Politechnika Poznańska
Wydział Informatyki i Zarządzania
Instytut Informatyki

Praca dyplomowa magisterska

UCZENIE PRZEZ DEMONSTRACJĘ NA PODSTAWIE INFORMACJI OBRAZOWEJ W ŚRODOWISKU 3D

Wojciech Kopeć, 101675

Promotor
dr inż. Krzysztof Dembczyński

Poznań, 2017 r.

Tutaj przychodzi karta pracy dyplomowej;
oryginał wstawiamy do wersji dla archiwum PP, w pozostałych kopiach wstawiamy ksero.

Spis treści

Rozdział 1

Wstęp

Rozdział 2

Wstęp teoretyczny

2.1 Wstęp

W poniższym rozdziale przedstawione zostaną uczenie ze wzmocnieniem i uczenie przez demonstrację, stojące za nimi motywacje oraz podstawy teoretyczne. Omówione będą najważniejsze metody oraz aktualny stan wiedzy. Na końcu zaprezentowane zostanie środowisko symulacyjne wykorzystane w poniższej pracy oraz słownik używanych pojęć.

2.2 Proces decyzyjny Markowa

Środowisko, w którym porusza się program jest matematycznie zamodelowane jako proces decyzyjny Markowa *(ang. Markov decision process, MDP)* [Bellman, 1954]. Oznacza to, że środowisko ma w każdym momencie czasu określony stan i umożliwia wykonanie określonych akcji, za które agent może otrzymać pozytywną lub negatywną nagrodę. Rezultatem wykonania akcji w danym stanie jest przejście do następnego stanu. Zakłada się, że nowy stan jest zależny tylko od stanu poprzedniego i wykonanej akcji. Takie środowisko ma właściwość braku pamięci*lub inaczej, właśność Markowa*. Celem agenta jest zgromadzenie nagród o jak największej sumie wartości. Im bardziej odległe w przyszłości nagrody, tym mniej są wartościowe (są dyskontowane).

Formalnie, proces decyzyjny Markowa definiujemy jako krotkę $(S, A, T(\cdot,\cdot), R(\cdot,\cdot), \gamma)$, gdzie:

- S jest skończonym zbiorem możliwych stanów środowiska,
- A_s jest skończonym zbiorem możliwych akcji wykonywalnych w stanie s,
- $T_a(s, s')$ - funkcja przejść, która reprezentuje prawdopodobieństwo trafienia do stanu s' po wykonaniu akcji a w stanie s,
- $R_a(s, s')$ - funkcja nagrody, która określa nagrodę (lub wartością oczekiwaną nagrody, obie mogą być negatywne) otrzymywaną po wykonaniu akcji a w stanie s i trafieniu na skutek tego do stanu s',
- $\gamma \in [0, 1]$ jest współczynnikiem dyskontowym, obniżającym wartość nagród uzyskanych w przyszłości.

Celem jest maksymalizacja zdyskontowanej sumy nagród

$$\sum_{t=0} \gamma^t R(s_t, s_{t+1}),$$

gdzie kolejne t są kolejnymi momentami czasowymi. Ponadto:

- Polityką (strategią) π, realizowaną przez agenta, nazywamy funkcję $\pi : S \to A$, która określa, jak agent powinien się zachować w danym stanie w celu osiągnięcia maksymalnej możliwej nagrody.

- Funkcja użyteczności $U(s)$ lub wartości *(ang. Value)* $V(s)$ określa maksymalną oczekiwaną nagrodę, jaką agent może osiągnąć znajdując się stanie s i postępując dalej zgodnie z aktualną polityką. Poniższe równanie oparte jest na równaniu Bellmana [Bellman, 1954].

$$U(s) = V(s) = \max_{a \in A(s)} \sum_{s'} T_a(s, s')(R_a(s, s') + \gamma U(s'))$$

- Funkcja Q $Q(s, a)$ określa maksymalną oczekiwaną nagrodę, jaką agent może osiągnąć wykonując w stanie s akcję a i postępując dalej zgodnie z aktualną polityką.

$$Q(s, a) = \sum_{s'} T_a(s, s')(R_a(s, s') + \gamma \max_{a' \in A(s)} Q(s', a'))$$

W badanym problemie środowisko VizDoom jest *częściowo obserwowalnym procesem decyzyjnym Markowa*, co oznacza, że stan obserwowany przez agenta nie zawiera pełnej informacji o środowisku. Środowisko to jest stochastyczne, co oznacza że skutki działań agenta nie są deterministyczne - wielokrotne wykonanie tej samej akcji w tym samym stanie może przynieść różne rezultaty.

Znając funkcje przejść możliwe jest iteracyjne określenie optymalnej polityki działania agenta.

2.3 Uczenie ze wzmocnieniem

W przypadku, kiedy funkcja przejść $T_a(s, s')$ albo funkcja nagród $R_a(s, s')$ nie jest znana (albo wielkość przestrzeni stanów S sprawia, że analityczne podejście nie jest możliwe), mamy do czynienia z *uczeniem ze wzmocnieniem (ang. Reinforcement learning, RL)*.

Intuicyjnie, uczenie ze wzmocnieniem opisuje sytuację, w której agent porusza się w nieznanym środowisku. Na podstawie obserwowanych rezultatów swoich działań buduje wiedzę o środowisku, pozwalającą na określenie strategii działania optymalną dla danej wizji środowiska. Postępując według tej strategii zdobywa dalsze doświadczenia, pozwalające dalej uaktualniać wiedzę o środowisku i politykę agenta.

2.3.1 Uczenie ze wzmocnieniem a uczenie nadzorowane

W rozdziale 2.2 określono politykę π jako funkcję $\pi : S \to A$, określającą optymalną akcję do wykonania a dla każdego stanu s. Odpowiednikiem w uczenia nadzorowanego byłoby określenie π jako klasyfikatora $S \to A$. Do uczenia ze wzmocnieniem nie stosuje się jednak technik uczenia nadzorowanego, ponieważ, bez rozwiązania całego problemu, nigdy nie są znane „poprawne" akcje a dla danych stanów s. Mimo, że środowisko dostarcza czasem informacji zwrotnej w postaci nagród lub kar, to wykonanie danej akcji w danym stanie jest najczęściej konsekwencją całej poprzedniej sekwencji ruchów - określenie, który z ruchów w sekwencji był faktycznie kluczowy dla uzyskania określonego rezultatu jest nietrywialne.

Przykładowo, w partii szachów zwycięski ruch jest najczęściej konsekwencją konkretnych zagrań lub błędów popełnionych wiele ruchów wcześniej - zadaniem algorytmu uczącego jest określenie, które z ruchów były decydujące dla końcowego wyniku. Ruchy, które doprowadzają do zwycięstwa

mogą krótkodystansowo przynosić straty (np. poświęcenie figury) - algorytm uczący powinien zrozumieć, że mimo negatywnej informacji zwrotnej dany ruch był porządany.

2.3.2 Zalety i zastosowanie

Uczenie ze wzmocnieniem jest narzędziem, które świetnie sprawdza się w sytuacjach, w których środowisko jest zbyt skomplikowane, żeby analitycznie znaleźć optymalną politykę działania. Dzięki temu, że model przejść i model nagród opisane są rozkładami prawdopodobieństwa, a nie stałymi zależnościami, uczenie ze wzmocnieniem radzi sobie bez problemu z modelowaniem zachowań w bardzo niepewnym świecie. Dzięki współczynnikowi dyskontowemu γ, możliwe jest balansowanie pomiędzy optymalizacją krótko i długoterminowych zysków.

Najważniejsza jest jednak możliwość działania bez żadnej wiedzy i silnych założeń na temat środowiska, a którym znajduje się agent. RL zakłada brak wiedzy o modelu świata, a wszystkie informacje czerpane są z doświadczeń na temat interakcji ze środowiskiem. Jest to kluczowe założenie, ponieważ dla wielu praktycznych problemów, które adresuje RL stworzenie dokładnego modelu świata jest niemożliwe (np. stworzenie dokładnego matematycznego modelu aerodynamiki i zachowania śmigłowca), albo mimo znajomości modelu matematycznego przestrzeń możliwych stanów jest zbyt wielka, by analitycznie otrzymać rozwiązanie (np. szachy, go).

Uczenie ze wzmocnieniem stosuje się z powodzeniem do sterowania robotami [Mataric, 1994] i windami[Crites and Barto, 1996], grania w gry planszowe (backgammon[Tesauro, 1992], go [Silver et al., 2016], warcaby [Samuel, 1959]) i gry komputerowe [Mnih et al., 2015].

2.4 Metody

Podejścia stosowane do uczenia ze wzmocnieniem możemy podzielić na trzy rodzaje, w zależności od typu informacji na której bazuje agent [Jaskowski, 2016]. [POPRAWIĆ NA ORYGINALNĄ REFERENCJĘ]

1. Agent z polityką - uczy się polityki $\pi : S \to A$. Przykłady:
   - algorytmy ewolucyjne,
   - uczenie przez demonstację.
2. Agent z funkcją użyteczności U. Przykłady:
   - adaptatywne programowanie dynamiczne *(ang. adaptative dynamic programming, ADP)*,
   - metoda różnic czasowych *(ang. temporal difference learning, TDL)*.
3. Agent z funkcją użyteczności Q. Przykłady:
   - Q-learning,
   - SARSA *(ang. State, Action, Reward, State, Action)*

Poniżej przedstawione zostaną najważniejsze metody.

2.4.1 Metoda różnic czasowych

Metoda różnic czasowych *(ang. Temporal difference learning, TDL)* [Sutton, 1988] opiera się na uaktualnianiu stanu wiedzy agenta na podstawie różnicy pomiędzy spodziewanym a zaobserwowanym wynikiem.

Agent trafiający do stanu s' po wykonaniu akcji a w stanie s może uaktualnić stan swojej wiedzy: $U(s) \leftarrow U(s) + \alpha(R(s,s') + \gamma U(s') - U(s))$ lub $Q(s,a) \leftarrow Q(s,a) + \alpha(R(s,s') + \gamma(\max_{a'} Q(s',a') - Q(s,a)))$, gdzie α jest współczynnikiem prędkości uczenia. Jeżeli α w odpowiedni sposób zmniejsza się w czasie, to TDL gwarantuje zbieżność do optimum globalnego [Jaskowski, 2016].[POPRAWIĆ NA ORYGINALNĄ REFERENCJĘ]

2.4.2 Funkcja U, Q i SARSA

- Funkcja U (patrz: rozdział 2.2) opisuje użyteczność stanu,
- Funkcja Q (patrz: rozdział 2.2) opisuje użyteczność wykonania danej akcji w danym stanie,
- SARSA stanowi wariację metody Q. W Q-learningu wartość funkcji Q jest aktualizowa na podstawie wartości Q dla najlepszej akcji do wykonywania w stanie s' ($\gamma \max_{a'} Q(s',a') - Q(s,a)$)), natomaist w SARSIE na podstawie wykonanej przez agenta akcji ($\gamma(Q(s',a') - Q(s,a))$)), czyli przebytej przez agenta trajektorii $s \rightarrow a \rightarrow s' \rightarrow a'$. Aktualizacja TD w SARSA-ie wygląda następująco:

  $$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s,s') + \gamma(Q(s',a') - Q(s,a)))$$

  SARSA może dla niektórych problemów zachowywać się nieznacznie lepiej niż Q-learning, ale w większości przypadków będzie się uczyła wolniej bez wpływu na jakość agenta.

Mimo podobnych wzorów i definicji nauka funkcji Q ma jedną, diametralną przewagę na nauką funkcji U - funkcja Q nie wymaga znajomości modelu świata do wyboru najlepszej akcji do wykonania. Zbiór dostępnych akcji A jest znany agentowi. Przy wyborze najlepszej akcji a w stanie s:

- Agent z funkcją Q wybiera akcję $a = argmax_{a \in A} Q(s,a)$.
- Agent z funkcją U wybiera akcję, która maksymalizuje $U(s')$ - wartość stanu, do którego trafi agent: $a = argmax_{a \in A} \sum_{s'} T_a(s,s')U(s')$. Obliczenie tego wyrażenia wymaga znajomości modelu przejść $T_a(\cdot,\cdot)$, czyli modelu świata. Można przyjąć, że dla trudniejszych i realnych problemów model świata nie jest dostępny.

Z tego powodu większość wiodących rozwiązań w dziedzinie uczenia ze wzmocnieniem oparta jest na Q-learningu. Dalsza część pracy przyjmuje Q-learning jako obowiązującą metodę rozwiązywania problemu uczenia ze wzmocnieniem.

Niezależnie od metody, dla sensownej wielkości problemów uczenie ze wzmocnieniem jest wymagające obliczeniowo i czasowo. Mimo wspomagających agenta technik, nauka sprowadza się najczęściej do interakcji ze środowiskiem metodą prób i błędów - potrzeba wiele prób i błędów, zanim agent zacznie pojmować zasady rządzące środowiskiem w którym się znajduje, a potem dużo dalszych zanim znajdzie dla danego środowiska satysfakcjonująco skuteczną politykę działania. [DOKOŃCZYĆ?]

2.5 Aproksymatory i głębokie sieci Q

Zdefiniowana w podrozdziale 2.4 przykładowa reguła aktualizacji wartości funkcji Q wygląda następująco:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s,s') + \gamma(max_{a'} Q(s',a') - Q(s,a)))$$

Wynika z niej, że po wykonaniu ruchu wartość funkcji Q dla poprzedniego stanu aktualizujemy na podstawie otrzymanej nagrody i wartości funkcji Q dla stanu aktualnego. Oznacza to, że dla każdego stanu, który analizujemy, konieczna jest znajomość jego wartości funkcji Q dla wszystkich możliwych akcji. Oznacza to, że dla dokładnego przedstawienia funkcji $Q(s, a)$ konieczne jest zapamiętanie $|S| \cdot |A|$ wartości. Co więcej, aby uzyskać sensowne wartości tej funkcji konieczne jest odwiedzenie każdego ze stanów wiele razy, zanim aktualizowana stopniowo wartość funkcji Q będzie bliska prawdziwej. Wiele z tych stanów jest też bardzo podobnych do siebie nawzajem, więc wiedza wyniesiona dla jednego stanu powinna się w pewien sposób generalizować na podobne stany.

Backgammon, jedna z gier planszowych służących jako benchmark algorytmów uczenia ze wzmocnieniem, ma 10^{20} możliwych stanów, a szachy 10^{40}. Jeden obraz 90x60 pixeli w skali szarości, używany jako zapis stanu w problemie rozwiązywanym w ramach poniższej pracy może przyjąć 256^{5400} różnych kombinacji. Wiele realnych problemów opisanych jest wartościami ciągłymi, nie dyskretnymi, a praktyczna liczba ich możliwych stanów rośnie wykładniczo wraz ze wzrostem dokładności pomiaru.

2.5.1 Aproksymatory funkcji Q

Rozważanie i zapamiętanie każdego stanu z osobna dla bardziej skomplikowanych problemów jest niemożliwe i niepraktyczne ze względu na liczbę możliwych stanów i podobieństwo wielu stanów. Rozwiązaniem jest wykorzystanie *aproksymatora funkcji Q* - niestablicowanej, parametrycznej funkcji pary (stan,akcja) $\hat{Q}_\theta(s, a)$, gdzie θ jest wektorem parametrów funkcji.

Aproksymator (za [Jaskowski, 2016]):

- musi być łatwo obliczalny,
- kompresuje dużą przestrzeń stanów w znacznie mniejszą przestrzeń parametrów,
- uogólnia wiedzę na temat podobnych stanów.
- w większości przypadków przyspiesza uczenie w stosunku do wersji stablicowanej ze względu na uogólnianie wiedzy

Jako jedne z pierwszych i prostszych aproksymatorów stosowano funkcje liniowe, opierające się na ręcznie zdefiniowanych cechach: $\hat{Q}_\theta(s, a) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$, gdzie wektor $x = (x_1, x_2, \ldots, x_n)$ jest wektorem cech. Przykładem zastosowania może być gra w warcaby, opisana w [Samuel, 1959]. Zaletami liniowego aproksymatora są prostota i łatwość interpretacji, a także szybkość obliczania i nauki. Dalszym krokiem było wykorzystanie sieci neuronowych jako aproksymatorów w grze w Backgammona [Tesauro, 1992]. W pierszej wersji algorytmu wykorzystano ręcznie zaprojektowane cechy, w kolejnych wykorzystano prawie surową informację o rozkładzie pionków na planszy. Sieć neuronowa jest bardziej skomplikowanym i trudniejszym do nauczenia aproksymatorem, ale jest w stanie zamodelować znacznie bardziej złożone funkcje.

2.5.2 Uczenie na podstawie surowych danych obrazowych - Atari 2600

Jednym z największych przełomów uczenia ze wzmocnieniem ostatnich lat była praca [Mnih et al., 2015], w której autorzy wykorzystali głębokie sieci neuronowe do stworzenia agenta potrafiącego grać na ludzkim poziomie w klasyczne gry z Atari 2600, wykorzystując jako reprezentację stanu jedynie surowy zapis obrazu 2D. Dotychczas, jak w poprzednich przykładach, algorytmy uczenia ze wzmocnieniem opierały się na manualnie stworzonej reprezentacji stanów. W [Mnih et al., 2015]

pokazano, że możliwe jest stworzenie rozwiązania, które samo będzie potrafiło ekstrahować wysokopoziomowe cechy z niskopoziomowych danych. Zaproponowana architektura, jak również pomysłowe usprawnienia zwiększające stabilność uczenia zaproponowane w artykule, a opisane w rozdziale 2.6 stanowią obecnie podstawę i punkt odniesienia dla większości dalszych badań na temat uczenia ze wzmocnieniem.

Jako aproksymator funkcji Q wykorzystano głęboką sięć neuronową. Z tego powodu opisywane podejście określa się często skrótem DQN *(ang. Deep Q Network)*, czyli głęboka sieć Q. Zastosowana architektura wygląda następujaco:

[RYSUNEK SIECI]

Analogicznie jak w obecnie stosowanych architekturach rozpoznawania obrazu, pierwsze warstwy sieci to warstwy konwolucyjne, które wykrywają kolejno nisko i wysokopoziomowe cechy obrazu. Dalsze warstwy, w pełni połączone, łączą informacje z warstw konwolucyjnych we wnioski na temat stanu świata, na podstawie których następne warstwy mogą określić wartość funkcji Q.

2.6 Q-learning - usprawnienia

Skuteczność i stabilność Q-learningu może zostać drastycznie polepszona dzięki zastosowaniu następujących technik.

2.6.1 Pamięć powtórek

Szkielet uczenia ze wzmocnieniem opiera się na zbieraniu doświadczeń i uaktualnianiu na ich podstawie stanu wiedzy agenta. W praktyce, doświadczenia zbierane bezpośrenio po sobie są silnie skorelowane - przykładowo agent uczący się na podstawie obrazu jazdy samochodem w kolejnych klatkach widzi niemal identyczne obrazy i wykonuje najczęsciej te same akcje. Oznacza to, że aktualizowanie wiedzy agenta na podstawie nowych doświadczeń, czy to pojedynczo czy w paczkach, będzie skutkować funkcją obciążoną w kierunku tych, nowych doświadczeń.

Aby temu zapobiec w [Mnih et al., 2015] zaproponowano metodę pamięci powtórek *(ang. Replay memory)*. Metoda opiera się na zapamiętywaniu znacznej ilości najnowszych doświadczeń. Po każdym kroku nowe doświadczenia dodawane są do pamięci (w przypadku braku miejsca zastępując najstarsze), a następnie z pamięci wybierana jest losowa próbka doświadczeń, na podstawie których aktualizowana jest wiedza agenta. Dzieki tej technice dane użyte do nauki przez agenta są nieskorelowane i niezależne. Dodatkowo, dzięki dostępowi do starszych danych agent jest mniej podatny na obniżanie jakości gry na skutek krótkotrwałych spadków wyników.

Dalsze rozszerzenia metody mają na celu np. priorytetyzowanie używania do nauki najważniejszych doświadczeń [Schaul et al., 2015].

2.6.2 Zamrażanie docelowej sieci

Podobnie jak pamięć powtórek, zamrażanie docelowej sieci *(ang. Target network freezeing / fixed target network)* zostało zaproponowane w [Mnih et al., 2015] i służy zmniejszeniu skutków obciążenia rozkładu danych uczących zebranych przez agenta, a wynikającego ze sposobu zbierania próbek. Zamrażanie sieci zakłada utrzymywanie dwóch funkcji Q - starej i nowej. Agent działa na podstawie nowej funkcji, ale wartości Q „docelowych" stanów używanych do aktualizacji wartości Q (2.4.1) pobierane są ze starej funkcji. Co jakiś czas do starej funkcji Q przepisywana jest nowa funkcja Q.

Technika ma na celu zniwelowanie oscylacji i ustabilizowanie zachowań agenta. Dzięki wykorzystaniu „zamrożonych" wartości do nauki funkcji Q zerwane jest sprzężenie zwrotne pomiędzy zebranymi danymi a wartościami docelowymi.

2.6.3 Kształtowanie

W wielu zadaniach stawianych przed uczeniem ze wzmocnieniem osiągnięcie celu jest bardzo trudne, a agent dostaje nagrody dopiero po osiągnięciu stanów terminalnych, albo na zaawansowanym etapie zadania. Agent uczący się na podstawie prób, błędów i losowych akcji nie jest najczęściej w stanie wykonać wystarczająco dużej części zadania, żeby dostać informację zwrotną w postaci nagrody, a więc nie ma jak się uczyć lub uczenie następuje bardzo wolno.

Kształtowanie *(ang. Shaping)* ([Mataric, 1994]) zakłada sztuczne wprowadzenie do środowiska dodatkowych nagród, które agent będzie dostawał po wykonaniu etapów pośrednich zadania. Przykładowo, przy grze w szachy, w której agent dostaje nagrodę tylko za wygraną lub przegraną (1 lub -1) można byłoby wprowadzić nagrodę 0.1 za zbijanie figur przeciwnika.

Kształtowanie wymaga możliwości ingerencji w środowisko albo percepcję agenta (rozpoznawanie, kiedy agent powinien dostać sztuczną nagrodę i ingerowanie w odczyty nagrody dokonywane przez agenta). Co ważniejsze, wymaga wiedzy eksperckiej na temat zadania wykonywanego przez agenta (możliwość określenia sensownych etapów zadania, na których agent miałby dostać sztuczną nagrodę) i wiedzy na temat środowiska, w którym agent się porusza (wysokość sztucznej nagrody musi być dopasowana do prawdziwych nagród, które może dostawać agent). Dodatkowo, kroki określone przez eksperta mogą wymuszać nieoptymalną politykę działania i powstrzymać agenta przed odkryciem optymalnych strategii.

2.7 Eksploracja

W uczeniu ze wzmocnieniem agent posiada umiejętność uczenia się na podstawie odbytych doświadczeń. Na początku każdej nauki jest jednak zupełnie nieświadomy zasad świata, w którym się znajduje i nie jest w stanie podejmować sensownych działań. Konieczna jest metoda pozwalająca na zdobywanie nowych doświadczeń przy jednoczesnej możliwości szlifowania i ulepszania opracowanych wcześniej przez agenta sposobów. Proces nakładania agenta do zbadania nieznanych jeszcze obszarów przestrzeni stanów nazywany jest eksploracją.

Wybieranie odpowiednich proporcji pomiędzy eksploracją nowych stanów a zgłębianiem znanych jest problemem nietrywialnym i przedmiotem wielu badań.

2.7.1 Algorytm e-zachłanny

Podstawowym i często używanym podejściem do eksploracji jest algorytm $\epsilon - zachanny$, w którym agent z zadanym prawdopodobieństwem ϵ zamiast akcji optymalnej względem aktualnej polityki wykonuje akcję losową. Takie zachowanie jest mało wydajne, szczególnie kiedy optymalne zachowanie agenta wymaga zaplanowania złożonych lub dalekosiężnych planów.

2.7.2 Algorytm R-max

Prostym, ale skutecznym i posiadającym teoretyczne gwarancje zbieżności algorytmem jest zaproponowany w [Brafman and Tennenholtz, 2002] R-max, realizujący ideę optymizmu wobec niepewności. Podstawą R-maxa jest optymistyczna inicjalizacja – przed rozpoczęciem uczenia funkcja aproksymacyjna powinna zwracać maksymalną nagrodę dla wszystkich stanów i akcji. W

ramach działania agent będzie uaktualniał (czyli obniżał) spodziewaną nagrodę w odwiedzonych stanach.

Największa spodziewana nagroda będzie zwracana dla zachowań, które agent odkrył już jako zyskowne i dla zachowań jeszcze nieodkrytych (dla których funkcja aproksymacyjna nie jest jeszcze poprawiona). Ten prosty zabieg powoduje, że algorytmy uczenia ze wzmocnieniem naturalnie balansują pomiędzy eksploracją i intensyfikacją przeszukiwania bez dodatkowych modyfikacji.

Od strony teoretycznej zaletą R-maxa jest duża ogólność zastosowania – algorytm wymaga spełnienia bardzo luźnych założeń, badany proces nie musi być nawet procesem decyzyjnym Markowa.

2.7.3 Przewidywanie przejść za pomocą autoenkodera

W [Stadie et al., 2015] autorzy zaproponowali rozwiązanie, które pozwala ocenić, w jakim stopniu odwiedzony stan jest dla agenta nowością. Opiera się ono na stworzeniu generatora, którego zadaniem jest przewidywanie, jaki stan osiągnie agent po wykonaniu danej akcji w danym stanie. Predykcja porównywana jest z faktycznie osiągniętym stanem, a wielkość błędu jest wyznacznikiem nowości stanu – im większy błąd predykcji, tym bardziej nieznany stan, za co przyznawana jest większa nagroda eksploracyjna. Jak większość opisywanych publikacji, w [Stadie et al., 2015] rozwiązywano problem uczenia agenta grania w gry zręcznościowe na podstawie surowego obrazu z wykorzystaniem Q-learningu i głębokich sieci neuronowych.

Pierwszą kwestią do rozwiązania przy implementacji pomysłu jest metryka pozwalająca określić podobieństwo stanów. Próby predykcji wartości konkretnych pikseli opisane przez autorów nie przyniosły efektów, generując tylko szum. Zamiast tego trenowano autoenkoder oparty o głęboką sieć neuronową i wykorzystano jedną z ukrytych warstw o mniejszej liczbie jednostek tej sieci jako enkoder stanu, który przenosi surowy obraz do przestrzeni o znacznie mniejszej liczbie parametrów. Za miarę podobieństwa między stanami przyjęto odległość kartezjańską parametrów uzyskanych z zakodowania dwóch stanów. Zakodowane stany używane były do wytrenowania właściwego, prostszego aproksymatora, na podstawie błędu którego określano nowość stanu. Dla każdego przejścia między stanami przyznawano sztuczną nagrodę zależną od nowości odwiedzonego stanu.

Potencjalnym problemem związanym z tym podejściem jest to, że Q-learning stara się nauczyć funkcji, która jest niestacjonarna. Autorzy piszą, jednak, że w praktyce nie stanowiło to problemu.

2.7.4 Bootstrapowane DQN

Inną taktykę dywersyfikacji przeszukiwania przy wykorzystaniu głębokiej sieci neuronowej zaprezentowano w [Osband et al., 2016]. Podobnie jak w [Stadie et al., 2015] uczono sieć funkcji Q, jednak zamiast pojedynczej funkcji Q trenowano jednocześnie K funkcji Q, przy czym każda trenowana była tylko na podzbiorze przykładów uzyskanym za pomocą techniki bootstrapingu. Każda funkcja Q reprezentowana była przez jedną K „głów" wspólnej wielopoziomowej sieci.

Dla każdego z epizodów wybierana losowo była jedna głowa – funkcja Q i przez cały epizod agent kierował się polityką optymalną dla tej funkcji Q.

Dzięki temu zabiegowi każda z sieci Q była nauczona na podstawie nieco różnych doświadczeń i prezentowała nieco inną politykę działania. Nowe informacje o porządanych zachowaniach były prędzej czy później propagowane do każdej z głów, ale jednocześnie różnorodność zachowań była wystarczająca, żeby utrzymać eksplorację.

Autorzy raportują spowolnienie uczenia o zaledwie 20% w stosunku do normalnej, pojedynczej sieci Q, ale w przeprowadzonych w ramach tej pracy eksperymentach uczenie było znacznie wolniejsze.

2.8 Uczenie przez demonstację

Uczenie ze wzmocnieniem jest bardzo kosztowne obliczeniowo, a trudniejsze problemy wymagają przejścia setek tysięcy lub milionów klatek nauki. Znaczną część tego czasu program spędza na początkowym poznawaniu możliwości i zasad rządzących środowiskiem, albo żmudnym, stopniowym poprawianiu suboptymalnych zachowań, ewoluujących powoli w skuteczną politykę agenta.

Ludzie i zwierzęta, których zachowanie często stanowi inspirację i motywację dla nowych rozwiązań algorytmicznych, potrafią odtwarzać czynności i zachowania na podstawie samej obserwacji wykonania danej czynności przez innych. Czerpiąc z tego przypadku, wskazane jest konstruowanie algorytmów, które będą potrafiły uczyć się zaawansowanych polityk nie na bazie milionów prób i błędów, ale na bazie obserwacji kilku, lub nawet jednego, powtórzenia wykonania docelowego zadania. Taki cel stawiany jest przed uczeniem przez demonstrację.

W większości przypadków ekspert jest człowiekiem, który potrafi wykonać zadanie postawione przed agentem i potrafi sterować agentem w celu jego wykonania. Wykorzystanie ludzkiego eksperta pociąga za sobą poważne konsekwencje:

- **czas ludzkiego eksperta może być znacznie droższy niż czas maszynowy** - jeżeli dla badanego problemu dostępne jest wiarygodne i wydajne środowisko symulacyjne, to wykorzystanie klasycznego uczenia ze wzmocnieniem może być bardziej opłacalne niż zdobywanie prezentacji ludzkiego eksperta,
- **zachowanie ludzkiego eksperta może być nieoptymalne** - sposób realizacji zadania przez ludzkiego eksperta może być suboptymalny. Uczenie ze wzmocnieniem wyprzedziło ludzkich ekspertów w wielu dziedzinach, na przykład w grze w szachy, backgammona i go. Opierając się na samym odtwarzaniu zachowań eksperta agent nie osiągnie lepszych niż on wyników.
- **zachowanie ludzkiego eksperta może być niespójne** - realizacje tego samego zadania przez ludzkiego eksperta cechuje najczęściej pewna wariancja. Co więcej ekspert odpytywany o zachowania dla tego samego zadania podczas prezentacji w różnych warunkach może prezentować niespójne ze wcześniejszymi zachowania. W przypadku częściowo obserwowalnych problemów decyzyjnych ekspert może mieć dostęp do większej ilości informacji niż agent, i podświadomie korzystać z nich podczas prezentacji.
- **konieczne jest zrozumienie „idei" za zachowaniami eksperta** - jak w każdym problemie uczenia, konieczna jest generalizacja wiedzy. W uczeniu przez demonstację duży nacisk kładziony jest na użycie niewielkiej ilości danych uczących do zrozumienia skomplikowanych zadań, przez co odporność na przeuczenie i umiejętność generalizacji są wyjątkowo istotne.

Omawiane poniżej publikacje wykorzystują często ekspertów komputerowych - za ekspertów służą oddzielne, niezależne programy (przykładowo znające model środowiska i potrafiące dokonywać optymalnych decyzji). Przy nauce od komputerowych ekspertów nie występują opisane powyżej problemy, (w szczególności koszt i niespójność), co znacznie ułatwia i przyspiesza badania nad algorytmami uczenia przez demonstrację. Z drugiej strony, prawidłowości zaobserowane przy eksperymentach z ekspertem komputerowym mogą nie zachodzić przy użyciu ludzkiego eksperta i odwrotnie. Eksperci komputerowi mogą się sprawdzić w sytuacji, w której jeden wielozadaniowy i uniersalny agent uczy się wykonywania wielu zadań na podstawie prezentacji wielu wyspecjalizowanych do danego zadania programów.

Celem uczenia przez demonstrację nie jest uzyskiwanie lepszych wyników niż uczenie ze wzmocnieniem, ale uzyskiwanie ich dużo szybciej. [DO DYSKUSJI]

2.8.1 Kopiowanie zachowań

Najprostszym podejściem do uczenia przez demonstrację, nazywanym kopiowaniem zachowań *(ang. Behavioral cloning)* i opisanym między innymi w [Schaal, 1999], jest traktowanie problemu jak każdego innego problemu uczenia nadzorowanego. W kopiowaniu zachowań, w przeciwieństwie do maksymalizowania nagrody agenta w uczeniu ze wzmocnieniem, minimalizowana jest różnica pomiędzy polityką wyuczonego agenta a polityką eksperta.

To podejście zakłada jednak, jak każda metoda uczenia nadzorowanego, że dane uczące i testowe są niezależne i mają jednakowy rozkład, podczas gdy przy uczeniu przez demonstrację nauczona polityka ma bezpośredni wpływ na osiągane później stany, na podstawie których dana polityka będzie sprawdzana - intuicyjnie, trajektorie eksperta będą przedstawiały dobre zachowania i będą odwiedzały tylko dobre stany leżące na ścieżce optymalnej polityki. Gdy klasyfikator popełni błąd w odwzorowywaniu polityki eksperta, agent najprawdopodobniej trafi do stanu nieodwiedzonego przez eksperta i w którym nie będzie wiedział jak się zachować. Z dużym prawdopodobieństwem oznacza to popełnianie następnych błędów, ponieważ uczeń nie miał jak nauczyć się „podnoszenia się" po błędach. Jak dowiedziono w [Ross and Bagnell, 2010] wynikający z tego błąd rośnie kwadratowo w stosunku do czasu trwania epizodów.

2.8.2 Uczenie w przód

Pierwszym podejściem opisywanym przez [Ross and Bagnell, 2010] jest uczenie w przód *(ang. Forward Training)*. Podejście opiera się na przeprowadzeniu kilku powtórzeń uczenia, gdzie w każdym kroku następuje uczenie się jednej polityki w jednym, konkretnym, momencie. Jeżeli uczenie będzie przeprowadzone po kolei dla każdego kolejnego kroku w czasie, to próbka uzyskanych stanów, na których prowadzone jest dalsze uczenie odpowiada dystrybucji stanów testowych, a algorytm może odpytać eksperta o właściwe działanie w osiągniętych stanach, dzięki czemu ekspert ma okazję zaprezentować jak „podnosić się" po popełnieniu błędów przez klasyfikator. Powyższe podejście działa tylko dla zadań o skończonym horyzoncie czasowym, wymaga dużej interakcji z ekspertem i możliwości zrestartowania stanu środowiska i dokładnego odtworzenia uzyskanego wcześniej stanu, co w wielu przypadkach nie jest możliwe do zrealizowania.

2.8.3 Iterowany probabilistyczny mieszający algorytm

W celu wyeliminowania tych ograniczeń [Ross et al., 2010] proponują iterowany probabilistyczny mieszający algorytm *(ang. Stachastic Mixing Iterative Learning, SMILe)*. Opierając się na algorytmie iterowania polityki, algorytm w każdym kroku stosuje nową stochastyczną politykę wybierając z zadanym prawdopodobieństwem pomiędzy wykonywaniem polityki wyuczonej w poprzednim kroku i konstruowanej w danej iteracji nowej polityki. Prawdopodobieństwo wyboru nowej polityki jest niewielkie. Algorytm zaczyna od dokładnego wykonywania akcji eksperta. W każdej kolejnej iteracji prawdopodobieństwo odpytania eksperta jest coraz niższe i zbiega do 0.

Opisane rozwiązanie zostało z powodzeniem przetestowane na przykładzie grania w proste gry, gdzie danymi wejściowymi były surowe dane obrazowe z ekranu. Wadą tego podejścia jest brak odrzucania nieskutecznych polityk podczas iteracji, co może prowadzić do niestabilnych wyników.

Wykorzystanie analogicznego rozwiązania proponują [Bengio et al., 2015]. Ich propozycja zakłada wybieranie z prawdopodobieństwem ϵ polityki eksperta i z prawdopodobieństwem $1 - \epsilon$ polityki wyuczonej. Początkowa wartość ϵ powinna wynosić 1, aby klasyfikator mógł nauczyć się odtwarzać politykę eksperta. Wraz z postępem nauki ϵ powinno stopniowo maleć do 0, aby klasyfikator miał szanse nauczyć się stanów nieodwiedzonych przez eksperta.

2.8.4 Agregacją zbioru danych

W kolejnej publikacji [Ross et al., 2010] prezentują nowe podejście, nazwane agregacją zbioru danych *(ang. Dataset Aggregation, DAgger)*. W uproszczeniu, podejście to jest następujące. W pierwszej iteracji algorytm zbiera dane testowe stosując politykę pokazaną przez eksperta, po czym trenuje klasyfikator odwzorowujący zachowanie eksperta na danym zbiorze danych. W każdej kolejnej iteracji algorytm stosuje politykę wygenerowaną w poprzedniej iteracji i dodaje dane uzyskane podczas jej stosowania do zbioru danych, po czym trenuje klasyfikator tak, by odwzorowywał zachowanie eksperta na całym zbiorze danych. Podobnie jak w poprzednim algorytmie, żeby przyspieszyć uczenie na pierwszych etapach algorytmu, dodano opcjonalną losową możliwość odpytania eksperta o decyzję dla wybranego stanu. Uzyskane z pomocą tej metody wyniki są wyraźnie lepsze od wyników metody SMILEe.

Metodę testowano z wykorzystaniem klasyfikatorów SVM na przykładzie gry wyścigowej 3D Mario Cart (ludzki ekspert wybiera pomiędzy skręcaniem w lewo i w prawo) oraz z komputerowym ekspertem na przykładzie gry Mario.

2.8.5 Podążanie za ekspertem a przewyższanie eksperta

Dla wielu praktycznych problemów polityka eksperta może nie być optymalna. Algorytm, który stara się tylko i wyłącznie odwzorować politykę eksperta będzie generował w takiej sytuacji nieoptymalne wyniki, które w wielu praktycznych sytuacjach mogą znacznie odbiegać od optimum. Prostym rozwiązaniem tego problemu przedstawionym w [Chang et al., 2015] jest stosowanie e-zachłannej strategii – w każdym ruchu algorytm może wybrać z małym prawdopodobieństwem e wykonanie losowej akcji zamiast akcji optymalnej według wyuczonej polityki. Dzięki temu algorytm może znaleźć lokalne optimum bliskie polityce eksperta. Warto zauważyć, że wymusza to posługiwanie się całościową nagrodą (kosztem) wykonania zadania jako celem optymalizacji, w przeciwieństwie do prostszego minimalizowania różnicy pomiędzy wynikami wyuczonej polityki a polityki eksperta.

2.9 Odwrócone uczenie ze wzmocnieniem

Użycie kopiowania zachowań i wywodzących sie z niego metod pozwala na wytrenowanie agenta odruchowego[SPRAWDZIĆ], który dla danego stanu będzie potrafił określić optymalną akcję na jeden krok do przodu. Taki agent zna optymalną politykę, ale nie jest świadomy jej powodów. Odwrócone uczenie ze wzmocnieniem *(ang. Inverse reinforcement learning, IRL)* opiera się na założeniu, że w ogólności optymalna polityka agenta nie stanowi najlepszego i najbardziej zwięzłego opisu zadania podstawionego przed agentem - najbardziej precyzyjnym opisem, pozwalającym na większą dowolność i adaptację jest znajomość funkcji nagród $R_a(\cdot,\cdot)$.

Oczywiście, bez modelu środowiska funkcja $R_a(\cdot,\cdot)$ nie jest dostępna, dlatego zadaniem postawionym przed odwróconym uczeniem ze wzmocnieniem jest odtworzenie $R_a(\cdot,\cdot)$ na podstawie dostarczonych trajektorii eksperta.

Zadanie odwróconego uczenia ze wzmocnieniem jest znacznie trudniejsze od zwykłego uczenia ze wzmocnieniem. Przede wszystkim, IRL musi się zmierzyć z 2 problemami:

- Niejasność $R_a(\cdot,\cdot)$ - w większości przypadków, dla danych trajektorii eksperta istnieje nieskończenie wiele pasujących $R_a(\cdot,\cdot)$. Formalnie oznacza to, że IRL nie ma zdefiniowanego poprawnego rozwiązania.

- Złożoność obliczeniowa - samo uczenie ze wzmocnieniem jest bardzo wymagające obliczeniowo. W odwróconym uczeniu ze wzmocnieniem, sprawdzanie $R_a(\cdot,\cdot)$ uzyskanych w kolejnych krokach wymaga każdorazowego rozwiązywania problemu uczenia ze wzmocnieniem na podstawie aktualnej funkcji $R_a(\cdot,\cdot)$, co oznacza, że IRL wymaga większego o rząd wielkości kosztu obliczeniowego niż RL.

Przykładowe zastosowania odwróconego uczenia ze wzmocnieniem to parkowanie samochodu [Abbeel et al., 2008], nawigacja na podstawie obrazów satelitarnych [Ratliff et al., 2006] albo wykonywanie ewolucji helipkopterem[Coates et al., 2009].

2.10 Środowisko VizDoom

Środowisko VizDoom, przedstawione w [Kempka et al., 2016], jest narzędziem do testowania algorytmów sterowania na podstawie surowych danych o obrazie 3D. Środowisko bazujące na klasycznej grze Doom, w której gracz widzi trójwymiarowy świat z perspektywy pierwszej osoby i strzela do potworów. W stosunku do nauki w środowisku 2D, takim jak Atari 2600 [Mnih et al., 2015], nauka w środowisku 3D jest wielkim krokiem naprzód i stanowi znacznie lepsze przybliżenie nauki w realnym świecie.

VizDoom oferuje wygodny interfejs, który doskonale wpisuje się w standardowy szkielet metod uczenia ze wzmocnieniem. VizDoom potrzebuje mało zasobów, może działać bez środowiska graficznego, jest wydajny, pozwala na uruchamianie wielu instancji równolegle oraz na wygodne tworzenie nowych scenariuszy dopasowanych do potrzeb konkretnych problemów badawczych.

Vizdoom udostępnia interfejs programistyczny dla Pythona, Javy, C++ i Lua. Preferowany i najbardziej rozwijany jest Python.

Przykładowe minimalne użycie środowiska VizDoom przedstawione jest poniżej (prezentowany kod jest zmodyfikowanym przykładem *scenarios.py* dołączonym do środowiska):

```
from vizdoom import DoomGame

game = DoomGame()

game.load_config("../../scenarios/basic.cfg")

game.set_screen_resolution(ScreenResolution.RES_640X480)
game.set_window_visible(True)
game.init()

# Creates all possible actions depending on how many buttons there are.
actions = prepare_actions(game.get_available_buttons_size())

episodes = 10

for i in range(episodes):
    game.new_episode()
    while not game.is_episode_finished():
        # Gets the state and possibly to something with it
        state = game.get_state()
        # Makes a random action and save the reward.
        reward = game.make_action(choose(state, actions))
        new_state  = game.get_state()
        learn(state,game.get_last_action(), reward, new_state)
```

2.11 Słownik pojęć

W dalszej częsci pracy używane będą następujące terminy:

- *agent/gracz* - program sterujący agentem poruszającym się w badanym środowisku,
- *ekspert* - człowiek lub zewnętrzny program, który zna optymalną (lub bliską optymalnej) politykę działania i potrafi udzielić agentowi informacji na temat realizacji tej polityki i oceniać politykę agenta,
- *epizod/gra* - ciąg interakcji agenta lub eksperta z otoczeniem, od stanu początkowego do stanu terminalnego.
- *klatka/krotka/doświadczenie* - pojedyncza informacja s, a, s', r zebrana przez agenta lub eksperta w ramach jednego kroku interakcji ze środowiskiem.
- *ocena/odpytanie eksperta* - zarządanie przez program agenta, aby ekspert określił jaką decyzję podjąłby w danym stanie.
- *trajektoria* - uporządkowana lista krotek pochodząca z jednego epizodu.
- *uczeń* - agent (w algorytmach uczenia przez demonstrację), najczęściej w kontekście odpytywania eksperta.

Rozdział 3

Zaimplementowane podejścia

W poniższym rozdziale przedstawione zostaną zaimplementowane i analizowane w pracy podejścia.

Każde z rozwiązań działa w ramach wspólnego szkieletu, bazującego na przykładowych rozwiązaniach towarzyszących środowisku VizDoom. Dzieki temu możliwe jest bezpośrednie porównanie zachowania różnych podejść przy zmianie tylko kluczowych algorytmów przy zachowaniu niezmienności pozostałych czynników.

3.1 Wykorzystane scenariusze

Eksperymenty przeprowadzono na następujących scenariuszach, reprezentujących łatwy, sredni i wysoki poziom trudności.

Podstawowy (ang. Basic)

Sceneria składa się z prostokątnego pomieszczenia. Agent startuje w jednym końcu pomieszczenia, po środku ściany, a w losowym miejscu pod przeciwległą ścianą znajduje się pojedynczy, nieruchomy przeciwnik.

Agent może strzelać do przeciwnika oraz poruszać się w prawo lub w lewo. Agent ma ograniczoną amunicję i dostaje punkt za trafienie przeciwnika.

Strategia optymalna polega na przesunięciu się w kierunku przeciwnika i oddaniu do niego pojedynczego strzału.

Rysunek 3.1: Scenariusz podstawowy

Obrona środka (ang. Defend the center)

Sceneria składa się z kolistej areny. Agent znajduje się na środku areny, a na jej krańcach losowo pojawiają się przeciwnicy, którzy poruszają się w stronę agenta, a po dotarciu do niego zadają mu obrażenia. Są dwa rodzaje przeciwników różniących się wyglądem i szybkością poruszania.

Agent może strzelać i kręcić się wokół własnej osi w lewo i prawo. Agent ma ograniczoną amunicję i dostaje punkty za każde trafienie przeciwnika.

Strategia optymalna polega na kręceniu się w jedną strone w kółko, ignorowaniu odległych przeciwników i strzelaniu do bliskich, priorytetyzując szybszych przeciwników. Ignorowanie dalekich przeciwników jest konieczne, żeby w czasie strzelania do nich inni przeciwnicy nie zaszli agenta od tyłu - optymalna strategia wymaga zajmowania się najpierw najbliższym zagrożeniem.

Rysunek 3.2: Scenariusz obrona środka

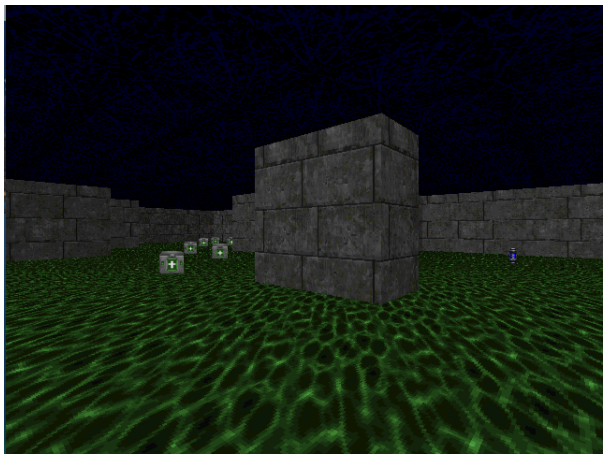
Trudne zbieranie apteczek (ang. Health gathering supreme)

Sceneria składa się z labiryntu, którego podłoga jest pokryta kwasem. Agent startuje w losowym miejscu labiryntu. W tym scenariuszu nie ma ruchomych przeciwników. Na podłodze labiryntu pojawiają się losowo apteczki, które dodają agentowi punkty życia i miny, które zabierają agentowi punkty życia. Kwas na podłodze nieustannie odbiera agentowi punkty życia.

Agent może poruszać się do przodu, na ukos w prawo lub w lewo oraz kręcić się wokół własnej osi w prawo lub w lewo. Agent dostaje punkty za pozostawanie przy życiu - im dłużej potrwa gra, tym większą sumę punktów zdobędzie agent.

Strategia optymalna polega na chodzeniu po labiryncie, niewchodzeniu na miny i zbieraniu apteczek, preferując kierowanie się do dużych skupisk apteczek. Wskazane jest unikanie zbierania pojedynczych, odizolowanych apteczek, gdyż liczba punktów życia uzyskana z takiej apteczki może być mniejsza niż liczba punktów życia straconych na dotarcie do apteczki. Wskazane jest niepozostawanie w jednym obszarze labiryntu, ponieważ nowe apteczki mogą nie pojawiać się wystarczająco szybko, żeby utrzymać agenta przy życiu.

Co istotne, w tym scenariuszu bardzo często optymalna decyzja nie jest jasna - sensownie działający ekspert i agent mogą często wybierać pomiędzy wieloma poprawnymi drogami i zachowaniami.



Rysunek 3.3: scenariusz trudne zbieranie apteczek

3.2 Q-learning

Q-learning, opisany dokładniej w 2.4.2, jest najpopularniejszą metodą uczenia ze wzmocnieniem. Agent uczy się wartości funkcji Q odwiedzanych par stan → akcja. Wartość funkcji Q dla danego stanu odpowiada zdyskontowanej sumie nagród, jaką można uzyskać po trafieniu do danego stanu, przy założeniu wykonywania dalej optymalnych akcji. Agent rozpoczyna działając losowo, a potem uczy się łącząc działania losowe z działaniami optymalnymi według aktualnego stanu wiedzy i obserwując ich rezultaty.

Problemem Q-learningu jest konieczność długotrwałego uczenia agenta, a odpowiedzią na ten problem jest uczenie z ekspertem. Wyniki Q-learningu będą stanowiły punkt odniesienia dla wyników innych metod.

3.2.1 Implementacja

Zastosowana implementacja opiera się na głębokiej sieci neuronowej jako aproksymatorze funkcji Q. Jest wyposażona w pamięć powtórek *(ang. replay memory)* (patrz 2.6.1), a do eksploracji używana jest metoda ϵ-zachłanna (patrz 2.7.1). Schemat sieci neuronowej jest następujący.

[SCHEMAT SIECI]

3.2.2 Zachowanie

Uczenie agenta trwa długo. Zastosowane proste rozwiązanie jest w stanie nauczyć się poprawnego działania dla scenariusza Podstawowy (ang. Basic). Dla Obrona środka (ang. Defend the center) agent zatrzymuje się w optimum lokalnym, polegającym na częstym strzelaniu i kręceniu się w kółko bez przejmowania się konkretnymi przeciwnikami, a jego zachowanie nie wygląda sensownie. W Trudne zbieranie apteczek (ang. Health gathering supreme) agent nie potrafi sensownie poruszać się po labiryncie.

Zastosowanie bardziej wyrafinowanych metod uczenia ze wzmocnieniem, których implementacja wykracza poza zakres tej pracy, pozwala poprawić wyniki na trudniejszych scenariuszach. Dla Obrona środka (ang. Defend the center) możliwe jest uzyskanie wyników bliskim maksymalnym, ale do osiągnięcia sensownego zachowania w Trudne zbieranie apteczek (ang. Health gathering supreme) konieczna jest na przykład pomoc z zewnątrz programu w postaci techniki kształtowania *(ang. shaping)* (2.6.3). Mimo dobrych wyników, metody te ciągle wymagają przeanalizowania

wielkich ilości klatek - wymagane liczby klatek są kilka rzędów wielkości większe niż liczba potrzebna do uczenia z ekspertem.

3.2.3 Wnioski

Q-learning potrzebuje dużo czasu, żeby uzyskać przywoicie zachowującego się agenta, ma również problem z wychodzeniem z optimów lokalnych polityki zachowań (szczególnie przy zastosowanej prostej implementacji). Dobrze sprawdza się za to w ewolucyjnym, stopniowym poprawianiu zachowań agenta.

3.3 Kopiowanie zachowań

Kopiowanie zachowań (ang. *Behavioral Cloning*) stanowi najbardziej podstawowe podejście do uczenia przez demonstację. Na podstawie zebranych trajektorii eksperta uczony jest klasyfikator, który przyjmując na wejściu stan s ma za zadanie przewidzieć, jaką akcję a wykonałby w danej sytuacji ekspert.

Mimo że dla wielu problemów Kopiowanie zachowań jest nieskuteczne (powody opisane są w 2.8.1), na wyniki osiągane na badanych scenariuszach VizDoom są bardzo zadawalające.

Kopiowanie zachowań jest podstawą bardziej zaawansowanych technik opisanych w dalszych punktach.

Ważną różnicę w stosunku do Q-learningu stanowi fakt, że czas trwania metody ogranicza się w znaczącej częsci do czasu zbierania przykładowych trajektorii przez eksperta. Czas trenowania klasyfikatora na zebranych danych powinien być pomijalny w stosunku do czasu zbierania. Oznacza to, że uczenie agenta za pomocą Kopiowania zachowań, wliczając w to zbieranie przykładowych trajektorii, trwa znacznie krócej niż za pomocą np. Q-learningu, w którym, dla wielu praktycznych problemów, agent musi grać przez przez wiele milionów klatek, by osiągnąć zadawalające wyniki.

3.3.1 Implementacja

Kopiowanie zachowań sprowadza się do klasycznego problemu uczenia nadzorowanego z wieloma klasami (możliwymi akcjami), z których tylko jedna etykieta na raz jest poprawna. Dane wejściowe stanowią obrazy przedstawiające stan, wynikiem jest etykieta akcji, którą należy wykonać. Jako klasyfikatora użyto głębokiej sieci neuronowej, o artchitekturze bazującej na architekturze z rozdziału 3.2.

Różnica pomiędzy architekturami sporowadza się do uczenia i interpretacji wyników. Przy Q-learningu sieć musi przewidywać wartość funkcji Q dla wszystkich akcji, a przy Kopiowaniu zachowań wystarczy określenie najbardziej pasującej akcji. Wykonane i przewidywane akcje są zakodowane za pomocą *one-hot encoding*, a wynik uzyskiwany jest przez zastosowanie funkcji *softmax* na wartościach q z architektury Q-learningu. Schemat sieci wygląda następująco.

[SCHEMAT SIECI]

Co istotne, dla większości scenariuszy problem uczenia stan → akcja charakteryzuje się niezbalansowanym zbiórem danych. Akcje „strzelaj" występują znacznie rzadziej niż akcje ruchu. Oznacza to, że klasyfikator naiwnie nauczony na niezmienionym zbiorze danych, mimo dobrej teoretycznej trafności, będzie zupełnie nieskuteczny (np. nie wybierze nigdy akcji „strzelaj").

Problem ten został rozwiązany przez zrównoważenie zbioru danych przy użyciu metody *oversampling* która polega na wielokrotnym uwzględnieniu w zbiorze uczącym przykładów mniej licznych klas w taki sposób, żeby liczność przykładów dla każdej z klas w uzyskanym zbiorze danych była podobna. W zastosowanej implementacji dla każdej akcji przetrzymywany jest oddzielny zbiór

danych, a użyte do uczenia próbki składają się w równych proporcjach z przykładów zastosowania każdej akcji.

Warto zauważyć, że architektura Q-learningu wymaga, żeby każda możliwa akcja była zdefiniowana oddzielnie, łącznie z akcjami stanowiącymi złożenie innych, podstawowych akcji. Przykładowo akcje „lewo", „prosto" i „lewo i prosto" są dla modelu zupełnie niezwiązane, mimo że często można byłoby stosować je zamiennie. W przypadku Kopiowania zachowań możliwe byłoby stworzenie klasyfikatora stan → akcja, który jest jednocześnie klasyfikatorem binarnym dla każdej podstawowej akcji z osobna. Taki klasyfikator mógłby zamiast wybierać pomiędzy „lewo", „prosto" i „lewo i prosto" zdecydować „lewo" - tak i „prosto" - tak, uzyskując „lewo i prosto". Jednakże takie rozwiązanie nie zostało w tej pracy zbadane.

3.3.2 Techniczna implementacja

Zbieranie danych zostało zrealizowane za pomocą trybu SPECTATOR środowiska VizDoom, pozwalającemu agentowi obserować grę człowieka. Podczas gry eksperta zapisywane są stany, akcje i nagrody dla każdej kolejnej klatki. Trajektoria eksperta serializowana jest do pliku za pomocą narzędzia *pickle* dostępnego dla języka python.

Eksperta gra przy rozdzielczości 640x480 pikseli, i takiej wielkości obrazy zapisywane są do pliku z trajektorią. Konsekwencją są bardzo duże rozmiary plików (3GB dla 6 tysięcy klatek). Obrazy nie są zmniejszane przed zapisem, żeby umożliwić swobodne manipulowanie wielkością obrazów używanych do uczenia klasyfikatora, bez konieczności generowania nowych trajektorii eksperta przy innych ustawieniach obrazu.

Tryb SPECTATOR ustawiany jest w następujący sposób.

```
game.set_window_visible(True)
game.set_mode(Mode.SPECTATOR)
```

Trajektoria eksperta zbierana i zapisywana jest następująco.

```
    game.new_episode()
    while not game.is_episode_finished():
        state = game.get_state()
        game.advance_action()
        next_state = game.get_state()
        last_action = game.get_last_action()
        reward = game.get_last_reward()
        isterminal = game.is_episode_finished()

        print("State #" + str(state.number))
        print("Game variables: ", state.game_variables)
        print("Action:", last_action)
        print("Reward:", reward)
        print("====================")
        memory.append((state.screen_buffer,last_action, next_state.screen_buffer, reward,
     isterminal))
```

Zapis trajektorii do pliku wygląda następująco.

```
with open('recorder_episode.pkl', 'wb') as f:
    pickle.dump(memory, f, 2)
```

3.3.3 Zachowanie

Eksperymenty były prowadzone na scenariuszach Trudne zbieranie apteczek (ang. Health gathering supreme) i Obrona środka (ang. Defend the center). W obu przypadkach kopiowanie

zachowań nauczone tylko na podstawie 3 trajektorii eksperta (6 tysięcy klatek) osiągało wizualnie sensowne zachowanie agentów i zaskakująco dobre wyniki.

Dla Obrona środka (ang. Defend the center) agentowi zdarzało się strzelać w nieodpowiednim momencie lub nie strzelać, kiedy było to potrzebne. Częste było też nieoptymalne zachowanie w postaci strzelania do odległych przeciwników, podczas gdy inni przeciwnicy mogli podkraść się za plecy agenta zabijając go i kończąc grę.

W tym scenariuszu zwiększanie liczby trajektorii eksperta użytych do uczenia zwiększało wyniki agenta, który w dużej częsci gier osiągał wyniki bliskie maksymalnym i tylko sporadycznie dawał się na początku gry zajść od tyłu, co skutkowało pojedynczymi niskimi wynikami.

Dla Trudne zbieranie apteczek (ang. Health gathering supreme) agentowi często zdarzało się blokować w rogach labiryntu, wpadając w nieskończoną pętlę akcji. Problem i rozwiązanie zostało opisane w 3.6. Po wyelimininowaniu problemu agent zachowywał się wizualnie sensownie i osiągał przyzwoite wyniki. Problemem jest tylko nauczenie agenta omijania min.

Na początku pracy ze scenariuszem ludzki Ekspert uznał miny za mniejsze apteczki i nie zauważył spadku życia po wejściu w nie. Wyniki uzyskiwane przez eksperta wchodzącego czasami w miny były tylko nieznacznie lepsze od wyników agenta nauczonego na podstawie tych trajektorii.

W tym scenariuszu zwiększanie liczby trajektorii eksperta użytych do uczenia nie polepszało wyników agenta.

3.3.4 Wnioski

W badanych scenariuszach Kopiowanie zachowań osiąga znacznie lepsze wyniki, niż sugerowałaby literatura i uzyskuje je w ciągu ułamka czasu potrzebnego klasycznym metodom uczenia ze wzmocnieniem. Uzyskani agenci w większości przypadków zachowują się sensownie, chociaż czasem popełniają systematyczne błędy. Kopiowanie zachowań wydaje się świetnym punktem startowym dla VizDooma i wydaje się wskazane, żeby inne metody rozszerzały to podejście, zamiast je zastępować.

3.4 Q-learning z ekspertem

Q-learning jest potężną metodą, która jest zdolna osiągać wyniki znacznie prześcigające ludzkie. Jej wadą jest natomiast znaczna ilość czasu i klatek działania, których potrzeba zanim agent zacznie zachowywać się sensownie. Uczenie z ekspertem, z drugiej strony, może na podstawie trajektorii eksperta szybko uzyskać przyzwoicie zachowującego się agenta, którego osiągi są niższe lub równe osiągom użytego eksperta.

Wobec tych dwóch przeciwstawnych podejść dobrym pomysłem wydaje się próba wykorzystania agenta uzyskanego na pomocą uczenia z ekspertem jako stanu początkowego dla Q-learningu - dzięki temu pomijane jest czasochłonne początkowe rozpoznawanie podstaw mechaniki środowiska, a metoda rozwija tylko za pomocą uczenia ze wzmocnieniem przyzwoitego już wcześniej agenta.

Podstawowym problemem jest niekompatybilność wyjść obu metod - klasyfikator z 3.3 potrafi określić najlepszą akcję dla danego stanu, ale te informacja nie pozwala na pełne odtworzenie wartości Q poszczególnych stanów. Wartości Q dla stanów odwiedzonych przez eksperta można odtworzyć na podstawie jego trajektorii (albo przybliżyć, jeżeli przejścia i nagrody nie są deterministyczne), ale niemożliwe będzie odtworzenie wartości Q dla akcji niewybranych przez eksperta - należy za to zagwarantować, że ich Q wartości będą niższe niż Q wartości wybranych akcji.

[UZUPEŁNIĆ]

3.5 DAgger

Podejście Agregacji Zbioru Danych (ang. Dataset Aggregation) [Ross et al., 2010] zostało opisane we wcześniejszym rozdziale. Kluczowym założeniem metody jest odpytywanie eksperta o właściwe działanie w stanach, które nie były wcześniej przez niego pokazane (i nie należą do „poprawnych" trajektorii), a które zostały odwiedzone przez agenta na skutek jego nieoptymalnego zachowania.

W rzeczywistości, dla bardziej skomplikowanych zadań, odpytywanie eksperta o decyzję dla każdego odwiedzonego przez agenta stanu jest niepraktyczne. Ocenianie wielu kolejnych stanów może być drogie i nużące dla eksperta, co może przekładać się na obniżoną jakość decyzji. Ocena dokonywana przez eksperta może też w praktyce różnić się w zależności od tego, czy ekspert napotkał dany stan podczas normalnego działania, czy podczas oceny pojedynczych, wyrwanych z kontekstu stanów.

Aby zminimalizować ten problem, konieczne jest określenie mniejszego podzbioru stanów, dla których potrzebna jest ocena eksperta.

3.5.1 Implementacja

Zastosowana implementacja jest rozszerzeniem 3.3. Pierwszym krokiem jest załadowane przygotowanych wcześniej trajektorii eksperta do pamięci agenta (zestawu danych).

Następnie agent rozpoczyna działanie, bazując na swoim aktualnym stanie wiedzy. Po wystąpieniu określonych warunków, definiujących potrzebę odpytania eksperta, działanie programu zostaje wstrzymane, a sterowanie przekazane jest do eksperta. Aby dostosować się do ograniczeń ludzkiego eksperta, po przekazaniu sterowania program przechodzi w tryb synchroniczny - przed każdą kolejną klatką czeka na reakcję eksperta. Ubocznym skutkiem tej implementacji jest pomijanie akcji „nic nie rób", która jest wykonywana dopiero po wciśnięciu dedykowanego klawisza.

Po wystąpieniu określonych warunków, definiujących koniec potrzeby odpytywania eksperta, wszystkie stany i akcje odwiedzone w trakcie danej demonstacji dodawane są do pamięci agenta (trajektoria może być dodana do pamięci z większą wagą niż początkowe prezentacje - w przeciwnym wypadku dodanie nowych danych mogłoby nie być odczuwalne). Agent aktualizuje klasyfikator akcji na podstawie rozszerzonego zestawu danych, po czym przejmuje sterowanie od eksperta i wraca do normalnego działania bazując na uaktualnionym stanie wiedzy.

Po ponownym wystąpieniu określonych warunków, kontrola może ponownie zostać przekazana do eksperta.

3.5.2 Przekazywanie sterowania

Jednym z najważniejszych problemów jest zdefiniowanie, kiedy przekazywać sterowanie pomiędzy agentem a ekspertem. Wybór sposobu będzie decydował o tym, jak często ekspert będzie odpytywany i na ile istotna będzie uzyskana wiedza. Sprawdzone zostały trzy następujące sposoby.

Losowe przekazanie sterowania

1. Przed wykonaniem każdej akcji agent z bardzo małym prawdopodobieństwem może zdecydować o przekazaniu sterowania ekspertowi.

2. Po każdej akcji eksperta program z większym prawdopodbieństwem może zdecydować o przekazaniu sterowania do agenta.

Losowe przekazywanie sterowania jest niepraktyczną metodą - dla analizowanych problemów agent nie potrzebuje pomocy eksperta przez większość czasu, więc losowo wybrane momenty przekazania sterowania w ogromnej większości nie dostarczają istotnej informacji. Zaletą jest natomiast automatyczność decyzji - program podczas gry agenta może działać w przyspieszonym tempie.

Analiza niepewności sieci

1. Przed wykonaniem każdej akcji sprawdzana jest jej niepewność sieci dla danego stanu [REF]. W przypadku wystąpienia zadanej liczby kolejnych niepewnych akcji sterowanie przekazywane jest do eksperta.

2. Po każdej akcji eksperta sprawdzana jest akcja, którą wykonałby agent. Jeżeli przez zadaną liczbę kolejnych kroków agent postąpiłby identycznie jak ekspert, to sterowanie wraca do agenta.

Analiza niepewności sieci jest skuteczniejsza niż losowe przekazywanie sterowania. Wybrane tym sposobem okna działania eksperta częściej pokrywają się z oknami niepoprawnego działania agenta. W dalszym ciągu skuteczność metody nie jest zadowalająca - przyjęta miara niepewności powoduje, że agent może przekazać sterowanie do Eskperta w obliczu sytuacji, dla której więcej niż jedna akcja jest sensowna. Porównywanie akcji agenta i eksperta przez zadaną liczbę kroków jest skuteczne dla problemów z niewielką liczbą akcji, ale nieskuteczne w sytuacji, w której podobny efekt można uzyskać za pomocą różnych sekwencji kroków (przykładowo dojście do danego punktu za pomocą permutacji akcji „lewo", „prosto" i „lewo i prosto"). Podobnie jak przy losowym podejściu, dzięki automatycznemu działaniu możliwe jest działanie programu w przyspieszonym tempie poczas gry agenta. [NIEPEWNOŚĆ DO DYSKUSJI]

Decyzja eksperta

1. ekspert obserwuje działanie agenta. ekspert przejmuje sterowanie kiedy uzna, że agent trafił do niepożądanego stanu.

2. Kiedy ekspert uzna, że agent nie jest już w niepożądanym stanie może oddać sterowanie agentowi.

Decyzja eksperta jest najskuteczniejszą metodą i jest używana w dalszych eksperymentach. Ekspert może sam stwierdzić, kiedy działanie agenta jest niezgodne z pożądanym, maksymalizując skuteczność odpytywania eksperta. Oczywiście, ekspert musi spędzić więcej czasu obserwując działanie agenta, ale obserwacja jest dużo mniej uciążliwa (a zatem tańsza), niż prezentowanie. Problemem w niektórych sytacjach jest możliwość rozróżnienia, kiedy agent zachowa się niepożądanie i należałoby przejąć sterowanie - w wielu sytuacjach ekspert reaguje zbyt późno, żeby demonstracja była skuteczna.

3.5.3 Techniczna implementacja

Architektura sieci neuronowej jest identyczna z architekturą zastosowaną w 3.3.

Implementacja przekazywania sterowania do eksperta w środowisku VizDoom byłaby wymagająca i czasochłonna, dlatego zastosowano znacznie prostsze, chociaż mniej eleganckie rozwiązanie.

Przy odpytywaniu eksperta o akcje program oczekuje na następny znak, który pojawi się na standardowym strumieniu wejścia programu (następny znak wpisany w konsoli). Wybrane znaki są przypisane do indeksów wybranych akcji, wpisanie nieznanego znaku powoduje wybranie akcji o indeksie 0, czyli „nic nie rób".

```
    def get_expert_action(self):
        fd = sys.stdin.fileno()
        old_settings = termios.tcgetattr(fd)
        try:
            tty.setraw(sys.stdin.fileno())
            move = sys.stdin.read(1)
        finally:
            termios.tcsetattr(fd, termios.TCSADRAIN, old_settings)
        if move == 'j':
            return 4
        if move == 'l':
            return 2
        if move == 'a':
            return 1

        if move == 'i':
            return 1
        if move == 'u':
            return 5
        if move == 'o':
            return 3
        return 0
```

Przy metodzie decyzji eksperta konieczne jest asynchroniczne przetwarzanie działania eksperta. Program nie może oczekiwać na działanie eksperta, ale kiedy ekspert zarząda przekazania sterowania nastepna akcja powinna być już wykonywana przez niego.

W tym celu wykorzystano bibliotekę PyKeyboardEvent, która umożliwia reagowanie na systemowe informacje o wciśnięciu bądź puszczeniu klawiszy klawiatury. Poniższa klasa wywołuje zadaną funkcję po wciśnięciu lub puszczeniu zadanych klawiszy.

```
from __future__ import print_function
from pykeyboard import PyKeyboardEventwojciech_kopec_101675.pdf


class KeyMonitor(PyKeyboardEvent):
    def __init__(self, keys, keypress_handler):
        PyKeyboardEvent.__init__(self)
        self.keypress_handler = keypress_handler
        self.keys = set(keys)

    def tap(self, keycode, character, press):
        if character in self.keys:
            self.keypress_handler(character, press)
```

Wywoływana funkcja znajduje się poniżej. Klawisz 'p' przekazuje sterowanie pomiędzy ekspertem i agentem. Klawisze ',' i '.' zwalniają i przyspieszają działanie programu podczas gry agenta.

```
    def __toggle_user_input(self, character):
        if character == 'p':
            if self.expert_mode:
                self.learn_all()
            self.expert_mode = not self.expert_mode
            print ("Expert toggled: " + str(self.expert_mode))
        elif character == '.':
            self.framerate+=5
            print ("Framerate: " + str(self.framerate))
        elif character == ',':
            self.framerate -= 5
            print ("Framerate: " + str(self.framerate))
        return True
```

3.5.4 Zachowanie

Eksperymenty były prowadzone na przede wszystkim scenariuszu Trudne zbieranie apteczek (ang. Health gathering supreme). Początkowe trajektorie eksperta były wygenerowane zgodnie z opisem w rozdziale 3.6.

Dla każdego z badanych scenariuszy uwzględnianie fragmentów trajektorii zaprezentowanych przez eksperta w trakcie gry obniża początkowo wyniki. Na skutek nauczenia się niespójnych zachowań eksperta agent zachowuje się mniej płynnie i częściej wpada w nieskończone pętle ruchów (przykładowo obracanie się naprzemian w lewo i w prawo w rogu labiryntu), co prowadzi do osiągania niższych wyników.

W scenariuszu Trudne zbieranie apteczek (ang. Health gathering supreme) głównym problemem agenta opisanego w rodziale Kopiowanie zachowań jest nieomijanie min i celem zastosowania podejścia DAgger jest wyeliminowanie tego problemu. Za każdym razem, kiedy agent zbliża się do min ekspert przejmuje kontrolę i omija miny bądź wybiera inną ścieżkę. Pary stan $\rightarrow$ akcja uzyskane w ten sposób są dodawane do pamięci dziesięciokrotnie.

Podczas pierwszych epizodów nauki wyniki uzyskiwane przez agenta zauważalnie się obniżają, a problem wchodzenia na miny nie jest wyeliminowany.

Następne epizody nauki powoli poprawiają wyniki agenta, przywracając je do poziomu wyjściowego lub nieznacznie go przewyższającego. Agent rzadziej wchodzi w miny, ale problem w dalszym ciągu pozostaje obecny.

Kolejne epizody nauki doprowadzają do przeuczenia - wyniki obniżają się, a agent regularnie wpada w nieskończone pętle ruchów. Wchodzenie w miny nie zostaje wyelimininowane. Pogorszenie zachowania agenta może wynikać ze znużenia eksperta, a co za tym idzie zmiany jego zachowań i pogorszenia jego decyzji.

3.5.5 Wnioski

Dla wypróbowanych problemów DAgger nie wydaje się być skuteczny. W VizDoomie decyzje podejmowane przez ludzkiego eksperta są bardziej skomplikowane niż w Mario Cart, przedstawianym w publikacji, co, na skutek niespójności przedstawianych przez eksperta zachowań, zamiast do podwyższenia wyników agenta prowadzi do obniżania jego skuteczności. Zastosowane głębokie sieci neuronowe mogą też znacznie skuteczniej uogólniać wiedzę zdobytą podczas pierwszej prezentacji eksperta niż prostsze klasyfikatory SVM, a co za tym idzie nawet bez użycia DAggera agent potrafi znaleźć sensowne wyjście z większości sytuacji. Uzyskiwanie oceny eksperta jest uciążliwe i kosztowne.

3.6 Świadomie prezentujący ekspert

W sekcji 3.3 opisano agenta budującego klasyfikator (stan $\rightarrow$ akcja) na podstawie trajektorii zebranych podczas gryeksperta. Uzyskany agent zachowywał się sensownie, ale problem stanowiło między innymi blokowanie się w rogach labiryntu i wchodzenie na miny. Główną praktyczną wadą metody 3.5, która miała na celu zaradzenie temu, jest niespójność zachowań eksperta podczas pierwszej (ciągłej) prezentacji i zachowań podczas krótkich prezentacji podczas gry agenta oraz uciążliwość obserwacji i przejmowania sterowania od agenta w trakcie gry.

Problem wchodzenia w ściany, dla przykładu, jest łatwo zauważalny podczas obserwacji działania agenta. Oczywistym jest też powód jego występowania - ekspert, w przeciwnieństwie do agenta, pamięta jak dotarł do danego stanu i znajdując się w rogu pamięta, w którą stronę powinien z niego wyjść. Badani agenci mogą pamiętać tylko kilka ostatnich odwiedzonych klatek i

nie pamiętają swoich trajektorii. Dlatego klasyfikator nauczony na trajektoriach eksperta nie ma wystarczającej informacji żeby rozróżnić konieczność wychodzenia z rogu obracając się w prawo bądź w lewo.

Rozwiązaniem jest powtórne zebranie trajektorii eksperta, kładąc przy prezentacji nacisk na zachowywanie się w sposób spójny i ułatwiający klasyfikatorowi skuteczną naukę. Możliwe jest też pokazywanie rozwiązań sytuacji, które wcześniej sprawiały klasyfikatorowi problem, w celu pokazania poprawnego zachowania w danej sytuacji.

Oczywiście, takie zachowanie eksperta skutkuje uzyskiwaniem przez niego nieoptymalnych wyników, a co za tym idzie wyniki możliwe do osiągnięcia przez idealnie odwzorowującego agenta też są niższe. W praktyce różnica pomiędzy wynikami eksperta i agenta powinna się zmniejszyć dzięki świadomej prezentacji, skutkując wyższymi wynikami osiąganymi przez agenta.

3.6.1 Techniczna implementacja

Architektura sieci neuronowej jest identyczna z architekturą zastosowaną w 3.3.

3.6.2 Zachowanie

Eksperymenty były prowadzone na scenariuszach Trudne zbieranie apteczek (ang. Health gathering supreme) i Obrona środka (ang. Defend the center).

W scenariuszu Obrona środka (ang. Defend the center) ekspert podczas świadomej prezentacji powstrzymywał się od strzelania do odległych przeciwników i świadomie preferował strzelanie do szybszych przeciwników. Świadoma prezentacja zmniejszyła liczbę niepotrzebnych strzałów nauczonego agenta.

W scenariuszu Trudne zbieranie apteczek (ang. Health gathering supreme) ekspert podczas świadomej prezentacji zawsze wychodził z rogów obracając się w tę samą stronę i w miarę możliwości odwracał się od tras z minami. Będąc otoczony przez miny wybierał trasę jak najbardziej odległą od nich. Świadoma prezentacja prawie całkowicie wyeliminowała wpadanie w nieskończone pętle ruchów w rogach. W niektórych sytuacjach zdarzało się, że agent zawracał za to w ciasnych, ale możliwych do przejścia korytarzach - było to zachowanie wyraźnie nieoptymalne, ale bez zauważalnego wpływu na osiągane wyniki. Niestety, świadoma prezentacja nie wyeliminowała wchodzenia w miny. Wynik punktowy agenta zwiększył się istotnie po zastosowaniu świadomej prezentacji.

To, jak ważna jest świadoma prezentacja widoczne było przy zwiększaniu wielkości trajektorii eksperta użytych do nauki klasyfikatora. Dla Obrona środka (ang. Defend the center), który jest prostszym scenariuszem i dla którego zysk ze świadomej prezentacji był mniej zauważalny, zwiększanie liczby trajektorii uczących prowadziło do wyższych wyników. Dla bardziej skomplikowanego Trudne zbieranie apteczek (ang. Health gathering supreme) agent nauczony na podstawie małej liczby trajektorii świadomego eksperta przewyższał agenta nauczonego na większej liczbie trajektorii nieświadomego eksperta i agenta nauczonego na mieszance trajektorii.

3.6.3 Wnioski

Dla bardziej skomplikowanych scenariuszy świadoma prezentacja eksperta jest prostym i bardzo skutecznym sposobem eliminowania części oczywistych błędów popełnianych przez agenta. Dla niektórych problemów i sytuacji może wypełniać zadanie postawione przed 3.5 w wygodniejszy i bardziej naturalny sposób. Świadoma prezentacja nie jest formalną metodą, a raczej wytyczną.

Dzięki temu można ją z powodzeniem stosować w połączeniu z innymi technikami uczenia z ekspertem.

3.7 Wnioski z porównania metod

Rozdział 4

Wyniki eksperymentalne

4.1 Gromadzenie trajektorii eksperta

Do zbierania trajektorii eksperta przeznaczony jest dedykowany program napisany w języku python, *spectator.py*, opisany wcześniej w rozdziale 3.3.2. Program powstał poprzez modyfikację programu o tej samej nazwie dołączonego do środowiska VizDoom.

4.1.1 Konfiguracja VizDoom

Instancja VizDoom inicjalizowana jest z następującymi ustawieniami:

```
game.set_window_visible(True)
game.set_mode(Mode.SPECTATOR)
```

Tryb SPECTATOR, wbrew intuicji, umożliwia komputerowi obserwowanie gry człowieka - eksperta. W tym trybie gra reaguje na wejście z klawiatury i myszki, jak przy normalnej grze w Dooma oraz umożliwia odczytywanie aktualnego stanu, ostatnio wykonanej akcji i wartości ostatnio otrzymanej nagrody. Przejście do następnego stanu jest osiągane poprzez wywołanie metody *game.advance_action()*. Tryb SPECTATOR jest synchroniczny, co oznacza że gra oczekuje na wywołanie *game.advance_action()* przed przejściem do następnego stanu (zatrzymanie przetwarzania w programie skutkuje zatrzymaniem przetwarzania w grze). Nie oznacza to jednak, że w tym trybie gra oczekuje z wykonamiem każdej klatki, aż gracz wykona jakiś ruch - brak ruchu ze strony gracza również interpretowany jest jako ruch.

```
        game.advance_action()
        next_state = game.get_state()
        last_action = game.get_last_action()
        reward = game.get_last_reward()
```

Obrazy z gry zapisywane i przetwarzane są nie w kolorze, ale w skali szarosci. Dzięki temu zajętość pamięciowa każdego ze stanów jest trzykrotnie mniejsza niż kolorowego odpowiednika. Według autorów platformy VizDoom i według krótkich eksperymentów przeprowadzonych w ramach tej pracy spadek jakości agentów w stosunku do kolorowego trybu jest pomijalny.

```
game.set_screen_format(ScreenFormat.GRAY8)
game.set_screen_resolution(ScreenResolution.RES_640X480)
```

Obraz oglądany i zapisywany jest w rozdzieczości 640 na 480 pikseli. Użycie mniejszego obrazu podczas gry sprawiłoby, że mniejsze szczegóły byłyby trudne do rozróżnienia, co znacząco utrudniałoby działanie ludzkiego gracza. Obrazy są zapisywane w pełnej rozdzielczości pomimo, że użyte algorytmy wykorzystują obraz w znacznie mniejszym rozmiarze. Dzięki temu raz zebrane

trajektorie eskperta mogą być przeskalowywane do wielu różnych rozmiarów z minimalną stratą jakości obrazu. Skutkiem ubocznym jest bardzo duża wielkość plików z zapisanymi trajektoriami eksperta (do 4GB dla pliku z zapisem 6 tysięcy króków gry).

4.1.2 Zapis i odczyt trajektorii

Zapis

W każdym kroku działania programu *spectator.py* odczytywane są wartości:

- stanu początkowego
- wykonanej akcji
- stanu, w którym znalazł się agent po wykonaniu akcji
- otrzymanej nagrody
- informacji, czy dany stan jest końcowy dla danego epizodu

Te pięć wartości łączone jest w jedną krotkę, a każda krotka dodawana jest do listy krotek.

Program *spectator.py* pozwala ekspertowi odbyć zadaną liczbę epizodów, po czym serializuje listę krotek z zapisem trajektorii do pliku za pomocą biblioteki *pickle*.

Dla scenariuszy *Trudne zbieranie apteczek* i *Obrona środka* na raz zbierane są trajektorie 3 epizodów. Przy poprawnej grze każdy z epizodów składa się z 1,5 do 2 tysięcy kroków, których odbycie zajmuje od jednej do półtorej minuty. Pięć minut ciągłej prezentacji pozwala na wygenerowanie danych wystarczających do nauki i jest wystarczająco krótkim czasem, żeby nie znużyć eksperta. Dla scenariusza *Podstawowego* na raz zbierane są trajektorie 10 epizodów. Ten scenariusz jest krótki w porównaniu do poprzednich, a każdy epizod trwa zaledwie kilka sekund.

Zestaw danych uczących

Dla każdego scenariusza (*Podstawowy*, *Obrona środka* i *Trudne zbieranie apteczek*) i dla każdego rodzaju eksperta (*Zwykły ekspert* i *Świadomie prezentujący ekspert*) wygenerowano po 5 zbiorów trajektorii (czyli odpowiednio po 50 i 15 epizodów dla scenariusza *Podstawowego* i pozostałych.

Odczyt

Program agenta otrzymuje listę plików z paczkami trajektorii do wykorzystania i informację o maksymalnej liczbie klatek, jaką ma wczytać. Program przetwarza pliki w losowej kolejności. Z każdego pliku trajektorie odczytywane są po kolei, aż limit odczytanych klatek nie zostanie osiągnięty lub aż wszystkie pliki nie zostaną przetworzone. Dane w ramach jednego pliku nie są przetwarzane w losowej kolejności, ponieważ:

- agent mógłby wyciągać dodatkowe informacje na podstawie przebytej sekwencji stanów (np. odtwarzać wartości q)
- liczba kroków z poszczególnymi akcjami może być silnie niezbalansowana, a losowe próbkowanie trajektorii mogło by skutkować zestawem danych zupełnie pozbawionym informacji o niektórych akcjach

Odczytane klatki zapisywane są bezpośrednio do pamięci powtórek agenta, po uprzednim przeskalowaniu obrazów-stanów do zadanej rozdzielczości.

Na każdym etapie uczenia program pomija klatki, w których nie została wykonana żadna akcja. To zachowanie opiera się na założeniu, że optymalne zachowanie agenta można osiągnąć bez wykonywania ruchu „czekania", czyli braku ruchu, a klatki z akcją „czekania" pojawiające się w trajektoriach eksperta są artefaktami spowodowanymi nieoptymalną grą eksperta lub nieoptymalnym interfejsem zbierania danych eksperta.

4.2 Kopiowanie zachowań - wyniki

4.3 Świadomie prezentujący ekspert - wyniki

4.4 DAgger - wyniki

4.5 Inne podejścia - wyniki

4.6 Zestawienie wyników

Rozdział 5

Wnioski i perspektywy rozwoju

Literatura

[Abbeel et al., 2008] Abbeel, P., Dolgov, D., Ng, A. Y., and Thrun, S. (2008). Apprenticeship learning for motion planning with application to parking lot navigation. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, September 22-26, 2008, Acropolis Convention Center, Nice, France*, pages 1083–1090.

[Bellman, 1954] Bellman, R. (1954). The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60(6):503–515.

[Bengio et al., 2015] Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099.

[Brafman and Tennenholtz, 2002] Brafman, R. I. and Tennenholtz, M. (2002). R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231.

[Chang et al., 2015] Chang, K., Krishnamurthy, A., Agarwal, A., III, H. D., and Langford, J. (2015). Learning to search better than your teacher. *CoRR*, abs/1502.02206.

[Coates et al., 2009] Coates, A., Abbeel, P., and Ng, A. Y. (2009). Apprenticeship learning for helicopter control. *Commun. ACM*, 52(7):97–105.

[Crites and Barto, 1996] Crites, R. and Barto, A. (1996). Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 8*, pages 1017–1023. MIT Press.

[Jaskowski, 2016] Jaskowski, W. (2016). Uczenie ze wzmocnieniem — generalizacja i zastosowania (materiały wykładowe).

[Kempka et al., 2016] Kempka, M., Wydmuch, M., Runc, G., Toczek, J., and Jaskowski, W. (2016). Vizdoom: A doom-based AI research platform for visual reinforcement learning. *CoRR*, abs/1605.02097.

[Mataric, 1994] Mataric, M. J. (1994). Reward functions for accelerated learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 181–189. Morgan Kaufmann.

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[Osband et al., 2016] Osband, I., Blundell, C., Pritzel, A., and Roy, B. V. (2016). Deep exploration via bootstrapped DQN. *CoRR*, abs/1602.04621.

[Ratliff et al., 2006] Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. (2006). Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 729–736, New York, NY, USA. ACM.

[Ross and Bagnell, 2010] Ross, S. and Bagnell, D. (2010). Efficient reductions for imitation learning. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 661–668, Chia Laguna Resort, Sardinia, Italy.

[Ross et al., 2010] Ross, S., Gordon, G. J., and Bagnell, J. A. (2010). No-regret reductions for imitation learning and structured prediction. *CoRR*, abs/1011.0686.

[Samuel, 1959] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229.

[Schaal, 1999] Schaal, S. (1999). Is imitation learning the route to humanoid robots?

[Schaul et al., 2015] Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. *CoRR*, abs/1511.05952.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[Stadie et al., 2015] Stadie, B. C., Levine, S., and Abbeel, P. (2015). Incentivizing exploration in reinforcement learning with deep predictive models. *CoRR*, abs/1507.00814.

[Sutton, 1988] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44.

[Tesauro, 1992] Tesauro, G. (1992). Temporal difference learning of backgammon strategy. In *Machine Learning Proceedings 1992*, pages 451 – 457. Morgan Kaufmann, San Francisco (CA).

Instytut Informatyki, Wydział Informatyki
Politechnika Poznańska

Skład przy użyciu systemu LaTeX.

BibTeX:

```
@mastersthesis{ key,
    author = "Wojciech Kopeć ",
    title = "{Uczenie przez demonstrację na podstawie informacji obrazowej w środowisku 3D }",
    school = "Poznan University of Technology",
    address = "Pozna{\'n}, Poland",
    year = "2017",
}
```