NEURAL INFORMATION PROCESSING SYSTEMS

Multi-Agent Reinforcement Learning is A Sequence Modeling Problem

Muning Wen[1,2], Jakub Grudzien Kuba[3], Runji Lin[4], Weinan Zhang[1], Ying Wen[1], Jun Wang[2,5], Yaodong Yang[6]
[1]Shanghai Jiao Tong University, [2]Digital Brain Lab,
[3]University of Oxford, [4]Institute of Automation, Chinese Academy of Science,
[5]University College London, [6]Institute for AI, Peking University

Prequels:

- Settling the Variance of Multi-Agent Policy Gradient (NeurIPS 2021).
- Trust Region Policy Optimisation in Multi-Agent Reinforcement Learning (ICLR 2022).

Motivation of Multi-Agent Transformer

NEURAL INFORMATION PROCESSING SYSTEMS

- How to abstract multi-agent decision making as a sequence modeling problem and leverage the prosperous development of the sequence models

- Building the bridge between MARL and sequence models so that the modeling power of modern sequence models, e.g. the Transformer, can be unleashed for MARL.

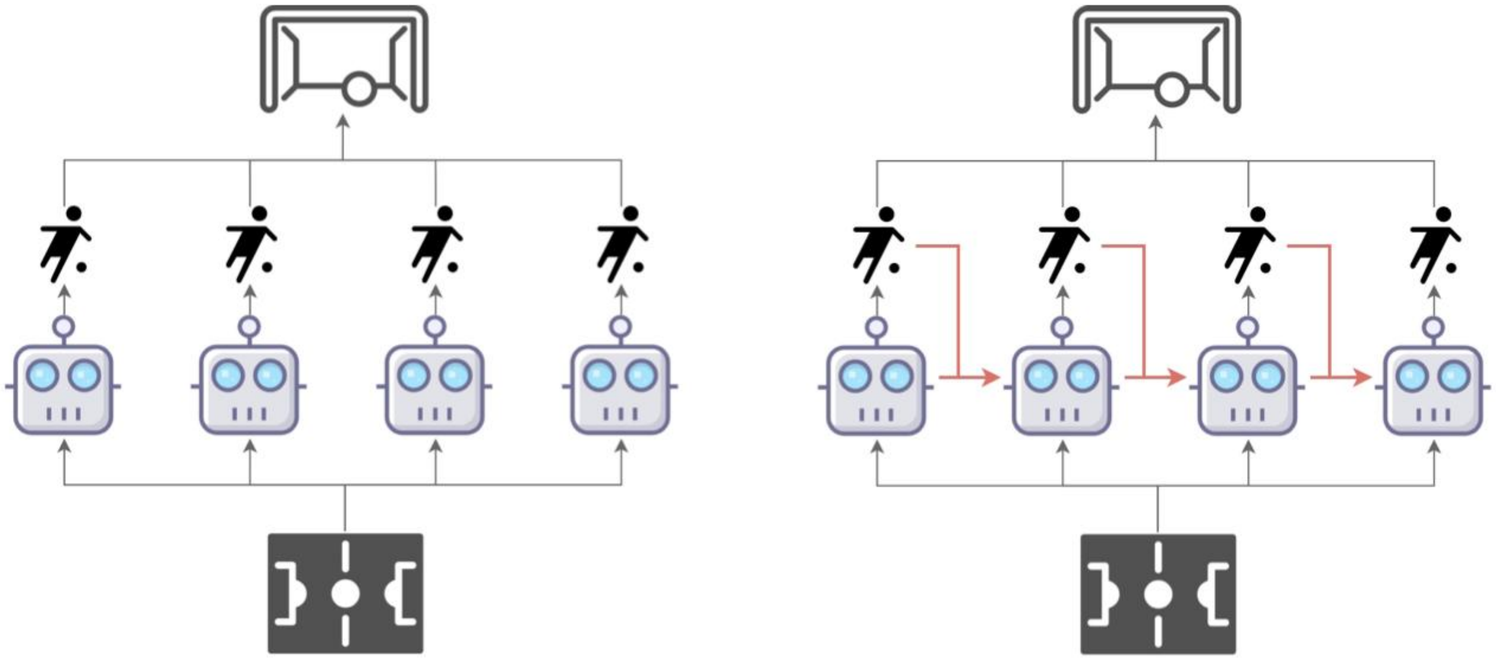
- Guarantee the monotonic performance improvement of joint policy with acceptable sample efficiency and training speed.

- More information could be found in our project website (https://sites.google.com/view/multi-agent-transformer) or github(https://github.com/PKU-MARL/Multi-Agent-Transformer).

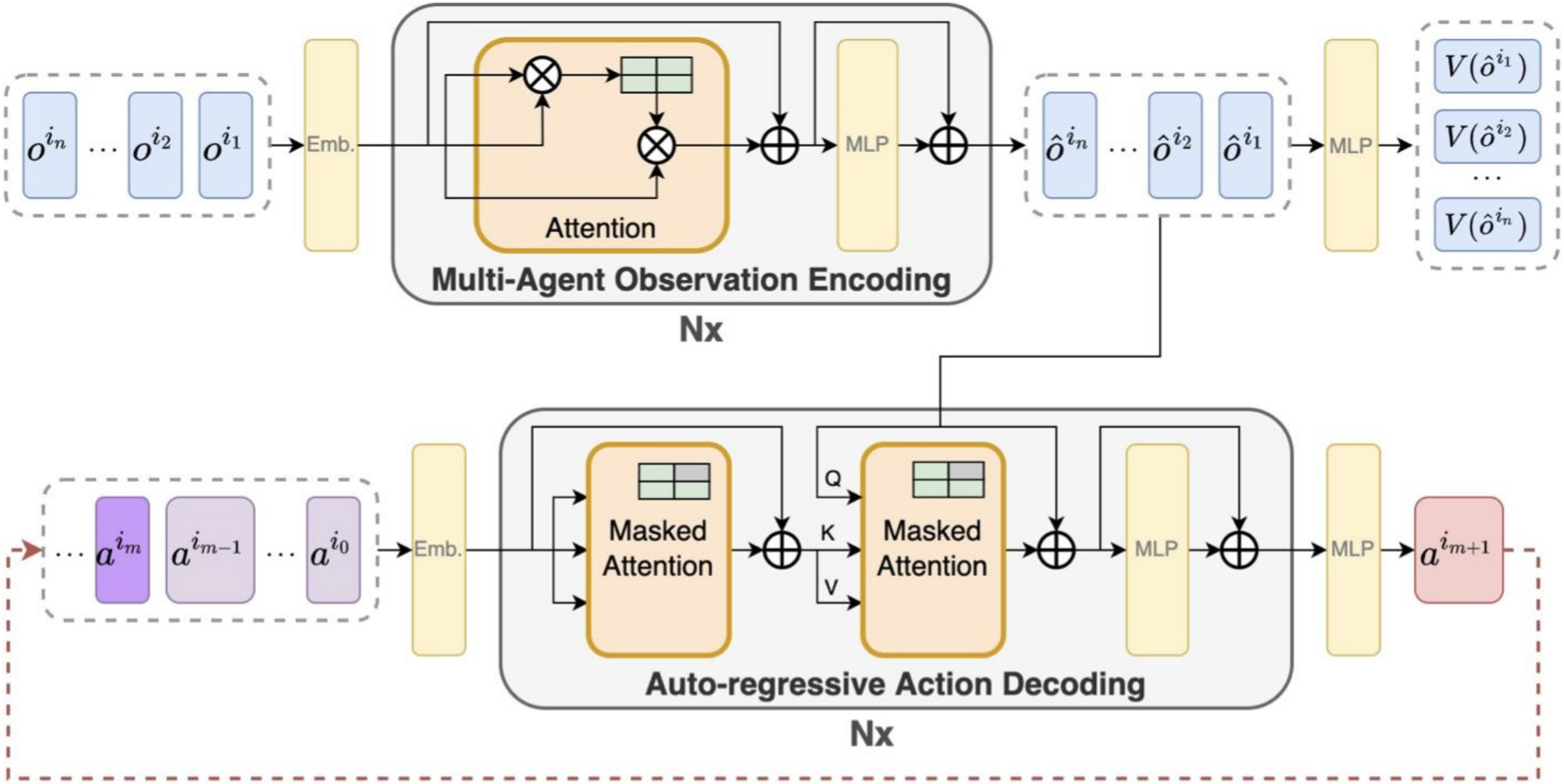
The Multi-Agent Sequential Decision Paradigm

Theorem 1 (Multi-Agent Advantage Decomposition [13]). *Let $i_{1:n}$ be a permutation of agents. Then, for any joint observation $\boldsymbol{o} \in \boldsymbol{\mathcal{O}}$ and joint action $\boldsymbol{a} = \boldsymbol{a}^{i_{1:n}} \in \boldsymbol{\mathcal{A}}$, the following equation always holds with no further assumption needed,*

$$A_{\boldsymbol{\pi}}^{i_{1:n}}\left(\boldsymbol{o}, \boldsymbol{a}^{i_{1:n}}\right) = \sum_{m=1}^{n} A_{\boldsymbol{\pi}}^{i_m}\left(\boldsymbol{o}, \boldsymbol{a}^{i_{1:m-1}}, a^{i_m}\right).$$



Multi-Agent Transformer



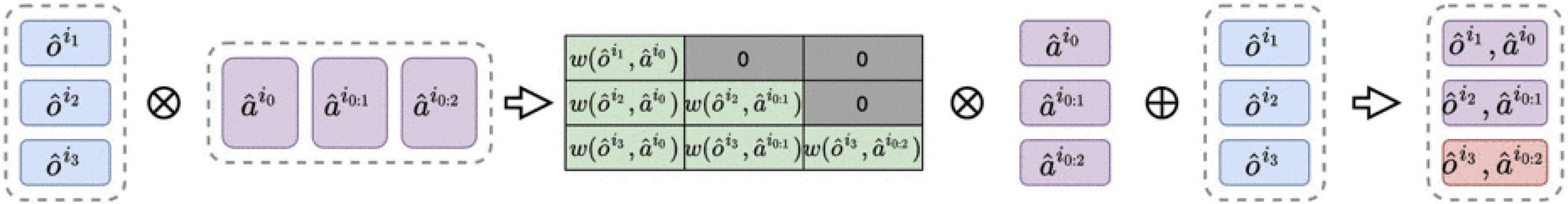
Multi-Agent Transformer

$$L_{\text{Encoder}}(\phi) = \frac{1}{Tn} \sum_{m=1}^{n} \sum_{t=0}^{T-1} \Big[R(\mathbf{o}_t, \mathbf{a}_t) + \gamma V_{\bar{\phi}}(\hat{\mathbf{o}}_{t+1}^{i_m}) - V_{\phi}(\hat{\mathbf{o}}_t^{i_m}) \Big]^2, \quad (4)$$

$$L_{\text{Decoder}}(\theta) = -\frac{1}{Tn} \sum_{m=1}^{n} \sum_{t=0}^{T-1} \min \Big(\mathrm{r}_t^{i_m}(\theta) \hat{A}_t, \mathrm{clip}(\mathrm{r}_t^{i_m}(\theta), 1 \pm \epsilon) \hat{A}_t \Big), \quad (5)$$

$$\mathrm{r}_t^{i_m}(\theta) = \frac{\pi_{\theta}^{i_m}(\mathrm{a}_t^{i_m} | \hat{\mathbf{o}}_t^{i_{1:n}}, \hat{\mathbf{a}}_t^{i_{1:m-1}})}{\pi_{\theta_{\text{old}}}^{i_m}(\mathrm{a}_t^{i_m} | \hat{\mathbf{o}}_t^{i_{1:n}}, \hat{\mathbf{a}}_t^{i_{1:m-1}})},$$

An example of how the attention work in decoder.



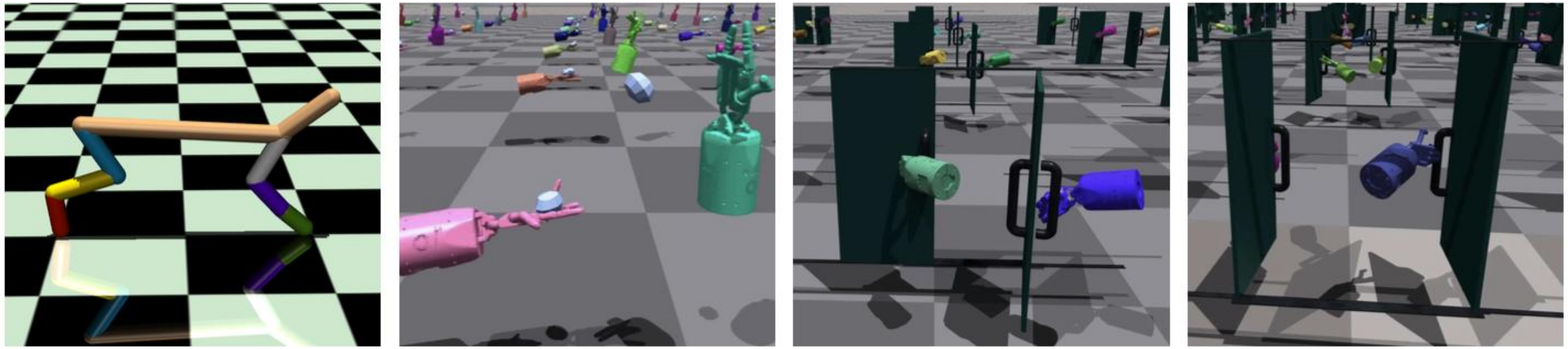
Single-task Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Table 1: Performance evaluations of win rate and standard deviation on the SMAC benchmark.

Task	Difficulty	MAT	MAT-Dec	MAPPO	HAPPO	QMIX	Steps
1c3s5z	Easy	**100.0**(2.4)	**100.0**(0.4)	**100.0**(2.2)	97.5(1.8)	96.9(1.5)	2e6
MMM	Easy	**100.0**(2.2)	98.1(2.1)	95.6(4.5)	81.2(22.9)	91.2(3.2)	2e6
2c vs 64zg	Hard	**100.0**(1.3)	95.9(2.3)	**100.0**(2.7)	90.0(4.8)	90.3(4.0)	5e6
3s vs 5z	Hard	**100.0**(1.7)	**100.0**(1.3)	**100.0**(2.5)	91.9(5.3)	92.3(4.4)	5e6
3s5z	Hard	**100.0**(1.9)	**100.0**(3.3)	72.5(26.5)	90.0(3.5)	84.3(5.4)	3e6
5m vs 6m	Hard	**90.6**(4.4)	83.1(4.6)	88.2(6.2)	73.8(4.4)	75.8(3.7)	5e6
8m vs 9m	Hard	**100.0**(3.1)	95.0(4.6)	93.8(3.5)	86.2(4.4)	92.6(4.0)	5e6
10m vs 11m	Hard	**100.0**(1.4)	**100.0**(2.0)	96.3(5.8)	77.5(9.7)	95.8(6.1)	5e6
25m	Hard	**100.0**(1.3)	86.9(5.6)	**100.0**(2.7)	0.6(0.8)	90.2(9.8)	2e6
27m vs 30m	Super Hard	**100.0**(0.7)	95.3(2.2)	93.1(3.2)	0.0(0.0)	39.2(8.8)	1e7
MMM2	Super Hard	**93.8**(2.6)	91.2(5.3)	81.8(10.1)	0.3(0.4)	88.3(2.4)	1e7
6h vs 8z	Super Hard	**98.8**(1.3)	93.8(4.7)	88.4(5.7)	0.0(0.0)	9.7(3.1)	1e7
3s5z vs 3s6z	Super Hard	**96.5**(1.3)	85.3(7.5)	84.3(19.4)	82.8(21.2)	68.8(21.2)	2e7

Single-task Experiments



(a) HalfCheetah (b) CatchOver2Underarm (c) DoorOpenInward (d) DoorCloseOutward

Figure 3: Demonstrations of the Bi-DexHands and the HalfCheetah environments.

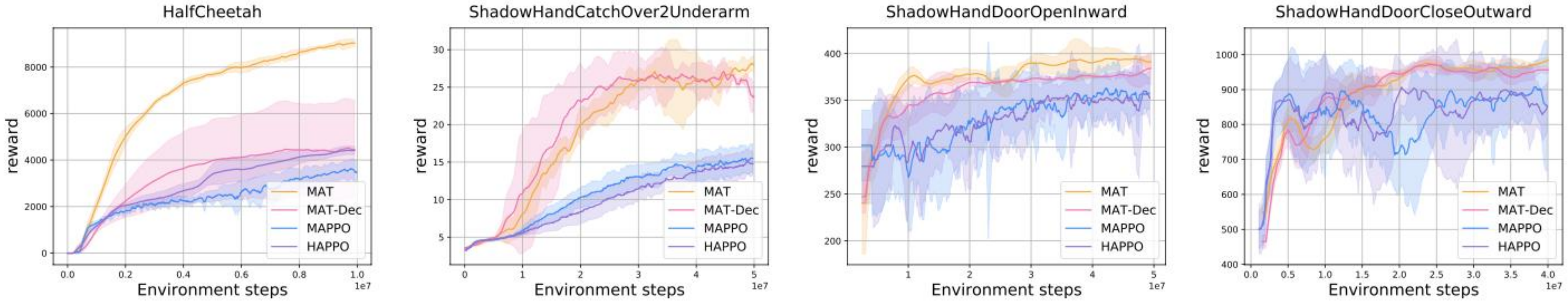


Figure 4: Performance comparisons on the Multi-Agent MuJoCo and the Bi-DexHands benchmarks.

Thank you for your attention!