DOI 10.37882/2223–2966.2022.10.14

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ НА ОСНОВЕ МОДЕЛИ ИЕРАРХИЧЕСКОЙ ТЕМПОРАЛЬНОЙ ПАМЯТИ

REINFORCEMENT LEARNING BASED ON HIERARCHICAL TEMPORAL MEMORY MODEL

G. Kanonir

Summary. Modern reinforcement learning methods have a number of limitations imposed by the used artificial neural networks paradigm with a point neuron model. The use of the latest achievements of neuroscience within the theory of intelligence «The Thousand Brains Theory of Intelligence», as well as the application of the machine learning model «Hierarchical Temporal Memory» (HTM), which implements some aspects of this theory, have the potential both to develop already established reinforcement learning methods, and to create new approaches for solving this problem. The purpose of this work is to identify the prospects for using the HTM machine learning model in reinforcement learning.

Keywords: biologically-plausible machine learning methods, reinforcement learning, hierarchical temporal memory.

Канонир Георгий
Аспирант
Университет ИТМО (Санкт-Петербург)
kanonirs@gmail.com

Аннотация. Современные методы обучения с подкреплением имеют ряд ограничений, наложенных использующейся парадигмой искусственных нейронных сетей с точечной моделью нейрона. Использование последних достижений нейронаук в рамках теории интеллекта «The Thousand Brains Theory of Intelligence», а также применение модели машинного обучения "Иерархическая Темпоральная Память" (Hierarchical Temporal Memory, HTM), которая реализует некоторые аспекты данной теории, имеют потенциал как для развития уже устоявшихся методов обучения с подкреплением, так и для создания новых подходов решения этой задачи. Целью данной работы является выявление перспектив применения модели машинного обучения HTM в обучении с подкреплением.

Ключевые слова: биологически-правдоподобные методы машинного обучения, обучение с подкреплением, иерархическая темпоральная память.

Введение

Обучение с подкреплением является направлением машинного обучения, в рамках которого моделируется взаимодействие агента с некоторой средой, в которой данный агент находится. Целью обучения является определение оптимальной стратегии принятия решений, основываясь на (возможно, неполных) наблюдениях среды и некотором сигнале вознаграждения, который косвенно даёт отложенную оценку принятых решений и таким образом ставит перед агентом задачу [1].

Для принятия эффективных решений агенту необходимо оценивать свое текущее состояние, основываясь на предыдущем опыте, т.к. именно оно определяет, какие решения агент может принять в данный момент и принятие какого решения может быть наиболее выгодно. В большинстве реальных задач пространство состояний, а в ряде случаев и пространство действий настолько велики, что единственным выходом является применение аппроксимирующих методов.

Современные методы обучения с подкреплением используют парадигму искусственных нейронных сетей с точечной моделью нейрона для выполнения таких аппроксимаций. Это налагает ряд ограничений на подобные методы, включая слабую устойчивость к шуму во входных данных [2], низкую эффективность хранения информации в модели, приводящей к появлению проблемы катастрофического забывания и невозможности непрерывного обучения [3], а также низкую эффективность самого процесса обучения в целом [4]. Кроме того, в процессе обучения с подкреплением обычно используется буфер, обновление которого тоже представляет непростую задачу, для хранения некоторого подмножества данных, полученных от взаимодействия агента с окружающей средой. В реальных задачах получение такого опыта может быть очень затратным и трудоёмким, а кроме того, и сам процесс обучения может быть неэффективным при отсутствии достаточного количества и разнообразия опыта.

В настоящее время достигнуты огромные успехи при решении некоторых задач с помощью методов машинного обучения в целом и методов обучения с подкреплением в частности [5, 6], тем не менее наиболее продвинутыми агентами остаются живые существа, обладающие мозгом, а точнее говоря, те из них, у которых

есть неокортекс, представляющий собой отдел мозга, отвечающий за интеллект.

Одной из наиболее перспективных теорий интеллекта, учитывающей данные исследований принципов строения и функционирования мозга, является теория интеллекта «The Thousand Brains Theory of Intelligence», разрабатываемая Дж. Хоукинсом и его коллегами [7] из компании Numenta. Помимо этого, исследователи и инженеры компании занимаются разработкой модели машинного обучения Иерархическая Темпоральная Память (Hierarchical Temporal Memory, HTM), постепенно внедряя свои наработки в стремлении реализовать теорию интеллекта Дж. Хоукинса в виде вычислительной модели [8].

Модель HTM является нейронной сетью, но использует более сложную и приближенную к естественному модель нейрона, а также имеет более сложную организацию нейронов, формирующих нейронную сеть. Основной компонентой модели HTM является пространственно-темпоральная память, обладающая способностью прогнозирования, формирование которой происходит за счет преобразования входных данных в распределенное разряженное представление и помещение этого представления в темпоральный контекст. Основополагающим в модели HTM является использование свойства разряженности и «активных дендритов», что и является главным фактором, позволяющим избежать проблем, возникающим при использовании традиционных искусственных нейронных сетей с точечной моделью нейрона [9, 10, 11].

Обзор литературы по теме исследования

Алгоритм HTM имеет две фазы — пространственное группирование, во время которого данные на входе преобразуются в распределённое разряженное представление, и формирование / использование темпоральной памяти, трансформирующее полученное представление в новое, но уже учитывающее не только пространственные закономерности во входных данных, но и темпоральный контекст, т.е. темпоральные закономерности в потоке данных. В [12] авторы предлагают новый алгоритм обучения с подкреплением с использованием только пространственного группировщика модели HTM и анализируют / оценивают его на задаче о многоруком бандите.

Поскольку модель HTM является не только биологически правдоподобной, но и биологически ограниченной моделью, то волне закономерно её использование для моделирования биологических конструктов и механизмов, которые в мозге живых организмов играют важную роль в процессе обучения управляемого своего рода сигналом вознаграждения. В [13] авторы моделируют взаимодействие неокортекса и дофаминергических нейронов для вычисления ошибки предсказания вознаграждения, используемой, например, в обучении на основе временных различий.

Также модель HTM использовалась для проектирования и реализации биологически правдоподобных агентов. В магистерской работе [14] содержатся размышления автора о проектировании такого агента с использованием темпоральной памяти HTM, а в магистерской работе [15] рассмотрено уже не только проектирование, но и практическая реализация агента — неигрового персонажа (non-player character, NPC) в трехмерной среде. Решение построено на использовании модели HTM, алгоритме TD-Lambda и руководствуясь нейробиологическими исследованиями о структурах, принципах и механизмах мозга, связанных с выбором действий и управлением моторикой тела. Биологически обоснованная архитектура агента моделирует взаимодействие неокортекса, базальной ганглии и моторных нейронов, которые получают визуальную информацию от сенсоров агента, обрабатывают её с целью построения модели окружающей среды и вырабатывают стратегию, максимизирующую ожидаемый доход основываясь на сигнале вознаграждения. Концептуально схожий, но иначе реализованный подход предложен в [16] для обучения с подкреплением для решения задачи поиска агентом ресурсов в среде-лабиринте.

Нахождение оптимальной стратегии взаимодействия агента со средой в обучении с подкреплением основано на полученном опыте, но в реальных задачах получение опыта необходимого для обучения может быть очень затратным и трудоёмким, а кроме того, и сам процесс обучения может быть крайне неэффективным без наличия достаточного количества и разнообразия опыта. Помимо использования опыта реального взаимодействия со средой возможно использование модели для имитации окружающей среды и порождения имитационного опыта. Используя планирование можно использовать такой опыт для порождения / улучшения стратегии или более целенаправленного исследования среды. В [17] авторы рассматривают задачу, формализуемую как Марковский Процесс Принятия Решений, в которой целью агента в среде-лабиринте является прибытие в целевое состояние. В процессе взаимодействия со средой агент использует темпоральную память HTM для запоминания переходов между состояниями с помощью выбранных действий, а также состояний, достижение которых привело к получению вознаграждения. Для выполнения поставленной задачи агент пробует определить план

действий с ограниченным радиусом планирования для достижения состояния с возможным вознаграждением и действует согласно ему или же действует произвольно для дальнейшего исследования среды, если определить план не удалось.

Задача распознавания визуальных образов не является типичной задачей для обучения с подкреплением, но легко преобразуется в таковую для случая, когда изображение в каждый момент времени предоставляется не полностью, а лишь частично, и необходимо совершить движение называемое саккадой для получения доступа к другой части изображения. В такой постановке задачи целью является определение стратегии саккад для быстрого и эффективного распознавания изображения. В бакалаврской работе [18] проводится анализ модели биологически ограниченного агента [15] и его адаптация для решения задачи распознавания визуальных образов. Такая же задача рассматривается в [19], но в данной работе во время обучения применяется epsilon-жадный подход и метод Монте-Карло во время которого, темпоральная память HTM запоминает траектории движения и позже их воспроизводит. В свою очередь похожий метод обучения уже использовался ранее и сравнивался с алгоритмом Q-обучения в бакалаврской работе [20], но на значительно более простой искусственной выборке.

В [21] авторы предлагают новый алгоритм обучения с подкреплением на основе темпоральной памяти HTM, которая в данной работе используется как замена Q-функции. Также в работе представлены экспериментальные результаты, отражающие превосходство предложенного алгоритма над алгоритмом Q-обучения при решении задачи CartPole.

В большинстве рассмотренных работ, посвященных использованию модели HTM в обучении с подкреплением, осуществляется попытка использования только элементов данной модели в комбинации с более традиционными подходами для обретения полезных свойств, но также присутствуют работы, авторы которых стремятся создать самостоятельную модель для обучения с подкреплением на основе модели HTM. Последние представляют наибольший интерес, но даже они оставляют вопрос выбора / построения архитектуры открытым, поскольку в настоящее время HTM не является полностью завершенной моделью машинного обучения, а выбор наиболее предпочтительного действия на основе хранящегося в модели сенсорно-моторного опыта малоизучен.

Дискуссия

Одной из основных функциональных особенностей модели HTM является осуществление прогнозирования, формирующее смещение в пользу представлений всех возможных входных образов учитывая темпоральный контекст. Следовательно, представляется возможным два полноценных варианта использования модели HTM в обучении с подключением:

1. использование в качестве "буфера воспроизведения опыта" при обучении устоявшихся методов обучения с подкреплением или для генерации имитационного опыта, который можно использовать для более целенаправленного изучения среды агентом;
2. использование для формирования пространственно-темпоральной памяти агента, способной не только хранить и обобщать сенсорно-моторный опыт агента, но также вырабатывать оптимальную стратегию поведения агента (при наличии дополнительного модуля, осуществляющего оценку состояний памяти и способного формировать дополнительное смещение, способствующего выработке оптимальной стратегии поведения и выбору наиболее предпочтительного действия в каждый момент времени).

Заключение

Применение модели машинного обучения HTM имеет огромный потенциал для развития методов обучения с подкреплением как за счёт решения известных проблем, так и с помощью добавления новых возможностей (прим., непрерывное обучение). Кроме того, развитие направления обучения с подкреплением с применением биологически-правдоподобных принципов и методов может потенциально привести к лучшему пониманию работы мозга и интеллекта как следствия его функционирования.

ЛИТЕРАТУРА

1. Саттон Р.С., Барто Э.Г. Обучение с подкреплением. Монография. — 2020.
2. Liu M. et al. Analyzing the noise robustness of deep neural networks //2018 IEEE Conference on Visual Analytics Science and Technology (VAST). — IEEE, 2018. — С. 60–71.
3. Goodfellow I.J. et al. An empirical investigation of catastrophic forgetting in gradient-based neural networks //arXiv preprint arXiv:1312.6211. — 2013.
4. Thompson N.C. et al. The computational limits of deep learning //arXiv preprint arXiv:2007.05558. — 2020.

5. Vinyals O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning //Nature. — 2019. — Т. 575. — № . 7782. — С. 350–354.
6. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold //Nature. — 2021. — Т. 596. — № . 7873. — С. 583–589.
7. Hawkins J.A thousand brains: A new theory of intelligence. — Hachette UK, 2021.
8. Hawkins, J. et al. Biological and Machine Intelligence. URL: https://numenta.com/resources/biological-and-machine-intelligence/ (дата обращения: 17.08.2022).
9. Grewal K. et al. Going beyond the point neuron: Active dendrites and sparse representations for continual learning //bioRxiv. — 2021.
10. Iyer A. et al. Avoiding Catastrophe: Active Dendrites Enable Multi-Task Learning in Dynamic Environments //Frontiers in neurorobotics. — 2022. — Т. 16.
11. Hunter K., Spracklen L., Ahmad S. Two sparsities are better than one: unlocking the performance benefits of sparse–sparse networks //Neuromorphic Computing and Engineering. — 2022. — Т. 2. — № . 3. — С. 034004.
12. Struye J., Mets K., Latré S. HTMRL: Biologically Plausible Reinforcement Learning with Hierarchical Temporal Memory //arXiv preprint arXiv:2009.08880. — 2020.
13. Choi H. et al. Reward hierarchical temporal memory //The 2012 International Joint Conference on Neural Networks (IJCNN). — IEEE, 2012. — С. 1–7.
14. Otahal M. Architecture of Autonomous Agent Based on Cortical Learning Algorithms: Modeling human brain & mind: дис. — Czech Technical University in Prague, 2013.
15. Sungur A.K. Hierarchical temporal memory based autonomous agent for partially observable video game environments: дис. — Middle East Technical University, 2017.
16. Dzhivelikian E. et al. Intrinsic motivation to learn action-state representation with hierarchical temporal memory //International Conference on Brain Informatics. — Springer, Cham, 2021. — С. 13–24.
17. Kuderov P., Panov A.I. Planning with Hierarchical Temporal Memory for Deterministic Markov Decision Problem //Proceedings of the 13th International Conference on Agents and Artificial Intelligence. — 2021. Т. 2, — С. 1073–1081.
18. Heyder J. Hierarchical Temporal Memory Software Agent: In the light of general artificial intelligence criteria. — 2018.
19. Nugamanov E., Panov A.I. Hierarchical Temporal Memory with Reinforcement Learning //Procedia Computer Science. — 2020. — Т. 169. — С. 123–131.
20. Gomez A.S. Hierarchical Temporal Memory as a reinforcement learning method: дис. — University of Manchester, 2016.
21. Koffi T.Y. et al. A Novel Reinforcement Learning Algorithm Based on Hierarchical Memory //2020 International Conference on Internet of Things and Intelligent Applications (ITIA). — IEEE, 2020. — С. 1–5.

ИТМО