به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

شبکه های عصبی مصنوعی و یادگیری عمیق

تمرین امتیازی

اردیبهشت ۱۴۰۱

فهرست سوالات

سوال ۱ – Lunar Lander ۳

پیوست ۴

سوال ۱ – Lunar Lander

در این سوال از شما میخواهیم با استفاده از روش Deep Q-Learning برای محیط LunarLander[1] یک عامل طراحی کرده و آموزش دهید که بتواند این مساله را حل نماید. برای حل تمرین notebook قرار داده شده در تمرین را با دقت مطالعه کرده و بخش های خواسته شده را تکمیل نمایید.

توجه: تا حد امکان ساختار notebook را حفظ کرده و از کپی کردن کد های آماده از اینترنت خودداری نمایید.

الف) محیط LunarLander را مطالعه کرده و به صورت خلاصه ویژگی های آن را شرح دهید. ویژگی های مد نظر عبارتند از مشخصات فضای حالت/ مشخصات فضای عمل/ سیستم پاداش

ب) عملکرد عامل را با رسم پاداش تجمعی در هر episode برای batch size های ۳۲, ۶۴ و ۱۲۸ بررسی کنید. تنها برای بهترین حالت به ازای episode های ۵۰, ۱۰۰, ۱۵۰, ۲۰۰ و ۲۵۰ فیلمی از عملکرد عامل تهیه کنید.

نکته: 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 به انتخاب بهترین حالت علاوه بر معیار سرعت همگرایی به پاداش بهینه معیار regret را نیز به صورت **شهودی** بررسی کنید.

ج) عملکرد مدل DQN و DDQN را با رسم پاداش تجمعی در هر episode و به ازای batch size برابر مقایسه کنید. برای هر دو مدل به ازای episode های ۱۰۰ و ۲۵۰ فیلمی از عملکرد مدل تهیه کنید.

نکته: هر بار آموزش عامل با استفاده از gpu های رایگان google colab بین ۱۰-۱۵ دقیقه زمان لازم خواهد داشت.

نکته: برای تهیه خروجی نمی توانید از سرویس google colab استفاده نمایید و می بایست checkpoint های مدل را دانلود کرده و روی سیستم خود فیلم ها را تهیه کنید.

[1] https://gym.openai.com/envs/LunarLander-v2/

پیوست

Algorithm 1 Deep Q-learning with Experience Replay

Initialize replay memory $\mathcal{D}$ to capacity N
Initialize action-value function Q with random weights
for episode $= 1, M$ **do**
  Initialise sequence $s_1 = \{x_1\}$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ **do**
    With probability ϵ select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$
    Execute action a_t in emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{for non-terminal } \phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$
  end for
end for

شکل ۱: شبه کد الگوریتم DQN

Basic Q-Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(R_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

Double Q-Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(R_{t+1} + \gamma Q'(s_{t+1}, a) - Q(s_t, a_t))$$

estimated/expected Q-value

$$a = \max_a Q(s_{t+1}, a)$$

$$q_{estimated} = Q'(s_{t+1}, a)$$

شکل ۲: تفاوت مدل DDQN با DQN

نکات:

- مهلت تحویل این تمرین ۶ خرداد است.
- گزارش را در قالب تهیه شده که روی صفحه درس در Elearn بارگذاری شده، بنویسید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفاً تمامی نکات و فرض‌هایی که برای پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- در گزارش خود برای تصاویر زیرنویس و برای جداول هم بالانویس اضافه کنید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست. اما باید نتایج بدست آمده را گزارش و تحلیل کنید.
- دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند. بنابراین هرگونه نتیجه و یا تحلیلی که در شرح سوال از شما خواسته شده است را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- در صورت مشاهدهٔ تقلب امتیاز تمامی افراد شرکت‌کننده در آن، ۱۰۰- لحاظ می‌شود.
- برای انجام تمرین ها و مینی پروژه ها، تنها زبان برنامه نویسی مجاز Python است.
- استفاده از کدهای آماده برای تمرین‌ها به‌هیچ‌وجه مجاز نیست. اما برای مینی‌پروژه‌ها فقط برای قسمت‌هایی از کد و به عنوان راهنمایی برای پیاده‌سازی، می‌توانید از کدهای آماده استفاده کنید.
- نحوهٔ محاسبه تاخیر به این شکل است: مهلت ارسال بدون جریمه تا تاریخ اعلام شده و پس از آن به ازای هر روز ۵ درصد نمره کسر خواهد شد و حداکثر تا یک هفته امکان ارسال با تاخیر وجود، پس از بازه تاخیر نمره تکلیف صفر خواهد شد.
- لطفاً گزارش، فایل کدها و سایر ضمایم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمایید.

Extra_[Lastname]_[StudentNumber].zip

- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌های زیر با دستیاران آموزشی مربوطه آقایان مرصاد اصالتی و محمدحسین حاجی کاظم نیلی در تماس باشید:

mersad.esalati@gmail.com

mohammad.nili@ut.ac.ir