# Алгоритм PPO с обменом сообщениями для эффективного управления железнодорожным трафиком

Константин Махнев

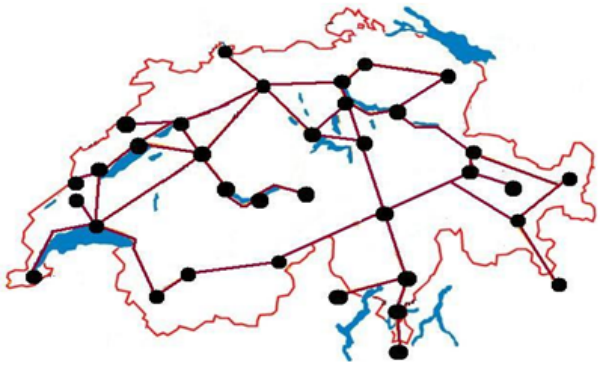
Научный руководитель: Алексей Шпильман
Научный консультант: Олег Свидченко

НИУ ВШЭ — Санкт-Петербург

25 мая 2021 г.

# Задача составления железнодорожного расписания

- Требуется на графе железнодорожной сети за минимальное суммарное время довести набор поездов до заданных целей.
- Поезда при этом не занимают одновременно один и тот же сегмент железнодорожных сетей.



# Виртуальное окружение Flatland[1]

Виртуальное окружение Flatland позволяет генерировать случайные железнодорожные графы и наборы маршрутов на них на двумерной карте, а также позволяет с дискретным временем симулировать и визуализировать составленное расписание.

[1] Erik Nygren et al.. "Flatland Challenge: Multi Agent Reinforcement Learning on Trains". In: 2019.

# Существующие решения

- Классические подходы основываются на эвристических подходах, таких как приоритезированное планирование, а также применении локальных оптимизаций[2]
- Лучшее решение Flatland Challenge в 2019 году, основанное на обучение с подкреплением использует независимых A2C агентов[3]
- Компания Deutsche Bahn совместно с Instadeep представили PPO агента с централизованным критиком, объединяющим наблюдения с использованием архитектуры трансформер

[2] Jonas Wälter. "Existing and novel Approaches to the Vehicle Rescheduling Problem (VRSP)". In: 2020.

[3] Roost D. et al. "Improving Sample Efficiency and Multi-Agent Communication in RL-based Train Rescheduling". In: 2020.

# Обучение с подкреплением


Proximal Policy Optimization(PPO)[4] — алгоритм, параллельно обучающий политику, выбирающую действие, а также критика, оценивающего ожидаемую награду в состоянии

[4] John Schulman et al.. "Proximal Policy Optimization Algorithms". In: 2017.

# Мотивация

- Обучение с подкреплением может выступить в качестве замены эвристикам, используемым в классических подходах
- Существующие на данный момент решения, основанные на обучении с подкреплением не используют подходов мультиагентного обучения с подкреплением, за исключением централизованного критика, немасштабируемого на большое число агентов

# Обмен сообщениями

На основе наблюдений агентов с помощью нейронной сети строятся сообщения
Эти сообщения передаются другим агентам и при выборе ими действия добавляются к их наблюдению.

- DIAL[5] сообщения передаются через дифференцируемый канал и оптимизируются вместе с политикой
- MAAC[6] сообщения от других агентов объединяются при помощи механизма внимания

[5] Jakob N. Foerster et al.. "Learning to Communicate with Deep Multi-Agent Reinforcement Learning". In: 2016.

[6] Jiechuan Jiang et al.. "Actor-Attention-Critic for Multi-Agent Reinforcement Learning". In: 2018.

# Цели и задачи

**Цель:** разработать алгоритм, оптимизирующий железнодорожную транспортировку, методами мультиагентного обучения с подкреплением с обменом сообщениями

**Задачи:**

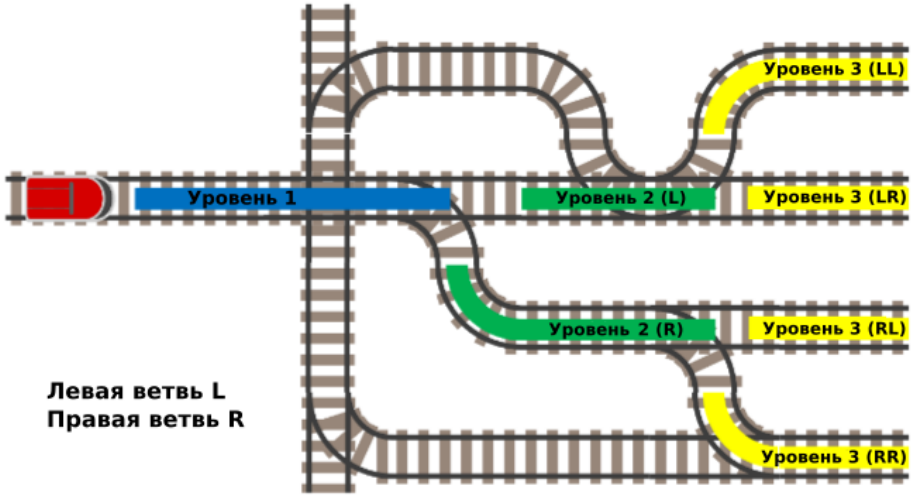
1. Свести задачу к обучению с подкреплением, определив наблюдения для агентов
2. Реализовать алгоритм обучения с подкреплением, использующий обмен сообщениями между агентами
3. Протестировать алгоритм на окружении Flatland и сравнить его с существующими решениями

# Окружение для тестирования

Для сравнения предложенных модификаций, в дальнейшем используются случайно сгенерированное окружение, размера $25 \cdot 25$ клеток, и использующее 10 поездов
При сравнительно большом разнообразии возможных ситуаций время обучения не слишком велико

| | 5 поездов | 10 поездов | 20 поездов |
|---|---|---|---|
| Время обучения | $\approx$ 14 часов | $\approx$ 44 часа | $\approx$ 140 часов |

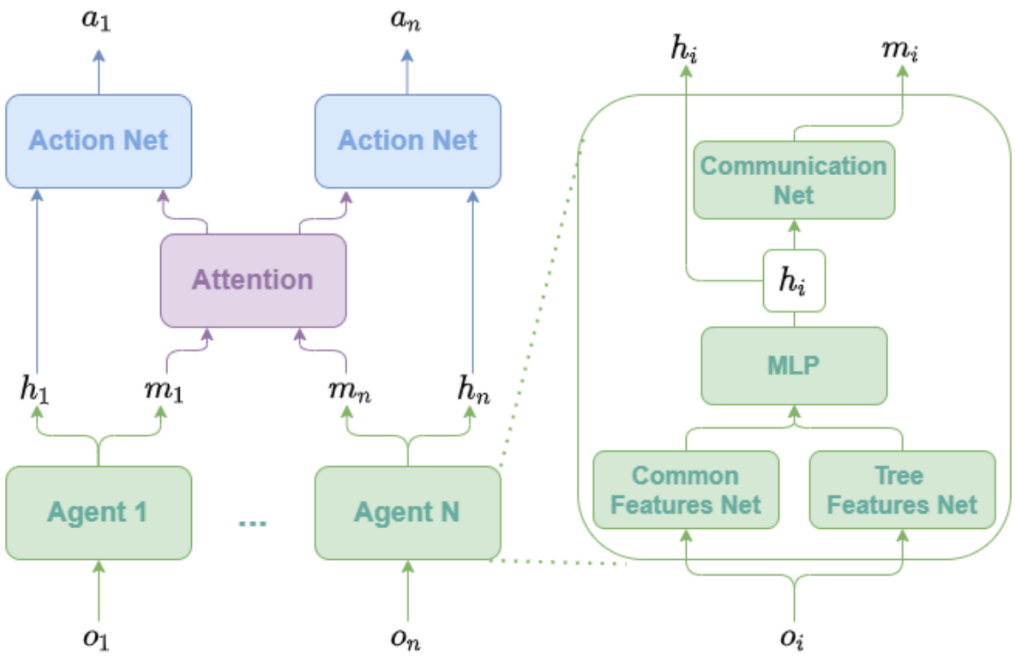
# Дерево наблюдений




| | Доля добравшихся поездов % | | Среднее время в пути | |
|---|---|---|---|---|
| | 10 поездов | 20 поездов | 10 поездов | 20 поездов |
| Глобальное наблюдение | $43.9 \pm 1.8$ | – | $210.5 \pm 34.8$ | – |
| Дерево наблюдений[7] | $81.8 \pm 1.2$ | $57.0 \pm 2.1$ | $107.4 \pm 4.7$ | $215.9 \pm 10.9$ |
| Модифицированное дерево наблюдений | $85.2 \pm 1.1$ | $60.7 \pm 2.4$ | $100.9 \pm 4.6$ | $204.6 \pm 10.4$ |

[7] Flatland-RL : Multi-Agent Reinforcement Learning on Trains. "S. Mohanty et al.". In: 2020.

# Архитектура агента




$o, a, m$ и $h$ соответственно обозначают наблюдения, действия, сообщения и выделенные признаки

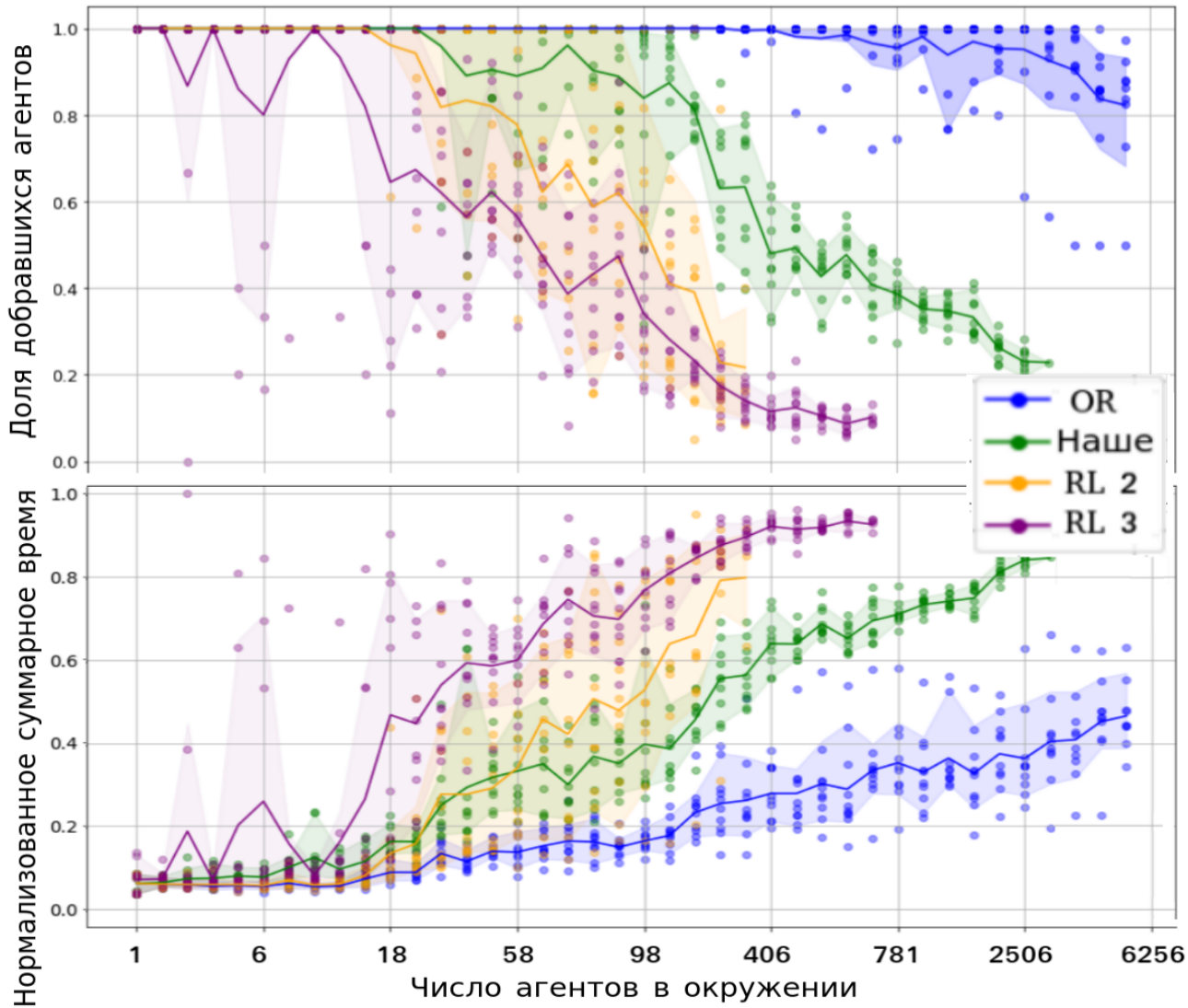
- Рекурсивные сети для обработки наблюдений
- Обмен сообщениями
- Механизм внимания для объединения сообщений

# Сравнение модификаций в архитектуре агента

| | Доля добравшихся поездов % | | Среднее время в пути | |
|---|---|---|---|---|
| | 10 поездов | 20 поездов | 10 поездов | 20 поездов |
| Без модификаций архитектуры | $85.2 \pm 1.1$ | $60.7 \pm 2.4$ | $100.9 \pm 4.6$ | $204.6 \pm 10.4$ |
| Рекурсивные сети | $88.2 \pm 0.9$ | $70.6 \pm 2.8$ | $92.3 \pm 5.4$ | $165.1 \pm 9.5$ |
| Обмен сообщениями без внимания | $89.1 \pm 1.0$ | $65.9 \pm 2.9$ | $87.0 \pm 5.3$ | $189.6 \pm 13.2$ |
| Обмен сообщениями + внимание | $92.4 \pm 0.9$ | $72.2 \pm 2.7$ | $83.5 \pm 4.6$ | $188.4 \pm 14.8$ |
| Сообщения + внимание + рекурсивные | $93.1 \pm 0.7$ | $81.3 \pm 2.4$ | $72.2 \pm 2.8$ | $136.6 \pm 8.6$ |
| Классический подход | $100.0 \pm 0.0$ | $100.0 \pm 0.0$ | $68.5 \pm 1.1$ | $115.5 \pm 3.4$ |

В таблице приведены результаты привнесения модификаций в архитектуру агента.
Доверительный интервал указан с уровнем доверия 99%

# Сравнение лучших решений Flatland Challenge



# Результаты

- Реализован алгоритм оптимизации железнодорожной транспортировки, основанный на обучении с подкреплением. Применение рекурсивых сетей и механизма обмена сообщениями, позволило повысить долю добирающихся поездов с 0.85 до 0.93 на окружении с 10 поездами
- Решение было протестировано на окружении Flatland и заняло 1 место среди решений, использующих обучение с подкреплением, превзойдя следующее решение по доле добравшихся агентов в среднем на 17.6% на наборе окружений из Flatland Challenge 2020
- По результатам соревнования подготовлена статья[8], включающая наше решение и доступная на arxiv.org

[8] Florian Laurent et al.. "Flatland Competition 2020: MAPF and MARL for EfficientTrain Coordination on a Grid World". In: 2021.