| (ICICML) کنفرانس بین المللی 2023 پردازش تصویر، بینایی کامپیوتری و یادگیری ماشین
979-8-3503-3141-7/23/31.00$ ©2023 IEEE | DOI: 10.1109/ICICML60161.2023.10424889

تقطیر دانش و پاداش شکل دهی دیپ یادگیری تقویتی

یو می*
گروه مهندسی برق و کامپیوتر
دانشگاه ایالتی اوهایو
کلمبوس، اوهایو، 42310، ایالات متحده آمریکا
*mei.103@osu.edu

خلاصه- تقطیر دانش تکنیکی است که دانش را از یک مدل پیچیده تر، که به عنوان مدل معلم نامیده می شود، به یک مدل کمتر پیچیده، به نام مدل دانش آموز، تقطیر می کند تا نیازها را کاهش دهد و کارایی مدل دوم را بهبود بخشد. علیرغم دستاوردهای اجرای سیاست‌هایی که از طریق این روش‌های (DRL)، متعدد یادگیری تقویتی عمیق پیشرفته در مقیاس بزرگ به دست می‌آیند به دلیل نبود منابع کافی مانع می‌شود. تقطیر دانش مدل‌های یادگیری عمیق تقویتی هنوز به اندازه تقطیر دانش سایر مدل‌های یادگیری عمیق شناخته شده مورد توجه قرار نگرفته است. این مقاله مروری بر تقطیر دانش ارائه می‌کند و اجزای مختلف، روش‌شناسی و کاربردهای بالقوه آن را مورد بحث DRL قرار می‌دهد. این یک تجزیه و تحلیل جامع از کارهای مرتبط را ارائه می دهد و پیشرفت های ایجاد شده در این زمینه را برجسته می کند. یک روش پایه برای تقطیر دانش یادگیری تقویتی عمیق با ارزیابی اثربخشی آن از طریق آزمایشات پیشنهاد شده است. پیاده‌سازی، Reward Shaping، علاوه بر این، یکی دیگر از تکنیک‌های بهینه‌سازی و با تقطیر دانش مقایسه می‌شود و بینش‌هایی را در مورد کاربرد آن‌ها در حوزه‌های مختلف ارائه می‌دهد. نتایج پتانسیل تقطیر دانش و شکل‌دهی پاداش را به عنوان تکنیک‌های قدرتمند برای افزایش عملکرد عوامل یادگیری تقویتی نشان می‌دهد.

کلمات کلیدی-یادگیری عمیق؛ یادگیری تقویتی؛ تقطیر دانش؛ عملکرد از دست دادن

مقدمه

روشهای (KD) و تقطیر دانش (DRL) هر دو یادگیری تقویتی عمیق محبوب در یادگیری ماشین هستند. یادگیری تقویتی عمیق با سایر روش های یادگیری عمیق متفاوت است زیرا خط مشی به روز شده شامل یک شبکه عصبی ها DRL یکی از ساده ترین Deep Q-Network (DQN) .[عمیق است [1،2 ،است. ایده آن شبیه به مینی بچ است. به‌جای به‌روزرسانی هر بار یک اقدام دنباله‌ای از حالت، اقدام و پاداش را به عنوان داده ورودی ذخیره می‌کند و DQN سپس از روش ضرر هوبر برای محاسبه پاداش مورد انتظار استفاده می‌کند. ضرر با استفاده از پاداش انتظار محاسبه می شود که در نهایت برای به روز رسانی شبکه استفاده می شود.

به عنوان یک تکنیک قوی برای حل مسائل پیچیده DRL ،در سال های اخیر تصمیم گیری ظاهر شده است [3]. با این وجود، موفقیت یادگیری تقویتی عمیق به شدت به توانایی آموزش مدل‌های بزرگ و محاسباتی گران‌قیمت بستگی دارد، که اغلب به منابع و زمان قابل توجهی نیاز دارد. این محدودیت در سناریوهایی که منابع محاسباتی محدود هستند یا تصمیم گیری در زمان واقعی مورد نیاز است، چالش هایی ایجاد می کند که می تواند در بسیاری از کاربردهای عملی بازدارنده باشد [4]. تقطیر دانش یک راه حل امیدوارکننده برای رسیدگی ارائه می دهد

979-8-3503-3141-7/23/31.00$ ©2023 IEEE

این چالش با تسهیل انتقال شبکه سیاست از یک مدل پیچیده به یک مدل ساده تر، هدف این مقاله ترکیب این دو .[5،6] DRL در نتیجه کاهش نیازهای محاسباتی

مفهوم یادگیری ماشینی با هم، و بررسی نحوه کارکرد آنها در شرایط مختلف است.

که به عنوان تقطیر سیاست نیز شناخته می شود، DQN، علاوه بر اصلاح تکنیک های بهینه سازی مختلفی برای یادگیری تقویتی وجود دارد. شکل دادن به RS پاداش یکی از آنها است که برای تسریع روند آموزش استفاده می شود. هدف تسریع یادگیری استراتژی‌های مؤثر با تغییر سیگنال‌های پاداش است، بنابراین به عامل کمک می‌کند ساختار و ویژگی‌های محیط را بهتر درک کند و در عین حال مقایسه RS راهنمایی صریح‌تری ارائه دهد. این مقاله کارایی تقطیر دانش را با می‌کند و این دو روش را برای دستیابی به نتایج بهتر ترکیب می‌کند.

II. کار مرتبط

این بخش ادبیات موجود در مورد تقطیر دانش و شکل‌دهی پاداش را در حوزه یادگیری تقویتی عمیق مرور می‌کند. تفاوت‌ها و شباهت‌های بین این تکنیک‌ها نیز مورد بررسی قرار می‌گیرد و بینشی در مورد نقاط قوت و محدودیت‌های آنها ارائه می‌کند.

[7] (OD3) مطالعه قبلی روشی به نام یادگیری تقویتی عمیق روی دستگاه را برای انتقال موثر دانش مقطر برای کنترل دستگاه های لبه در سیستم های انتقال دانش و فشرده OD3 محاسباتی لبه محدود با منابع معرفی کرد. هدف سازی مدل های خط مشی به طور همزمان در طول آموزش بر روی دستگاه های را با پیاده سازی آن بر روی OD3 لبه با تقطیر دانش است. این مقاله عملکرد یک سیستم تعبیه شده تجاری با منابع سخت افزاری محدود ارزیابی کرد. نتایج علیرغم استفاده از یک شبکه خط مشی بسیار OD3 تجربی نشان داد که کوچکتر، به عملکرد قابل مقایسه با راه حل های مبتنی بر ابر دست یافت. علاوه بر این، زمان تمرین برای خط مشی لبه به طور قابل توجهی در مقایسه با تمرین را در سیستم‌های محاسباتی لبه‌ای OD3 از ابتدا کاهش یافت. این نتایج اثربخشی که منابع محدود هستند ثابت کرد و بر اهمیت نه تنها تقطیر سیاست کارآمد بلکه بر حفظ منابع نیز تأکید کرد.

یادگیری تقویتی در حوزه های مختلف به کار گرفته شده است، اما تحقیقات کنونی اغلب فاقد رویه های لازم برای اجرای مؤثر یادگیری تقویتی، مانند شکل دادن به پاداش، یادگیری برنامه درسی، و سایر تکنیک ها برای تجزیه وظایف پیچیده به اجزای کوچکتر هستند. برای سهولت تلاش‌های مهندسی، این مطالعه استاندارد Q اثربخشی این تکنیک‌ها را در تقویت یک عامل یادگیری عمیق بررسی کرد [8]. برجسته‌ترین تکنیک‌ها سپس در سه محیط مختلف، در کنار و اولیه اعمال شدند DQN یک عامل

انجام شد. نتایج اولیه نشان داد که RL آزمایش‌هایی برای ارزیابی تأثیر آن‌ها بر تقسیم عمدی کار به وظایف فرعی و از پیش تنظیم اقدامات عامل بسیار است. خوبی به نظر می رسید در بهبود عملکرد عامل موثر است، به ویژه زمانی که اجرای این تکنیک ها دشوار نبود. بنابراین، ترکیب نیز رویکرد RS با KD، سایر ترفندهای تقطیر، مانند

$$\mathcal{L} =^{1}{}_{|��|}\sum{}_{(�������������",��)}������\mathcal{L}(��)(2)_1$$

رویکرد 1 (1

$$\mathcal{L}(��) = \{ ������������$$

$$|��| \leq 1 \quad {}_2��^2�������$$

|��|

$$-\frac{1}{2}\text{��������������������� ������� (3)}$$

سه بخش عمده در تقطیر دانش وجود دارد: مدل معلم، مدل دانش آموز و مدل مقایسه.

با پیکربندی 128 نورون که در سه لایه Deep Q-Network مدل معلم از توزیع شده اند استفاده می کند.

ساخته شده است که شامل 96 نورون است که DQN مدل دانش آموز با یک در سه لایه پخش شده اند که تقریباً سه چهارم اندازه مدل معلم است. شبکه از پیش آموزش دیده الگوی معلم به مدل دانش آموز به ارث می رسد.

علاوه بر این، داده های مدل اضافی برای مقایسه ثبت می شود. این مدل مقایسه به طور یکسان با مدل دانش آموز ساخته شده است و فرآیند آموزشی مشابه مدل معلم را طی می کند.

2) 2 رویکرد

که معادله T و دانش را نیز می توان با تغییر دمای DRL راندمان تقطیر داده شده را برآورده می کند تحت تأثیر قرار داد. معلم احتمالات هدف نرم$_{��}$��و احتمالات هدف نرم دانش آموز$_{��}$��را می توان از لاجیت معلم $_{��}$��و دانش آموز محاسبه کرد که با نشان داده می شوند$_{��}$��و متقابلا.

$$\text{-}^{����}{}_{��/��}\Sigma\ ��^{��}{}_{��/�����} \quad \text{�����} \quad \Sigma\ ��^{��}{}_{��/��}\ �� \quad)(4)$$

و DRL پیش‌فرض در عملکرد استاندارد نرم‌افزار 1 است. رفتار تقطیر نیز هنگامی که دما به مقادیر دیگر تغییر می‌کند کنترل می‌شود، Knowledge در حالی که سایر شرایط ثابت می‌مانند.

3) 3 رویکرد

در شرایط مختلف بررسی شده است. DRL در آزمایشات قبلی، رفتار تقطیر با این حال، آن را با سایر تکنیک های بهینه سازی مقایسه نشده است. شکل دادن به پاداش خطی یکی از آنهاست که با تنظیم یک موقعیت هدف، پاداش خروجی را به صورت خطی تغییر می دهد. هنگامی که موقعیت فعلی به موقعیت هدف CartPole-v1، نزدیکتر می شود، پاداش به تدریج افزایش می یابد. در محیط پاداش را می توان با معادله زیر نشان داد:

$$�� = (���_{��} - |��_0|)/(��_{��} -- �_{��})(5)$$

نشان دهنده موقعیت مشاهده شده است، در حالی که$_{��}$��موقعیت$_0$�� شروع است.$_{��}$��نقطه وسط بین نقطه شروع و نقطه پایان مثبت است. معادله $r = 1$.$_{��}$�� >فوق فقط زمانی کار می کند که$_{��}$��� < $_0$�� این معادله باید برای محیط .$r = 0$ وقتی$_0$��=��$_{��}$��، و در غیر این صورت های مختلف اصلاح شود، اما رابطه خطی همچنان ادامه دارد.

را قادر می‌سازد DRL یک ترفند مؤثر است که (RS) شکل‌دهی پاداش دانش دامنه را در فرآیند یادگیری ادغام کند. با این وجود، روش‌های موجود، مانند شکل‌دهی پاداش مبتنی بر پتانسیل، اغلب بر تابع پاداش شکل‌دهی ثابت تکیه به‌عنوان یک موضوع RS می‌کنند. این مقاله تحقیقاتی با فرمول‌بندی استفاده از بهینه‌سازی باینری [9] به این اشکال پرداخت. قسمت پایین شامل بهینه سازی خط با گرادیان حاصل از پارامترهای تابع وزن است، در حالی RS مشی با استفاده از که قسمت بالایی با هدف پارامترسازی و بهبود تابع وزن شکل دهنده برای دستیابی به بالاترین پاداش خروجی ممکن از طریق سه الگوریتم یادگیری با سایر تکنیک RS پیشنهادی است. این مقاله با موفقیت مزایای بالقوه همکاری های بهینه سازی را نشان داد.

روش .III

الف. سابقه

محیطی است که برای آزمایش ها و ارائه Gymnasium از CartPole-v1 داده ها استفاده می شود. گاری قصد دارد تا زمانی که ممکن است باقی بماند قبل از اینکه 2.4 واحد از مرکز دور شود یا قطب از روی آن بیفتد [10]. بنابراین، پاداش متناسب با مدت زمان است.

بر عملکرد مدل استفاده می‌شوند، اما

MountainCar-v0 و Pendulum-v1

$$������_{��} =^{1}{}_{��}(���_{��} - ��_{��})$$

نیز برای مطالعه تأثیر محیطهای مختلف

$$=^{1}{}_{��}(^{����}{}_{��/��}$$

آنها فقط در بخش بحث گنجانده شده اند.

شبکه های عصبی و الگوریتم ها دو جنبه جذاب یادگیری تقویتی عمیق هستند، زیرا می توان آنها را با تکنیک های مختلف بهینه سازی یا تقطیر تغییر DQN داد. رویکردهای زیر شامل هر دو سناریو هستند. تقطیر دانش روی یک را تغییر DRL الگوریتم Reward Shaping پایه اعمال می‌شود، در حالی که می‌دهد. توابع تلفات مختلف و دماهای مختلف برای مشاهده نحوه رفتار تقطیر دانش در شرایط مختلف مورد استفاده قرار می گیرند.

ب. روش

از طریق بهینه سازی r به حداکثر رساندن پاداش تجمعی DQN هدف ثابت بین 0 و 1، مجموع را همگرا می کند و γ، سیاست است. ضریب تنزیل عامل را تشویق می کند تا به پاداش های زودتر توجه بیشتری کند.

شبکه‌های عصبی به‌عنوان مونتاژکننده‌های مجدد توابع جهانی عمل می‌کنند و که می تواند بازده مورد انتظار را R→، را تقریب می‌کنند×عمل Q: State تابع تحت یک عمل خاص و یک وضعیت معین پیش بینی کند. با استفاده از این تابع، می توان سیاستی ایجاد کرد که پاداش را به حداکثر می رساند. در آزمایش اول، این کار از معادله بلمن استفاده می کند و خطای اختلاف زمانی را دریافت می کند.��:

$$�� = ��(��\text{،} ���) - (�� + ����������_{��}��(���''\text{،} ���))(1)$$

کمک می کند و مقاومت آن را در برابر δ اتلاف هوبر به به حداقل رساندن MSE نویزدار است افزایش می دهد، زیرا تلفات هوبر مشابه Q پرت زمانی که رفتار می کند. در MAE بین 1- و 1 است یا در غیر این صورت δ زمانی که از محاسبه می شود B یک دسته از انتقال حافظه تکرار نمونه:

هدف اصلی از شامل شدن دما، نرم کردن نتیجه از دست دادن است. دمای

نشان‌دهنده مدت y از توزیع داده‌ها و خط بهترین تناسب، که در آن محور تعداد نمونه‌ها را نشان می‌دهد. هر x ماندگاری سبد خرید در ثانیه است و محور دو محور دارای مرز بالایی 500 هستند.

1) 1 آزمایش

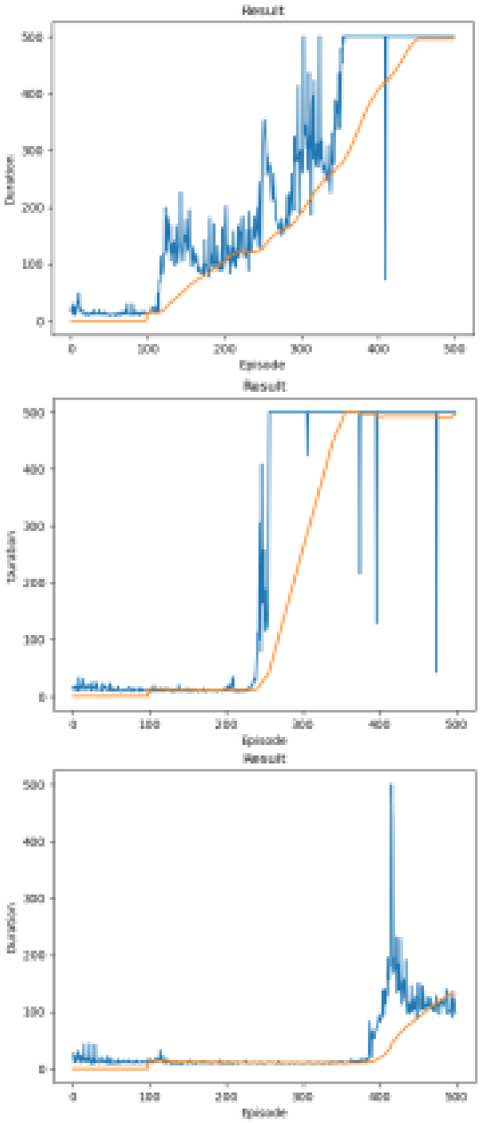
شکل 1 نتایج را پس از ارزیابی و مقایسه مدل معلم، مدل دانش آموز و مدل

آزمایش و نتیجه .IV

الف. آزمایشات

چهار آزمایش زیر از یادگیری تقویتی عمیق بر چهار جنبه مختلف تمرکز دارد: تقطیر دانش، تغییر تابع از دست دادن، تغییر دما، و شکل‌دهی پاداش. نتایج تجربی از طریق نمودار ارائه شده است

مقایسه مربوطه از نظر عملکرد آنها نشان می دهد، در حالی که سایر شرایط را بدون تغییر نگه می دارد.



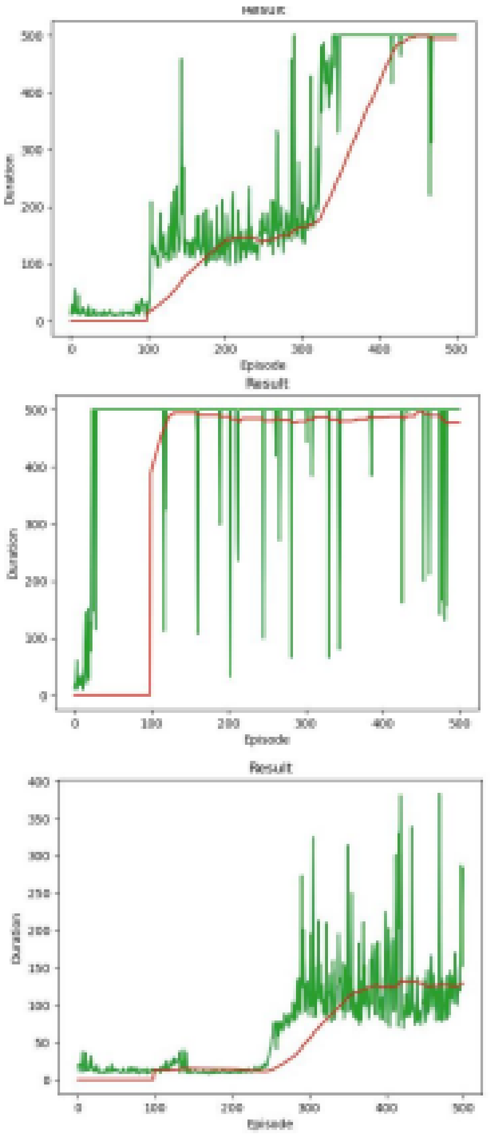
شکل 1. نتایج مدل های مقایسه معلم، دانش آموز و مقایسه در آزمایش 1 (شکل اعتبارات: اصلی)

مدت زمان بعد از 100 قسمت شروع به رشد می کند، در حدود 350 قسمت به بالاترین امتیاز خود یعنی 500 می رسد و تثبیت می شود.

بعد از آن. تقریباً 250 قسمت طول می کشد تا ماژول دانش آموز به بالاترین امتیاز خود دست یابد و در محیط به ثبات برسد. با این حال، ماژول مقایسه نمی تواند بالاترین امتیاز را در 500 قسمت به دست آورد.

2) آزمایش 2

هنگامی که تابع حداکثر دما درگیر باشد، عملکرد تلفات اصلاح می شود. همانطور که در شکل 2 نشان داده شده است تحت تأثیر DRL بنابراین عملکرد قرار می گیرد.

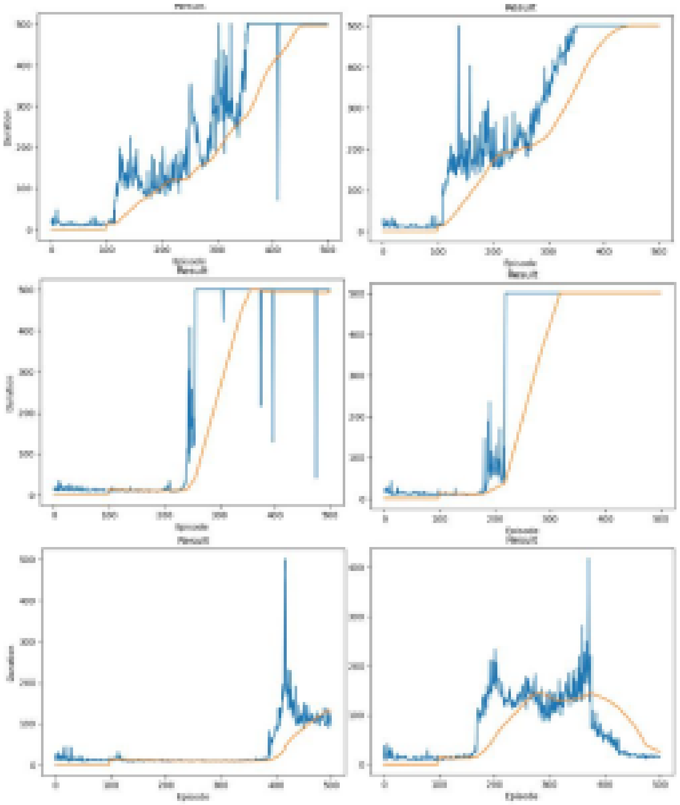


(شکل اعتبارات: اصلی) T=1 شکل 2. نتایج مدل های مقایسه معلم، دانش آموز و در آزمایش 2 با

الگوی معلم کارایی خود را اندکی افزایش داده است، در حالی که الگوی دانش آموز بهبود قابل توجهی دارد. دانش‌آموز در 150 قسمت به بالاترین امتیاز مدت زمان 500 می‌رسد و این مدت را ادامه می‌دهد و موارد پرت را کنار می‌گذارد. از سوی دیگر، مدل مقایسه قادر به دریافت میانگین 200 مدت زمان پس از 500 قسمت نیست، اگرچه زودتر از آزمایش قبلی شروع به افزایش می کند.

3) آزمایش 3

سپس دما به 5 و 10 تغییر می کند، در حالی که نمودارهای داده ها نتیجه مدل معلم، دانش آموز و مقایسه را در شکل 3 نشان می دهند.



شکل 4. نتیجه مدل های معلم، دانش آموز و مقایسه در آزمایش 4 با طرح اصلی و شکل دادن به پاداش (شکل اعتبارات: اصلی)

V. بحث

اجرای تقطیر دانش با موفقیت کارایی شبکه دانشجویی را در دو جنبه زیر افزایش می دهد:

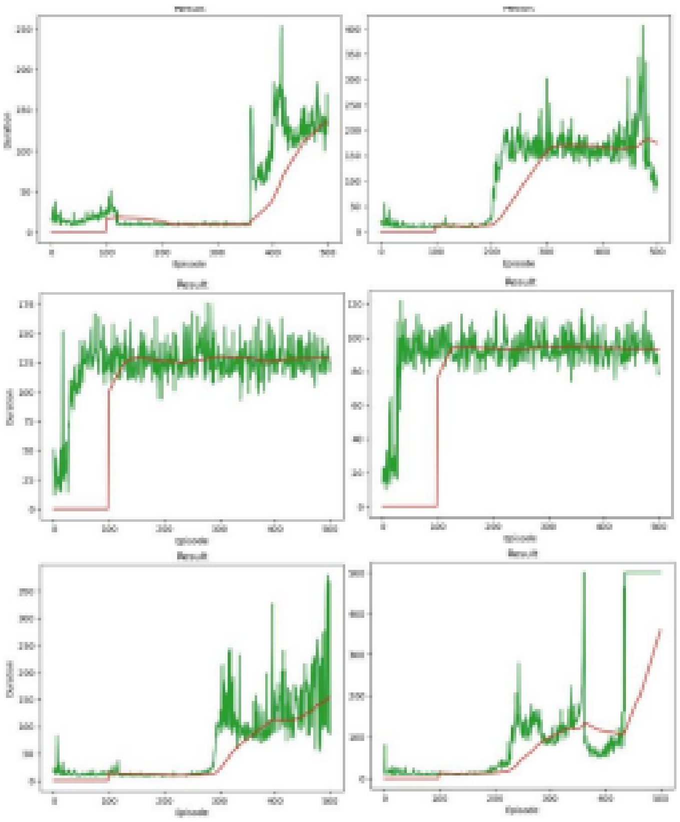
اول، مدل دانش‌آموز پایدارترین فرآیند آموزشی را نشان می‌دهد و به طور مداوم شبکه خود را در جهت دلخواه بهبود می‌بخشد. علاوه بر این، در مقایسه با مدل های دیگر، مدل دانش آموز نرخ یادگیری به طور قابل توجهی بالاتری را نشان می دهد.

دوم، تقطیر دانش نه تنها باعث صرفه جویی در زمان در طول آموزش می شود، بلکه دقت نتایج را نیز افزایش می دهد. ماژول دانشجویی بالاترین امتیاز را کسب می کند و از ماژول مقایسه ای که قادر به رسیدن به این سطح از عملکرد نبود عملکرد بهتری دارد.

از آنجایی که تقطیر هیچ تاثیری بر تابع تلفات ندارد، همچنان همانطور که انتظار می رود کار می کند و مدل دانشجویی عملکرد بهتری دارد. بدون شک، عملکرد پایدار تقطیر دانش، زمانی که تابع تلفات تغییر می‌کند، نسبت به سایر ترفندهای تقطیر مفیدتر است.

نتیجه آزمایش همچنین ثابت می کند که وقتی الگوریتم مدل معلم اصلاح می شود، تقطیر دانش کارایی خود را از دست نمی دهد. این ویژگی باعث می شود که تقطیر دانش در برخی از سناریوها مفیدتر از بهینه سازهای سنتی باشد.

دمای بالاتر می تواند روند تمرین را تسریع کند، اما به قیمت از دست دادن برخی جزئیات در داده های ویژگی است. برعکس، دمای پایین‌تر تمام داده‌های ویژگی را حفظ می‌کند



شکل 3. نتایج مدل های مقایسه معلم، دانش آموز و در آزمایش 3 با T=5 و 10 (شکل اعتبارات: اصلی)

مدل دانش آموز به دلیل تأثیر دمای بالا در مقایسه با مدل معلم شروع به کاهش جزئی در اوج امتیاز می کند، اما همچنان نسبت به مدل مقایسه بسیار بهتر عمل می کند و بالاترین سرعت رشد را حفظ می کند. به طور کلی، تقطیر دانش هنوز در دماهای مختلف قابل اعتماد است.

4) آزمایش 4

آخرین آزمایش شکل‌دهی پاداش خطی را به سه مدل اصلی اضافه می‌کند.

نتایج در شکل 4 نشان می‌دهد که شکل‌دهی پاداش قابلیت بهبود نرخ یادگیری را دارد، اما زمانی که تعداد کل قسمت‌ها نسبتاً کوچک باشد، به اندازه DRL تقطیر دانش مؤثر نیست.

با هم اعمال می شوند، مدل دانش آموز در اوایل 200 RS و KD وقتی قسمت به اوج امتیاز می رسد و ثابت می کند که همکاری این دو تکنیک عملکرد DRL را بیشتر افزایش می دهد.

مناسبی که به طور موثر این تعادل را برقرار می کند، مهم است.

DRL نمودارهای داده نشان می دهد که چگونه این دو روش بر کارایی به راحتی با RS و KD تأثیر می گذارد. موقعیت های مختلف کد باعث می شود

جزئیات اما منجر به کاهش سرعت تمرین می شود. نتایج آزمایش 3 نشان می دهد که با افزایش دما، مدت زمان شروع به رشد بسیار سریعتر کرد، اما مقدار پیک کاهش می یابد. یافتن تعادل مناسب بین جزئیات داده های ویژگی و سرعت تمرین برای دستیابی به فرآیند آموزشی بهینه بسیار مهم است. شناسایی دمای

نیز از طریق دماها و عملکردهای KD هم ترکیب شوند. سازگاری و اثربخشی تلفات مختلف بررسی می شود.

اگرچه به نظر می‌رسد تقطیر دانش عملکرد بهتری نسبت به شکل‌دهی پاداش به عنوان یک تکنیک RS نمی‌تواند جایگزین KD ،دارد CartPole در محیط توانایی KD .بهینه‌سازی شود، زیرا آنها حوزه‌های تخصصی متفاوتی دارند را می توان به RS .کاهش سخت افزار مورد نیاز در کنار بهبود کارایی را دارد طور مستقیم طراحی و پیاده سازی کرد و در زمان اضافی برای آموزش مدل دانش آموز صرفه جویی کرد.

(MountainCar-v0، تکرار آزمایش های قبلی در محیط های مختلف نتایج مشابهی را به همراه دارد. تقطیر دانش مدل های خبره (Pendulum-v1 بهتری را ایجاد می کند و تحت تأثیر تغییر تابع تلفات و دما قرار نمی گیرد. اثر منجر به بهبودهای قابل توجهی در فرآیند آموزش و عملکرد RS و KD ترکیبی کلی مدل می شود.

در آزمایش‌های بالا، تقطیر در یادگیری تقویتی عمیق در توابع مختلف دما و افت اجرا شده است، در حالی که شکل‌دهی پاداش نیز اعمال شده است. با این وجود، هنوز تکنیک های بهینه سازی و تقطیر بسیار زیادی وجود دارد، مانند گرادیان خط مشی، یادگیری برنامه درسی و سلسله مراتب دستی. زنجیره پیچیده تر از آن تکنیک ها عملکرد مدل را بیشتر بهبود می بخشد. رویکرد دیگر استفاده یا Double Deep Q Learning از یک مدل معلم پیشرفته‌تر است، مانند Deep Recurrent Q Learning. روشهای پیشرفته تقطیر مانند Transfer Learning نیز روش خوبی برای بررسی است.

اره. نتیجه

این کار مفهوم تقطیر دانش و شکل‌دهی پاداش اعمال شده بر روی مدل‌های یادگیری تقویتی عمیق را بررسی کرده است. کارهای مرتبط، روش‌شناسی و کاربردهای این تکنیک مورد بحث قرار گرفت. نتایج تجربی ثابت کرد که تقطیر دانش می تواند عملکرد شبکه سیاست را افزایش دهد. این یافته‌ها همچنین پتانسیل شکل‌دهی پاداش را به‌عنوان ابزاری ارزشمند برای بهبود کارایی مدل‌های یادگیری تقویتی عمیق برجسته کردند. علاوه بر این، این مقاله بینش هایی را در مورد کاربرد و چالش های بالقوه مرتبط با تقطیر دانش و یادگیری تقویتی ارائه کرده است. تحقیقات و کاوش بیشتر در این زمینه ها احتمالاً به الگوریتم های یادگیری تقویتی کارآمدتر و مؤثرتر و روش های تقطیر قوی تر منجر می شود.

منابع

[1] V. François-Lavet، P. Henderson، R. Islam، M. G. Bellemare و J. Pineau. مقدمه ای بر یادگیری تقویتی عمیق مبانی و روندها در یادگیری ماشین، جلد. 11 (3-4)، صص 219-354 2018.

[2] جی. گو، بی. یو، اس جی میبانک و دی. تائو. تقطیر دانش: نظرسنجی International Journal of Computer Vision, vol. 129، ص 1789-1819 2021.

[3] پی هندرسون، آر اسلام، پی باخمن، جی. پینو، دی. پرکاپ و دی. میگر. یادگیری تقویتی عمیق که مهم است. در مجموعه مقالات کنفرانس AAAI در مورد هوش مصنوعی، جلد 32 (1)، صفحات 1-8 2018.

[4] T. T. Nguyen، N. D. Nguyen و S. Nahavandi، یادگیری تقویتی عمیق برای سیستم های چند عاملی: بررسی چالش ها، راه حل ها و برنامه ها. معاملات IEEE در سایبرنتیک، جلد. 50 (9)، ص 3826-3839 2020.

[5] Wang, L., & Yoon, K. J. تقطیر دانش و یادگیری دانش‌آموز-معلم برای هوش بصری: یک بررسی و دیدگاه‌های جدید. تراکنش‌های IEEE در تحلیل الگو و هوش ماشینی، جلد. 44 (6)، ص 3048-3068 2021.

[6] P. Chen، S. Liu، H. Zhao و J. Jia، تقطیر دانش از طریق بررسی دانش. در مجموعه مقالات کنفرانس IEEE/CVF در مورد دید رایانه و تشخیص الگو، صفحات 5008-5017 2021.

[7] I. Jang، H. Kim، D. Lee، Y. S. Son، و S. Kim، انتقال دانش برای یادگیری تقویت عمیق روی دستگاه در سیستم‌های محاسباتی لبه محدود منابع. دسترسی IEEE، جلد. 8، ص 146588-146597 2020.

[8] A. Kanervisto، C. Scheller، Y. Schraner، و V. Hautamäki، ترفندهای یادگیری تقویتی تقطیر برای بازی های ویدیویی. در کنفرانس IEEE 2021 در مورد بازی ها، صفحات 01-04 2021.

[9] Y. Hu، W. Wang، H. Jia، Y. Wang، و همکاران. آموزش استفاده از پاداش‌های شکل‌دهی: رویکردی جدید در شکل‌دهی پاداش. پیشرفت‌ها در سیستم‌های پردازش اطلاعات عصبی، جلد. 33، صفحات 15931-15941 2020.

[10] CartPole-v1 نشانی اینترنتی: https://www.gymlibrary.dev/environments/classic_co ntrol/cart_pole/، آخرین دسترسی: 2023/09/24