بهبود یادگیری عامل یادگیرنده اجتماعی به کمک مشاهده رفتار عامل‌های دیگر

دانشجو: نیما زمان پور

استاد راهنما: مجید نیلی احمدآبادی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

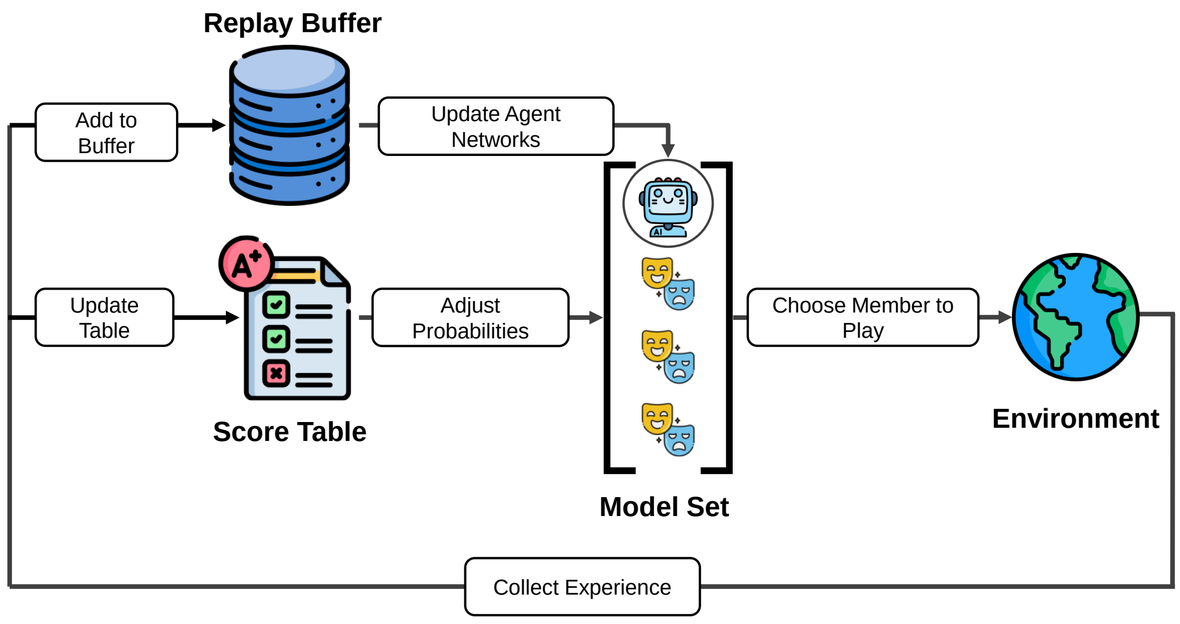
یادگیری تقویتی در زمینه آموزش عامل‌های هوشمند برای تصمیم‌گیری مستقل در محیط های پیچیده نوید زیادی نشان داده‌است. با این حال، مهمترین چالش این زمینه زمان بسیار زیاد مورد نیاز برای تعامل با محیط برای دریافت بازخورد و یادگیری سیاست بهینه می‌باشد. در محیط هایی که هزینه زمانی یا مالی کسب تجربه فردی کمتر از هزینه پردازش سخت افزاری است؛ وجود دیگر عامل‌های اجتماعی این فرصت را می‌دهد که از استخراج تجربه و دانش آن‌ها برای افزایش سرعت یادگیری استفاده کرد. در این پروژه قصد داریم بدون برقراری ارتباط و تنها با مشاهده رفتار عامل های اجتماعی حاضر در محیط، تجربه آنان را استخراج نموده و با ایجاد مدلی از آن عامل، و شناسایی سطح تخصص آن، به بهره‌برداری از دانش عامل برای تعامل با محیط به جای عامل کم‌تجربه ما و نیز تشکیل یک داده‌گان از تجربیات عامل متخصص بپردازیم. این روش گام بلندی در راستای افزایش سرعت یادگیری و کاهش تعاملات کم‌بهره برمی‌دارد.

مدل پیشنهادی

در این روش فرض می‌شود عامل ما در یک محیط به همراه سایر عامل‌های اجتماعی قرار گرفته‌است. عامل توانایی دیدن حالت و عمل بقیه عامل‌های اجتماعی را دارد. اما پاداش از دید او پنهان است. اولین قدم عامل جمع آوری و تشکیل یک دادگان از دوتایی های حالت-عمل است. دادگان به دو صورت آفلاین و آنلاین تهیه می‌شود. در حالت آفلاین فرض می‌شود که به اندازه کافی داده در اختیار هست که مدلی دقیقی از عامل ساخته شود. در حالت آنلاین فرض می‌شود عامل‌های اجتماعی همزمان با عامل ما در محیط شروع به تعامل کرده و تجربیات به تدریج تولید می‌شود.

سپس با استفاده از روش Behavior Cloning(BC) مدلی از هر کدام از عامل‌های اجتماعی تشکیل می‌دهیم. در روش آنلاین که دادگان بتدریج تکمیل می‌شود هر چند قسمت یکبار روش BC را اجرا می‌کنیم تا مدل آپدیت شود.

سپس تمامی مدل ها به همراه عامل ما تشکیل یک دسته می‌دهند. در هر قسمت یکی از اعضای این دسته انتخاب می‌شود و آن عضو در محیط تعامل می‌کند. پاداشی که در این قسمت بدست می‌آید متعلق به کل دسته است و به نام عامل ما ثبت می‌شود.

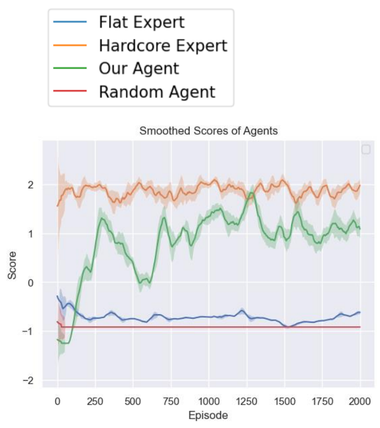


شکل ۱- دیاگرام روش پیشنهادی

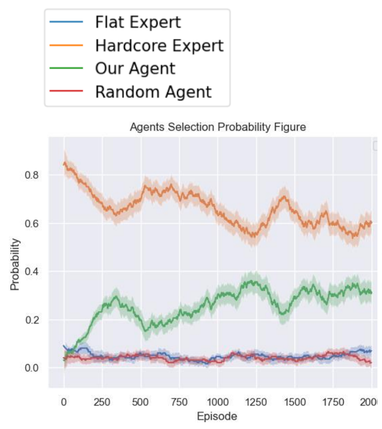
معیار انتخاب اعضا بر اساس امتیازی است که کسب می‌کنند. امتیاز هر عضو برابر میانگین ۲۰ قسمت قبلی است که در محیط تعامل داشته. این امتیاز ها در ابتدا نرمالایز شده و سپس با عبور از تابع SoftMax به احتمال تبدیل می‌شوند. در ابتدای هر قسمت از خروجی این تابع یک نمونه گرفته و عضو انتخابی در محیط تعامل می‌کند. از الگوریتم Soft Actor-Critic(SAC) به همراه Conservative Q-Learning(CQL) برای آموزش عامل استفاده می‌شود.

نتایج

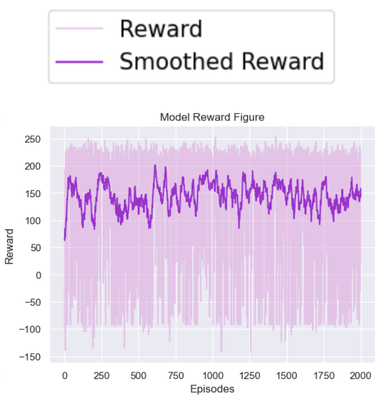
با انجام شبیه‌سازی‌ها در محیط BipedalWalker-v3 Hardcore کتابخانه gym، نتایج زیر بدست آمد:



نمودار امتیاز اعضای دسته

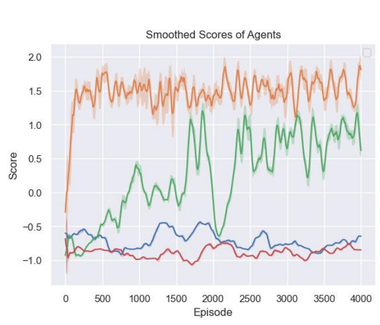


نمودار سهم انتخاب اعضای دسته

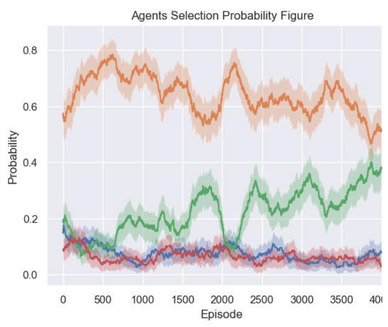


نمودار پاداش دسته

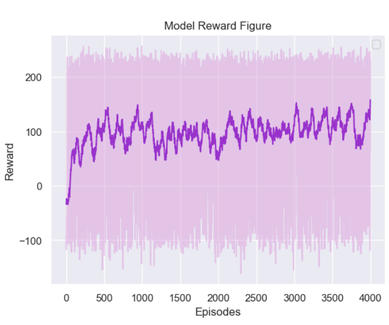
آزمایش ۱: عامل خبره، عامل خبره محیط دیگر، عامل رندوم (آفلاین)



نمودار امتیاز اعضای دسته

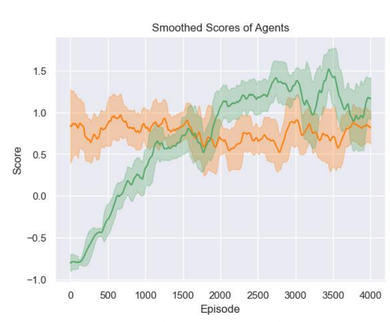


نمودار سهم انتخاب اعضای دسته

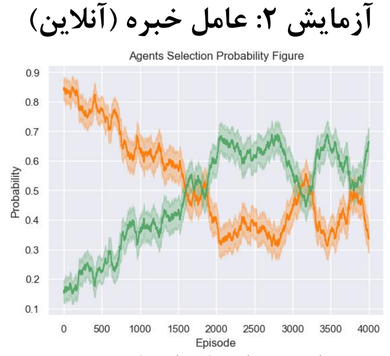


نمودار پاداش دسته

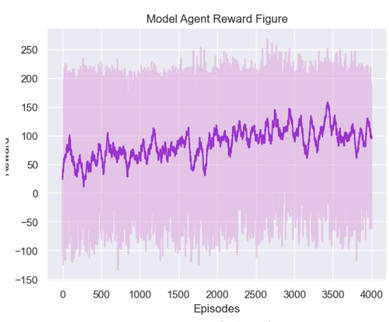
آزمایش ۲: عامل خبره (آنلاین)



نمودار امتیاز اعضای دسته



نمودار سهم انتخاب اعضای دسته



نمودار پاداش دسته

آزمایش ۳: عامل نیمه خبره، عامل خبره محیط دیگر، عامل رندوم (آفلاین)

جمع بندی

با بکارگیری این روش، بعد از گذشت زمان کمی از شروع آموزش عامل‌های خبره شناسایی شده و شانس بیشتری برای تعامل با محیط خواهند داشت. و نیز تمام تجربیات اعضای دسته به دادگان آموزش عامل ما اضافه می‌شود. بدین ترتیب عامل ما بجای شروع از صفر، مستقیما دادگان خود را با دیتا های باکیفیت و سودمند پر می‌کند. که باعث می‌شود عملکرد روش بسیار رضایت بخش باشد. و در تمام حالاتی که دانش قابل استخراجی در محیط وجود داشته باشد، موفق به بهره‌برداری از آن و سرعت بخشیدن به یادگیری تا چند برابر حالت تکی شده است.

مراجع اصلی

[1] T. Haarnoja, A. Zhou, P. Abbeel and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018.

[2] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft Actor-Critic Algorithms and Applications," arXiv:1812.05905, 2018.

[3] A. Kumar, A. Gupta, and S. Levine, "Conservative Q-Learning for Offline Reinforcement Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.