یادگیری ماشین

پاسخ مینی پروژه‌ی شماره ۳

نام و نام خانوادگی	مهدی خدابنده لو
شمارهٔ دانشجویی	۴۰۱۰۵۳۱۴
تاریخ	بهار ۱۴۰۳

فهرست مطالب

فهرست تصاویر

فهرست جداول

فهرست برنامه‌ها

colab
github

۱ پرسش یک

۱.۱ برای این مسئله یک بار با روش **qlearning** و یک بار با روش **deep q learning** عاملی را آموزش دهید.

۲.۱ پاداش تجمعی را برای هر دو عامل در طول زمان ترسیم کنید.چگونه عملکرد عامل در طول زمان بهتر می شود؟ میانگین پاداش در هر اپیزود را برای دو عامل در ۱۰۰۰ اپیزود مقایسه کنید.

شکل ۱ نمودار امتیاز کسب شده توسط عامل را در هر اپیزود در Qlearning نمایش می دهد. شکل ۲ نیز نمودار امیاز کسب شده را به صورت تجمعی برای عامل ذکر شده نمایش می دهد. پاداش میانگین در این حالت 9.755- است.

شکل ۳ نمودار امتیاز کسب شده توسط عامل را در هر اپیزود در Qlearning نمایش می دهد. شکل ۴ نیز نمودار امیاز کسب شده را به صورت تجمعی برای عامل ذکر شده نمایش می دهد. پاداش میانگین در این حالت 545.768- است.

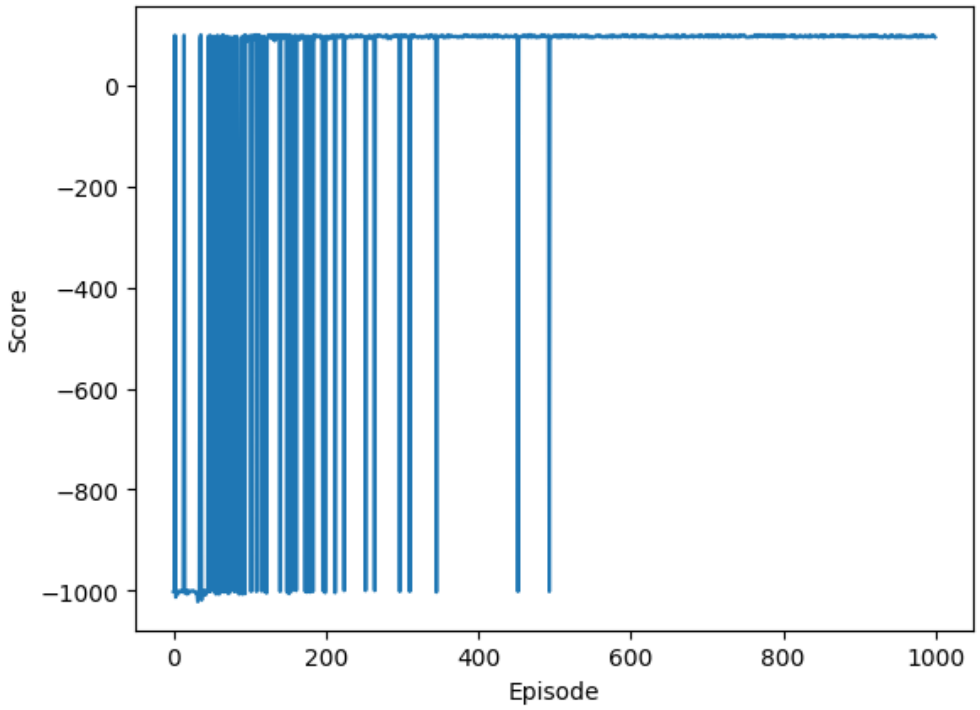
همان طور که مشاهده می شود Qlearning بسیار سریع تر از DQN عمل کرده است. در مسائل کوچک با تعداد حالات محدود استفاده از آسانQLearning تر و سریع تر است و نیاز به توان محاسباتی بالایی وجود ندارد، ولی در صورتی که مسئله پیچیده شود و تعداد حالات ممکن زیاد باشد استفاده از Qlearning مناسب نیست چون نیاز به یک Qmatrix با اندازه ی خیلی بزرگ خواهیم داشت از این رو توصیه می شود که در مسائل پیچیده از DQN استفاده شود چون در DQN به جای استفاده از یک ماتریس از قبل آماده یک شبکه عصبی داریم که با در یافت موقعیتی که عامل دارد یک Qvalue به ازای هر اکشن تولید می کند. در کل Qlearning برای این مسئله مناسب تر است. نکته ای که در آموزش عامل ها وجود دارد این است که به منظور آموزش بهتر نقطه ی شروع تصادفی انتخاب می شود.

۳.۱ چگونه اپسیلون بر فرآیند یادگیری تاثیر می گذارد؟

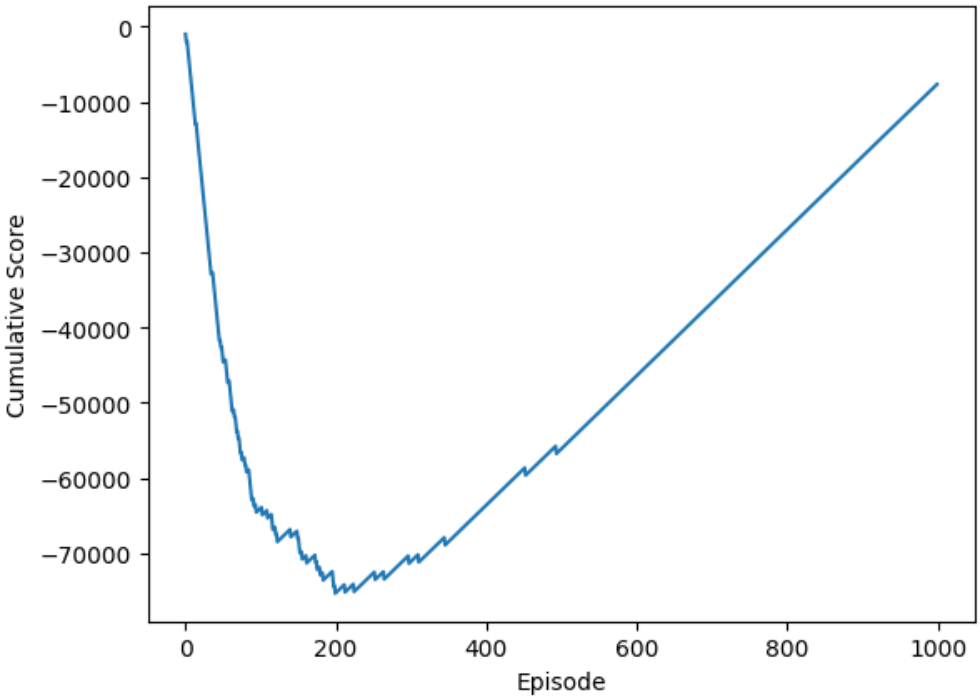
پارامتر اپسیلون میزان تصادفی بودن حرکت بعددی را مشخص می کند. ما در این جا این پارامتر را در ابتدا کم در نظر گرفتیم و به مرور مقدار آن را افزایش دادیم. در صورتی که این پارامتر کم باشد ممکن است عامل بهترین راه حل را پیدا نکند. زمانی که اپسیلون بالا باشد عامل دیر تر به نقطه ای می رسد که همیشه برنده است ولی بهتر آموزش می بیند این درحالی است که در صورتی که اپسیلون کم باشد ممکن است احتمال خطای عامل همیشه وجود داشته باشد و عامل بهترین راه را پیدا نکند.

۴.۱ چند اپیزود طول کشید تا عامل Qlearning به طور مداوم طلا را بدون افتادن در گودال یا خورده شدن پیدا کند؟ Qlearning و DQN را مقایسه کنید

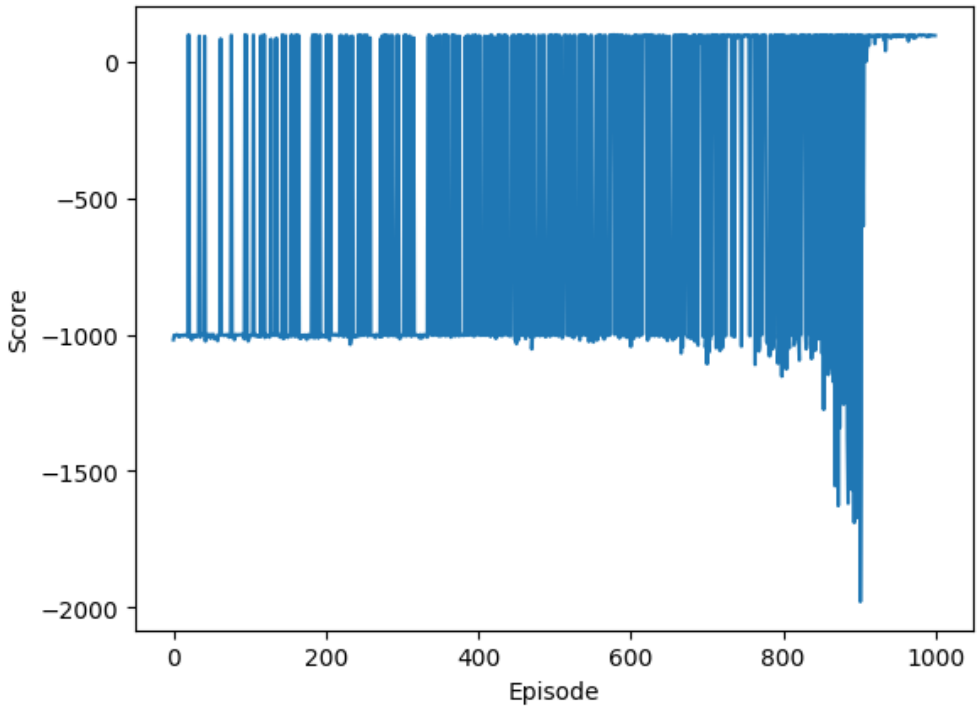
۳۴۶ اپیزود طول کشید که عامل با QLearning بدون باخت به طلا برسد این درحالی است که در برای عامل DQN ۹۰۶ اپیزودی طول کشید که عامل بدون باخت به طلا برسد. یکی از دلایل این که DQN دیر تر آموزش دید این است که شبکه برای آموزش بهتر شبمه عصبی، نرخ کاهش اپسیلون برای عامل DQN کم تر در نظر گرفته شد. همان طور که در قسمت های قبل هم ذکر شد Qlearning نسبت به DQN سریع تر است ولی مسئله ای که وجود دارد این است که Qlearning برای مسائل کوچک مناسب است و در صورتی که تعداد



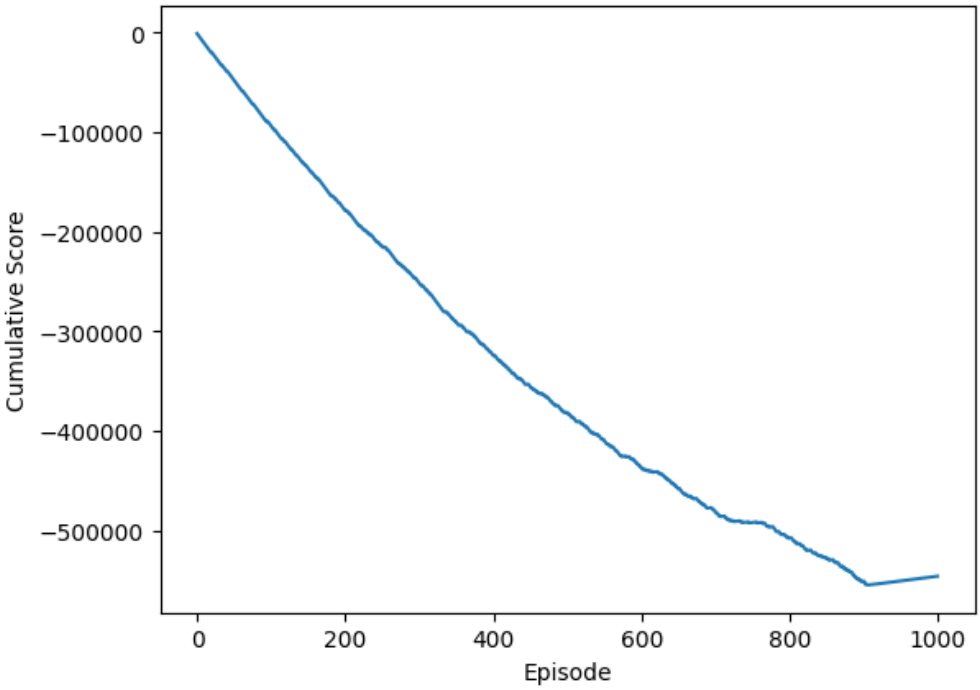
شکل ۱: نمودار امتیاز کسب شده در هر اپیزود توسط عامل در Qlearning



شکل ۲: نمودار امتیاز کسب شده تجمعی در هر اپیزود توسط عامل در QLearning



شکل ۳: نمودار امتیاز کسب شده در هر اپیزود توسط عامل در DQN



شکل ۴: نمودار امتیاز کسب شده تجمعی در هر اپیزود توسط عامل در DQN

حالات مسئله زیاد شود نیاز به یک جدول بزرگ خواهیک داشت پیاده سازی این الگوریتم را غیر ممکن می کند از این رو از DQN استفاده می کنیم که ورودی آن state است و در خروجی هر اکشن و Qvalue مربوط به آن را نمایش می دهد.

۵.۱ معماری مورد استفاده برای شبکه ی DQN را شرح دهید.

شبکه ی مورد استفاده دارای دو لایه ی تمام متصل است که هر لایه ۲۵۶ نورون است. ابعاد ورودی شبکه ۲ است(به تعداد بعد های (state و ابعاد خروجی ۴ است(به تعداد اکشن ها). همچنین بین دو لایه یک لایه فعال ساز ReLU وجود دارد. علت این که در انتهای لایه ی آخر فعال ساز وجود ندارد این است که ما انتظار داریم که این شبکه مقادیر Q را تخمین بزند و این مقدار می تواند مثبت یا منفی باشد در حالی که یک فعال ساز مثل ReLU خروجی بین ۰ تا ۱ دارد. علت این که از دو لایه استفاده کردیم این است که محیط بازی زیاد پیچیده نیست که از لایه های بیشتری استفاده کنیم.

۲ پرسش دو

۱.۲ در مورد محیط **Lunar Lander**مطالعه کرده و به صورت خلاصه ویژگی های آن را شرح دهید. ویژگی های مدنظر عبارتند از مشخصات فضای حالت، مشخصات فضای عمل و سیستم پاداش

مشخصات فضای حالت:

فضای حالت شامل هشت متغیر به شرح زیر است:

-مختصات جسم در x

-مختصات جسم در y

-سرعت جسم در راستای x

-سرعت جسم در راستای y

-زاویه

-سرعت زاویه ای

-تماس پایه ی چپ با زمین

-تماس پایه ی راست با زمین

مشخصات فضای عمل

فضای action شامل چهار عمل به شرح زیر است:

-کاری نکند

-روشن شدن موتور راست

-روشن شدن موتور چپ

مشخصات سیستم پاداش

-عامل با فرود موفق امتیاز مثبت دریافت می کند. هر چه عامل نزدیک به مرکز فرود بیاید امتیاز بیشتری دریافت می کند

-دریافت امتیاز منفی با برخورد با زمین(سقوط)

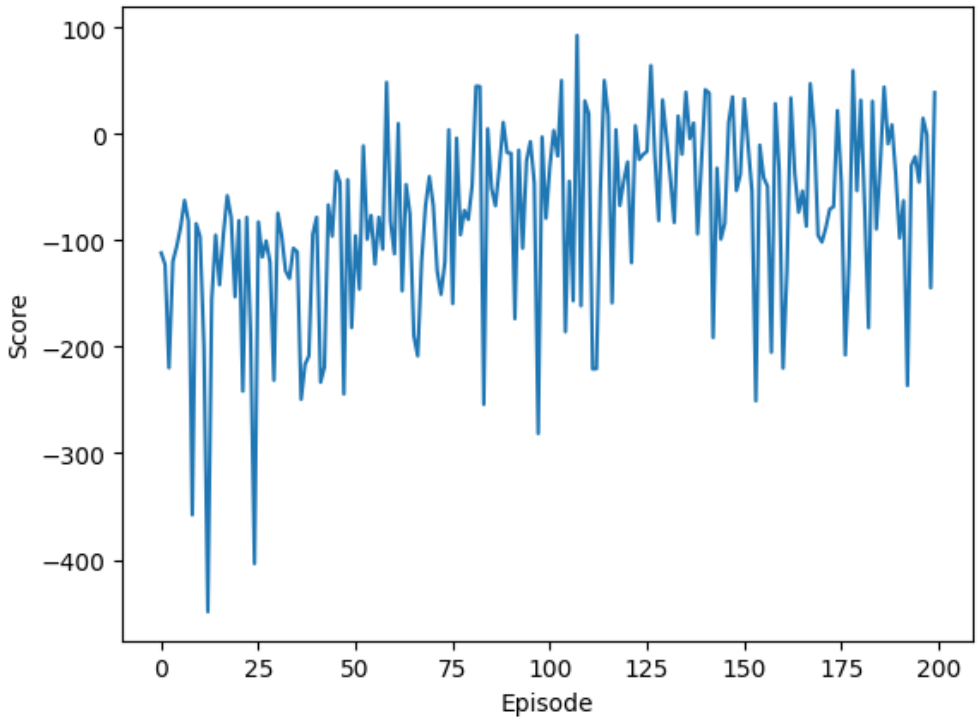
-دریافت امتیاز مثبت به ازای برخورد پایه ها به زمین

-دریافت امتیاز منفی به ازای مصرف سوخت

۲.۲ عملکرد عامل را با رسم پاداش تجمعی در هر episode و برای **batch size** های ۶۴ ،۳۲و ۱۲۸بررسی کنید. تنها برای بهترین حالت به ازای هایepisode ۲۰۰ ،۱۵۰ ،۱۰۰ ،۵۰و ۲۵۰ فیلمی ز عملکرد عامل تهیه کنید. در صورتی که عملکرد عامل به ازای هر سه مقدار **batch size**مشابه یکدیگر شد، یکی از آن ها را به دلخواه به عنوان بهترین حالت انتخاب کنید. در رابطه با انتخاب بهترین حالت علاوه بر معیار سرعت همگرایی به پاداش بهینهٔ معیار راregret نیز به صورت شهودی بررسی کنید

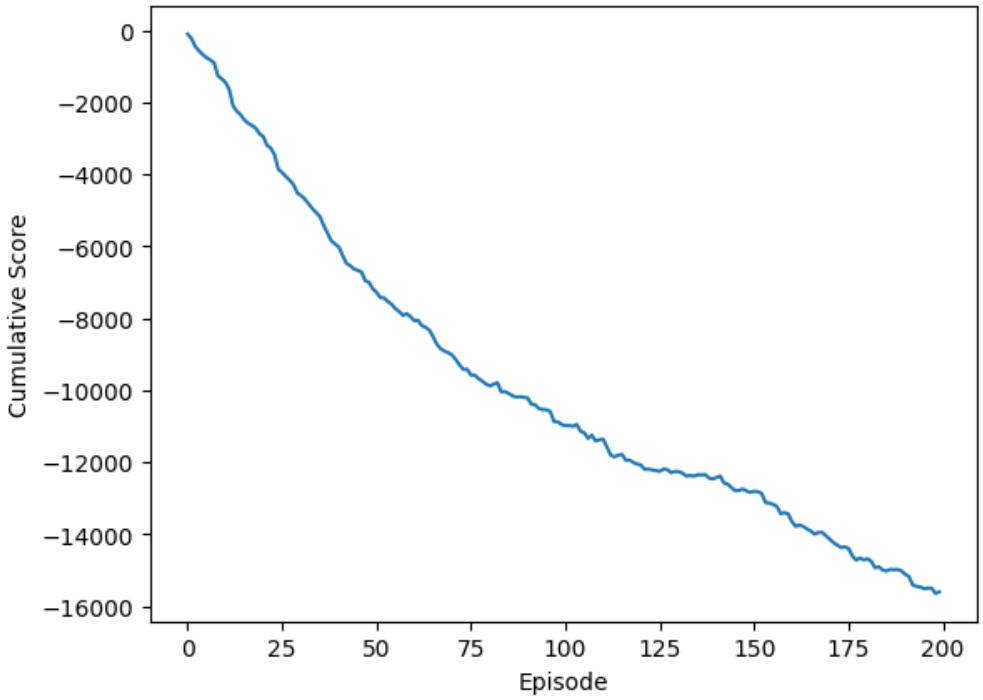
برای این سوال به علت طولانی شدن فرایند آموزش تنها موفق با آموزش عامل با batch ۳۲ و ۶۴ شدیم. همچنین تعداد اپیزود های آموزش برابر با ۲۰۰ در نظر گرفته شده است.

شکل ۵ و ۶ مربوط یه آموزش عامل با batch_size=۳۲ است. همان طور که مشخص است شیب امتیازات دریافت شده تقریبا مثبت است.



شکل ۵: نمودار امتیاز کسب شده تجمعی در هر اپیزود توسط عامل(batch_size = ۳۲)

episode۰
episode۱۰
episode۲۰
episode۳۰
episode۴۰
episode۵۰
episode۶۰



شکل ۶: نمودار امتیاز کسب شده تجمعی در هر اپیزود توسط عامل(batch_size = ۳۲)

episode۷۰
episode۸۰
episode۹۰
episode۱۰۰

۳.۲ عملکرد مدل وDQN راDDQN با رسم پاداش تجمعی در هر episode و به ازای **batch size** برابر مقایسه کنید. برای هر دو مدل به ازای episode های ۱۰۰و ،۲۵۰قیلمی از عملکرد مدل تهیه کنید

استفاده از DQN ممکن است منجر به overestimation شود از این رو سراغ DDQN می رویم. در DDQN اکشن با بیشترین مقدار Q توسط یک شبکه تعیین می شود و ارزیابی آن توسط یک شبکه ی دیگر انجام می شود.