Система прийняття рішень для навігації в середовищах з неповною інформацією

Виконав:

Студент 4-го курсу

Групи КА-53

Титаренко А.М.

Керівник:

д-р фіз.-мат. наук, професор

Касьянов П.О.

Дослідження

- Об’єкт дослідження
- Предмет дослідження
- Мета дослідження

Задача візуальної навігації

Застосування методів глибокого навчання з підкріпленням для вирішення задачі візуальної навігації

Розробка та навчання системи прийняття рішень для візуальної навігації

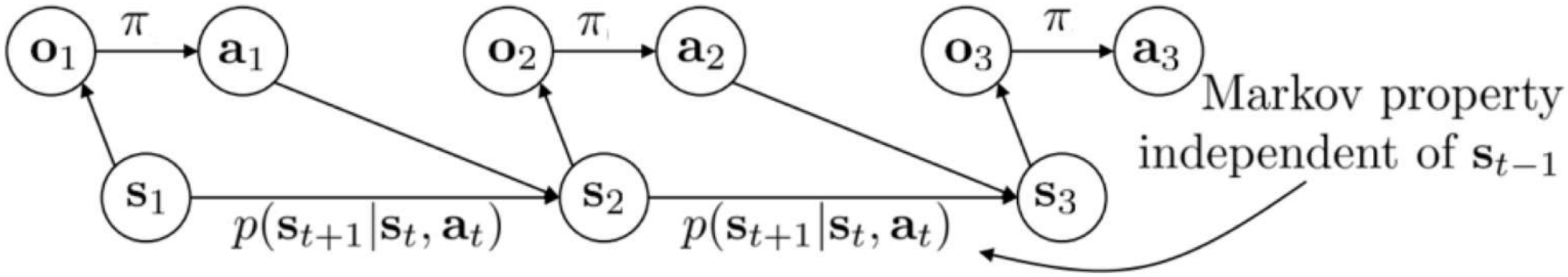
Актуальність

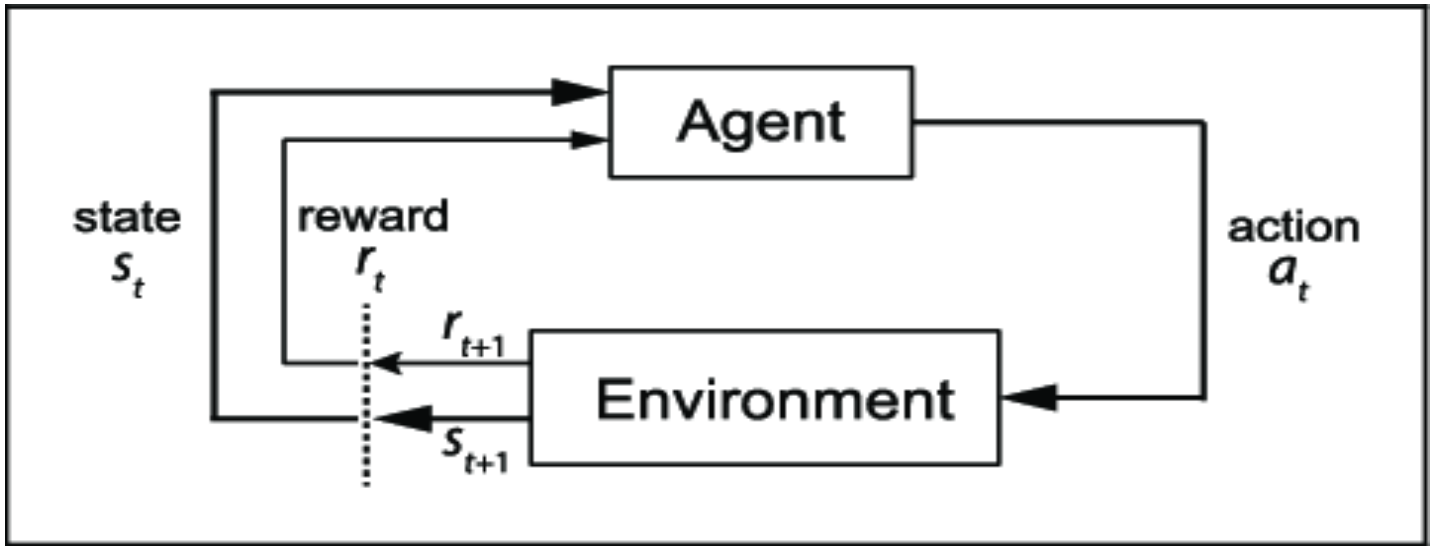
- Задача візуальної навігації наразі є предметом активних досліджень в машинному навчанні, комп’ютерному зорі та штучному інтелекті в цілому
- Якість рішення задачі впливає на розвиненість автономних роботизованих систем, що виконують доручення користувача у фізичному світі.
- Майже повна відсутність рішень на базі навчання з підкріпленням, що тільки набирає популярності через розвиток обчислювальної техніки та глибокого навчання

Постановка задачі

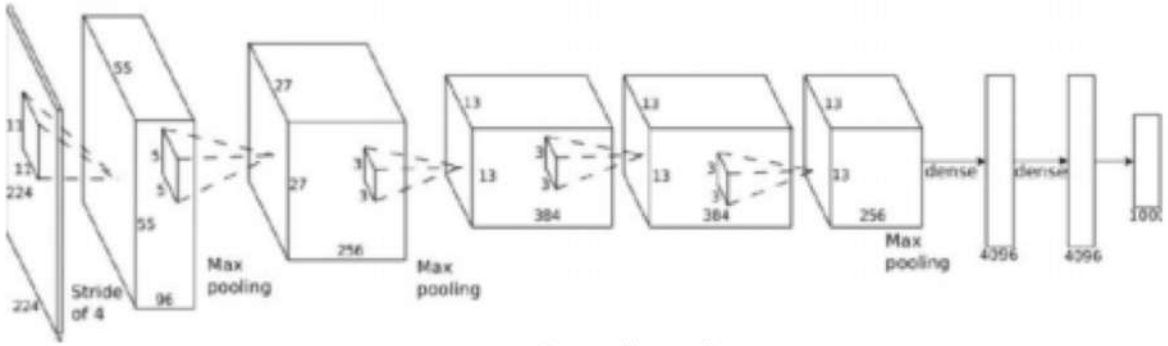
- Розробити систему для навчання та підтримки вибраних алгоритмів глибокого навчання з підкріпленням.
- Навчити агент, який на кожному кроці за візуальним сигналом (Зображення з камери та карта глибин) та інформацією про положення цілі (кут та відстань) приймав би рішення щодо руху, які б приводили до знаходження цілі за коротший час.
- Упевнитися в здатності агенту до пошуку в середовищах, відсутніх в наборі для тренування.

Модель взаємодії середовища і агенту





Глибоке навчання з підкріпленням



$\mathbf{o}_t$ $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ $\mathbf{a}_t$

$\mathbf{s}_t$ – стан

$\mathbf{o}_t$ – спостереження

$\mathbf{a}_t$ – дія

$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – функція-

$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – стратегії

Цільовий функціонал

$$\pi_\theta(\tau) = \pi_\theta(s_1, a_1, \dots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi_\theta(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)}[R(\tau)]$$

$$J(\theta) = \int \pi_\theta(\tau) R(\tau) d\tau$$

$$Q(s_t, a_t) = \sum_{t'=t}^{T} \mathbb{E}_{\pi_\theta}[R(s_{t'}, a_{t'})|s_t, a_t]$$

$$V(s_t) = \mathbb{E}_{a_t \sim \pi_\theta(a_t, s_t)}[Q(s_t, a_t)]$$

Існування оптимальної стратегії

- **Теорема** Для Марковських процесів прийняття рішень із скінченними просторами дій та станів існує оптимальна марковська інформаційна стратегія, яка максимізує цільовий функціонал (очікувану кумулятивну винагороду)

Пошук оптимальної стратегії методом градієнту стратегії

$$\nabla_\theta J(\theta) = \int \nabla_\theta \pi_\theta(\tau) R(\tau) d\tau = \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) R(\tau) d\tau =$$

$$= \mathbb{E}_{\tau \sim \pi_\theta(\tau)} [\nabla_\theta \log \pi_\theta(\tau) R(\tau)]$$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t | s_t) \right) \left(\sum_{t=1}^{T} R(s_t, a_t) \right) \right]$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t} | s_{i,t}) \right) \left(\sum_{t=1}^{T} R(s_{i,t}, a_{i,t}) \right) \right]$$

Алгоритм REINFORCE

Ініціалізувати довільними значеннями параметри стратегії π_θ

Повторювати:

Генерування набору прикладів, використовуючи стратегію $\pi_\theta(a_t|s_t)$

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}[(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|s_t))(\sum_{t=1}^{T}R(s_t,a_t))]$$

$$\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta)$$

Зниження дисперсії

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\left(\sum_{t\prime=t}^{T}R(s_{i,t\prime},a_{i,t\prime})\right)\right]$$

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}[\nabla_\theta \log \pi_\theta(\tau)\,[R(\tau)-b]]$$

Алгоритм Актор-Критик

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(a_{i,t}|s_{i,t}\right) A^{\pi_\theta}\left(s_{i,t}, a_{i,t}\right)\right]$$

$$A^{\pi_\theta}\left(s_{i,t}, a_{i,t}\right) = Q\left(s_{i,t}, a_{i,t}\right) - V\left(s_{i,t}\right)$$

Наближення

$$A^{\pi}(s_t, a_t) \approx R(s_t, a_t) + V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$$

Алгоритм Наближеної оптимізації стратегії (PPO)

$$\mathbb{E}_{(s_t,a_t)\sim p_\theta(s_t,a_t)}\left[\frac{\pi_{\theta'}(a_t|s_t)}{\pi_\theta(a_t|s_t)}A^{\pi_\theta}(s_t,a_t)\right]\rightarrow max$$

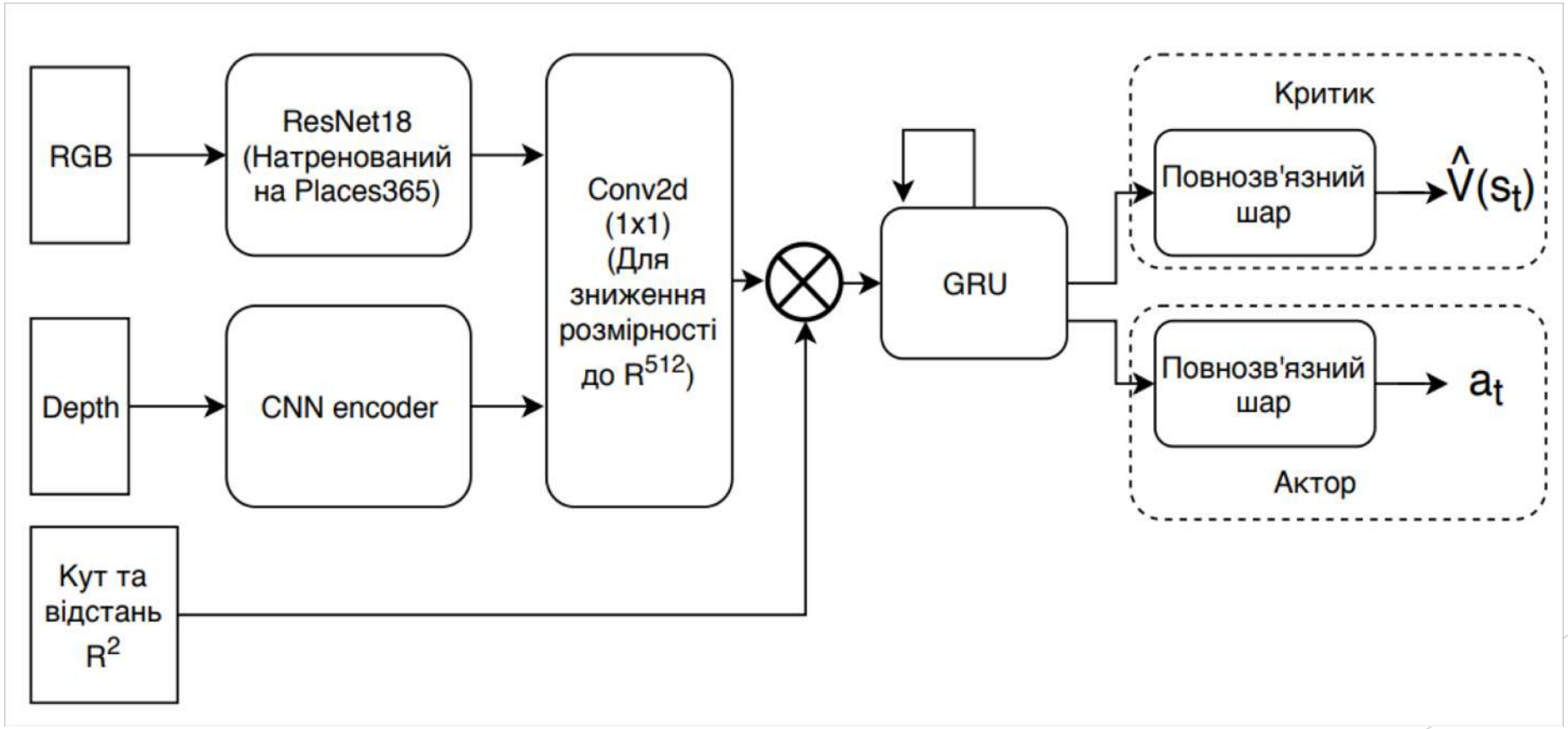
При обмеженні:

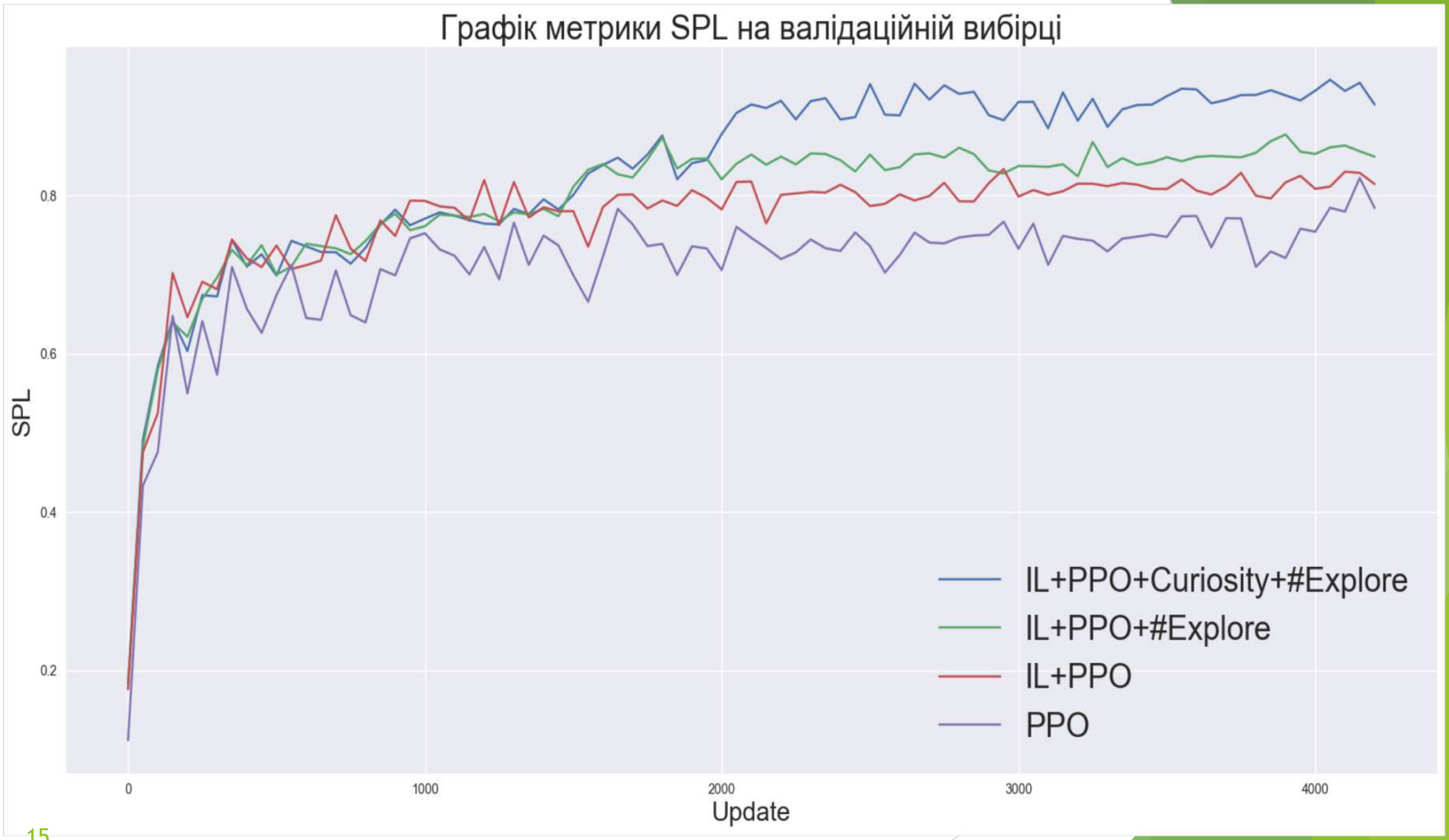
$$D_{KL}(\pi_{\theta'}(a_t|s_t)\|\pi_\theta(a_t|s_t))\leq\epsilon$$

Ідея: Робити невеликі кроки від стратегії до стратегії

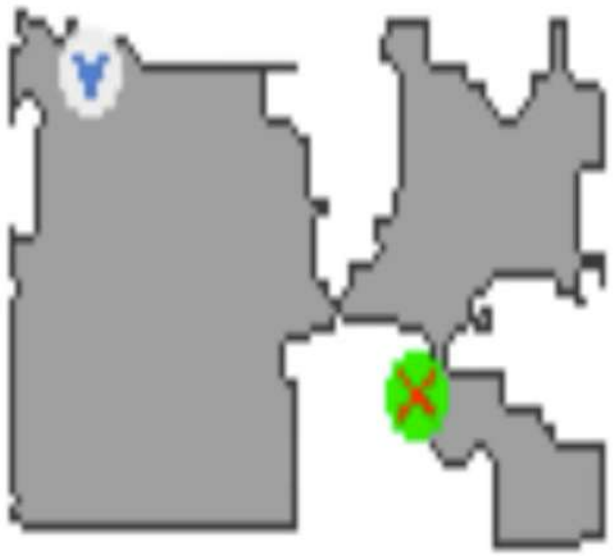
Набір сцен
Gibson

Архітектура мережі для наближення функції-стратегії




Графік метрики SPL на валідаційній вибірці
SPL
0.2
0.4
0.6
0.8
0
1000
2000
3000
4000
Update
IL+PPO+Curiosity+#Explore
IL+PPO+#Explore
IL+PPO
PPO

Приклади траєкторій



Висновки

У результаті було:

- Спроектовано архітектуру і процедури навчання моделі на базі алгоритму Наближеної оптимізації плану (PPO) із різними модифікаціями для заохочення дослідження агенту.
- Розроблено програмне забезпечення для навчання і валідації моделі, на базі якого проводилися експерименти
- Був проведений детальний аналіз отриманих моделей, їх ефективності в рамках поставленої задачі та їх здатності до узагальнення.

Перспективи досліджень

У подальшому планується:

- Дослідження більш складної задачі, в рамках якої агент отримує команди на природній мові, що описують об’єкт, який він повинен знайти
- Дослідження задачі, в якій агент шукає відповідь на задане питання у форматі «Чи стоїть стілець на кухні?», «Якого кольору шпалери у вітальні?»
- Розробка більш узагальненої бібліотеки для полегшення дослідження цих задач, та впровадження їх в реальні застосунки

Дякую за увагу!