



دانشگاه تهران  
پردیس دانشکده‌های فنی  
دانشکده برق و کامپیوتر



گزارش تمرین شماره ۵

درس یادگیری تعاملی

پاییز ۱۴۰۱

محمد سعادت

۸۱۰۱۹۸۴۱۰

## فهرست

چکیده.....	۳
سوالات تحلیلی.....	۴
۱-۱.....	۴
۲-۱.....	۴
۳-۱.....	۵
مسئله پیاده سازی.....	۷
بخش اول.....	۸
زیر بخش ۱.....	۸
زیر بخش ۲.....	۸
بخش دوم.....	۱۰
هدف سوال.....	۱۰
توضیح پیاده سازی.....	۱۰
نتایج و پاداش.....	۱۰
ارائه ویدیو (render).....	۱۱
پارامتر ها.....	۱۲
بخش سوم.....	۱۴
هدف سوال.....	۱۴
توضیح پیاده سازی.....	۱۴
نتایج.....	۱۴
بخش چهارم - امتیازی.....	۱۶
زیر بخش ۱.....	۱۶
زیر بخش ۲.....	۱۶
منابع.....	۱۶

## چکیده

---

با توجه به پیشرفت روز افزون مدل های یادگیری عمیق و همچنین نتایج قابل توجه ادغام این مدل ها با کاربردهای یادگیری تقویتی، در این تمرین به بررسی این الگوریتم ها و مدل ها خواهیم پرداخت. حل کردن مسائل یادگیری تقویتی با استفاده از شبکه های عصبی پیشینه ای قدیمی دارد و با توجه به پیشرفت سخت افزارهای محاسباتی در دو دهه اخیر، سرعت توسعه مدل های عمیق برای مسائل یادگیری تقویتی افزایش قابل ملاحظه ای داشته است. استفاده از شبکه های عصبی این امکان را به ما می دهد که از مسئله را با استفاده از یک مدل end to end حل کنیم. با توجه به این نکته کسب مهارت کارکردن با مدل های یادگیری عمیق و حل مسائل یادگیری تقویتی با استفاده از این مدل ها از مهارت های ضروری در زمینه یادگیری تقویتی می باشد. این تمرین مقدمه آشنایی ما با این مسائل را فراهم می کند.

## سوالات تحلیلی

### ۱-۱

الگوریتم پالیسی گردینت یک تکنیک یادگیری تقویتی برای بهینه سازی سیاست های کنترلی در محیط های پیچیده است. از بازخورد محیط برای بهبود یک سیاست با تنظیم آن استفاده می کند تا اکشن ها با expected reward بالاتر احتمال انتخاب شدن بیشتری داشته باشند. هدف، به حداکثر رساندن expected reward به دست آمده در هنگام پیروی از سیاست آموخته شده است.

الگوریتم پالیسی گردینت یک رویکرد تکرار سیاست است که در آن سیاست مستقیماً مدل سازی و بهینه سازی می شود تا به سیاست بهینه ای برسد که expected return را به حداکثر می رساند. این سیاست معمولاً با یک تابع پارامتر شده با توجه به  $\theta$ ،  $\pi_{\theta}(a|s)$  مدل می شود. مقدار تابع پاداش (هدف) به این سیاست بستگی دارد و سپس می توان الگوریتم های مختلفی را برای بهینه سازی برای بهترین پاداش اعمال کرد. این نوع الگوریتم ها یادگیری تقویتی model-free هستند. [3] [1]

روش های پالیسی گردینت نوعی از تکنیک های یادگیری تقویتی هستند که بر بهینه سازی سیاست های پارامتری با توجه expected return (پاداش تجمعی بلندمدت) با نزول گرادیان تکیه دارند. آنها از بسیاری از مشکلاتی که رویکردهای یادگیری تقویتی سنتی را مخدوش کرده اند، مانند فقدان تضمین یک value function، مشکل حل نشدنی ناشی از اطلاعات استیت نامشخص و پیچیدگی ناشی از استیت ها و اکشن های مستمر، رنج نمی برند. روش های پالیسی گردینت پارامتر سیاست را در هر مرحله در جهت برآورد گرادیان عملکرد با توجه به پارامتر سیاست به روزرسانی می کنند. این روش ها یک سیاست پارامتری را یاد می گیرند که امکان انجام اکشن ها را بدون مشورت با تخمین های action-value می دهد. روش پالیسی گردینت به طور مکرر وزن های شبکه سیاست را (با به روزرسانی های نرم) اصلاح می کند تا جفت های اکشن-استیت که منجر به بازده مثبت شده اند، و جفت های اکشن-استیت که منجر به بازده منفی شده اند، کمتر محتمل تر شوند. [4] [2]

### ۲-۱

• مزایا:

- الگوریتم های یادگیری تقویتی عمیق مزایای زیادی نسبت به الگوریتم های یادگیری تقویتی سنتی دارند، مانند یادگیری از ورودی های high-dimensional sensory، تعمیم بهتر به موقعیت های جدید و توانایی تصمیم گیری طولانی مدت. روش های یادگیری عمیق

همچنین مقیاس‌پذیری بهبود یافته و پایداری بهتر را هنگام برخورد با مجموعه داده‌های بزرگ ارائه می‌دهند. در نهایت، الگوریتم‌های یادگیری تقویت عمیق می‌توانند فضاهای حالت پیچیده و عملی را با دقت بیشتری نسبت به همتایان کم عمق خود نشان دهند.

○ استفاده از یادگیری عمیق در مسائل مربوط به high-dimensional state space بیشترین کاربرد را دارد. این بدان معناست که با یادگیری عمیق، یادگیری تقویتی به دلیل توانایی خود در یادگیری سطوح مختلف abstractions از داده‌ها، قادر به حل وظایف پیچیده‌تر با دانش قبلی پایین‌تر است. در الگوریتم‌های Deep RL پارامترها در طول زمان تنظیم می‌شوند و تعداد بهترین پارامترهای ممکن به سرعت با همگرا شدن مدل کاهش می‌یابد. [5] [7]

• معایب:

○ عیب اصلی الگوریتم‌های RL عمیق این است که برای درست کار کردن به مقدار زیادی داده و تجربه نیاز دارند. این می‌تواند پرهزینه و وقت‌گیر باشد، به خصوص زمانی که با وظایفی با پیچیدگی بالا سر و کار دارید. علاوه بر این، الگوریتم‌های RL عمیق هنوز هم مستعد بیش از حد overfitting هستند، که ممکن است منجر به عملکرد تعمیم ضعیف در نقاط داده نادیده شود. در نهایت، الگوریتم‌های RL عمیق نیز می‌توانند در همگرایی‌های خود ناپایدار باشند به دلیل مسائل چالش برانگیز بهینه‌سازی مرتبط با آنها.

○ Deep RL از نظر تئوری تضمین نمی‌شود که کار کند. می‌توان نشان داد که واگرا می‌شود، به این معنی که وزن‌ها تا بی‌نهایت منفجر می‌شوند. هرگز نمی‌توان تضمین داد که اوزان در یک کار جدید تا بی‌نهایت از هم جدا نمی‌شوند. این واگرایی در تنظیمات خطی زمانی اتفاق نمی‌افتد که از شبکه عمیق استفاده نمی‌شود، بلکه از بردار وزن‌ها استفاده می‌کند. [6]

○ Deep RL به زمان بسیار زیادی برای آموزش نیاز دارد. [6]

○ برای اینکه نسبت به سایر تکنیک‌ها بهتر عمل کند به حجم بسیار زیادی داده نیاز دارد. [8]

○ به دلیل مدل‌های پیچیده داده، آموزش بسیار گران است. [8]

### ۳-۱

بافرهای تجربه در یادگیری تقویتی عمیق (RL) برای ذخیره و پخش مجدد تجربیات گذشته استفاده می‌شوند. این تکنیک در خدمت این است که یک عامل را قادر می‌سازد تا با استفاده از داده‌های جمع

آوری شده از تعاملات قبلی با محیط، کارآمدتر و سریع تر یاد بگیرد. بافرهای تجربه فرآیند آموزش را با کاهش میزان تعاملات زمان واقعی که یک عامل باید انجام دهد، تسریع می کند، زیرا می تواند از حافظه استخراج کند و تجربیات گذشته را دوباره پخش کند. این روش همچنین sample efficiency را با اجازه دادن به آموزش مجدد مکرر بر اساس اطلاعاتی که قبلاً به دست آورده اید، به جای جمع آوری مداوم داده های جدید، بهبود می بخشد.

ایده اصلی پشت experience replay، ذخیره تجربیات گذشته و سپس استفاده از زیرمجموعه ای تصادفی از این تجربیات برای به روزرسانی Q-network است، نه تنها استفاده از آخرین تجربه. ما می توانیم با استفاده از یک بافر بزرگ از تجربیات گذشته خود و نمونه داده های آموزشی از آن، به جای استفاده از آخرین تجربه خود، از نوسان یا واگرایی فاجعه بار مقادیر عمل جلوگیری کنیم. این تکنیک بافر تکرار یا بافر تجربه نامیده می شود. بافر پخش مجدد شامل مجموعه ای از تاپل های تجربه  $(S, A, R, S')$  است. تاپل ها به تدریج به بافر اضافه می شوند که ما در حال تعامل با محیط هستیم. ساده ترین پیاده سازی یک بافر با اندازه ثابت است که داده های جدیدی به انتهای بافر اضافه می شود تا قدیمی ترین تجربه را از آن خارج کند. بافرهای بازپخش تجربه معمولاً برای الگوریتم های یادگیری تقویتی off-policy با گرفتن تمام نمونه های تولید شده توسط یک عامل در تعامل با محیطش و سپس ذخیره آنها برای استفاده مجدد بعدی اعمال می شوند و به روش یادگیری ما اجازه می دهد چندین بار روی آنها به روز شود. [9] [10]

## مسئله پیاده سازی

---

در این تمرین ما با محیط معروف highway کار خواهیم کرد. این محیط در کتابخانه gym پیاده سازی شده است و تسک های 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 drive در یک محیط دو بعدی می باشند.

## بخش اول

### زیر بخش ۱

- استیت: استیت به عنوان یک شبکه  $4 \times 4$  تعریف می شود که موقعیت فعلی ماشین ها و جاده خالی را نشان می دهد. هر خودرو موقعیت شروع منحصر به فرد خود را دارد که با عدد صحیح بین ۰ تا ۱۵ نشان داده می شود که نشان می دهد در حال حاضر در کدام مربع است. مربع ها خالی هستند مگر اینکه ماشینی آن را اشغال کند. استیت عامل نشان دهنده وضعیت Observations محیط، سرعت، وضعیت تصادف، اکشن انتخاب شده، پاداش ها، وضعیت خاتمه برنامه و وضعیت truncated می باشد.
- اکشن: شامل ۵ عمل تغییر لاین به چپ، انجام یک اکشن غیرقابل دسترس (عمل غیر مجاز) (انجام هیچ کاری) (IDLE)، تغییر لاین به راست، افزایش سرعت و کاهش سرعت می باشد.
- پاداش: به طور کلی بر دو ویژگی متمرکز شده است: یک وسیله نقلیه باید به سرعت در جاده پیشرفت کند. از برخورد اجتناب کنید. بنابراین، تابع پاداش اغلب از یک جمله سرعت و یک جمله برخورد تشکیل شده است:

$$R(s, a) = a \frac{v - v_{\min}}{v_{\max} - v_{\min}} - b \text{ collision}$$

که  $v_{\max}$   $v_{\min}$   $v$  به ترتیب سرعت فعلی، حداقل و حداکثر سرعت خودروی ego هستند و  $b$  و  $a$  دو ضریب هستند. در برخی محیط ها، وزن جریمه برخورد را می توان از طریق پارامتر  $\text{collision\_penalty}$  پیکربندی کرد.

### زیر بخش ۲

- اکشن: اکشن ها در این تسک در حالت پیش فرض از نوع DiscreteMetaAction هستند.
  - اکشن پیوسته: نوع ContinuousAction به عامل اجازه می دهد تا به طور مستقیم کنترل های سطح پایین سینماتیک خودرو، یعنی دریچه گاز  $a$  و زاویه فرمان  $\delta$  را تنظیم کند. (کنترل دریچه گاز و فرمان را می توان به ترتیب از طریق تنظیمات طولی و جانبی فعال یا غیرفعال کرد. بنابراین، فضای عمل می تواند یک بعدی یا دو بعدی باشد).
  - اکشن گسسته: DiscreteAction یک کوانتیزاسیون یکنواخت از ContinuousAction فوق است. پارامتر  $\text{actions\_per\_axis}$  اجازه می دهد تا مرحله کوانتیزاسیون را تنظیم کرد.



مشابه اکشن های پیوسته، محور طولی و جانبی را می توان به طور جداگانه فعال یا غیرفعال کرد.

○ Meta-Actions گسسته: نوع DiscreteMetaAction لایه ای از کنترل کننده های سرعت و فرمان را در بالای کنترل سطح پایین پیوسته اضافه می کند، به طوری که خودروی ego می تواند به طور خودکار جاده را با سرعت دلخواه دنبال کند. سپس، متا اکشن های موجود شامل تغییر خط هدف و سرعت است که به عنوان نقطه تنظیم برای کنترل کننده های سطح پایین استفاده می شود.

- استیت: استیت ها در این محیط در حالت پیش فرض گسسته هستند. برای observations در این محیط در حالت پیش فرض از Kinematics استفاده شده است. KinematicObservation یک آرایه  $V \times F$  است که فهرستی از  $V$  خودروهای مجاور را با مجموعه ای از ویژگی های اندازه  $F$ ، که در قسمت پیکربندی «features» فهرست شده اند، توصیف می کند. برای مثال:

Vehicle	$x$	$y$	$v_x$	$v_y$
ego-vehicle	5.0	4.0	15.0	0
vehicle 1	-10.0	4.0	12.0	0
vehicle 2	13.0	8.0	13.5	0
...	...	...	...	...
vehicle V	22.2	10.5	18.0	0.5

## بخش دوم

### هدف سوال

در این بخش قرار است مسئله را با استفاده از الگوریتم Deep Q-Learning حل کنیم. برای پیاده سازی الگوریتم ها مطابق شبه کد آن در مقاله داده شده عمل می کنیم: [11]

---

**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  $\epsilon$  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  $(y_j - Q(\phi_j, a_j; \theta))^2$  according to equation 3
    end for
end for
```

---

### توضیح پیاده سازی

کد های مربوط به پیاده سازی این قسمت در فایل `IL_HW5.ipynb/html` قسمت **Part 2 – Deep Q-Learning** قرار دارد.

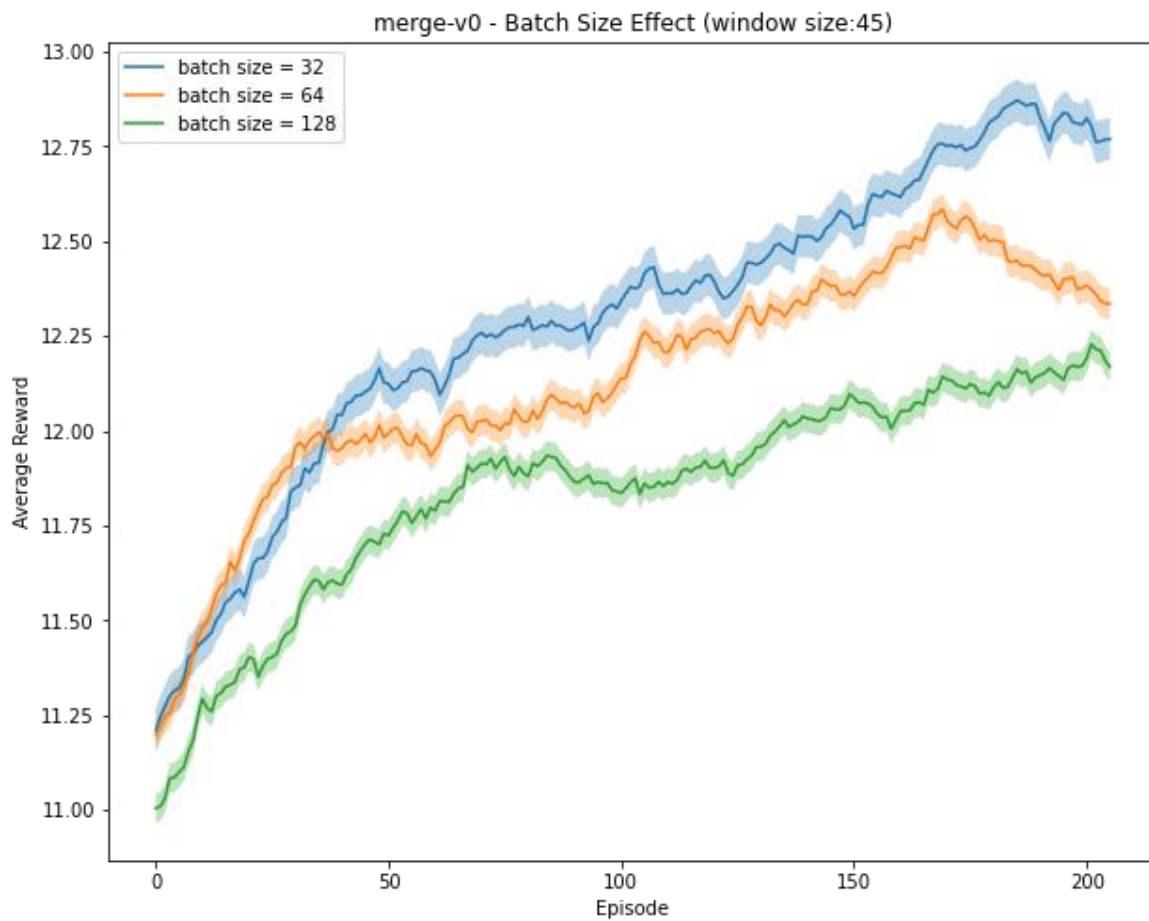
- در کلاس **ExperienceBuffer** بافر تجارب پیاده سازی شده است.
- در کلاس **DQN** قسمت شبکه عصبی الگوریتم پیاده سازی شده است.
- در کلاس **DeepQLearning** الگوریتم Deep Q-Learning پیاده سازی شده است.
- تابع `run_DeepQLearning_agent` الگوریتم Deep Q-Learning را با ۱۰ بار تکرار و به تعداد ۲۵۰ اپیزود اجرا می کند و پاداش های دریافت شده را بر می گرداند.

برای اجرا کد های این بخش سوال است تا کد های بخش ابتدایی و کد های این بخش را اجرا کنید.

### نتایج و پاداش

کد های مربوط به پیاده سازی این قسمت در فایل `IL_HW5.ipynb/html` قسمت **Part 2 – Reward** قرار دارد.

نمودار پاداش کسب شده توسط عامل در طول یادگیری برای ۱۰ بار اجرا الگوریتم به ازای ۲۵۰ اپیزود به همراه بازه اطمینان ۹۵ درصد برای مقادیر مختلف به عنوان `batch_size` به شکل زیر است:



## ارائه ویدیو (render)

کدهای مربوط به پیاده سازی این قسمت در فایل [IL\\_HW5.ipynb/html](https://github.com/IL-HW5.ipynb/html) قسمت **Part 2 – Render** قرار دارد.

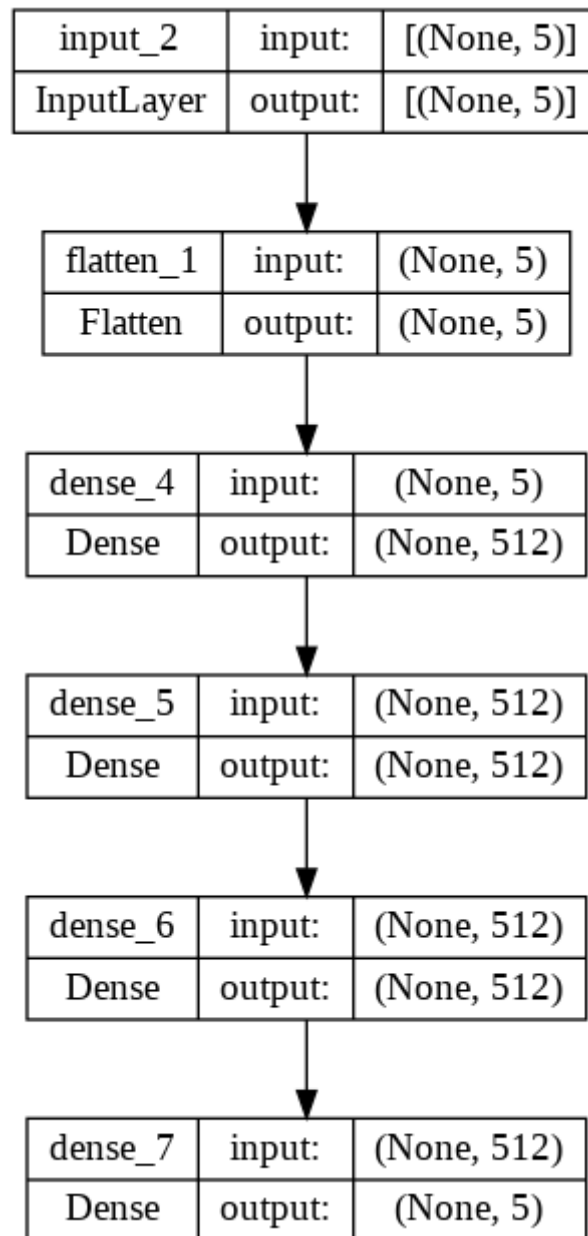
## پارامترها

کد های مربوط به پیاده سازی این قسمت در فایل *IL\_HW5.ipynb/html* قسمت **Part 2** – **Hyperparameters** قرار دارد.

مقادیر پارامترها به شرح زیر است:

Hyperparameters	Value
Optimizer	Adam
Optimizer learning rate	1e-4
Discount factor	0.99
Epsilon decay rate	0.97
Epsilon start	1
Epsilon minimum	0.01
Number of episodes	250
Number of replies	10

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 5)]	0
flatten_1 (Flatten)	(None, 5)	0
dense_4 (Dense)	(None, 512)	3072
dense_5 (Dense)	(None, 512)	262656
dense_6 (Dense)	(None, 512)	262656
dense_7 (Dense)	(None, 5)	2565
=====		
Total params: 530,949		
Trainable params: 530,949		
Non-trainable params: 0		



### هدف سوال

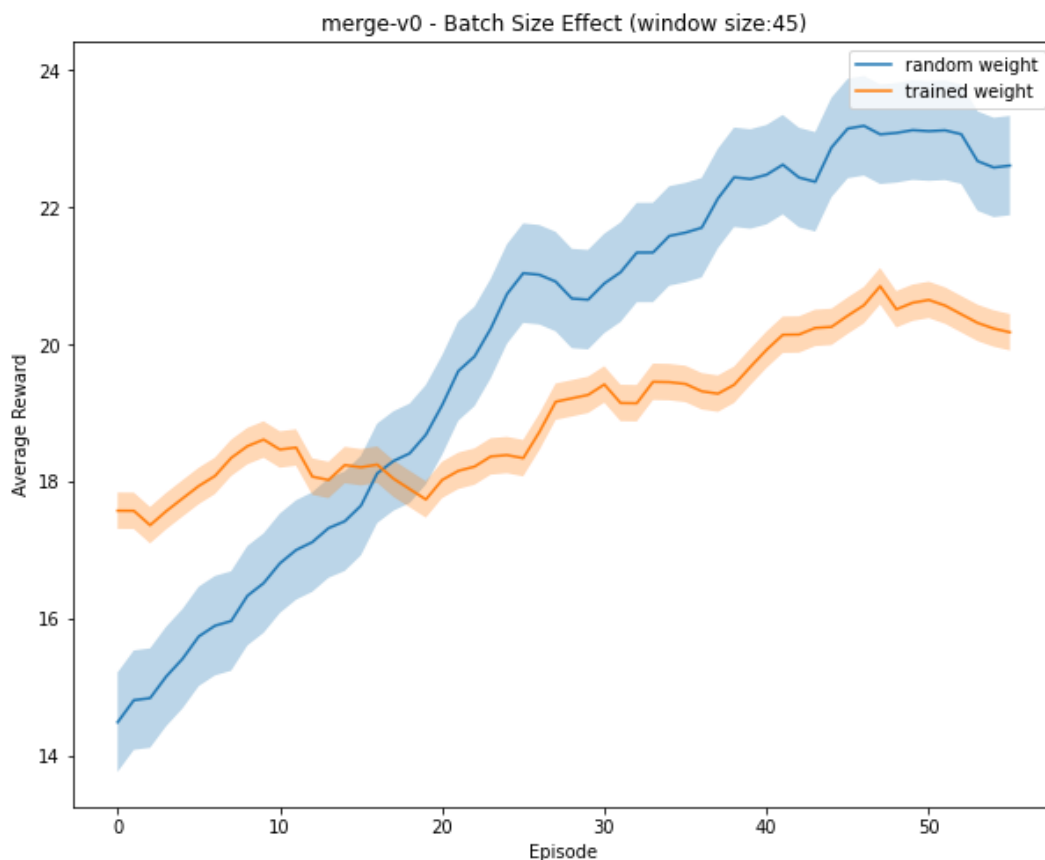
در این بخش قرار که تاثیر انتقال تجربه یاد گرفته شده در یک محیط به محیط دیگر را بررسی کنیم. برای این منظور ابتدا عامل را در محیط merge-v0 آموزش می دهیم. سپس در محیط highway-v0 در دو حالت عامل با وزن دهی رندم و عامل train شده، به بررسی سرعت یادگیری در محیط جدید می پردازیم.

### توضیح پیاده سازی

کد های مربوط به پیاده سازی این قسمت در فایل [IL\\_HW5.ipynb/html](#) قسمت **Part 3** قرار دارد. برای اجرا کد های این بخش سوال است تا کد های بخش ابتدایی و کد های این بخش را اجرا کنید.

### نتایج

نتایج حاصل از اجرا این بخش به شکل زیر است:



مشاهده می شود که سرعت یادگیری افزایش می یابد و در حالتی که از وزن های تسک قبلی استفاده کردیم، نسبت به حالت که عامل را با وزن دهی رندم اجرا کرده ایم، پاداش در اپیزود های اولیه بیشتر است. علت دریافت پاداش بیشتر در اپیزود های اولیه این است که عامل قبلاً یک یادگیری ای انجام داده است. در نتیجه در تسک جدید دیگر از حالت رندم یادگیری را آغاز نمی کند و استفاده از تجربه کسب شده در تسکی دیگر، منجر به افزایش پاداش در اپیزود های اولیه در تسک جدید می شود. همچنین سرعت یادگیری نیز افزایش می یابد زیرا به دلیل استفاده از تجربه یادگیری شده، نرخ همگرایی کاهش می یابد در نتیجه عامل در تعداد گام های کمتری در تسک جدید آموزش می بیند. استفاده از وزن دهی تسک قبلی آموزش داده شده در وزن دهی شبکه برای تسک جدید، باعث می شود تا دیگر از یک حالت رندم یادگیری در تسک جدید را آغاز نکنیم و با استفاده از تجربه یادگیری شده در تسک قبل، یادگیری در تسک جدید را با سرعت بیشتری انجام دهیم و روند یادگیری هدفمند تر و مبتنی بر تجربه کسب شده انجام می شود.

استفاده از وزن های آموزش دیده برای عامل در RL عمیق، سرعت یادگیری را افزایش می دهد، زیرا شبکه قبلاً با اطلاعات آزمایش قبلی seed شده است. این به عامل اجازه می دهد تا به جای شروع هر بار از ابتدا، بر آنچه قبلاً آموخته است، بسازد. علاوه بر این، عاملی که از وزن های train شده استفاده می کند، با داشتن تنوع بیشتر شرایط اولیه در مقایسه با عواملی که training را از یک مجموعه وزن های اولیه به طور تصادفی شروع می کنند، شانس بیشتری برای اجتناب از local optima دارند. در نهایت، این بدان معنی است که exploration کمتری مورد نیاز است و عامل می تواند سریعتر به سیاست های بهینه خود برسد.

استفاده از وزن های از پیش آموزش دیده در یادگیری تقویتی عمیق به عامل اجازه می دهد تا از نقطه شروع بهتری تعامل با محیط را آغاز کند. این «warm start» یادگیری را برای عامل آسان تر می کند، زیرا از قبل نمایش خوبی از ویژگی های خاصی مانند وظایف تشخیص شی یا پیش بینی حالت های مفید در یک محیط دارد. این امر میزان زمان آموزش و منابع محاسباتی مورد نیاز برای یادگیری کار را کاهش می دهد و در نتیجه سرعت یادگیری در RL عمیق را افزایش می دهد.

زیر بخش ۱

زیر بخش ۲

منابع

---

- 1- <https://medium.com/intro-to-artificial-intelligence/reinforce-a-policy-gradient-based-reinforcement-learning-algorithm-84bde440c816>
- 2- [http://www.scholarpedia.org/article/Policy\\_gradient\\_methods](http://www.scholarpedia.org/article/Policy_gradient_methods)
- 3- <https://lilianweng.github.io/posts/2018-04-08-policy-gradient/#policy-gradient>
- 4- <https://towardsdatascience.com/policy-gradient-methods-104c783251e0>
- 5- <https://viso.ai/deep-learning/deep-reinforcement-learning/#:~:text=The%20use%20of%20deep%20learning,levels%20of%20abstractions%20from%20data.>
- 6- <https://www.quora.com/What-is-the-downside-of-deep-reinforcement-learning-When-shouldnt-it-be-used>
- 7- <https://www.unite.ai/what-is-deep-reinforcement-learning/>
- 8- <https://www.rfwireless-world.com/Terminology/Advantages-and-Disadvantages-of-Deep-Learning.html#:~:text=Drawbacks%20or%20disadvantages%20of%20Deep%20Learning&text=%E2%9E%A8It%20requires%20very%20large,increases%20cost%20to%20the%20users.>
- 9- <https://towardsdatascience.com/deep-q-network-dqn-ii-b6bf911b6b2c>
- 10- <https://towardsdatascience.com/a-technical-introduction-to-experience-replay-for-off-policy-deep-reinforcement-learning-9812bc920a96>
- 11- <https://arxiv.org/pdf/1312.5602.pdf>
- 12-