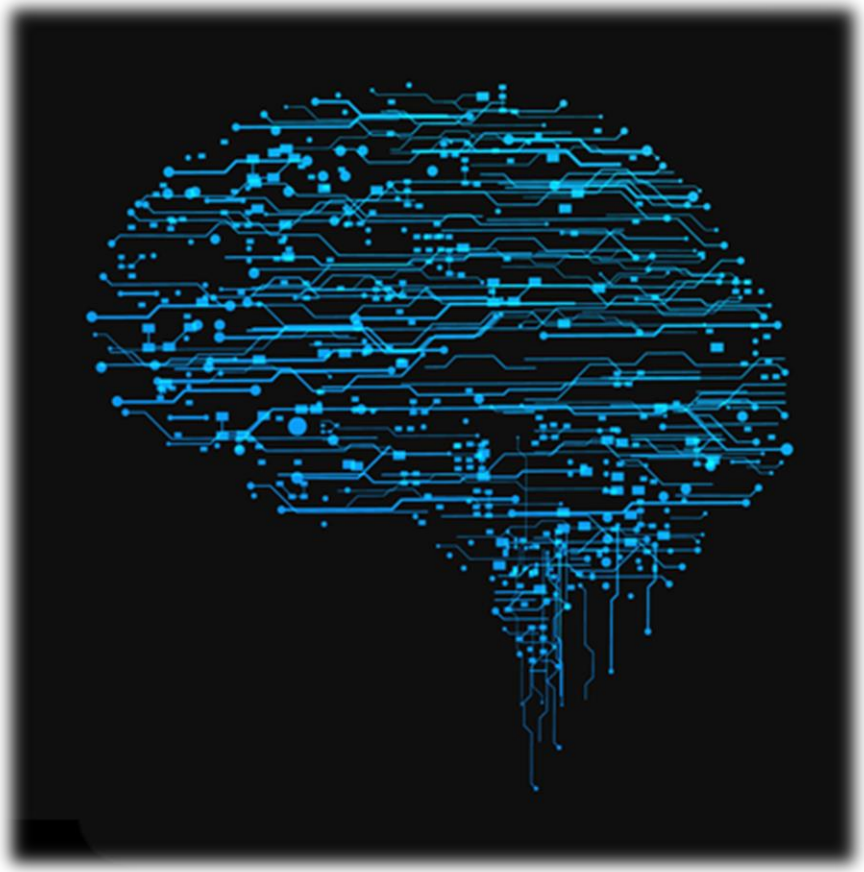
МЕТОДЫ ОБУЧЕНИЯ СЕМЕЙСТВА ГРАДИЕНТНОЙ ПОЛИТИКИ

И СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ VPG, TRPO, PPO

Подготовили студенты гр. 1308:

- Мельник Даниил
- Лепов Алексей

ВВЕДЕНИЕ



Цель исследовательской работы:

- рассмотреть семейство методов Policy Gradient;
- изучить и реализовать на практике алгоритмы VPG, TRPO, PPO;
- дать оценку работе алгоритмов.

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

Среда: окружение, в котором действует агент.

Агент: сущность, принимающая решения на основе информации из среды.

Состояние: информация о текущем состоянии среды, доступная агенту.

Награда: сигнал обратной связи от среды, определяющий успешность действий агента.

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ

Цель: максимизация накопленной суммы наград за период взаимодействия.

Обучение: процесс, в котором агент улучшает свои действия, оптимизируя награду.

Алгоритмы: Q-обучение, глубокое обучение с подкреплением и др.

СРЕДА И ЗАДАЧИ
CARTPOLE

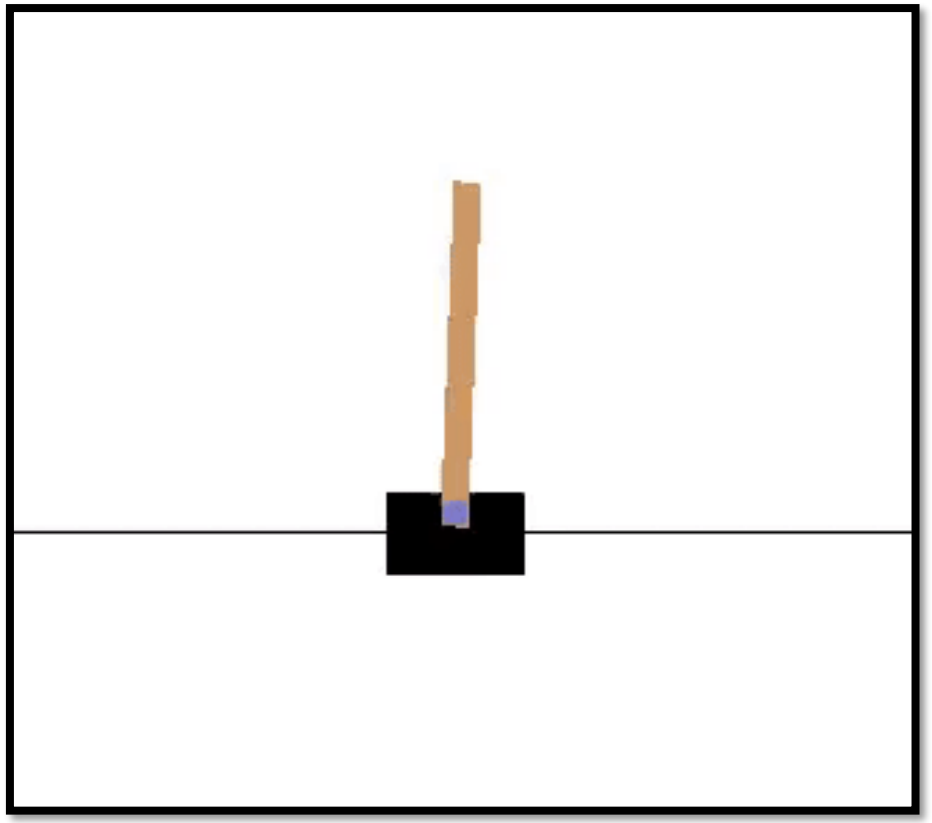
СРЕДА И ЗАДАЧИ CARTPOLE

Цель: удерживать шест в вертикальном положении на тележке.

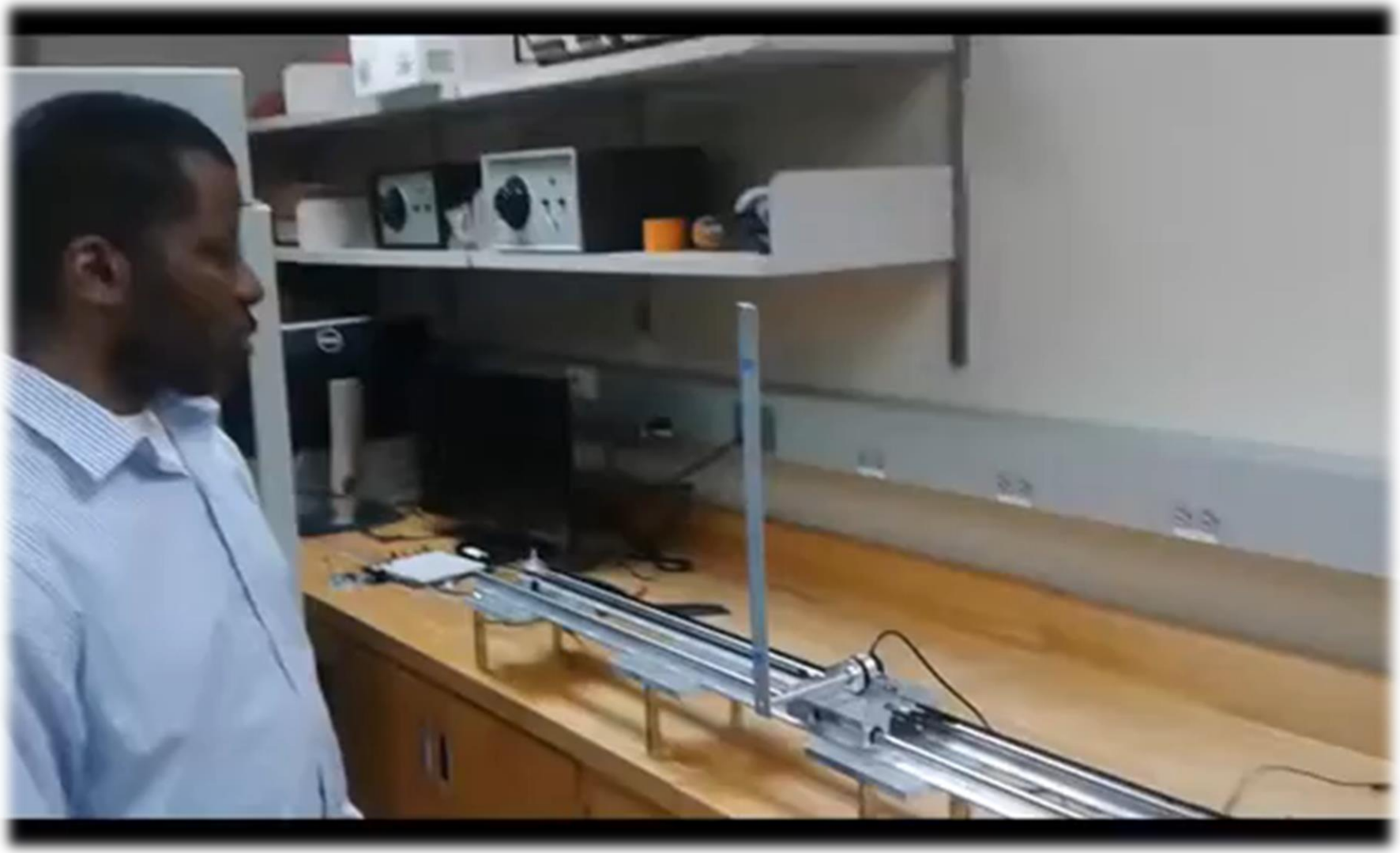
Агент имеет доступ к наблюдениям:

- положение,
- скорость тележки,
- угол наклона,
- скорость шеста.

Действия: двигаться влево или вправо.



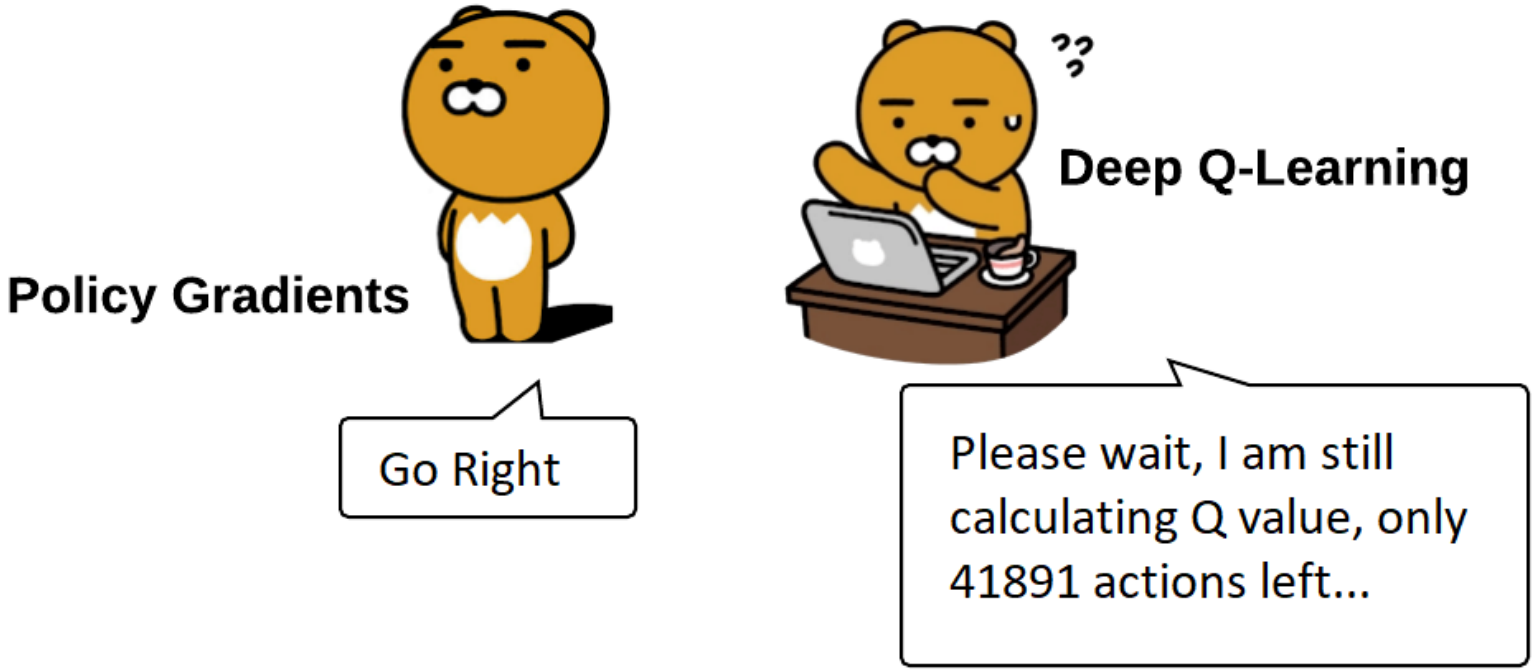
РЕАЛЬНЫЙ ПРОТОТИП CARTPOLE



VANILLA POLICY GRADIENT (VPG)

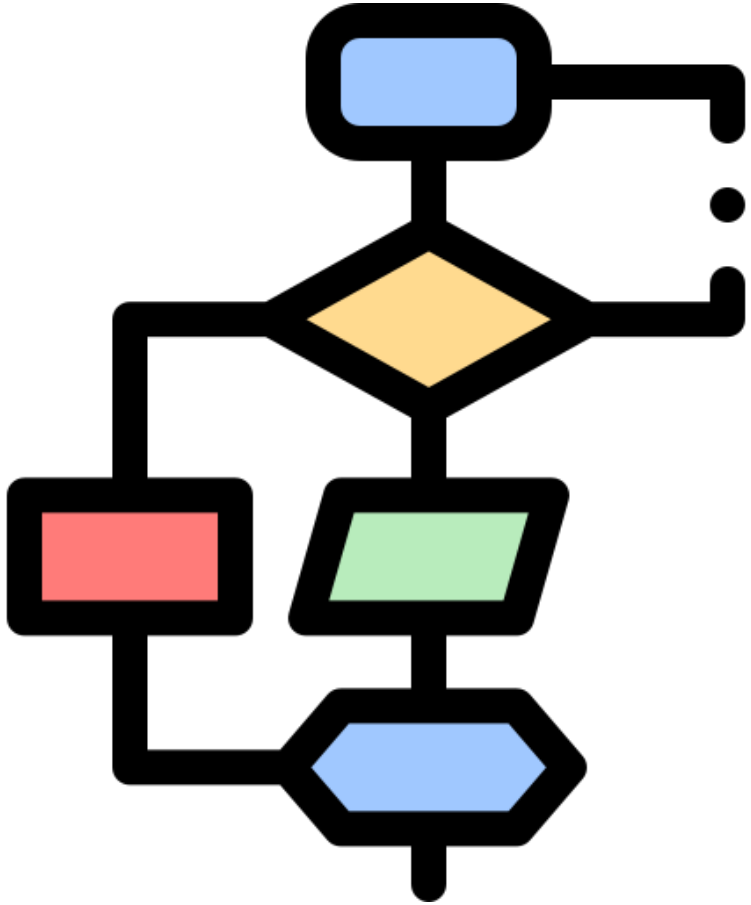
ОПИСАНИЕ МЕТОДА VPG

Метод VPG (Vanilla Policy Gradient) - простой и популярный метод обучения с подкреплением. Он использует градиентный спуск для обновления параметров политики и максимизации суммарной награды.



ОСНОВНЫЕ ШАГИ

1. Собрать данные
2. Вычислить функцию потерь
3. Вычислить градиент
4. Обновить параметры политики
5. Повторить шаги 1 - 4



ФОРМУЛА POLICY GRADIENT

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{k=1}^{n} \nabla \ln(P_k(a|s,\theta) \cdot Q_k(s,a)$$

- $J(\theta)$ - функция производительности агента, которую мы хотим максимизировать
- θ - параметры политики агента
- N - число эпизодов для оценки градиента
- $\nabla ln(P(a|s,\theta)$ - градиент логарифма вероятности действия a в состоянии s по отношению к θ
- $Q(s,a)$ - оценка ожидаемого вознаграждения для выполнения действия a в состоянии s

ФОРМУЛА POLICY GRADIENT

Идея: Обновляем параметры политики θ, используя градиент функции производительности, умноженный на оценку ожидаемого вознаграждения.

Процесс: Собираем опыт, оцениваем градиент и обновляем параметры политики.

Policy Gradient является основой для различных алгоритмов обучения с подкреплением.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

Преимущества:

- Простота реализации
- Концептуальная простота
- Гарантированная сходимость

Недостатки:

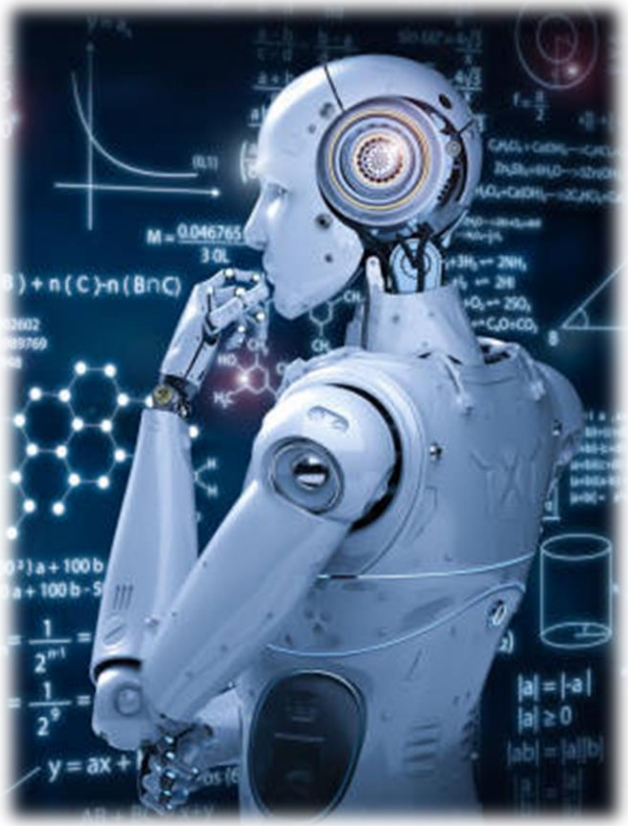
- Высокая дисперсия градиентов
- Отсутствие использования опыта
- Зависимость от гиперпараметров
- Отсутствие учета корреляции между действиями

TRUST REGION POLICY OPTIMIZATION (TRPO)

ОПИСАНИЕ МЕТОДА TRPO

Метод TRPO (Trust Region Policy Optimization) - алгоритм оптимизации политики обучения с подкреплением для задач с непрерывным пространством действий.

Основная идея TRPO - использование "области доверия", которая ограничивает изменение политики на каждом шаге.



ОБНОВЛЕНИЕ ПОЛИТИКИ TRPO

Процесс обновления политики включает:

- Сбор обучающих данных.
- Оценку ожидаемой награды для каждого состояния.
- Вычисление преимущества для каждого состояния.
- Вычисление градиента политики на основе оценки преимущества.
- Ограничение размера обновления политики с помощью "области доверия".
- Обновление политики с использованием ограниченного обновления.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

Преимущества	Недостатки
Гарантия монотонного улучшения политики	Более сложная реализация и вычислительно затратная оптимизация
Контроль степени изменения политики с помощью ограничения региона доверия	Может потребоваться больше времени для сходимости в сравнении с другими методами

PROXIMAL POLICY OPTIMIZATION (PPO)

ОПИСАНИЕ МЕТОДА PPO

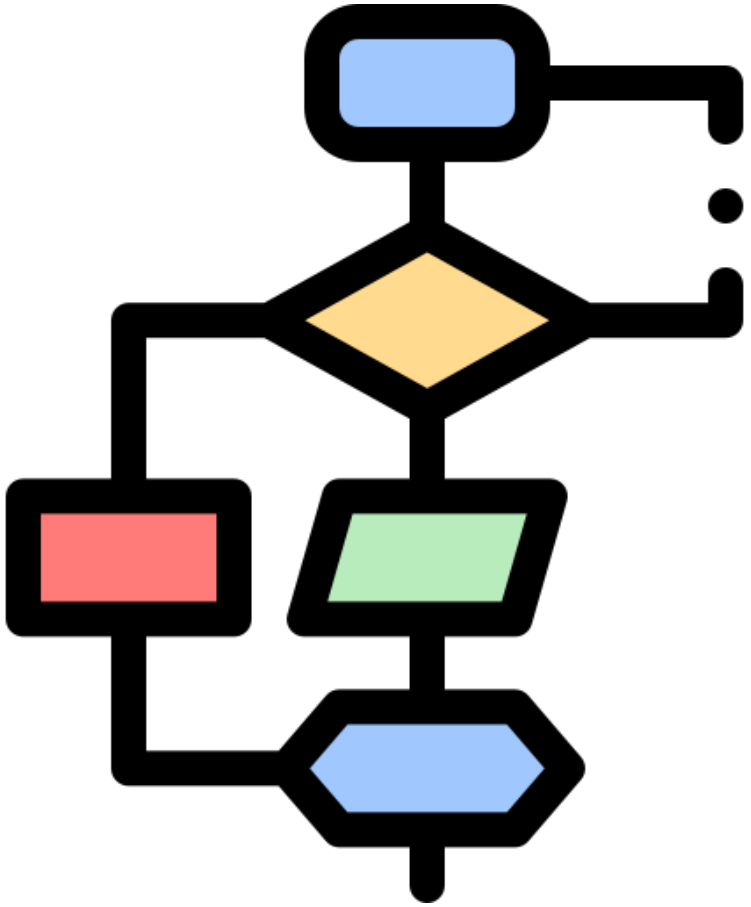
Проксимальная оптимизация политики (PPO) - алгоритм обучения с подкреплением для обучения агентов в задачах последовательных действий.

Он обновляет политику постепенно, основываясь на собранных данных, и использует функцию потерь с ограничением на размер изменений политики.

PPO обеспечивает стабильное обучение с подкреплением и контролирует величину изменений для избежания слишком больших скачков.

ОПИСАНИЕ МЕТОДА PPO

1. Сбор данных
2. Вычисление преимуществ
3. Вычисление функции потерь
4. Обновление политики
5. Итерационный процесс



ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

Преимущества	Недостатки
Более стабильная и безопасная оптимизация политики	Дополнительные вычислительные затраты для оценки отношения вероятностей действий.
Возможность контролировать величину обновления с помощью гиперпараметров	Не всегда гарантирует сходимость к оптимальной политике.

МЕТОДЫ ОБУЧЕНИЯ СЕМЕЙСТВА ГРАДИЕНТНОЙ ПОЛИТИКИ

И СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ VPG, PPO, TRPO

Подготовили студенты гр. 1308:

- Мельник Даниил
- Лепов Алексей