کنترل ربات انسان‌نما در حرکت بر روی سطوح ناهموار با روش‌های یادگیری تقویتی عمیق

نام و نام خانوادگی: امیرحسین منصوری استاد راهنما: دکتر حامد جلالی

Isfahan University of Technology

مقدمه

ربات‌های انسان‌نما با توانایی اجرای وظایف پیچیده در محیط‌های غیرقابل پیش‌بینی، می‌توانند نقش کلیدی در آینده فناوری و زندگی بشر ایفا کنند. نمونه‌هایی مانند Boston Dynamics Atlas و Tesla Optimus نمایانگر پیشرفت‌های قابل توجهی در سخت‌افزار ربات‌های انسان‌نما هستند، اما کنترل‌کننده‌های این ربات‌ها هنوز به‌طور کامل یا جزئی به صورت دستی برای وظایف خاص طراحی می‌شوند که نیازمند تلاش‌های مهندسی گسترده و پیچیده برای هر وظیفه جدید است.

یادگیری تقویتی عمیق به عنوان یکی از روش‌های نوین، به ربات‌ها امکان می‌دهد از طریق تجربه و تعامل با محیط، رفتارهای مقاوم و تطبیق‌پذیر بیاموزند. برخلاف کنترل‌کننده‌های سنتی، الگوریتم‌های یادگیری تقویتی به ربات‌ها کمک می‌کنند تا به طور خودکار و با انعطاف‌پذیری بالا در شرایط مختلف سازگار شوند.

در این پروژه، هدف یافتن یک الگوریتم یادگیری تقویتی مناسب به همراه تعریف بهینه فضای حالت، فضای عمل و تابع پاداش است، به طوری که بیشینه سازی پاداش و یافتن سیاست بهینه منجر به حرکت طبیعی و پایدار ربات بر روی سطوح مختلف شود.

مقدمات فنی

الگوریتم Proximal Policy Optimization

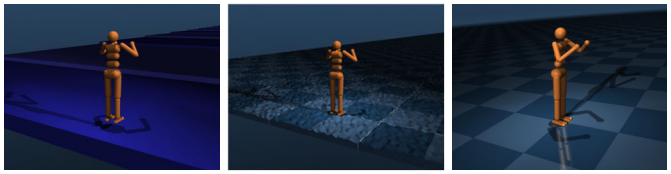
یک الگوریتم بدون مدل و مبتنی بر سیاست است که از سبک بازیگر-منتقد برای به‌روزرسانی سیاست‌ها استفاده می‌کند. مانند TRPO، هدف آن بهبود حداکثری سیاست‌ها بدون افت عملکرد است. در این پروژه، از نسخه Clipped Surrogate الگوریتم PPOاستفاده شده است که تابع هدف آن به صورت زیر تعریف می‌شود:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, CLIP(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$$

که در آن $r_t(\theta)$ برابر با $\frac{\pi_{\theta_{new}}(S_t|A_t)}{\pi_{\theta_{old}}(S_t|A_t)}$، $\hat{A}_t$ تابع مزیت و ϵ یک فراپارامتر است.

مدل ربات انسان‌نما

در این پروژه، از مدل آدمک ۲۷ درجه آزادی با ۲۱ موتور با کنترل گشتاور استفاده شده است و الگوریتم ها برروی چهار سطح صاف، سنگلاخی، شیب و پله‌ای آموزش دیده و ارزیابی شده‌اند.



شبیه‌سازی سناریوهای فوق در کتابخانه Mujoco MJX انجام شده است که قابلیت شبیه‌سازی دینامیکی سیستم را با استفاده از GPU فراهم می‌کند.

خلاصه روش‌ها

بیش از ۲۶۰ آزمایش برای یافتن ترکیب بهینه از روش‌ها، تابع پاداش، فضای مشاهده و عمل انجام شد که در شرایط مشابه، روش‌های زیر بهبود عملکرد سیستم را به همراه داشتند.

کاهش فضای عمل و فضای مشاهده

آزمایش‌ها نشان داد با حذف حالات‌های مربوط به آرنج و دو درجه آزادی از سه درجه آزادی شانه از فضای عمل و فضای مشاهده می توان ضمن بهبود ظاهری حرکت، پایداری ربات را بهبود بخشید.

زیان تقارن

اضافه کردن عبارت زیان تقارن به تابع الگوریتم PPO، با محدود کردن فضای جستجوی سیاست، منجر به همگرایی سریع‌تر و حرکات متقارن، طبیعی‌تر و بهینه‌تر می‌شود و با استفاده از رویکرد زیان تقارن، مدل به تولید حرکات متعادل و پایدار تشویق می‌گردد.

$$L_{symmetry} = \frac{1}{N_A}\sum_{i=0}^{N_A}(\pi(O_t) - \overline{\pi(\bar{O}_t)})^2$$

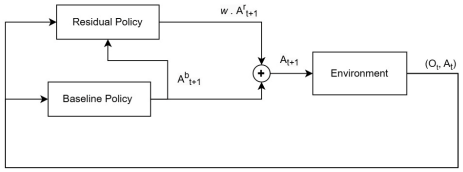
که در آن N_A اندازه فضای عمل، O_t فضای مشاهده و عملگر $\bar{}$ ، ورودی را با توجه به راست و چپ ربات قرینه می کند.

تقلیل فضای حالت سیاست به نیمی از دوره‌ی حرکت

فرض کنید که چرخه حرکت هر پا دقیقاً مشابه دیگری است. بنابراین، در نیمه اول چرخه، حرکت آموزش داده می‌شود و در نیمه دوم، با استفاده از قرینه‌سازی فضای مشاهده و خروجی شبکه سیاست، ادامه حرکت یادگرفته می شود. این روش با بهره‌گیری از تقارن حرکت و پارامترسازی چرخه با متغیر فاز، پیچیدگی فضای جستجوی سیاست را کاهش داده و به حرکات متوازن‌تر منجر می‌شود.

یادگیری باقی مانده

در روش یادگیری باقی‌مانده، از ترکیب دو عامل استفاده شده است، به‌طوری که عامل اول برای حرکت روی سطح صاف آموزش دیده و عامل دوم با یادگیری تفاوت‌های حرکت روی سطوح ناهموار، سیاست پایه را تقویت و بهبود می‌بخشد.

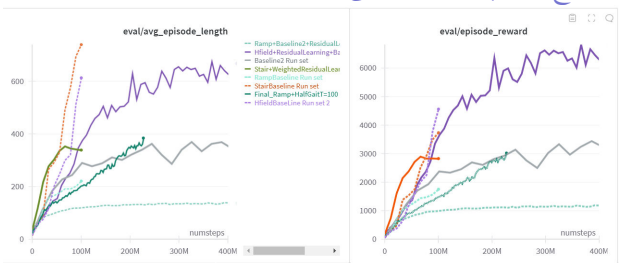


نتایج نهایی آزمایش‌ها

تنظیمات نسخه نهایی

Name	Dimension	Description
Action space	15	All exclude arm states
Observation space	59	Joint positions, velocities, and previous actions
Reward	3	Forward velocity reward, alive reward, control cost
Policy	(59, 32, 32, 32, 32, 30)	Feedforward neural network with 4 hidden layers of 32 units each

مقایسه نتایج نسخه پایه و نسخه نهایی



نتیجه‌گیری

طراحی سیستم کنترل ربات‌های انسان‌نما با استفاده از یادگیری تقویتی عمیق، چالشی است که نیازمند توازن دقیق میان المان‌های مختلف سیستم می‌باشد. در این پروژه، با بهره‌گیری از مقالات علمی روز و استخراج روش‌های قابل پیاده‌سازی با سخت‌افزار موجود، از یادگیری باقیمانده و زیان تقارن همراه با طراحی مناسب تابع پاداش و فضای مشاهده و عمل استفاده شد. این رویکرد، امکان تعمیم حرکت ربات از سطح صاف به سطوح ناهموار مانند سنگلاخی، پله‌ای و شیب را در تعداد تعاملات معقول با محیط فراهم کرد.

Slide 1

AM0 xfd

Amir Hossein Mansouri,
2024-09-17T17:06:04.213