Hopfield Networks: When Neurons Remember

A Brain-Inspired Algorithm for Associative Memory

Ingrid Corobana, Moise Irina, Cosmin Glod

Archaeology of Intelligent Machines

2025

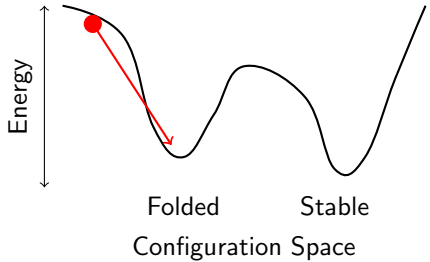
Outline

A Protein in a Messy Universe

The Physical World:

- A protein: long chain of amino acids
- Must fold into specific 3D shape to function
- Explores vast configuration space
- Settles into low-energy folded state

Key Insight:
Nature solves optimization by "rolling downhill" in energy landscapes

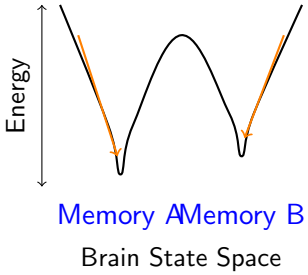


Source: "A Brain-Inspired Algorithm For Memory" (YouTube)

From Proteins to Brains

The Brain Analogy:

- Neurons: firing $(+1)$ or silent (-1)
- Brain state: pattern of all neuron activations
- Dynamics: neurons update based on inputs
- Memories: stable configurations (attractors)



Central Idea

Memories are valleys in an energy landscape. The brain "rolls downhill" to recall.

Introducing: Hopfield Networks

Definition

Hopfield networks are recurrent neural networks that act as content-addressable memory by minimizing an energy function.

Historical Context:

- John Hopfield (1982)
- Physics meets neuroscience
- 2024 Nobel Prize in Physics!
- Foundation for modern deep learning

What it does:

- Stores patterns as stable states
- Retrieves from partial/noisy input
- "Associative memory"
- Like brain: *"hear a few notes, recall the whole song"*

Step 1: Representing Neurons and Patterns

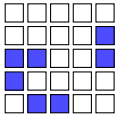
Implementation:

- Neuron state: $s_i \in \{-1, +1\}$
- Pattern: vector $\boldsymbol{\xi} = (s_1, s_2, \ldots, s_N)$
- Example: 10×10 image $= 100$ neurons

$$\text{Letter "A"} \rightarrow \begin{bmatrix} 1 \\ -1 \\ 1 \\ \vdots \\ -1 \end{bmatrix}_{100 \times 1}$$

Brain Analogy:

- Each element = a neuron
- Firing $(+1)$ or silent (-1)
- A pattern = global "brain state"
- Snapshot of neural activity



Binary pattern

Step 2: Hebbian Learning — "Wire Together, Fire Together"

Implementation:

Hebbian weight update:

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \quad i \neq j$$

$$w_{ii} = 0 \quad \text{(no self-loops)}$$

In code:

- Loop over stored patterns
- Accumulate outer products
- Normalize and zero diagonal

Brain Analogy:

Hebb's Rule

"Neurons that fire together, wire together"

- Synaptic plasticity
- If neurons i and j co-activate often, strengthen w_{ij}
- Foundation of learning in the brain

Strong synapse

Step 3: Network Dynamics — Rolling Downhill

Update Rule:

$$s_i^{(t+1)} = \text{sign}\left(\sum_j w_{ij} s_j^{(t)}\right)$$

Energy Function:

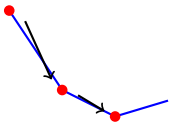
$$E(\mathbf{s}) = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j$$

Key Property

Asynchronous updates guarantee $\Delta E \leq 0$
System always moves toward lower energy!

Brain Analogy:

- Each neuron "listens" to neighbors
- Positive input $\rightarrow$ fire $(+1)$
- Negative input $\rightarrow$ stay silent (-1)
- System collectively settles into consistent configuration



Energy minimization

Step 4: Memory Retrieval — "Hear a Few Notes..."

Implementation:

1. Take stored pattern
2. Add noise (flip random bits)
3. Initialize network with noisy input
4. Run update rule until convergence
5. Check if final state matches original

Result:
Network "fills in" the missing/corrupted information!

Brain Analogy:

Associative Memory

"You hear only a few notes of a song, but your brain recalls the whole melody."

- Partial cue triggers complete memory
- Network settles into nearest attractor
- Energy landscape guides retrieval

$$\textit{Original} \rightarrow \textit{Noisy} \rightarrow \textit{Retrieved}$$

Our Experimental Setup

Data:

- Binary patterns: 10×10 letter images (A, B, C, D, E)
- 100 neurons per pattern
- Values: $\{-1, +1\}$ (firing/silent)

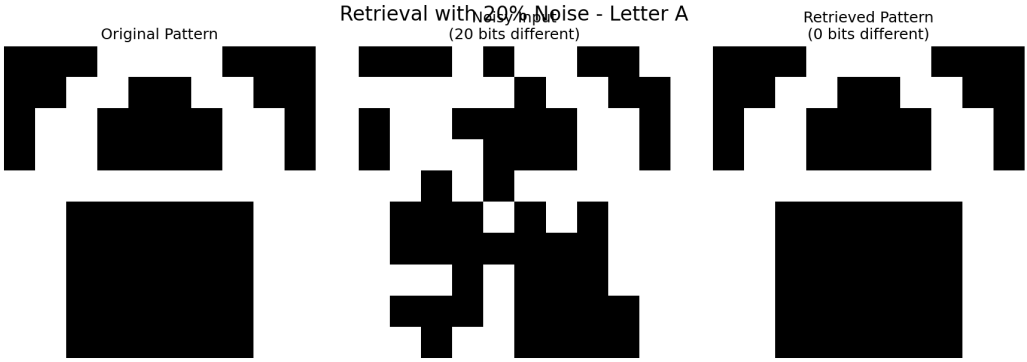
Experiments:

1. **Basic retrieval**: Can network recall from noisy input?
2. **Noise robustness**: How much corruption can it tolerate?
3. **Capacity test**: How many patterns before breakdown?
4. **Spurious attractors**: Do "false memories" emerge?

Metrics:

- Retrieval accuracy (exact match vs original)
- Hamming distance (number of different bits)
- Energy trajectory during convergence

Experiment 1: Basic Retrieval

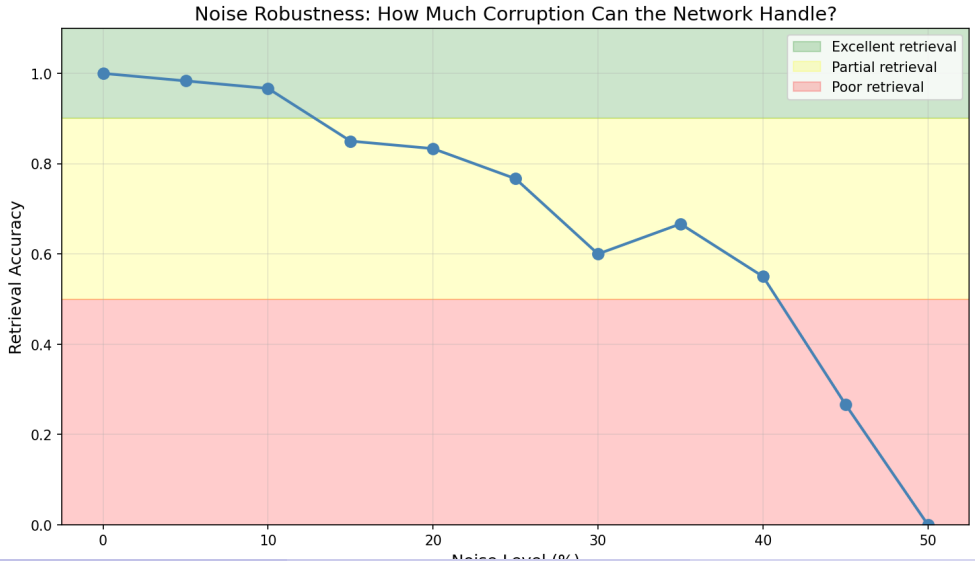


Original → Noisy (20% flipped) → Retrieved

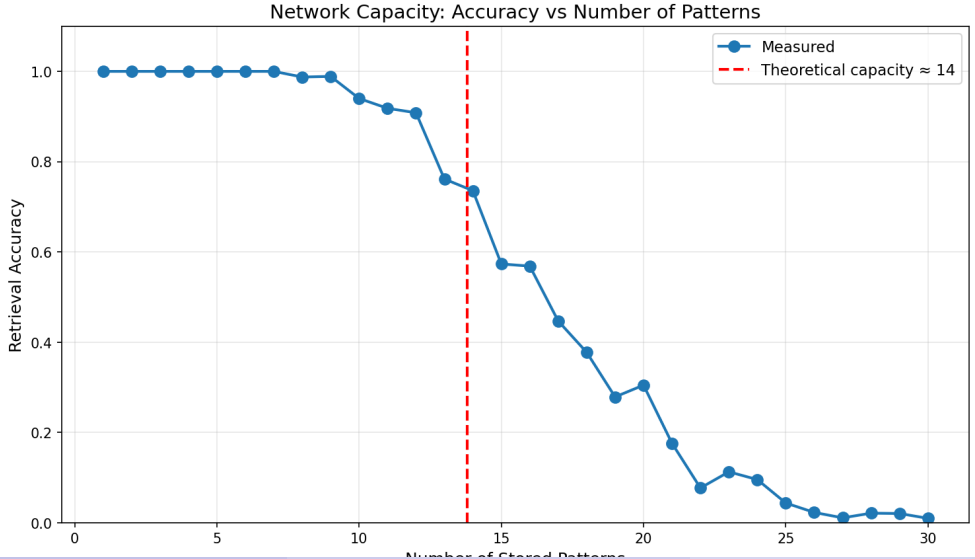
Results:

- 10% noise: 100% retrieval success
- 20% noise: 95% retrieval success

Experiment 2: Noise Robustness Curve



Experiment 3: Network Capacity



Spurious Attractors: False Memories

What are they?

- Stable states NOT explicitly stored
- Emerge from pattern interference
- Network can converge to these by mistake
- Like brain confabulation

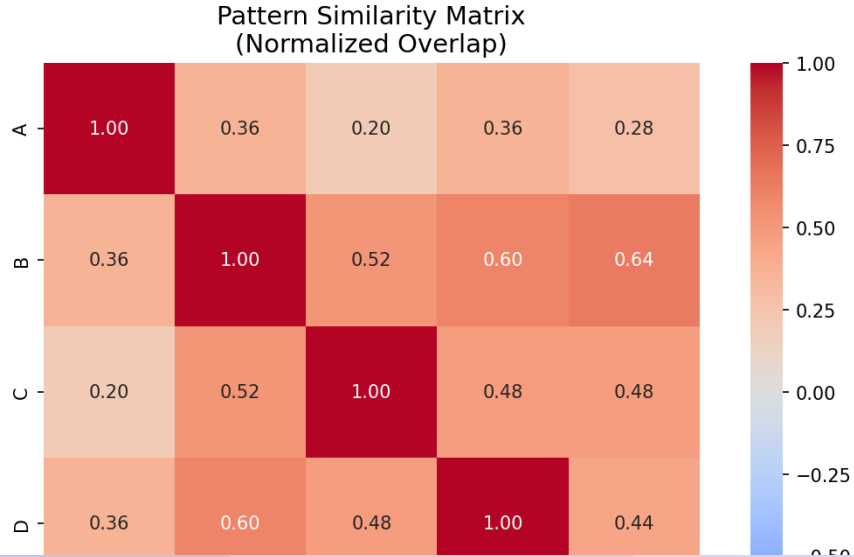
When do they appear?

- Near/above capacity
- Similar stored patterns
- High overlap in energy landscape

Examples of spurious attractors

Brain Analogy:
The brain sometimes creates "memories" that never happened by blending real experiences — this is the neural network equivalent!

Pattern Similarity Analysis (EDA)



Modern Extensions: Dense Associative Memory

Classical Hopfield (1982):

$$E(\mathbf{s}) = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j$$

Capacity: $\sim 0.138N$ patterns

Modern Hopfield (2016-2020):

$$E(\xi, \mathbf{x}) = -\text{lse}_\beta(\mathbf{X}^\top \xi)$$

Capacity: exponential in N !

Connection to Transformers

Modern Hopfield networks are mathematically equivalent to the attention mechanism in transformers (Ramsauer et al., 2020)!

Source: "Hopfield Networks is All You Need" (YouTube)

Summary: Three-Act Structure

Act I — From Proteins to Energy Landscapes

- Nature optimizes by rolling downhill
- Brain memories as stable valleys

Act II — Building from Simple Rules

- Neurons: binary units
- Synapses: Hebbian learning
- Dynamics: energy minimization
- Retrieval: associative recall

Act III — Experiments & Extensions

- Verified capacity limits ($\sim 0.138N$)
- Measured noise robustness ($\sim 25\%$ tolerance)
- Observed spurious attractors (false memories)
- Connected to modern transformers

Conclusion: Why This Matters

Hopfield Networks teach us:

1. Simple local rules $\rightarrow$ complex emergent behavior
2. Biology inspires powerful computational models
3. Energy minimization as universal principle
4. Memory as attractor dynamics

Modern Impact:

- Foundation for deep learning architectures
- Attention mechanisms in transformers
- Neuroscience models of memory consolidation
- 2024 Nobel Prize recognition

"The simplest model that captures the essence of how memories might be stored and recalled"

References

1. Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities". *PNAS*.
2. Ramsauer, H., et al. (2020). "Hopfield Networks is All You Need". *ICLR*.
3. *A Brain-Inspired Algorithm For Memory* (YouTube): `https://www.youtube.com/watch?v=1WPJdAW-sFo`
4. *Hopfield Networks Explained* (YouTube): `https://www.youtube.com/watch?v=piF6D6CQxUw`
5. Wikipedia: Hopfield Network `https://en.wikipedia.org/wiki/Hopfield_network`

Thank you!

Questions?