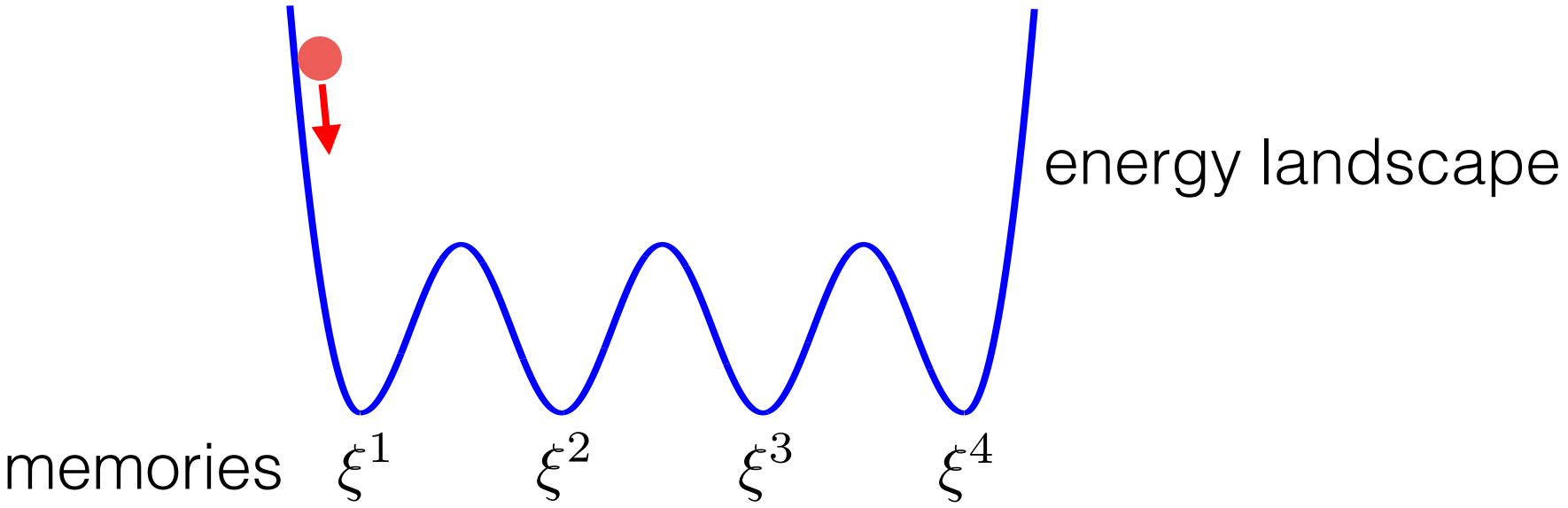
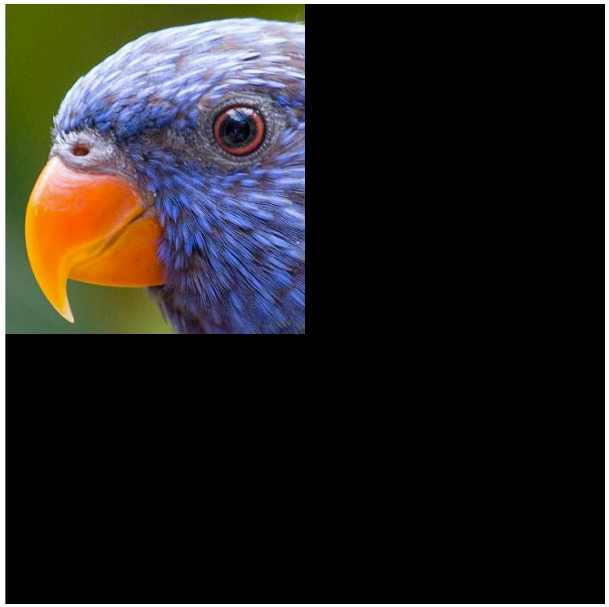
Modern Hopfield Networks in AI and Neuroscience

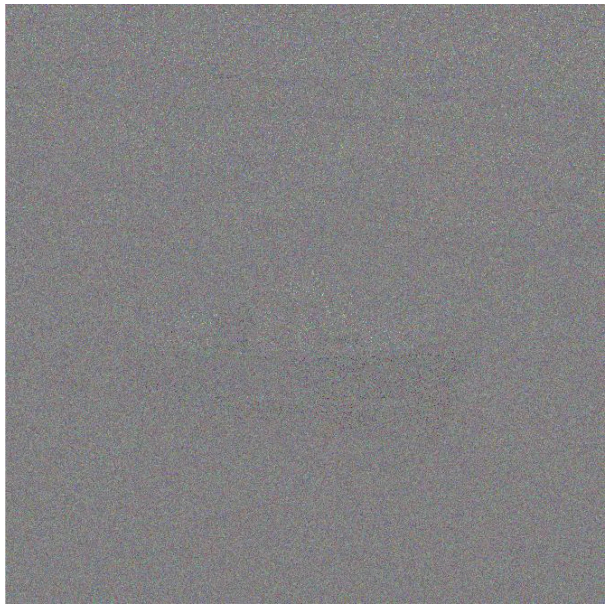
Dmitry Krotov
MIT-IBM Watson AI Lab
IBM Research

MIT-IBM
Watson
AI Lab

What is associative memory?







Standard Associative Memory
Classical Hopfield Network

$$E = -\sum_{i,j=1}^{N} \sigma_i T_{ij} \sigma_j \qquad T_{ij} = \sum_{\mu=1}^{K} \xi_i^\mu \xi_j^\mu$$

σ_i -dynamical variables

ξ_i^μ -memorized patterns

N -number of neurons

K -number of memories

$$E = -\sum_{\mu=1}^{K} \Big(\sum_{i=1}^{N} \xi_i^\mu \sigma_i \Big)^2$$

$$K^{\max} \approx 0.14N$$

Hopfield, PNAS, 1982

Dense Associative Memory
Modern Hopfield Network

$$E = -\sum_{\mu=1}^{K} F\Big(\sum_{i=1}^{N} \xi_i^\mu \sigma_i \Big)$$

$F(x) = x^n, \quad \text{with} \quad n \geq 2$

$F(x) = \exp(x)$

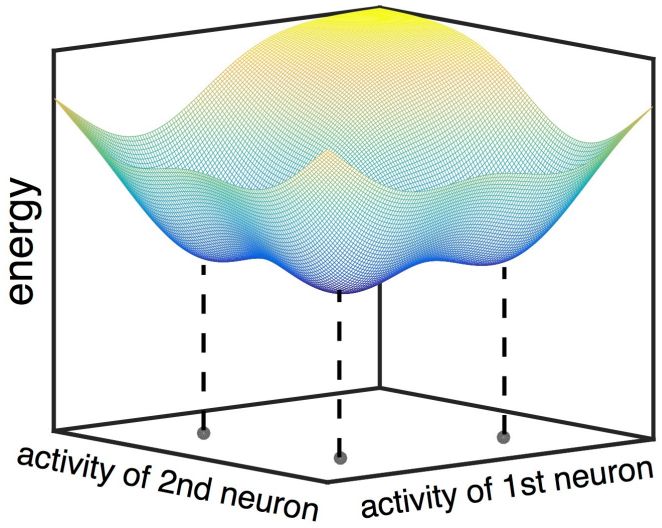
$$K^{\max} \approx \alpha_n N^{n-1}$$

$$K^{\max} \approx \exp(\alpha N), \quad \alpha < \ln(2)/2$$

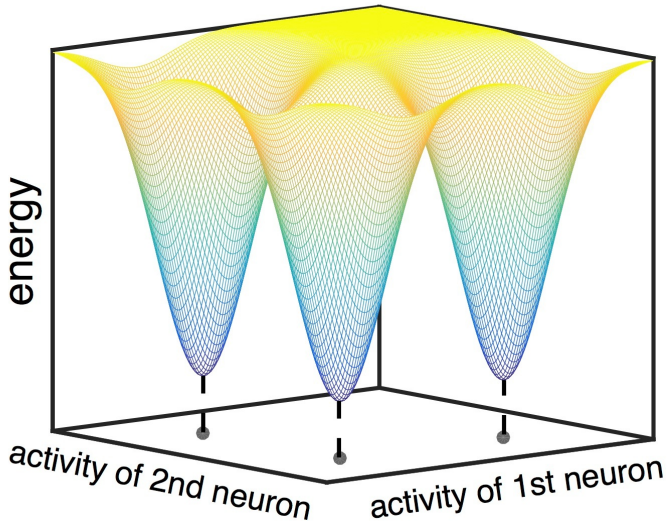
Krotov & Hopfield, NeurIPS, 2016
Demircigil et al., J.Stat.Phys.,2017

Why should this work?

Classical Hopfield Network



Modern Hopfield Network

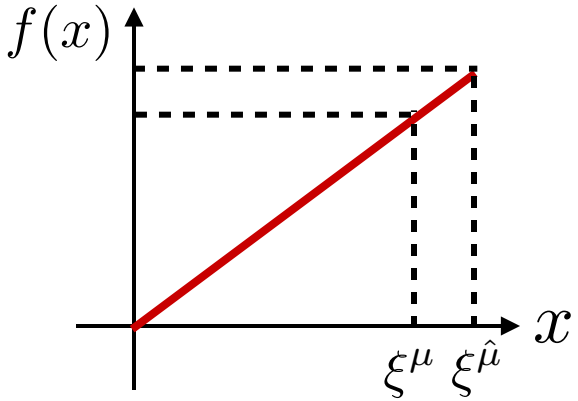


Why should this work?

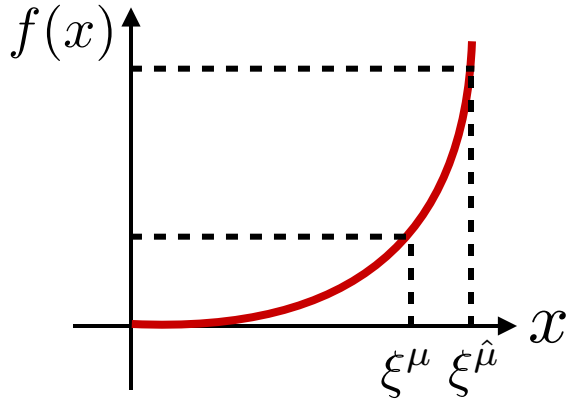
$$\sigma_i^{(t+1)} = Sign\Big[\sum_{\mu=1}^{K}\Big(F\Big(\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j^{(t)}\Big) - F\Big(-\xi_i^\mu + \sum_{j\neq i}\xi_j^\mu\sigma_j^{(t)}\Big)\Big)\Big]$$

$$\sigma_i^{(t+1)} = Sign\Big[\sum_{\mu=1}^{K}\xi_i^\mu f\Big(\sum_{j\neq i}\xi_j^\mu\sigma_j^{(t)}\Big)\Big] \qquad \text{where} \quad f(x) = \frac{dF}{dx}$$

classical Hopfield network



modern Hopfield network



What is a Modern Hopfield Network with one hidden layer in its most general form?

Published as a conference paper at ICLR 2021

L RGE SSOCI TIVE MEMORY PROBLEM IN NEURO-BIOLOGY ND M CHINE LE RNING

Dmitry Krotov
MIT-IBM Watson I Lab
IBM Research
krotov@i m.com

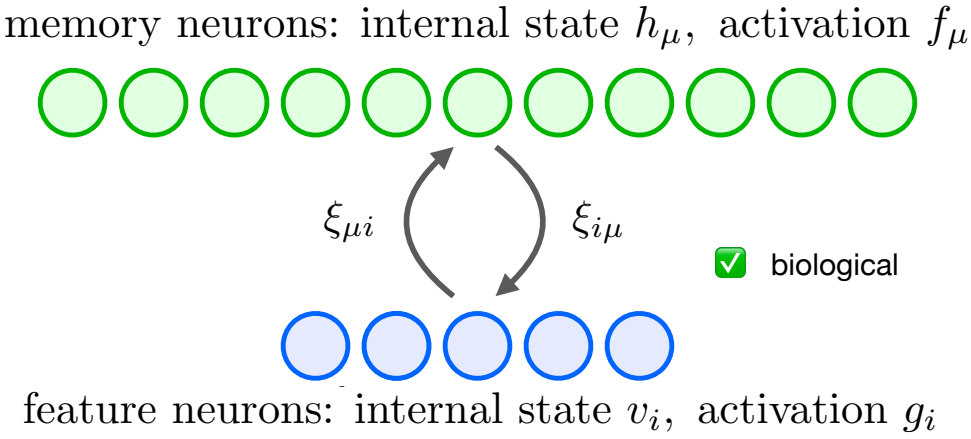
John Hopfield
Princeton Neuroscience Institute
Princeton University
hopfield@princeton.edu

BSTR CT

Dense ssociative Memories or modern Hopfield networks permit storage and reliable retrieval of an exponentially large (in the dimension of feature space) number of memories. t the same time, their naive implementation is non-biological, since it seemingly requires the existence of many-body synaptic junctions between the neurons. We show that these models are effective descriptions of a more microscopic (written in terms of biological degrees of freedom) theory that has additional (hidden) neurons and only requires two-body interactions between them. For this reason our proposed microscopic theory is a valid model of large associative memory with a degree of biological plausibility. The dynamics of our network and its reduced dimensional equivalent both minimize energy (Lyapunov) functions. When certain dynamical variables (hidden neurons) are integrated out from our microscopic theory, one can recover many of the models that were previously discussed in the literature, e.g. the model presented in "Hopfield Networks is ll You Need" paper. We also provide an alternative derivation of the energy function and the update rule proposed in the aforementioned paper and clarify the relationships between various models of this class.

Microscopic theory of modern Hopfield networks

$$\begin{cases} \tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu} f_\mu - v_i + I_i \\ \tau_h \frac{dh_\mu}{dt} = \sum_{i=1}^{N_f} \xi_{\mu i} g_i - h_\mu \end{cases}$$

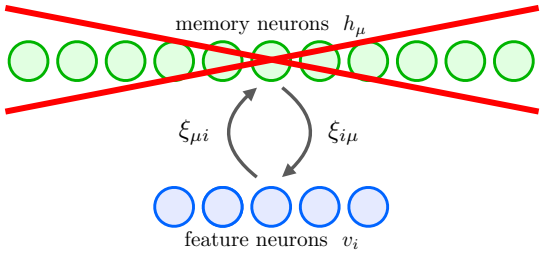


$$f_\mu = \frac{\partial L_h}{\partial h_\mu}, \quad \text{and} \quad g_i = \frac{\partial L_v}{\partial v_i}$$

$$E(t) = \Big[\sum_{i=1}^{N_f} (v_i - I_i) g_i - L_v \Big] + \Big[\sum_{\mu=1}^{N_h} h_\mu f_\mu - L_h \Big] - \sum_{\mu,i} f_\mu \xi_{\mu i} g_i$$

$$\frac{dE(t)}{dt} = -\tau_f \sum_{i,j=1}^{N_f} \frac{dv_i}{dt} \frac{\partial^2 L_v}{\partial v_i \partial v_j} \frac{dv_j}{dt} - \tau_h \sum_{\mu,\nu=1}^{N_h} \frac{dh_\mu}{dt} \frac{\partial^2 L_h}{\partial h_\mu \partial h_\nu} \frac{dh_\nu}{dt} \leq 0$$

Effective theory for feature neurons



	Model A	Model B	Model C
Lagrangian functions	$L_h = \sum_\mu F(h_\mu)$ $L_v = \sum_i \lvert v_i \rvert$	$L_h = \log\left(\sum_\mu e^{h_\mu}\right)$ $L_v = \frac{1}{2}\sum_i v_i^2$	$L_h = \sum_\mu F(h_\mu)$ $L_v = \sqrt{\sum_i v_i^2}$
energy	$E = -\sum_{\mu=1}^{N_h} F\left(\sum_i \xi_{\mu i}\sigma_i\right)$	$E = \frac{1}{2}\sum_{i=1}^{N_f} v_i^2 - \log\left(\sum_\mu \exp(\sum_i \xi_{\mu i} v_i)\right)$	$E = -\sum_\mu F\left(\sum_i \xi_{\mu i}\frac{v_i}{\sqrt{\sum_j v_j^2}}\right)$
effective update rule	$\tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu} f\left(\sum_{j=1}^{N_f} \xi_{\mu j}\sigma_j\right) - v_i$	$\tau_f \frac{dv_i}{dt} = \sum_{\mu=1}^{N_h} \xi_{i\mu}\mathrm{softmax}\left(\sum_{j=1}^{N_f} \xi_{\mu j} v_j\right) - v_i$	$\tau_f \frac{dv_i}{dt} = \sum_\mu \xi_{i\mu} f\left[\sum_j \xi_{\mu j}\frac{v_j}{\sqrt{\sum_k v_k^2}}\right] - v_i$

- Dense Associative Memory for pattern recognition
- On a model of associative memory with huge storage capacity

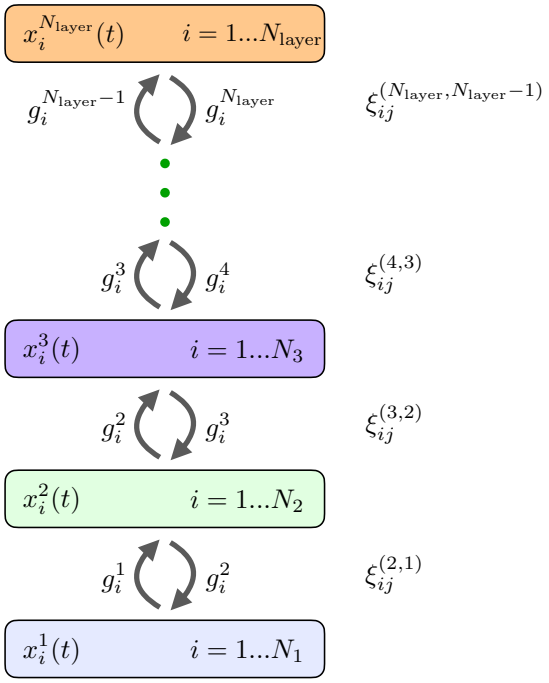
- Hopfield networks is all you need

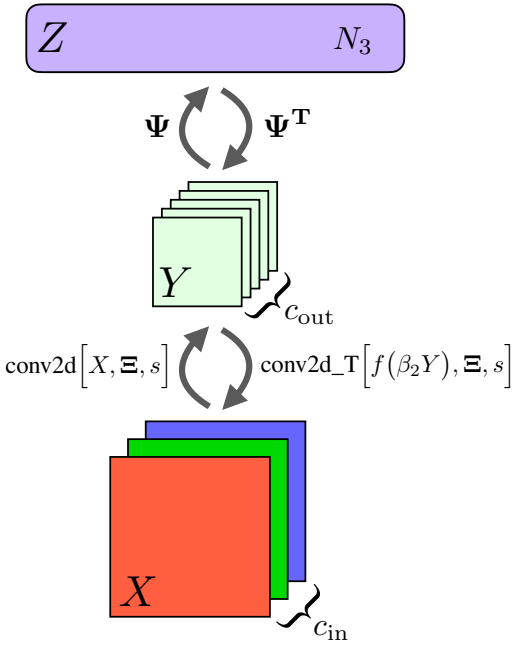
Hierarchical Associative Memory

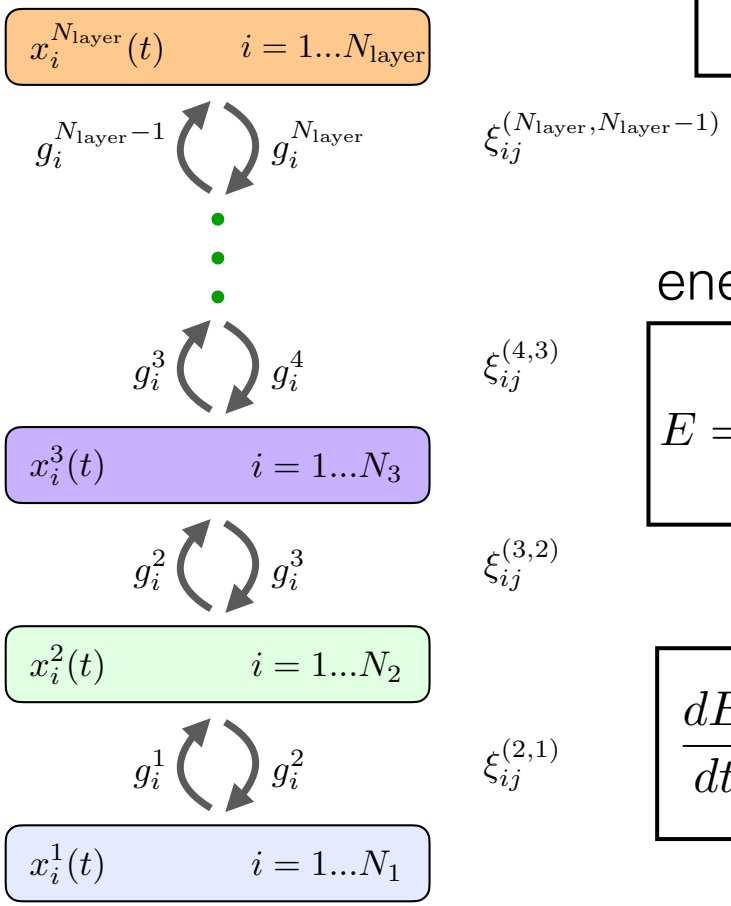
Dmitry Krotov
MIT-IBM Watson AI Lab
IBM Research
krotov@ibm.com

Abstract

Dense Associative Memories or Modern Hopfield Networks have many appealing properties of associative memory. They can do pattern completion, store a large number of memories, and can be described using a recurrent neural network with a degree of biological plausibility and rich feedback between the neurons. At the same time, up until now all the models of this class have had only one hidden layer, and have only been formulated with densely connected network architectures, two aspects that hinder their machine learning applications. This paper tackles this gap and describes a fully recurrent model of associative memory with an arbitrary large number of layers, some of which can be locally connected (convolutional), and a corresponding energy function that decreases on the dynamical trajectory of the neurons' activations. The memories of the full network are dynamically "assembled" using primitives encoded in the synaptic weights of the lower layers, with the "assembling rules" encoded in the synaptic weights of the higher layers. In addition to the bottom-up propagation of information, typical of commonly used feedforward neural networks, the model described has rich top-down feedback from higher layers that help the lower-layer neurons to decide on their response to the input stimuli.







$$g_i^A = \frac{\partial L^A}{\partial x_i^A}$$

definition of the activation function through the Lagrangian function

energy function

$$E = \sum_{A=1}^{N_{\text{layer}}} \left[\sum_{i=1}^{N_A} x_i^A g_i^A - L^A \right] - \sum_{A=1}^{N_{\text{layer}}-1} \sum_{i=1}^{N_{A+1}} \sum_{j=1}^{N_A} g_i^{A+1} \xi_{ij}^{(A+1,A)} g_j^A$$

$$\frac{dE}{dt} = -\sum_{A=1}^{N_{\text{layer}}} \tau_A \sum_{i,j=1}^{N_A} \frac{dx_j^A}{dt} \frac{\partial^2 L^A}{\partial x_j^A \partial x_i^A} \frac{dx_i^A}{dt} \leq 0$$

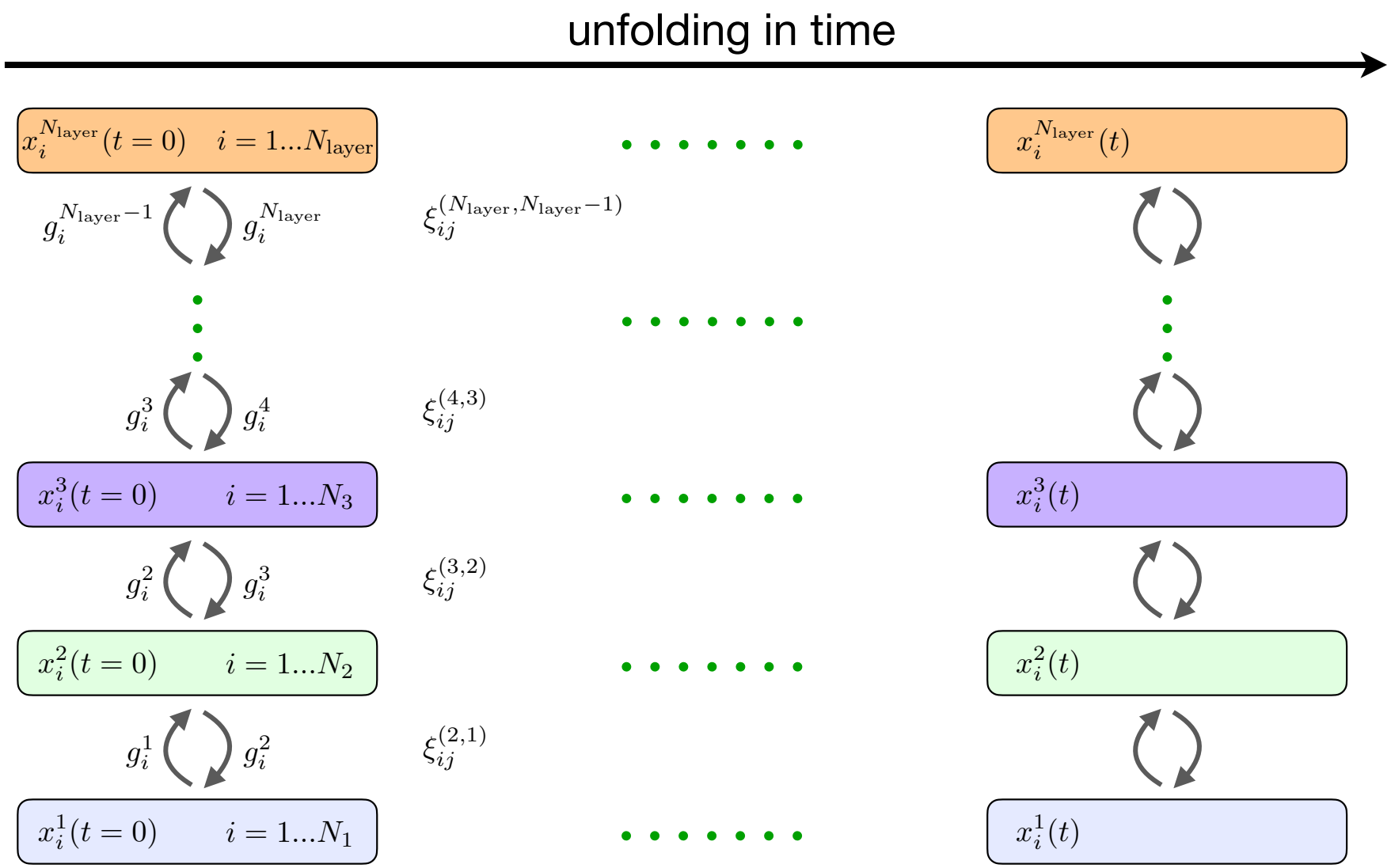
equations for neuron's state update

$$\tau_A \frac{dx_i^A}{dt} = \sum_{j=1}^{N_{A-1}} \xi_{ij}^{(A,A-1)} g_j^{A-1} + \sum_{j=1}^{N_{A+1}} \xi_{ij}^{(A,A+1)} g_j^{A+1} - x_i^A$$

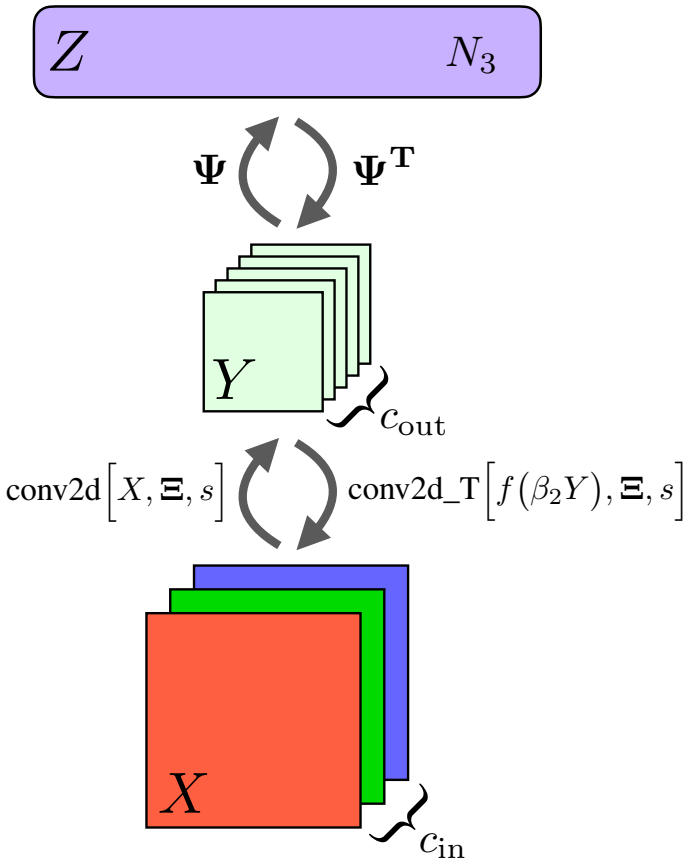
with boundary conditions

$$g_i^0 = 0, \quad \text{and} \quad g_i^{N_{\text{layer}}+1} = 0$$

How can we train such hierarchical networks?



Can we make Hopfield Networks convolutional?



$$\begin{cases} \tau_3 \dfrac{dZ}{dt} = \mathbf{\Psi}\ \text{flatten}\left[f\left(\beta_2 Y\right)\right] - Z \\ \tau_2 \dfrac{dY}{dt} = \text{reshape}_{[\tilde{L},\tilde{L},c_{\text{out}}]}\left[\mathbf{\Psi^T} p\left(\beta_3 Z\right)\right] + \text{conv2d}\left[X, \mathbf{\Xi}, s\right] - Y \\ \tau_1 \dfrac{dX}{dt} = \text{conv2d\_T}\left[f\left(\beta_2 Y\right), \mathbf{\Xi}, s\right] - X \end{cases}$$

$$E = \frac{1}{2}\sum_{x,y,\mu} X_{x,y,\mu}^2 + \sum_{\tilde{x},\tilde{y},\tilde{\mu}} Y_{\tilde{x},\tilde{y},\tilde{\mu}} f\left(\beta_2 Y_{\tilde{x},\tilde{y},\tilde{\mu}}\right) - \frac{1}{\beta_2}\sum_{\tilde{x},\tilde{y}} \log\left(\sum_{\tilde{\mu}} e^{\beta_2 Y_{\tilde{x},\tilde{y},\tilde{\mu}}}\right)$$
$$- \frac{1}{\beta_3}\log\left(\sum_{\alpha} e^{\beta_3 Z_\alpha}\right) - \sum_{\tilde{x},\tilde{y},\tilde{\mu}} f\left(\beta_2 Y_{\tilde{x},\tilde{y},\tilde{\mu}}\right) \text{conv2d}\left[X, \mathbf{\Xi}, s\right]_{\tilde{x},\tilde{y},\tilde{\mu}}$$

Modern Hopfield Networks and Attention for Immune Repertoire Classification

Michael Widrich[*] **Bernhard Sch fl**[*] **Milena Pavlović**[‡,§] **Hubert Ramsauer**[*]

Lukas Gruber[*] **Markus Holzleitner**[*] **Johannes Brandstetter**[*] **Geir Kjetil Sandve**[§]

Victor Greiff[‡] **Sepp Hochreiter**[*,†]

Günter Klambauer[*]
[*]ELLIS Unit Linz and LIT AI Lab,
Institute for Machine Learning,
Johannes Kepler University Linz, Austria
[†]Institute of Advanced Research in Artificial Intelligence (IARAI)
[‡]Department of Immunology, University of Oslo, Norway
[§]Department of Informatics, University of Oslo, Norway

Abstract

A central mechanism in machine learning is to identify, store, and recognize patterns. How to learn, access, and retrieve such patterns is crucial in Hopfield networks and the more recent transformer architectures. We show that the attention mechanism of transformer architectures is actually the update rule of modern Hopfield networks that can store exponentially many patterns. We exploit this high storage capacity of modern Hopfield networks to solve a challenging multiple instance learning (MIL) problem in computational biology: immune repertoire classification. Accurate and interpretable machine learning methods solving this problem could pave the way towards new vaccines and therapies, which is currently a very relevant research topic intensified by the COVID-19 crisis. Immune repertoire classification based on the vast number of immunosequences of an individual is a MIL problem with an unprecedentedly massive number of instances, two orders of magnitude larger than currently considered problems, and with an extremely low witness rate. In this work, we present our novel method DeepRC that integrates transformer-like attention, or equivalently modern Hopfield networks, into deep learning architectures for massive MIL such as immune repertoire classification. We demonstrate that DeepRC outperforms all other methods with respect to predictive performance on large-scale experiments, including simulated and real-world virus infection data, and enables the extraction of sequence motifs that are connected to a given disease class. Source code and datasets: *https://github.com/ml-jku/DeepRC*

HOPFIELD NETWORKS IS ALL YOU NEED

Hubert Ramsauer[*] **Bernhard Schäfl**[*] **Johannes Lehner**[*] **Philipp Seidl**[*]
Michael Widrich[*] **Thomas Adler**[*] **Lukas Gruber**[*] **Markus Holzleitner**[*]
Milena Pavlović[‡,§] **Geir Kjetil Sandve**[§] **Victor Greiff**[‡] **David Kreil**[†]
Michael Kopp[†] **Günter Klambauer**[*] **Johannes Brandstetter**[*] **Sepp Hochreiter**[*,†]
[*]ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University Linz, Austria
[†]Institute of Advanced Research in Artificial Intelligence (IARAI)
[‡]Department of Immunology, University of Oslo, Norway
[§]Department of Informatics, University of Oslo, Norway

ABSTRACT

We introduce a modern Hopfield network with continuous states and a corresponding update rule. The new Hopfield network can store exponentially (with the dimension of the associative space) many patterns, retrieves the pattern with one update, and has exponentially small retrieval errors. It has three types of energy minima (fixed points of the update): (1) global fixed point averaging over all patterns, (2) metastable states averaging over a subset of patterns, and (3) fixed points which store a single pattern. The new update rule is equivalent to the attention mechanism used in transformers. This equivalence enables a characterization of the heads of transformer models. These heads perform in the first layers preferably global averaging and in higher layers partial averaging via metastable states. The new modern Hopfield network can be integrated into deep learning architectures as layers to allow the storage of and access to raw input data, intermediate results, or learned prototypes. These Hopfield layers enable new ways of deep learning, beyond fully-connected, convolutional, or recurrent networks, and provide pooling, memory, association, and attention mechanisms. We demonstrate the broad applicability of the Hopfield layers across various domains. Hopfield layers improved state-of-the-art on three out of four considered multiple instance learning problems as well as on immune repertoire classification with several hundreds of thousands of instances. On the UCI benchmark collections of small classification tasks, where deep learning methods typically struggle, Hopfield layers yielded a new state-of-the-art when compared to different machine learning methods. Finally, Hopfield layers achieved state-of-the-art on two drug design datasets. The implementation is available at: https://github.com/ml-jku/hopfield-layers

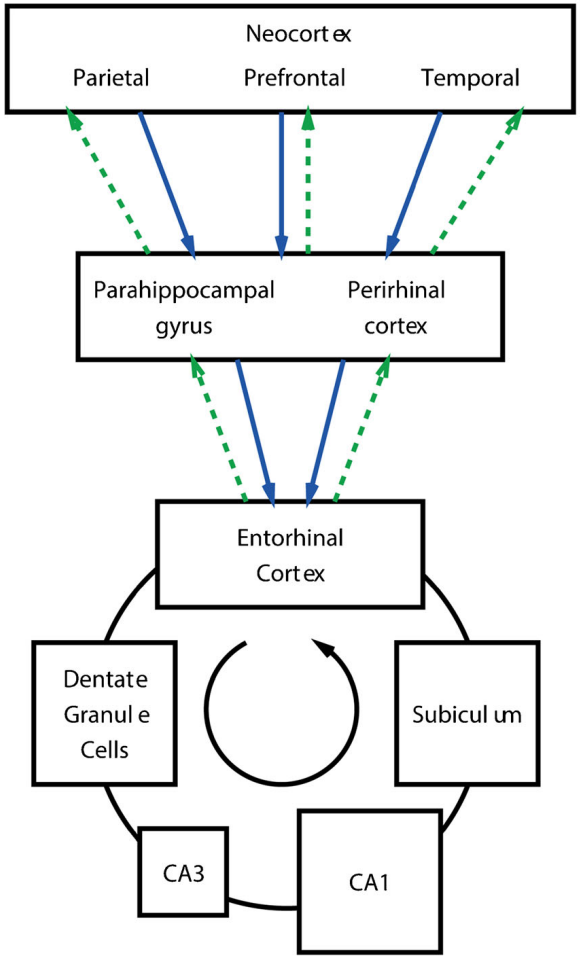
LARGE ASSOCIATIVE MEMORY PROBLEM IN NEUROBIOLOGY AND MACHINE LEARNING

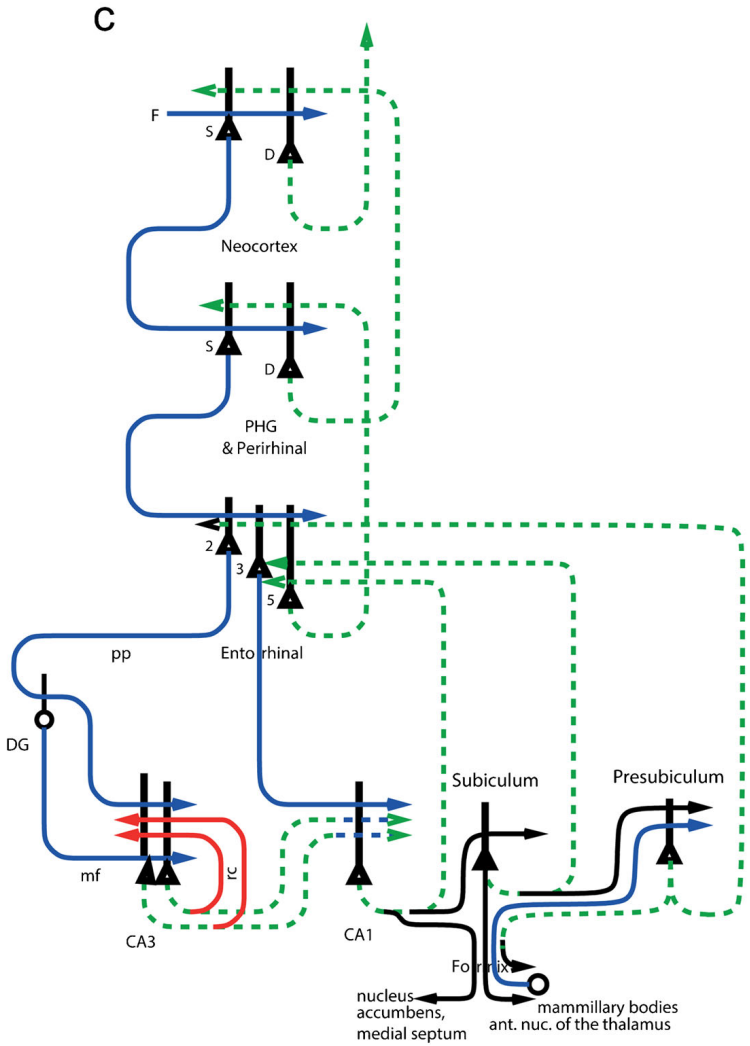
Dmitry Krotov
MIT-IBM Watson AI Lab
IBM Research
krotov@ibm.com

John Hopfield
Princeton Neuroscience Institute
Princeton University
hopfield@princeton.edu

ABSTRACT

Dense Associative Memories or modern Hopfield networks permit storage and reliable retrieval of an exponentially large (in the dimension of feature space) num-





Rolls, E.T. The storage and recall of memories in the hippocampo-cortical system. Cell Tissue Res 373, 577–604 (2018)

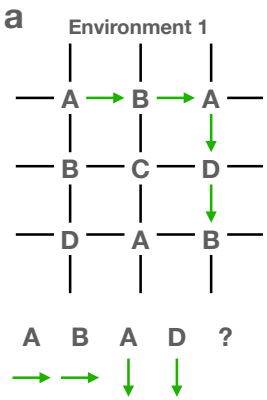
RELATING TRANSFORMERS TO MODELS AND NEURAL REPRESENTATIONS OF THE HIPPOCAMPAL FORMATION

James C.R. Whittington*
University of Oxford & Stanford University

Joseph Warren, Timothy E.J. Behrens
University of Oxford & University College London

ABSTRACT

Many deep neural network architectures loosely based on brain networks have recently been shown to replicate neural firing patterns observed in the brain. One of the most exciting and promising novel architectures, the Transformer neural network, was developed without the brain in mind. In this work, we show that transformers, when equipped with recurrent position encodings, replicate the precisely tuned spatial representations of the hippocampal formation; most notably place and grid cells. Furthermore, we show that this result is no surprise since it is closely related to current hippocampal models from neuroscience. We additionally show the transformer version offers dramatic performance gains over the neuroscience version. This work continues to bind computations of artificial and brain networks, offers a novel understanding of the hippocampal-cortical interaction, and suggests how wider cortical areas may perform complex tasks beyond current neuroscience models such as language comprehension.



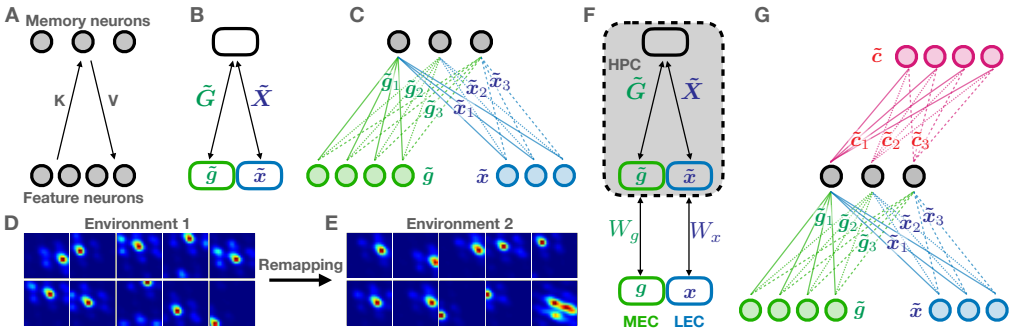


Figure 5: TEM-Transformer neural architecture. **(a)** Krotov & Hopfield (2020) describe a neurally plausible architectural instantiation the 'Hopfield networks is all you need' with a separation between 'feature' neurons (i.e. $\boldsymbol{h}$) and memory neurons (i.e. softmax($\boldsymbol{q}_t \boldsymbol{K}^T$). **(b-c)** This can be extended for TEM-t, but now the feature neurons are not all updated simultaneously, but only those across brain regions. **(d)** Memory neurons resemble hippocampal place cells and **(e)** remap randomly across environments. **(f)** A possible architecture where cortical neurons project to feature neurons in hippocampus which in turn project to memory neurons in hippocampus. **(g)** Additional brain regions can be included easily in this architecture with minimal increase in hippocampal neuron number.

Biological learning in key-value memory networks

Danil Tyulmankov*
Columbia University
dt2586@columbia.edu

Ching Fang*
Columbia University
ching.fang@columbia.edu

Annapurna Vadaparty
Columbia University
Stanford University
apvadaparty@gmail.com

Guangyu Robert Yang
Columbia University
Massachusetts Institute of Technology
yanggr@mit.edu

Abstract

In neuroscience, classical Hopfield networks are the standard biologically plausible model of long-term memory, relying on Hebbian plasticity for storage and attractor dynamics for recall. In contrast, memory-augmented neural networks in machine learning commonly use a key-value mechanism to store and read out memories in a single step. Such augmented networks achieve impressive feats of memory compared to traditional variants, yet their biological relevance is unclear. We propose an implementation of basic key-value memory that stores inputs using a combination of biologically plausible three-factor plasticity rules. The same rules are recovered when network parameters are meta-learned. Our network performs on par with classical Hopfield networks on autoassociative memory tasks and can be naturally extended to continual recall, heteroassociative memory, and sequence learning. Our results suggest a compelling alternative to the classical Hopfield network as a model of biological long-term memory.

Conclusions

- Modern Hopfield Networks have a large memory storage capacity, which scales significantly faster than linearly as a function of the number of feature neurons.
- MHN can be both continuous and binary.
- The easiest way to mathematically describe these networks is through the Lagrangian function.
- MHN can have many hidden layers with hierarchical representations.
- MHN can have structural architecture, with convolutions, attention, average pooling, etc.