Мультиагентное обучение с подкреплением с непрерывными действиями

Пшеницын А.С.
pshenitsyn.as@phystech.edu

Центр когнитивного моделирования МФТИ

2024 – Обзор исследований





Введение

Цель данного обзора - проанализировать современные подходы к мультиагентному обучению с подкреплением, включая методы координации и планирования пути. Рассмотреть как теоретические аспекты, так и практические применения.

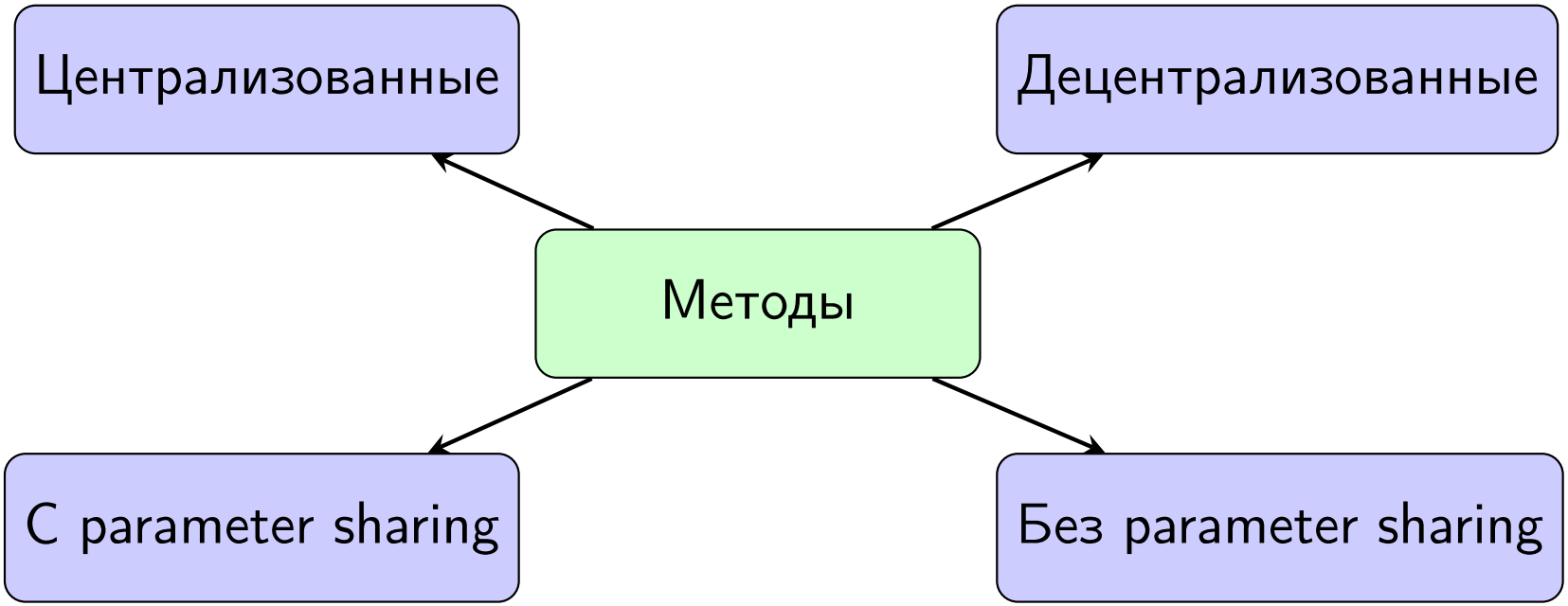
Постановка задачи

Мультиагентный процесс принятия решений задается как $M = \langle S, A, I, P, \vec{r}, \gamma \rangle$, где:

- S - пространство состояний,
- A - пространство действий для каждого агента,
- $I = \{1, \ldots, n\}$ - агенты,

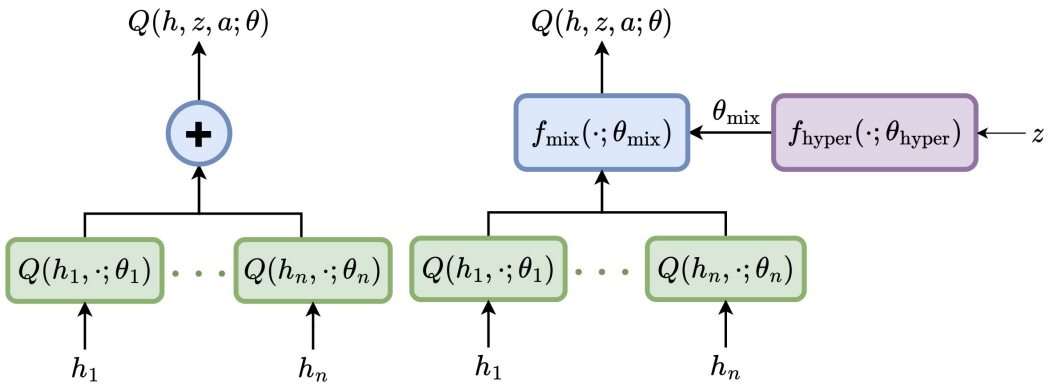
  В каждый момент времени агент $i \in I$ совершает действие $a^i \in A$. Совокупность действий всех агентов образует объединенное действие $\vec{a} \in A^n$
- $P = P(s'|s, \vec{a})$ - динамика среды, функция вероятности переходов между состояниями,
- $\vec{r} = \vec{r}(s, \vec{a})$ - функция награды для каждого агента, возможно, из одинаковых элементов
- γ - фактор дисконтирования для дисконтированной кумулятивной награды $\vec{R} = \sum_{t \geq 0} \gamma^t \vec{r}(s_t, \vec{a_t})$

Подходы к моделированию мультиагентных систем

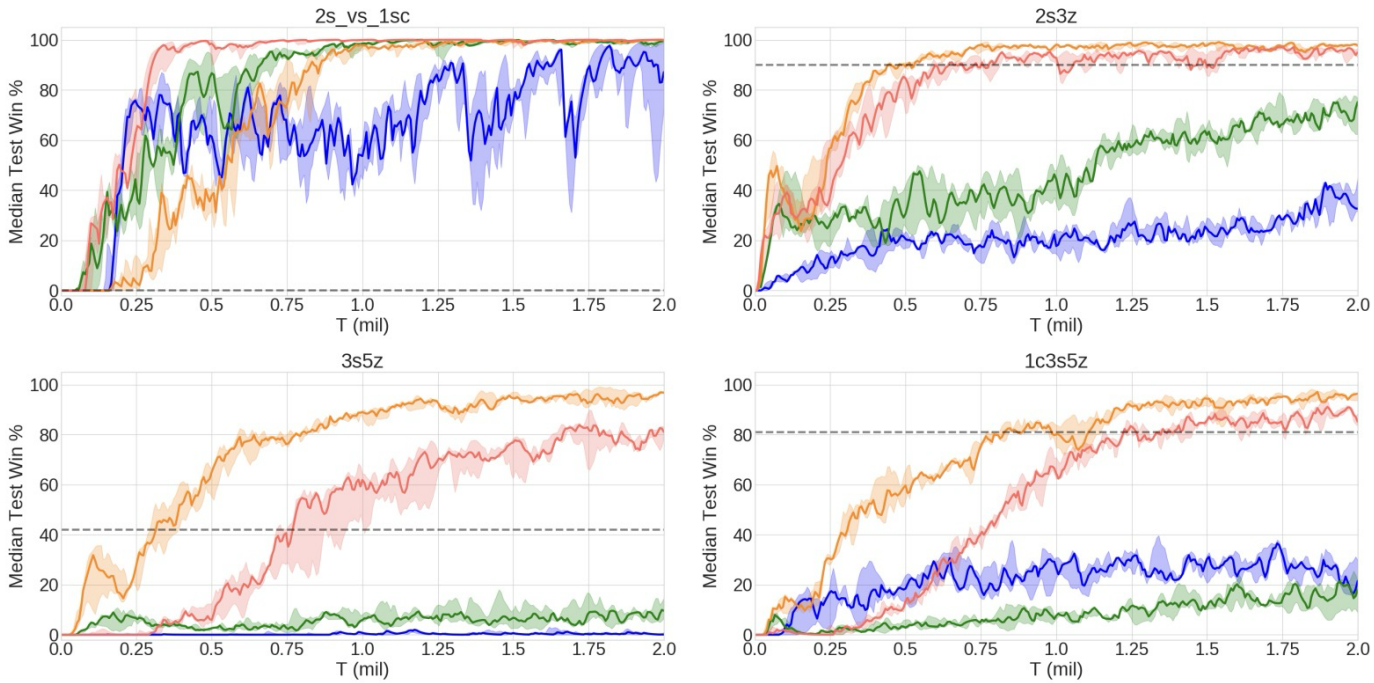


Централизованное обучение и децентрализованное исполнение

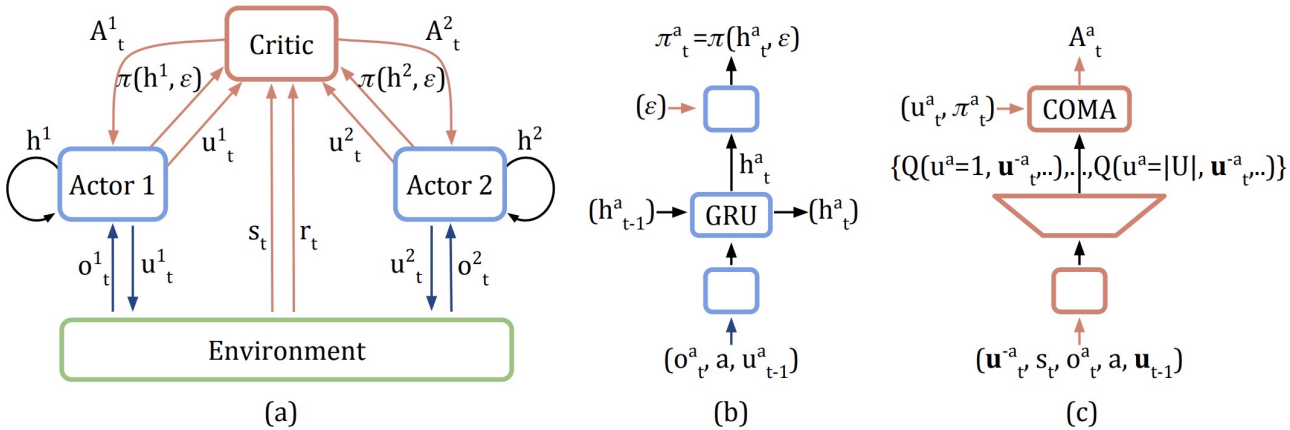
Методы, основанные на функции ценности

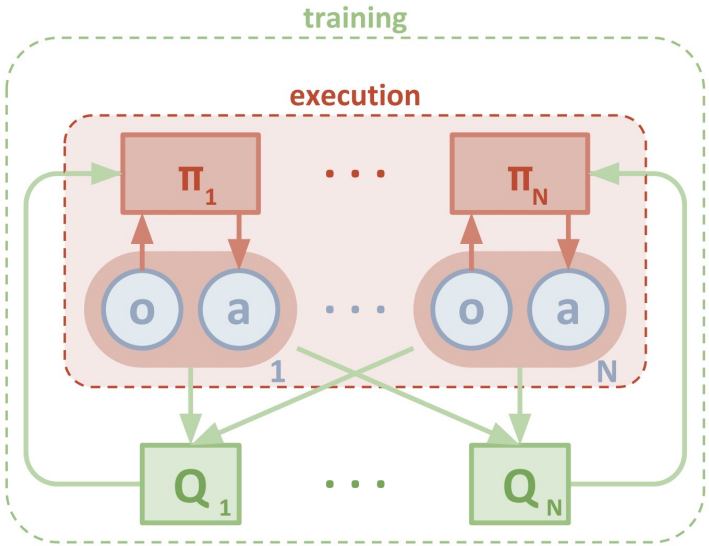


(a) VDN (b) QMIX

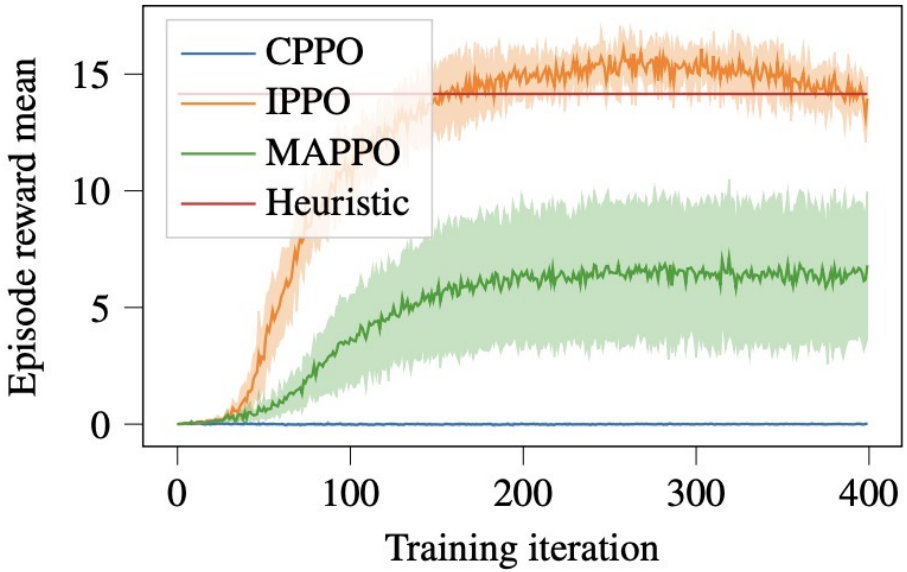


Методы, основанные на стратегии

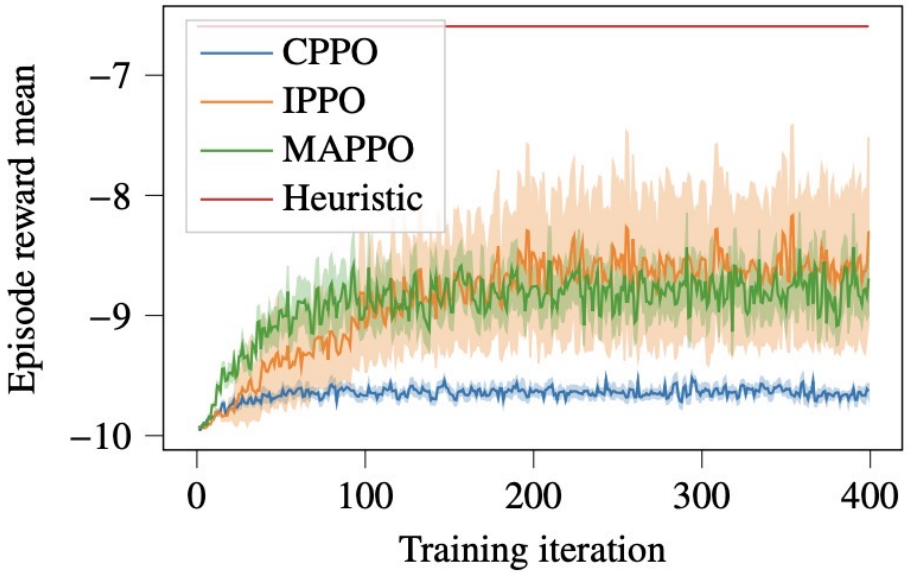




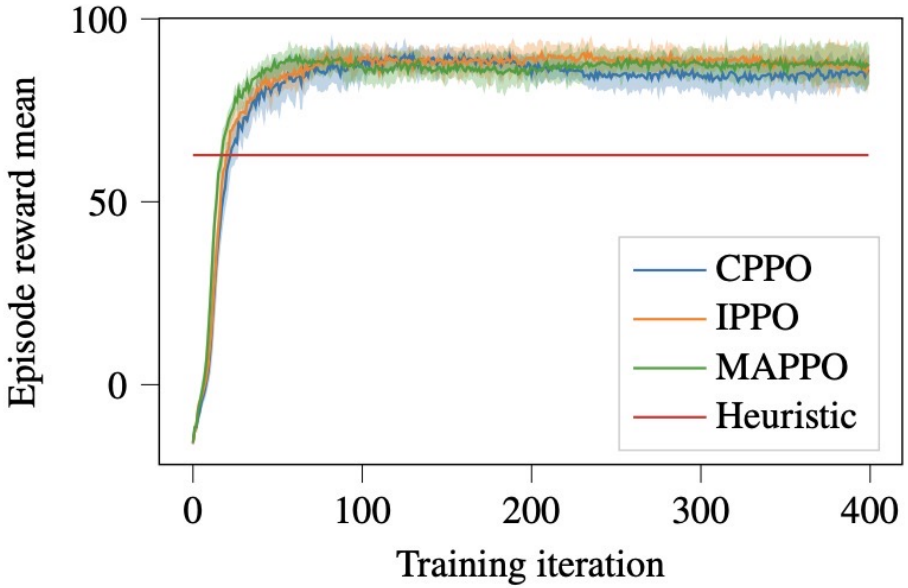
Мультиагентные модификации PPO

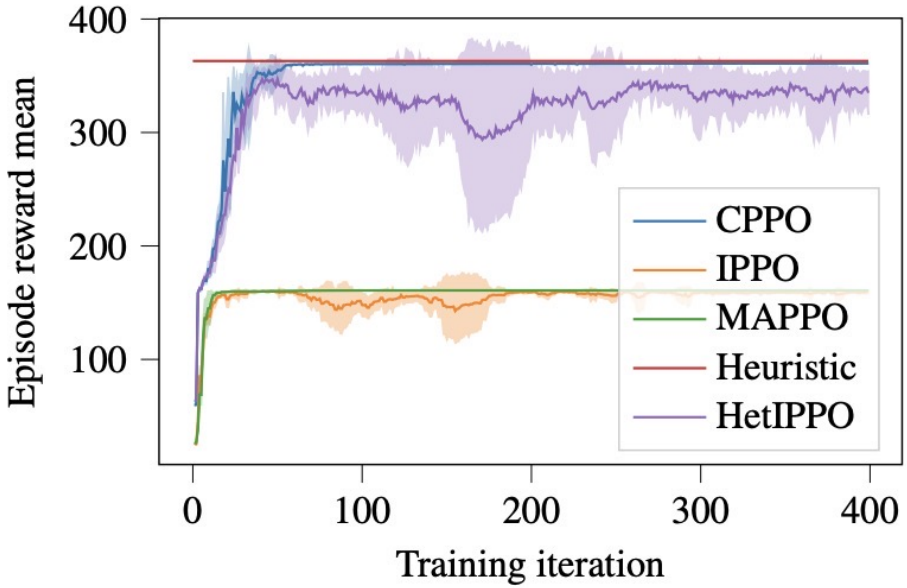


(a) Transport

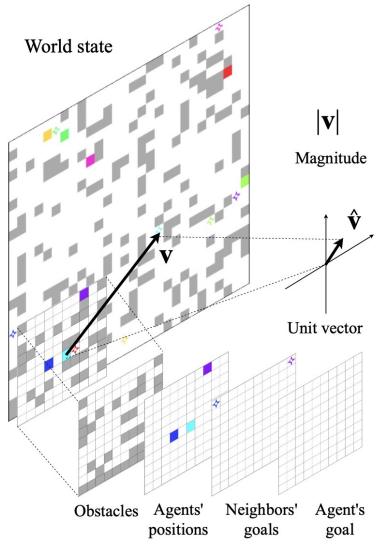


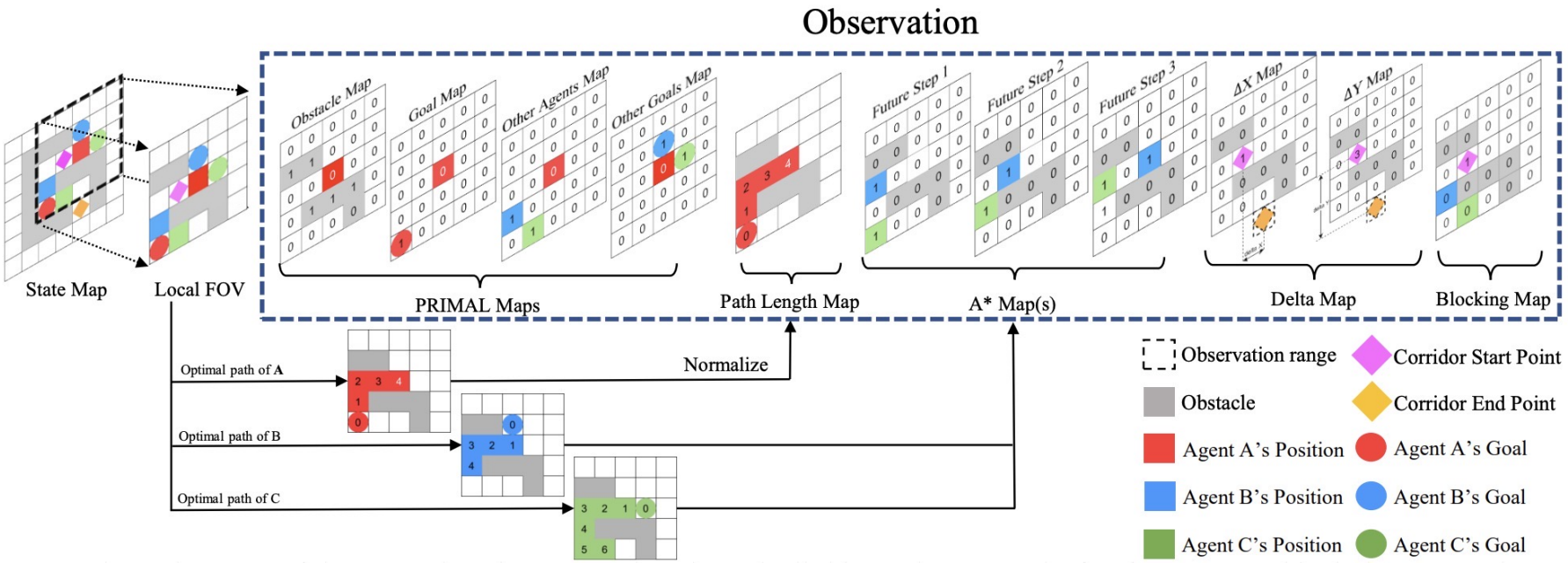
(b) Wheel



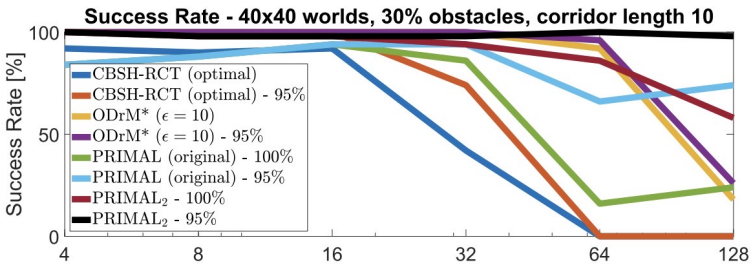


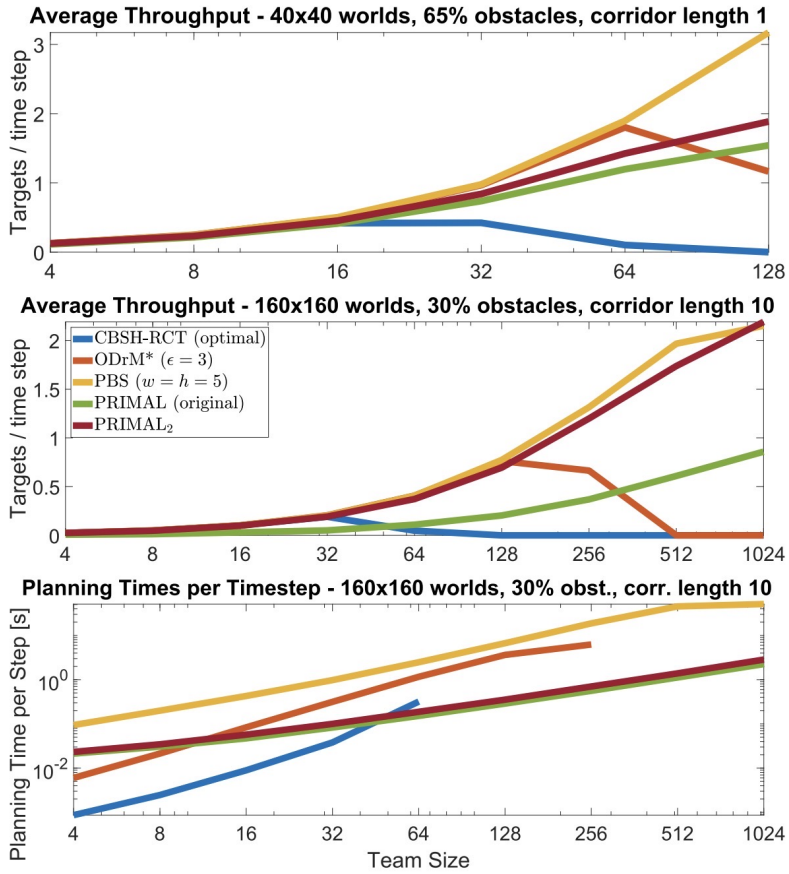
PRIMAL и PRIMAL$_2$



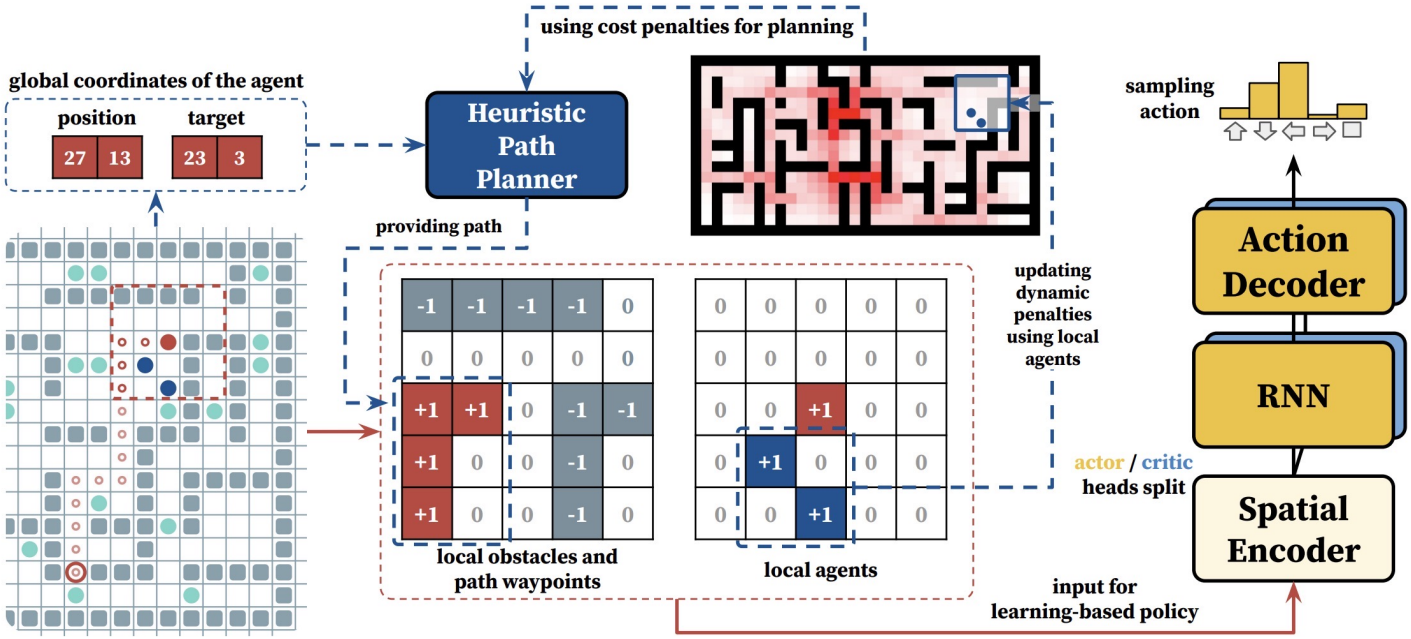


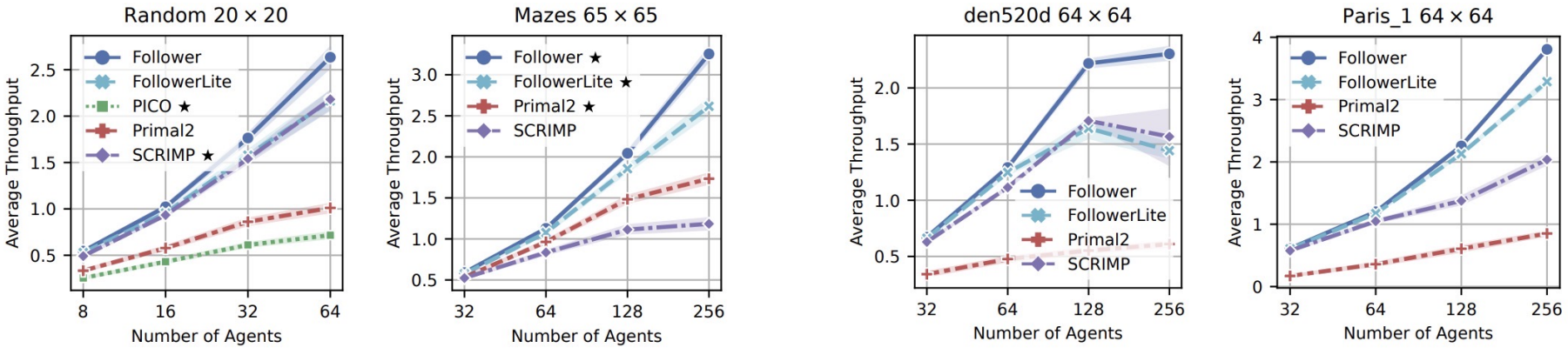
PRIMAL и PRIMAL$_2$



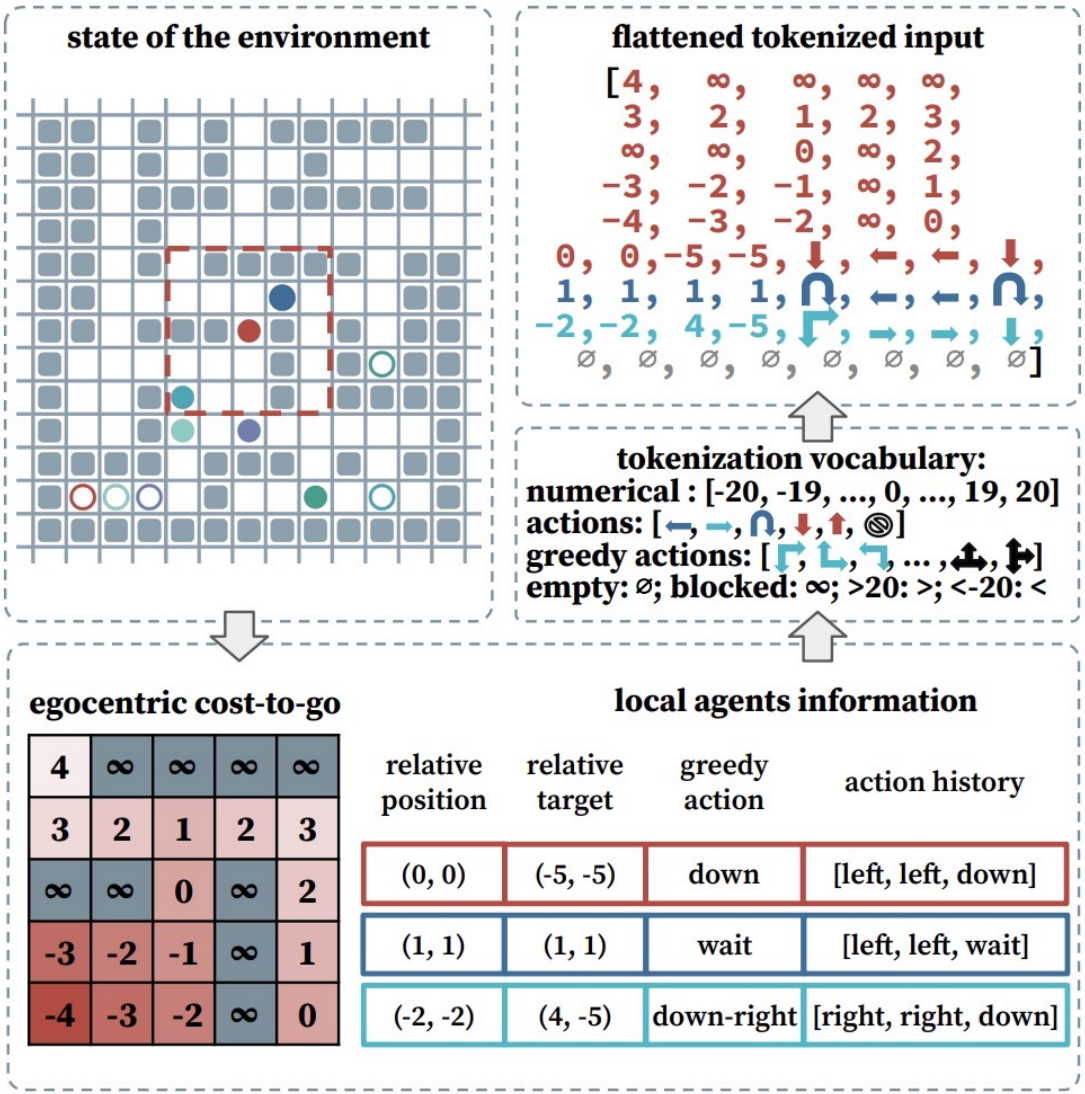


Follower

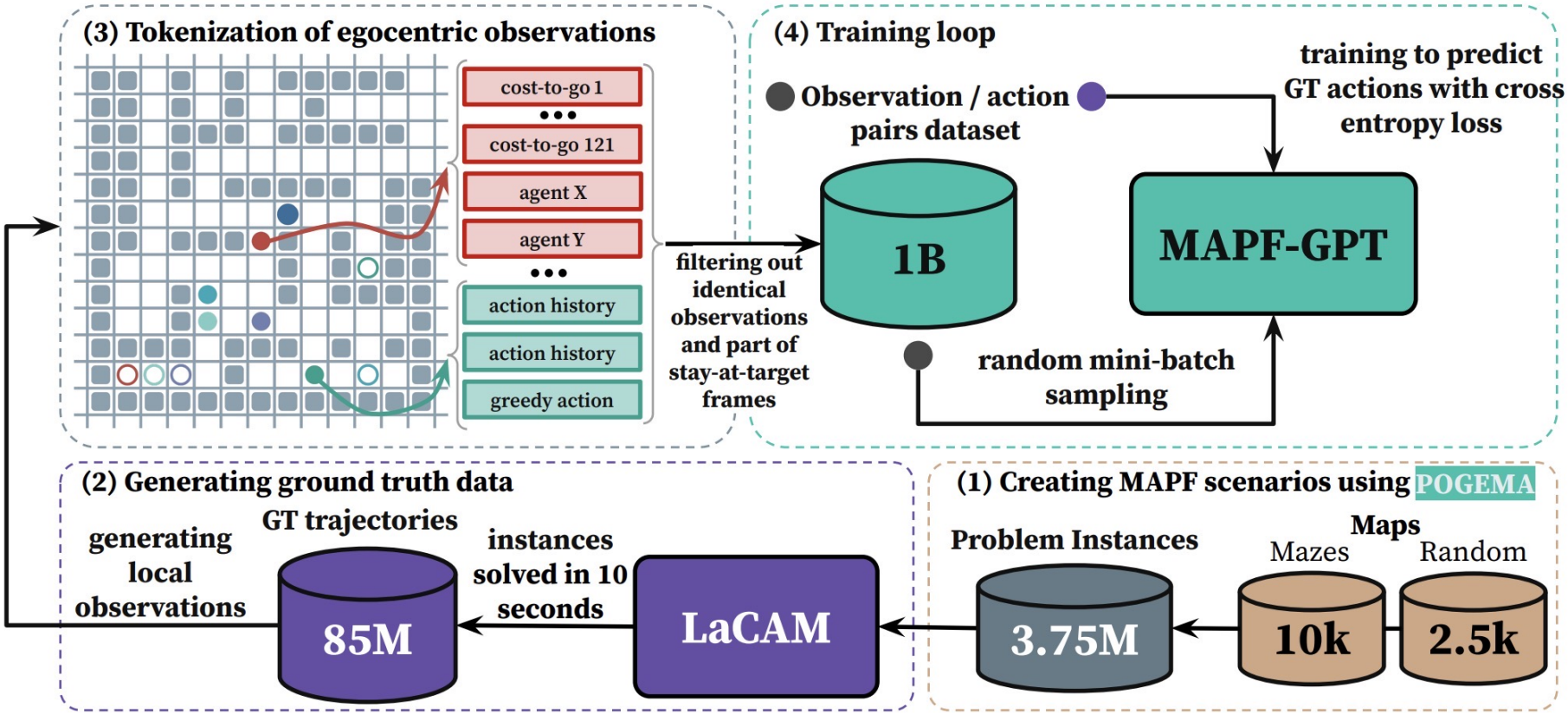

using cost penalties for planning
global coordinates of the agent
position
target
27
13
23
3
Heuristic Path Planner
providing path
updating dynamic penalties using local agents
sampling action
Action Decoder
RNN
actor / critic heads split
Spatial Encoder
local obstacles and path waypoints
local agents
input for learning-based policy


Random 20 × 20
Mazes 65 × 65
den520d 64 × 64
Paris_1 64 × 64
Follower
FollowerLite
PICO ★
Primal2
SCRIMP ★
Follower ★
FollowerLite ★
Primal2 ★
SCRIMP
Average Throughput
Number of Agents

MAPF-GPT


state of the environment
flattened tokenized input
[4, ∞, ∞, ∞, ∞,
3, 2, 1, 2, 3,
∞, ∞, 0, ∞, 2,
-3, -2, -1, ∞, 1,
-4, -3, -2, ∞, 0,
0, 0, -5, -5, ↓, ←, ←, ↓,
1, 1, 1, 1, ∩, ←, ←, ∩,
-2, -2, 4, -5, ↱, →, →, ↓,
∅, ∅, ∅, ∅, ∅, ∅, ∅, ∅]
tokenization vocabulary:
numerical : [-20, -19, ..., 0, ..., 19, 20]
actions: [←, →, ∩, ↓, ↑, ⊗]
greedy actions: [↱, ↳, ↰, ... , ↕, ↦]
empty: ∅; blocked: ∞; >20: >; <-20: <
egocentric cost-to-go
4 ∞ ∞ ∞ ∞
3 2 1 2 3
∞ ∞ 0 ∞ 2
-3 -2 -1 ∞ 1
-4 -3 -2 ∞ 0
local agents information
relative position
relative target
greedy action
action history
(0, 0)
(-5, -5)
down
[left, left, down]
(1, 1)
(1, 1)
wait
[left, left, wait]
(-2, -2)
(4, -5)
down-right
[right, right, down]

MAPF-GPT


(3) Tokenization of egocentric observations
cost-to-go 1
cost-to-go 121
agent X
agent Y
action history
action history
greedy action
filtering out identical observations and part of stay-at-target frames
(4) Training loop
Observation / action pairs dataset
1B
training to predict GT actions with cross entropy loss
MAPF-GPT
random mini-batch sampling
(2) Generating ground truth data
generating local observations
GT trajectories
85M
instances solved in 10 seconds
LaCAM
(1) Creating MAPF scenarios using POGEMA
Problem Instances
3.75M
Maps
Mazes
Random
10k
2.5k

Выводы

- **Ограничения архитектуры:**

  Модели могут не сходиться к оптимальной стратегии при использовании централизованного актор-критика с parameter sharing, особенно в условиях сложных взаимодействий между агентами.

- **Проблемы сходимости:**

  Децентрализованные архитектуры с большим числом параметров могут испытывать трудности с сходимостью.

- **Гибридные подходы:**

  Интеграция классических методов планирования с методами обучения с подкреплением, а также продуманный дизайн функции награды значительно улучшают сходимость и эффективность обучения в многокомпонентных системах.

Спасибо за внимание!

cogmodel.mipt.ru

pshenitsyn.as@phystech.edu