ISSN 0012-4966, Doklady Biological Sciences, 2022, Vol. 502, pp. 1–5. © The Author(s), 2022. This article is an open access publication.
Russian Text © The Author(s), 2022, published in Doklady Rossiiskoi Akademii Nauk. Nauki o Zhizni, 2022, Vol. 502, pp. 48–53.

Neural Network-Based Decoding Input Stimulus Data Based on Recurrent Neural Network Neural Activity Pattern

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

Presented by Academician A. G. Degermendzhi

Received July 20, 2021; revised October 31, 2021; accepted October 31, 2021

Abstract—The paper reports the assessment of the possibility to recover information obtained using an artificial neural network via inspecting neural activity patterns. A simple recurrent neural network forms dynamic excitation patterns for storing data on input stimulus in the course of the advanced delayed match to sample test with varying duration of pause between the received stimuli. Information stored in these patterns can be used by the neural network at any moment within the specified interval (three to six clock cycles), whereby it appears possible to detect invariant representation of received stimulus. To identify these representations, the neural network-based decoding method that shows 100% efficiency of received stimuli recognition has been suggested. This method allows for identification the minimum subset of neurons, the excitation pattern of which contains comprehensive information about the stimulus received by the neural network.

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

DOI: 10.1134/S001249662201001X

The possibility of reconstructing the content of data processed by the brain from the dynamic patterns of neural activity is the key task in the Neural Correlates of Consciousness (NCC) concept [1]. According to the current views and based on neurophysiological data [2–4], encoding task-relevant information in working memory is very dynamic since it is represented by widely varying patterns of neuronal activity.

It is known that coding the information about the external stimulus received by the recurrent artificial neural network (RNN) in the course of a delayed matching-to-sample (DMS) test is also a dynamic process [5]. Because in the present study the pause length between the acquisition of two stimuli was fixed, the only requirement was to reach the desired point in the RNN neural activity space by the time the second stimulus arrives [2]. If the pause length between the first and second stimuli is chosen randomly from a given interval, then the problem of how the information available for use at any moment during the pause can be stored in the RNN becomes much more challenging.

[a] *Institute of Biophysics, Siberian Branch, Russian Academy of Sciences, Krasnoyarsk, 660036 Russia*

[b] *Siberian Federal University, Krasnoyarsk, 660041 Russia*

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

The aim of the present work was to assess the possibility of identifying the stimulus received by RNN based on the neural activity pattern in the period when the network stores data about the stimulus in its working memory in the ready-to-respond state. The task which requires RNN storing information in the form of a neuronal activity pattern for a certain period of time is a DMS test.

Simple RNNs with two inputs and 25 internal neurons were used. The neuron number was determined empirically as the minimum number required to complete the task. In contrast to 20-neuron RNNs, the 25-neuron ones could be successfully trained (to the error as low as about 10^{-5}). Verification showed that 30-neuron RNNs were easier to train to pass the DMS test. However, 25-neuron RNNs are more convenient in terms of analysis, while focus on using the minimum possible neuron set is in line with the neural correlates approach.

The initial values for weight coefficients were chosen randomly from within the (–0.025; 0.025) range. RNN response $y_o^{(t)}$ at the time point t was recorded at the two output neurons:

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

where W_h, W_i, W_o are the weight coefficient matrices for the internal neurons, input, and output neurons, respectively; $x^{(t)}$ is the vector of input signals at the

time point t; $y_h^{(t)}$ and $y_h^{(t-1)}$ are vectors describing internal neuron excitation levels at the time points t and $t-1$. The $f_h(.)$ and $f_o(.)$ functions are activation functions for internal and output neurons, respectively. For the sake of simplicity, neuron displacements are omitted from the equations.

The activation function for internal neurons was sigmoidal (2a). The piecewise linear activation function (2b) for the output neurons was used to obtain an accurate 0/1 output signal.

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

The parameters of the activation functions (2) had the values $a = 0.1$ and $b = 1$, which were selected empirically for the fastest RNN training. The synapse modification step was set equal to 10^{-3}.

The RNN was trained using the error backpropagation algorithm. Since the structure of the trained network does not depend on the training algorithm [6, 7], its specific form is not important for the analysis of its functioning. The quadratic loss function was used:

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

where $y_i^{(t)}$ and $\delta_i^{(t)}$ the present and the required signals at the ith RNN output neuron at the time point t and N is the output neuron number.

RNN could receive one of the three input stimuli: A, (01); B, (10); and C, (11). Given that (00) is the absence of any stimuli, the full set of possible stimuli for a given number of inputs was used. The DMS test was conducted as follows. One of the randomly chosen stimuli (A, B, C) arrived to the RNN input at random time points. The stimulus was presented to the RNN as a single beat. Then a pause 3 to 6 beats long followed during which no signal arrived at the RNN input. The length of the pause was also determined randomly. Then a second stimulus was presented once, also chosen at random. The third beat after the second stimulus was the RNN response (10) or (01) which depended on whether the two acquired stimuli were the same or different. Then, after a relaxation period of at least 9 beats, the next training cycle began. Thus, the training sample was continuously generated during the training process, which allowed us to neglect the probability that significant fragments of the input stream of quasi-random events would be repeated.

The trained RNNs were subjected to the DMS test in the function mode in the same continuous quasi-random event stream ensuring non-reproducibility of the test signal sample.

To identify the stimulus received by the RNN, we used data on the network's neural activity during the pause between the first and second stimuli. During this period from the 3 to 6 beats after receiving the first stimulus, the RNN stored the information about this stimulus in the form of a neural activity pattern. The neural activity dynamics revealed high variability of excitation patterns in the interval between the stimuli with no clear signs of statics. To identify the stimulus in the function mode, the pause between the stimuli was set to be of maximum length (6 beats).

As a control, the centroid method [8] was used to identify the dynamic invariant of RNN neural activity during the information storage period. The activity of the RNN neurons at each moment of time was represented as a point in the multidimensional neuronal activity space with R^N dimensions, where N is the number of neurons in the RNN. By averaging the activity at four consecutive beats during the information storage period, the most likely location of the points corresponding to each of the three possible stimuli was calculated:

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

where $E_{t,n}^{\alpha}$ is the activity at the RNN neuron n at the time point t after receiving the stimulus $\alpha(A, B, C)$. In this case, activity values from the training sample were used. The three points thus obtained were the A, B, and C centroids, respectively. To identify the stimulus, squared Euclidean distances from each centroid to the points from the test sample were calculated:

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

where $E_{t,n}$ is the activity at the RNN neuron n at the time point t obtained from the test sample.

Identification of the stimulus, information about which was stored encoded in the RNN neural activity was carried out according to which of the three centroids was closest to the point in question:

$$\mathrm{Stim}_t = \begin{cases} A, & \text{if } \min(D_t^A, D_t^B, D_t^C) = D_t^A, \\ B, & \text{if } \min(D_t^A, D_t^B, D_t^C) = D_t^B, \\ C, & \text{if } \min(D_t^A, D_t^B, D_t^C) = D_t^C. \end{cases} \quad (6)$$

The resulting stimulus type $Stim_t$ was compared with the real one available for each test data set and based on this the accuracy of identification was evaluated.

Although in a number of cases the centroid method allowed correct identification of stimulus based on the neural activity pattern, its efficiency did not exceed 80%, which could be explained by the high signal variability (Fig. 1). At different points within the stimulus

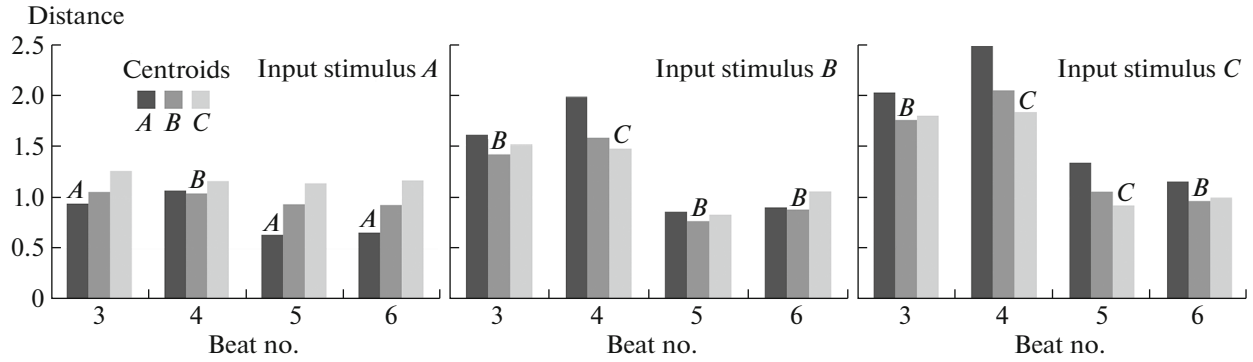

Fig. 1. Using the centroid method to identify RNN-derived stimuli. Columns, distances from the points in the neuronal activity space to each of the three centroids. The type of the input stimulus received by the RNN is indicated at the top of the diagram. Letters above each group of columns indicate the type of the stimulus identified in the corresponding beat based on the minimum distance to the centroids.

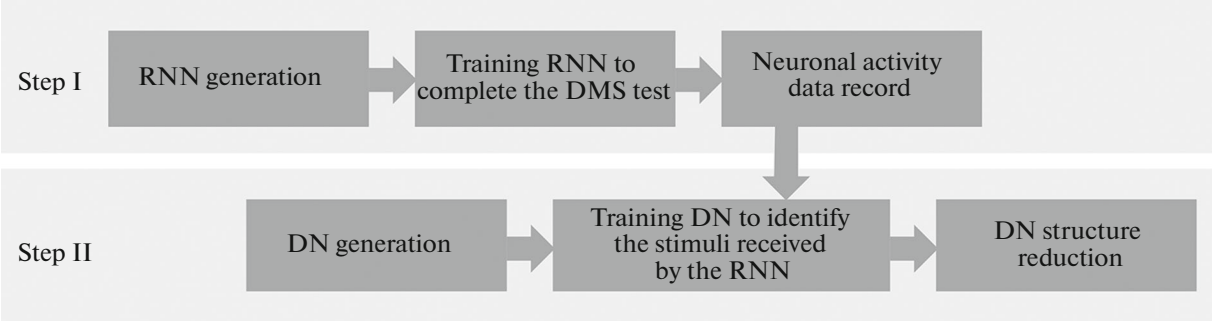

Fig. 2. DN training experiment outline.

storage period, the accuracy of identification using this method was different.

As the next step, the invariant representation storing the information about the received stimulus during four beats was extracted with the aid of an additional neural network, a neural network decoder (DN) (Fig. 2).

A single-layer neural network consisting of three neurons with linear response was used as the DN (2b). Each neuron had a modifiable synapse with each of the inputs, the number of which was equal to the number of neurons in the RNN. The DN produced 1 at one of the three neurons corresponding to the assigned stimulus and zeros at the others. The error backpropagation algorithm was used for training. The loss function was quadratic as it was in the previous case (3).

The input data for the DN were the neural activity of a particular RNN which was subjected to the DMS test. The neural activity of the RNN was recorded line by line. The line contained the activity of each of the 25 RNN neurons at a given moment in time (3, 4, 5, or 6 beats after the first stimulus), and the stimulus, the information about which was stored in the RNN at that time, was associated with it. For all the trained RNNs, 72 lines were recorded, which were distributed between the training and test samples randomly.

An individual DN needed to be trained for each trained RNN, indicating the uniqueness of the neural network's internal stimulus representation. The trained DNs decoded the stimuli represented by RNNs with 100% accuracy. The DN structure was further reduced, namely, the synapses with the lowest absolute values were sequentially equated to zero, and at each step, the DN was trained again until it reached its original performance level. This procedure was stopped when DN performance started to decline. As a result, a group of six or seven neurons was selected for each of the trained RNNs, whose activity was used to decode the received stimuli.

The neural activity invariants corresponding to the conditions for recognizing each of the three stimuli may be localized in the multidimensional space of dynamic patterns. Towards this end, the sets of randomly generated numbers imitating the activities of RNN neurons selected for decoding were applied to the inputs of trained DNs. Those sets of random numbers which were identified as corresponding to any of the stimuli by the DN were selected and considered as

points in the neural activity space representing the code of a certain stimulus.

Let us consider the structure of a particular DN as an example. This DN identified the six neurons from the original network numbered 2, 3, 8, 17, 20, 21 as significant and sufficient.

When the input data pass through the DN, the result of the calculation has the general form $w_2^\alpha x_2 + w_3^\alpha x_3 + w_8^\alpha x_8 + w_{17}^\alpha x_{17} + w_{20}^\alpha x_{20} + w_{21}^\alpha x_{21} = s^\alpha$, where w_i^α is a nonzero weight DN coefficient obtained after DN structure was reduced was reduced which links the i DN input with the neuron responsible for the recognition of the α stimulus, x_i is the activity of the neuron i in the original RNN, and $\alpha = A, B, C$. When $s^A \geq 1$, $s^B \leq 0$, and $s^C \leq 0$, the DN identifies the resulting data set as storing the information about the stimulus A. The datasets for the stimuli B and C are identified in the same way.

Linear polynomials which allow identifying the invariants for the RNN considered as an example are as follows:

$$0.158x_2 - 0.142x_3 + 0.461x_8 + 0.595x_{17} - 0.582x_{20} + 0.245x_{21} = s^A,$$

$$-0.244x_2 + 0.545x_3 - 0.079x_8 - 0.4x_{17} + 0.072x_{20} - 0.509x_{21} = s^B,$$

$$-0.243x_2 - 0.292x_3 - 0.827x_8 - 0.349x_{17} + 0.328x_{20} + 0.887x_{21} = s^C.$$

Principal component analysis showed that the points corresponding to invariant recognition property in the neuronal activity space form three compact clusters (see Fig. 3 for an example). Mapping the neuron activity onto the two-dimensional plane formed by the first and the second principal components is sufficient for recognition. The activity of one neuron among the six RNN neurons make insignificant contribution to the second principal component. This suggests that the activities of five neurons are sufficient for stimulus recognition for a given RNN, although such variant was not detected when reducing the DN structure. Therefore, principal component analysis or similar methods may be useful to finally minimize the number of stimulus coding invariant representations in the neural network.

Based on the results obtained in the present work, we may conclude that despite the dynamic nature of neuronal activity enabling the storage of information about the received stimuli the stimulus type may be

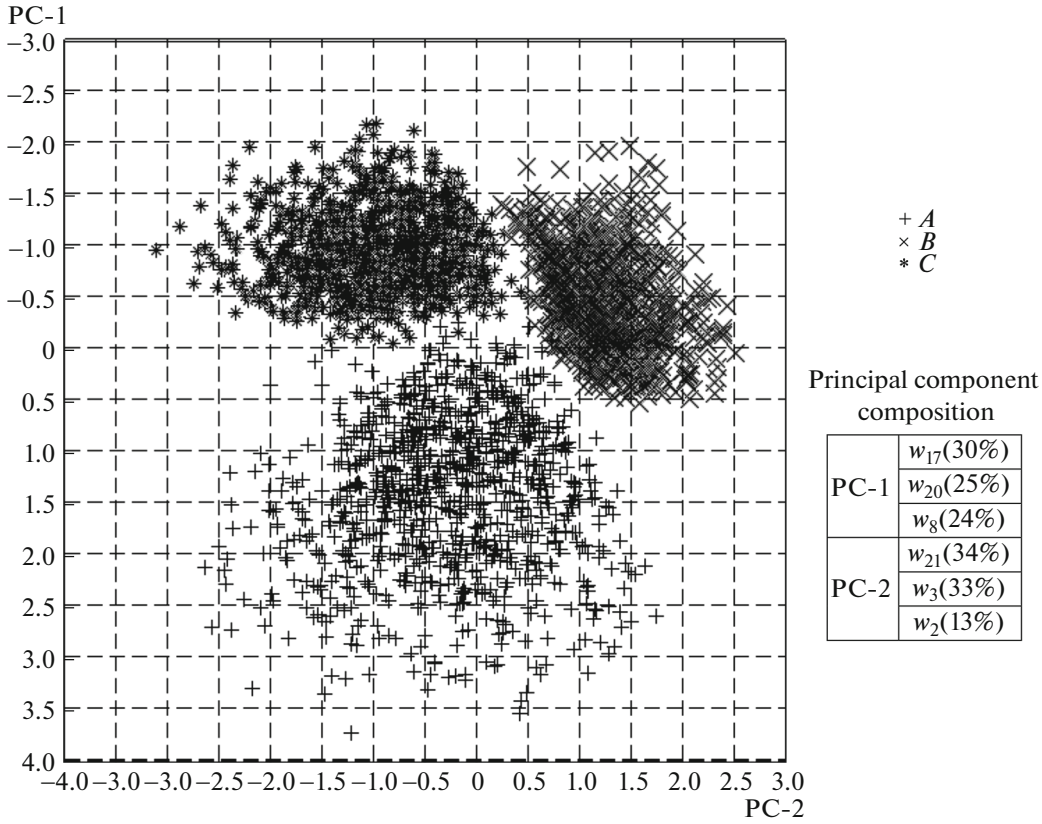

Fig. 3. Invariant configurations corresponding to stimulus recognition conditions after the principal component analysis.

identified based on neuronal activity patterns. Neural network–based decoding method proposed in the present work allows to identify the dynamic neuronal activity invariant which represents a given stimulus with 100% accuracy. In addition, this approach implies the identification of the minimum set of neurons and, consequently, the minimum neuronal activity required to solve the task set for the neural network; hence, this approach is in line with the concept of neural correlates [1].

FUNDING

The work was supported by the Russian Foundation for Basic Research, the Krasnoyarsk Krai Government, and the Krasnoyarsk Regional Science Foundation (project no. 20-41-240003).

COMPLIANCE WITH ETHICAL STANDARDS

The study does not contain any research involving animals or humans.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

OPEN ACCESS

REFERENCES

1. Crick, F. and Koch, C., *Nat. Neurosci.,* 2003, vol. 6, no. 2, pp. 119–126.
2. Meyers, E.M., D *J. Neurophysiol.,* 2018, vol. 120, no. 5, pp. 2260–2268.
3. Barak, O., Tsodyks, M., and Romo, R., J. *Neurosci.,* 2010, vol. 30, no. 28, pp. 9424–9430.
4. Stokes, M.G., Kusunoki, M., Sigala, N., et al., *Neuron,* 2013, vol. 78, no. 2, pp. 364–375.
5. Miconi, T., *Elife,* 2017, vol. 6б e20899.
6. Bartsev, S.I. and Bartseva, O.D., *Dokl. Biochem. Biophys.,* 2002, vol. 386, pp. 235–238.
7. Bartsev, S.I. and Bartseva, O.D., *Dokl. Biochem. Biophys.,* 2006, vol. 406, pp. 15–18.
8. Crowe, D.A., Averbeck, B.B., and Chafee, M.V., *J. Neurosci.,* 2010, vol. 30, no. 35, pp. 11640–11653.

Translated by E. Martynova