Rectified Lagrangian for Out-of-Distribution Detection in Modern Hopfield Networks

Ryo Moriai[*1], **Nakamasa Inoue**[*1], **Masayuki Tanaka**[1], **Rei Kawakami**[1],
Satoshi Ikehata[1, 3], **Ikuro Sato**[1,2]
[1] Institute of Science Tokyo, Japan
[2] Denso IT Laboratory, Inc. Japan
[3] National Institute of Informatics, Japan
moriai.ryo@d-itlab.titech.ac.jp, inoue@comp.isct.ac.jp

Abstract

Modern Hopfield networks (MHNs) have recently gained significant attention in the field of artificial intelligence because they can store and retrieve a large set of patterns with an exponentially large memory capacity. A MHN is generally a dynamical system defined with Lagrangians of memory and feature neurons, where memories associated with in-distribution (ID) samples are represented by attractors in the feature space. One major problem in existing MHNs lies in managing out-of-distribution (OOD) samples because it was originally assumed that all samples are ID samples. To address this, we propose the rectified Lagrangian (RegLag), a new Lagrangian for memory neurons that explicitly incorporates an attractor for OOD samples in the dynamical system of MHNs. RecLag creates a trivial point attractor for any interaction matrix, enabling OOD detection by identifying samples that fall into this attractor as OOD. The interaction matrix is optimized so that the probability densities can be estimated to identify ID/OOD. We demonstrate the effectiveness of RecLag-based MHNs compared to energy-based OOD detection methods, including those using state-of-the-art Hopfield energies, across nine image datasets.

1 Introduction

Associative memory models have been proposed to model memory retrieval in the brain through fixed-point search in an artificial neural network. Hopfield networks (Hopfield 1982, 1984) are classic examples, based on the idea of using recurrently connected neurons to store and retrieve memory patterns. Although these models are theoretically sound, they suffer limited memory capacity, as the number of distinct memory patterns is at most proportional to the dimension of the feature space. Recently, numerous studies have explored models with significantly increased memory capacity, the so-called modern Hopfield networks (MHNs) (Krotov and Hopfield 2016; Demircigil et al. 2017; Krotov and Hopfield 2018; Barra, Beccaria, and Fachechi 2018; Agliari and De Marzo 2020). Some of them are known to have an exponentially large memory capacity with respect to the feature dimension (Demircigil et al. 2017).

From a theoretical perspective, Krotov and Hopfield (2021) introduced a dynamical system that represents associative memory in a continuous time space based on two-body interactions between neurons. In their system, two Lagrangian functions, one for memory neurons and one for feature neurons, determine the model dynamics. When certain pairs of Lagrangian functions are chosen, the system is reduced to classical Hopfield networks (Hopfield 1982), dense associative memory (Krotov and Hopfield 2016; Demircigil et al. 2017), or the MHNs described in Ramsauer et al. (2021), indicating that new Lagrangian function designs could lead to new MHNs.

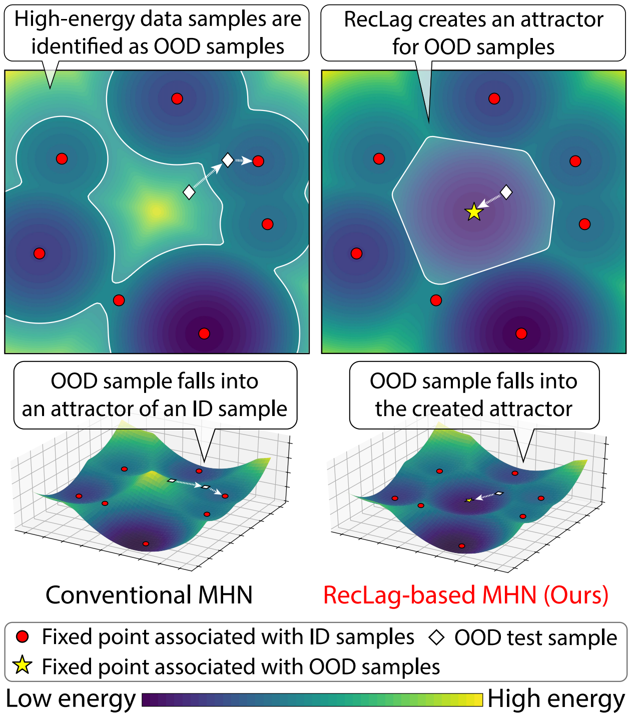

High-energy data samples are identified as OOD samples
RecLag creates an attractor for OOD samples
OOD sample falls into an attractor of an ID sample
OOD sample falls into the created attractor
Conventional MHN
RecLag-based MHN (Ours)
● Fixed point associated with ID samples
◇ OOD test sample
☆ Fixed point associated with OOD samples
Low energy
High energy

Figure 1: Visualization of energy landscapes. Existing methods identify samples with high Hopfield energy as OOD, though such samples fall into the attractors associated with ID samples. In contrast, our RecLag-based MHNs possess a dedicated attractor that specifically captures OOD samples.

While these studies have expanded the potential of MHNs

[*] Equal contribution

both theoretically and practically, one of the primary limitations of existing MHNs lies in managing out-of-distribution (OOD) samples. The dynamical system does find a fixed point for any test input; *i.e.*, an OOD sample is inevitably associated with one of the memorized in-distribution (ID) samples. Zhang et al. (2023) proposed an OOD-sample detection method based on the Hopfield energy function. However, they lack theoretical foundation explaining the relationship between the energy and the probability of the input/transient states. We are thus motivated to develop new MHNs equipping probability-aware OOD rejection functionality within the fixed-point search mechanism.

In this paper, we propose the rectified Lagrangian (RegLag), a new Lagrangian for memory neurons that creates an attractor for OOD samples in the dynamical system of MHNs, as shown in Fig. 1. RegLag introduces a rectified linear unit (ReLU) with a constant indicating the ID memory strength to the Lagrangian function of memory neurons. We theoretically show that 1) RecLag creates a trivial point attractor for any interaction matrix and 2) RecLag-based MHNs are reduced to vanilla MHNs when the ID memory strength is infinitely large, indicating our approach is a natural extension of existing MHNs. We further devise a training method for RecLag-based MHNs via probabilistic interaction, along with a probability density estimated for ID samples by optimizing the interaction matrix. Our contributions are summarized below:

1. We propose RecLag, a new Lagrangian for memory neurons. RecLag is designed to create a trivial point attractor for any interaction matrix, enabling OOD detection by identifying samples that fall into the attractor as OOD.
2. We propose a training method for RecLag-based MHNs having a probabilistic interaction between memory and feature neurons. We prove that samples with low probability fall into the special attractor created by RecLag.
3. We demonstrated the effectiveness of our approach in comparison with energy-based OOD detection methods, including those using state-of-the-art Hopfield energy functions (Zhang et al. 2023) on nine image datasets.

2 Related Work

Hopfield Networks. Hopfield networks (Hopfield 1982, 1984) are a type of artificial neural network with recurrent structures that model associative memory. Their development laid the foundation for later models such as Boltzmann machines (Ackley, Hinton, and Sejnowski 1985) and long short-term memory (Hochreiter and Schmidhuber 1997) in the latter part of the 20th century.

In recent years, MHNs, also known as dense associative memory (Krotov and Hopfield 2016), have been attracting attention because they can have an exponentially large memory capacity (Demircigil et al. 2017). Numerous studies have demonstrated the effectiveness of MHNs on various tasks including image classification (Fürst et al. 2021; Ota et al. 2023), immune repertoire classification (Widrich et al. 2020), tabular data classification (Schäfl et al. 2021), reaction template prediction (Seidl et al. 2022), predictive coding (Salvatori et al. 2021) and reinforcement learning (Widrich et al. 2021).

MHNs are formulated as dynamical systems described by analytical differential equations. Specifically, Ramsauer et al. (2021) generalized the energy function from discrete states to continuous states, and then Krotov and Hopfield (2021) formulated the dynamical system of MHNs with two-body differential equations. Follow-up studies, such as work on universal Hopfield networks (Millidge et al. 2022), have further generalized the dynamical system.

OOD Detection. OOD detection aims to identify data samples that deviate from the distribution of training data samples. This paper focuses on post hoc approaches, where the detection mechanism is applied after the model has been trained. One of the most well-known approaches is maximum softmax probability (MSP) scoring (Hendrycks and Gimpel 2017), which uses the highest softmax output score to identify OOD samples, under the assumption that ID samples yield higher MSP scores compared to OOD samples. To more precisely estimate the distribution of OOD samples, various enhancements and alternative post hoc methods have been proposed (Liang, Li, and Srikant 2018; Liu et al. 2020; Sun, Guo, and Li 2021; Sun and Li 2022; Shen et al. 2023; Chen et al. 2024). Among them, energy-based OOD detection approaches (Liu et al. 2020; Sun, Guo, and Li 2021) are related to this study in the sense that MHNs have a scalar-valued function associated with the network states, the so-called the Hopfield energy.

Most recently, several pioneering studies have demonstrated the effectiveness of Hopfield energy for OOD detection (Zhang et al. 2023; Hofmann et al. 2024). Their methods identify data samples with high Hopfield energy as OOD samples and achieve superior performance among energy-based OOD detection methods. However, from a theoretical perspective, every test sample, including an OOD sample, falls into one of the attractors representing a memory pattern associated with an ID data sample as the dynamical system of MHNs evolves over time. To address this problem, this paper explores MHNs that explicitly have an attractor for OOD samples.

3 Modern Hopfield Networks

3.1 Lagrangian-Based Dynamical System

Notation and Settings. This paper discusses MHNs with the Lagrangian-based dynamical system proposed by Krotov and Hopfield (2021). We denote the feature neurons as $v(t) \in \mathbb{R}^{N_V}$ and the memory neurons as $h(t) \in \mathbb{R}^{N_H}$, both at continuous time $t \in \mathbb{R}_{\geq 0}$, where $N_V, N_H \in \mathbb{N}$ are the numbers of neurons. The dynamical system is described by the following differential equations:

$$\tau_V \frac{dv_i(t)}{dt} = \sum_{\mu=1}^{N_H} \xi_{i\mu} f_\mu(h(t)) - v_i(t), \tag{1}$$

$$\tau_H \frac{dh_\mu(t)}{dt} = \sum_{i=1}^{N_V} \xi_{\mu i} g_i(v(t)) - h_\mu(t), \tag{2}$$

where $\xi \in \mathbb{R}^{N_H \times N_V}$ is an interaction matrix representing the strength of synapses, $f : \mathbb{R}^{N_H} \to \mathbb{R}^{N_H}$ and $g : \mathbb{R}^{N_V} \to$

$\mathbb{R}^{N_V}$ are activation functions, and $\tau_V, \tau_H \in \mathbb{R}$ are constants that determine the dynamics of neurons. The activation functions are determined by the Lagrangians $L_H : \mathbb{R}^{N_H} \to \mathbb{R}$ and $L_V : \mathbb{R}^{N_V} \to \mathbb{R}$ such that

$$f(h) = \frac{\partial L_H(h)}{\partial h}, \quad g(v) = \frac{\partial L_V(v)}{\partial v}, \tag{3}$$

where $h \in \mathbb{R}^{N_H}$ and $v \in \mathbb{R}^{N_V}$. The energy function is then given by

$$\begin{aligned} E(v,h) = &\sum_{i=1}^{N_V} v_i g_i(v) - L_V(v) + \sum_{\mu=1}^{N_H} h_\mu f_\mu(h) \\ &- L_H(h) - \sum_{\mu,i} f_\mu(h) \xi_{\mu i} g_i(v). \end{aligned} \tag{4}$$

Note that this energy monotonically decreases; that is, we have $dE(v(t),h(t))/dt \leq 0$ along the trajectory of the dynamical system when the Hessian matrices of the Lagrangians are positive semi-definite.

Lagrangians. If we suppose a fixed interaction matrix ξ, then the model dynamics are defined by the choice of the Lagrangians. For example, when the Lagrangian functions are given by the additive functions

$$L_H(h) = \sum_{\mu=1}^{N_H} \sigma(h_\mu), \quad L_V(v) = \sum_{i=1}^{N_V} |v_i|, \tag{5}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinear function, the energy function reduces to

$$E(v) = -\sum_{\mu=1}^{N_H} \sigma\left(\sum_{i=1}^{N_V} \xi_{\mu i} \cdot \mathrm{sgn}(v_i)\right). \tag{6}$$

under the adiabatic limit $\tau_V \gg \tau_H$ when ξ is a symmetric matrix. This energy function is identical to that of dense associative memory (Krotov and Hopfield 2016). Further, when $\sigma(x) = x^2$, it reduces to the energy function of the classical Hopfield network (Hopfield 1982).

Recently, Krotov and Hopfield (2021) introduced the following Lagrangians:

$$L_H(h) = \frac{1}{\beta}\log\left(\sum_{\mu=1}^{N_H} \exp\left(\beta h_\mu\right)\right), L_V(v) = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2, \tag{7}$$

where $\beta \in \mathbb{R}_{\geq 0}$ is a constant. Under the adiabatic limit and when $\beta = 1$, the energy function reduces to

$$E(v) = -\log\left(\sum_{\mu=1}^{N_H} \exp\left(\sum_{i=1}^{N_V} \xi_{\mu i} v_i\right)\right) + \frac{1}{2}\sum_{i=1}^{N_V} v_i^2. \tag{8}$$

This energy function is identical to that of the MHNs proposed by Ramsauer et al. (2021).

3.2 Energy-Based OOD Detection

Let us consider classification problems and denote the number of ID classes for training as C. The goal of OOD detection is to identify data samples that do not belong to any of the C classes. Zhang et al. (2023) proposed using the energy function of MHNs for OOD detection. Specifically, they introduced two energy functions: modern Hopfield energy (MHE) and simplified Hopfield energy (SHE). MHE is obtained by replacing the interaction matrix ξ in Eq. (8) with a class-specific pattern matrix $S^c \in \mathbb{R}^{d \times N}$ and by omitting the second term as follows:

$$\mathrm{MHE}(\tilde{v}) = -\log\left(\sum_{\mu=1}^{d} \exp\left(\sum_{i=1}^{N} S^c_{\mu i} \tilde{v}_i\right)\right), \tag{9}$$

where $\tilde{v} \in \mathbb{R}^d$ is a test pattern, $c \in \{1, 2, \cdots, C\}$ is the classification result of $\tilde{v}$ obtained from a pre-trained classification model, d is the hidden dimension, and N is the number of stored patterns. SHE is a Taylor approximation of MHE, but is more effective than MHE at detecting OOD. It is defined as

$$\mathrm{SHE}(\tilde{v}) = \frac{1}{d}\sum_{\mu=1}^{d}\sum_{i=1}^{N} S^c_{\mu i} \tilde{v}_i. \tag{10}$$

OOD samples can be detected by applying a threshold to these energy functions. However, as the dynamical system of MHNs evolves over time, every test sample falls into an attractor associated with an ID data sample, indicating a lack of theoretical consistency.

4 Rectified Lagrangian

This section introduces RecLag, a Lagrangian function that creates a point attractor for OOD samples in the dynamical system of HMNs. As shown in Figure 2, RecLag creates a point attractor in the feature space. This attractor is designed to exist for any interaction matrix ξ, enabling OOD detection by identifying data samples that fall into it as OOD.

4.1 Definition

To incorporate a point attractor for OOD samples in the dynamical system, we propose a minimal yet effective modification to the Lagrangian function of memory neurons. Specifically, we introduce an *inverse memory strength constant* γ, which determines the strength of ID samples stored in memory, with a $\max$ function to screen out negative values, which is applied in the same way as ReLU. The proposed RecLag is defined as follows.

Definition 1. *We define RecLag as*

$$L_H(h) = \max\left(\frac{1}{\beta}\log\left(\frac{1}{\gamma}\sum_{\mu=1}^{N_H} \exp\left(\beta h_\mu\right)\right), 0\right), \tag{11}$$

where $\beta, \gamma \in \mathbb{R}_{\geq 0}$ *are constants.*

4.2 Existence of a Trivial Point Attractor

With the dynamical system using RecLag L_H in Eq. (11) for memory neurons and the Lagrangian L_V in Eq. (7) for feature neurons, Theorem 1 shows that there exists a trivial point attractor at the origin of the feature space for any interaction matrix.

Theorem 1. *Suppose that activation functions f and g in the dynamical system of Eqs. (1,2) are given by the derivatives* *of RecLag L_H in Eq. (11) and the Lagrangian L_V in Eq. (7), respectively. For any interaction matrix $\xi \in \mathbb{R}^{N_H \times N_V}$, a trivial point attracting set $A = \{\mathbf{0}\}$ exists at the origin $\mathbf{0} \in \mathbb{R}^{N_V}$ in the feature space when $\gamma > N_H$ under the adiabatic limit $\tau_V = dt$.*

Sketch of proof. With RecLag, writing the differential equations of the dynamical system in finite differences with $\frac{dv_i}{dt} \simeq \frac{v_i^{(k+1)} - v_i^{(k)}}{\Delta t}$ and $\tau_V = \Delta t$ gives the following update rule for feature neurons:

$$v_i^{(k+1)} = \chi\Big(G(v^{(k)})\Big) \sum_{\mu=1}^{N_H} \xi_{i\mu} \mathrm{softmax}\left(\beta \sum_{j=1}^{N_V} \xi_{\mu j} v_j^{(k)}\right), \quad (12)$$

where $k \in \mathbb{N}$ is a discrete time step, and

$$G(v) = \log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp\left(\beta \sum_{j=1}^{N_V} \xi_{\mu j} v_j\right)\right), \quad (13)$$

$$\chi(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases}. \quad (14)$$

When $v^{(k)} = \mathbf{0}$, we have $\chi(G(v^{(k)})) = 0$, and thus we have $v^{(k+1)} = \mathbf{0}$. This shows that $\mathbf{0} \in \mathbb{R}^{N_V}$ is a fixed point of the dynamical system in the feature space. Further, with the epsilon neighborhood of the origin $U_\epsilon = \{u : \|u\|_2 < \epsilon\}$, we have $\chi(G(u)) = 0$ for every $u \in U_\epsilon$ if ϵ is small enough. This shows that $A = \{\mathbf{0}\}$ is an attracting set for every fixed interaction matrix ξ. A full proof is given in Appendix A. □

4.3 Reduction to Vanilla MHNs

Along with the existence of the trivial point attractor, it is also worth noting the limit where it disappears. Theorem 2 shows that RecLag-based MHNs reduce to vanilla MHNs when the memory strength of ID samples is infinitely large, that is, when the inverse memory strength constant $\gamma \to 0$. This theoretical result indicates that our approach is a natural extension of MHNs. A proof is given in Appendix B.

Theorem 2. *Let v_{A} and v_{B} be feature neurons of a vanilla MHN and a RecLag-based MHN, respectively. Suppose $v_{\mathrm{A}}^{(0)} = v_{\mathrm{B}}^{(0)}$. For every $\epsilon > 0$, there exists a small $\gamma > 0$ such that $\sup_k \|v_{\mathrm{A}}^{(k)} - v_{\mathrm{B}}^{(k)}\|_2 < \epsilon$.*

4.4 Visualization and Discussion

Visualization. Figure 2 compares the energy distributions of a vanilla MHN and a RecLag-based MHN, where each red point indicates a fixed point $\xi_\mu \in \mathbb{R}^{N_V}$ at a local minimum of the energy function. As shown, RecLag creates an attractor at the origin of the feature space. This attractor is associated with OOD samples as described in the next section. The 3D visualization of these energy functions is shown in Figure 1 with trajectories of a test sample (white diamond-shaped point) over time. As shown, with the vanilla MHN, the test sample falls into one of the attractors even if it is an OOD sample. In contrast, with the RecLag-based MHN, the same test sample falls into the created attractor, indicating that none of the memory patterns are associated with it. This shows that the RecLag-based MHN can explicitly manage OOD samples in the dynamical system.

Memory Strength. The size of the created attractor increases as the inverse memory strength constant γ increases. Consequently, the number of samples identified as OOD samples also increases with γ. This indicates that γ can serve as a threshold parameter that adjusts the sensitivity of RecLag-based MHNs to OOD samples. In practice, to draw a receiver operating characteristic (ROC) curve, one could vary γ to generate different true positive rates (TPRs) and false positive rates (FPRs) for OOD detection.

5 Training via Probabilistic Interaction

This section discuss the basin of the attractor created by RecLag, and proposes a method for training the interaction matrix with ID samples. Because the basin obviously involves $B_0 = \{v : G(v) < 0\}$, as shown in the sketch of the proof for Theorem 1, we introduce a method to train the interaction matrix via probabilistic interaction, by which data samples with low probability density values fall into B_0.

5.1 Probabilistic Interaction

The probabilistic interaction explicitly chooses a single memory neuron for each input feature during training in a probabilistic manner. This creates a cycle of interaction between feature neurons and memory neurons in the following two steps. First, given an input feature $x \in \mathbb{R}^{N_V}$, a memory neuron is sampled as $\mu \sim p_H(\mu|x)$, where $\mu \in \{1, 2, \cdots, N_H\}$ is an index of memory neurons and $p_H(\mu|x)$ is a pre-defined conditional probability mass distribution. Second, given an index μ, an output feature $y \in \mathbb{R}^{N_V}$ is sampled as $y \sim p_V(y|\mu)$, where $p_V(y|\mu)$ is a pre-defined conditional probability density distribution.

Because this interaction can be understood as the stochastic feedforward neural network (SFNN) proposed by Tang and Salakhutdinov (2013), which samples an index of neurons in a hidden layer, we train the interaction matrix using the training method for SFNN. Specifically, given a set of ID data samples $\mathcal{D} \subset \mathbb{R}^{N_V}$, the interaction matrix ξ is trained to maximize the sum of probability products:

$$P = \sum_{x \in \mathcal{D}} p_V(x|\mu) p_H(\mu|x). \quad (15)$$

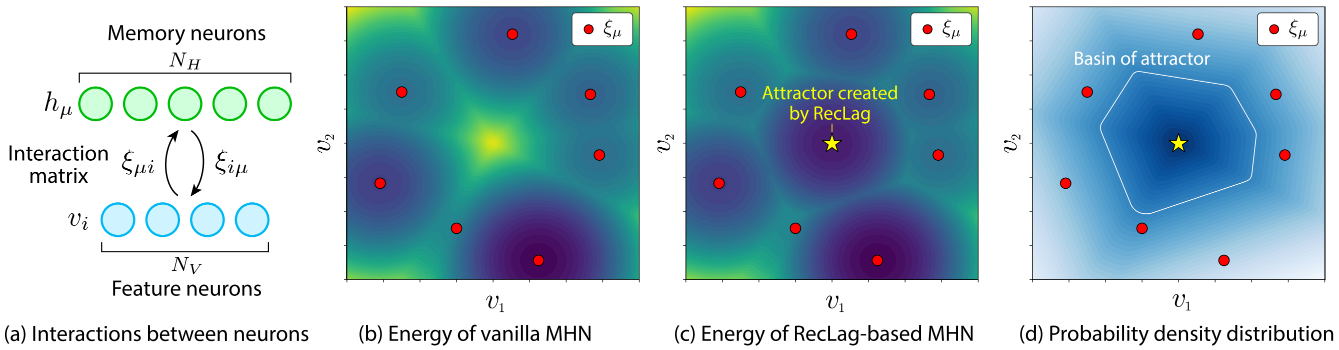
Here, the distribution $p_H(\mu|x)$ is computed through the joint probability distribution described in the next subsection. The distribution $p_V(x|\mu)$ is used only for training, and thus we use a Gaussian distribution following Tang and Salakhutdinov (2013):

$$p_V(x|\mu) = \frac{1}{\sqrt{(2\pi)^{N_V} |\Sigma|}} \exp\left(-\frac{1}{2}(x - \xi_\mu)^\top \Sigma^{-1} (x - \xi_\mu)\right) \quad (16)$$

where Σ is a learnable covariance matrix.

5.2 Attracting Probability

Interestingly, there exists a joint probability distribution $p_H(x, \mu)$ that relates the SFNN and the basin B_0. Specifically, Definition 2 provides the joint probability distribution,



(a) Interactions between neurons (b) Energy of vanilla MHN (c) Energy of RecLag-based MHN (d) Probability density distribution

Figure 2: (a) MHN with an interaction matrix $\xi_{\mu i}$ between memory neurons h_μ and feature neurons v_i. (b) Energy distribution of a vanilla MHN using the Lagrangians in Eq. (7). (c) Energy distribution of a RecLag-based MNH. The point attractor in Theorem 1 created by RecLag is marked by the yellow star. (d) Probability density distribution in Eq. (17). Data samples with low probability density values fall into the created attractor.

by which the conditional probability for the SFNN is computed as $p_H(\mu|x) = p_H(x,\mu)/\sum_\mu p_H(x,\mu)$, and data samples with low probability density values fall into the basin.

Definition 2. *Let X be a continuous random variable of feature neurons over $\mathbb{R}^{N_V}$, and let M be a discrete random variable of the index of hidden neurons over $\{1, 2, \cdots, N_H\}$. We define the joint probability distribution function as*

$$p_H(X = x, M = \mu) = \frac{1}{Z} \exp\left(\beta \sum_{j=1}^{N_V} \xi_{\mu j} x_j\right). \quad (17)$$

Here, Z is a normalization constant given by

$$Z = \sum_{\mu=1}^{N_H} \int_{\mathcal{S}} \exp\left(\beta \sum_{j=1}^{N_V} \xi_{\mu j} x_j\right) dx, \quad (18)$$

where $\mathcal{S} \in \mathbb{R}^{N_V}$ is a sufficiently large hypersphere to cover all data samples.

5.3 OOD Detection

Finally, Theorem 3 shows that the probability density distribution $p_H(x) = \sum_\mu p_H(x,\mu)$ explicitly models the distribution of ID samples and that all data samples with a probability density lower than δ fall into the attractor created by RecLag. Therefore, OOD samples can be detected by evaluating $p_H(\tilde{v})$ given a test sample $\tilde{v} \in \mathbb{R}^{N_V}$. A proof is given in Appendix C.

Theorem 3. *The basin $B_0 = \{v : G(v) < 0\}$ is identical to the set of points that have low probability density values. In other words, a threshold δ exists such that*

$$B_0 = \{x : p_H(X = x) < \delta\}. \quad (19)$$

Visualization. Figure 2(d) shows the probability density function, where the basin boundary is drawn in white.

6 Experiments

We focus on evaluating OOD detection performance of our proposed method along with strong baselines in this work.

6.1 Experimental Settings

Datasets. Eleven image datasets were used to conduct OOD detection experiments: CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), SVHN (Netzer et al. 2011), LSUN-C (Yu et al. 2015), LSUN-R (Yu et al. 2015), iSUN (Xu et al. 2015), Places365 (Zhou et al. 2017), DTD (Cimpoi et al. 2014), TinyImageNet (TIN) (Deng et al. 2009), SUN (Xiao et al. 2010), and iNaturalist (Van Horn et al. 2018). The CIFAR-10 or CIFAR-100 dataset was used as the ID dataset, and the other nine datasets were used as OOD datasets.

Evaluation Measure. We used FPR95 as the primary evaluation measure, which is the FPR of OOD samples when the TPR for ID samples is 95.0%. ROC curves and the area under the curve (AUC) are also reported.

Baselines. We chose five baseline methods: MSP scoring (Hendrycks and Gimpel 2017), energy-based detection (Energy) (Liu et al. 2020), rectified activations applied to energy (ReAct) (Sun, Guo, and Li 2021), MHE (Zhang et al. 2023), and SHE (Zhang et al. 2023). Note that the last four methods are energy-based OOD detection methods, with MHE and SHE being state-of-the-art using MHNs. Also note that these methods, including ours, process representations from a frozen encoder. For a fair comparison, we use the same encoder in each experiment. Another type of ODD detection methods (such as Zhang et al. (2023)) that jointly optimize encoder and OOD module in a specific fashion is excluded.

Neural networks. Three image classification networks were used: ResNet18 (He et al. 2016), ResNet34 (He et al. 2016), and WideResNet40-2 (WRN40-2) (Zagoruyko and Komodakis 2016). They were trained on an ID dataset using cross-entropy loss. The OOD benchmark was conducted with no dynamics simulations (no extra computational costs compared to the baselines). Other implementation details are given in Appendix D.

6.2 Experimental Results

Comparison With Energy-Based Methods. Table 1 shows the OOD detection performance on the nine OOD datasets

	Method	SVHN	LSUN-C	LSUN-R	iSUN	Places	DTD	TIN	SUN	iNaturalist	Average
ResNet18	MSP	76.34	27.52	36.54	34.84	20.55	30.65	45.82	22.89	12.62	34.19
	Energy	56.05	8.10	11.60	9.10	3.18	16.98	25.47	3.27	3.47	15.25
	ReAct	59.47	7.57	12.52	10.13	2.93	16.86	27.61	3.27	3.80	16.02
	MHE	17.59	9.20	7.68	4.74	0.33	8.96	15.86	**0.00**	2.35	7.41
	SHE	**17.45**	9.22	7.69	4.77	0.33	8.99	15.84	**0.00**	2.38	7.41
	RecLag	18.12	**6.40**	**4.60**	**2.67**	**0.28**	**6.82**	**12.09**	**0.00**	**1.68**	**5.85**
		± 2.02	**± 0.25**	**± 0.12**	**± 0.47**	**± 0.02**	**± 0.13**	**± 0.25**	**±0.00**	**± 0.04**	**± 0.24**
ResNet34	MSP	59.86	28.26	32.06	31.69	33.61	43.28	45.56	32.43	32.95	37.74
	Energy	30.51	6.84	9.43	8.47	9.32	23.74	25.16	8.99	10.86	14.81
	ReAct	45.86	14.37	14.09	13.28	15.83	29.73	31.60	15.53	11.98	21.36
	MHE	6.20	6.17	4.40	2.94	2.34	14.32	15.86	0.54	4.91	6.41
	SHE	6.14	6.20	4.45	3.01	2.36	14.32	15.93	0.54	4.92	6.43
	RecLag	**5.19**	**5.60**	**2.85**	**2.11**	**2.31**	**12.04**	**11.71**	**0.33**	**4.14**	**5.14**
		± 0.24	**± 0.07**	**± 0.05**	**± 0.05**	**± 0.03**	**± 0.07**	**± 0.23**	**± 0.11**	**± 0.08**	**± 0.08**
WRN40-2	MSP	41.52	44.43	38.47	39.70	33.84	35.80	51.52	34.88	27.69	38.65
	Energy	15.35	17.77	14.98	17.45	10.58	19.71	36.75	9.54	8.95	16.79
	ReAct	18.83	19.93	18.25	20.68	11.98	21.67	42.02	11.44	13.26	19.78
	MHE	5.40	14.60	12.03	11.48	2.90	10.99	27.28	**0.82**	1.83	9.70
	SHE	**5.25**	14.39	13.18	12.39	2.83	10.98	28.35	**0.82**	1.84	10.00
	RecLag	5.75	**7.37**	**8.44**	**8.01**	**2.63**	**9.75**	**22.62**	1.06	**1.67**	**7.47**
		± 0.12	**± 0.18**	**± 0.17**	**± 0.15**	**± 0.05**	**± 0.10**	**± 0.34**	± 0.09	**± 0.05**	**± 0.85**

Table 1: OOD detection performance as FPR95(%) ↓ with CIFAR-10 images being ID samples. Our RecLag-based MHN (RecLag) is compared with MSP (Hendrycks and Gimpel 2017), Energy (Liu et al. 2020), ReAct (Sun, Guo, and Li 2021), MHE (Zhang et al. 2023), and SHE (Zhang et al. 2023). For the proposed RecLag the trimmed means and standard deviations (following ± symbols) over 11 trials with the largest and the smallest ones being trimmed are reported.

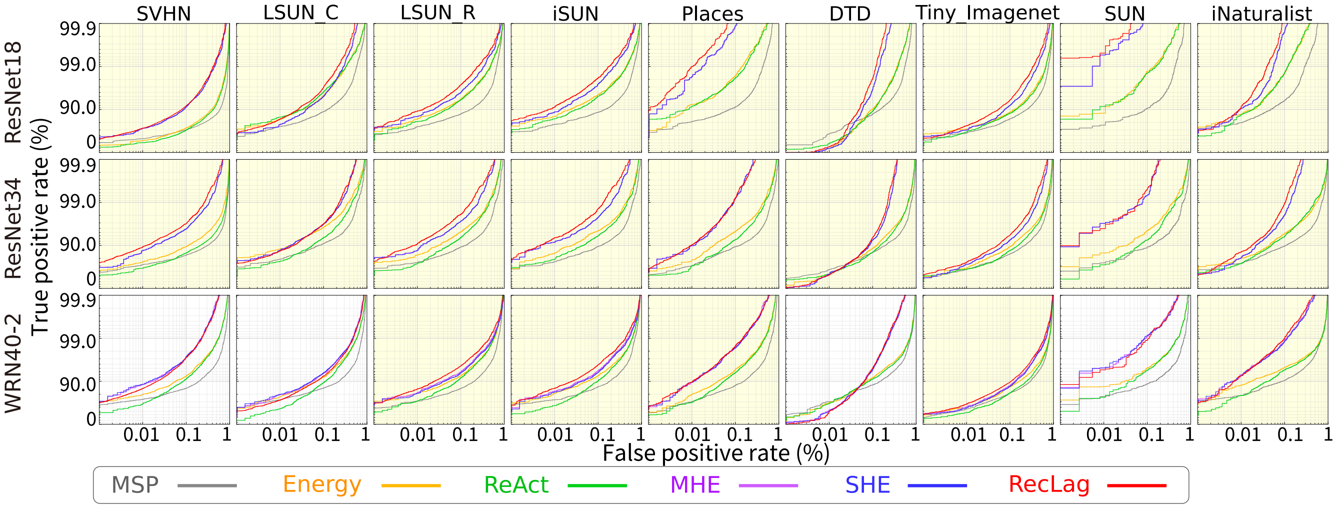


Figure 3: ROC curves (log-scale) and AUC↑. Yellow background indicates that RecLag performed the best in terms of AUC.

with the three neural networks trained on the CIFAR-10 dataset. As shown, our RecLag-based MHN (RecLag) achieved the best average performance across all neural networks. This demonstrates the effectiveness of our approach, which incorporates an attractor for OOD samples in post-hoc OOD detection scenarios.

ROC Curves. Figure 3 reports the ROC curves with the AUC values. As shown, RecLag exhibited the best AUC value in 23 out of 27 comparisons (highlighted with the yellow background), indicating its consistent superiority in OOD detection performance.

In-Distribution Data. To investigate how OOD detection performance is affected when a neural network is trained on a more complex task, Table 2 shows the results for WRN40-2 trained on CIFAR-100. As shown, the OOD performance decreases for all methods compared to those in Table 1. This is because the variance of features in ID samples increased, making OOD detection more challenging. However, even in this case, our RecLag-based MHN outperformed the other methods. This result indicates that the relative effectiveness

Method	SVHN	LSUN-C	LSUN-R	iSUN	Places	DTD	TIN	SUN	iNaturalist	Average
MSP	73.51	67.42	88.65	86.73	66.61	81.79	83.06	73.57	72.27	77.07
Energy	66.00	54.96	82.88	82.23	56.55	78.85	77.49	66.21	70.86	70.67
ReAct	61.33	52.73	83.36	83.48	53.61	74.92	77.27	62.67	66.29	68.41
MHE	16.24	41.21	67.61	56.08	9.99	40.80	61.79	10.35	17.22	35.70
SHE	16.15	41.07	67.78	56.42	**9.91**	40.37	61.89	**10.08**	**16.90**	35.61
RecLag	**15.50** ± **3.09**	**39.94** ± **0.80**	**65.28** ± **3.96**	**55.67** ± **4.52**	11.57 ± 0.54	**39.18** ± **1.13**	**59.02** ± **4.43**	12.17 ± 0.64	19.29 ± 1.29	**35.29** ± **1.93**

Table 2: OOD detection performance as FPR95(%) ↓ with CIFAR-100 images being ID samples. WRN40-2 arch. was used. For other descriptions, see the caption of Table 1.

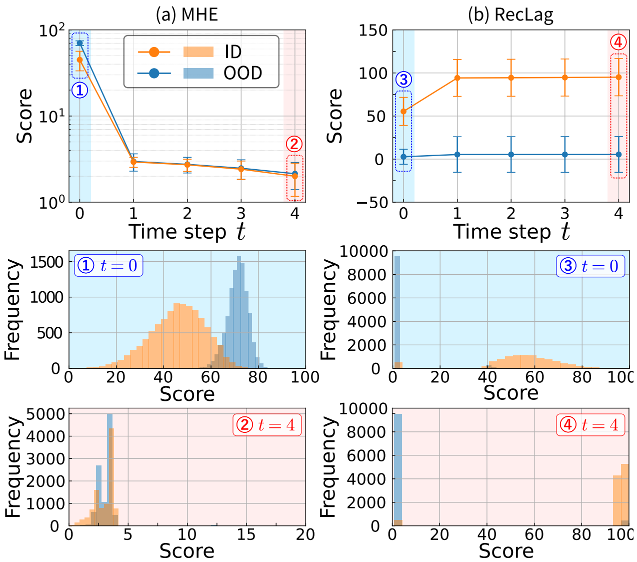


Figure 4: Comparison detection scores over time on LSUN-R. ResNet18 trained on CIFAR-10 was used.

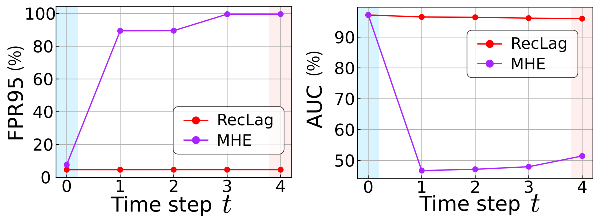


Figure 5: FPR95 ↓ and AUC ↑ over time on LSUN-R.

of our approach is robust against differences in ID samples.

Time Evolution. Figure 4 analyzes how the detection scores change as the dynamical system of MHNs evolves over time. With MHE, OOD samples have higher energy scores than ID samples at time $t = 0$; however, the scores decrease over time, making it almost impossible to distinguish between ID samples and OOD samples at a discrete time step of 4. In contrast, RecLag-based MHN can distinguish between ID samples and OOD samples even after the score converges, thereby maintaining OOD detection performance over time as shown in Figure 5. This demonstrates that our approach successfully managed OOD samples within the dynamical system of MHNs.

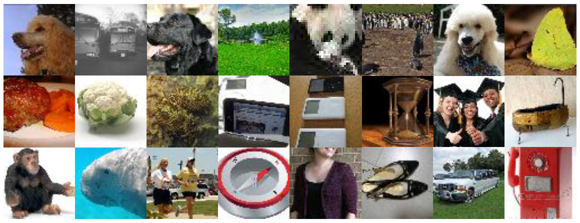


Figure 6: OOD samples incorrectly identified as ID samples. ID: CIFAR-100, OOD: TIN.

Visual Analysis. Figure 6 analyzes the OOD samples incorrectly identified as ID samples. As shown, images of animals, people, and foods were difficult to detect as OOD. This study focused on post-hoc OOD detection, but in the future, it would be interesting to simultaneously train MHNs and classification networks to further improve the performance.

7 Conclusion

We proposed the RecLag function, a specially designed Lagrangian to equip MHNs with OOD rejection functionality. In our method, the interaction matrix is optimized so as to compute probability densities, which are used to determine ID/OOD. Theoretically, RecLag-based MHNs reduces to vanilla MHNs when the ID memory strength is infinitely large; therefore, the proposed method is a natural extension of existing MHNs. Experiments on nine image datasets demonstrated the effectiveness of our approach, surpassing energy-based OOD detection methods.

Limitation. While this work introduced a new Lagrangian for memory neurons, the Lagrangian for feature neurons remains underexplored. Similar to previous works, we used the activation function g that takes the simplest form in Euclidean space, $g_i(v) = v_i$, because recent deep learning efforts often assume that the feature space is Euclidean. Investigating new Lagrangians in other non-linear feature spaces, such as spherical or hyperbolic space, might be promising.

Future Work. In future work, we will focus on generalizing RecLag for structured memory patterns such as hierarchical memory patterns. Applications to regression tasks are also intriguing. We believe this work has opened up new avenues for exploring the potential of MHNs.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP22H03642 and DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo). We thank Toshihiro Ota for fruitful discussion.

References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1): 147–169.

Agliari, E.; and De Marzo, G. 2020. Tolerance versus synaptic noise in dense associative memories. *The European Physical Journal Plus*, 135(11): 1–22.

Barra, A.; Beccaria, M.; and Fachechi, A. 2018. A new mechanical approach to handle generalized Hopfield neural networks. *Neural Networks*, 106: 205–222.

Chen, J.; Zhang, T.; Zheng, W.-S.; and Wang, R. 2024. TagFog: Textual Anchor Guidance and Fake Outlier Generation for Visual Out-of-Distribution Detection. In *Proc. AAAI Conference on Artificial Intelligence*, 1100–1109.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3606–3613.

Demircigil, M.; Heusel, J.; Löwe, M.; Upgang, S.; and Vermet, F. 2017. On a model of associative memory with huge storage capacity. *Springer Journal of Statistical Physics*, 168(2): 288–299.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Fürst, A.; Rumetshofer, E.; Tran, V.; Ramsauer, H.; Tang, F.; Lehner, J.; Kreil, D.; Kopp, M.; Klambauer, G.; Bitto-Nemling, A.; et al. 2021. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. *arXiv preprint arXiv:2110.11316*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hendrycks, D.; and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. International Conference on Learning Representations (ICLR)*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hofmann, C.; Schmid, S.; Lehner, B.; Klotz, D.; and Hochreiter, S. 2024. Energy-based Hopfield Boosting for Out-of-Distribution Detection. In *Proc. ICML AMHN Workshop*.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8): 2554–2558.

Hopfield, J. J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10): 3088–3092.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009, University of Toronto.

Krotov, D.; and Hopfield, J. 2018. Dense associative memory is robust to adversarial inputs. *Neural computation*, 30(12): 3151–3167.

Krotov, D.; and Hopfield, J. J. 2016. Dense Associative Memory for Pattern Recognition. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Krotov, D.; and Hopfield, J. J. 2021. Large Associative Memory Problem in Neurobiology and Machine Learning. In *Proc. International Conference on Learning Representations (ICLR)*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proc. International Conference on Learning Representations (ICLR)*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 21464–21475.

Millidge, B.; Salvatori, T.; Song, Y.; Lukasiewicz, T.; and Bogacz, R. 2022. Universal Hopfield Networks: A General Framework for Single-Shot Associative Memory Models. *arXiv preprint arXiv:2202.04557*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Ota, T.; Sato, I.; Kawakami, R.; Tanaka, M.; and Inoue, N. 2023. Learning with Partial Forgetting in Modern Hopfield Networks. In *Proc. International Conference on Artificial Intelligence and Statistics*, 6661–6673.

Ramsauer, H.; Schäfl, B.; Lehner, J.; Seidl, P.; Widrich, M.; Gruber, L.; Holzleitner, M.; Adler, T.; Kreil, D.; Kopp, M. K.; Klambauer, G.; Brandstetter, J.; and Hochreiter, S. 2021. Hopfield Networks is All You Need. In *Proc. International Conference on Learning Representations (ICLR)*.

Salvatori, T.; Song, Y.; Hong, Y.; Sha, L.; Frieder, S.; Xu, Z.; Bogacz, R.; and Lukasiewicz, T. 2021. Associative Memories via Predictive Coding. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 3874–3886.

Schäfl, B.; Gruber, L.; Bitto-Nemling, A.; and Hochreiter, S. 2021. Hopular: Modern Hopfield Networks for Tabular Data. *arXiv preprint arXiv:2206.00664*.

Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; and Klambauer, G. 2022. Improving Few-and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. *Journal of Chemical Information and Modeling*.

Shen, M.; Bu, Y.; Sattigeri, P.; Ghosh, S.; Das, S.; and Wornell, G. 2023. Post-hoc Uncertainty Learning Using a Dirichlet Meta-Model. In *Proc. AAAI Conference on Artificial Intelligence*, 9772–9781.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, 144–157.

Sun, Y.; and Li, Y. 2022. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Proc. European Conference on Computer Vision (ECCV)*, 691–708.

Tang, C.; and Salakhutdinov, R. R. 2013. Learning Stochastic Feedforward Neural Networks. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The iNaturalist Species Classification and Detection Dataset. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8769–8778.

Widrich, M.; Hofmarcher, M.; Patil, V. P.; Bitto-Nemling, A.; and Hochreiter, S. 2021. Modern Hopfield Networks for Return Decomposition for Delayed Rewards. In *Deep RL Workshop NeurIPS 2021*.

Widrich, M.; Schäfl, B.; Pavlović, M.; Ramsauer, H.; Gruber, L.; Holzleitner, M.; Brandstetter, J.; Sandve, G. K.; Greiff, V.; Hochreiter, S.; and Klambauer, G. 2020. Modern Hopfield Networks and Attention for Immune Repertoire Classification. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 18832–18845.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN Database: Large-scale Scene Recognition from Abbey to Zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 3485–3492. IEEE.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. Technical Report 1504.06755, arXiv preprint.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *Proc. British Machine Vision Conference (BMVC)*, 87.1–87.12.

Zhang, J.; Fu, Q.; Chen, X.; Du, L.; Li, Z.; Wang, G.; Liu, X.; Han, S.; and Zhang, D. 2023. Out-of-Distribution Detection based on In-Distribution Data Patterns Memorization with Modern Hopfield Energy. In *Proc. International Conference on Learning Representations (ICLR)*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6): 1452–1464.

Appendix A. Proof of Theorem 1

Theorem 1 shows that a point attractor exists at the origin of the feature space. We first define the attracting set A and its basin B, and then provide a proof of Theorem 1.

Definition 3. *Let $\mathcal{X}$ be a Hausdorff space and $\varphi(t, x)$ be a dynamical system on $\mathcal{X}$, where $t \in \mathbb{N}$ is a time index, $x \in \mathcal{X}$ is an initial point, and all $\varphi(t, \cdot) : \mathcal{X} \to \mathcal{X}$ are continuous functions. We say that a closed set $A \subset \mathcal{X}$ is an attracting set if there exists a neighborhood U of A that satisfies the following two conditions.*

(a) There exists T such that $\bigcap_{t \geq T} \{\varphi(t, x) : x \in U\} = A$.

(b) There exists T such that, for every neighborhood V of A, $t \geq T \Rightarrow \{\varphi(t, x) : x \in U\} \subset V$.

We define the basin B of attraction of A as $B = \bigcup_{t \geq 0} \{x : \varphi(t, x) \in U\}$.

Theorem 1. *Suppose that activation functions f and g in the dynamical system of Eqs. (1, 2) are given by the derivatives of RecLag L_H in Eq. (11) and the Lagrangian L_V in Eq. (7), respectively. For any interaction matrix $\xi \in \mathbb{R}^{N_H \times N_V}$, a trivial point attracting set $A = \{\mathbf{0}\}$ exists at the origin $\mathbf{0} \in \mathbb{R}^{N_V}$ in the feature space when $\gamma > N_H$ under the adiabatic limit $\tau_V = dt$.*

Proof. With RecLag, the activation function f_ν is given by

$$f_\nu(h) = \frac{\partial}{\partial h_\nu} \max\left(\log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right)^{\frac{1}{\beta}}, 0 \right) \tag{20}$$

$$= \chi\left(\log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right)^{\frac{1}{\beta}} \right) \cdot \frac{\partial}{\partial h_\nu} \log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right)^{\frac{1}{\beta}} \tag{21}$$

$$= \chi\left(\frac{1}{\beta} \log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right) \right) \cdot \frac{1}{\beta} \left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right)^{-1} \cdot \frac{\beta}{\gamma} \exp(\beta h_\nu) \tag{22}$$

$$= \chi\left(\log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right) \right) \cdot \left(\sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right)^{-1} \cdot \exp(\beta h_\nu) \tag{23}$$

$$= \chi\left(\log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp(\beta h_\mu) \right) \right) \cdot \mathrm{softmax}_\nu(\beta h) \tag{24}$$

where

$$\chi(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases}. \tag{25}$$

Under the adiabatic limit, *i.e.*, when the dynamics of memory neurons is changing rapidly, we have

$$h_\mu = \sum_{j=1}^{N_V} \xi_{\mu j} v_j. \tag{26}$$

Thus, we obtain

$$\text{Eq. (24)} = \chi\left(\log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp\left(\beta \sum_{j=1}^{N_V} \xi_{\mu j} v_j \right) \right) \right) \cdot \mathrm{softmax}_\nu\left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j \right) \tag{27}$$

$$= \chi(G(v)) \cdot \mathrm{softmax}_\nu\left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j \right). \tag{28}$$

where

$$G(v) = \log\left(\frac{1}{\gamma} \sum_{\mu=1}^{N_H} \exp\left(\beta \sum_{j=1}^{N_V} \xi_{\mu j} v_j \right) \right). \tag{29}$$

The differential equation in Eq. (1) is then written by

$$\begin{aligned}
\tau_V \frac{dv_i(t)}{dt} &= \sum_{\mu=1}^{N_H} \xi_{i\mu} f_\mu(h(t)) - v_i(t) && (30)\\
&= \sum_{\mu=1}^{N_H} \xi_{i\mu} \chi(G(v(t))) \operatorname{softmax}_\mu \left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j(t) \right) - v_i(t) && (31)\\
&= \chi(G(v(t))) \sum_{\mu=1}^{N_H} \xi_{i\mu} \operatorname{softmax}_\mu \left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j(t) \right) - v_i(t). && (32)
\end{aligned}$$

To derive the update rule, we consider the first-order Taylor approximation

$$v_i(t + \Delta t) = v_i(t) + \frac{dv_i(t)}{dt} \Delta t, \tag{33}$$

where Δt is a small time step. From Eq. (32), we have

$$v_i(t + \Delta t) = v_i(t) + \frac{\Delta t}{\tau_V} \left(\chi(G(v(t))) \sum_{\mu=1}^{N_H} \xi_{i\mu} \operatorname{softmax}_\mu \left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j(t) \right) - v_i(t) \right). \tag{34}$$

Therefore, when $\tau_V = \Delta t$, we have

$$v_i(t + \Delta t) = \chi(G(v(t))) \sum_{\mu=1}^{N_H} \xi_{i\mu} \operatorname{softmax}_\mu \left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j(t) \right). \tag{35}$$

This yields the update rule with discrete time steps $k \in \mathbb{N}$ as follows:

$$v_i^{(k+1)} = \chi \left(G(v^{(k)}) \right) \sum_{\mu=1}^{N_H} \xi_{i\mu} \operatorname{softmax}_\mu \left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} v_j^{(k)} \right). \tag{36}$$

Finally, we show that $A = \{\mathbf{0}\}$ is an attracting set for every fixed ξ. Suppose that $\mathcal{X} = \mathbb{R}^{N_V}$ is the feature space. We consider the Euclidean distance $d(x, y) = \|x - x'\|_2$ between two points $x, x' \in \mathcal{X}$. Clearly, with the topology induced by the open balls

$$U_\epsilon(x) = \{x' \in \mathcal{X} : d(x, x') < \epsilon\} \;\; (\epsilon > 0), \tag{37}$$

the space $\mathcal{X}$ is a Hausdorff space. The dynamic system $\varphi(k, x)$ is then given by

$$\varphi(k, x) = \begin{cases} x & (k = 0) \\ \chi\left(G(\varphi(k, x))\right) \sum_{\mu=1}^{N_H} \xi_{i\mu} \operatorname{softmax}_\mu \left(\beta \sum_{j=1}^{N_V} \xi_{\cdot j} \varphi_j(k, x) \right) & (k > 0) \end{cases}. \tag{38}$$

Below, we show that the two conditions (a) and (b) in Definition 3 are satisfied.
Proof of (a). Let $U = U_\epsilon(\mathbf{0})$ be an open ball with

$$\epsilon = \frac{1}{N_V \beta \, \Xi} \log \frac{\gamma}{N_H}, \quad \Xi = \max_{\mu, j} |\xi_{\mu j}|, \tag{39}$$

where $\gamma > N_H$. For every $x \in U$, we have

$$
\begin{aligned}
G(\varphi(0,x)) &= \log\left(\frac{1}{\gamma}\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}x_j\right)\right) && (40)\\
&= \log\left(\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}x_j\right)\right) - \log\gamma && (41)\\
&\leq \log\left(N_H \max_\mu \exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}x_j\right)\right) - \log\gamma && (42)\\
&\leq \max_\mu\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}x_j\right) - \log\frac{\gamma}{N_H} && (43)\\
&\leq N_V\beta\max_\mu\max_j\left(\xi_{\mu j}x_j\right) - \log\frac{\gamma}{N_H} && (44)\\
&\leq N_V\beta\;\Xi\;\|x\|_\infty - \log\frac{\gamma}{N_H} && (45)\\
&\leq N_V\beta\;\Xi\;\|x\|_2 - \log\frac{\gamma}{N_H} && (46)\\
&< N_V\beta\;\Xi\;\epsilon - \log\frac{\gamma}{N_H} && (47)\\
&= 0. && (48)
\end{aligned}
$$

Thus, when $T = 1$, we have

$$\bigcap_{t\geq T}\{\varphi(t,x) : x \in U_\epsilon\} = \bigcap_{t\geq T}\{\mathbf{0}\} = A. \tag{49}$$

Proof of (b). Suppose that $V = \{x : d(x,\mathbf{0}) < \epsilon'\}$ is a neighborhood of A. With the open ball $U = U_\epsilon(\mathbf{0})$ defined by Eq. (39) and when $T = 1$, we have

$$\{\varphi(t,x) : x \in U\} = A \subset V, \tag{50}$$

when $t \geq T$.

This shows that A is an attracting set when $\gamma > N_H$ for every fixed ξ. □

Appendix B. Proof of Theorem 2

Theorem 2. *Let v_A and v_B be feature neurons of a vanilla MHN and a RecLag-based MHN, respectively. Suppose that $v_\mathrm{A}^{(0)} = v_\mathrm{B}^{(0)}$. For every $\epsilon > 0$, a small $\gamma > 0$ exists such that $\sup_k \|v_\mathrm{A}^{(k)} - v_\mathrm{B}^{(k)}\|_2 < \epsilon$.*

Proof. The update rules for v_A and v_B are given by

$$
\begin{aligned}
v_{\mathrm{M},i}^{(k+1)} &= \sum_{\mu=1}^{N_H}\xi_{i\mu}\,\mathrm{softmax}_\mu\left(\beta\sum_{j=1}^{N_V}\xi_{\cdot j}v_{\mathrm{M},j}^{(k)}\right), && (51)\\
v_{\mathrm{R},i}^{(k+1)} &= \chi\left(G(v_\mathrm{B}^{(k)})\right)\sum_{\mu=1}^{N_H}\xi_{i\mu}\,\mathrm{softmax}_\mu\left(\beta\sum_{j=1}^{N_V}\xi_{\cdot j}v_{\mathrm{R},j}^{(k)}\right). && (52)
\end{aligned}
$$

Let $0 < \delta < 1$ be a small constant and

$$\gamma = \delta\min_k\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}v_{\mathrm{M},j}^{(k)}\right). \tag{53}$$

When $v_{\mathrm{B}}^{(k)} = v_{\mathrm{A}}^{(k)}$, we have

$$\chi(G(v_{\mathrm{B}}^{(k)})) = \chi\left(\log\left(\frac{1}{\gamma}\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}v_{\mathrm{R},j}^{(k)}\right)\right)\right) \tag{54}$$

$$= \chi\left(\log\left(\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}v_{\mathrm{M},j}^{(k)}\right)\right) - \log\gamma\right) \tag{55}$$

$$= \chi\left(\log\frac{\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}v_{\mathrm{M},j}^{(k)}\right)}{\min_k\sum_{\mu=1}^{N_H}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}v_{\mathrm{M},j}^{(k)}\right)} - \log\delta\right) \tag{56}$$

$$= 1, \tag{57}$$

and we have

$$\|v_{\mathrm{A}}^{(k+1)} - v_{\mathrm{B}}^{(k+1)}\|_2^2 = \sum_{i=1}^{N_V}\left(\left(1-\chi\left(G(v_{\mathrm{B}}^{(k)})\right)\right)\sum_{\mu=1}^{N_H}\xi_{i\mu}\,\mathrm{softmax}_\mu\left(\beta\sum_{j=1}^{N_V}\xi_{\cdot j}v_{\mathrm{M},j}^{(k)}\right)\right)^2 \tag{58}$$

$$= \left(1-\chi\left(G(v_{\mathrm{B}}^{(k)})\right)\right)^2\sum_{i=1}^{N_V}\left(\sum_{\mu=1}^{N_H}\xi_{i\mu}\,\mathrm{softmax}_\mu\left(\beta\sum_{j=1}^{N_V}\xi_{\cdot j}v_{\mathrm{M},j}^{(k)}\right)\right)^2 \tag{59}$$

$$= 0. \tag{60}$$

This assumption gives us $v_{\mathrm{B}}^{(0)} = v_{\mathrm{A}}^{(0)}$, and thus, for every $\epsilon > 0$,

$$\sup_k \|v_{\mathrm{A}}^{(k)} - v_{\mathrm{B}}^{(k)}\|_2 = 0 < \epsilon \tag{61}$$

□

Appendix C. Proof of Theorem 3

Theorem 3. *The basin $B_0 = \{v : G(v) < 0\}$ is identical to the set of points that have low probability density values, i.e., a threshold δ exists such that*

$$B_0 = \{x : p_H(X = x) < \delta\}. \tag{62}$$

Proof. With the joint probability distribution $p_H(X = x, M = \mu)$ given by Definition 2, the marginal distribution $p_H(X = x)$ is given by

$$p_H(X = x) = \sum_{\mu=1}^{N_H}\frac{1}{Z}\exp\left(\beta\sum_{j=1}^{N_V}\xi_{\mu j}x_j\right) = \frac{\gamma}{Z}\exp(G(x)). \tag{63}$$

Therefore, for fixed values of ξ and γ, we have

$$\delta = \frac{\gamma}{Z} \tag{64}$$

satisfying

$$B_0 = \{v : G(v) < 0\} = \{v : \exp(G(v)) < 1\} = \{x : p_H(X = x) < \delta\}. \tag{65}$$

This shows that the basin is a set of data samples that have a probability density lower than δ. □

Appendix D. Implementation details

Three image classification networks were used: ResNet-18 (He et al. 2016), ResNet-34 (He et al. 2016), and WideResNet40-2 (Zagoruyko and Komodakis 2016). Each network was trained on an ID dataset using cross-entropy loss for 200 epochs with an SGD momentum optimizer. The initial learning rate was set to 0.1, and it was decayed by a factor of 0.1 at 100 and 150 epochs. The batch size was set to 128. Random cropping and horizontal flipping were used to augment the training images. The dimension of the output representation was 512, and thus the number of feature neurons was set as $N_V = 512$. During the training of the interaction matrix, normalization was applied to the output representations so that the L2 norms are 10.0. The interaction matrix was trained for 100 epochs by following the training method for SFNN proposed by Tang and Salakhutdinov (2013), where the number of memory neurons is set as $N_H = 250$, the inverse temperature parameter β is set to 5.0, and the number of samples for Monte Carlo approximation is set to 5. The objective function was computed using the input features as targets, as described in Eq. (15). The OOD datasets were prepared following Shen et al. (2023), with all images resized to 32×32. We used the official implementation of Energy (Liu et al. 2020), ReAct (Sun, Guo, and Li 2021), MHE and SHE (Shen et al. 2023) to report their results.