시야 범위와 다양한 가치 분해 기반 방법들에 적용되었을 때에도 강건한 성능을 보일까?**

- 에이전트의 시야 범위를 점점 더 좁혔음에도 불구하고, 성능이 크게 저하되지 않음 → 시야 제약에도 강건함
- 효율적인 통신이 가능한 MASIA가 다양한 MARL 방법론에서도 큰 성능 개선이 이뤄지는 것을 확인



SMAC의 1o10b_vs_1r 시나리오에 대한 실험 결과

Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.

Methods

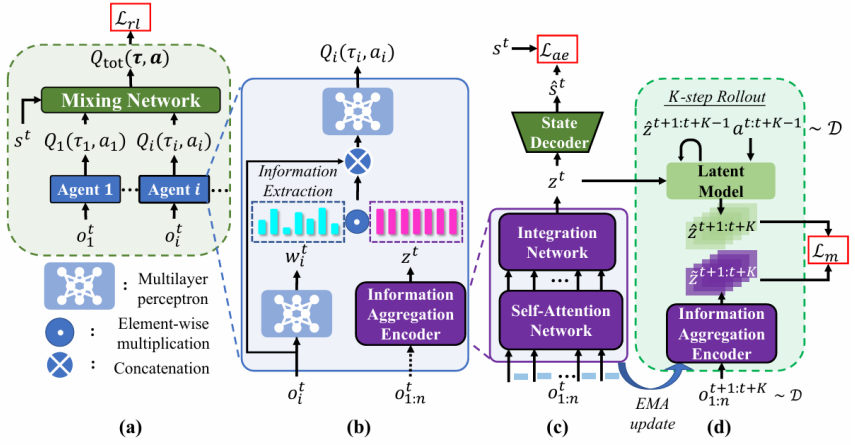
Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ 연구 배경

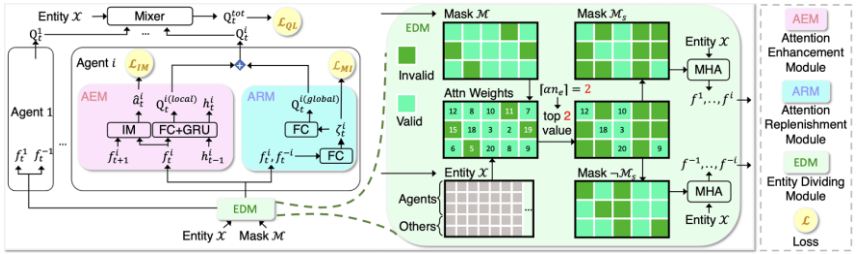
- Partial Observability 때문에 각 에이전트는 제한된 시야만 확보 가능
- 시야 범위가 너무 좁으면, 협력에 필요한 정보를 놓치고, 너무 넓으면 불필요한 정보가 과도하게 포함

  → **Sight range dilemma**
- 기존 연구들은 글로벌 정보를 사용하는 셀프 어텐션 기반 통신 방법을 통해 해당 딜레마를 개선

  → 어느 정도 완화는 했지만, **현실 세계에서는 글로벌 정보 획득이 어려우며 얻는데 비용**이 많이 요구됨



MASIA



CAMA

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **문제 정의**

- 그렇다면 **글로벌 정보를 사용하는 통신 시스템 없이 sight range dilemma를 직접 해결**할 수 있는 방법은 없는가?
- 학습 과정에서 에이전트의 **개별 시야 범위만을 제어**하여, 해당 딜레마를 해결하는 방법이 연구됨

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning

Wei-Chen Liao
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
wcl.cs11@nycu.edu.tw

Ti-Rong Wu
Institute of Information Science, Academia Sinica
Taipei, Taiwan
tirongwu@iis.sinica.edu.tw

I-Chen Wu
Department of Computer Science, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
icwu@cs.nycu.edu.tw(correspondence)

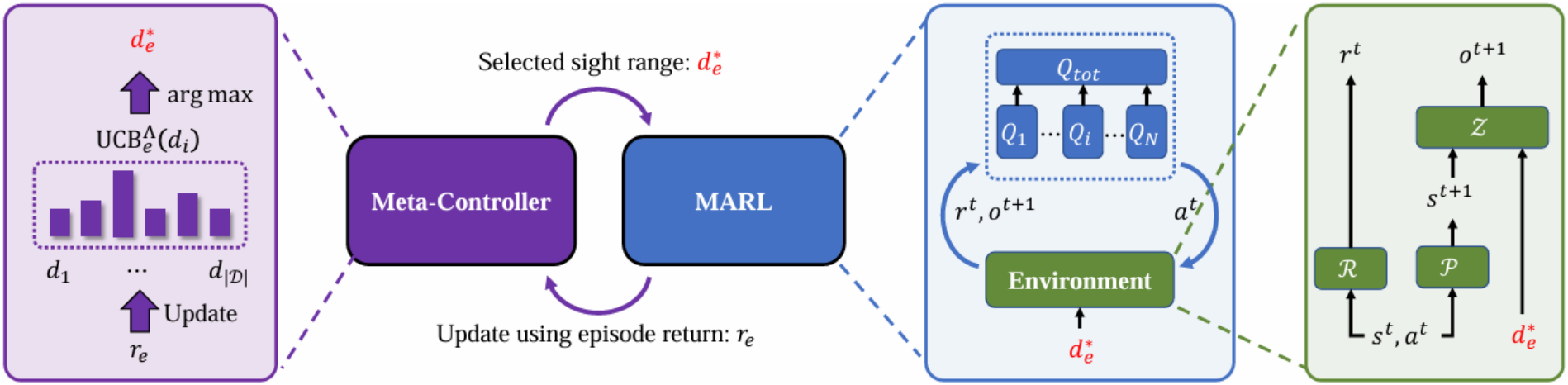
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ DSR

- DSR을 **어떠한 MARL 방법론도 다 적용 가능** → DSR = module
- 동적으로 주어지는 sight range에 따라, **observation function을 수정**
- **meta-controller**를 통해 어떤 sight range가 가장 적합한지 결정
- **이를 통해, 최적의 시야 범위를 환경에게 전달**



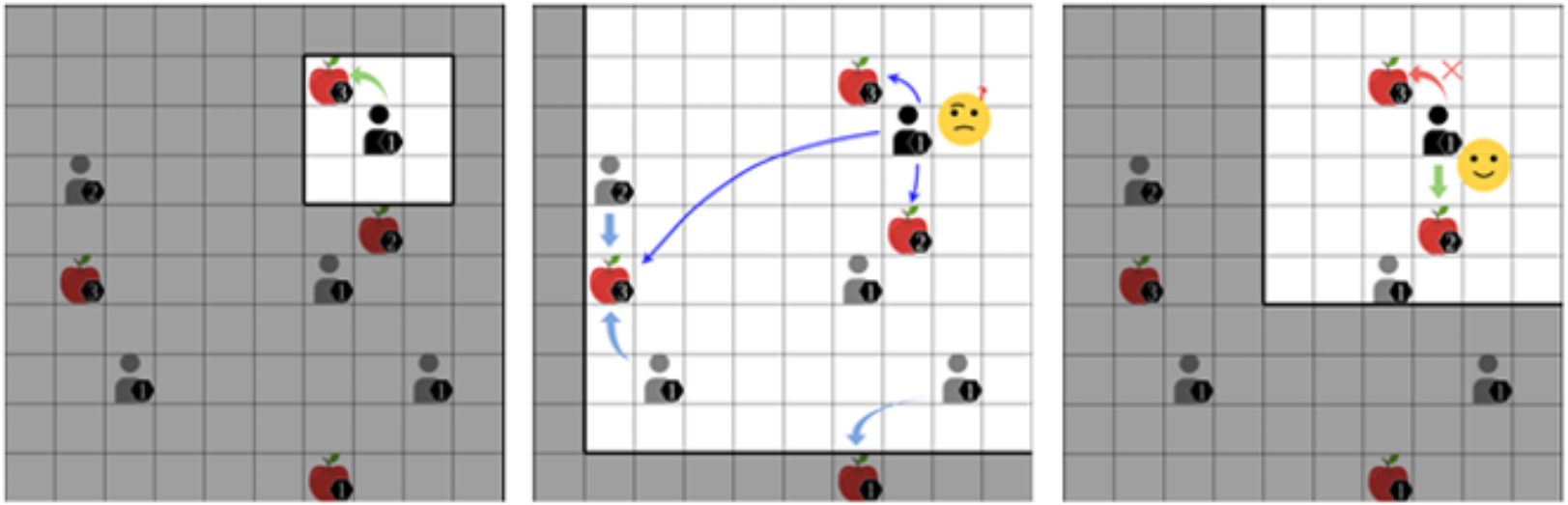
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ DSR – observation function 수정?

- 기존의 observation function $o_t^i = Z(s_t, n_i)$는 환경의 전체 상태 s_t와 특정 에이전트 n_i를 입력받아, 에이전트가 가질 수 있는 부분적 정보 o_t^i를 반환
- DSR에서는 observation function에 시야 범위 제약을 주기 위해 $o_t^i = Z(s_t, n_i, d)$로 수정
- 예시처럼, 실제 응용에서는 시야 범위가 반드시 에이전트를 중심으로 한 고정된 거리는 아님
- 자율주행에서는 서로 다른 d 값이 서로 다른 범위와 모양을 가지는 센서 설계를 의미할 수도 있음
- 즉, **목표는 observation function에 d를 제약함으로써, 학습 과정 동안 적절한 시야 범위 d를 찾는 것**



$d = 1$ $d = 6$ $d = 3$

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ Upper Confidence Bound (UCB)

- 의사결정 문제에서 **탐험(exploration)**과 **활용(exploitation)**의 균형을 맞추기 위해 널리 사용되는 방법
- **경험적 평균 보상**과 **상한 신뢰 구간**을 함께 고려하여, 가장 높은 예상 보상을 가진 행동을 선택
  - $\hat{X}_t(a_i)$: 시점 t까지 행동 a_i가 선택되었을 때의 경험적 평균 보상
  - $N_t(a_i)$: 행동 a_i가 선택된 횟수
  - c: 탐험과 활용을 조절하는 하이퍼 파라미터

탐험 유도

$$UCB_t(a_i) = \hat{X}_t(a_i) + c \times U_t(a_i) = \hat{X}_t(a_i) + c \times \sqrt{\frac{\log t}{N_t(a_i)}}$$

활용 유도

$$\hat{X}_t(a_i) = \frac{1}{N_t(a_i)} \sum_{j=1}^{t} r_j(a_i) \cdot \mathbf{1}[a_j = a_i] = \begin{cases} 1 & \text{if } a_j = a_i \\ 0 & \text{if } a_j \neq a_i \end{cases}$$

j번째 시점에서 선택된 행동

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ Sliding-window Upper Confidence Bound (SW-UCB)

- 기존의 **UCB는 비정상 환경**을 고려하지 못함 → Multi-agent System에서 발생하는 문제 중 하나 = 비정상성
- 따라서, 비정상 환경을 고려할 수 있는 **비정상 UCB 알고리즘이** 개발됨
- UCB와 다른 점 → 첫 번째 항 (경험적 평균 보상) → 전체 이력이 아닌, 최근 관측값들로 한정된 윈도우로 한정

$$UCB_t^{SW}(a_i) = \hat{X}_t^{SW}(a_i) + c \times U_t^{SW}(a_i),$$

$$= \frac{1}{N_t(a_i, w)} \sum_{j=t-w+1}^{t} r_j(a_i) \mathbb{1}_{\{UCB_j^{SW}=i\}}$$

$$+ c \times \sqrt{\frac{\log \min(t, w)}{N_t(a_i, w)}}, \text{ and}$$

$$N_t(a_i, w) = \sum_{j=t-w+1}^{t} \mathbb{1}_{\{UCB_j^{SW}=i\}},$$

슬라이딩 윈도우 내에서 행동 a_i가 선택된 횟수

슬라이딩 윈도우 내에서 SW-UCB가 선택한 행동이 i인가?

초기에는 t < w → logt를 사용
충분히 시점이 쌓이면 t > w → log w 사용
→ 최근 w개의 기록만을 기준으로 탐험 압력을 유지하기 위함

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **DSR – meta-controller?**

- **Sight range dilemma를 직접적으로 해결**하는 부분으로, 각 에피소드의 시작 시 **SW-UCB를 적용**하여 최적의 시야 범위를 도출함
- 결과적으로, **meta-controller Λ**를 통해 **최적의 시야 범위를 조정 및 선택**하고, 그 동안 MARL 알고리즘은 선택된 시야 범위 내에서 **전역 가치 함수 Q^{π}를 최대화**하는 데 집중함
- 실행 단계에서는 단순히 슬라이딩 윈도우 내 평균 리턴이 가장 큰 시야 범위($argmax(r(d_i))$)를 선택

$$UCB_e^{\Lambda}(d_i) = \hat{X}_e(d_i) + c \times U_e(d_i)$$

$$= \frac{1}{N_e(d_i, w)} \sum_{j=e-w+1}^{e} r_j(d_i) \mathbb{1}_{\{UCB_j^{\Lambda}=i\}}$$

$$+ c \times \sqrt{\frac{\log \min(e, w)}{N_e(d_i, w)}},$$

e: 에피소드

j: 에피소드 인덱스

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

- ❖ **DSR – training algorithm?**
  - ❖ 계층적 최적화 진행 → **(상위 레벨) 시야 범위 최적화**와 **(하위 레벨) 행동 정책 학습 문제**가 계층적으로 분리되어 동시에 최적화됨

Algorithm 1 Dynamic Sight Range Selection (DSR)

1: **Input:** Set of sight ranges $\mathcal{D}$, sliding window size w, constant c, total number of training episodes E, total number of training steps per episode T
2: **Output:** The best sight range d^*
3: **for** episode $e = 1$ to E **do**
4: Meta-controller selects sight range d_e^*:

$$d_e^* = \arg\max_{d_i \in \mathcal{D}} \hat{X}_e(d_i) + c \times U_e(d_i)$$

5: Generate one episode with a modified observation function $\mathcal{Z}(s, n_i, d_e^*)$
6: Obtain the episode return r_e
7: Train the episode by any MARL algorithm
8: **if** $d_{e-t} \neq d_e^*$ **then**

$$N_e(d_e^*, w) = N_e(d_e^*, w) + 1$$

9: **end if**
10: Record reward r_e for d_e^*
11: **end for**

에피소드 시작 시 meta-controller를 통해 최적의 시야 범위 선택

선택된 시야 범위에 따라 관측값이 수정되고, 수정된 관측값이 환경과의 상호작용 동안 에이전트가 사용하는 입력이 됨

윈도우 내에서 선택된 시야 범위의 수가 갱신됨

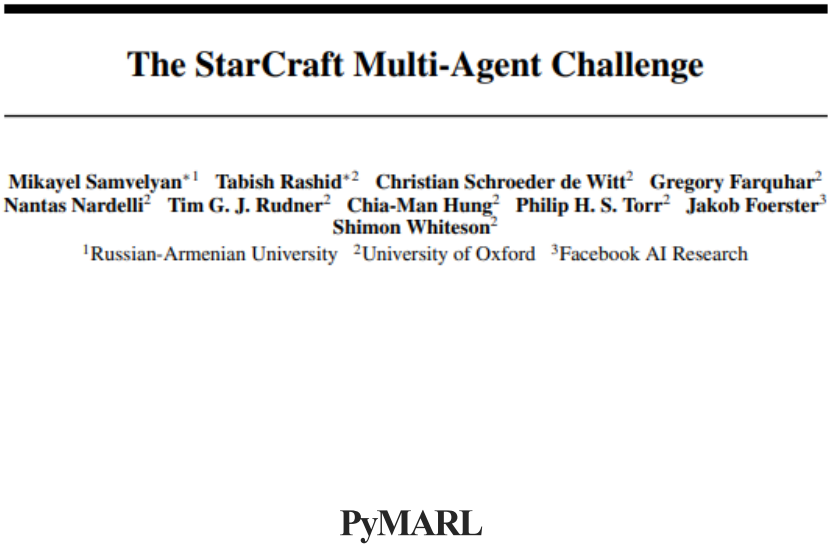
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

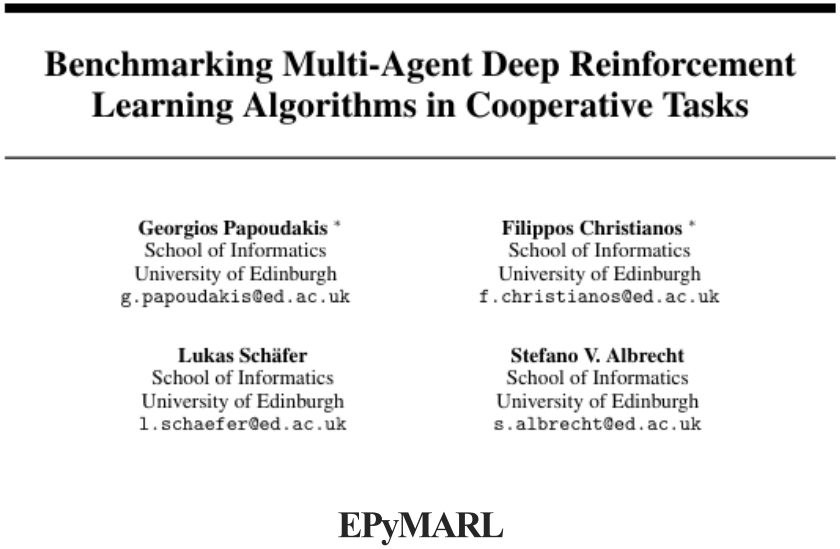
❖ **실험을 위해 사용한 오픈소스 벤치마킹 프레임워크**

- PyMARL (2019, AAMAS): SMAC 환경에서 QMIX, VDN, QTRAN 실험을 위해 만든 코드 프레임워크
- EPyMARL (2021, arXiv): PyMARL을 확장해서 다양한 환경과 더 많은 최신 알고리즘 실험이 지원된 코드 프레임워크


The StarCraft Multi-Agent Challenge

Mikayel Samvelyan[*1] Tabish Rashid[*2] Christian Schroeder de Witt[2] Gregory Farquhar[2]
Nantas Nardelli[2] Tim G. J. Rudner[2] Chia-Man Hung[2] Philip H. S. Torr[2] Jakob Foerster[3]
Shimon Whiteson[2]
[1]Russian-Armenian University [2]University of Oxford [3]Facebook AI Research

PyMARL


Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks

Georgios Papoudakis *
School of Informatics
University of Edinburgh
g.papoudakis@ed.ac.uk

Filippos Christianos *
School of Informatics
University of Edinburgh
f.christianos@ed.ac.uk

Lukas Schäfer
School of Informatics
University of Edinburgh
l.schaefer@ed.ac.uk

Stefano V. Albrecht
School of Informatics
University of Edinburgh
s.albrecht@ed.ac.uk

EPyMARL

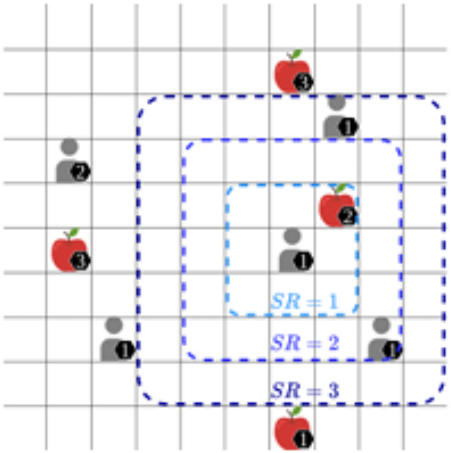
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

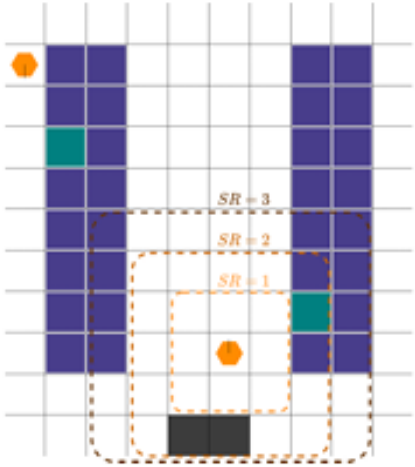
Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ 실험에 사용한 환경

- Level-Based Foraging (LBF): 음식을 모으기 위해 다중 에이전트 협력 행동을 평가하기 위해 설계된 격자 기반 환경
- Multi-Robot Warehouse (RWARE): 협력하는 로봇들이 창고 내에서 물품을 운반하는 과제를 수행하는 격자 기반의 다중 에이전트 환경
- StarCraft Multi-Agent Challenge (SMAC): Starcraft 2에서의 아군(에이전트)이 적군을 섬멸하는 것이 목표인 다중 에이전트 환경



Level-Based Foraging (LBF)



Multi-Robot Warehouse (RWARE)

StarCraft Multi-Agent Challenge (SMAC)

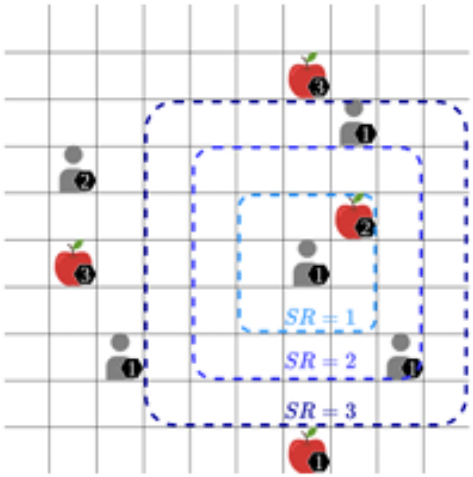
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ 실험에 사용한 환경 - LBF

- 에이전트들이 격자 세계에서 음식을 모으기 위해 협력해야 하는 환경
- 에피소드 리턴 값은 맵에 존재하는 총 음식 점수 대비 실제로 수집된 음식 비율 (0~1 사이)
- 시야 범위는 에이전트의 현재 위치를 중심으로 관측할 수 있는 격자 거리로 정의
- 아래 그림에서, d=10이면 에이전트가 맵 전체를 관측할 수 있음을 의미 → 글로벌 상태에 접근하는 것과 동일
- 시야 범위 집합은 두 가지 설정 사용: D = {2,4,**6**} / D = {2,4,6,8,**10**}



Level-Based Foraging (LBF)
(10X10-5p-4f-coop)

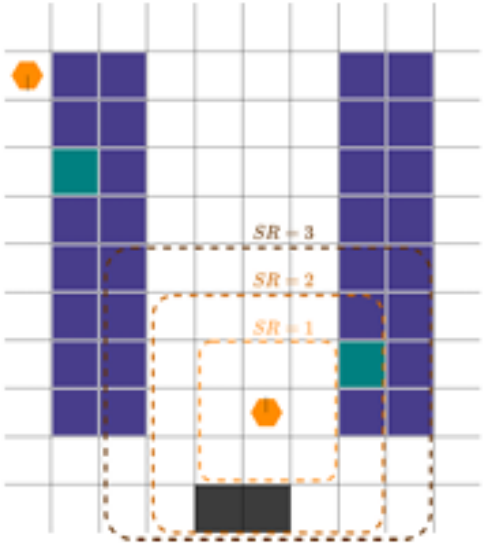
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ 실험에 사용한 환경 - RWARE

- 에이전트들이 요청된 선반(청록색으로 표시된 격자)을 찾아 지정된 위치(그림 하단의 검은색 격자)로 옮긴 뒤, 선반을 비어 있는 위치로 되돌려 놓는 작업을 반복
- 성공적인 운송에는 +1의 보상을 줌
- 해당 환경은 10X11 (tiny) 격자 맵과 10X20 (small) 격자 맵, 두 가지 맵 크기를 사용하며, 두 명의 에이전트 사용
- 시야 범위 집합은 두 가지 설정 사용: D = {1,2,**3**} / D = {1,2,3,4,**5**}



Multi-Robot Warehouse (RWARE)

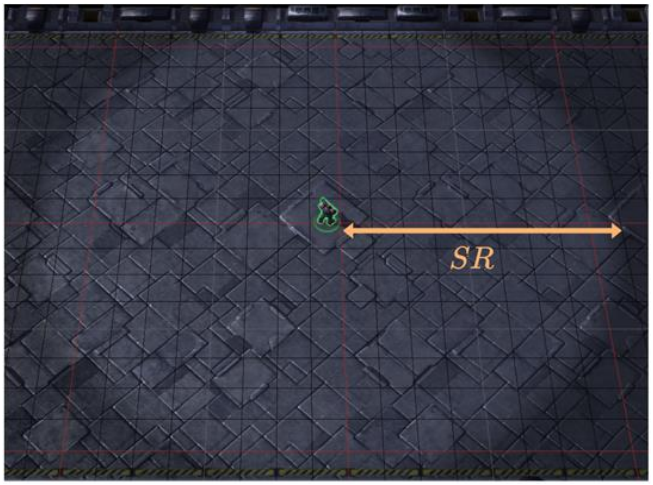
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ 실험에 사용한 환경 - SMAC

- 다중 에이전트 강화학습을 위한 가장 잘 알려진 벤치마크 환경
- 각 에이전트는 고유한 유닛을 제어하고, 목표는 적 유닛을 전투에서 물리치는 것
- 시야 범위는 각 유닛을 중심으로 한 가시 반경으로 정의됨
- 실험에서 사용한 맵: 5m_vs_6m, 8m_vs_9m, 10m_vs_11m, 3s_vs_5z, 3s5z, MMM2
- 시야 범위 집합은 세 가지 설정 사용: D = {3,6,**9**} / D = {3,6,9,12,**15**} / D = {3,6,9,12,15,18,**21**}
- LBF와 RWARE는 환경이 제공하는 글로벌 상태가 모든 정보 포함하지 않음 → 다른 환경과 동등한 설정을 위해 에이전트들의 관측값을 연결하여 글로벌 상태를 구성하는 변형된 SMAC 환경을 사용



Definition of Sight Range

StarCraft Multi-Agent Challenge (SMAC)

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **Experiment ① - 다양한 환경 설정에서 DSR의 효과는 어떻게 될까?**

- 다양한 환경 설정에서 DSR의 효과를 확인하기 위해 QMIX와 QMIX (w/ DSR) 성능 비교 진행
- 5개의 서로 다른 시드에 대한 평균 결과이며, LBF와 RWARE은 평균 테스트 리턴을 사용하고 SMAC에서는 평균 테스트 승률을 사용

		w/o DSR	w/ DSR (ours)
LBF	10x10-4p-2f-coop-10s	0.769 ± 0.384	**0.925 ± 0.064**
	10x10-4p-2f-coop-6s	**0.972 ± 0.037**	0.957 ± 0.040
	10x10-4p-2f-10s	0.988 ± 0.005	**0.993 ± 0.004**
	10x10-4p-2f-6s	**0.998 ± 0.004**	0.987 ± 0.010
	10x10-4p-4f-coop-10s	0.338 ± 0.339	**0.798 ± 0.050**
	10x10-4p-4f-coop-6s	0.277 ± 0.194	**0.772 ± 0.068**
RWARE	tiny-2ag-5s	1.486 ± 1.361	**4.762 ± 4.702**
	tiny-2ag-3s	5.846 ± 1.426	**11.900 ± 6.474**
	small-2ag-5s	**0.074 ± 0.148**	0.050 ± 0.100
	small-2ag-3s	**0.182 ± 0.359**	0.036 ± 0.072

해당 환경에서 사용한 기본 시야 범위

d=3s → {1,2,3}

SMAC	5m_vs_6m-9s	0.102 ± 0.028	**0.128 ± 0.066**
	5m_vs_6m-15s	0.080 ± 0.100	**0.140 ± 0.055**
	5m_vs_6m-21s	0.054 ± 0.079	**0.112 ± 0.053**
	8m_vs_9m-9s	0.452 ± 0.114	**0.508 ± 0.077**
	8m_vs_9m-15s	0.234 ± 0.287	**0.504 ± 0.087**
	8m_vs_9m-21s	0.120 ± 0.190	**0.472 ± 0.078**
	10m_vs_11m-9s	0.402 ± 0.220	**0.540 ± 0.061**
	10m_vs_11m-15s	0.122 ± 0.202	**0.546 ± 0.127**
	10m_vs_11m-21s	0.186 ± 0.262	**0.680 ± 0.066**
	3s_vs_5z-9s	**0.716 ± 0.143**	0.676 ± 0.065
	3s_vs_5z-15s	0.498 ± 0.409	**0.660 ± 0.081**
	3s_vs_5z-21s	0.292 ± 0.260	**0.712 ± 0.110**
	3s5z-9s	**0.854 ± 0.086**	**0.854 ± 0.051**
	3s5z-15s	**0.808 ± 0.094**	0.770 ± 0.075
	3s5z-21s	**0.784 ± 0.156**	0.736 ± 0.069
	MMM2-9s	0.690 ± 0.122	**0.736 ± 0.076**
	MMM2-15s	0.572 ± 0.113	**0.648 ± 0.094**
	MMM2-21s	0.190 ± 0.266	**0.714 ± 0.101**

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

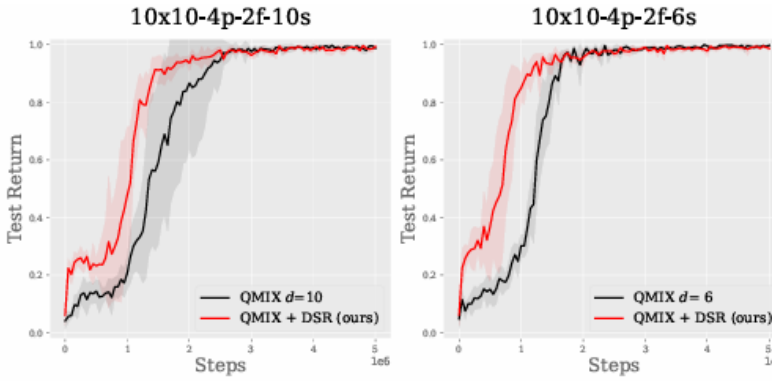
❖ **Experiment ① - 다양한 환경 설정에서 DSR의 효과는 어떻게 될까?**

- 다양한 환경 설정에서 DSR의 효과를 확인하기 위해 QMIX와 QMIX (w/ DSR) 성능 비교 진행
- 5개의 서로 다른 시드에 대한 평균 결과이며, LBF와 RWARE은 평균 테스트 리턴을 사용하고 SMAC에서는 평균 테스트 승률을 사용

		w/o DSR	w/ DSR (ours)
LBF	10x10-4p-2f-coop-10s	0.769 ± 0.384	**0.925 ± 0.064**
	10x10-4p-2f-coop-6s	**0.972 ± 0.037**	0.957 ± 0.040
	10x10-4p-2f-10s	0.988 ± 0.005	**0.993 ± 0.004**
	10x10-4p-2f-6s	**0.998 ± 0.004**	0.987 ± 0.010
	10x10-4p-4f-coop-10s	0.338 ± 0.339	**0.798 ± 0.050**
	10x10-4p-4f-coop-6s	0.277 ± 0.194	**0.772 ± 0.068**
RWARE	tiny-2ag-5s	1.486 ± 1.361	**4.762 ± 4.702**
	tiny-2ag-3s	5.846 ± 1.426	**11.900 ± 6.474**
	small-2ag-5s	**0.074 ± 0.148**	0.050 ± 0.100
	small-2ag-3s	**0.182 ± 0.359**	0.036 ± 0.072

해당 환경에서 사용한 기본 시야 범위

d=3s → {1,2,3}



Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **Experiment ① - 다양한 환경 설정에서 DSR의 효과는 어떻게 될까?**

- 다양한 환경 설정에서 DSR의 효과를 확인하기 위해 QMIX와 QMIX (w/ DSR) 성능 비교 진행
- 5개의 서로 다른 시드에 대한 평균 결과이며, LBF와 RWARE은 평균 테스트 리턴을 사용하고 SMAC에서는 평균 테스트 승률을 사용

		w/o DSR	w/ DSR (ours)
LBF	10x10-4p-2f-coop-10s	0.769 ± 0.384	**0.925 ± 0.064**
	10x10-4p-2f-coop-6s	**0.972 ± 0.037**	0.957 ± 0.040
	10x10-4p-2f-10s	0.988 ± 0.005	**0.993 ± 0.004**
	10x10-4p-2f-6s	**0.998 ± 0.004**	0.987 ± 0.010
	10x10-4p-4f-coop-10s	0.338 ± 0.339	**0.798 ± 0.050**
	10x10-4p-4f-coop-6s	0.277 ± 0.194	**0.772 ± 0.068**
RWARE	tiny-2ag-5s	1.486 ± 1.361	**4.762 ± 4.702**
	tiny-2ag-3s	5.846 ± 1.426	**11.900 ± 6.474**
	small-2ag-5s	**0.074 ± 0.148**	0.050 ± 0.100
	small-2ag-3s	**0.182 ± 0.359**	0.036 ± 0.072

10X11 (tiny) /10X20 (small)

SMAC	5m_vs_6m-9s	0.102 ± 0.028	**0.128 ± 0.066**
	5m_vs_6m-15s	0.080 ± 0.100	**0.140 ± 0.055**
	5m_vs_6m-21s	0.054 ± 0.079	**0.112 ± 0.053**
	8m_vs_9m-9s	0.452 ± 0.114	**0.508 ± 0.077**
	8m_vs_9m-15s	0.234 ± 0.287	**0.504 ± 0.087**
	8m_vs_9m-21s	0.120 ± 0.190	**0.472 ± 0.078**
	10m_vs_11m-9s	0.402 ± 0.220	**0.540 ± 0.061**
	10m_vs_11m-15s	0.122 ± 0.202	**0.546 ± 0.127**
	10m_vs_11m-21s	0.186 ± 0.262	**0.680 ± 0.066**
	3s_vs_5z-9s	**0.716 ± 0.143**	0.676 ± 0.065
	3s_vs_5z-15s	0.498 ± 0.409	**0.660 ± 0.081**
	3s_vs_5z-21s	0.292 ± 0.260	**0.712 ± 0.110**
	3s5z-9s	**0.854 ± 0.086**	**0.854 ± 0.051**
	3s5z-15s	**0.808 ± 0.094**	0.770 ± 0.075
	3s5z-21s	**0.784 ± 0.156**	0.736 ± 0.069
	MMM2-9s	0.690 ± 0.122	**0.736 ± 0.076**
	MMM2-15s	0.572 ± 0.113	**0.648 ± 0.094**
	MMM2-21s	0.190 ± 0.266	**0.714 ± 0.101**

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **Experiment ① - 다양한 환경 설정에서 DSR의 효과는 어떻게 될까?**

- 다양한 환경 설정에서 DSR의 효과를 확인하기 위해 QMIX와 QMIX (w/ DSR) 성능 비교 진행
- 5개의 서로 다른 시드에 대한 평균 결과이며, LBF와 RWARE은 평균 테스트 리턴을 사용하고 SMAC에서는 평균 테스트 승률을 사용

		w/o DSR	w/ DSR (ours)
LBF	10x10-4p-2f-coop-10s	0.769 ± 0.384	**0.925 ± 0.064**
	10x10-4p-2f-coop-6s	**0.972 ± 0.037**	0.957 ± 0.040
	10x10-4p-2f-10s	0.988 ± 0.005	**0.993 ± 0.004**
	10x10-4p-2f-6s	**0.998 ± 0.004**	0.987 ± 0.010
	10x10-4p-4f-coop-10s	0.338 ± 0.339	**0.798 ± 0.050**
	10x10-4p-4f-coop-6s	0.277 ± 0.194	**0.772 ± 0.068**
RWARE	tiny-2ag-5s	1.486 ± 1.361	**4.762 ± 4.702**
	tiny-2ag-3s	5.846 ± 1.426	**11.900 ± 6.474**
	small-2ag-5s	**0.074 ± 0.148**	0.050 ± 0.100
	small-2ag-3s	**0.182 ± 0.359**	0.036 ± 0.072

SMAC	5m_vs_6m-9s	0.102 ± 0.028	**0.128 ± 0.066**
	5m_vs_6m-15s	0.080 ± 0.100	**0.140 ± 0.055**
	5m_vs_6m-21s	0.054 ± 0.079	**0.112 ± 0.053**
	8m_vs_9m-9s	0.452 ± 0.114	**0.508 ± 0.077**
	8m_vs_9m-15s	0.234 ± 0.287	**0.504 ± 0.087**
	8m_vs_9m-21s	0.120 ± 0.190	**0.472 ± 0.078**
	10m_vs_11m-9s	0.402 ± 0.220	**0.540 ± 0.061**
	10m_vs_11m-15s	0.122 ± 0.202	**0.546 ± 0.127**
	10m_vs_11m-21s	0.186 ± 0.262	**0.680 ± 0.066**
	3s_vs_5z-9s	**0.716 ± 0.143**	0.676 ± 0.065
	3s_vs_5z-15s	0.498 ± 0.409	**0.660 ± 0.081**
	3s_vs_5z-21s	0.292 ± 0.260	**0.712 ± 0.110**
	3s5z-9s	**0.854 ± 0.086**	**0.854 ± 0.051**
	3s5z-15s	**0.808 ± 0.094**	0.770 ± 0.075
	3s5z-21s	**0.784 ± 0.156**	0.736 ± 0.069
	MMM2-9s	0.690 ± 0.122	**0.736 ± 0.076**
	MMM2-15s	0.572 ± 0.113	**0.648 ± 0.094**
	MMM2-21s	0.190 ± 0.266	**0.714 ± 0.101**

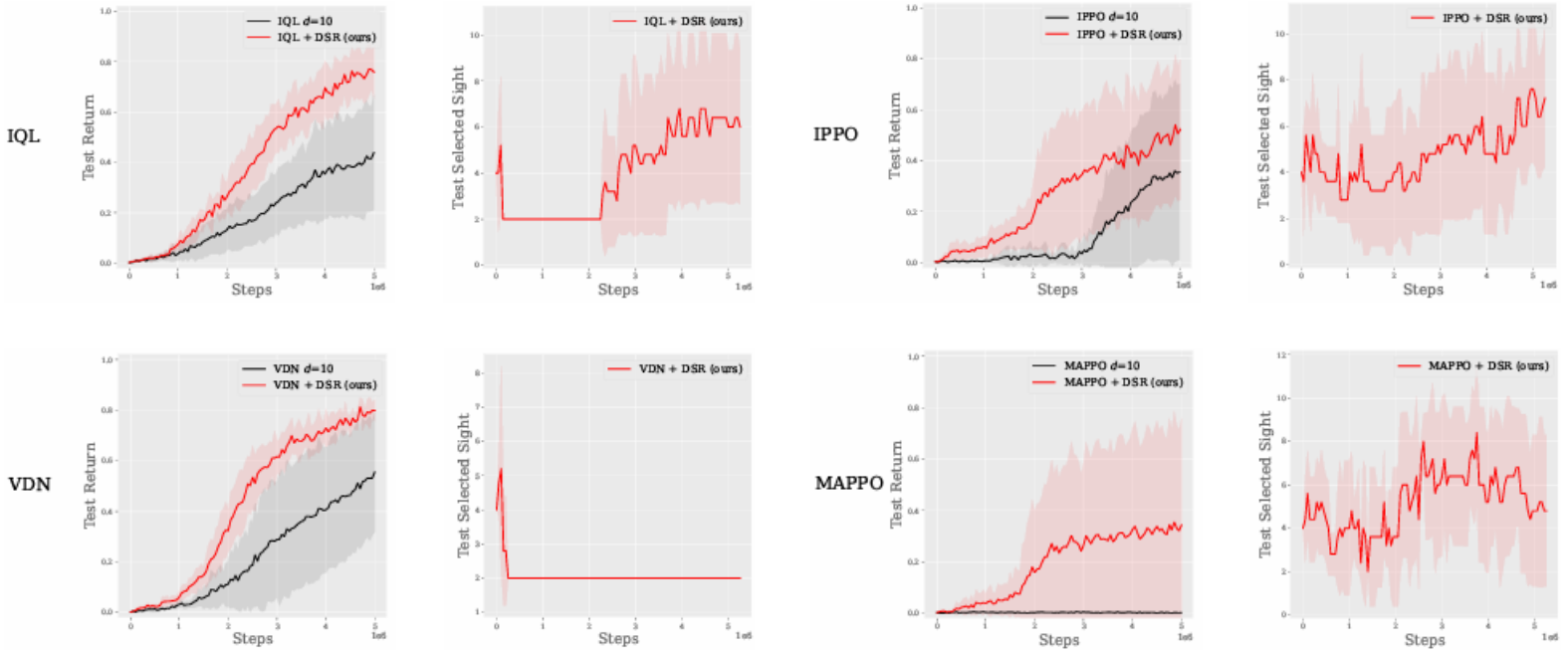
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **Experiment ② - 다양한 MARL 알고리즘에서의 DSR의 적용 효과는?**

- IQL, VDN IPPO, MAPPO에 대한 LBF환경(10X10-4p-4f-coop-10s)에서 일반성을 검증하는 실험을 수행
- 전반적으로, 시야 범위 딜레마는 모든 알고리즘에서 여전히 나타남
- **MARL 방법론을 수정 없이도 쉽게 통합이 가능하며, 학습 성능 향상할 수 있다는 점을 발견함**



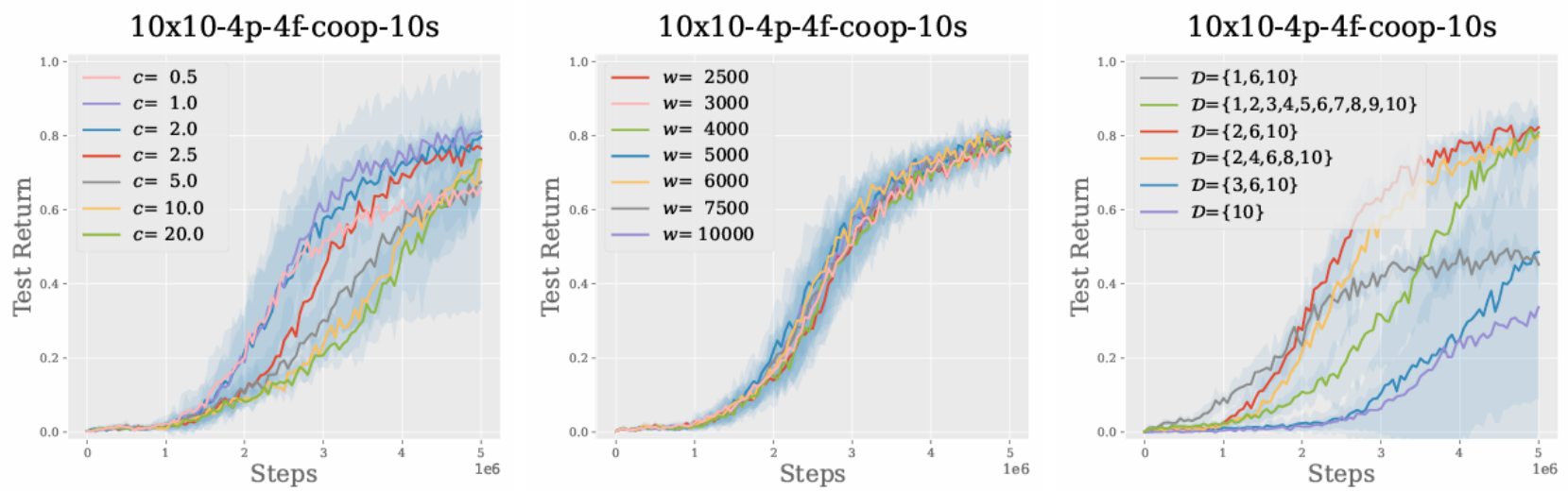
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **Experiment ③ - DSR에서 사용하는 다양한 하이퍼 파라미터의 효과는?**

- 탐색 상수 c, 슬라이딩 윈도우 크기 w, 그리고 meta-controller가 선택할 수 있는 다양한 시야 범위 조합에 대해 확인
- 이러한 결과를 통해, 적절한 시야 범위 집합 D를 설계하는 것이 중요함을 확인



(a) Exploration coefficient c. **(b) Sliding window size w.** **(c) Different set of sight range $\mathcal{D}$.**

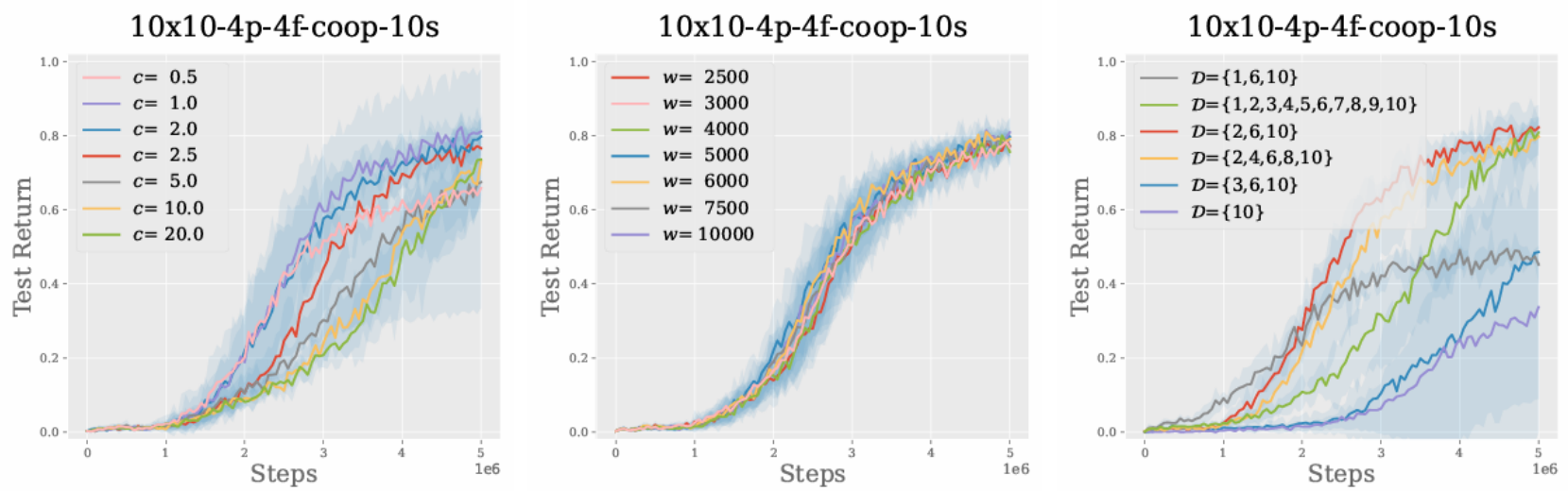
Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Methods

Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning (2025, AAMAS)

❖ **Experiment ③ - DSR에서 사용하는 다양한 하이퍼 파라미터의 효과는?**

- 탐색 상수 c, 슬라이딩 윈도우 크기 w, 그리고 meta-controller가 선택할 수 있는 다양한 시야 범위 조합에 대해 확인
- 이러한 결과를 통해, 적절한 시야 범위 집합 D를 설계하는 것이 중요함을 확인



(a) Exploration coefficient c. **(b) Sliding window size w.** **(c) Different set of sight range $\mathcal{D}$.**

Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Conclusion

- MASIA (2022, NeurIPS)
  - 에이전트 간의 정보를 집계하고, 그 중 의미 있는 정보만 추출함으로써 시야기 제한된 상황에서도 안정적인 협력이 가능하게 함
- DSR (2025, AAMAS)
  - 환경과 학습 단계에 따라 에이전트별 최적 시야 범위를 동적으로 선택하도록 설계되어, 시야가 너무 좁거나 넓을 때 발생하는 학습 비효율성을 완화함

Reference

1. Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55(2), 895-943.
2. Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly de composable value functions via communication minimization. In International Conference on Learning Representations, 2020.
3. Guan, C., Chen, F., Yuan, L., Wang, C., Yin, H., Zhang, Z., & Yu, Y. (2022). Efficient multi-agent communication via self-supervised information aggregation. Advances in Neural Information Processing Systems, 35, 1020-1033.
4. Liao, W. C., Wu, T. R., & Wu, I. (2025). Dynamic Sight Range Selection in Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2505.12811.

Thank you