ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. НАУКИ О ЖИЗНИ, 2022, том 502, с. 48–53

УДК 577.38; 004.81

НЕЙРОСЕТЕВОЕ ДЕКОДИРОВАНИЕ ИНФОРМАЦИИ О ВНЕШНЕМ СТИМУЛЕ ПО ПАТТЕРНУ НЕЙРОННОЙ АКТИВНОСТИ РЕКУРРЕНТНОЙ НЕЙРОННОЙ СЕТИ

© 2022 г. С. И. Барцев[1,2,*], П. М. Батурина[2], Г. М. Маркова[2]

Представлено академиком РАН А.Г. Дегерменджи
Поступило 20.07.2021 г.
После доработки 31.10.2021 г.
Принято к публикации 31.10.2021 г.

Работа посвящена оценке возможности восстановления информации, полученной искусственной нейронной сетью, по паттерну нейронной активности. В ходе теста отложенного сравнения с образцом, усложненного варьирующей длительностью паузы между получением стимулов, простая рекуррентная нейронная сеть формирует динамические паттерны возбуждения, с помощью которых обеспечивается хранение информации о полученном стимуле. Информация, хранящаяся в данных паттернах, может быть использована нейронной сетью в любой момент времени из заданного интервала (3–6 тактов), следовательно, может быть выделено инвариантное представление полученного стимула. Для выделения данных представлений предложен метод нейросетевого декодирования со 100% эффективностью идентификации полученных стимулов. Метод позволяет выделить минимальное подмножество нейронов, в паттерне возбуждения которых содержится исчерпывающая информация о стимуле, полученном нейронной сетью.

Ключевые слова: тест отложенного сравнения с образцом, нейронная активность, динамическое кодирование, классификация паттернов нейронной активности

DOI: 10.31857/S2686738922010048

Возможность реконструкции содержания информации, которая обрабатывается мозгом, по динамическим паттернам нейронной активности составляет ключевую задачу концепции нейронных коррелятов сознания (НКС) [1]. Согласно современным представлениям и нейрофизиологическим данным [2–4], кодирование информации, релевантной решаемой организмом задаче, информации в рабочей памяти является очень динамичным, поскольку представлено широко варьирующими паттернами нейронной активности.

[1] *Институт биофизики Сибирского отделения Российской академии наук – обособленное подразделение Федерального исследовательского центра "Красноярский научный центр Сибирского отделения Российской академии наук", Красноярск, Россия*

[2] *Федеральное государственное автономное образовательное учреждение высшего образования "Сибирский федеральный университет", Красноярск, Россия*

**e-mail: BartsevSI@ibp.ru*

Известно, что кодировка информации о внешнем стимуле, полученном рекуррентной искусственной нейронной сетью (РНС) в ходе теста отложенного сравнения с образцом (ОСО), также является динамичной [5]. Поскольку в данной работе длительность паузы между получением двух стимулов была фиксирована, то единственным требованием являлось достижение нужной точки в пространстве нейронной активности РНС к моменту поступления второго стимула [2]. Если же продолжительность паузы между первым и вторым стимулами выбирается случайно из заданного интервала, то вопрос о том, как в РНС хранится информация, доступная для использования в любой момент паузы, встает гораздо острее.

Цель настоящей работы – оценить возможность идентификации стимула, полученного РНС, по паттерну нейронной активности в период, когда сеть хранит информацию о стимуле в рабочей памяти в состоянии готовности к ответу. Задача, которая требует от РНС хранения информации в паттерне нейронной активности в течение некоторого времени, – тест ОСО.

Использовались простые РНС, имеющие два входа, 25 внутренних нейронов. Данное число нейронов было определено эмпирически как минимально необходимое для решения задачи. В отличие от 20-нейронных, 25-нейронные обучались задаче успешно (до величины ошибки порядка 10^{-5}). Проверка показала, что 30-нейронные РНС легче обучаются прохождению теста ОСО. Однако 25-нейронные РНС удобнее для анализа, а ориентация на использование минимального набора нейронов лежит в русле подхода нейронных коррелятов.

Начальные значения весовых коэффициентов выбирались случайным образом из диапазона (–0.025; 0.025). Отклик РНС $y_o^{(t)}$ в момент времени t регистрировался на двух выходных нейронах:

$$\begin{aligned} y_h^{(t)} &= f_h(W_h \cdot y_h^{(t-1)} + W_i \cdot x^{(t)}) \\ y_o^{(t)} &= f_o(W_o \cdot y_h^{(t)}), \end{aligned} \quad (1)$$

где W_h, W_i, W_o – матрицы весовых коэффициентов внутренних нейронов, входов и выходных нейронов соответственно; $x^{(t)}$ – вектор входных сигналов в момент времени t; векторы $y_h^{(t)}$ и $y_h^{(t-1)}$ описывали уровни возбуждения внутренних нейронов в моменты времени t и t-1. Функции $f_h(.)$ и $f_o(.)$ – функции активации внутренних и выходных нейронов соответственно. Для простоты в уравнениях опущены смещения нейронов.

Активационная функция внутренних нейронов имела сигмоидный вид (2a). Кусочно-линейная функция активации (2b) выходных нейронов использовалась для получения точного выходного сигнала 0/1.

$$\begin{aligned} a)\quad & f_h(x) = \frac{1}{2}\left(\frac{x}{a + |x|} + 1\right), \\ b)\quad & f_o(x) = \begin{cases} 0, & if \quad x \le 0, \\ b \cdot x, & if \quad x > 0 \ \& \ x < 1, \\ 1, & if \quad x \ge 1. \end{cases} \end{aligned} \quad (2)$$

Параметры функций активации (2) имели значения $a = 0.1$, $b = 1$, подобранные эмпирически для наиболее быстрого обучения РНС. Шаг модификации синапсов задавался равным 10^{-3}.

Обучение РНС проводилось с помощью алгоритма обратного распространения ошибки. Поскольку структура обученной сети не зависит от алгоритма обучения [6, 7], то его конкретный вид не имеет значения для анализа ее функционирования. Использовалась квадратичная функция потерь:

$$C = \frac{1}{2}\sum_{i=1}^{N}(y_i^{(t)} - \delta_i^{(t)})^2, \quad (3)$$

где $y_i^{(t)}$ и $\delta_i^{(t)}$ – имеющийся и требуемый сигналы на i-м выходном нейроне РНС в момент времени t, N – количество выходных нейронов.

На вход РНС мог поступить один из трех стимулов: A – (01), B – (10) и C – (11). Учитывая, что (00) – отсутствие стимула, задействован полный набор возможных стимулов для данного количества входов. Тест ОСО проводился следующим образом. На вход РНС в случайные моменты времени поступал один из стимулов (A, B, C), также выбранный случайно. Стимул предъявлялся РНС в течение одного такта. Затем следовала пауза, когда на вход РНС ничего не поступало, продолжительностью от 3 до 6 тактов. Длительность паузы в заданном интервале тоже определялась случайным образом. Далее однократно предъявлялся второй стимул, также выбранный случайно. На третьем такте после получения второго стимула РНС выдавала отклик (10) или (01), в зависимости от того, были полученные стимулы одинаковы или различны. Затем после периода релаксации длительностью не менее 9 тактов начинался следующий цикл обучения. Тем самым учебная выборка постоянно генерировалась в процессе обучения, что позволило пренебречь вероятностью того, что значительные фрагменты входного потока квазислучайных событий будут повторяться.

Обученные РНС проходили тест ОСО в режиме функционирования в таком же непрерывном квазислучайном потоке событий, обеспечивающем невоспроизводимость выборки.

Для идентификации стимула, полученного РНС, использовались данные по нейронной активности сети во время паузы между поступлениями первого и второго стимула. В данный период, с 3 по 6 такт после получения первого стимула, РНС хранила информацию об этом стимуле в форме паттерна нейронной активности. Динамика нейронной активности показала большую вариативность паттернов возбуждений в интервале между поступлениями стимулов и отсутствие явных признаков статичности. Для идентификации стимула в режиме функционирования между поступлением стимулов устанавливалась пауза максимальной длительности (6 тактов).

В качестве контроля, для выделения динамического инварианта нейронной активности РНС в период хранения информации о полученном стимуле использовался метод центроидов [8]. Активность нейронов РНС в каждый момент време-

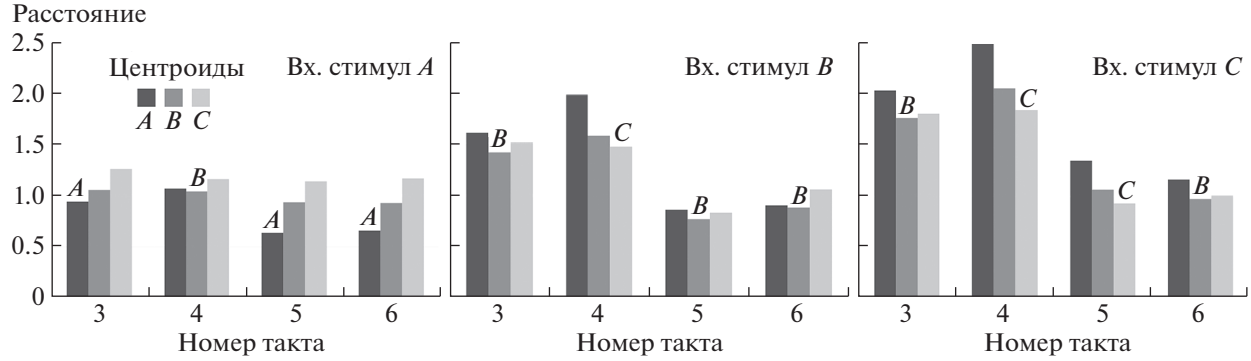

Рис. 1. Применение метода центроидов для идентификации стимулов, полученных РНС. Столбцы – расстояния от точек в пространстве нейронной активности до каждого из трех центроидов. В верхней части диаграмм указано, какой входной стимул получила РНС. Буквы над каждой группой столбцов показывают, какой стимул был идентифицирован на данном такте по минимальному расстоянию до одного из центроидов.

Рис. 2. Схема эксперимента по обучению ДН.

ни представлялась как точка в многомерном пространстве нейронной активности размерности R^N, где N – количество нейронов РНС. Путем усреднения значений активности на четырех последовательных тактах в период хранения информации о первом полученном стимуле вычислялось наиболее вероятное расположение точек, соответствующих каждому из трех возможных стимулов:

$$\overline{E}_t^{\alpha} = \frac{1}{4}\sum_{t=3}^{6} E_{t,n}^{\alpha}, \tag{4}$$

где $E_{t,n}^{\alpha}$ – значение активности на n-м нейроне РНС в момент времени t после получения стимула α(A, B, C). В данном случае использовались значения активности из обучающей выборки. Полученные таким образом три точки являлись A, B и C-центроидами соответственно. Для идентификации стимула вычислялся квадрат евклидового расстояния от каждого центроида до точек из тестовой выборки:

$$D_t^{\alpha} = \sum_{t=1}^{N} (\overline{E}_n^{\alpha} - E_{t,n})^2, \tag{5}$$

где $E_{t,n}$ – значение активности на n-м нейроне РНС в момент времени t, взятое из тестовой выборки.

Идентификация стимула, информация о котором хранилась в нейронной активности РНС, проводилась в соответствии с тем, какой из трех центроидов оказался наиболее близким к рассматриваемой точке:

$$Stim_t = \begin{cases} A, & if \quad \min(D_t^A, D_t^B, D_t^C) = D_t^A, \\ B, & if \quad \min(D_t^A, D_t^B, D_t^C) = D_t^B, \\ C, & if \quad \min(D_t^A, D_t^B, D_t^C) = D_t^C. \end{cases} \tag{6}$$

Полученный вид стимула $Stim_t$ сопоставлялся с реальным, заведомо известным для каждого на-

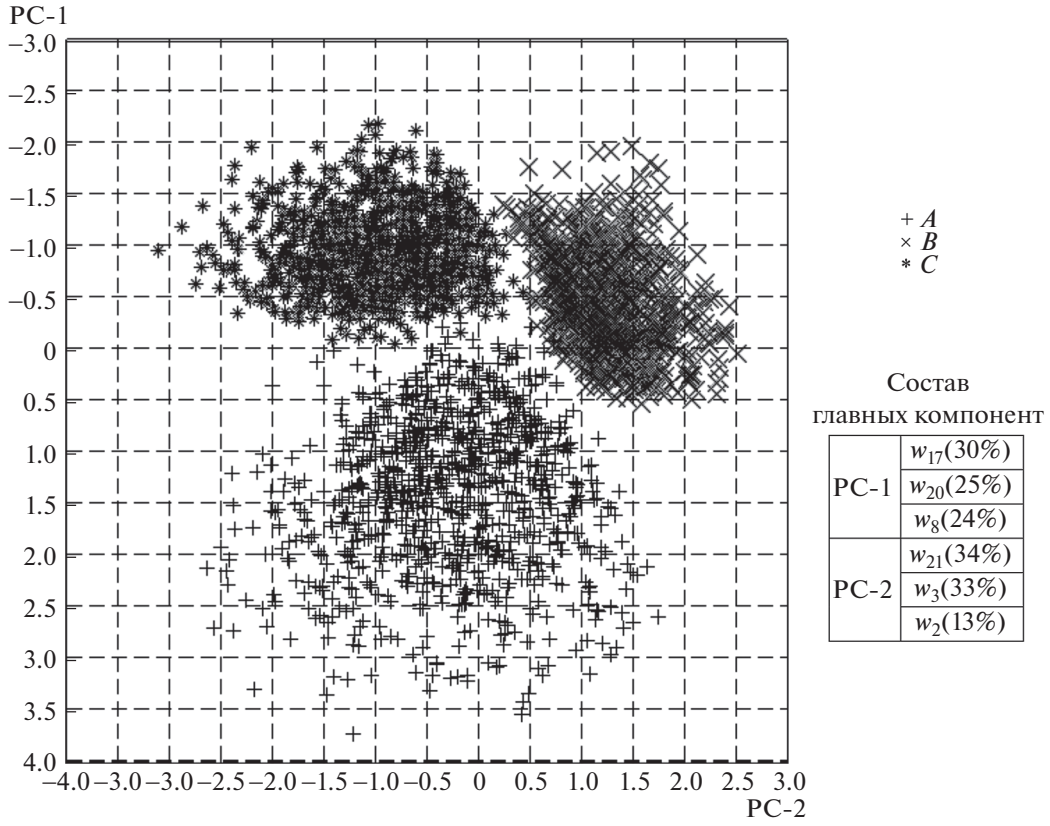

Рис. 3. Конфигурации инвариантов, соответствующие условиям распознавания стимулов, после обработки методом главных компонент.

бора тестовых данных, и на основании этого оценивалась точность идентификации.

Хотя в ряде случаев метод центроидов позволил верно идентифицировать стимул по паттерну нейронной активности, его эффективность не превышала 80%, что объясняется высокой вариабельностью сигнала (рис. 1). В разные моменты времени в период хранения стимула точность идентификации данным методом менялась.

На следующем этапе работы инвариантное представление, хранящее информацию о полученном стимуле в течение четырех тактов, выделялось с помощью дополнительной нейронной сети – нейросетевого декодера (ДН) (рис. 2).

В качестве ДН использовалась однослойная нейронная сеть из трех нейронов с линейной характеристикой (2b). Каждый нейрон имел модифицируемый синапс с каждым из входов, число которых равнялось количеству нейронов РНС. ДН выдавал единицу на одном из трех нейронов, соответствующем приписанному стимулу, и нули на остальных. Для обучения использовался алгоритм обратного распространения ошибки. Функция потерь ДН также являлась квадратичной (3).

В качестве входных данных для ДН использовалась нейронная активность конкретной РНС, проходивших тест ОСО. Нейронная активность РНС записывалась построчно. Строка состояла из значений активности каждого из 25 нейронов РНС в данный момент времени (3, 4, 5 или 6 такт после получения первого стимула) и ей ставился в соответствие стимул, информация о котором хранилась в РНС в это время. Для всех обученных РНС было записано по 72 строки, которые распределялись между обучающей и тестовой выборкой случайным образом.

Для каждой обученной РНС требовалось обучать отдельный ДН, что указывает на индивидуальность внутренней репрезентации стимулов у нейронной сети. Обученные ДН декодировали репрезентируемые РНС стимулы с точностью 100%. Далее проводилась редукция структуры ДН: синапсы с наименьшими по модулю значениями последовательно приравнивались нулю, и на каждом шаге ДН доучивался до исходного качества функционирования. Процедура прекращалась, когда качество функционирования ухудшалось. В результате для каждой из обученных РНС была выделена группа из 6–7 нейронов, по значениям активности которых обеспечивалось декодирование поступивших стимулов.

Инварианты нейронной активности, соответствующие условиям распознавания каждого из трех стимулов, могут быть локализованы в многомерном пространстве динамических паттернов. Для этого на входы обученного ДН подавались наборы случайно сгенерированных чисел, имитирующих значения активности нейронов РНС, выделенных для декодирования. Те наборы случайных чисел, которые ДН классифицировал как соответствующие какому-либо из стимулов, отбирались и рассматривались как точки в пространстве нейронной активности, представляющие кодировку конкретного стимула.

В качестве примера рассмотрим структуру конкретного ДН. Этот ДН выделил как значимые и достаточные 6 нейронов исходной сети под номерами 2, 3, 8, 17, 20, 21.

При прохождении входных данных через ДН результат вычисления имеет общий вид $w_2^\alpha x_2 + w_3^\alpha x_3 + w_8^\alpha x_8 + \quad w_{17}^\alpha x_{17} + w_{20}^\alpha x_{20} + w_{21}^\alpha x_{21} = s^\alpha$, где w_i^α – ненулевой весовой коэффициент ДН после редукции структуры, связывающий i-й вход ДН с нейроном, ответственным за распознавание стимула α, x_i – значение активности i-го нейрона исходной РНС, $\alpha = A, B, C$. При условии, что $s^A \geq 1$, $s^B \leq 0$, $s^C \leq 0$, ДН идентифицирует полученный набор данных как соответствующий хранению информации о стимуле A. Аналогично идентифицируются наборы для стимулов B и C.

Линейные многочлены, которые позволяют выявить инварианты для РНС, рассматриваемой в качестве примера, имеют вид:

$$0.158x_2 - 0.142x_3 + 0.461x_8 +$$
$$+ 0.595x_{17} - 0.582x_{20} + 0.245x_{21} = s^A,$$
$$-0.244x_2 + 0.545x_3 - 0.079x_8 -$$
$$- 0.4x_{17} + 0.072x_{20} - 0.509x_{21} = s^B,$$
$$-0.243x_2 - 0.292x_3 - 0.827x_8 -$$
$$- 0.349x_{17} + 0.328x_{20} + 0.887x_{21} = s^C.$$

Метод главных компонент показал, что точки в пространстве нейронной активности, соответствующие инвариантному свойству распознавания, разделяются на три компактных кластера (пример на рис. 3). При этом для распознавания достаточно отображения значений активности нейронов на двумерную плоскость первой и второй главных компонент. Значения активности одного из шести нейронов РНС оказывают незначительный вклад в формирование второй главной компоненты. Этот факт позволяет предположить, что для распознавания стимула для данной РНС достаточно значений активности пяти нейронов, но такой вариант не был обнаружен в ходе редукции структуры ДН. Следовательно, для окончательной минимизации представлений инварианта кодирования стимулов в нейронной сети может быть полезен метод главных компонент или аналогичные ему.

На основе полученных результатов можно заключить, что, несмотря на динамический характер нейронной активности, обеспечивающей хранение информации о полученных стимулах, по паттернам нейронной активности можно идентифицировать конкретный вид стимула. Предложенный в работе метод нейросетевого декодирования позволяет выделить динамический инвариант нейронной активности, репрезентирующий данный стимул, со 100% точностью. Кроме того, данный подход предполагает выявление минимального набора нейронов и, соответственно, минимальной нейронной активности, необходимых для решения поставленной перед нейронной сетью задачи, и тем самым данный подход лежит в русле концепции нейронных корелятов [1].

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследование выполнено при финансовой поддержке РФФИ, Правительства Красноярского края и Красноярского краевого фонда науки в рамках научного проекта № 20-41-240003.

СПИСОК ЛИТЕРАТУРЫ

1. *Crick F., Koch C.* A framework for consciousness // Nat. Neurosci. 2003. V. 6. № 2. P. 119–126.
2. *Meyers E.M.* Dynamic population coding and its relationship to working memory // J. Neurophysiol. 2018. V. 120. № 5. P. 2260–2268.
3. *Barak O., Tsodyks M., Romo R.* Neuronal population coding of parametric working memory // J. Neurosci. 2010. V. 30. № 28. P. 9424–9430.
4. *Stokes M.G., Kusunoki M., Sigala N., et al.* Dynamic coding for cognitive control in prefrontal cortex // Neuron. 2013. V. 78. № 2. P. 364–375.
5. *Miconi T.* Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks // Elife. 2017. V. 6. P. e20899.
6. *Барцев С.И., Барцева О.Д.* Симметрии структуры и эквифинальность эволюционных исходов в простых нейросетевых моделях // ДАН. 2002. Т. 386, № 1. С. 114–117.
7. *Барцев С.И., Барцева О.Д.* Функционально-инвариантный подход к проблеме уникальности биологических систем: простая нейросетевая модель // ДАН. 2005. Т. 405. № 4. С. 1–4.
8. *Crowe D.A., Averbeck B.B., Chafee M.V.* Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex // J. Neurosci. 2010. V. 30. № 35. P. 11640–11653.

NEURAL NETWORK-BASED DECODING OF INFORMATION ABOUT INPUT STIMULUS BY NEURAL ACTIVITY OF RECURRENT NEURAL NETWORK

S. I. Bartsev[a,b,#], P. M. Baturina[b], and G. M. Markova[b]

[a] *Biophysics Institute of the Siberian Branch of the RAS – Division of Federal Research Center "Krasnoyarsk Scientific Center of the Siberian Branch of the RAS", Krasnoyarsk, Russian Federation*

[b] *Siberian Federal University, Krasnoyarsk, Russian Federation*

[#] *e-mail: BartsevSI@ibp.ru*

Presented by Academician of the RAS A.G. Degermendzhi

We assess the possibility to recover information received by artificial neural network via inspecting neural activity pattern. Simple recurrent neural network forms dynamic excitation patterns for storing information about input stimulus during the delayed match to sample test with variable duration of pause between received stimuli. Information stored in these patterns can be used by neural network at any moment of time within specified interval (3–6 clock cycles), therefore it is possible to detect invariant representation of received stimulus. To identify these representations, we suggest the neural network-based decoding method that shows 100% efficiency of received stimuli recognition. This method allows to identify the minimum subset of neurons, the excitation pattern of which contains comprehensive information about the stimulus received by neural network.

Keywords: delayed match to sample test, neural activity, dynamic coding, classification of neural activity patterns