Środowisko wieloagentowe uczenia przez wzmacnianie inspirowane ekosystemami naturalnymi

Autor: Bartłomiej Tarcholik

Promotor: dr hab. Adrian Horzyk, prof. AGH

Cel pracy

Głównym celem pracy była implementacja inspirowanego naturalnymi ekosystemami środowiska wieloagentowego uczenia przez wzmacnianie, w którym agenci reprezentujący ryby walczą o przetrwanie.

Praca miała na celu położenie podwalin pod większy projekt, który umożliwiłby symulację rzeczywistych środowisk wodnych oraz pozwoliłby na badania nad zachowaniem organizmów w nich żyjących.

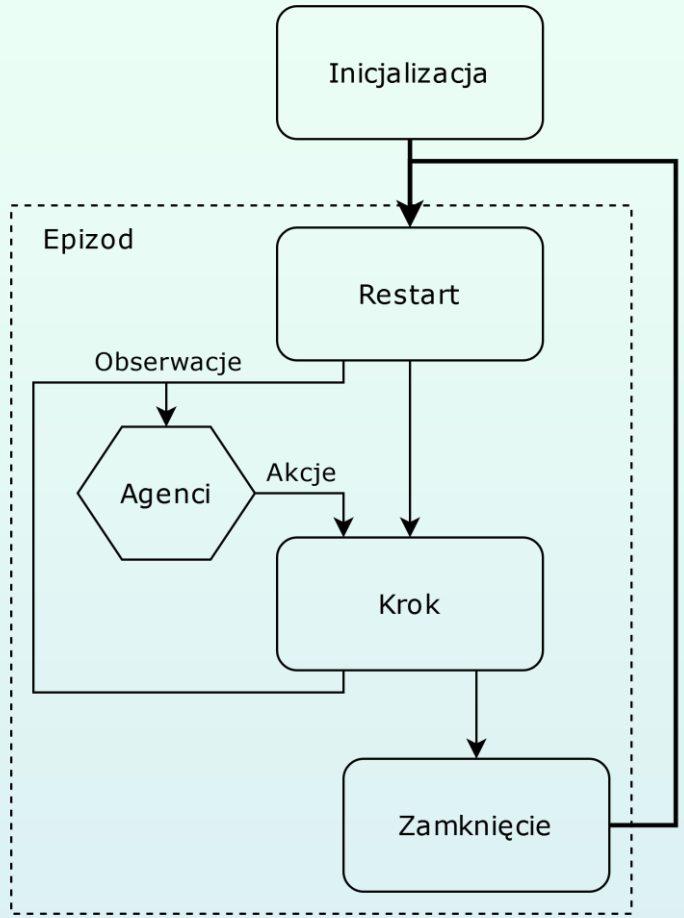
Założenia projektu

- Skalowalność – mapa może mieć dowolną wielkość, nie wpływa to na złożoność obliczeń
- Zmienne parametry środowiska – możliwość regulacji startowej populacji, ilości pożywienia, nagród czy zasięgu wzroku
- Nagrody i kary oparte na potrzebach życiowych – żywienie, energia na ruch, śmierć i odnoszenie ran
- Obserwacje inspirowane wzrokiem i rozumieniem przestrzennym

Cykl działania środowiska

Środowisko składa się z 4 głównych funkcji:

- Inicjalizacja – wczytanie pliku mapy, inicjalizacja stałych
- Restart – odnowienie danych agentów, wygenerowanie mapy i pozycji agentów oraz pożywienia
- Krok – wykonanie akcji każdego agenta, obliczenie nagród i kar za te akcje
- Zamknięcie – usunięcie okien wizualizacji



Rys. 1: Cykl działania środowiska

Działanie agenta

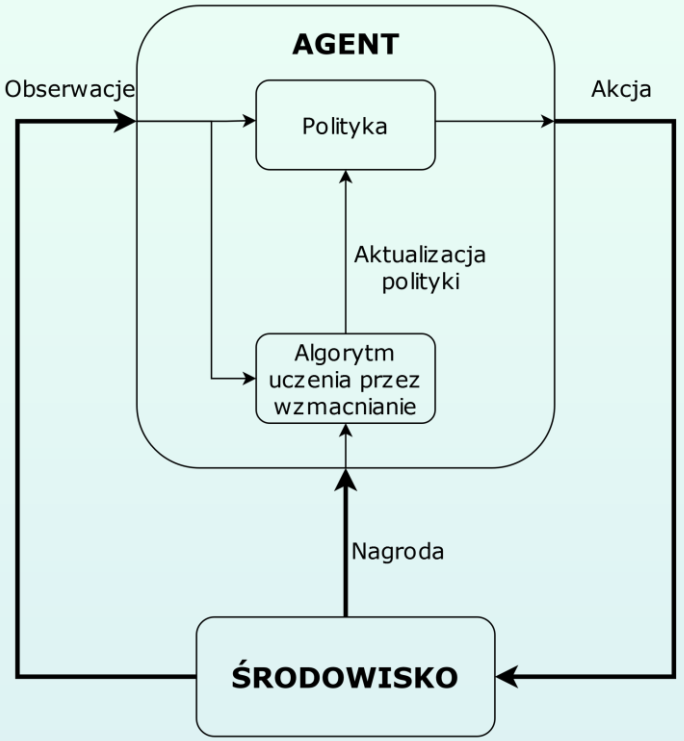
Nagrody i kary

Algorytm uczenia motywowany jest maksymalizacją nagród:

- Za bycie blisko pożywienia
- Za zdobycie pożywienia

Oraz minimalizacją kar:

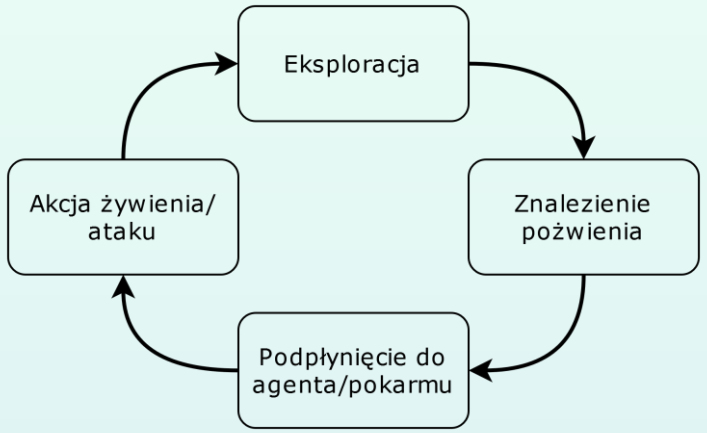
- Za ruch
- Za wykonanie nielegalnej akcji
- Za odniesienie obrażeń lub śmierć z głodu/ran



Rys. 2: Przepływ danych w trakcie szkolenia

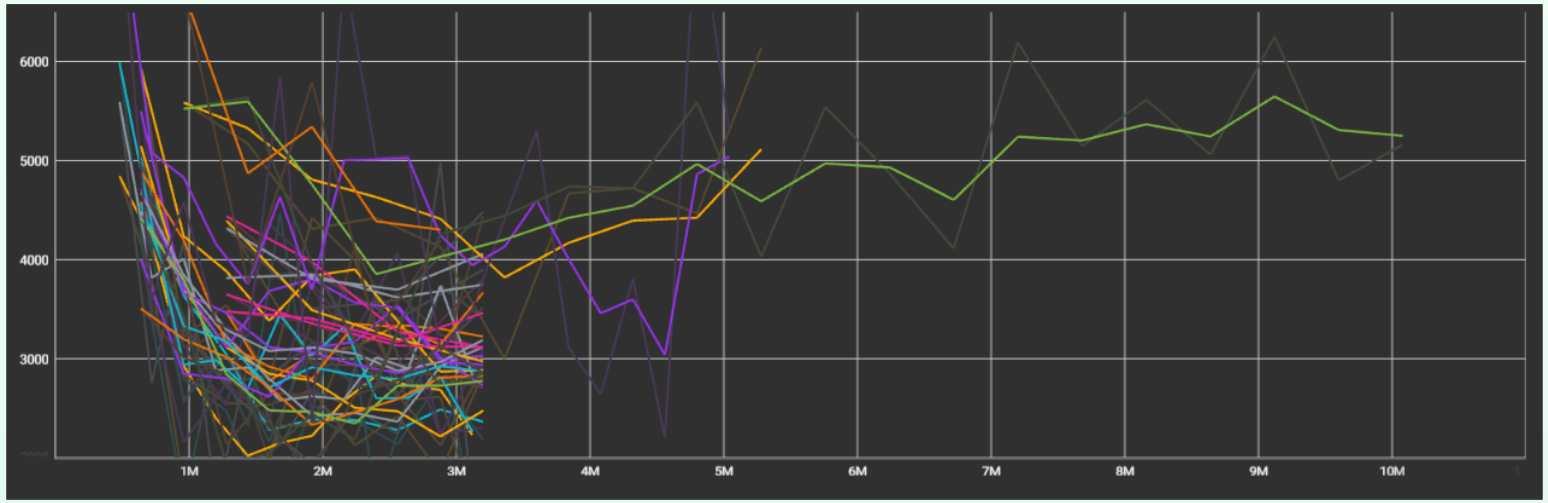
Cykl zachowania agenta

- Główną motywacją agenta jest zdobywanie punktów pożywienia, co przedłuża jego życie.
- Zdobywać je można poprzez jedzenie pokarmu lub atakowanie innych agentów.
- Utrata punktów życia lub pożywienia oznacza śmierć agenta.



Rys. 3: Cykl zachowania agenta

Wnioski ze szkolenia

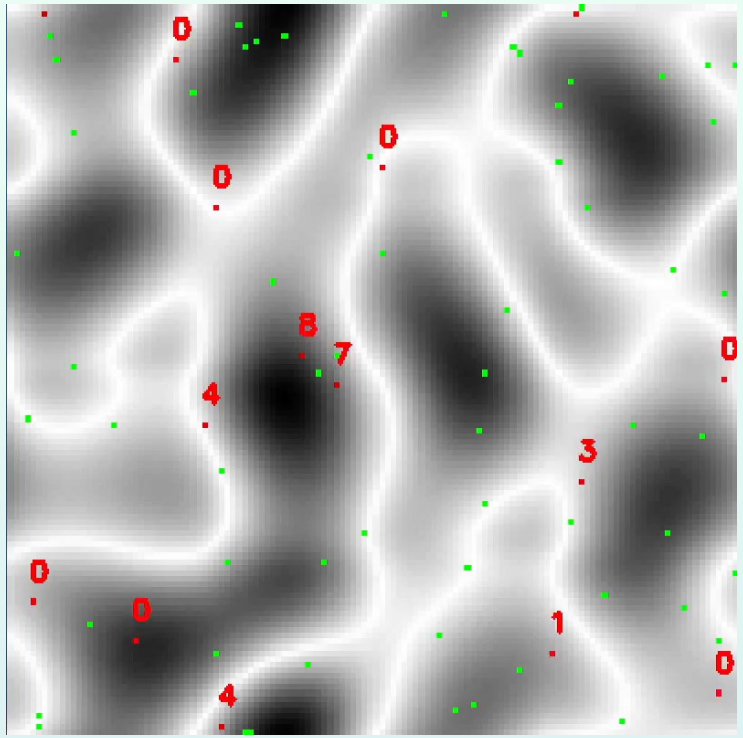


Rys. 4: Przebieg value loss dla drugiej partii szkolenia

- Wyszkolono ponad 100 polityk używając algorytmu PPO w implementacji Stable-Baselines3
- Szkolenie przebiegało w partiach, każda następna wyciągała wnioski z poprzedniej
- Optymalną długością uczenia było 3-5 milionów kroków
- Najważniejszymi parametrami był współczynnik dyskontowy oraz ilość analizowanych na raz przez algorytm danych

Przykładowa wyszkolona polityka

- Agenci świadomie podpływają w okolice pożywienia
- Relatywnie skutecznie wykonują akcję żywienia
- Okazyjnie atakują innych agentów
- Powtarzają akcje gdy nie są w stanie znaleźć lepszego rozwiązania



Rys. 5: Wizualna reprezentacja działania

Podsumowanie i przyszłość projektu

Osiągnięcia pracy

- Zaimplementowano w pełni sprawne środowisko MARL zgodne z aktualnymi standardami
- Umożliwiono zmianę parametrów środowiska przez użytkownika
- Algorytm środowiska symuluje interakcję organizmów żywych
- Wytrenowano ponad 100 polityk, wyłoniono oraz opisano 3 najlepsze polityki zachowania

Przyszłość projektu

- Możliwość rozszerzenia o nową funkcjonalność agentów i środowiska
- Podwaliny pod pełną symulację dowolnego środowiska wodnego, np. interakcji różnych gatunków w akwarium lub jeziorze
- Osobna maska akcji pozwoliłaby na kompletną eliminację problemu akcji nielegalnych