Politechnika Poznańska
Wydział Informatyki i Zarządzania
Instytut Informatyki

Praca dyplomowa magisterska

IMITATION LEARNING

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 teoretyczny

1.1 Uczenie przez demonstrację a uczenie nadzorowane

Najprostszym podejściem do uczenia przez demonstrację jest traktowanie go jak każdego innego problemu uczenia nadzorowanego, przy czym w przeciwieństwie do minimalizowania kosztu działania agenta minimalizowana jest różnica pomiędzy pomiędzy polityką wyuczonego agenta a polityką eksperta. Najprostsze podejście zakłada jednak, 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. Jak dowiedziono w [RGB10] wynikający z tego błąd rośnie kwadratowo w stosunku do czasu trwania epizodów – gdy klasyfikator popełni błąd w odwzorowywaniu polityki eksperta najprawdopodobniej trafi do stanu nieodwiedzonego przez eksperta, co z dużym prawdopodobieństwem oznacza popełnianie następnych błędów, ponieważ uczeń nie miał jak nauczyć się „podnoszenia się" po błędach.

Jednym ze sposobów radzenia sobie z tym problemem jest wprowadzanie małych zmian podczas iteracji polityki, dzięki czemu rozkład stanów dla nowej polityki jest bliski staremu. Idea polega na zaczynaniu od polityki całkowicie identycznej z polityką eksperta i stopniowym przechodzeniu na politykę wyuczoną. Aby to osiągnąć można wymagać, aby podczas uczenia uczeń mógł w każdej chwili zapytać eksperta, jakie akcje ekspert podjąłby w danym stanie. Dany układ wymaga większej interakcji, ale może być zrealizowany dla wielu z praktycznych przykładów wykorzystania uczenia przez demonstrację.

Pierwszym podejściem opisywanym przez [RGB10] jest uczenie w przód. 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 systemu i dokładnego odtworzenia uzyskanego wcześniej stanu, co w wielu przypadkach nie będzie możliwe do zrealizowania.

W celu wyeliminowania tych ograniczeń [RGB10] proponują Iterowany Probabilistyczny Mieszający algorytm. 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,

przy czym prawdopodobieństwo wyboru nowej polityki jest niewielkie. Algorytm zaczyna od dokładnego wykonywania akcji eksperta. W każdej kolejnej iteracji algorytmu prawdopodobieństwo odpytania eksperta jest coraz niższe i zbiega się do 0. Opisane rozwiązanie zostało z powodzeniem przetestowane na przykładzie grania w proste gry, gdzie danymi wejściowymi był obraz z ekranu. Autorzy zdecydowali się na klasyfikator wybierający konkretne akcje dla danego stanu, zamiast częściej używanego w uczeniu ze wzmocnieniem klasyfikatora odwzorowującego funkcję kosztu. Wadą tego podejścia jest brak odrzucania nieskutecznych polityk podczas iteracji, co może prowadzić do niestabilnych wyników.

Wykorzystanie analogicznego rozwiązania proponują [BVJS15]. Ich propozycja zakłada wybieranie z prawdopodobieństwem e polityki eksperta i z prawdopodobieństwem $1 - e$ polityki wyuczonej. Początkowa wartość e powinna wynosić 1, aby klasyfikator mógł nauczyć się odtwarzać politykę eksperta. Wraz z postępem nauki e powinno stopniowo maleć do 0, aby klasyfikator miał szanse nauczyć się stanów nieodwiedzonych przez eksperta.

W kolejnej publikacji [RGB10] prezentują nowe podejście, nazwane Agregacją Zbioru Danych. 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 kolejne iteracji algorytm stosuje politykę wygenerowaną w poprzedniej iteracji i dodaje dane uzyskane podczas jej stosowania do zbioru danych, po czym trenuje klasyfikator by odwzorowywał zachowanie eksperta na całym zbiorze danych. Podobnie jak w poprzednim algorytmie, żeby przyspieszyć uczenie na pierwszych etapach algorytmu, dodano opcjonalną możliwość odpytania eksperta o jego wybór akcji. Uzyskane z pomocą tej metody wyniki są wyraźnie lepsze od wyników uzyskanych za pomocą metody opisanej w poprzednim paragrafie.

1.2 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 [CKA+15] 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.

1.3 Eksploracja

Podstawowym i często używanym podejściem do eksploracji jest wspomniany wcześniej e-zachłanny algorytm, w którym agent z zadanym prawdopodobieństwem e zamiast akcji optymalnej względem aktualnej polityki wykonuje akcję losową. Takie zachowanie jest nieskuteczne, kiedy optymalne zachowanie agenta wymaga zaplanowania złożonych lub dalekosiężnych planów.

Prostym, ale skutecznym i posiadającym teoretyczne gwarancje zbieżności algorytmem jest zaproponowany w [BT02] 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 spodzie-

wana 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.

W [SLA15] autorzy zaproponowali rozwiązanie, które pozwala ocenić, w jakim stopniu odwiedzony stan jest dla agenta nowością. Opiera się ono na stworzeniu aproksymatora, 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 [SLA15] 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 głęboką sieć neuronową do przewidywania następnego stanu i wykorzystano jedną z ukrytych warstw tej sieci o mniejszej liczbie jednostek 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. Zakodowanymi stany używane były do wytrenowania właściwego, prostszego aproksymatora, na podstawię błędy którego określano nowość stanu. Dla każdego przejścia między stanami przyznawano bonusową nagrodę zależną od nowości. 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.

Innym taktykę dywersyfikacji przeszukiwania przy wykorzystaniu głębokiej sieci neuronowej zaprezentowano w [OBPR16]. Podobnie jak w [SLA15] 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.

Rozdział 2

Experymenty

2.1 Scenariusze

Eksperymenty przeprowadzono na następujących scenariuszach.

2.1.1 Basic

Sceneria składa się z prostokątnego pomieszczenia. Agent jest w jednym końcu pomieszczenia, a w losowym miejscu pod przeciwległą ścianą jest pojedynczy, nieruchomy przeciwnik. Agent może atakować i poruszać się bokiem w lewo i prawo. Strategia optymalna polega na przesunięciu się w kierunku przeciwnika i oddaniu pojedynczego strzału.

2.1.2 Defend the center

Sceneria składa się z kolistej areny. Agent jest na środku areny, a na jej krańcach losowo pojawiają się przeciwnicy, którzy poruszają się w stronę agenta, a po dotarciu do niego atakują. Agent może atakować i kręcić się w okół własnej osi w lewo i prawo. Strategia optymalna polega na kręceniu się w kółko, ignorowaniu odległych przeciwników i strzelaniu do biskich.

2.2 Bootstrapowane DQN - prosta implementacja

Celem tego eksperymentu było zaimplementowanie prostej wersji Bootstrapowanej DQN opisanej w [OBPR16] i sprawdzenie jej zachowania pod względem kierowania eksploracją i modelowania niepewności agenta.

2.2.1 Implementacja

Zaproponowana w [OBPR16] wersja ostateczna BDQN opiera się na jednej głębokiej sieci neuronowej, w której najwyższe warstwy „rozdzielały się" na K róznych „głów" sieci, gdzie każda głowa odpowiadała jednej funkcji Q. Każda z funkcji Q jest uczona na podstawie osobnego próbkowanego zestawu danych, do którego z zadanym prawdopodobieństwem p dołączane są wygenerowane podczas uczenia próbki. Na początku każdego epizodu losowo wybierana jest aktywna głowa, która w danym epizodzie będzie służyła za funkcję Q - dzięki temu zachowanie agenta w ramach każdego epizodu jest nieco różne, ale spójne.

Uproszczona wersja BDQN wspomniana w artykule i zaimplementowana w tym eksperymencie zakłada wykorzystanie K niezależnych sieci zamiast jednej sieci z wieloma „głowami". Warto zwrócić uwagę, że to uproszczona implementacja dokładniej realizuje bootstraping, zapewniając

niezależność sieci - w łączonej sieci górne warstwy uczyły się na próbkowanych danych, natomiast dolne uczyły się na pełnym zbiorze - kosztem niezależności znacznie skrócono czas uczenia.

Miarę pewności sieci, mierzoną dla każdej podjętej decyzji, przyjęto jako liczbę głów zgadzających się z decyzją aktywnej głowy podzieloną przez liczbę głów K. Ostateczna miara pewności dla danego epizodu była uśrednioną wartością pewności wszystkich decyzji podjętych w tym epizodzie.

2.2.2 Ustawienia

Eksperyment przeprowadzono przy użyciu scenariusza *basic*. Punktem odniesienia był przykładowy agent dostarczony przez autorów VizDooma, opierający się na DQN. Kod agenta (a w szczególności architektura sieci neuronowej) posłużyła za podstawę eksperymentalnego agenta wykorzystującego BDQN. Dzięki temu jedynym zmiennym elementem w eksperymencie było badane bootstrapowanie danych dla sieci.

Eksperymenty przeprowadzano dla liczby podsieci $K = 5, 10$ i prawdopodobieństwa uwzględnienia próbki $p = 0.5, 0.75, 0.9, 1$. Agenci uczli się przez 20 epok po 2000 epizodów każda. Celem eksperymentu było wstępne zbadanie przydatności BDQN i zaobserwowanie ogólnych trendów, dlatego badania nie były wielokrotnie powtarzane.

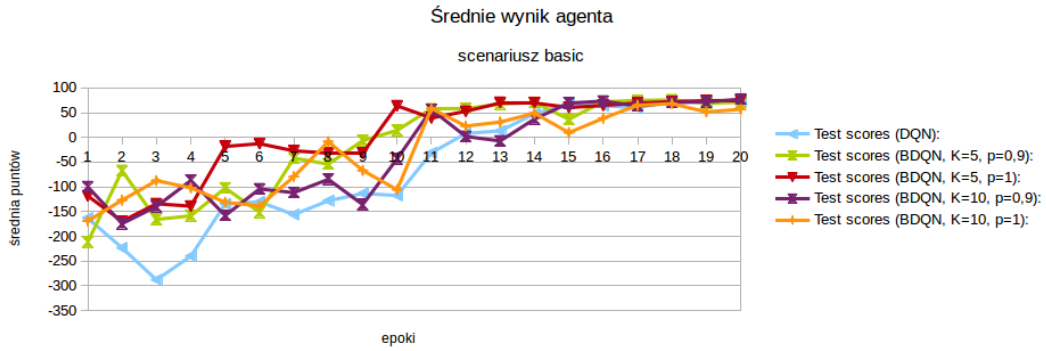
2.2.3 Wyniki

Eksperyment wykazał, że BDQN może uczyć się znacznie szybciej (w kontekscie liczby epizodów, nie czasu) niż zwykła DQN. Dodatkowo przyjęta miara pewności sieci wyraźnie koreluje z wynikami uzyskiwanymi przez agenta.

Jak widać na wykresie 2.1, wyniki BDQN przekraczają granicę 0 i dochodzą do ostatecznych wartości kilka epok wcześniej niż DQN. Wszystkie rozwiązania zbiegają się do porównywalnych wyników - empiryczna analiza zachowań agenta wskazuje, że nie jest to jeszcze zachowanie optymalne, ale sensowne. Akcje agenta sugerują niedokładność percepcji, co może wskazywać na zbyt prostą lub zbyt małą sieć. Zaimplementowane BDQN działa szybciej biorąc pod uwagę epoki, ale nie czas. Prostota implementacji sprawia, że odbycie każdej epoki zajmuje BDQN do trzech razy więcej czasu niż DQN. Spodziewane jest, że docelowa implementacja będzie porównywalna z DQN.

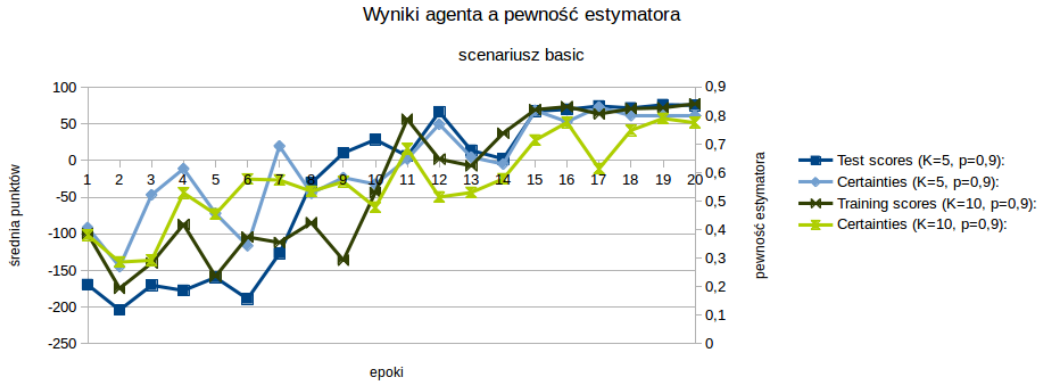
Ciekawe jest zachowanie BDQN dla różnych zestawów parametrów. Autorzy [OBPR16] osiągali najlepsze wyniki dla $p = 0.5$ i rosnące nieznacznie wraz ze wzrostem K. Badany agent dla $p = 0.5$ osiągał gorsze wyniki niż DQN. Jest to zrozumiałe zachowanie - w docelowej implementacji część sieci uczona jest w praktyce na wszystkich danych, podczas gdy badane rodzielne sieci dla $p = 0.5$ są w każdym momencie nauczone dwa razy mniejszym zestawie danych niż analogiczny agent DQN. Warto zwrócić uwagę, że dzięki losowej inicjalizacji wag nawet dla $p = 1$ uzyskane podsieci nie są identyczne.

Wyniki gorsze dla $K = 10$ niż dla $K = 5$ są natomiast sprzeczne z oczekiwaniami, przy czym to zachowanie nie wydaje się istotne przy badaniu na najprostszym scenariuszu.



Rysunek 2.1: Średnie wyniki agentów dla scenariusza *basic*

Jak widać na wykresie 2.2, pewność estymatora wyraźnie rośnie w miarę uczenia i koreluje z wynikami uzyskiwanymi przez agentów.



Rysunek 2.2: Średnie wyniki a pewność estymatora dla scenariusza *basic*

2.2.4 Wnioski

Eksperyment z użyciem prostej implementacji BDQN i prostego scenariusza *basic* wykazał, że BDQN ma potencjał prowadzenia skutecznej eksploracji samo w sobie, a co ważniejsze, umożliwia sensowne estymowanie stopnia pewności sieci, co stanowi podstawę do bardziej zaawansowanych technik kierowania eksploracją.

Literatura

[BT02] R. I. Brafman, M. Tennenholtz. R-MAX—a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[BVJS15] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099, 2015.

[CKA+15] Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, John Langford. Learning to search better than your teacher. *CoRR*, abs/1502.02206, 2015.

[OBPR16] Ian Osband, Charles Blundell, Alexander Pritzel, Benjamin Van Roy. Deep exploration via bootstrapped DQN. *CoRR*, abs/1602.04621, 2016.

[RGB10] Stéphane Ross, Geoffrey J. Gordon, J. Andrew Bagnell. No-regret reductions for imitation learning and structured prediction. *CoRR*, abs/1011.0686, 2010.

[SLA15] Bradly C. Stadie, Sergey Levine, Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *CoRR*, abs/1507.00814, 2015.

Instytut Informatyki, Wydział Informatyki
Politechnika Poznańska

Skład przy użyciu systemu LaTeX.

BibTeX:

```
@mastersthesis{ key,
    author = "Wojciech Kopeć \and ",
    title = "{Imitation Learning}",
    school = "Poznan University of Technology",
    address = "Pozna{\'n}, Poland",
    year = "2017",
}
```