تمرین اول

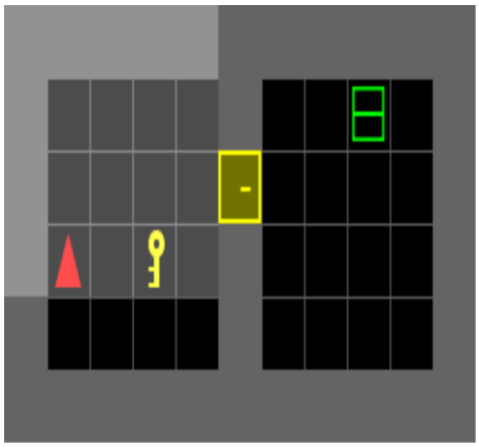
یادگیری تقویتی عمیق

سیاست در یادگیری

مدرس: دکتر آرمین سلیمی بدر

فروردین ۱۴۰۳

تشریح مسئله



در این تمرین می‌خواهیم دو نوع رویکرد یادگیری سیاست، یعنی یادگیری On-policy و یادگیری Off-policy از خانواده الگوریتم‌های یادگیری تقویتی را با یکدیگر بررسی نماییم. برای این منظور، از الگوریتم‌های مبتنی بر توابع ارزش SARSA و DQN به‌عنوان نماینده‌های این دو نوع رویکرد یادگیری استفاده می‌کنیم. به‌جهت آنکه عملکرد این دو الگوریتم را بررسی کنیم، دو عامل را به‌صورت مجزا در یک محیط مشبک (شکل روبه‌رو) آموزش می‌دهیم تا توانایی رسیدن به هدف را پیدا کنند.

در این محیط، عامل توانایی برداشتن اشیاء را دارد. عامل (مثلث ▼) باید جعبه‌ای (مربع نصف شده ◫) را که در اتاق دیگر در پشت درب قفل شده (▣) قرار می‌گیرد، بردارد. ازاین‌رو، عامل ابتدا باید یاد گیرد که کلید (⚿) را یافته و آن را بردارد، سپس درب را باز کند، و در انتها جعبه در اتاق دیگر را بردارد؛ بنابراین، عامل باید با انجام سلسه‌ای از کارها، بهینه‌سازی خود را برای به اتمام رساندن وظیفه محوله انجام دهد. توجه کنید که رنگ کلید، درب و جعبه با توجه به Seed که در کد تعریف می‌شود متغیر خواهد بود.

آماده‌سازی و پیش‌نیازها

علاوه‌بر کتابخانه‌ی Gymnasium، باید کتابخانه Minigrid را نصب کنید که از اینجا قابل دسترس است. محیطی که باید در آن تمرین را انجام دهید با نام Unlock Pickup از اینجا قابل در دسترس است. پیش از شروع به حل مسئله، مستندات این کتابخانه را به‌طور کامل مطالعه کنید تا با امکاناتی این کتابخانه برای حل محیط در اختیار شما قرار می‌دهد؛ آشناییت کامل پیدا کنید.

سه ویدیو آموزشی به‌جهت آشنایی و کار با کتابخانه Gymnasium، پیاده‌سازی الگوریتم DQN و حل محیط‌های این کتابخانه تهیه و در کانال تلگرامی درس منتشر شده است که از اینجا نیز قابل دسترسی هستند.

خواسته‌ها

بخش اول

۱- دو عامل را به‌صورت مجزا با استفاده از الگوریتم‌های SARSA و DQN آموزش دهید تا مسئله را حل کنند. علل موفقیت یا عدم‌موفقیت و سرعت همگرایی در حل مسئله را برای هر دو الگوریتم با ارائه **توضیحات کافی** بررسی کنید.

۲- نمودارهای پاداش‌های دریافتی توسط عامل ($\boldsymbol{Reward}$)، میزان خطای شبکه ($\boldsymbol{Loss}$)، مقدار اپسیلون در سیاست اپسیلون-حریصانه ($\boldsymbol{\epsilon - greedy}$) را رسم کنید و با **توضیحات کافی** ارتباط آن‌ها را با یکدیگر تحلیل نمایید. در انتها، عملکرد هر دو الگوریتم را با یکدیگر قیاس کنید.

۳- پس از اینکه مدل DQN را با فراپارامترهای بهینه آموزش دادید، آموزش خود را با نرخ‌های معقول و مختلف از فراپارامترهای ذیل برای حداقل ۳ بار دیگر تکرار کنید (فقط برای DQN) و گزارش دهید که تغییر هر یک از هایپرپارامترها چه تأثیری در روند آموزش گذاشته است. در انتها نتایج آزمایش‌های خود را در یک جدول ارائه کرده و نتایج حاصله را قیاس کنید. فراپارامترها عبارت‌اند از:

- Learning Rate
- Discount Factor
- Target Network Update Frequency

۴- وزن‌های نهایی (بهترین وزن‌ها) مدل‌های آموزش داده SARSA و DQN خود را ذخیره کنید و حتماً برای آن‌ها تابع تست تعریف کنید تا امکان آزمون مجدد مدل‌ها مهیا باشد.

بخش دوم

۱- به‌جای استفاده از سیاست اپسیلون-حریصانه در DQN، سیاست رفتاری بولتزمان را اعمال کنید. توضیح دهید که چه تغییراتی در فرآیند کاوش ایجاد شده است. سرعت همگرایی این دو سیاست رفتاری را بر اساس نمودارها مقایسه کنید.

۲- پارامتر دما در رابطه‌ی سیاست بولتزمان را همانند روش اپسیلون-حریصانه به‌صورت تدریجی کاهشی پیاده‌سازی کنید. سپس بررسی کنید که آیا تغییر در پارامتر دما تأثیری در سرعت همگرایی مدل ایجاد می‌کند؟ این آزمایش را حداقل ۳ بار با پارامترهای مختلف تکرار کنید و نتایج را گزارش دهید.

بخش سوم

ازآنجایی‌که یادگیری الگوریتم SARSA مبتنی بر سیاست است، استفاده چندباره از داده‌های قبلی (تجربیات) برای آموزش سیاست صحیح نیست. این محدودیت را در عمل بررسی کنید. بدین منظور، یک حافظه تجربه برای این الگوریتم تعریف کنید و تجربیات عامل را در آن ذخیره کنید. سپس از این داده‌ها برای آموزش استفاده کنید. نتایج حاصله از این آزمایش را بررسی و در گزارش بیان کنید (**موفقیت** در این آزمایش فاقد اهمیت است).

نکات

- استفاده از کتابخانه‌ی Stable Baselines برای این تمرین مجاز نیست.
- استفاده از کد پیاده‌سازی شده در ویدیو بلامانع است.

موارد تحویل

۱- یک ویدیو کم حجم (Gif, Mp4, ...) از اجرای موفقیت‌آمیز مدل‌های DQN و SARSA از بخش اول در زمان **تست** تهیه و در فایل ارسالی ضمیمه کنید.

۲- اسکریپت پایتون هر دو الگوریتم را به‌صورت مجزا همراه فایل‌های دیگر ارسال نمایید.

۳- گزارش کار خود را به‌طور آراسته و منظم، با موارد خواسته‌شده و توضیحات کامل ارسال کنید. حتماً نام، نام خانوادگی و شماره دانشجویی خود را در گزارش و فایل زیپ ارسالی ذکر کنید.

مهلت تحویل: تا پایان روز یکشنبه ۱۴۰۳/۰۲/۰۹

موفق باشید.