محیط های ناشناخته

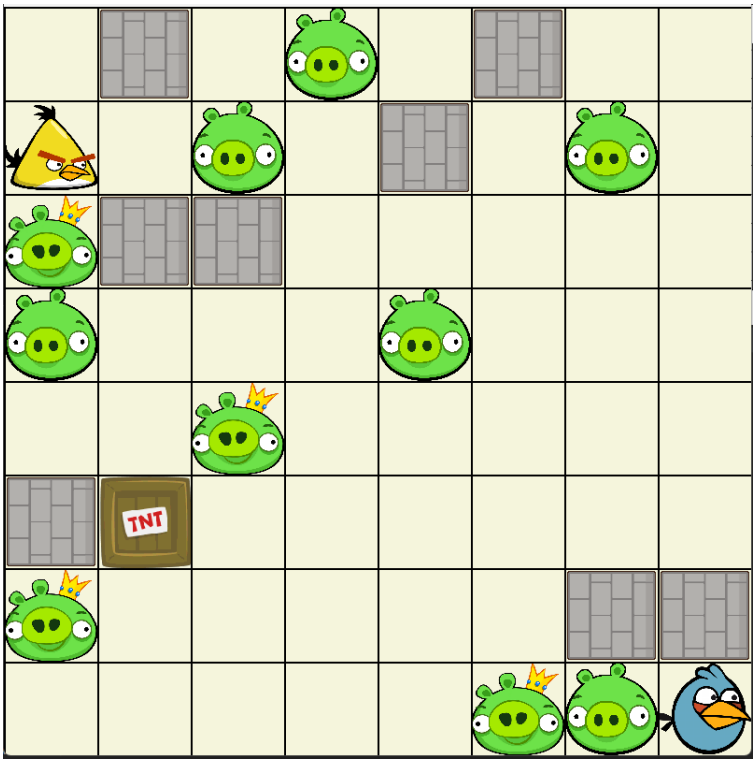
در این بخش باید الگوریتمی را پیاده‌سازی کنید تا بتواند در محیط ناشناخته فعالیت کند و یک سیاست برای حل مسئله استخراج کند. نام این محیط Unknown Angry Birds است.

۱-۲-معرفی محیط

یک محیط Grid به ابعاد 8x8 هستیم. در این محیط باید پرنده زرد که در ابتدای بازی در موقعیت (0, 0) یعنی بالا-چپ قرار دارد به پرنده آبی برسد که در موقعیت (7, 7) یعنی پایین-راست قرار دارد. پرنده زرد دارای چهار کنش **بالا، پایین، چپ و راست** می‌باشد. در صورت رسیدن به پرنده آبی **٤٠٠ امتیاز** کسب خواهید کرد و بازی **خاتمه** می‌یابد. هر کنش پرنده زرد نیز **یک امتیاز منفی** در پی خواهد داشت.

- در این محیط **۸ مانع** قرار دارند که عامل نمی‌تواند از آنها عبور کند. این موانع در ابتدای ایجاد محیط بصورت تصادفی در خانه‌ها قرار می‌گیرند.
- در این محیط **۸ خوک** وجود دارد که در صورت برخورد پرنده زرد با آنها عامل **۲۵۰ امتیاز مثبت** دریافت کرده و آنها را از بین می‌برد. موقعیت قرارگیری این ۸ خوک در خانه‌ها در ابتدای ایجاد محیط بطور تصادفی مشخص می‌شود.
- در این محیط **۲ ملکه‌خوکی** قرار دارند که در صورت برخورد عامل با هر ملکه، **٤٠٠ امتیاز منفی** دریافت می‌کند. موقعیت این ملکه‌ها نیز در ابتدای ایجاد محیط بطور تصادفی تعیین می‌شود.
- در این محیط یک TNT قرار دارد که در صورت برخورد با آن عامل **۲۰۰۰ امتیاز منفی** دریافت کرده و بازی **خاتمه** می‌یابد. موقعیت TNT بطور تصادفی در ابتدای تعیین می‌شود.
- عامل فقط **۱۵۰ کنش** فرصت دارد تا به پرنده آبی برسد. در صورتیکه تعداد کنش‌های عامل بیشتر از این تعداد شود، **۱۰۰۰ امتیاز منفی** دریافت کرده و بازی **خاتمه** می‌یابد.
- مانند بخش قبلی عامل دارای کنش‌های تصادفی است. کنش در نظر گرفته شده برای عامل با احتمال خاصی انجام خواهد شد که به این کنش، کنش اصلی گفته می‌شود. کنش‌اصلی دارای دو کنش همسایه است که آنها نیز احتمال خاصی دارند و ممکن است یکی از آنها به جای کنش اصلی انجام شود. توجه کنید که توزیع احتمالاتی کنش‌ها در یک بازی ثابت ولی در بازی مختلف می‌تواند فرق داشته باشد. کنش اصلی و کنش‌های همسایه آن به شرح زیر هستند:

کنش اصلی	کنش همسایه	کنش همسایه
بالا	چپ	راست
پایین	چپ	راست
چپ	بالا	پایین
راست	بالا	پایین



شکل۲-۱ محیط *Unknown Angry Birds*

۲-۲-مراحل پیاده‌سازی

۱. در ابتدا باید با استفاده از الگوریتم‌های موجود در محیط‌های ناشناخته، یک سیاست در محیط استخراج کنید. انتخاب الگوریتم برعهده شما می‌باشد. توصیه می‌شود الگوریتم‌های مختلف را پیاده‌سازی کرده و عملکرد هر کدام را بررسی کنید.

۲. تفاوت مهم این بخش نسبت به بخش قبلی در این است که عامل معمولاً نمی‌تواند تنها با یک دور (episode) از فعالیت در محیط سیاست بهینه را استخراج کند. به همین دلیل لازم عامل دورهای متعددی از فعالیت در محیط را سپری کند و دانش استخراج شده از آنها را با هم ترکیب کند تا یادگیری تحقق پیدا کند. به عنوان مثال اگر دانش استخراج شده از محیط در قالب یک جدول Q باشد، عامل باید در ابتدای هر دور جدول ذخیره شده قبلی را بازیابی کرده، در حین فعالیت در آن دور آن را بروزرسانی کرده و دوباره در پایان دور آن را ذخیره کند، تا در نهایت به یک سیاست خاص همگرا شود.

۳. حال باید در حین اجرای الگوریتم اطلاعاتی را استخراج کنید تا بتوانیم همگرایی الگوریتم شما را بررسی کنیم. بعد از هر episode که جدول Q* بروزرسانی شد، معیار زیر را محاسبه کنید. نام آن Value Difference می‌باشد.

$$Value\ Diffrence\ = \sum_{s\in S}\sum_{a\in A}\left|Q^{(k+1)}(s,a) - Q^{(k)}(s,a)\right|$$

توجه کنید که این فرمول بر روی جدول Q* قبل از بروزرسانی (اشاره به اندیس k) و پس از بروزرسانی در یک episode (اشاره به اندیس k+1) را در نظر می‌گیرد. شما در هر episode از اجرای الگوریتم باید این مقدار را محاسبه کرده و ذخیره کنید و در نهایت نمودار تغییرات آنرا رسم کنید. توجه کنید در صورت همگرایی الگوریتم شما، این معیار باید به مرور کاهش یابد. زمانیکه Value Difference کمتر از یک مقدار مثل ∈ بشود به معنای آن است که می‌توانیم اجرای الگوریتم را متوقف کنیم. . مقدار ∈ یک عدد بسیار کوچک (به عنوان مثال 0.01 یا 0.001) است.

۴. در راستای ارزیابی سیاست استخراج شده نیاز است که روی یک نقشه، کنشی که سیاست برای آن حالت پیشنهاد داده است را نشان دهید. یک نقشه به ابعاد 8x8 بسازید و روی هر حالت کنش پیشنهادی سیاست را نمایش دهید.

۵. پس از همگرایی الگوریتم، با توجه به سیاست استخراج شده در ماتریس *Q عامل باید به فعالیت در محیط بپردازد.

۲-۳-ارزیابی

در ابتدا الگوریتم شما مورد بررسی قرار گرفته و صحت و شیوه پیاده‌سازی آن ارزیابی خواهد شد. سپس همگرایی الگوریتم شما با توجه به نموداری که در بخش ۲-۲ معرفی شد بررسی خواهد شد.
پس از اجرای الگوریتم و بررسی همگرایی آن، عملکرد عامل برای فعالیت در محیط بر اساس سیاست استخراج شده مورد ارزیابی قرار می‌گیرد. مانند بخش قبلی عامل در تعدادی محیط و در هر محیط ۵ دور فعالیت خواهد کرد. هر دور از فعالیت عامل با رسیدن به یکی از شرایط پایان بازی متوقف شود و که میانگین امتیازاتش در همه دورها محاسبه می‌شود. در نهایت نمره شما براساس برآیند امتیازاتی که عامل در محیط‌های مختلف کسب کرده است، مطابق با جدول زیر مشخص خواهد شد.

میانگین امتیاز دریافتی	نمره
mean score > 1100	100 %
mean score > 850	70 %
mean score > 600	50 %
برخورد با TNT (در صورتیکه سیاست استخراج شده عامل را به سمت TNT هدایت کند)	0 %

۲-۴-بخش امتیازی

پیاده‌سازی یک شبکه‌عصبی مبتنی بر Deep QLearning با استفاده از کتابخانه‌های Pytorch یا TensorFlow. ورودی شبکه‌عصبی باید حالتی از محیط باشد و لایه خروجی شبکه‌عصبی باید دارای ۴ خروجی باشد(به تعداد کنش‌ها). عامل برای تصمیم‌گیری در انتخاب کنش در هر حالت باید از این شبکه استفاده کند.

موفق باشید