

Обучение с подкреплением по большому количеству человеческих демонстраций

Сазанович Никита Валерьевич, БПМ151

Научный руководитель:

Даниел Куденко, доктор компьютерных наук



ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

7 июня 2019

Происходит на марковском процессе принятия решений $\langle S, A, R, T, \gamma \rangle$.

Цель — нахождение политики $\pi : \mathcal{S} \rightarrow \mathcal{A}$ с максимальной кумулятивной дисконтированной наградой:

$$\mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(s_k, \pi(s_k), s_{k+1}) \right]$$

Демонстрации:

$\{s_0, a_1, r_1, s_1, \dots, s_{t-1}, a_t, r_t, s_t\}, \{s_0, a_1, s_1, \dots, s_{t-1}, a_t, s_t\}.$

Их неструктурированная форма: $\{(s_i, a_i)\}_{i=1}^N.$

Проблема эффективного использования опыта:

- OpenAI Five получил более чем 11000 лет суммарного игрового опыта Dota 2¹
- AlphaStar использовал 200 лет суммарного игрового опыта StarCraft II².

Тема соревнования MineRL³ для NeurIPS 2019.

¹OpenAI. OpenAI Five. 2018. <https://blog.openai.com/openai-five/>.

²AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. / O. Vinyals [et al.]. 2019. <https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/>.

³The MineRL Competition on Sample Efficient Reinforcement Learning using Human Priors. / W. H. Guss [et al.] // arXiv e-prints. 2019. Apr. arXiv:1904.10079.

- Изменение наград исходя из потенциалов действий⁴:

$$\mathcal{R}'(s, a, s', a') = \mathcal{R}(s, a, s') + (\gamma\Phi(s', a') - \Phi(s, a))$$

- Демонстрации для генерации потенциалов⁵.
- Демонстрации как средство исследования среды⁶.
- Использование демонстраций в алгоритме глубокого Q-обучения для предобучения⁷.

⁴Wiewiora E., Cottrell G. W., Elkan C. Principled Methods for Advising Reinforcement Learning Agents. // ICML. 2003.

⁵Reinforcement Learning from Demonstration through Shaping. / T. Brys [et al.] // IJCAI. 2015.

⁶Wang Z., Taylor M. E. Improving Reinforcement Learning with Confidence-Based Demonstrations. // IJCAI. 2017.

⁷Deep Q-learning From Demonstrations. / T. Hester [et al.]. 2018.

Мотивация

В литературе не исследован вопрос о влиянии объема демонстраций на производительность алгоритмов.

Цель

Исследовать выгоду использования неструктурированных демонстраций большого объема от многих людей для алгоритма глубокого Q-обучения (DQN) и среды Dota 2.

Задачи:

1. Разработать среду для обучения в игре Dota 2
2. Получить датасет демонстраций
3. Реализовать DQN с использованием демонстраций
4. Исследовать выгоду использования различного числа демонстраций.

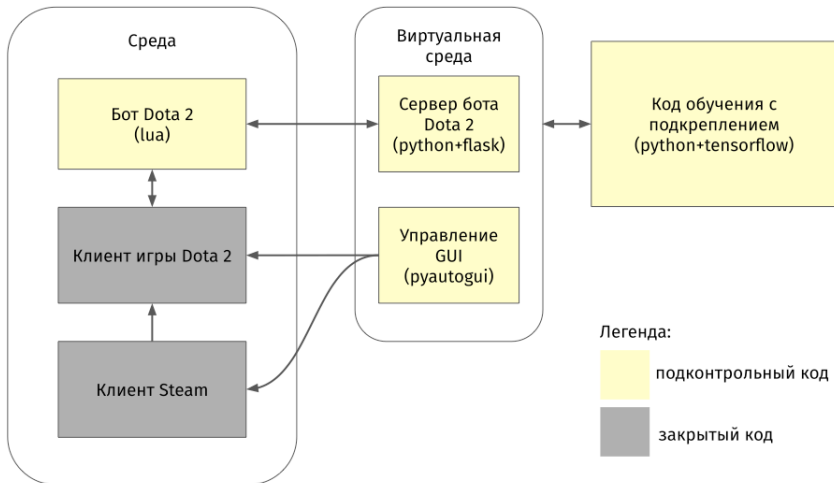
Среда Dota 2



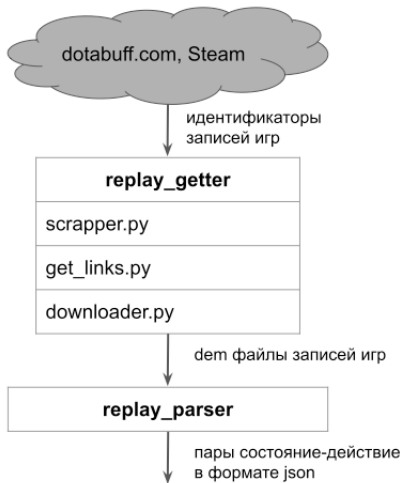
Параметризация:

- $S = \mathbb{R}^{18}$, $|A| = 11$, $R = 1$ при атаке юнита оппонента и 0 иначе.

Архитектура среды для обучения



Извлечение демонстраций



- Реализовано в работе⁸.
- Добавлено восстановление недостающей информации через Dota Bot Scripting API.
- Взяты 254 игры.
- Содержат 404129 пар состояние-действие.

⁸Парадовский Ю., Куденко Д. Crowd-sourced AI на примере игры Dota 2. 2018. СПбАУ РАН.

- Функция потерь DQN:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim D} \left[\left(r + \gamma Q(s', \arg \max_{a'} Q(s', a'; \theta_i); \theta_i^-) - Q(s, a; \theta_i) \right)^2 \right]$$

- Автоматически сгенерирована функция потенциалов⁹:

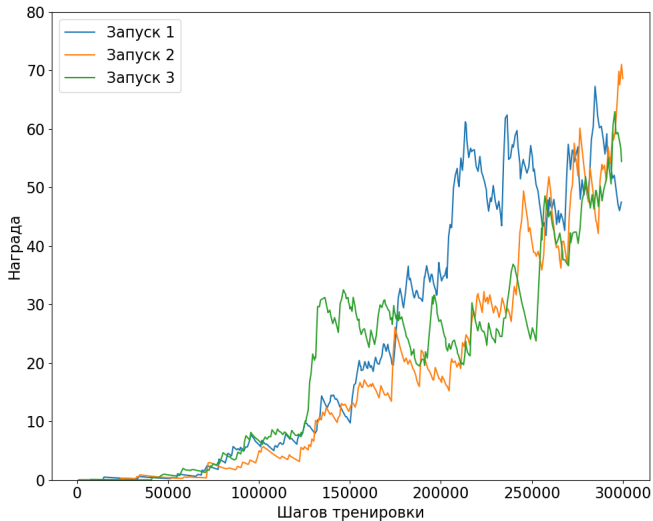
$$\begin{aligned} \Phi^D(s, a) &= \max_{(s^d, a^d) | a = a^d} g(s, s^d, \Sigma) \\ g(s, s^d, \Sigma) &= e^{(-\frac{1}{2}(s-s^d)\Sigma^{-1}(s-s^d))} \end{aligned}$$

- Объединение демонстраций на основе функции подобия состояний g и коэффициента максимальной близости K .

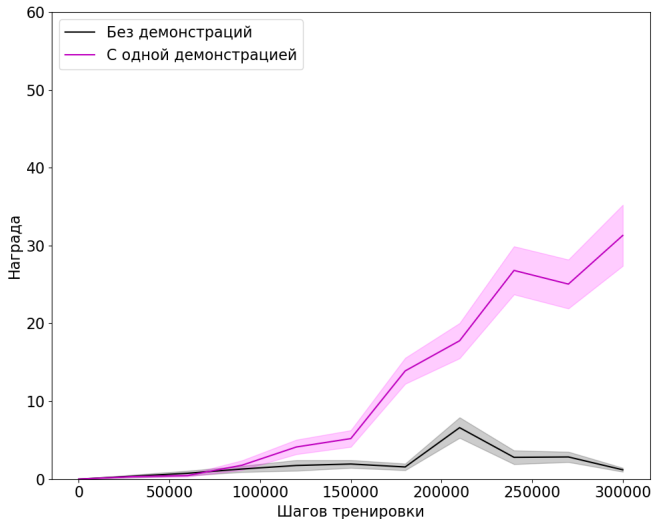
⁹Reinforcement Learning from Demonstration through Shaping. / T. Brys [et al.] // IJCAI. 2015.

Условия проведения

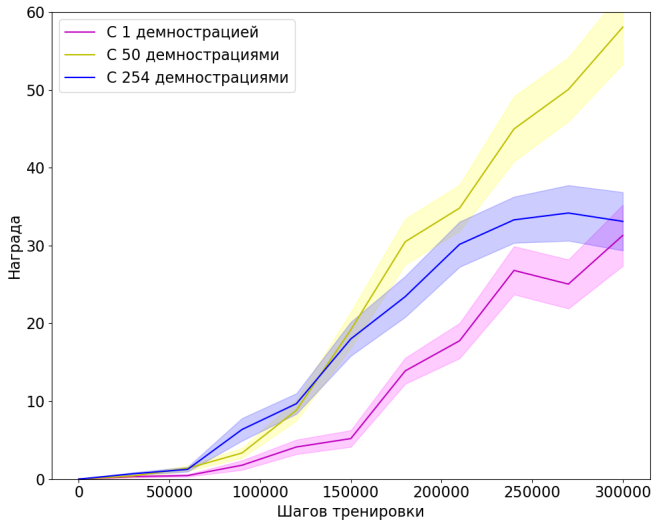
Обучение в течение 300000 шагов. Один запуск длится 12 часов.



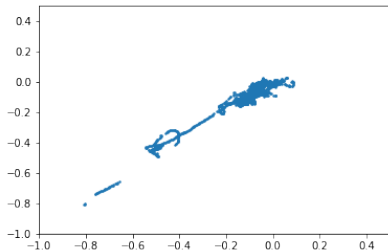
Использование неструктурированных демонстраций



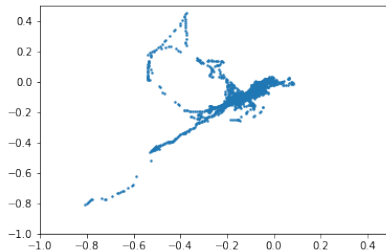
Объем демонстраций и производительность



Противоречивость демонстраций



(a) Траектория первого эксперта



(b) Траектория второго эксперта

Заключение

Выводы:

- Неструктурированные демонстрации позволяют увеличить среднюю награду DQN в Dota 2 с 1.19 до 31.30
- Оптимальным в моих условиях количеством оказалось не одна и не все демонстрации (средние награды 31.30, 58.07, 33.11 для 1, 50, 254 демонстраций).

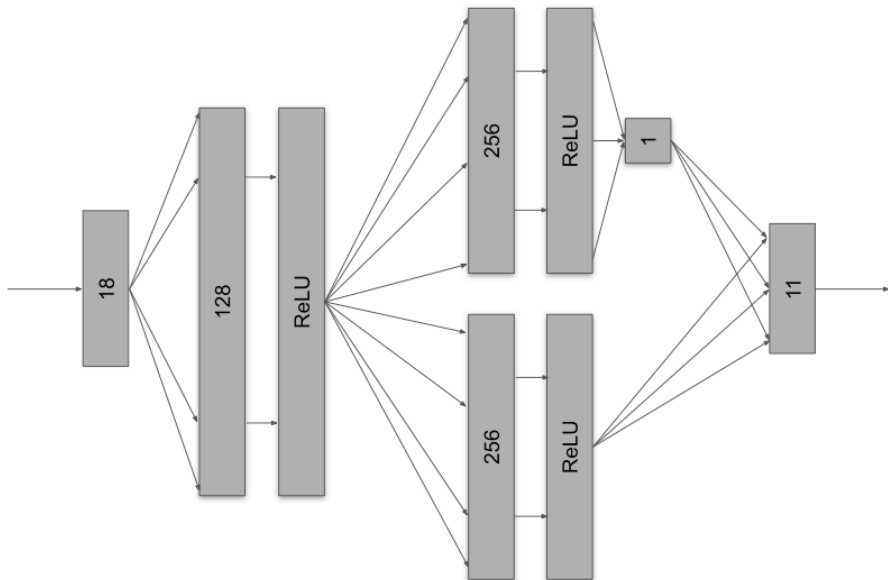
Продолжения:

- Рассмотрение большего набора значений количества демонстраций (четыре запуска одной конфигурации требуют 48 часов)
- Исследование методов объединения демонстраций
- Проверка гипотезы оптимальности на других средах и/или с другими алгоритмами.

Алгоритм объединения демонстраций

```
def generate_merged_demo():
    merged_demo = []
    for demo in demos:
        for demo_state, demo_action in demo:
            similar = False
            for state, action in merged_demo:
                sim = get_states_similarity(
                    demo_state, state)
                if sim > K:
                    similar = True
                    break
            if not similar:
                merged_demo.append(
                    DemoStateActionPair(
                        demo_state, demo_action))
```

Архитектура модели



Параметры обучения

метод оптимизации	Adam
коэффициент скорости обучения	$1e-3$
размер буфера опыта	100000
коэффициент ϵ -жадного исследования	0.1
частота обучения сети	4
частота синхронизации целевой сети обучения	1000
коэффициент дисконтирования	0.999
размер тренировочного блока	32
α приоритизированного буфера	0.6
β_0 приоритизированного буфера	0.4
ϵ приоритизированного буфера	$1e-6$