



## مینی پروژه شماره چهار

در انجام این مینی پروژه حتماً به نکات زیر توجه کنید:

- موعد تحویل این مینی پروژه، ساعت ۱۸:۰۰ روز جمعه ۱۵ تیرماه ۱۴۰۳ است.
- برای این مینی پروژه ملزم به ارائه گزارش متنی شامل توضیحات کامل هر قسمت هستید. هم گزارش و هم کدهای خود را در گیت‌هاب و سامانه دانشگاه بارگذاری کنید.
- برای گزارش لازم است که پاسخ هر سوال و زیربخش هایش به ترتیب و به صورت مشخص نوشته شده باشند. بخش زیادی از نمره به توضیحات دقیق و تحلیل‌های کافی شما روی نتایج بستگی خواهد داشت.
- لازم است که در صفحه اول گزارش خود لینک پوشه گیت‌هاب و گوگل کولب مربوط به مینی پروژه خود را در حالتی که دسترسی Public دارد به اشتراک گذاشته باشید. دفترچه‌کد گوگل کولب باید به صورت منظم و با بخش‌بندی مشخص تنظیم شده باشد، و خروجی سلول‌های اجرا شده قابل مشاهده باشد. در گیت‌هاب هم برای هر مینی پروژه یک پوشه مجزا ایجاد کنید.
- هر جا از دفترچه‌کد گوگل کولب شما نیاز به فراخوانی فایلی خارج از محیط داشت، مطابق آموزش‌های ارائه شده ملزم هستید از دستور `gdown` استفاده کنید و مسیرهای فایل‌ها را طوری تنظیم کنید که صرفاً با اجرای سلول‌های کد، امکان فراخوانی و خواندن فایل‌ها توسط هر کاربری وجود داشته باشد.
- در تمامی مراحل تعریف داده و مدل و هر جای دیگری که مطابق آموزش‌های ویدیویی و به لحاظ منطقی نیاز است، Random State را برابر با دو رقم آخر شماره دانشجویی خود در نظر بگیرید.
- استفاده از ابزارهای هوشمند (مانند ChatGPT) در کمک‌گرفتن برای بهبود کدها مجاز است؛ اما لازم است تمام جزئیات مواردی که در خروجی‌های مختلف گزارش خود عنوان می‌کنید را به خوبی خوانده، درک و تحلیل کرده باشید. استفاده از این ابزارهای هوشمند در نوشتن گزارش و تحلیل‌ها ممنوع است.
- در جاهایی که با توجه به دو رقم آخر شماره دانشجویی خود محدود به انتخاب عدد، متغیر و یا داده‌ای خاص شده‌اید، برای تست‌های اضافه‌تر و نمایش بهبود در نتایج خود، مجاز هستید از مقادیر دیگر هم استفاده کنید. ۱۵ تا ۲۰ درصد از نمره هر سوال به بهترین نتایج کسب‌شده اختصاص خواهد یافت.
- رعایت نکات بالا به حرفه‌ای‌تر شدن شما کمک خواهد کرد و اهمیتی معادل مطالب درسی فراگرفته شده دارد؛ بنابراین، در صورت عدم رعایت هریک از این نکات، گزارش شما تصحیح نخواهد شد.

به انتخاب خود به یکی از پرسش‌های یک یا دو پاسخ دهید.

### ۱ پرسش یک: حل دنیای Wumpus

**Wumpus World** یک مسئله کلاسیک در هوش مصنوعی و یادگیری تقویتی است که شامل یک محیط مبتنی بر شبکه است که در آن یک عامل باید برای یافتن طلا حرکت کند و در عین حال از خطراتی مانند چاله‌ها و **Wumpus** اجتناب کند.

- اهداف پیمایش در شبکه **Grid**: عامل باید یاد بگیرد که به طور موثر در شبکه حرکت کند.
- اجتناب از خطرات: عامل باید یاد بگیرد که از چاله‌ها و **Wumpus** اجتناب کند.

- جمع آوری طلا: عامل باید طلا را پیدا کرده و جمع آوری کند.
- کشتن Wumpus: عامل می تواند برای کشتن Wumpus تیری شلیک کند و آن را به عنوان تهدید از بین ببرد.
- راه اندازی محیط شبکه: یک شبکه  $4 \times 4$  که در آن هر سلول می تواند خالی باشد، حاوی یک گودال، Wumpus یا طلا باشد.
- فضای اکشن ها: حرکت به بالا، پایین، چپ، راست.
- یک فلش را در هر یک از چهار جهت (بالا، پایین، چپ، راست) شلیک کنید (امتیازی).
- تصورات: Wumpus در شبکه با هر تغییر اکشن به اندازه یک خانه در راستای چپ، راست، بالا یا پایین حرکت می کند (امتیازی).
- فضای Reward:

- $100+$  برای گرفتن طلا
- $1000-$  برای افتادن در گودال یا خورده شدن توسط Wumpus
- $50+$  برای کشتن Wumpus (امتیازی)
- $1-$  برای هر حرکت
- تعریف محیط: یک شبکه  $4 \times 4$  با موقعیت های دلخواه برای چاله ها، Wumpus و طلا ایجاد کنید. حالت اولیه و حالت های ممکن را بعد از هر عمل تعریف کنید.
- تنظیم پارامترها:
  - نرخ یادگیری:  $0.1$
  - ضریب تخفیف:  $0.9$
  - نرخ اکتشاف: از  $1.0$  شروع می شود و در طول زمان کوچک میشود.

با توجه به موارد کلی گفته شده راجع به مسئله، موارد زیر را پاسخ دهید.

آ. برای این مسئله یک بار با روش Q-learning و یک بار با روش Deep Q-learning عاملی را طراحی کرده و آموزش دهید.

ب. عملکرد Policy:

- پاداش تجمعی را در اپیزودها برای هر دو عامل Q-learning و DQN ترسیم کنید. چگونه عملکرد عامل در طول زمان بهبود می یابد؟
- میانگین پاداش در هر اپیزود را برای هر دو عامل پس از  $1000$  اپیزود مقایسه کنید. کدام الگوریتم عملکرد بهتری داشت؟

ج. بحث کنید که چگونه نرخ اکتشاف اپسیلون بر فرآیند یادگیری تأثیر می گذارد. وقتی اپسیلون بالا بود در مقابل وقتی کم بود چه چیزی را مشاهده کردید؟

د. کارایی یادگیری:

- چند اپیزود طول کشید تا عامل Q-learning به طور مداوم طلا را بدون افتادن در گودال یا خورده شدن توسط Wumpus پیدا کند؟

- کارایی یادگیری Q-learning و DQN را مقایسه کنید. کدام یک Policy بهینه را سریعتر یاد گرفت؟

ه. معماری شبکه عصبی مورد استفاده برای عامل DQN را شرح دهید. چرا این معماری را انتخاب کردید؟

## ۲ پرسش دو: Deep Q-Learning برای محیط Lunar Lander

در این سوال از شما می خواهیم با استفاده از روش Deep Q-Learning برای محیط Lunar Lander یک عامل طراحی کرده و آموزش دهید که بتواند مسأله را حل کند. برای حل این تمرین می توانید از این دفترچه استفاده کرده و در صورت نیاز آن را تکمیل کرده و تغییر دهید. اگر از دفترچه ارايه شده و کدهای آماده استفاده می کنید، لازم است که تمام اجزای کد را با دقت توضیح دهید. دقت داشته باشید که هدف، توضیح روند عملکردی کد است، نه تحلیل خشک و لغتی برآمده از ChatGPT!

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

ب. عملکرد عامل را با رسم پاداش تجمعی در هر episode و 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. در صورتی که عملکرد عامل به ازای هر سه مقدار batch size مشابه یکدیگر شد، یکی از آن‌ها را به دلخواه به عنوان بهترین حالت انتخاب کنید. در رابطه با 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. هر بار آموزش عامل با استفاده از GPU های رایگان گوگل کولب حدوداً بین ۱۰ تا ۱۵ دقیقه زمان لازم خواهد داشت. برای تهیه خروجی می‌توانید checkpoint های مدل را دانلود کرده و روی سیستم خود فیلم‌ها را تهیه کنید.

---

**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**  
    Initialize sequence  $s_1 = \{x_1\}$  and preprocessed sequence  $\phi_1 = \phi(s_1)$   
    **for**  $t = 1, T$  **do**  
        With probability  $\epsilon$  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: شبکه‌د 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

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(R_{t+1} + \gamma \overbrace{Q'(s_{t+1}, \hat{a})}^{\text{estimated/expected Q-value}} - Q(s_t, a_t))$$

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

$$Q_{\text{estimated}} = Q(s_{t+1}, \hat{a})$$

شکل ۲: تفاوت مدل DDQN با DQN

منابع

[1] <https://github.com/MJAHMADEE/MachineLearning2024W>