Мультизадачное обучение интеллектуальных агентов на скрытых представлениях

И.Н. Аглюков
Высшая Школа Экономики
Москва, Россия
ildariwe@gmail.com

К.В. Святов
Ульяновский государственный технический университет
Ульяновск, Россия
k.svyatov@ulstu.ru

С.В. Сухов
УФИРЭ им. В. А. Котельникова РАН
Ульяновск, Россия
ssukhov@ulireran.ru

***Аннотация*—Рассматривается проблема объединения опыта нескольких агентов при мультизадачном глубоком обучении с подкреплением. Обмен опытом осуществляется путем дистилляции знаний между предобученными агентами-учителями и учеником. Буфер опыта, используемый для хранения накопленных состояний, преобразуется в скрытые представления автокодировщика, что снижает требования к используемым вычислительным ресурсам.**

Ключевые слова— обучение с подкреплением, катастрофичное забывание, автокодировщик, дистилляция знаний.

1. Введение

Показывая поразительные результаты в решении различных частных задач, при попытках мультизадачного обучения, глубокое обучение с подкреплением сильно страдает от проблемы “катастрофической интерференции” [1]. Градиенты, привносимые сразу несколькими задачами, выступают в роли шума, который мешает нормальному обучению агента. При последовательном обучении агента на нескольких задачах в нескольких окружениях предыдущий опыт взаимодействия со средами (окружениями) может храниться в так называемом буфере опыта. При этом возникает другой ряд проблем, связанный с обучением агента без взаимодействия со средой на ранее полученном опыте (при "офлайн обучении"). Здесь очевидной проблемой является ограниченность полученного опыта и отсутствие возможности исследовать среду в поисках областей пространства состояний, дающих большее вознаграждение [2]. Другой не менее важной проблемой является проблема хранения состояний, так как многие подходы обучения с подкреплением основаны на постепенных инкрементальных улучшениях поведения интеллектуального агента, происходящих за счет многократного прохождения по накопленному буферу, а решение проблемы катастрофической интерференции требует хранения максимально возможного количества состояний для процесса передачи знаний.

В нашей работе был разработан метод, способный в значительной мере сократить потери при многозадачном обучении с подкреплением без взаимодействия со средой (Рис. 1). Также был опробован способ сокращения объема необходимой памяти для хранения буфера опыта путем сжатия состояний во внутреннее представление нейронной сети.

2. Описание метода передачи знаний

На первом этапе мы обучаем агентов-учителей с помощью немного модифицированного алгоритма глубокого Q обучения (Dueling Deep Q learning) [3]. В качестве сред для обучения агентов были выбраны среды эмулирующие игры Atari. Во время обучения мы сохраняли наборы состояний для дальнейшего процесса передачи знаний между учителями и учеником (копией одного из агентов учителей). Сама передача знаний осуществлялась посредством популярного в глубоком обучении метода “Дистилляции” [4], что позволило нам свести проблему обучения с подкреплением к проблеме обучению с учителем. Агентам-учителям подается набор состояний из их собственных буферов опыта; тот же набор подается ученику. Затем значения функции полезности (Q-функции) учителей и ученика пропускаются через функцию softmax, тем самым получая вероятность того или иного действия. Полученные значения подаются в функции потерь для каждой задачи и по сумме результатов функций потерь происходит расчет градиента и оптимизация сети с помощью алгоритма обратного распространения ошибки.

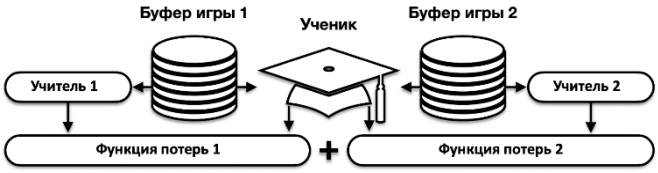

Рис. 1. Диаграмма процесса передачи знаний

Описанный алгоритм апробирован на наборе игр Atari с разным количеством действий (4 действия - Breakout, Atlantis, 6 действий - Pong, Up and Down, DemonAttack) и различными функциями потерь. Наши исследования показали, что агент ученик набирает сопоставимое с учителем количество очков, но немного уступает учителю (Таблица 1).

Таблица 1. Результаты переноса знаний

Учитель 1 **Учитель 2**	**Учитель**	**Ученик на части буфера**	**Ученик на скрытых представлениях**
Atlantis **Breakout**	$(2,7\pm1,3)\times10^5$ 212±45	$(2,3\pm1,8)\times10^5$ 200±53	$(1,8\pm0,2)\times10^5$ 8,4±2
Demon Attack **Up and Down**	9251±3301 6607±917	6169±2599 4526±328	9284±2987 4893±471
Pong **Up and Down**	15,9±1,43 6608±918	15,54±0,86 6336±886	15,43±0,78 6397±686

Средние набранные очки агентов за тест

Процесс тестирования агента ученика состоит из 10 циклов по 12 500 состояний. Такое количество состояний примерно соответствует тридцати минутам игры в

реальном времени. Ограничение по времени тестирования было продиктовано тем, что агент может зациклиться, и игра будет продолжаться бесконечно, но агент не будет набирать очки. Мы брали среднее количество очков за один цикл, а потом среднее за все циклы.

Одной из проблем «офлайн обучения» является отсутствие возможности поместить весь сохранённый объем знаний учителя в оперативную память. Как следствие, буфер для передачи знаний не покрывает даже весь спектр состояний, которые были доступны учителю при первичном обучении. Одно из решений – использовать для обучения лишь часть буфера (Таблица 1). Другое возможное решение этой задачи – перевод содержимого буфера в сжатое представление. Для сжатия состояний мы воспользовались автокодировщиком [5], который позволил сократить потребляемую память приблизительно в двадцать семь раз путем перевода изображений во внутреннее представление нейронной сети.

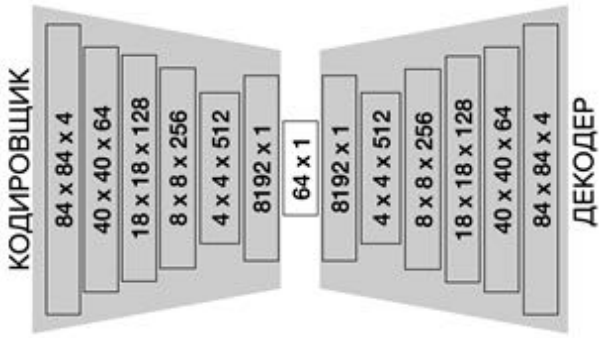

Рис. 2. Схема автокодировщика

Для каждой среды был обучен автокодировщик, сжимающий четыре кадра состояния в 64 цифры с плавающей точкой (Рис. 2). Автокодировщик обучался на всем буфере каждой среды, а затем все сохраненные состояния были закодированы и сохранены на диске. При использовании автокодировщика алгоритм процесса передачи знаний имеет лишь небольшое изменение в части получения состояния, а именно – декодирование из скрытого представления набора состояний для последующей дистилляции знаний. Результаты обучения агента-ученика с помощью скрытых представлений показаны в Таблице 1. Так как известные ограничения автокодировщика делают изображение мутным, такой способ подходит не ко всем средам и в наших экспериментах в игре Breakout была потеряна часть информации, на которую, по-видимому, опирался агент учитель. Это привело к плохим результатам при передаче знаний в данном конкретном случае.

3. Заключение

Мы предложили и протестировали разновидность метода клонирования поведения с помощью дистилляции. Результаты передачи знаний представлены в Таблице 1. Разработанный нами метод успешно справляется с проблемой катастрофического забывания и может использоваться при многозадачном обучении с подкреплением.

Использование набора скрытых признаков позволяет уменьшить объем используемой памяти для хранения буфера опыта, но увеличивает время обучения агента, так как приходится тратить время на декодирование набора состояний для дальнейшей дистилляции знаний. Стоит отметить и тот факт, что такой метод имеет свои ограничения ввиду особенностей автокодировщика и требует дальнейшего совершенствования подхода.

Благодарности

Исследование выполнено при финансовой поддержке Российского Фонда Фундаментальных Исследований и Правительства Ульяновской области (проект № 18-47-732006).

Литература

[1] Teh, Y.W. Distral: Robust multitask reinforcement learning / Y.W. Teh, V. Bapst, W.M. Czarnecki, J. Quan, J. Kirkpatrick, R. Hadsell, N. Heess, R. Pascanu // Advances in Neural Information Processing Systems. – 2017. – P. 4496-4506.

[2] Levine, S. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems / S. Levine, A. Kumar, G. Tucker, J. Fu // ArXiv preprint: 2005.01643, 2020.

[3] Wang, Z. Dueling Network Architectures for Deep Reinforcement Learning / Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas // PMLR. – 2016. – Vol.48. – P.1995-2003.

[4] Hinton, G. Distilling the knowledge in a neural network / G. Hinton, O. Vinyals, J. Dean // ArXiv preprint: 1503.02531, 2015.

[5] Ballard, D. Modular learning in neural networks / D. Ballard // Proc. of the sixth National conf. on Artificial intelligence. – 1987. – Vol. 1. – P. 279-284.