به نام خدا

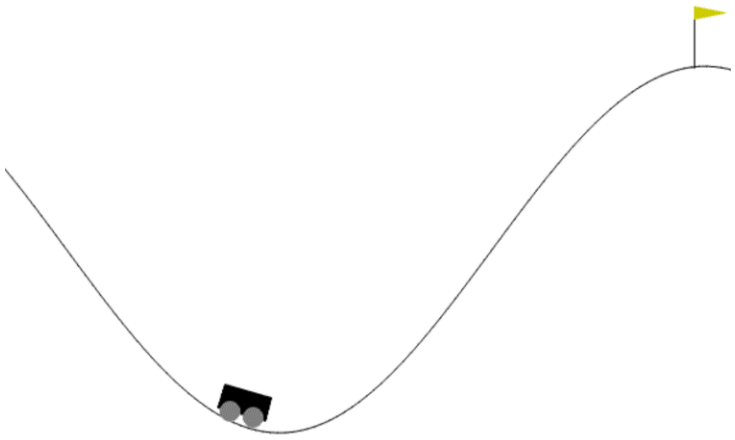
نام درس: یادگیری عمیق	نام دانشکده : دانشکده برق و کامپیوتر
تمرین اختیاری مبحث یادگیری تقویتی	ترم ۴۰۲۱
استاد : دکتر سمانه حسینی	نام طراح : یاسین حمزوی
	تاریخ تحویل: ۳۰ دی

۱- معرفی تمرین :

در این تکلیف، شما قرار است از الگوریتم یادگیری تقویتی DQN برای حل مسئله mountain-car در محیط gym استفاده کنید. مسئله mountain-car یک مسئله کلاسیک یادگیری تقویتی است که در آن یک ماشین در یک شیب سینوسی مانند قرار دارد و باید با استفاده از شتاب‌دهنده‌های چپ و راست خود به قله شیب برسد. این مسئله به دلیل شیب تند و ناهموار خود، یک مسئله دشوار برای حل با الگوریتم‌های یادگیری تقویتی محسوب می‌شود. هدف این تکلیف، آشنایی دانشجویان با الگوریتم DQN و کاربرد آن در حل مسائل یادگیری تقویتی است. همچنین، این تکلیف به دانشجویان کمک می‌کند تا مهارت‌های خود را در زمینه پیاده‌سازی الگوریتم‌های یادگیری تقویتی در محیط‌های واقعی بهبود بخشند.

۲- توضیحات محیط mountain car :

مسئله تصمیم‌گیری مارکف mountain car یک مسئله تصمیم‌گیری مارکف قطعی است که شامل یک ماشین قرار گرفته در پایین یک دره سینوسی به صورت تصادفی می‌باشد و تنها اعمال ممکن شتاب‌هایی هستند که می‌توانند به ماشین در هر دو جهت چپ و راست اعمال شوند. هدف این مسئله تصمیم‌گیری مارکف، شتاب دادن استراتژیک ماشین برای رسیدن به حالت هدف در بالای تپه سمت راست است. دو نسخه از این مسئله در محیط شبیه‌سازی gym وجود دارد: یکی با اعمال گسسته و دیگری با اعمال پیوسته. این نسخه، نسخه‌ای با اعمال گسسته است.



شکل ۱- یک نمونه تصویر از محیط Mountain Car

۲-۱-اعمال ممکن:

در این محیط، سه عمل وجود دارد که عبارتند از :

- شتاب به چپ
- بدون اعمال شتاب
- شتاب به راست

۲-۲- پاداش:

هدف این است که عامل هر چه سریع‌تر به پرچم هدف در سمت راست محیط برسد. بدین منظور خود محیط به ازای هر گام زمانی که سپری شود، یک پاداش منفی معادل ۱- اعمال می‌کند. این پاداش هر بار و توسط فراخوانی دستور env.step(action) در خروجی این دستور برگردانده می‌شود. علاوه بر این شما باید هر بار که عامل توانست به پرچم برسد، یک پاداش مثبت بزرگ برای عامل در نظر بگیرید.

۲-۳- محیط قابل مشاهده:

- موقعیت ماشین در محور افقی x که در بازه [-1.2,0.6] محدود شده است.
- سرعت عامل که در بازه [-0.07,0.07] محدود شده است

۲-٤- موقعیت و سرعت شروع :

محل شروع حرکت عامل از یک تابع یکنواخت به صورت تصادفی و در بازه [-0.6,-0.4] تعیین می‌شود و سرعت اولیه عامل همیشه صفر است.

۲-۵ پایان هر اپیزود:

در صورت رخ دادن هر یک از شرایط زیر، اپیزود به پایان می‌رسد:

- موقعیت مکانی عامل در محور افقیx ، بزرگتر یا مساوی 0.5 شود.(رسیدن به پرچم هدف)
- زمانی که ۲۰۰ گام زمانی سپری شده باشد.

توضیحات بیشتر از این محیط در لینک زیر قرار دارد:

https://www.gymlibrary.dev/environments/classic_control/mountain_car/

۳- توضیحات الگوریتم DQN :

الگوریتم DQN عمدتاً شبیه الگوریتم یادگیری Q می‌باشد. تنها تفاوت این است که به جای نگاشت دستی جفت‌های حالت-عمل به مقادیر Q متناظرشان، از شبکه‌های عصبی استفاده می‌کنیم. از آنجایی که محیط در اینجا قطعی است ، ما هم برای سادگی فرض می‌کنیم که معادلات بلمن به صورت قطعی بوده و از رابطه‌ی زیر برای به روز رسانی مقادیر Q استفاده می‌کنیم:

$$Q^{\pi}(s,a) = r + \gamma Q^{\pi}(s', \pi(s'))$$

٤-توضیحات مربوط به Replay Memory :

در این تمرین ما از حافظه Replay Memory برای آموزش DQN خود استفاده خواهیم کرد. این حافظه، گذارهایی را که عامل مشاهده می کند، ذخیره می کند و به ما امکان می دهد این داده ها را بعداً دوباره استفاده کنیم. با نمونه برداری تصادفی از آن، گذارهایی که یک batch را تشکیل می دهند، از هم گسسته می شوند. نشان داده شده است که این ترفند ، باعث ثبات و بهبود بسیاری در روش آموزش DQN می شود.

هر گذاری در Buffer به صورت یک لیست به صورت زیر ذخیره خواهد شد:

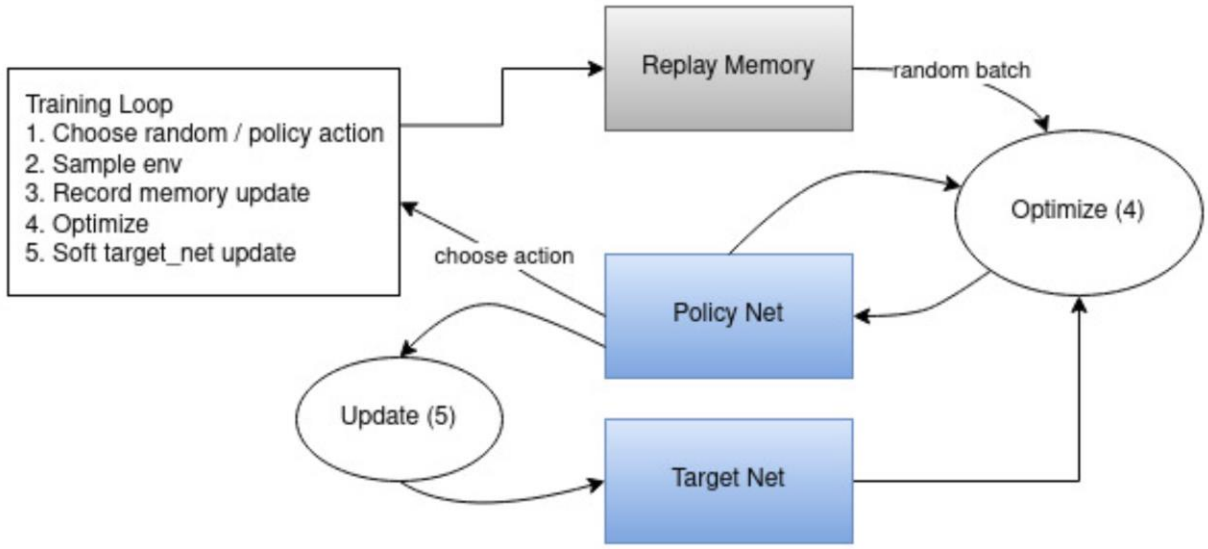
[currentState, bestAction, reward, new_state, done]

که در آن، متغیر done در حقیقت تعیین می‌کند که آیا اپیزود پایان یافته است یا خیر و به صورت متغیر boolean تعریف میشود.

٥-شماتیک کلی یادگیری:

اعمال یا به صورت تصادفی انتخاب می‌شوند یا بر اساس یک سیاست عمل و در گام بعدی یک نمونه از محیط شبیه‌سازی gym به دست می‌آید. ما نتایج را در حافظه بازپخش(Replay Buffer) ثبت می‌کنیم و

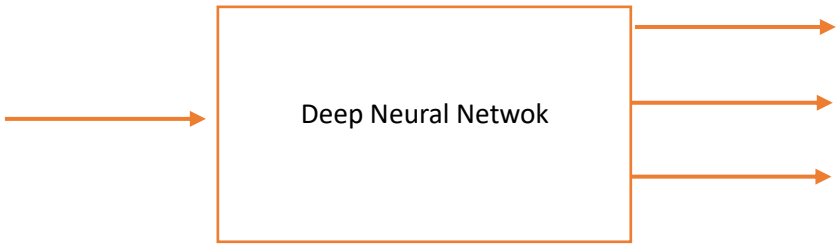
همچنین مرحله بهینه‌سازی را در هر تکرار اجرا می‌کنیم. الگوریتم بهینه‌سازی، یک دسته تصادفی از حافظه بازپخش را برای آموزش سیاست جدید انتخاب می‌کند. شبکه target_net نیز در بهینه‌سازی برای محاسبه مقادیر Q مورد انتظار استفاده می‌شود و به‌روزرسانی وزن‌های آن در هر گام انجام می‌شود(که در کد ارائه شده ، به صورت آماده نوشته شده است و لازم نیست شما آن‌را اعمال کنید.)



شکل۲ – شماتیک کلی یادگیری

٦-شماتیک شبکه عمیق عصبی:

این شبکه عمیق عصبی شامل چند لایه Dense می‌باشد. ورودی شبکه به صورت اندازه هر state و خروجی شبکه شامل سه خروجی برای مقادیر Q برای هر یک از اعمال می‌باشد.



شکل۳- شماتیک کلی شبکه عمیق عصبی

۷- تست مدل:

بعد از اجرای کامل فایل، در هر اپیزودی که عامل توانسته باشد به هدف برسد ، یک فایل خروجی از مدل گرفته شده است.

برای تست مدل خودتان ، شما باید آدرس فایلی که در آخرین اپیزود ذخیره شده است را در فایل تست قرار داده شده در تکلیف ، قرار دهید . فایل تست ، مدل شما را برای ۲۰ اپیزود مختلف اجرا گرفته و از اپیزودی که بهترین امتیاز را گرفته باشد ، یک فایل gif. تولید می‌کند. هدف این است که در هنگام تست ، تعداد اپیزودهایی که به هدف رسیده باشند زیاد بوده و همراه با امتیاز قابل قبول باشند.

در نهایت تعداد دفعات موفقیت و بهترین امتیازی که مدل شما به آن رسیده به همراه فایل خروجی مدل‌ها و فایل خروجی gif باید گزارش شود. همین طور توجه شود که نمره دهی براساس نتایج گزارش شده توسط مدل‌های به دست آمده شما می‌باشد و به صورت رقابتی خواهد بود لذا مسئله مهم، نحوه تعیین تابع پاداش و تعیین پارامترها و سایر عوامل خواهد بود.

نکات بسیار کلیدی:

- یک فایل پایتون در تکلیف قرار داده شده است و از شما خواسته شده است که کدهای خود را در قسمت های خالی و با توجه به توضیحات آن قسمت که به صورت کامنت نوشته شده ، بنویسید.
- از مرحله تست خود فیلم کوتاه تهیه کنید و در آن موارد لازم را توضیح دهید.
- حتما باید فایل های زیر در هنگام تحویل گزارش، ارائه شوند:
  - مدل‌های خروجی
  - تعداد دفعات موفقیت آمیز در هنگام تست
  - فایل خروجی gif
  - فایل کدهای نوشته شده در فایل Hw_DQN_training
  - نمره دهی بر اساس امتیاز های به دست آمده از مدل هاست. پس باید سعی کنید پاداش های مناسب تری تعریف کنید تا بهترین نتایج را بگیرید.

نکات تحویل تکلیف:

- همانطور که قبلا هم اطلاع داده شد شما مجاز هستید در طول ترم تا ۸ روز تاخیر در تحویل کل تکالیف داشته باشید.
- دانشجویان می توانند در حل تکالیف با دوستان خود مشورت نمایند اما در نهایت هرکس موظف است تکلیف را به صورت فردی، انجام و تحویل دهد. لذا در صورت مشاهده تکالیف کپی بین دانشجویان نمره تمامی افراد شرکت کننده در آن صفر خواهد بود.
- توضیحات شما در فرآیند تصحیح از اهمیت ویژه ای برخوردار است لطفا تمامی نکات و فرضهایی که برای پیاده سازیها و محاسبات خود در نظر میگیرید را در گزارش ذکر کنید.
- در صورت داشتن هرگونه سوال میتوانید از طریق ایمیل یا اکانت تلگرام زیر با دستیار آموزشی مربوطه در ارتباط باشید.

Telegram : @yasinhmv / Email:yasinhamzavi1998@gmail.com