PIC: Permutation-Invariant Critic for Multi-Agent Deep Reinforcement Learning

Iou-Jen Liu*, Raymond A. Yeh*, Alexander G. Schwing
University of Illinois Urbana-Champaign

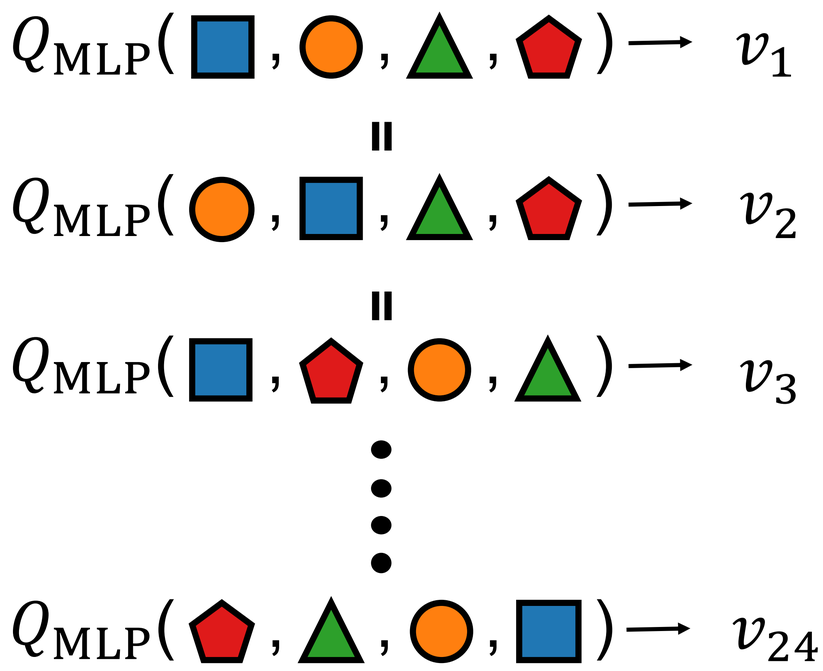
1. Introduction

Background

- Multi-agent training suffers from non-stationary environment
- Current frameworks use a centralized MLP critic to address this non-stationary environment issue

Problem

- In an environment with N homogeneous agents, permuting agent positions results in an identical environment
- However, MLP critic doesn't assign the same value to the **N!** permuted inputs
- MLP critic needs to learn **permutation invariance** from data
- Consequence: low **sample efficiency.** A lot of data is needed to learn permutation invariance, particularly when the number of agents is large



2. Approach

Our Work: Permutation-Invariant Critic

- We map all N! input permutations to the same value by default, i.e., without seeing any data
- Achieves significantly better data efficiency and scalability

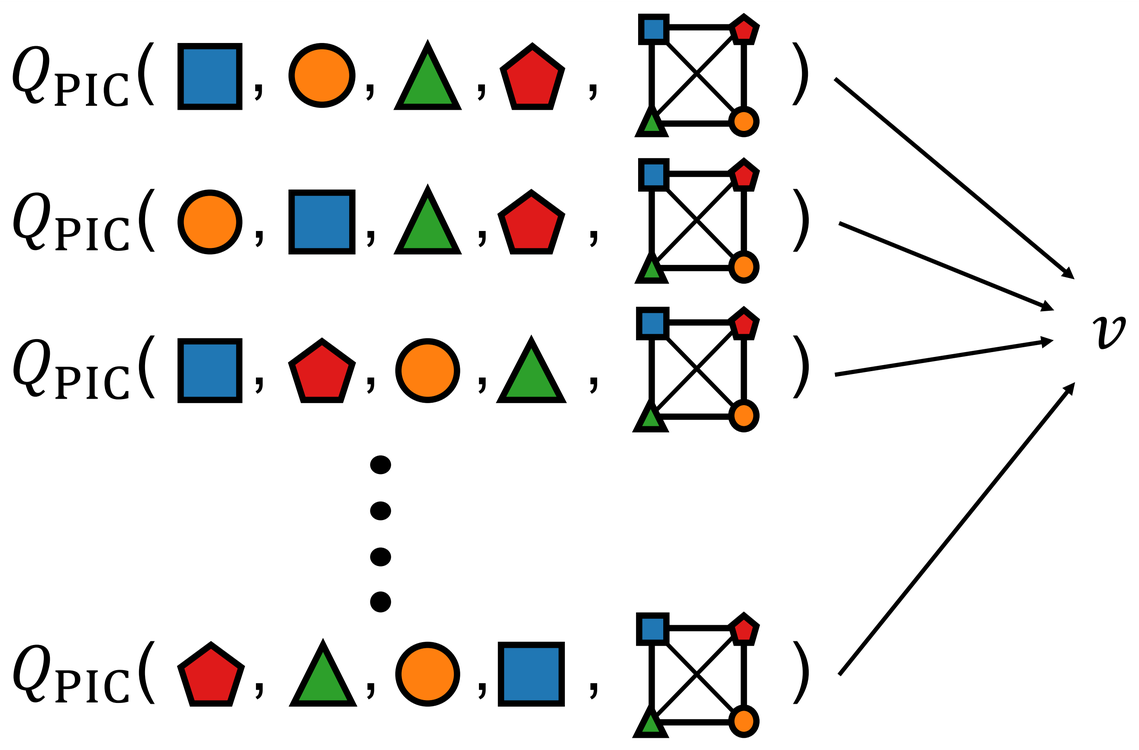
Approach

- Represent agents via a graph convolution neural network

$$Q_{\text{PIC}}(z^t) \coloneqq f_v \circ f_{\max} \circ f_{\text{GCN}}^{(L)} \circ \cdots \circ f_{\text{GCN}}^{(1)}(z^t)$$

Notation

- f_v: fully connected layer which outputs the Q value
- f_{max}: max pooling layer (permutation invariant)
- $f_{\text{GCN}}^{(l)}$: l-th graph convolution layer (permutation invariant)
- z^t: agents' actions and observations at time t



3. Scalability

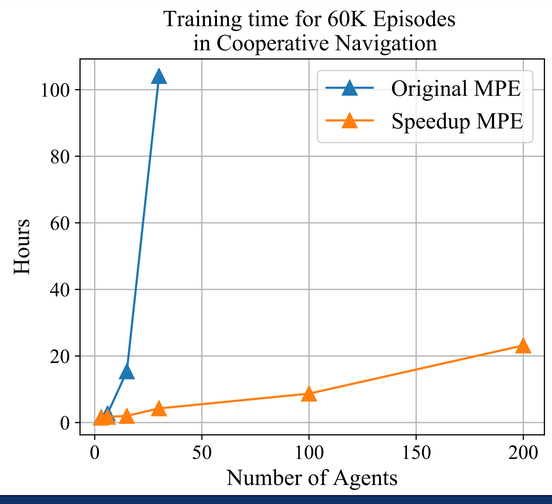
Improved Multiple-Particle Environment (MPE)

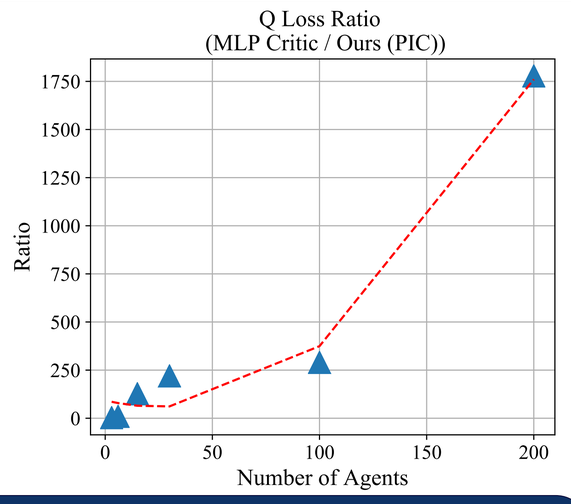
Training 200 agents:

- Original MPE: > 1 week
- Our improved MPE: 20 hours

Q-value Loss

- PIC better approximates the Q-value function than MLP critic
- Improvement increases with the number of agents





4. Experimental Results (see paper for additional results)

Comparison with MLP Critic

- PIC outperforms MLP critic with/without data augmentation
- PIC has better sample efficiency
- PIC scales to N = 200 agents

	Cooperative Push				Heterogeneous Navigation			
	N=3	N=6	N=15	N=30	N=8	N=16	N=30	N=100
MLP	-0.171	-0.561	-2.538	-3.499	-0.398	-3.410	-12.94	-121.5
MLP+Data Aug.	-0.171	-0.672	-2.645	-3.761	-0.683	-3.479	-12.81	-130.0
Our PIC	-0.155	-0.401	-2.231	-3.117	-0.398	-1.825	-6.293	-94.32

	Cooperative Navigation				Prey & Predator			
	N=15	N=30	N=100	N=200	N=6	N=15	N=30	N=100
MLP	-6.489	-20.72	-128.1	-502.3	0.026	3.982	-0.377	19.89
MLP+Data Aug.	-6.280	-21.20	-128.0	-509.9	-0.024	4.198	-0.093	28.74
Our PIC	-1.999	-11.36	-71.49	-436.2	0.176	10.13	6.662	99.39

(x1000)