Shaping Latent Geometry with Noise-Injected Hopfield Dynamics

Wooyul Jung[1], Youngseok Joo[1],
Dohyun Yu[2], Suhyung Choi[1], Byoung−Tak Zhang[1]

[2]College of Liberal Studies, Seoul National University

[1]Interdisciplinary Program in Artificial Intelligence, Seoul National University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Motivation.
Representation learning often incidentally acquires geometric structure e.g., class-mean collapse late in training or consistent eigenspectrum decay across biological and artificial systems. Such phenomena suggest that useful latent geometry emerges spontaneously but is not explicitly controlled.
Prior work explicitly shapes latent spaces by (i) controlling local sensitivity and (ii) contrastive objectives, and (iii) explicit priors. Yet these approaches do not exploit attractor dynamics as a direct mechanism for organizing representations.

Approach.
We introduce noise-injected Hopfield retrieval layers inserted between network layers to explicitly sculpt latent spaces. By tuning two parameters # of retrieval iterations, and inverse temperature, we control how strongly representations contract, interpolating from discrete attractors to smooth continuous manifolds.

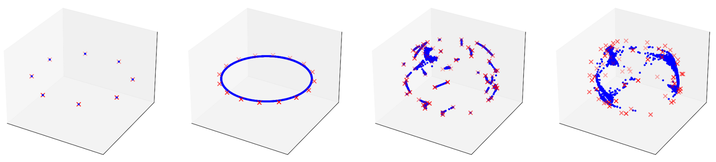
Effect.
These layers organize intermediate features into stable clusters, naturally improving noise robustness and overall representation quality.

Hopfield Retrieval layer

Modern Hopfield Dynamics.
We formulate a stochastic Hopfield update that introduces noise-driven attractor dynamics into latent representations. Given a set of unit-norm memory vectors $\{e_i\}_{i=1}^{P}$with $\|e_i\|_2 = 1$ and memory matrix $C = [e_1, \cdots, e_P] \in \mathbb{R}^{D\times P}$, the query vector $z_t \in \mathbb{R}^D$is updated as follows. At each iteration t, Gaussian noise $\xi_t \sim \mathcal{N}(0, \sigma^2 I_D)$ is injected and the perturbed vector is normalized to unit norm. This process corresponds to gradient descent on the Modern Hopfield energy function.

$$z_{t+1} = C\ \mathrm{softmax}\left(\beta\, C^\top \hat{z}_t\right). \quad \hat{z}_t = \frac{z_t + \xi_t}{\|z_t + \xi_t\|_2}.$$

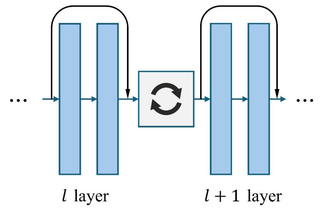
$$E_{\mathrm{MHN}}(z) = -\frac{1}{\beta}\log\sum_{i=1}^{P}\exp\left(\beta\, z^\top e_i\right).$$



P=8, β=100 P=16, β=10 P=64, β=100 P=64, β=10

Noise-Injected Hopfield Retrieval (NIHR).
We implement these dynamics as NIHR, a general and fully differentiable module that can be inserted between arbitrary network layers.
NIHR performs iterative refinement, pulling representations toward Hopfield energy minima and thereby enforcing local geometric structure in the latent space.

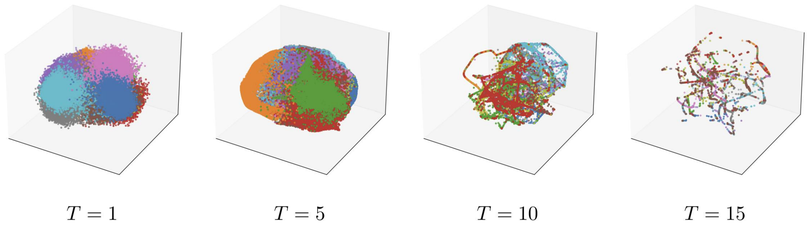
$$z_{t+1} = \mathcal{H}_\sigma(z_t), \quad t = 0, \ldots, T-1.$$



Structural evolution under Autoencoder framework

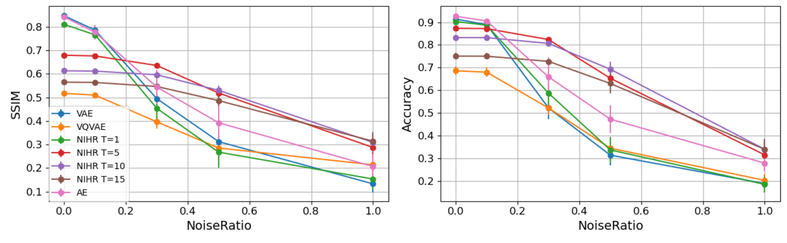
We insert NIHR between the encoder and decoder to study how Hopfield retrieval reshapes latent organization. This allows us to analyze how structured refinement influences reconstruction robustness and representation quality.

PCA Visualization.
we project latent codes onto their top principal components after T = 1, 5, 10, 15 refinement steps. As T increases, diffuse latents progressively contract into clear, stable clusters, demonstrating how NIHR sharpens latent geometry over iterations.



$T = 1$ $T = 5$ $T = 10$ $T = 15$

Robustness under Gaussian Noise.
We conduct controlled experiments on MNIST with additive Gaussian noise of varying standard deviation σ to simulate input degradation. Two metrics are used: the structural similarity index (SSIM) for reconstruction quality, and linear probing accuracy for assessing the discriminative power of learned representations.



Evaluation under Gaussian input corruption. Left: SSIM scores for reconstruction fidelity. Right: Linear probing accuracy for representation quality. Models include AE, VAE, VQ-VAE, and NIHR with $T \in \{1, 5, 10, 15\}$.

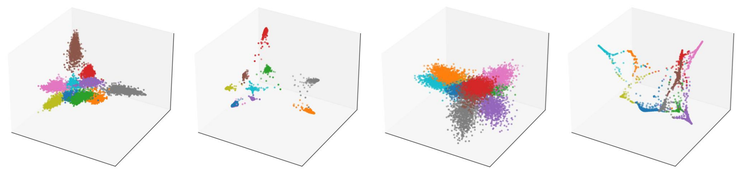
Robust Classification under Corruptions

We integrate NIHR into ResNet-18 as a latent refinement layer to improve robustness under input corruptions. The insertion point is chosen empirically: after the first residual block for MNIST and after the last residual block for CIFAR-10.

Evaluation under Corruption.
Robustness is measured on MNIST-C and CIFAR-10-C, while all models are trained only on clean data with no augmentation or corruptions. The resulting classification accuracies on corrupted datasets are summarized below.

Dataset	Model	Clean (0)	s1	s2	s3	s4	s5	Overall
MNIST-C	ResNet	99.47 ± 0.05	–	–	–	–	–	72.13 ± 2.19
	ResNet+NIHR	$\mathbf{99.50 \pm 0.02}$	–	–	–	–	–	$\mathbf{80.56 \pm 2.2}$
CIFAR-C	ResNet	$\mathbf{84.78 \pm 0.18}$	$\mathbf{75.69 \pm 0.68}$	$\mathbf{70.39 \pm 0.92}$	64.99 ± 1.03	58.51 ± 1.21	48.98 ± 1.04	$\mathbf{63.71 \pm 0.96}$
	ResNet+NIHR	81.81 ± 0.64	73.93 ± 1.25	69.28 ± 1.29	$\mathbf{64.58 \pm 1.38}$	$\mathbf{58.73 \pm 1.38}$	$\mathbf{51.00 \pm 1.94}$	63.50 ± 1.31
CIFAR-C (Gaussian)	ResNet	$\mathbf{84.78 \pm 0.18}$	70.10 ± 1.74	55.40 ± 2.93	41.91 ± 4.14	36.36 ± 4.63	31.93 ± 4.86	47.14 ± 3.53
	ResNet+NIHR	81.81 ± 0.64	$\mathbf{72.81 \pm 2.30}$	$\mathbf{63.83 \pm 3.41}$	$\mathbf{54.47 \pm 4.21}$	$\mathbf{49.75 \pm 4.59}$	$\mathbf{45.99 \pm 4.75}$	$\mathbf{57.37 \pm 3.63}$

PCA Visualization.
We visualized how the intermediate representations of ResNet-18 evolve under latent attractor dynamics and observed that they become increasingly coherent as they are pulled toward the attractor.



MNIST w/o NIHR MNIST w/ NIHR CIFAR-10 w/o NIHR CIFAR-10 w/ NIHR

Contributions

NIHR is a general, differentiable module that integrates into diverse architectures without modification. Its noise-driven Hopfield retrieval forms clear attractor basins that organize intermediate representations, improving robustness to noise and input corruptions. We also plan to extend this work to RNNs and explore the connections among Hopfield networks, Transformers, RNNs, and other ANNs in terms of attractor dynamics.

References

Arash Vahdat, Karsten Kreis, and Jan Kautz (2021). “Score-based generative modeling in latent space.” In: Advances in Neural Information Processing Systems 34, pp. 11287–11302.

Hubert Ramsauer et al. (2020). “Hopfield networks is all you need.” In: arXiv preprint arXiv:2008.02217.