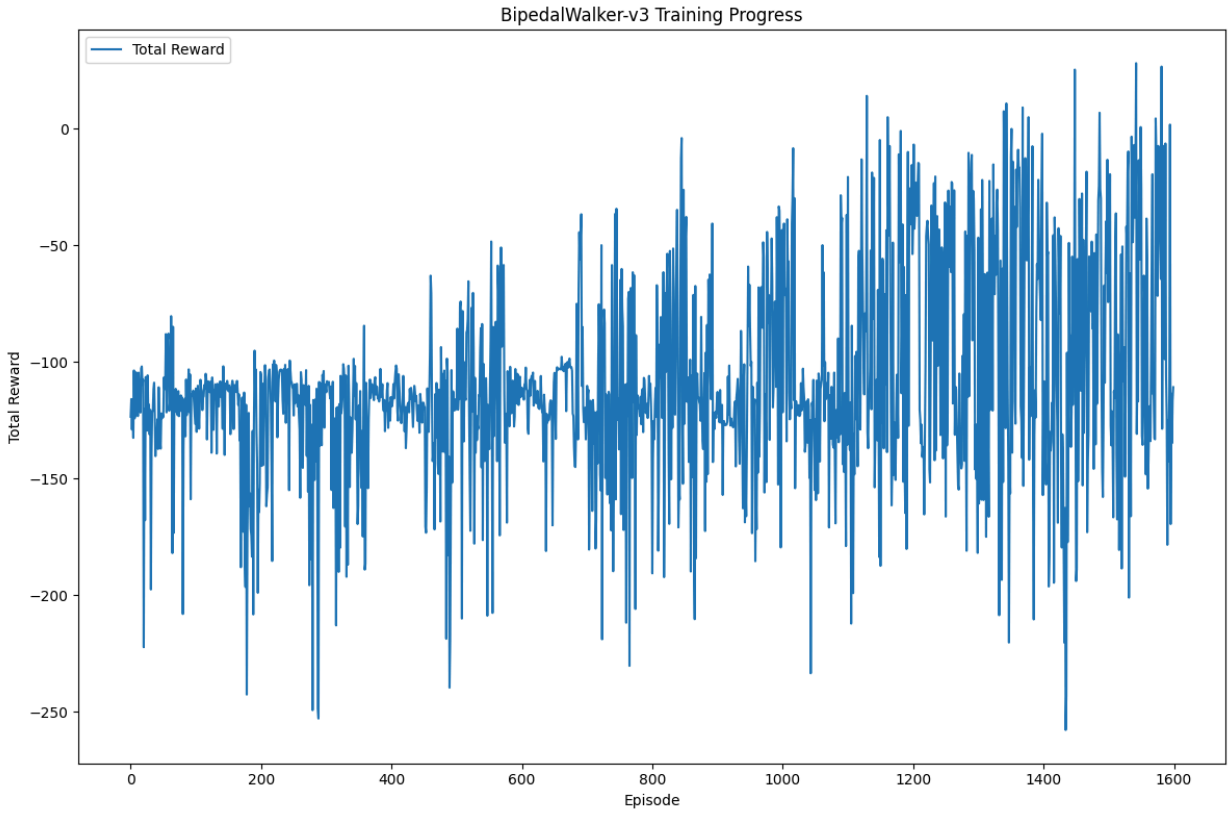
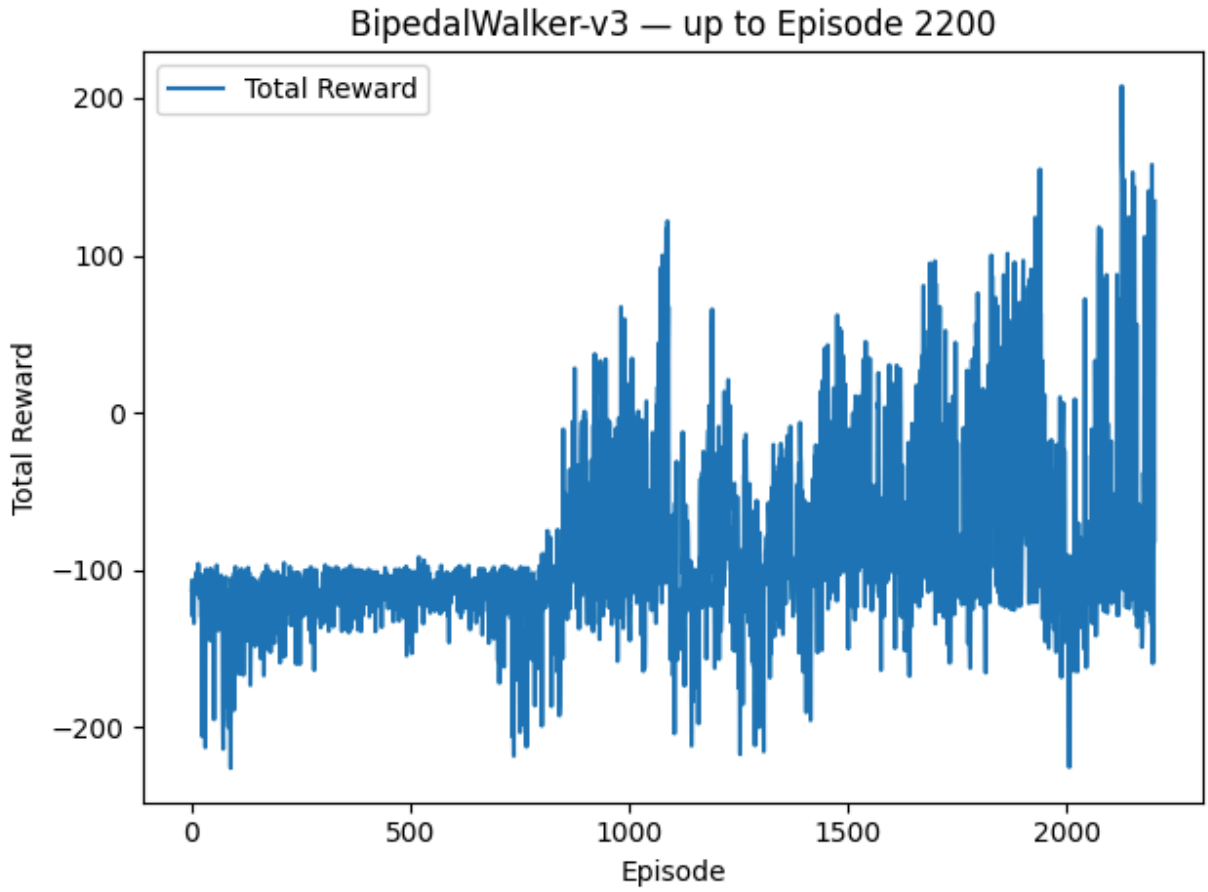
خب درود
چند تا experiment ران کردیم یکی dqn ، دیگری double dqn و d3qn که شبکه عصبی dueling double q network هستش.

برای dqn عادی داریم:



در بررسی روند آموزش الگوریتم DQN بر روی محیط BipedalWalker-v3، می‌توان سه بازه‌ی زمانی متمایز را شناسایی کرد که هر کدام ویژگی‌های خاص خود را دارند. در قسمت ابتدایی آموزش، یعنی از اپیزود ۰ تا حدود ۳۰۰، پاداش‌ها نسبتاً ثابت و در سطح پایین باقی مانده‌اند (در حدود ۱۱۰-)، که نشان می‌دهد عامل یادگیرنده هنوز در مرحله‌ی اکتشاف قرار دارد و موفق نشده سیاست مؤثری برای حل مسئله یاد بگیرد. این موضوع طبیعی است، چرا که در مراحل اولیه، عامل هنوز تجربه کافی برای درک دینامیک محیط و تشخیص رفتارهای بهینه را به‌دست نیاورده است. در بازه‌ی میانی بین اپیزودهای ۳۰۰ تا حدود ۱۰۰۰، نوسانات زیادی در مقادیر پاداش مشاهده می‌شود. با اینکه در برخی اپیزودها شاهد بهبود نسبی در عملکرد هستیم، اما به دلیل ناپایداری در فرآیند یادگیری، پاداش‌ها همچنان پراکندگی زیادی دارند. این ناپایداری ممکن است به دلیل عدم استفاده از ساختارهایی نظیر شبکه‌ی هدف (target network) باشد، چرا که در نبود آن، مقادیر Q-value با همان شبکه‌ای که در حال آموزش است تخمین زده می‌شوند و این می‌تواند منجر به نوسانات شدید در گرادیان‌ها و مقادیر پیش‌بینی شده شود. در

نهایت، در اپیزودهای پایانی یعنی از ۱۰۰۰ تا ۱۶۰۰، به‌تدریج نشانه‌هایی از یادگیری موفق مشاهده می‌شود؛ برخی پاداش‌ها به مقدار نزدیک به صفر یا حتی بالاتر از آن می‌رسند که می‌تواند حاکی از آن باشد که عامل در برخی موارد موفق به تکمیل مسیر شده است. با این حال، همچنان نوسانات بسیار زیاد هستند و در اپیزودهایی شاهد افت شدید عملکرد تا حدود ۲۰۰- نیز هستیم. این موضوع نشان می‌دهد که گرچه روند یادگیری آغاز شده و به سمت بهبود حرکت می‌کند، اما پایداری کامل هنوز حاصل نشده و مدل نیازمند تقویت بیشتر و استفاده از تکنیک‌های پیشرفته‌تر برای بهبود پایداری در آموزش است.

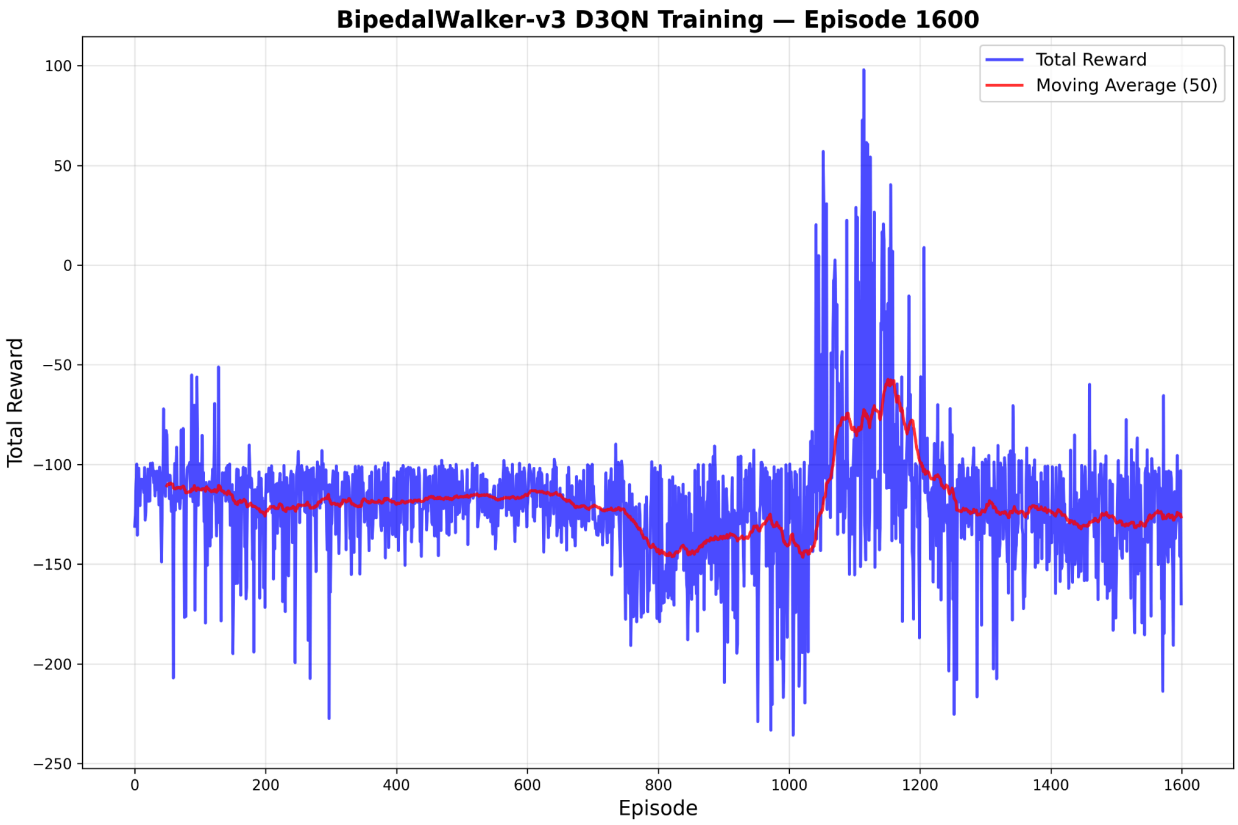


نموداری که ارائه شده نمایانگر عملکرد الگوریتم Double DQN در محیط `BipedalWalker-v3` است و در مقایسه با نسخه‌ی ساده‌ی DQN، نشان‌دهنده‌ی پیشرفت چشمگیر در ثبات و کیفیت یادگیری است. در اپیزودهای ابتدایی (تا حدود اپیزود ۷۰۰)، روند آموزش همچنان در فاز اکتشاف قرار دارد و مقادیر پاداش عمدتاً منفی و نسبتاً ثابت هستند که در این مرحله

امری طبیعی به شمار می‌آید. اما نکته‌ی قابل توجه، کاهش محسوس نوسانات شدید نسبت به DQN ساده است که نشان از کنترل بهتر فرآیند یادگیری دارد.

از حدود اپیزود ۹۰۰ به بعد، نمودار افزایش تدریجی و قابل‌توجهی را در پاداش‌ها نشان می‌دهد. این به معنای آن است که عامل یادگیرنده به تدریج توانایی راه رفتن مؤثر را کسب می‌کند. در اپیزودهای بعدی، به‌ویژه پس از اپیزود ۱۵۰۰، الگوریتم در برخی اپیزودها موفق به کسب پاداش‌هایی بالای صفر و حتی نزدیک به ۲۰۰ شده است؛ موضوعی که بیانگر تسلط نسبی عامل بر محیط و توانایی در طی مسیر به صورت موفقیت‌آمیز است.

علت اصلی عملکرد بهتر Double DQN نسبت به نسخه‌ی پایه، در شیوه‌ی تخمین مقادیر Q نهفته است. در این الگوریتم، با جدا کردن مرحله‌ی انتخاب و ارزیابی عمل در Q-value estimation، از بروز خطای بیش‌برآوردی (Overestimation) جلوگیری می‌شود. همین موضوع باعث شده تا فرآیند به‌روزرسانی مقادیر Q به شکلی پایدارتر و دقیق‌تر انجام گیرد.



در نمودار مربوط به آموزش الگوریتم D3QN در محیط BipedalWalker-v3، روند یادگیری در طول ۱۰۰۰ اپیزود به‌صورت دقیق قابل مشاهده است. در ابتدا، پاداش‌ها منفی هستند و عامل یادگیرنده عملکرد ضعیفی دارد که مطابق انتظار در مراحل اولیه آموزش است. با این حال، نکته‌ی مهمی که در مقایسه با الگوریتم‌های قبلی مانند DQN و Double DQN قابل توجه است، این است که با گذر زمان، گرچه مقادیر پاداش همچنان منفی باقی مانده‌اند، اما به‌تدریج از لحاظ آماری متمرکزتر شده و واریانس آن‌ها کاهش یافته است؛ به‌طوری‌که از اپیزودهای حدود ۲۰۰ به بعد، نوسانات شدید پاداش کمتر شده‌اند و الگوریتم به نوعی پایداری نسبی دست یافته است. برخلاف نسخه‌های قبلی، در اینجا الگوریتم موفق شده است در پایان دوره‌ی آموزش، بهبود قابل توجهی در میانگین پاداش‌ها ایجاد کند و در اپیزودهای پایانی شاهد افزایش ناگهانی و چشمگیر پاداش‌ها به سمت مقادیر مثبت هستیم که نشان‌دهنده یادگیری موفقیت‌آمیز عامل است. با این حال، در همین نقطه‌ی اوج عملکرد، یک سقوط شدید و ناگهانی رخ می‌دهد که به وضوح در نمودار دیده می‌شود؛ این پدیده که به آن سقوط عملکرد (performance collapse) گفته می‌شود، یکی از چالش‌های شناخته‌شده در آموزش با الگوریتم‌های تقویتی پیچیده به‌ویژه در محیط‌هایی مانند BipedalWalker است. علت آن معمولاً به ناپایداری در به‌روزرسانی Q-value یا وابستگی بیش از حد به تجارب اخیر در حافظه بازپخش مربوط می‌شود. در مجموع، این نمودار گویای آن است که D3QN نسبت به نسخه‌های قبلی عملکردی منسجم‌تر و هدفمندتر دارد، اما همچنان نیاز به بهبودهایی در پایداری بلندمدت و جلوگیری از افت ناگهانی عملکرد در مراحل پیشرفته آموزش دیده می‌شود.

اها راستی یچیو یادم رفت اینکه حافظه با الویت داشتیم و هربار بر اساس td الویت بندی میشه و خب البته در نهایت جواب نداد :(و im sad