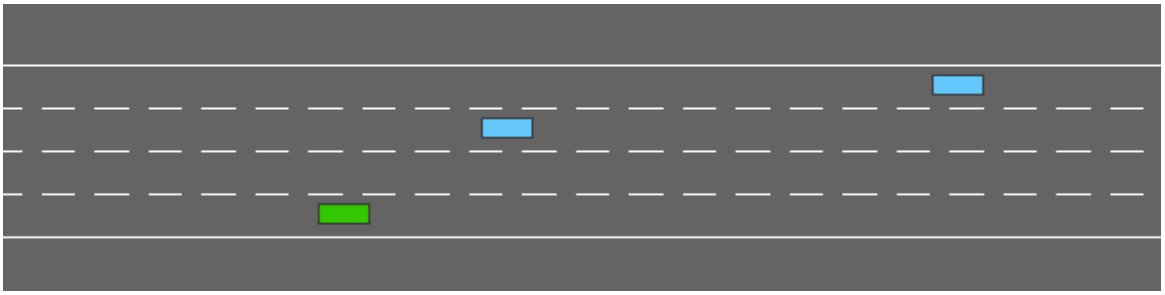
با توجه به پیشرفت روز افزون مدل‌های یادگیری عمیق و همچنین نتایج قابل توجه ادغام این مدل‌ها با کاربردهای یادیگیری تقویتی، تمرین پنجم به بررسی این الگوریتم‌ها و مدل‌ها خواهد پرداخت. حل کردن مسائل یادگیری تقویتی با استفاده از شبکه‌های عصبی پیشینه‌ای قدیمی دارد و با توجه به پیشرفت سخت‌افزارهای محاسباتی در دو دهه اخیر، سرعت توسعه مدل‌های عمیق برای مسائل یادگیری تقویتی افزایش قابل ملاحظه‌ای داشته است. استفاده از شبکه‌های عصبی این امکان را به ما می‌دهد که از مسئله را با استفاده از یک مدل end to end حل کنیم. با توجه به این نکته کسب مهارت کارکردن با مدل‌های یادگیری عمیق و حل مسائل یادگیری تقویتی با استفاده از این مدل‌ها از مهارت‌های ضروری در زمینه یادگیری تقویتی می‌باشد. این تمرین مقدمه آشنایی شما با این مسائل را فراهم می‌کند و طبیعتا کسب مهارت‌های بیشتر در این زمینه نیاز مطالعه و تمرین بیشتر خواهد بود.

سوالات تحلیلی

- به طور مختصر توضیح دهید هدف الگوریتم پالیسی گردینت چیست.
- یک مورد از فواید و معایب الگوریتم‌های Deep RL را توضیح دهید.
- یک مورد از دلایل استفاده از بافر تجارب را نام ببرید.

محیط مورد استفاده در سوال پیاده‌سازی

در این تمرین شما با محیط معروف highway کار خواهید کرد. این محیط در کتابخانه gym پیاده‌سازی شده است و تسک‌های مختلفی از جمله تسک پارکینگ، عبور از چهار راه و ... روی این محیط تعریف شده است. مسائل موجود در این محیط مربوط به تسک autonomous driving در یک محیط دو بعدی می‌باشند.



تصویر ۱ – محیط highway-v0

برای استفاده است این محیط با توجه به حجم زیاد محاسبات توصیه ما این است که با استفاده از محیط گوگل کولب، از قطعه کد زیر جهت نصب و import کردن کتابخانه‌ها مورد نیاز استفاده کنید.

```
! pip install gym
! pip install highway-env
import gym
import highway_env
env = gym.make("highway-v0")
```

جهت آشنایی بیشتر با نحوه کارکرد این محیط می‌توانید از ریپازیتوری گیتهاب خود کتابخانه استفاده کنید.

بخش اول – آشنایی با محیط مسئله

با توجه به توضیحات و لینک فراهم شده محیط را در مود highway-v0 اجرا کنید.

- اکشن‌ها، استیت‌ها و پاداشی که عامل از محیط دریافت می‌کند را شرح دهید. برای مثال توضیح دهید که استیت عامل نشان دهنده چه خصیصه‌هایی از محیط است.
- در مورد پیوستگی و گسسته بودن استیت‌ها و اکشن‌های این محیط تحقیق کنید.

بخش دوم – الگوریتم حل

در این بخش هدف پیاده‌سازی یکی از دو الگوریتم policy gradient و Deep Q-learning [1] می‌باشد.

- یکی از دو الگوریتم را بدون استفاده از کتابخانه stable_baselines3 پیاده‌سازی کنید. برای پیاده‌سازی ترجیحا از کتابخانه Pytorch استفاده کنید. استفاده از کتابخانه tensorflow نیز بلامانع است. عامل شما باید تسک merge را از محیط highway-env یاد بگیرد.
- پس از یادگیری عامل نمودار پاداش کسب شده در طول یادگیری توسط عامل در چند ران مختلف را در گزارش خود قرار دهید. نمودار مورد نظر باید شامل بازه اطمینان ۹۵ درصد باشد.

- ارائه ویدیو (render) از چند اپیزود تست عامل پس از یادگیری نیز دارای نمره مثبت می‌باشد. برای رندر گرفتن توصیه می‌کنیم جهت درگیر نشدن با مشکلات گوگل کولب وزن‌های مدل خود را ذخیره کنید و این کار را به صورت local روی کامپیوتر خود انجام دهید. جهت رندر گرفتن هم می‌توانید از خود توابع کتابخانه highway-env استفاده کنید.
- پارامترهای مورد استفاده خود را در گزارش در یک جدول بیان کنید.

بخش سوم – انتقال تجربه با استفاده از transfer learning

در این بخش هدف این است که تاثیر انتقال تجربه یاد گرفته شده در یک محیط به محیط دیگر را بررسی کنید. با استفاده از الگوریتم DQN و یا policy gradient که در بخش قبلی آن را پیاده‌سازی نمودید، عامل را با وزن‌دهی رندوم در تسک highway-v0 از کتابخانه گفته شده train کنید لازم به ذکر است که معماری شبکه استفاده شده شما باید با قبل یکسان باشد.

- حال نتایج این پیاده‌سازی را با حالتی که به جای وزن‌دهی رندوم از وزن‌های نهایی تسک merge برای شبکه استفاده کردید، مقایسه کنید. بررسی کنید که آیا سرعت یادگیری افزایش پیدا می‌کند و در صورت مثبت بودن جواب دلیل این افزایش سرعت در یادگیری چیست؟

بخش چهارم – امتیازی

به عنوان نمره امتیازی می‌توانید از دو مورد زیر یک مورد را انتخاب کرده و پیاده سازی کنید.

- الگوریتم DQN با حالت Prioritized Experience Replay را پیاده‌سازی کنید. توضیحات مربوط به این الگوریتم را می‌توانید در مقاله [2] پیدا کنید. همچنین بیان اضافه کردن این ویژگی چه سودی برای الگوریتم DQN بدون این ویژگی دارد.

- الگوریتم DQN و یا policy gradient را با استفاده از image observation و استفاده از شبکه‌های CNN پیاده سازی کنید. در این حالت ورودی مدل شما فریم‌های تصویری محیط خواهد بود. همچنین تفاوت observation و state را بیان کرده و مسئله‌ای که استفاده از observation می‌تواند داشته باشد و راه‌حل‌های ممکن را ذکر کنید.

در هر دو حالت نتایج را باید با نتایج بخش دوم مقایسه کنید. برای مقایسه می‌توانید از نمودارهای مناسب استفاده کنید. آیا اختلاف به صورت significant می‌باشد؟ برای ادعای خود می‌توانید از تست‌های آماری استفاده کنید.

مراجع

[1] V. Mnih *et al.*, “Playing Atari with deep reinforcement learning,” *arXiv [cs.LG]*, 2013

[2] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” *arXiv [cs.LG]*, 2015.

نکات پیاده‌سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز شنبه **۸ بهمن ۱۴۰۱** ماه خواهد بود.
- در رسم نمودارها حتماً باید title، axis label و grid داشته باشد و مقادیر به‌صورت گویا نمایش داده شود.
- پیاده‌سازی تنها با پایتون قابل‌قبول است.
- حجم گزارش شما هیچ‌گونه تأثیری در نمره نخواهد داشت و تحلیل و نمودارهای شما بیشترین ارزش را دارد.
- گزارش خود را در قالب آپلود شده در سامانه نوشته و ارسال کنید.
- انجام این تمرین به‌صورت یک‌نفره می‌باشد.
- لطفاً گزارش، فایل کدها و سایر ضمائم موردنیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمایید.

HW5_[Lastname]_[StudentNumber].zip

- در صورت وجود سؤال و یا ابهام می‌توانید تنها از طریق رایانامه زیر با دستیاران آموزشی در ارتباط باشید:

amir.mesbah@ut.ac.ir
banafshehkarimian@ut.ac.ir