Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 419-431.
URL: http://www.matbio.org/2013/Lakhman_8_419.pdf

УДК:519.7

Механизмы кратковременной памяти в целенаправленном поведении нейросетевых агентов

©2013 Лахман К.В.[1], Бурцев М.С.[2]

Лаборатория нейроинтеллекта и нейроморфных систем
Курчатовский НБИКС-центр, НИЦ "Курчатовский институт"
Россия, Москва, пл. Академика Курчатова, 1

Аннотация. Современные методы машинного обучения не позволяют достичь того уровня адаптивности, который наблюдается в поведении животных в сложных средах с множеством целей. Данное обстоятельство диктует необходимость исследования общих принципов формирования сложных управляющих систем, позволяющих обеспечивать эффективное целенаправленное поведение. Нами была разработана оригинальная модель нейроэволюции агентов в стохастической среде с иерархией целей. В работе проведен анализ эволюционной динамики поведенческих стратегий агентов. Результаты анализа показали, что в процессе эволюции возникают нейросетевые контроллеры, позволяющие агентам хранить информацию в кратковременной памяти за счет различных нейродинамических механизмов и использовать ее в поведении с альтернативными действиями. При исследовании нейронального обеспечения поведения агентов мы обнаружили, что группы нейронов могут отвечать за разные этапы поведения.

Ключевые слова: *целенаправленное поведение, альтернативное поведение, кратковременная память, нейроэволюция, среды с множественными целями, рекуррентные нейронные сети.*

ВВЕДЕНИЕ

Исследование механизмов обучения и поддержания эффективного поведения в стохастических средах со сложной иерархией целей является одним из ключевых направлений в изучении принципов нейрональной обработки информации и разработке систем био-подобного искусственного интеллекта.

Большинство существующих методов машинного обучения не способны обеспечить выработку приемлемого решения в средах с иерархией целей [1]. Одним из наиболее популярных подходов в этой области является обучение с подкреплением (ОП) [2] и его модификации. Традиционно ОП-алгоритмы использовались в проблемных средах с одной целью. Тем не менее на данный момент ведутся исследования в направлении адаптации ОП-подхода к проблемам в сложных многоцелевых средах. Временная абстракция над атомарными действиями агента может быть представлена в качестве *опций* [3]. Опции обобщают примитивные действия и позволяют строить политики поведения на основе более абстрактных блоков. Отдельная опция определяется

[1] klakhman@gmail.com
[2] burtsev.m@gmail.com

множествами возможных начальных и конечных состояний, а также внутренним распределением атомарных действий. В рамках данного подхода *temporal-difference (TD) сети* используются для выработки общей спецификации основных целей обучения [4]. Последней разработкой в направлении абстракции в ОП является архитектура *Horde* [5], спроектированная с целью повышения качества извлечения общих знаний о задаче при взаимодействии со средой. В сущности, *Horde* – это популяция независимых ОП-агентов, каждый из которых специализирован на конкретном аспекте общей проблемы.

Если последовательность действий представлена в виде некоторой опции, тогда эта опция сама может быть использована в качестве атомарной единицы для составления поведенческих стратегий более высокого уровня. Эффективность подобного подхода исследуется в области *иерархического ОП* [1, 6]. Иерархическое ОП предназначено для выработки "поведенческих модулей", которые могут быть скомбинированы и использованы в дальнейшем для решения различных задач.

Основной проблемой на пути эффективного применения подхода, основанного на опциях, в многоцелевых средах является отсутствие понимания того, как опции могут быть составлены самим алгоритмом обучения, а не непосредственно заданы проектировщиком. Другим препятствием является невозможность априорного задания подходящей оценочной функции для алгоритма в целом ряде малоформализованных задач. Для решения последней проблемы на данный момент предпринимаются попытки использовать эволюционные алгоритмы для автоматической генерации значений подкрепления [7].

Существующие альтернативные методы для непосредственной генерации целенаправленных поведенческих последовательностей [8, 9] в большинстве случаев не могут быть эффективно использованы в средах со значительным количеством целей, а также в тех ситуациях, когда агенту необходимо самостоятельно находить цели. Одним из наиболее широко используемых подходов к разработке адаптивных контроллеров в многоцелевых средах на данный момент является нейроэволюция [10–12].

Выработка альтернативного поведения, приводящего к достижению нескольких альтернативных целей, требует реализации немарковского процесса принятия решений и не может быть выполнена с помощью подходов обучения с подкреплением, так как они по своей сути являются марковскими [13]. Другими словами, для обеспечения возможности выбора различных действий в одном и том же состоянии среды агенту необходимо обладать памятью о своих предыдущих действиях. Потенциальные механизмы поддержания кратковременной памяти широко исследованы в области рекуррентных нейронных сетей с точки зрения реверберации сигнала [14] и воспроизведения последовательностей [15]. В соответствии с основным предположением, ключевым механизмом обеспечения кратковременной памяти является реверберация нейрональной активности [16]. Тем не менее нейрональные принципы нелинейной интеграции сенсорной информации и внутреннего состояния сети, которая необходима для целенаправленного поведения, до сих пор не до конца поняты. Вопрос о том, какую роль играет кратковременная память в адаптивном поведении с альтернативными действиями, также недостаточно исследован.

Целью нашего исследования является разработка алгоритма обучения общего назначения для сред с множественными конкурирующими целями. Создание данного алгоритма проводится на основе нейробиологических теорий функциональных систем [17] и селекции нейрональных групп [18]. Мы предполагаем, что поведение агента формируется под действием двух процессов. Первым процессом является эволюция, которая отбирает нейросетевые структуры, обеспечивающие первичный

репертуар врожденного поведения. Для моделирования эволюции нейронных сетей мы применяем эволюционный механизм дупликации с последующей дивергенцией функций дуплицировавших частей [19] – в нашем случае нейронов. Вторым процессом является обучение, которое происходит в течение жизни агента и служит для рекомбинации и расширения врожденного репертуара стратегий.

В данной статье мы представляем *только* нейроэволюционную часть модели. Поведение агента в модели управляется рекуррентной нейронной сетью. Мы моделировали эволюцию агентов в средах с иерархией конкурирующих целей и изучали появление кратковременной памяти, а также ее роль в целенаправленном поведении.

ОПИСАНИЕ МОДЕЛИ

Среда с иерархией целей

В данной статье мы представляем простую абстрактную модель многоцелевой среды. Среда определена как n^{env}-мерный гиперкуб. Каждая размерность гиперкуба может интерпретироваться как некоторый признак внешнего мира. Текущее состояние среды представлено бинарным вектором:

$$\mathbf{E}\left(t\right) = \left(e_1\left(t\right), \cdots, e_{n^{\text{env}}}\left(t\right)\right), \quad e_i\left(t\right) \in \left\{0, 1\right\}. \tag{1}$$

В каждый момент дискретного времени агент может изменить состояние одного бита этого вектора на противоположное. В среде заданы конкурирующие цели, определенные как множества последовательных изменений битов вектора состояния среды:

$$\mathbf{g_i} = \left(\left(n_1, q_1\right), \ldots, \left(n_{k_i}, q_{k_i}\right)\right), \tag{2}$$

где n_j – это номер целевого бита вектора состояния, q_j – это целевое значение бита, k_i – это сложность i-ой цели (количество необходимых для ее достижения действий). Таким образом, цель определена как специфическая последовательность действий. Цели различной сложности заданы в среде, формируя в совокупности разветвленную иерархическую структуру. Предложенное определение позволяет легко генерировать вложенные структуры целей путем добавления действий к уже существующей подцели. Схематичное изображение введенной модели среды и целей в ней приведено на рис. 1. Для оценки сложности конкретной среды мы ввели показатель плотности целей:

$$C_{\text{GD}} = \sum_{i=1}^{N_{\text{A}}} 2^{-k_i'} \left(\frac{1}{2n^{\text{env}}}\right)^{k_i}, \tag{3}$$

где N_{A} – это количество целей в среде, k_i' – это количество уникальных битов вектора состояния, которые должны быть изменены в процессе достижения цели, учитывая тот факт, что один бит может быть изменен несколько раз. Второй множитель в (3) обозначает вероятность совершения конкретной последовательности k_i действий (в рамках данного приближения мы предполагаем, что все действия равновероятны). В то время как первый множитель отражает размер области среды, из которой возможно начать выполнение i-ой цели без совершения лишних действий (каждый бит, значимый для i-ой цели, находится в “обратной позиции”). Данная оценка учитывает не только количество целей, но также и сложность этих целей для агента. Значение, обратное плотности целей, мы будем называть сложностью среды:

$$C_{\text{D}} = \frac{1}{C_{\text{GD}}}. \tag{4}$$

Таким образом, чем сложнее среда для агента, тем больше значение сложности C_{D}.

С каждой целью в среде ассоциирована награда для агента. По определению это значение прямо пропорционально сложности соответствующей цели. Когда агент выполняет последовательность действий, соответствующую некой цели, награда за достижение этой цели добавляется к текущей общей награде агента. После этого награда за данную цель сбрасывается до нуля и линейно восстанавливается до исходного значения за время T_{rec}. Общая награда, накопленная в течении всего времени жизни, влияет на репродуктивный успех агента.

Среда в модели может быть либо детерминированной, либо недетерминированной (стохастической). Во втором случае изменения вектора состояния могут происходить не только из-за действий агента, но и случайно с фиксированной вероятностью.

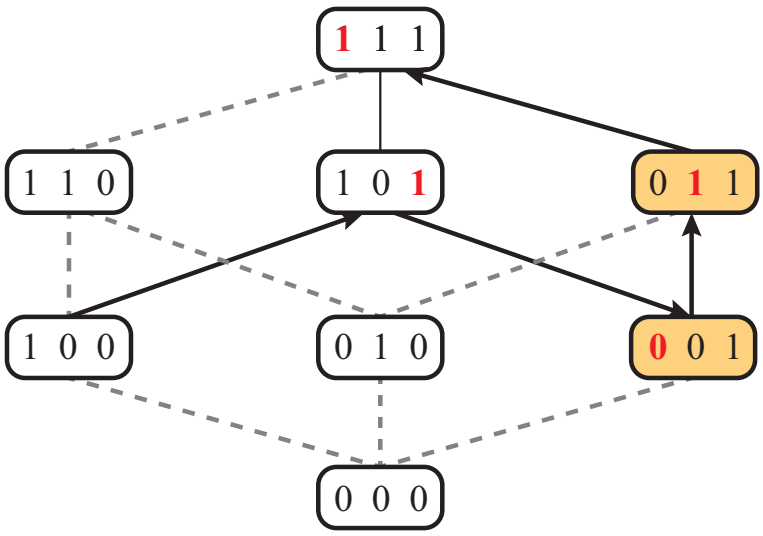

Рис. 1. Среда-гиперкуб размерности 3 (пунктирными линиями показаны потенциально возможные переходы между состояниями среды). На схеме представлен пример поведения агента в среде — последовательные переходы, начиная с состояния $(1,0,0)$, в процессе которого он достигает две цели: $((3,1),(1,0))$ и $((3,1),(1,0),(2,1))$. Желтым цветом выделены моменты, когда агенту начисляется награда.

Поведение агента и эволюционный алгоритм

Поведение агента в среде управляется искусственной нейронной сетью (ИНС) произвольной топологии, которая состоит из входного, выходного и некоторого количества скрытых слоев. Сеть построена на основе нейронов МакКаллока–Питтса с положительной логистической активационной характеристикой. Нейрон может быть либо активен, когда значение на его выходе выше порога, установленного на уровне 0.5 во всех запусках, либо неактивен, если выход ниже порога. Сигнал проходит по синапсу к пост-синаптическому нейрону, только если пре-синаптический нейрон активен. Текущий вектор состояния среды подается непосредственно на нейроны входного слоя сети. В рамках скрытых слоев возможно формирование рекуррентных связей. В любой момент времени пара наиболее активных нейронов выходного слоя кодирует действие агента. Каждая комбинация двух выходных нейронов ответственна за перевод конкретного бита среды в конкретное состояние (0 или 1). Агент может совершать неэффективные действия в тех ситуациях, когда пытается изменить состояние бита на то, в котором это бит уже находится.

Популяция агентов эволюционирует в среде-гиперкубе, описанной в предыдущем разделе. Каждый индивид независимо помещается в среду для оценки своей приспособленности. В течение фиксированного времени агент действует в среде, достигая цели и накапливая награду. Агент не обладает внешней информацией о моментах достижения целей или общем значении накопленной награды. Популяция имеет фиксированный размер, и для каждого агента вероятность стать родителем для агента в следующей популяции прямо пропорциональна накопленной награде:

$$V_{\text{total}} = \sum_{t} \sum_{i \in G(t)} v_i \times \min\left(\frac{t - t_i^{\text{last}}}{T_{\text{rec}}}, 1\right), \tag{5}$$

где $G\,(t)$ — это множество целей, достигнутых на шаге времени t, v_i — это награда за i-ую цель, t_i^{last} — это шаг времени, на котором агент в предыдущий раз достигал i-ую цель.

Для моделирования эволюции управляющих ИНС мы использовали нейроэволюционный алгоритм, основанный на дупликации нейронов. Данный алгоритм похож на широко известный *NEAT* [20], так как он реализует эволюцию топологий нейронных сетей, но делает это более естественным способом за счет дупликации нейронов со всеми связями. Таким образом, мы используем мутацию *дупликация нейрона* вместо мутации *добавление вершины* в алгоритме NEAT (рис. 2). Дуплицировавший нейрон наследует от своего родителя полную структуру входящих и исходящих синаптических связей. Входящие связи дочернего и родительского нейронов сохраняют свои веса, а веса исходящих связей делятся пополам как для “родителя”, так и для “потомка”. Таким образом, пост-синаптические нейроны получают такой же уровень сигнала, как и до процедуры дупликации. Два нейрона, в общем, реализуют прошлую функцию, но позже в эволюции их веса мутируют независимо, что может приводить к функциональной дивергенции. Для оптимизации эволюции путем динамического уменьшения размерности пространства поиска мы также ввели мутацию *удаление связи*, которая выполняется таким же образом, как и мутация *добавления связи*. Мы не использовали оператор скрещивания, так как предварительные результаты моделирования не показали значительной разницы между двумя версиями эволюционного алгоритма (со скрещиванием и без).

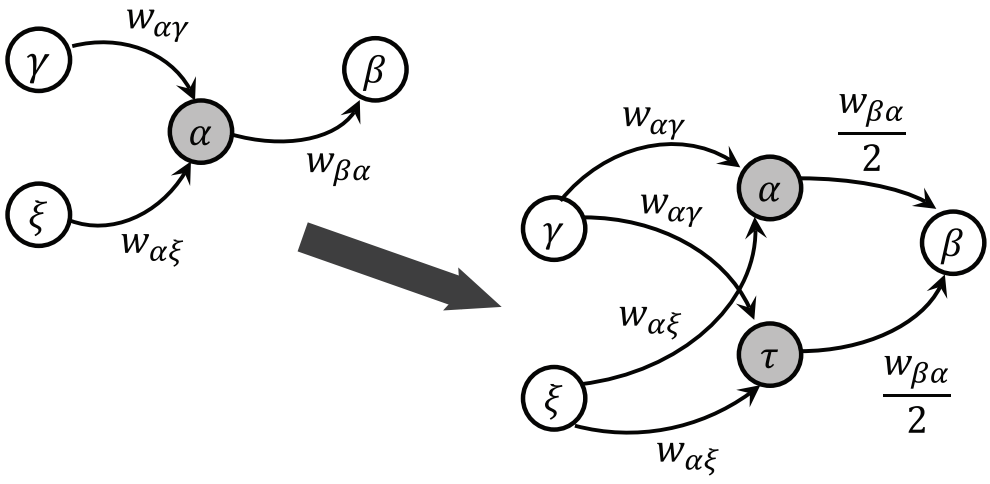

Рис. 2. Схематичное представление мутации *дупликация нейрона*. В процессе дупликации “дочерний” нейрон τ наследует всю структуру синапсов от “родительского” нейрона α. $w_{\alpha\beta}$ обозначает синаптический вес связи между нейронами α и β.

Структуры сетей агентов в первой популяции в начале эволюции состоят из входного и выходного слоев и одного интернейрона, из которого в результате дупликации будут

появляться все будущие интернейроны. Этот исходный интернейрон полностью связан с входным слоем (входящими связями) и с выходным слоем, при этом исходная сеть не содержит прямых связей между входными и выходными нейронами.

Мы использовали следующие параметры модельной среды во всех запусках: размерность среды-гиперкуба $n^{\text{env}} = 8$ бит; продолжительность жизни агента $T_{\text{life}} = 250$ шагов времени; время восстановления награды за достижения цели $T_{\text{rec}} = 30$ шагов времени; вероятность случайного изменения бита вектора состояния $P_{\text{stoch}} = 0.0085$ (для каждого бита); награда, ассоциированная с каждой целью, была определена как $10k_i$ (пропорциональна сложности цели). Эволюционный алгоритм запускался с параметрами: размер популяции $N_{\text{p}} = 250$ агентов; длительность эволюции $T_{\text{ev}} = 5000$ поколений; вероятность мутации синаптического веса $P_{\text{m}} = 0.6$ (для каждого синапса); дисперсия значения мутации синаптического веса $D_{\text{m}} = 0.08$; вероятность добавления синапса $P_{\text{addsyn}} = 0.1$ (для всей сети); вероятность удаления синапса $P_{\text{delsyn}} = 0.05$ (для всей сети); вероятность дупликации нейрона $P_{\text{dup}} = 0.007$ (для всей сети).

Вероятности мутаций были выбраны после обширного предварительного исследования, ставящего цель максимизировать эффективность алгоритма. Продолжительность жизни агента, размер популяции и длительность эволюции были ограничены доступными вычислительными ресурсами. Тем не менее, в подавляющем большинстве запусков средняя накопленная награда по популяции выходила на плато задолго до остановки моделирования. Мы не проводили тщательного анализа зависимости поведения модели от остальных параметров. Не смотря на это, мы ожидаем, что изменение размерности среды не должно качественно поменять поведение модели, в то время как изменения времени восстановления наград и общей схемы определения награды за достижение цели могут существенно повлиять на наблюдаемое поведение. Влияние степени стохастичности среды кратко обсуждается в разделе с экспериментальными результатами.

ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Эволюция поведения

На первом этапе мы исследовали, как плотность целей в среде влияет на эффективность эволюции поведения в детерминированных и стохастических средах. Мы выбрали 18 значений плотности целей, распределенные от 0.005 до 0.15, и сгенерировали 20 случайных сред с различными структурами целей для каждого значения плотности. Для получения достаточной статистики мы запускали эволюционный алгоритм 10 раз для каждой среды в условиях детерминированности и стохастичности. После этого мы определяли лучшую популяции (с точки зрения средней накопленной награды) в каждом запуске, и тестировали поведения каждого агента этой популяции из всех $2^{n^{\text{env}}}$ начальных состояний в *детерминированной* версии среды. Награды, накопленные агентами в течение этого теста, усреднялись для подсчета средней награды для каждого запуска. Таким образом, для каждого значения плотности целей мы получили 200 оценок эффективности алгоритма.

Мы ожидали, что в случае большей плотности целей в среде агент должен достигать большего их количества в сравнении со случаями меньшей плотности. Результаты, представленные на рис. 3, подтверждают это предположение. В качестве базового уровня для абсолютной оценки эффективности эволюции мы приводим результаты моделирования случайных агентов, для которых вероятности совершения всех действий одинаковы.

Агенты, эволюционировавшие в стохастических средах, были значительно

более успешны (рис. 3) с точки зрения средней накопленной награды, чем агенты, эволюционировавшие в детерминированных средах. Необходимо напомнить, что данные, показанные на рис. 3, были получены в ходе тестовой оценки на детерминированных версиях сред. Данные тестирования "детерминированных" и "стохастических" агентов в условиях стохастических сред (не приведены) показывают такие же результаты. Таким образом, стохастичность среды способствует отбору агентов с более устойчивым поведением, позволяющим достигать большее количество целей из разных начальных состояний. Тем не менее, с ростом вероятности случайных изменений в среде можно наблюдать резкий спад эффективности эволюции (данные не приведены), вследствие дестабилизации любого возможного поведения, вызванной слишком частыми изменениями в среде.

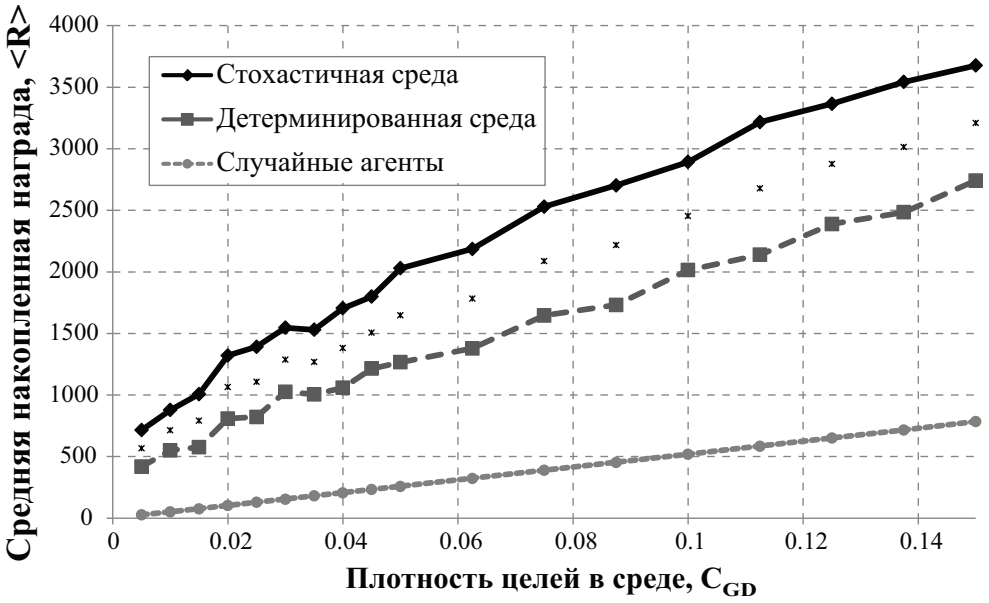

Рис. 3. Зависимость средней накопленной награды от плотности целей в среде (каждый отсчет - это усреднение по 20 средам и 10 запускам эволюции в каждой среде, $*$ – t-критерий, $p = 0.01$ для сравнения средних наград в двух версиях сред).

В дальнейшем мы проводили анализ поведенческих последовательностей агентов в средах с небольшим значением плотности целей ($C_{\mathrm{GD}} = 0.03$) более детально. Обычно выработанное эволюцией поведение состоит из двух фаз: предварительная фаза схождения к какому-то циклу действий, и затем последовательное его выполнение. Для конкретного агента обе этих фазы зависят от начального состояния среды, из которого была запущена жизнь агента. Мы будем называть повторяющуюся последовательность действий или достигнутых целей поведенческой стратегией или циклом.

Результаты моделирования показывают, что резкое увеличение средней накопленной награды в популяции обычно сопровождается появлением и "соревнованием" большого числа новых стратегий. Соответствующие периоды наиболее интересны для понимания принципов эволюции целенаправленного поведения. На рис. 4 мы приводим пример такого периода.

Можно наблюдать "соревнование" между различными поведенческими стратегиями на протяжении ограниченного эволюционного периода (менее 1000 поколений на рис. 4). Примечательно, что новые стратегии появляются как расширение уже выработанных циклов. Расширение может быть реализовано не только как простое добавление новой части к старой (как в случае развития цикла $[86, 28, 29, 30]$

в $[161, 162, 86, 28, 29, 30]$), но и как сложная компиляция двух стратегий (как в образовании цикла $[161, 162, 86, 28, 29, 30, 117, 126, 127, 161, 28, 31, 32]$). В конце концов одна из стратегий выигрывает "соревнование" и доминирует в популяции.

На нижней части рис. 4 можно видеть, что средняя награда все еще возрастает, в то время как максимальная не претерпевает существенных изменений и не появляется новых успешных стратегий. Для объяснения этого мы отслеживали размер области притяжения доминантного поведения сразу после завершения "соревнования" между стратегиями, которое было описано в предыдущем абзаце, и обнаружили, что на протяжении этой стадии эволюция приводит к увеличению количества состояний, из которых успешная стратегия может выполняться. Другой причиной такого эффекта может служить стабилизация стратегии в эволюционном смысле, то есть повышение робастности к негативным мутациям. В общем, и мы обнаружили это во многих других симуляциях, в одном запуске может быть несколько эволюционных "всплесков",

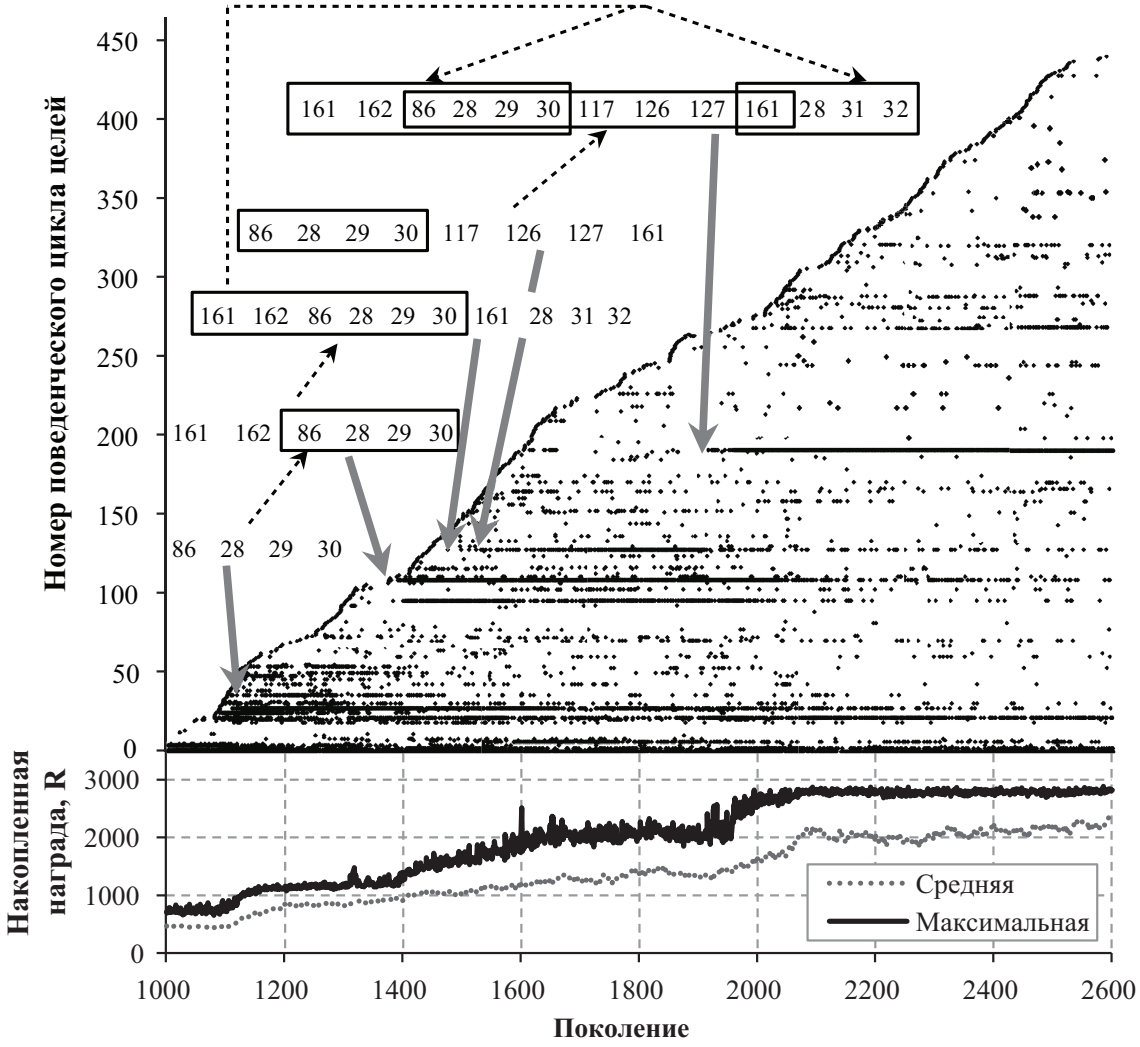

Рис. 4. Появление и отбор поведенческих стратегий в эволюции. На схеме приведены последовательности целей для наиболее успешных стратегий. Каждая точка на верхнем графике соответствует присутствию поведенческой стратегии в соответствующем поколении; светлые стрелки указывают на поколение, в котором соответствующая стратегия начинает доминировать в популяции; черные рамки указывают на включение старого поведенческого цикла в новый. Нижний график отображает динамику максимальной и средней накопленной награды по популяции.

разделенных медленной адаптацией.

Альтернативное поведение и кратковременная память

Анализ поведения, выработанного в результате эволюции, показывает, что агенты приобретают способность хранить информацию в кратковременной памяти. Подтверждение этому следует из того факта, что агенты могут выполнять стратегии, основанные на совершении альтернативных действий из одного и того же состояния среды в разные моменты времени. Диаграмма переходов состояний для типичного полученного поведения с тремя альтернативными действия показана на рис. 5,А. Соответствующее поведение обеспечивается нейронной сетью, состоящей из 30 нейронов, из которых только 15 являются интернейронами, и 611 синаптических связей (фактически данная сеть является полносвязной).

Возможность совершать альтернативные действия означает, что агент “принимает во внимание” предыдущую историю поведения. Данный феномен обеспечивается реверберацией сигнала в нейронной сети за счет рекуррентных связей. Наиболее

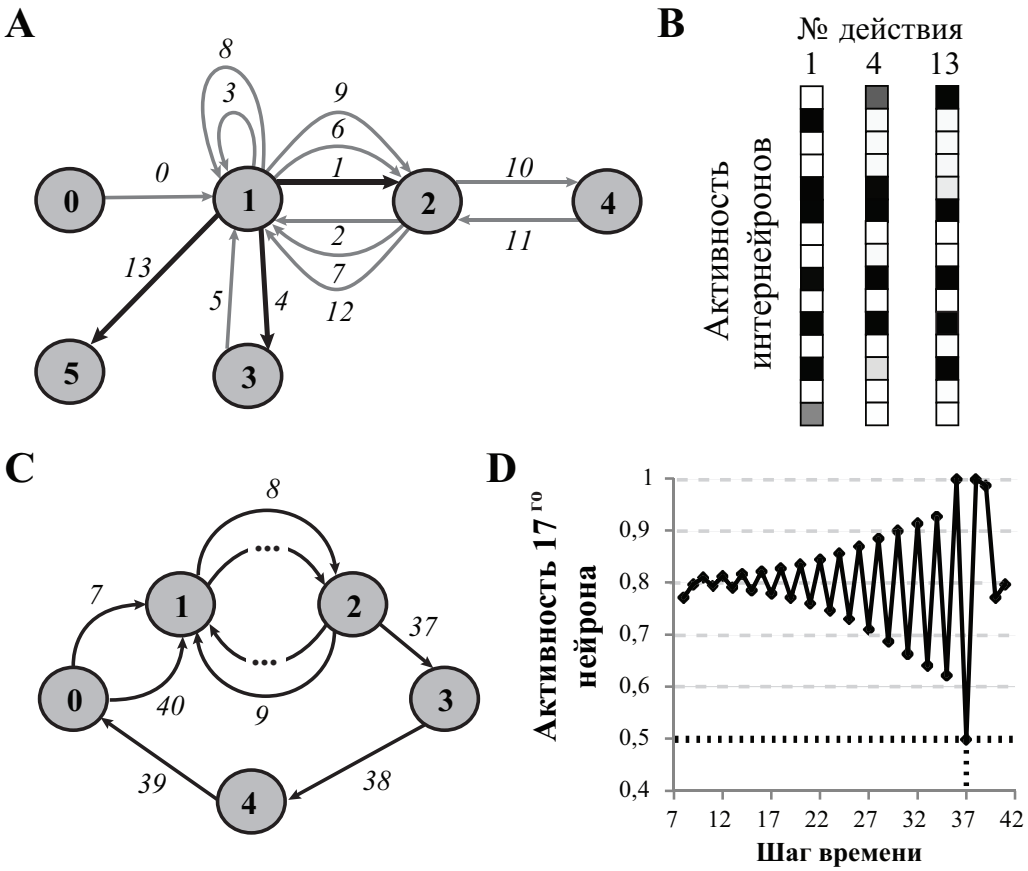

Рис. 5. Нейрональные механизмы альтернативного поведения. **A)** Пример альтернативного поведения (состояния среды обозначены кругами; переходы/действия обозначены стрелками и пронумерованы последовательно, альтернативные действия выделены черным цветом). **B)** Активность интернейронов во время выполнения трех альтернативных действий (черный цвет – максимальный уровень активности соответствующего нейрона; белый – нулевая активность). **C)** Пример альтернативного поведения, основанного на медленном нейродинамическом процессе. **D)** Динамика выхода нейрона, который ответственен за реализацию первой части поведения на рис. C) – последовательные переходы между первым и вторым состояниями.

глубокая кратковременная память, которую нам удалось обнаружить, составляет как минимум 10 предыдущих состояний. Такой вывод был сделан после рассмотрения последовательностей состояний, предшествующих двум альтернативным действиям. Нижняя граница глубины памяти может быть определена как количество переходов до первого отличающегося состояния в этих последовательностях.

Мы проанализировали нейрональную динамику, обеспечивающую выбор альтернативных действий (рис. 5,B). Только небольшая часть нейронов значительно меняет выход и влияет на принятие решений в состояниях, ассоциированных с "выбором", в то время как большинство нейронов сохраняют тот же уровень активности (не смотря на то, что выход нейрона может быть непрерывно распределен от 0 до 1).

Способность использовать кратковременную память делает возможным значительно более сложное поведение. Примером такой поведенческой стратегии может служить последовательное чередование двух циклов действий. Такая стратегия позволяет целям, которые были достигнуты на первом цикле, восстанавливать свою значимость, в то время как агент проходит по второму циклу, что увеличивает его репродуктивный успех.

Примитивное альтернативное поведение в модели может быть также реализовано с помощью медленного нейродинамического процесса (рис. 5,C-D). В этом случае "память", которая необходима для выполнения альтернативных действий, может достигать 30 прошлых состояний. Механизм лежащий в основе этого типа поведения заключается в осцилляторной динамике выхода нейрона. Пока выход нейрона выше порога и он активен, агент выполняет первую часть поведения. В какой-то момент времени выход опускается ниже порога и пост-синаптический нейрон, контролирующий альтернативное действие, перестает "вытормаживаться".

Поведенческие типы нейронов

Для определения вклада отдельных нейронов в поведение агента мы исследовали изменение его эффективности после удаления нейронов. Средняя и максимальная накопленные награды были подсчитаны после удаления интер-нейронов по одному (со всеми входящими и выходящими связями) и запуска агента из всех начальных состояний. Этот анализ позволил выделить три типа нейронов:

1. удаление критически влияет как на среднюю, так и на максимальную награды;
2. удаление не оказывает никакого эффекта на успешность поведения ("бесполезные" нейроны);
3. удаление нейрона не влияет на максимальную награду, но значительно снижает среднюю.

Более того, данные группы нейронов также различаются по вероятности спайка (т.е. нахождения в активном состоянии) — очень высокая для нейронов первого типа, низкая (около нуля) для второго типа и средняя (менее 0.5) для третьего типа.

Более глубокий анализ нейронов, принадлежащих третьему типу, позволил определить, что они ответственны за схождение к поведенческому циклу. В качестве примера можно рассмотреть агента, поведение которого сходится к одной стратегии из всех начальных состояний. После удаления одного нейрона третьего типа поведение разбивается на области притяжения нескольких стратегий, включая исходный цикл (рис. 6). Важным свойством этого множества стратегий является то, что все они могут быть найдены в поведении агентов более ранних поколений. Этот факт позволяет

утверждать, что удаленный нейрон изменял поведение агента таким образом, чтобы он мог выполнять более успешную стратегию в большем количестве ситуаций. Мы также можем утверждать, что нейроны первого типа ответственны за поддержание основной стратегии, так как их активность необходима для выполнения поведенческого цикла.

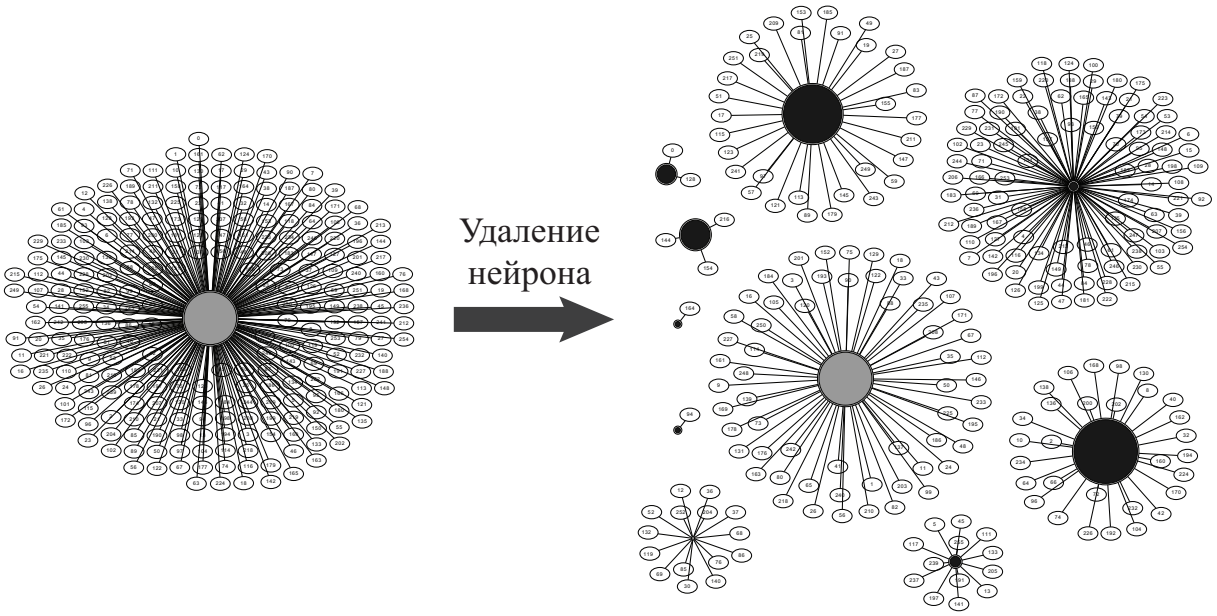

Рис. 6. Слева: поведение агента сходится к одной стратегии (серый круг) из всех начальных состояний (белые овалы). Справа: области притяжения поведенческих стратегий (черные круги) после удаления одного нейрона третьего типа. Размер области притяжения изначальной стратегии значительно уменьшился. Размер кругов пропорционален длине соответствующего цикла действий.

ЗАКЛЮЧЕНИЕ

В рамках представленного исследования мы изучили эволюцию рекуррентных нейронных сетей, которые контролируют поведение агентов в многоцелевой среде. На первом этапе мы проанализировали общую динамику модели. Как мы и ожидали, эффективность агентов положительно коррелирует с плотностью целей в среде. С другой стороны, неожиданно выяснилось, что в условиях стохастических сред эволюционируют агенты с более гибкими и стабильными поведенческими стратегиями, которые способны достигать большего количества целей по сравнению с агентами в той же среде, но без “шума”. Этот феномен крайне необычен для алгоритмов машинного обучения [13], так как в случае недетерминированных сред эффективность обычно ухудшается.

Детальное исследование эволюционной динамики позволило определить, что эволюция агентов состоит из двух фаз: 1) быстрое появление/развитие новых поведенческих стратегий и 2) их распространение по популяции. Образование новых стратегий обычно выполняется с помощью интеграции уже существующих.

На уровне индивидуального поведения мы показали, что эволюции вырабатывают стратегии, основанные на альтернативных действиях. Это происходит благодаря появлению способности оперировать кратковременной памятью и за счет этого выбирать действия, принимая по внимание предыдущую история поведения.

Эволюция выработала два различных механизма реализации альтернативных действий: первый основан на интеграции сенсорной информации и внутреннего

сигнала, который реверберирует по обратным связями; второй основан на медленном нейродинамическом осцилляторном процессе. В действительности оказывается, что нет необходимости привлекать синаптическую пластичность в течение "жизни" агента для того, чтобы наблюдать эффективное использование кратковременной памяти. Важно отметить, что появление способности оперировать кратковременной памятью появляется в нашей модели без каких-либо искусственных предпосылок к этому в строении эволюционного алгоритма.

Поведение животных строится за счет трех основных процессов: эволюция, развитие и обучение. Эволюционная часть крайне важна не только для генерации исходных структур контроллеров агентов, но и для обеспечения системы "оценок" для поведения, приобретенного в процессе самообучения в течение жизни. Таким образом, следующей стадией представленного исследования будет введение алгоритма обучения. Основным механизмом данного алгоритма будет детекция на нейрональном уровне проблем для всего организма, расположенного в среде. Мы планируем решать эту проблему связывания поведенческого и нейронального уровней с помощью концептуальных подходов теории функциональных систем [17] и нейронального дарвинизма [18].

Это исследование частично поддержано Российским Фондом Фундаментальных Исследований (РФФИ) – проект 13-04-01273а. Результаты работы были получены с использованием вычислительных ресурсов МВК НИЦ "Курчатовский институт" (http://computing.kiae.ru).

СПИСОК ЛИТЕРАТУРЫ

1. Botvinick M.M., Niv Y., Barto A.C. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*. 2009. V. 113. P. 262–280.
2. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
3. Sutton R.S., Precup D., Singh S. Etween MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*. 1999. V. 112. P. 181–211.
4. Sutton R.S., Rafols E.J., Koop A. Temporal abstraction in temporal-difference networks. In: *Proceedings of NIPS-18*. MIT Press, 2006. P. 1313–1320.
5. Sutton R.S., Modayil J., Delp M., Degris T., Pilarski P.M., White A., Precup D. Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In: *The 10th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2011. V. 2. P. 761–768.
6. Barto A.G., Mahadevan S. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*. 2003. V. 13. N. 1–2. P. 41–77.
7. Satinder S., Lewis R.L., Barto A. G. Where do rewards come from? In: *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, 2009. P. 2601–2606.
8. Sandamirskaya Y., Schöner G. An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*. 2010. V. 23. N. 10. P. 1164–1179.
9. Komarov M.A., Osipov G.V., Burtsev M.S. Adaptive functional systems: Learning with chaos. *Chaos*. 2010. V. 20. N. 4. P. 045119.
10. Floreano D., Mondada F. Automatic creation of an autonomous agent: genetic evolution of a neural-network driven robot. In: *Proceedings of the third international conference on Simulation of adaptive behavior : from animals to animats 3*. MIT

Press, 1994. P. 421–430.
11. Floreano D., Dürr P., Mattiussi C. Neuroevolution: from architectures to learning. *Evolutionary Intelligence*. 2008. V. 1. P. 47–62.
12. Schrum J., Miikkulainen R. Evolving multimodal networks for multitask games. *IEEE Transactions on Computational Intelligence and AI in Games*. 2012. V. 4. N. 2. P. 94–111.
13. Kaelbling L.P., Littman M.L., Moore A.W. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*. 1996. V. 4. P. 237–285.
14. Hochreiter S., Informatik F.F., Bengio Y., Frasconi P., Schmidhuber J. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: *Field Guide to Dynamical Recurrent Networks*. Eds. Kolen J., Kremer S. IEEE Press, 2001.
15. Botvinick M.M., Plaut D.C. Short-term memory for serial order: A recurrent neural network model. *Psychological Review*. 2006. V. 113. P. 201–233.
16. Grossberg S. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*. 1973. V. 52. N. 3. P. 213–257.
17. Anokhin P. *Biology and Neurophysiology of the Conditioned Reflex and Its Role in Adaptive Behavior*. Pergamon Press, 1974.
18. Edelman G. *Neural Darwinism: The Theory of Neuronal Group Selection*. Basic Books, 1987.
19. Taylor J.S., Raes J. Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics*. 2004. V. 38. P. 615–643.
20. Stanley K.O., Miikkulainen R. Evolving neural networks through augmenting topologies. *Evolutionary Computation*. 2002. V. 10. N. 2. P. 99–127.

Материал поступил в редакцию 21.05.2013, опубликован 26.08.2013.