Associative Memory as the Core of Intelligence in Technology and Evolution

Tomaso Poggio

Center for Brains, Minds, and Machines, CSAIL, McGovern Institute, MIT

January 18, 2026

Abstract

A rarely acknowledged but deep conceptual continuity runs from the earliest associative memories in theoretical neuroscience to the core mechanisms of modern artificial intelligence. This paper argues that associative memory is not an obsolete concept but the central computational primitive underlying both biological and artificial intelligence. We trace a continuous line connecting (i) early formal models of associative memories in the 1960s and 70s; (ii) the role of monosynaptic reflexes and episodic binding in the evolution of animal intelligence; (iii) the emergence of Kernel Machines, radial basis function (RBF) and hyperBF networks as powerful nonlinear associative memories; (iv) hippocampal indexing and scaffolding models; and (v) the attention mechanism in modern transformers, which can be interpreted mathematically as a high-dimensional associative memory (a form of learnable, metric-based hyper-RBF network). This unified perspective suggests that associative memory is not merely a component of intelligence but its origin and structural foundation.

1 Introduction

Associative memory—the ability to bind one pattern to another and later retrieve the bound information from a full or partial cue—is among the earliest and most fundamental ideas in neural computation. Yet, despite its centrality, it is often treated as a historical curiosity: a precursor to modern deep networks rather than a principle that continues to shape state-of-the-art AI systems and the evolution of biological intelligence.

The core claim of this review is that associative memory is *the primitive operation from which intelligence emerges*. Every major leap in both biological evolution and artificial intelligence can be understood as a refinement, extension, or composition of associative mechanisms. We argue that:

(a) the simplest adaptive behaviors in early animals—monosynaptic reflexes linking sensory stimuli to motor actions—are the earliest form of associative memory;

(b) the hippocampal formation implements a hybrid associative system (sparse indexing, autoassociation, and heteroassociation) that provides the "scaffold" enabling cortex to learn structured knowledge;

(c) classical models of associative memory in the 1960s–1990s (Longuet-Higgins, Willshaw, Kohonen, Palm, Poggio) led to kernel and radial basis function (RBF) / hyperBF networks that realize flexible nonlinear associative mappings;

(d) the attention operation in transformers is *itself* a form of associative memory: a learned, differentiable, high-dimensional mapping between queries and keys that retrieves values via similarity, and is formally equivalent to a normalized hyperBF network;

(e) modern deep learning works so well not despite but *because* it rediscovered this fundamental associative mechanism in a powerful compositional form.

Recognizing this lineage reveals a unifying perspective on intelligence. Associative memories constitute the bridge between:

- the earliest nervous systems capable of learning;
- the hippocampus as the biological engine of episodic memory and scaffolding;
- classical theoretical models of neural computation (including RBF / hyperBF architectures);
- and the architecture of modern large-scale AI systems such as transformers.

The argument is that associative memory is the *ancestor* of modern deep learning in both an evolutionary and a technological sense. More strongly, it is the computational foundation upon which intelligent systems—natural and artificial—are built.

In the sections that follow, we trace this thread from biology to computation, from early neural models to contemporary AI, and from simple reflex arcs to large-language models. Far from being a relic of early theoretical neuroscience, associative memory is the conceptual core around which the modern understanding of intelligence is beginning to coalesce.

2 Historical Background

The modern concept of associative memory—a system that stores pairs of patterns (x^μ, y^μ) such that a later presentation of x^μ retrieves y^μ—can be traced to the theoretical neuroscience of the late 1960s and early 1970s. The early models were explicitly concerned with how neural circuits could store large sets of associations efficiently, robustly, and with simple learning rules. Although the terminology has evolved, the mathematical structures developed in this period anticipate the key computational mechanisms of modern attention-based architectures.

We review the major foundational models, emphasizing their mathematical form and their conceptual relationship to today's deep networks.

2.1 The Willshaw–Longuet-Higgins Associative Memory (1969–1970)

In one of the earliest formal treatments of associative memory, Willshaw and Longuet-Higgins considered binary patterns $x^\mu, y^\mu \in \{0,1\}^N$ and proposed a sparse heteroassociative memory with the learning rule

$$W_{ij} = \bigvee_{\mu=1}^{P} x_i^\mu y_j^\mu, \tag{1}$$

where $\vee$ denotes the logical OR. At retrieval time, the system computes

$$\hat{y}_j = \bigvee_{i:x_i=1} W_{ij}, \tag{2}$$

effectively implementing a binary OR over stored associations.

This model is notable for its:

- *sparse, binary coding*, which allows high capacity,
- *one-shot learning*,
- *feedforward retrieval* with no dynamics.

While extremely simple, it anticipates the key components of key–value attention: a stored collection of (x^μ, y^μ) pairs, and a retrieval process that aggregates over all stored pairs that match the input.

2.2 Kohonen's Linear Associative Memory (1972–1977)

Teuvo Kohonen introduced a linear algebraic formulation of associative memory, which stores input–output pairs (x^μ, y^μ) via the correlation matrix

$$W = \sum_{\mu=1}^{P} y^\mu (x^\mu)^\top. \tag{3}$$

Retrieval is simply

$$\hat{y} = Wx. \tag{4}$$

This model is a direct mathematical precursor to:

- kernel regression,
- radial basis function networks,
- and modern attention (as we discuss later).

Indeed, the attention mechanism

$$\mathrm{Attn}(q, K, V) = \mathrm{softmax}(qK^\top)V$$

reduces to the Kohonen memory when the softmax is replaced by an identity (or threshold) function and K and V correspond to stored x^μ and y^μ.

2.3 Nonlinear Associative Recall (1975)

An important but often underappreciated contribution in the early development of associative memories is due to Poggio (1973, 1975), who formulated the associative recall problem as a general *operator estimation problem* between two finite sets of input–output vectors. In contrast to the earlier binary and linear models, this approach introduced *nonlinear polynomial associative mappings*, anticipating the flexibility of kernel methods, RBF networks, and ultimately the metric-based associative mechanisms underlying modern attention systems.

In Poggio's formulation, one is given two matrices of paired data

$$X \in \mathbb{R}^{m\times n}, \qquad Y \in \mathbb{R}^{r\times n},$$

whose columns are input vectors x^j and output vectors y^j respectively. The goal is to find an operator Θ such that

$$Y \approx \Theta(X), \tag{5}$$

in the least-squares sense. The key result of the 1975 paper is that the optimal nonlinear associative mapping can be written as a polynomial expansion of degree k:

$$Y = L_0 + L_1(X) + L_2(X, X) + \cdots + L_k(X, \ldots, X), \tag{6}$$

where each L_ℓ is an ℓ-linear symmetric operator. Written explicitly, the expansion takes the form

$$Y_{ij} = (L_0)_{ij} + \sum_{\alpha_1}(L_1)_{i\alpha_1}X_{\alpha_1 j} + \sum_{\alpha_1,\alpha_2}(L_2)_{i\alpha_1\alpha_2}X_{\alpha_1 j}X_{\alpha_2 j} + \cdots + \sum_{\alpha_1\cdots\alpha_k}(L_k)_{i\alpha_1\cdots\alpha_k}\prod_{t=1}^{k}X_{\alpha_t j}. \tag{7}$$

A central theorem of the paper gives the optimal L_1 as:

$$L_1 = YX^\dagger, \tag{8}$$

where $X^\dagger$ is the Penrose Moore–Penrose generalized inverse. This recovers Kohonen's linear associative memory as a special case.

The nonlinear corrections are computed iteratively. Defining the residual after removing the approximation up to order $k-1$ as

$$E_{k-1} = Y - \sum_{\ell=0}^{k-1} L_\ell(X, \ldots, X),$$

the optimal kth-order correction is given by:

$$L_k = E_{k-1}C_k^\dagger, \tag{9}$$

where C_k is the "k-way" lifted tensor of inputs defined by

$$(C_k)_{\alpha_1\cdots\alpha_k, j} = \prod_{t=1}^{k} X_{\alpha_t j}.$$

Several ideas emerged from this work:

- **Polynomial associative mappings** provide universal approximation capability (via the Stone–Weierstrass theorem).
- **Linear associative memory** (Kohonen, Willshaw) emerges as the degree–1 case.
- The approach introduces the notion of *generalized correlations* and *generalized convolutions* anticipating the structure of later kernel methods and RBF networks (polynomial kernel discussed extensively by Vapnik were derived in Poggio, 1973).
- The method works for arbitrary real-valued data and arbitrary continuous mappings, generalizing associative memory models of the period.
- The formalism reveals connections between associative memory, nonlinear system identification, and what were later called "kernel machines" (1990s) and "attention mechanisms" (2017–present).

Technically, polynomial associative memory preceded by more than a decade the kernel/RBF approach (Poggio and Girosi, 1989), and thus represents one of the earliest formulations of a flexible, learnable, nonlinear associative mapping. This places it conceptually *much closer* to the modern transformer attention rule

$$\mathrm{Attn}(q, K, V) = \mathrm{softmax}(qK^\top)\, V,$$

than the later Hopfield networks, which were purely autoassociative (they lacked heteroassociative structure) and were defined as a dynamical system rather than as functional relation.

In summary, Poggio's 1973–1975 nonlinear associative recall theory provides a general mathematical framework unifying:

- linear associative memories,
- multilayer polynomial memories,
- holographic and convolutional associative schemes,
- and the later kernel/RBF/hyperBF architectures.

Within the historical development of associative memory, it marks the transition from simple discrete models to general nonlinear function approximation—a conceptual leap whose modern counterpart is the transformer.

Implicit introduction of polynomial kernels. Viewed from the later perspective of kernel methods and reproducing kernel Hilbert spaces, the polynomial expansion in (6) can be reinterpreted as an explicit feature map associated with a *polynomial kernel*. Consider the degree–k feature map

$$\Phi_k : \mathbb{R}^m \to \mathbb{R}^{D_k},$$

whose coordinates are all monomials of degree up to k,

$$\Phi_k(x) = \big\{ x_{i_1} x_{i_2} \cdots x_{i_\ell} : 0 \le \ell \le k \big\}.$$

The polynomial expansion (6) can be written as a *linear* map in this feature space,

$$Y \approx W\,\Phi_k(X),$$

for some matrix W collecting the coefficients of the multilinear operators L_ℓ. The corresponding kernel

$$K_k(x, x') = \langle \Phi_k(x), \Phi_k(x') \rangle$$

is precisely a polynomial kernel. In the simplest homogeneous case (ignoring L_0) one obtains

$$K_k(x, x') = (x^\top x')^k,$$

while including L_0 leads to the inhomogeneous family

$$K_k(x, x') = (1 + x^\top x')^k.$$

In this sense, the 1975 nonlinear associative recall framework implicitly introduces polynomial kernels well before they became standard in the support vector machine and kernel methods literature. The later development of RBF and hyperBF networks can be seen as replacing the polynomial kernel by Gaussian or more general metric-based kernels, but the underlying idea of representing functions via inner products in a high-dimensional feature space is already present in this early work.

2.4 Palm's Correlation Matrix Memories (1980s)

Günther Palm introduced a mathematically rigorous theory of correlation-based memories. For binary $\{-1, +1\}$ patterns, he defined the storage matrix

$$W = \frac{1}{P} \sum_{\mu=1}^{P} y^\mu (x^\mu)^\top, \tag{10}$$

with retrieval again given by

$$\hat{y} = \operatorname{sign}(Wx).$$

This introduced:

- capacity bounds,
- error-correcting properties,
- and connections to coding theory.

Palm's work formalized the theoretical performance of heteroassociative memories and clarified the relationship to Hebbian learning.

2.5 Radial Basis Function Networks

In the late 1980s, Broomhead and Lowe introduced a framework for function approximation using *radial basis functions (RBFs)*, which can be interpreted as nonlinear, metric-based associative memories:

$$f(x) = \sum_{\mu=1}^{P} \alpha_\mu \, \phi(\|x - x^\mu\|) \, . \tag{11}$$

Here the stored "keys" x^μ and the learned coefficients α_μ define a smooth mapping from inputs to outputs.

The later *hyperBF* networks (Poggio and Girosi, 1990) generalized this formulation by allowing a learnable metric matrix A_μ for each basis function:

$$f(x) = \sum_{\mu=1}^{P} \alpha_\mu \, \phi\big((x - x^\mu)^\top A_\mu (x - x^\mu)\big) \, . \tag{12}$$

This is strikingly close to modern attention mechanisms, where the similarity metric between query q and key k^μ is learned through a dot-product or bilinear form and modulates the contribution of the associated value.

2.6 Kanerva's Sparse Distributed Memory (1984–1988)

Kanerva proposed a high-dimensional associative memory based on sparse binary addresses $h^\mu \in \{0,1\}^D$. A memory write distributes a value vector y^μ into all "hard locations" within a Hamming ball of radius r around the address h^μ. Formally, if H is the set of hard locations, then

$$\forall h \in H : d(h, h^\mu) < r \implies M(h) \mathrel{+}= y^\mu. \tag{13}$$

Retrieval is the sum of values at all locations within the same Hamming ball:

$$\hat{y} = \sum_{h \in H : d(h,x) < r} M(h). \tag{14}$$

Kanerva's architecture is conceptually close to the dentate gyrus–CA3 system in the hippocampus: sparse high-dimensional codes, random projections, and redundant storage across many locations.

2.7 Hopfield Networks (1982) and Their Limited Relation to Transformers

John Hopfield introduced an autoassociative memory storing patterns $\xi^\mu \in \{-1,+1\}^N$ in a symmetric recurrent network:

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu, \qquad w_{ij} = w_{ji}. \tag{15}$$

The dynamics

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

performs discrete coordinate descent on the energy function:

$$E(s) = -\frac{1}{2}\sum_{i \neq j} w_{ij} s_i s_j. \tag{16}$$

Historically, Hopfield networks became synonymous with associative memories, but they represent a relatively narrow subclass:

- purely autoassociative (no $x^\mu \mapsto y^\mu$ mapping),
- reliant on symmetric weights (not biologically plausible),
- limited capacity ($0.138\,N$ patterns),
- and lacking the content-addressable key–value structure central to transformers.

In contrast, the earlier heteroassociative models (Willshaw, Kohonen, Palm, Poggio) and the later RBF/hyperBF networks bear a much stronger formal resemblance to the modern attention mechanism. Transformers rely on *metric-based key–query similarity* and *value aggregation*, which are absent from classical Hopfield models.

Thus, although Hopfield networks played an influential role in theoretical neuroscience, the core mathematical ideas underlying transformers trace far more directly to the linear and nonlinear associative memories developed from the late 1960s through the 1990s.

3 Modern Parallels: Transformers as Associative Memories

The rise of the Transformer architecture has reshaped modern deep learning, yet its core computational mechanism—the attention block—is mathematically and conceptually an associative memory. This connection is rarely highlighted, but it represents the culmination of a long intellectual lineage originating with early heteroassociative memories (Willshaw–Longuet-Higgins), linear correlation memories (Kohonen, Palm), nonlinear polynomial associative recall (Poggio 1973–1975), and later kernel-based and HyperBF architectures (Poggio–Girosi). Recent work has now shown that a single attention head is formally equivalent to a *normalized HyperBF network*, and that an entire Transformer block is a composition of associative memory modules.

This section develops this connection in detail.

3.1 Attention as Kernel-Based Associative Recall

At the heart of a Transformer is the dot-product attention mechanism:

$$\text{Attn}(q, K, V) = \text{softmax}\left(\frac{qK^\top}{\sigma^2}\right) V. \tag{17}$$

Here:

- $q \in \mathbb{R}^d$ is the query,
- $K = (k_1^\top, \ldots, k_L^\top)^\top$ is the matrix of keys,
- $V = (v_1^\top, \ldots, v_L^\top)^\top$ is the matrix of values.

The attention output is a weighted sum of values:

$$\hat{y} = \sum_{i=1}^{L} w_i(q)\, v_i, \qquad w_i(q) = \frac{\exp(q \cdot k_i/\sigma^2)}{\sum_j \exp(q \cdot k_j/\sigma^2)}. \tag{18}$$

This is precisely the form of a normalized kernel-based associative memory (Nadaraya–Watson estimator):

$$f(x) = \sum_i \frac{c_i\, K(\|x - t_i\|^2)}{\sum_j K(\|x - t_j\|^2)}.$$

The conceptual parallel is immediate:

$$x \leftrightarrow q, \quad t_i \leftrightarrow k_i, \quad c_i \leftrightarrow v_i.$$

Attention is therefore a smooth, high-dimensional, metric-based associative memory.

3.2 HyperBF Equivalence: Attention as a Normalized HyperBF

Recent work (e.g. CBMM Memo 143) shows that if $\|q\| = \|k_i\| = 1$ (or more generally after removing constant offsets in the exponent), then the softmax of the dot product is exactly equivalent to a Gaussian RBF kernel. Formally, one can rewrite the attention weights as

$$\mathrm{Attn}(q, k_i, v_i) = \sum_i \frac{\exp(-\|q - k_i\|^2/(2\sigma^2))}{\sum_j \exp(-\|q - k_j\|^2/(2\sigma^2))}\, v_i. \tag{19}$$

This is exactly the "normalized HyperBF" form:

$$\phi(x) = \sum_{i=1}^{N} \frac{c_i\, K(\|x - t_i\|_W^2)}{\sum_j K(\|x - t_j\|_W^2)}. \tag{20}$$

Thus, an attention head is *identical* to a HyperBF unit:

$$\boxed{\text{Attention Head} = \text{Normalized HyperBF Network.}}$$

The Transformer's most distinctive component is therefore a kernel-based associative memory.

3.3 Multi-Head Attention as Parallel Associative Memories

A single attention head performs one associative mapping. Multi-head attention applies several such mappings in parallel:

$$\mathrm{MHA}(Q,K,V) = [\mathrm{Attn}_1(Q,K,V), \ldots, \mathrm{Attn}_H(Q,K,V)]\, W_O.$$

Each head has:

$$q_h = QW_{Q,h}, \quad k_h = KW_{K,h}, \quad v_h = VW_{V,h},$$

and hence each head defines its own:

- learned metric on the input space,
- associative memory addressing scheme,
- value space.

Thus the Transformer does not contain "one associative memory." It contains *many*, each specialized to a different subspace, mirroring the way biological memory systems (e.g., hippocampal place cells, grid cells, concept cells) decompose representations.

3.4 The MLP Block as a Second Associative Memory

The feedforward network (MLP block) in a Transformer layer can also be expressed as a HyperBF operator:

$$\phi(x) = \sum_i K\big(\|x - k_i\|_W^2\big)\, v_i,$$

where the "centers" k_i are learned. Thus the two parts of a Transformer layer are:

Self-Attention	**HyperBF with Learned Centers**
keys = projected inputs	keys = learned centers
queries = projected inputs	queries = layer inputs
values = projected inputs	values = learned outputs

Hence a Transformer layer is a *composition of two associative memories*:

$$\text{Transformer Layer} = \underbrace{\mathrm{HyperBF}_{\text{sample}\to\text{sample}}}_{\text{Self-Attention}} \circ \underbrace{\mathrm{HyperBF}_{\text{sample}\to\text{center}}}_{\text{Feedforward Layer}}.$$

This unifies the architecture under a single computational abstraction: the associative memory operator.

3.5 Experimental Evidence from Homogeneous HyperBF Transformers

The formal equivalence between attention heads and normalized HyperBF units would be of limited interest if it did not also support practical architectures. Recent work (see CBMM Memo 143, "A Homogeneous Transformer Architecture") constructs a vision architecture in which *all* nonlinear components of a standard Vision Transformer (ViT)—both the self–attention blocks and the MLP feedforward blocks—are replaced by HyperBF modules. The resulting "homogeneous" architecture can be written schematically as

$$y = x + \mathrm{HyperBF}\Big(\mathrm{LayerNorm}\big(x + \mathrm{HyperBF}(\mathrm{LayerNorm}(x))\big)\Big), \tag{21}$$

where each $\mathrm{HyperBF}(\cdot)$ denotes a multi-head HyperBF operator playing the role usually played by attention or by the MLP block.

In this homogeneous architecture, the only nonlinearity is provided by HyperBF units; there are no ReLUs or softmax-based attention blocks. Despite this drastic change, experiments on standard image classification benchmarks show that the HyperBF-only models closely match the performance of a baseline ViT with comparable depth, width and number of heads. For example, on CIFAR10, CIFAR100, and Tiny ImageNet, a 4-block, 4-head HyperBF network trained under the same conditions as a ViT yields accuracies that differ by less than two percentage points:

Dataset	ViT Acc. (%)	HyperBF Acc. (%)	Gap (%)
CIFAR10	75.61	75.12	0.49
CIFAR100	49.90	48.30	1.60
Tiny ImageNet	32.03	31.14	0.89

Thus, a carefully designed network in which attention and MLP blocks are both replaced by HyperBF layers can reproduce the characteristic behavior and performance of a Vision Transformer on nontrivial vision tasks. This provides concrete empirical evidence that:

(i) the essential computational role of attention and MLP blocks can be implemented by metric-based associative memories (HyperBF units);

(ii) the transformer block is, in practice as well as theory, a composition of associative memory modules rather than a fundamentally different type of computation.

In combination with the formal derivations above, these experiments strongly support the view that modern transformers can be reinterpreted as deep networks built from a small number of canonical associative operations.

3.6 Relation to Classical Associative Memories

The attention mechanism extends classical associative memories in several important ways:

- **It is heteroassociative**, mapping queries to values (unlike Hopfield).

- **It is metric-based**, like RBF and HyperBF networks.
- **It uses learned projections** W_Q, W_K, W_V which define the associative metric and embedding.
- **It is differentiable and end-to-end trainable**, unlike early discrete memories.
- **It is high-dimensional and compositional**, enabling flexible retrieval.
- **It is multi-head**, providing modular, reusable sub-associations.

In this sense, Transformers follow directly from the classical theory:

$$\text{Willshaw / Kohonen / Palm} \;\Rightarrow\; \text{RBF / HyperBF} \;\Rightarrow\; \text{Attention.}$$

3.7 Transformers as the Culmination of Associative Memory Theory

Seen through this lens, the Transformer is not a break from earlier neural models but the culmination of their central idea:

Modern deep learning works because it rediscovered associative memory in a powerful parametric form.

Transformers combine:

- sparse compositional structure (as in biological systems),
- metric-based high-dimensional retrieval (as in HyperBF),
- modular multi-head specialization,
- and stacked associative mappings forming deep architectures.

Thus, associative memory is not a historical footnote: it is the operating principle of modern AI. It is also likely the operating principle of biological intelligence, from monosynaptic sensorimotor associations to hippocampal episodic binding.

4 Evolutionary Perspective

From an evolutionary standpoint, associative memory is not an abstract computational operation but the most primitive and biologically plausible mechanism for linking sensory patterns to motor actions. Long before animals possessed cortex, or even multi-layered neural architectures, early nervous systems relied on simple one-to-one associations: a particular visual, tactile, or chemical stimulus would trigger an innate motor program such as escape, withdrawal, attraction, or feeding. These *monosynaptic reflex arcs* constitute the earliest form of heteroassociation:

$$x_{\text{sensory}} \longmapsto y_{\text{motor}}.$$

They are the biological analog of the (x^μ, y^μ) pairs stored in the first heteroassociative memory models (Willshaw, Longuet-Higgins, Kohonen).

Over evolutionary time, such hardwired associations became *plastic*. Hebbian and heterosynaptic mechanisms—documented experimentally as early as Kandel–Tauc (1965)—allowed the strength of the sensory→motor mapping to change with experience. These plastic reflexes represent the next step in the evolutionary ladder: a learned, adaptive heteroassociative mapping. In modern animals, this basic motif survives in every spinal cord and brainstem circuit.

A crucial point is that these systems scale. Once evolution has a reliable mechanism for forming and modifying associations, it can compose them into multi-stage circuits. Reflex chains become behavior sequences; sequences become multi-area sensorimotor loops; and eventually, the same principles give rise to hierarchical architectures such as:

$$\text{retina} \rightarrow \text{thalamus} \rightarrow \text{primary cortex} \rightarrow \text{association cortex}.$$

Associative Circuits as Evolutionary Precursors to Optimization

The evolutionary leap from simple associative bindings to multi-layer adaptive circuits requires not only feedforward connections but also feedback pathways capable of evaluating and adjusting intermediate computations. This idea is directly reflected in the *ascending* (feedforward) and *descending* (feedback) cortical pathways that are anatomically ubiquitous.

A striking illustration of how such a system could arise through gradual evolutionary refinement is provided by the recently proposed "Self-Assembling Learning" (SAL) circuit (Liao et al., 2025) :contentReference[oaicite:1]index=1. This model shows that a biologically realistic motif involving four synaptic connections between ascending and descending streams can, under entirely local plasticity rules, self-organize to implement a form of stochastic gradient descent (SGD). The circuit begins with random bidirectional connectivity and, through development and experience, converges to a structured system that optimizes multi-stage computations.

In other words, a multi-layer optimization circuit can *emerge spontaneously* from:

- random initial wiring,
- local Hebbian and heterosynaptic plasticity,
- and reciprocal ascending/descending pathways.

This provides a concrete mechanistic example of how evolution could discover hierarchical learning systems without requiring explicit genetic specification of multi-layer backpropagation-like algorithms.

From Reflexes to Multi-Stage Adaptive Hierarchies

The SAL circuit supports a speculative but biologically grounded evolutionary narrative:

(a) **Reflexes (single synapse).** Hardwired sensorimotor associations appear first. They are one-shot heteroassociations.

(b) **Plastic reflex circuits (local learning).** Heterosynaptic and Hebbian mechanisms allow these arcs to adapt with experience.

(c) **Two-stage loops (ascending + descending pathways).** Evolution introduces crude feedback pathways, enabling multi-step refinement of output. These are precursors of optimization over short time windows.

(d) **Self-assembly of multi-stage learning circuits.** As shown by the SAL model, random upstream/downstream pathways with plastic cross-links can self-organize into effective SGD-like optimization circuits capable of hierarchical learning.

(e) **Cortical scaffolding.** The hippocampus provides episodic heteroassociation and indexing; cortex receives repeated, structured input from hippocampus and gradually internalizes stable associations (the "scaffolding" effect).

(f) **Compositionality + associative memory = intelligence.** Once multi-layer circuits can be optimized and once their representations can be stabilized in cortex, all of the machinery required for high-level cognition—planning, abstraction, symbolic reasoning, and linguistic structure—can emerge from deep stacks of associative memories.

Associative Memory as an Evolutionary Attractor

The unified picture is that evolution repeatedly converges toward *associative architectures* for both biological and computational reasons:

- Associative memories require no weight symmetry (unlike Hopfield or strict backprop).
- They operate with local plasticity (Hebbian or heterosynaptic).
- They tolerate random initial connectivity and can self-organize.
- They naturally scale to multi-layer forms when paired with feedback pathways.
- They support both fast modulation (short-term recurrent computation) and slow learning (long-term consolidation).

This makes associative memory an *evolutionary attractor*: the simplest neuronal mechanism that can scale upward into the sophisticated multi-stage optimization circuits found in cortex and, as we argue in this paper, in modern artificial intelligence systems such as transformers.

5 Conclusion

From their inception in the 20th century to their role in today's cutting-edge deep learning models, associative memories have proven to be a powerful and versatile concept. By understanding their historical roots and evolutionary plausibility, we gain a deeper appreciation of their enduring significance in the field of artificial intelligence.

References

[1] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222(5197):960–962, 1969.

[2] T. Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, 100(4):353–359, 1972.

[3] T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics*, 19(4):201–209, 1975.

[4] G. Palm. On associative memory. *Biological Cybernetics*, 36(1):19–31, 1980.

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

[6] P. Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

[7] T. Poggio and F. Girosi. A theory of networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[8] L. M. Talamini and S. M. Meeter. A naturalistic model of human episodic memory and the hippocampal-cortical interface. *Reviews in the Neurosciences*, 15(4):241–260, 2004.

[9] A. Vaswani, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] T. Poggio. A Homogeneous Transformer Architecture. *CBMM Memo*, No. 143, MIT, 2023.

[11] E. R. Kandel and L. Tauc. Heterosynaptic facilitation in neurones of the abdominal ganglion of Aplysia. *The Journal of Physiology*, 181(1):1–27, 1965.

[12] I. Liao, et al. Self-Assembling Learning (SAL): A biologically plausible circuit for optimization. *arXiv preprint (pending/cited)*, 2025.