Bridging ANNs and Neuroscience: Insights from RNNs and CNNs

Summary: Artificial Neural Networks (ANNs) have become indispensable in machine learning and share a strong connection with neuroscience [1]. This study explores the relationship between ANNs and neuroscience from two perspectives. First, we demonstrate the efficiency of Recurrent Neural Networks (RNNs) by applying them to various neuroscience tasks. Second, we leverage insights from neuroscience, particularly methods used to analyze biological neurons, to study the functions of individual neurons in Convolutional Neural Networks (CNNs). This approach proves to be valuable for enhancing our understanding of neural networks.

1 From ANNs to Neuroscience: Using RNNs for Neuroscience Tasks

Recurrent Neural Networks (RNNs) are widely used in neuroscience to model cognitive, motor, and navigational systems. Unlike their machine learning counterparts, RNNs in neuroscience are often formulated as continuous-time dynamical systems to better capture the temporal dynamics observed in biological neural networks. In contrast to the discrete-time framework commonly used in machine learning [2], where the state at time step t is derived from the previous state $t-1$, neuroscience models are governed by continuous-time equations [3]:

$$\tau\frac{\mathbf{dr}}{\mathbf{dt}} = -\mathbf{r(t)} + \mathbf{f}\left(\mathbf{W_r r(t)} + \mathbf{W_x x(t)} + \mathbf{b_r}\right) \quad (1) \qquad \mathbf{r(t+\Delta t)} \approx \mathbf{r(t)} + \frac{\mathbf{\Delta t}}{\tau}\left[-\mathbf{r(t)} + \mathbf{f}\left(\mathbf{W_r r(t)} + \mathbf{W_x x(t)} + \mathbf{b_r}\right)\right] \quad (2)$$

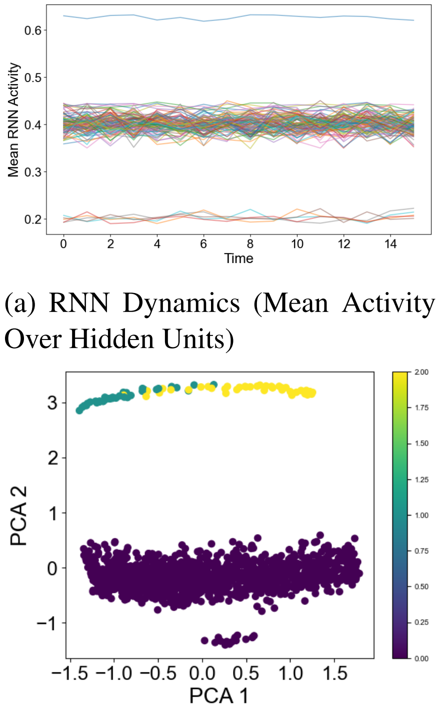
Here, τ represents the timescale of a single unit. Equation 1 describes the continuous-time dynamics, while equation 2 shows its discretization using the Euler method with a small time step Δt, where $\Delta t < \tau$.

To train our model, we used NeuroGym [4], a toolkit for training RNNs on neuroscience tasks, utilizing supervised learning. Notably, we employed mutual information loss [5](Test Acc: 99.32%), which is more relevant to neuroscience applications, instead of traditional cross-entropy loss. This is because mutual information is often used to quantify information transfer between neurons, providing a more biologically plausible measure of neural activity.

Experiment 1 (Visualizing RNN Activity)**.** This experiment explores the temporal dynamics of RNNs by visualizing hidden unit activity over time [6]. By analyzing the stability, variability, and dimensionality of RNN activity, we aim to understand how the network processes sequential data. These findings highlight the RNN's ability to learn task-specific features and emulate biologically plausible neural dynamics, especially when incorporating neuroscience-inspired loss functions and continuous-time models.

1. Temporal Stability (Figure 1a): Activity stabilizes after an initial transient phase, demonstrating the network's capacity to reach steady-state dynamics under repeated inputs.
2. Hidden Unit Variability (Figure 1a): Hidden unit trajectories exhibit diverse patterns, suggesting specialized roles in input processing and output generation, akin to functional diversity in biological neurons.
3. Dimensionality Reduction (Figure 1b): PCA [7] reduces the high-dimensional activity into clusters, potentially reflecting distinct cognitive or behavioral states encoded by the RNN.

These results underscore the RNN's ability to capture task-relevant features while modeling biologically plausible neural dynamics, highlighting the importance of neuroscience-inspired architectures.

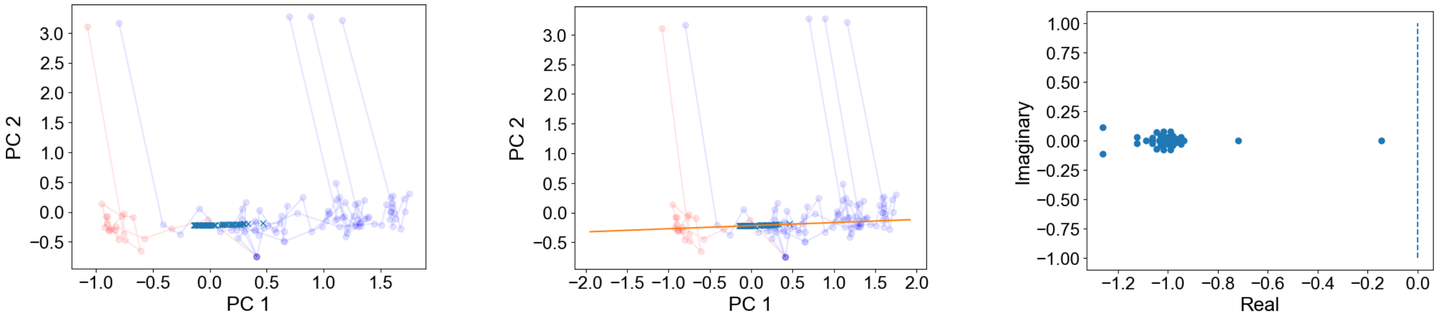


(a) RNN Dynamics (Mean Activity Over Hidden Units)

(b) PCA of RNN Activity

Figure 1: Visualization of RNN Activity

Experiment 2 (Dynamical System Analysis of RNNs)**.** In this experiment, we explore the dynamics of RNNs by analyzing their fixed points and local behavior. Fixed points, defined by $\mathbf{F}(\mathbf{r}) = 0$, are identified via gradient-based optimization that minimizes $||\mathbf{F}(\mathbf{r})||^2$ [8]. The local dynamics around these fixed points are approximated by the linear system $\frac{d\Delta\mathbf{r}}{dt} = J(\mathbf{r}_{\text{ss}})\Delta\mathbf{r}$, where $J(\mathbf{r}_{\text{ss}})$ is the Jacobian matrix evaluated at the fixed point. Eigenvalue analysis of $J(\mathbf{r}_{\text{ss}})$ determines stability: negative eigenvalues indicate stable fixed points, while positive eigenvalues signal instability. After 10,000 optimization steps, fixed points are visualized in PCA space (Figure 2a), revealing clusters of stable states and potential transitions. Figure 2b depicts the attractor dynamics, where trajectories (colored lines) converge to these fixed points, illustrating the network's tendency to settle in stable states. Eigenvalue analysis in Figure 2c highlights the dominant directions in the state space, with the largest eigenvalue's eigenvector defining the line attractor that dictates the network's response to perturbations. These analyses provide insights into RNN stability, memory encoding, and state transitions, shedding light on how the network evolves towards stable attractors in response to both inputs and internal dynamics.



(a) Fixed Points

(b) Activity Trajectories

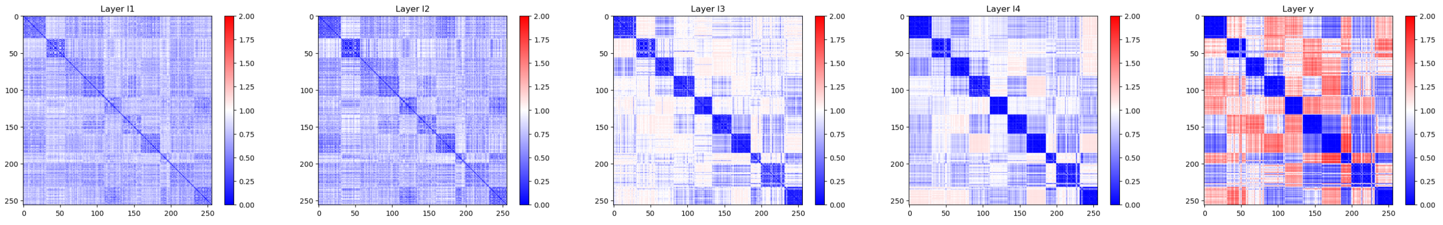
(c) Jacobian Eigenvalues

Figure 2: Dynamical System Analysis of RNN

2 From Neuroscience to ANNs: Understanding CNNs with Neuroscience Methods

Understanding artificial neural networks (ANNs) in both machine learning and neuroscience can often be challenging. However, methods used to study biological neural networks can provide useful insights for analyzing the performance of ANNs. For instance, to explore the link between neurons and behaviors, researchers may inactivate certain neurons to observe the effects [9]. Similarly, disconnecting two groups of neurons can help reveal their relationship. In this section, we explore neuroscience-inspired methods to study the functions of neurons in ANNs [10]. Specifically, we focus on using a Convolutional Neural Network (CNN) trained on the MNIST dataset.

Experiment 3 (Connection in neural activities)**.** In this experiment, we analyze two neuron groups (N_1 and N_2) by comparing their activities (S_1 and S_2) under the same D task conditions. Using neuroscience-inspired techniques, we construct $D \times D$ dissimilarity matrices (Figure 3) to study neuronal relationships and processing, similar to methods used for biological neural populations. Layers 1 and 2 (Figures 3a, 3b) capture basic features with high similarity, while deeper layers (Figures 3c, 3d) form distinct clusters, reflecting complex feature separation. The output layer (Figure 3e) shows clear classification boundaries.



(a) Layer 1

(b) Layer 2

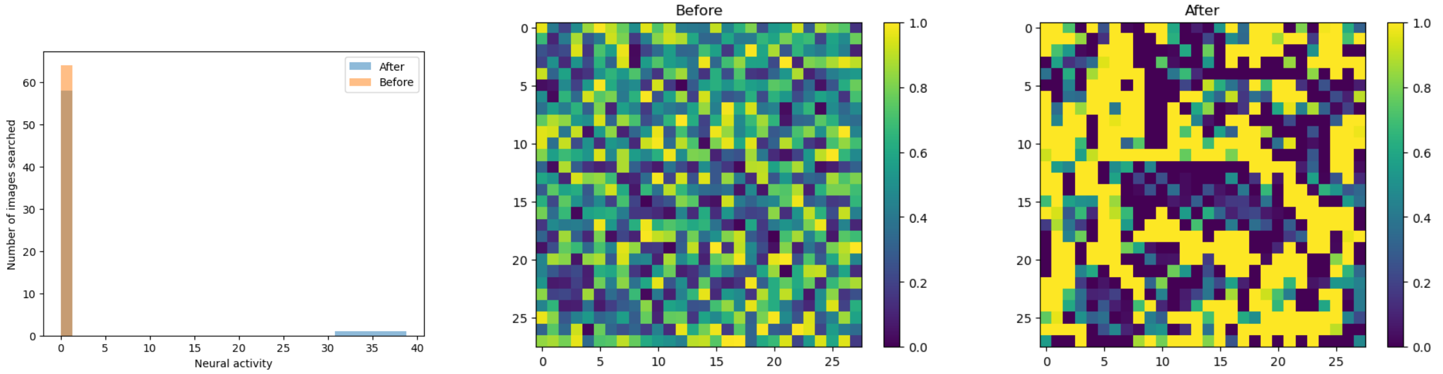
(c) Layer 3

(d) Layer 4

(e) Layer Y

Figure 3: Visualizations of connection activities across different layers

Experiment 4 (Prediction of S_2)**.** In this experiment, we use linear regression to predict matrix S_2 from a linear transformation of S_1 [11]. Inspired by neuroscience methods that study neuron responses by stimulating specific neurons, we optimized synthetic images to maximize the activation of a specific neuron in Layer 4 of the CNN. Figure 4 illustrates the process: images are initially random (Figure 4b), and an Adam optimizer adjusts their pixel values to enhance the neuron's activation (Figure 4c). Before optimization, activations are low and random, but after optimization, certain images strongly activate the neuron (Figure 4a), revealing the features it responds to most. This neuroscience-inspired approach helps uncover the neuron's learned preferences and interpret the network's internal representations.



(a) Activity Distribution

(b) Randomly Initialized Image

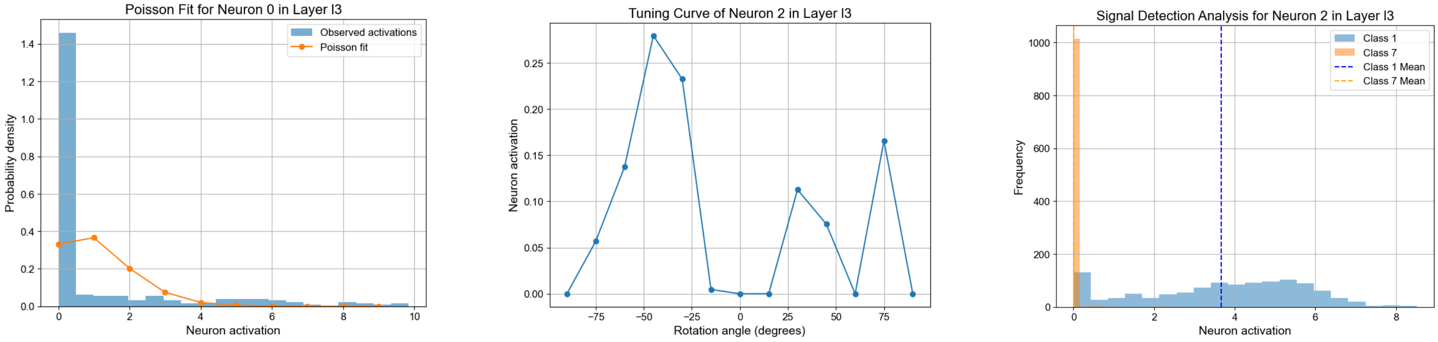
(c) Optimized Synthetic Image

Figure 4: Visualization of Predictions for S_2

Experiment 5 (Neural Feature Encoding and Decoding Analysis)**.** This experiment consists of two parts: neural encoding and decoding [12]. In the encoding analysis, the activation distribution of Neuron 2 in Layer 3 was fitted to a Poisson distribution, similar to methods used in neuroscience when examining neural response patterns. The results (Figure 5a) show that the observed activations align well with the expected distribution, emphasizing the neuron's role in encoding input information. In the decoding analysis, a logistic regression model was used to classify labels based on Layer 3 activations, achieving an accuracy of 94.23%. This demonstrates Layer 3's ability to capture meaningful, task-relevant features for classification.

Experiment 6 (Rotation Tuning Analysis of Neuronal Response)**.** This experiment analyzes the tuning curve of Neuron 2 in Layer 3 of the neural network by measuring its responses to rotated images. Sharing the similarity that how tuning curves are used in neuroscience to study orientation selectivity [13], we plotted the neuron's activation as a function of the rotation angle in Figure 5b. The results reveal distinct peaks at specific angles, showing that the neuron is highly selective to certain orientations. This indicates the neuron's role in encoding rotational features and provides insights into its functional behavior in processing visual patterns.

Experiment 7 (Signal Detection Analysis for Neural Discriminability)**.** This experiment takes inspiration from neuroscience by performing a signal detection analysis [14] for Neuron 2 in Layer 3. The activations of this neuron were compared for inputs from Class 1 and Class 7. Figure 5c shows our result. The histogram reveals distinct distributions, with the two classes showing clearly separated means. A d-prime value of 2.52 was calculated, indicating the neuron's strong ability to differentiate between these classes. This demonstrates how the neuron contributes to feature separation and supports the network's overall classification performance.



(a) Poisson Fit

(b) Tuning Curve

(c) Signal Detection

Figure 5: Three Further Experiments

References

[1] G. R. Yang and X.-J. Wang, “Artificial neural networks for neuroscientists: a primer,” *Neuron*, vol. 107, no. 6, pp. 1048–1070, 2020.

[2] J. L. Elman, “Finding structure in time,” *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[3] H. R. Wilson and J. D. Cowan, “Excitatory and inhibitory interactions in localized populations of model neurons,” *Biophysical journal*, vol. 12, no. 1, pp. 1–24, 1972.

[4] M. Molano-Mazon, J. Barbosa, J. Pastor-Ciurana, M. Fradera, R.-Y. Zhang, J. Forest, J. del Pozo Lerida, L. Ji-An, C. J. Cueva, J. de la Rocha, *et al.*, “Neurogym: An open resource for developing and sharing neuroscience tasks,” 2022.

[5] P. Dey, A. Khan, G. Saha, and R. K. Pal, “Mirnn: A mutual information augmented recurrent neural network framework for reconstruction of gene regulatory networks,” in *2024 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8, IEEE, 2024.

[6] A. Karpathy, J. Johnson, and L. Fei-Fei, “Visualizing and understanding recurrent networks,” *arXiv preprint arXiv:1506.02078*, 2015.

[7] H. Hotelling, “Analysis of a complex of statistical variables into principal components.,” *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[8] D. Sussillo and O. Barak, “Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks,” *Neural computation*, vol. 25, no. 3, pp. 626–649, 2013.

[9] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang, “Task representations in neural networks trained to perform many cognitive tasks,” *Nature neuroscience*, vol. 22, no. 2, pp. 297–306, 2019.

[10] A. S. Andalman, V. M. Burns, M. Lovett-Barron, M. Broxton, B. Poole, S. J. Yang, L. Grosenick, T. N. Lerner, R. Chen, T. Benster, *et al.*, “Neuronal dynamics regulating brain and behavioral state transitions,” *Cell*, vol. 177, no. 4, pp. 970–985, 2019.

[11] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, “Performance-optimized hierarchical models predict neural responses in higher visual cortex,” *Proceedings of the national academy of sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[12] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, “Neural encoding and decoding with deep learning for dynamic natural vision,” *Cerebral cortex*, vol. 28, no. 12, pp. 4136–4160, 2018.

[13] P. Seriès, P. E. Latham, and A. Pouget, “Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations,” *Nature neuroscience*, vol. 7, no. 10, pp. 1129–1135, 2004.

[14] W. Van Drongelen, *Signal processing for neuroscientists*. Academic press, 2018.