Стратегічний захист: застосування навчання підкріплення до ігор Tower Defense з оцінкою DQN

Автор: Шпіганович Владислав Олегович

Студент БП ПМ-4

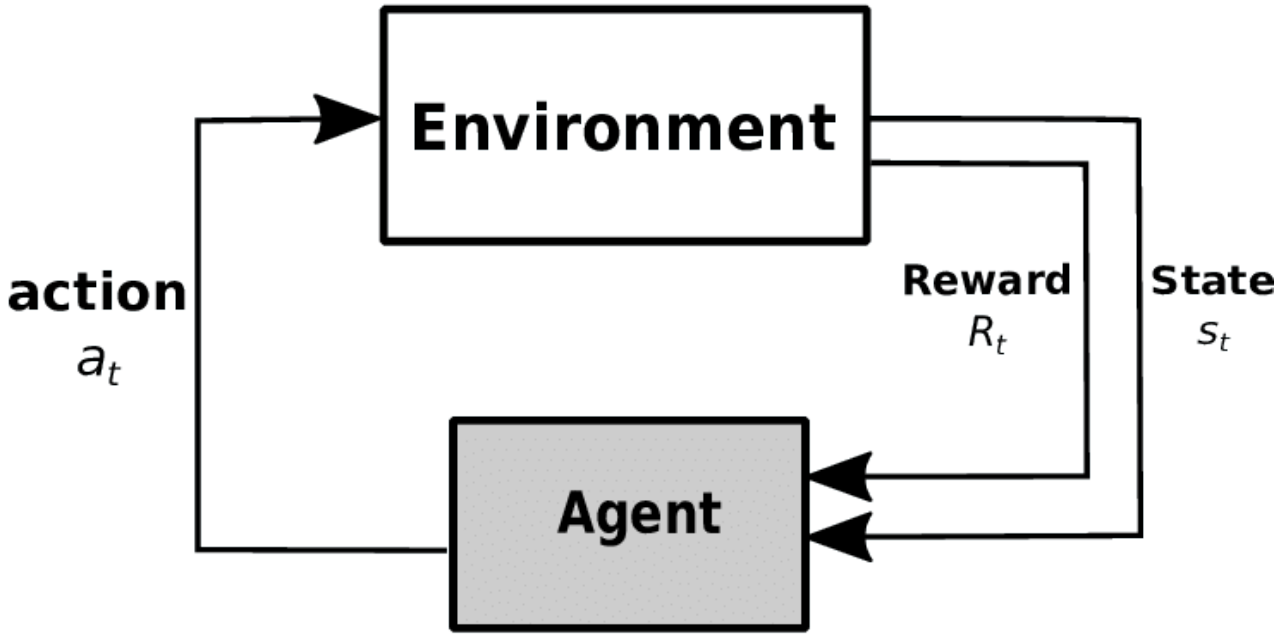
Науковий керівник: к.ф.-м. н. Козеренко С. О.

Постановка задачі

1. Створення RL середовища у стилі Tower Defense

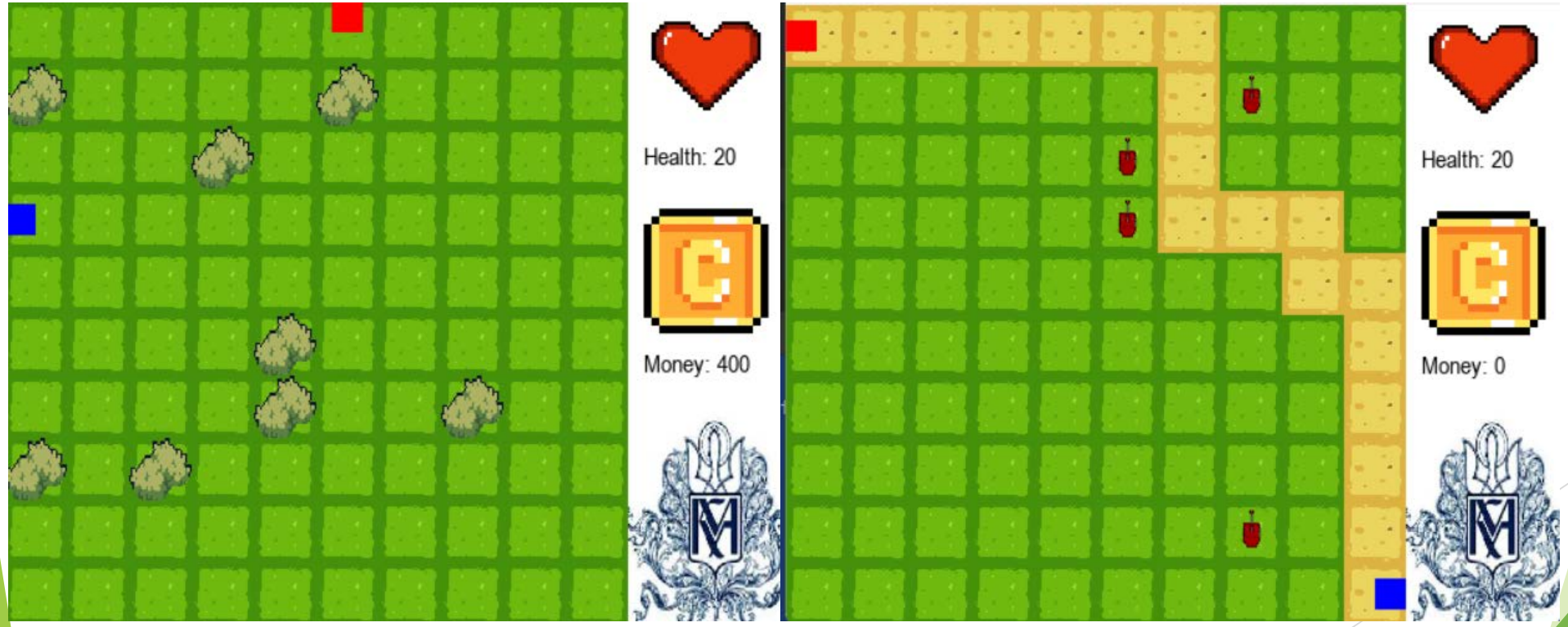
2. Тренування моделі для гри

3. Аналіз результатів



Реалізовані середовища

- Типи: пустий світ, зі сталим шляхом, з перешкодами (від 7 до 12)
- Розмір сітки: 10x10; початкові значення здоров’я/монет: 20/400



Правила середовища

- Гра починається з 20 життів та 200 монет
- Є 4 види ворогів
- Агент робить 2 дії в секунду
- Монети заробляються за вбивство противників і дається різна кількість в залежності від типу
- Супротивники з’являються один раз в секунду випадково зі списку хвилі
- Їх список — випадковий або наперед заданий в налаштуваннях
- Шлях до кінцевої точки рахується через поле потоку
- Агенту даються нагороди в залежності від подій що відбулись між його вибором дій

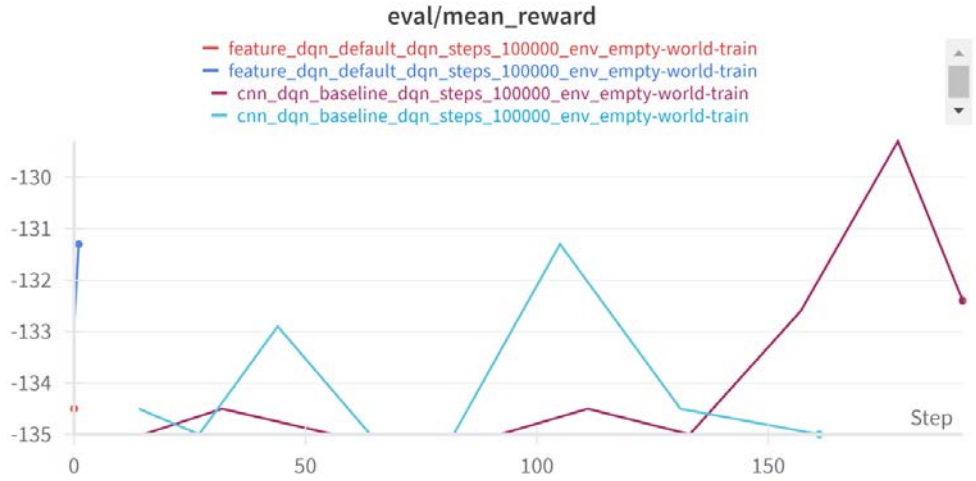
Простори дій та спостережень

- Дії: дискретний простір розмірністю 2*$tower\_types$ x $width$ x $height$.
- Спостереження:
  1. Купа з 4 зображень у сірому відтінку розмірами 200x250
  2. Матриця $< width, height, f_c >$, де
     $$f_c =< t_{cell}, t_{tower}, level_{tower}, count_{enemy}, health_{enemies} >$$

Таким чином, для зроблених середовищ матриця приймала розмірність 10x10x5.

Експерименти - DQN

- Всі виконувалися у Python, використовуючи бібліотеку stable-baselines з застосуванням відеокарти RTX 4090
- DQN – провалились. Великі негативні нагороди,
- Спробували застосувати PPO з застосуванням маскування дій

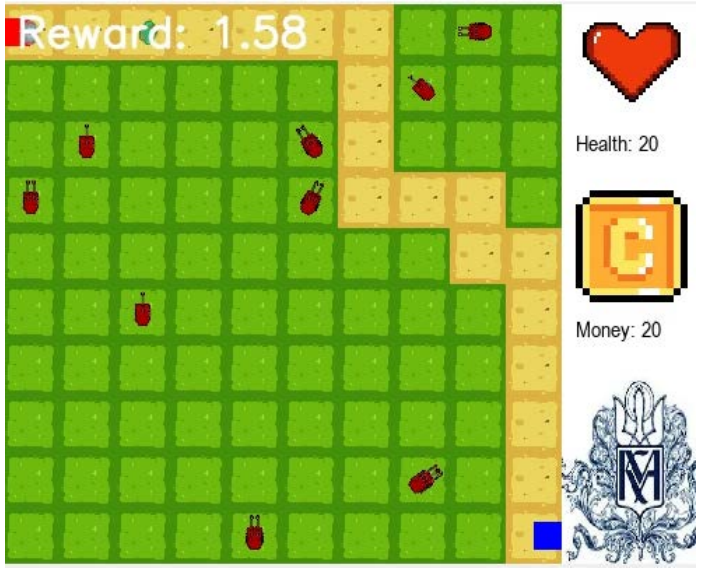


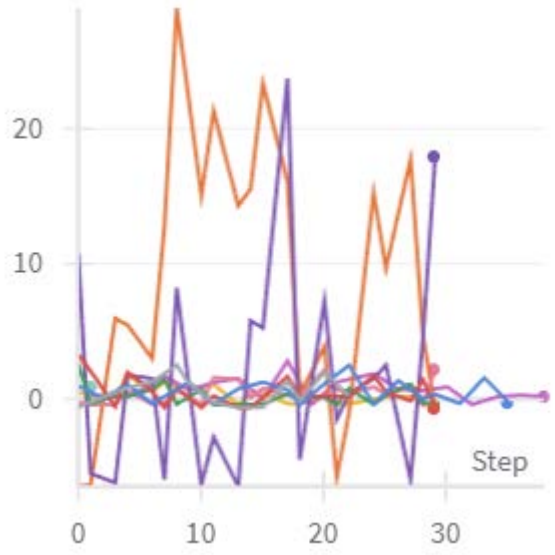
PPO

- Для середовища з дорогою агент інколи може випадково обрати непогані позиції і виграти, але залежить від згенерованого положення дороги.

  Але є багато сумнівних ходів

- Навчання провалюється на середовищах пустого світу та з перешкодами





Аналіз

Як DQN, так і PPO мають набори дій, які схильні робити незалежно від стану середовища. Значення гіперпарпараметрів, які відповідають за дослідження середовища, не допомогли.

Найімовірнішою причиною є високий дискретний простір дій у 201 значення. З цього випливають наступні проблеми:

1. Дослідження – стає важко ефективно дослідити всі можливі дії через їх кількість та факт того, що більшість часу вони не доступні.

2. Мережі важко вивчити репрезентації між діями та нагородами через велику кількість дій. Окрім цього, нагороди надходять з часом, що теж ускладнє процес.

Потенційні рішення: зменшення простору дій або декомпозиція дії та використання ієрархічної моделі з декількома головами

Майбутня робота

- Робота над покращенням архітектури моделей для алгоритмів для
- Добавлення нових видів суперників, веж, ділянок карти.
- Створення нових середовищ

Висновки

- 1. Було зроблено середовище для навчання з підкріпленням на основі жанру ігор Tower Defense
- 2. На основі цього середовища були випробувані алгоритми DQN та PPO
- 3. Був зроблений аналіз результатів

Дякую за увагу!