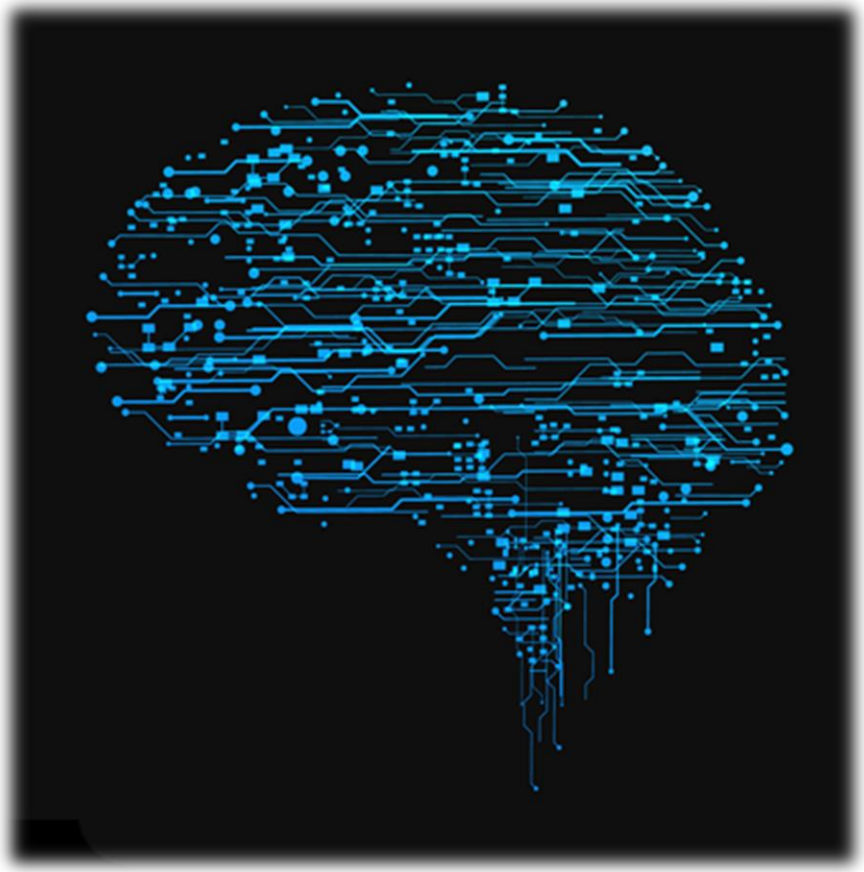
МЕТОДЫ ОБУЧЕНИЯ СЕМЕЙСТВА ГРАДИЕНТНОЙ ПОЛИТИКИ

И СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ VPG, PPO, TRPO

Подготовили студенты гр. 1308:

- Мельник Даниил
- Лепов Алексей

ВВЕДЕНИЕ



Цель исследовательской работы:

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

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 и скорость шеста.

Действия: двигаться влево или вправо.

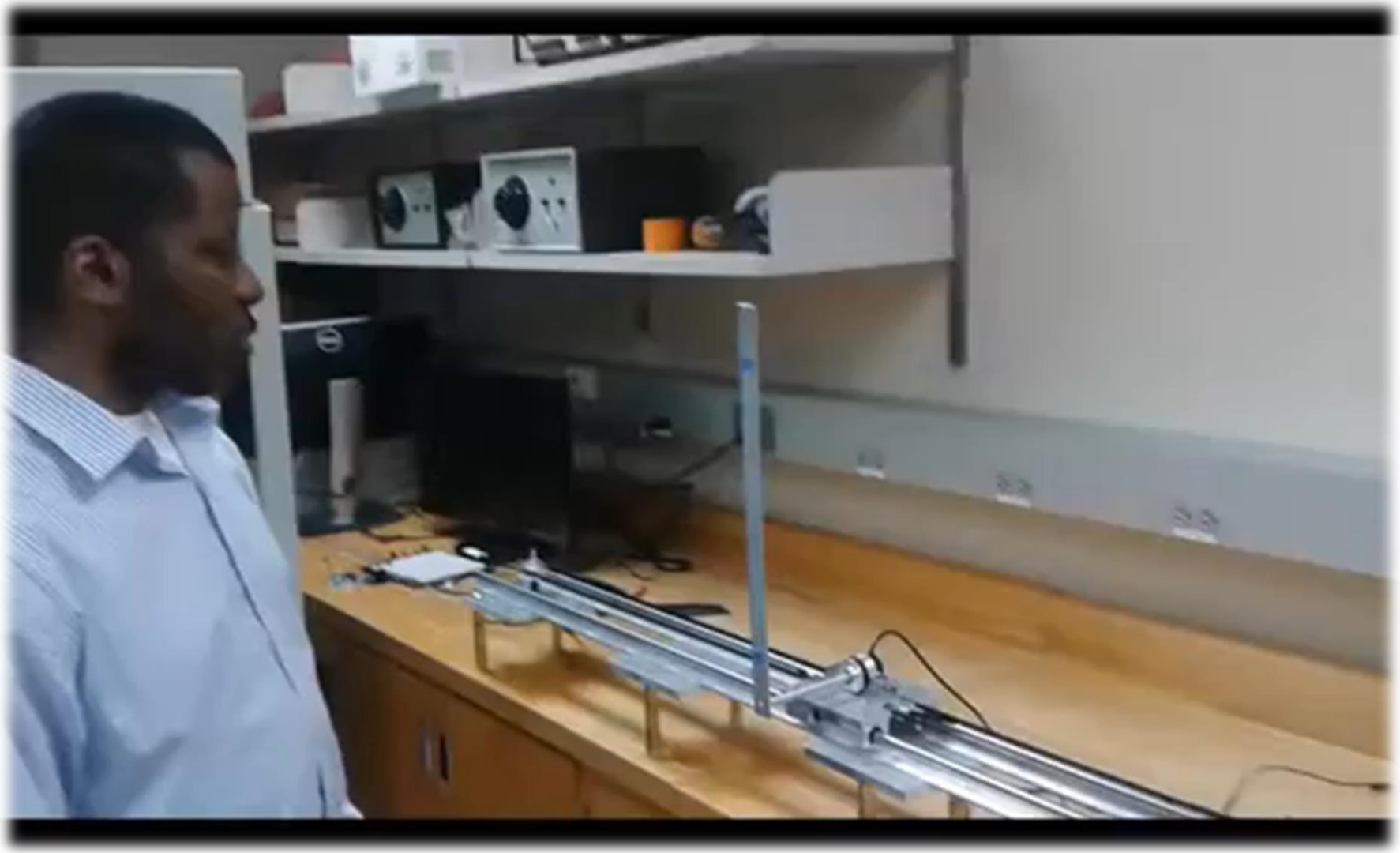
Policy Gradient: оптимизация параметров политики.

Градиентный спуск для обновления параметров.

Без моделирования функции ценности состояний или действий.

Эффективное обучение в средах с большим пространством состояний и действий.

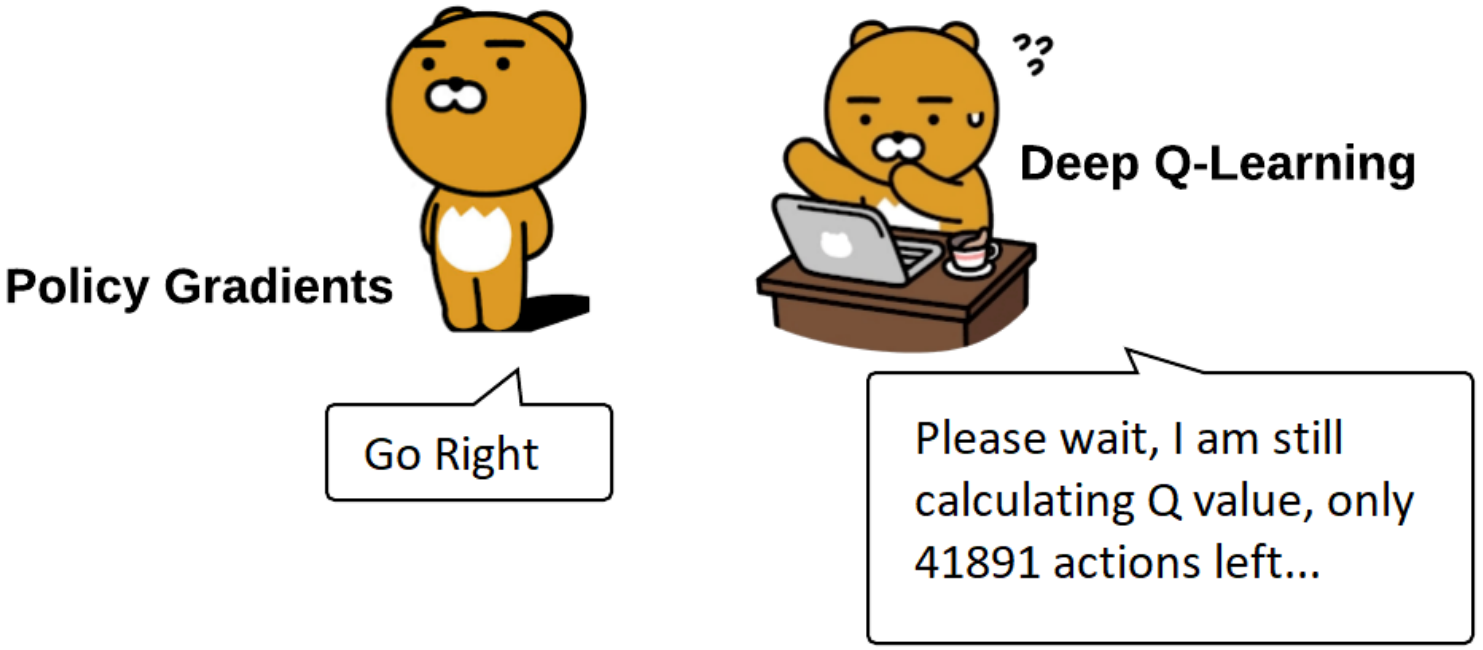
РЕАЛЬНЫЙ ПРОТОТИП CARTPOLE



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.



ФОРМУЛА POLICY GRADIENT

$$\nabla J(\theta) \approx 1/N * \Sigma[1 \text{ to } N] \nabla \log(P(a|s, \theta)) * Q(s, a)$$

- $J(\theta)$ - функция производительности агента, которую мы хотим максимизировать
- θ - параметры политики агента
- N - число эпизодов для оценки градиента
- $\nabla \log(P(a|s, \theta))$ - 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.

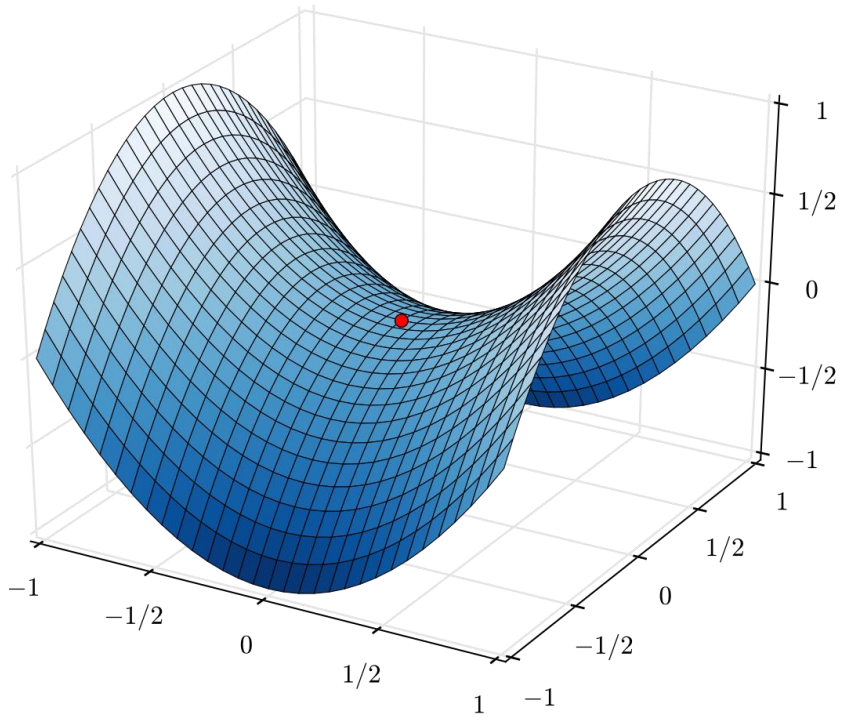
Policy Gradient является основой для различных алгоритмов обучения с подкреплением.

ФОРМУЛА ОБНОВЛЕНИЯ ПАРАМЕТРОВ

Формула обновления параметров варианта политики по градиентам (VPG) выглядит следующим образом:

$$\Delta\theta = \alpha * \nabla_\theta J(\theta)$$

- $\Delta\theta$: изменение параметров политики
- α: скорость обучения (learning rate)
- ∇_θ: градиент по параметрам политики
- $J(\theta)$: ожидаемая награда (reward)



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

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.

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

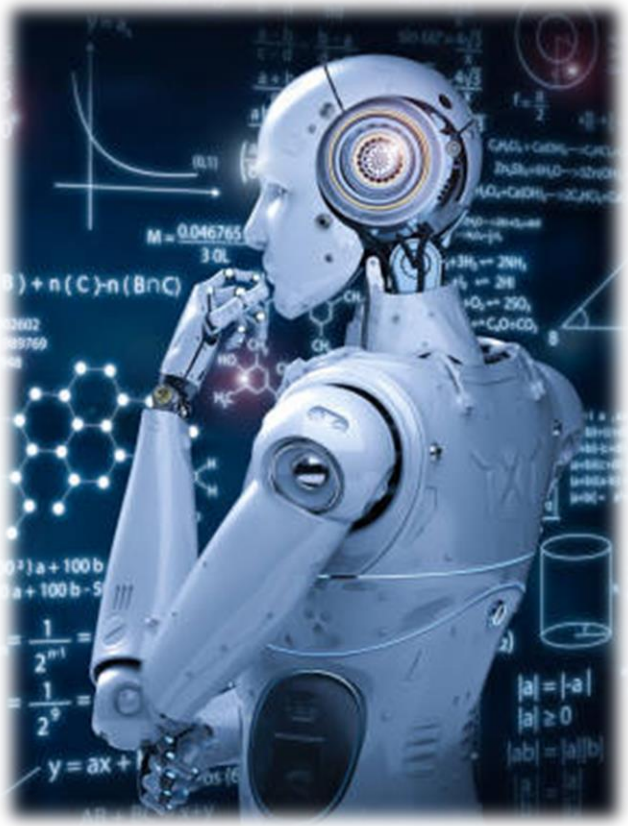
Преимущества	Недостатки
Более стабильная и безопасная оптимизация политики	Дополнительные вычислительные затраты для оценки отношения вероятностей действий.
Возможность контролировать величину обновления с помощью гиперпараметров	Не всегда гарантирует сходимость к оптимальной политике.

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

МЕТОДЫ ОБУЧЕНИЯ СЕМЕЙСТВА ГРАДИЕНТНОЙ ПОЛИТИКИ

И СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ VPG, PPO, TRPO

Подготовили студенты гр. 1308:

- Мельник Даниил
- Лепов Алексей