https://doi.org/10.1038/s42254-023-00595-y

A new frontier for Hopfield networks

Dmitry Krotov

Over the past few years there has been a resurgence of interest in Hopfield networks of associative memory. Dmitry Krotov discusses recent theoretical advances and their broader impact in the context of energy-based neural architectures.

Hopfield networks of associative memory, famously formalized by John Hopfield more than forty years ago[1,2], had an enormous impact on multiple disciplines including statistical physics, neuroscience, and machine learning. The core idea, that the associative memory capability can be described by the energy descent dynamics (Fig. 1), found numerous applications in the study of complex systems with rugged landscapes. This idea also inspired the development of restricted Boltzmann machines, which were instrumental in the early days of deep learning. In cognitive and neurosciences, Hopfield networks established a conceptual formalization of the notion of memory.

Traditional Hopfield networks

Mathematically, the state of the Hopfield network is described by an N_f-dimensional vector of features $\mathbf{x}$, which can be either binary or continuous. The temporal evolution of this state vector is governed by an energy function, which has local minima located at a set of K_mem memory vectors $\boldsymbol{\xi}^\mu$ representing the patterns that the network stores in its weights (each pattern is an N_f-dimensional vector). When presented with an initial prompt that resembles one of the memory vectors, the energy descent dynamics finds the most similar memory vector based on the similarity between the initial prompt and the set of available memory patterns. The traditional Hopfield network[1] successfully solves this task if the number of memory patterns K_mem is small. Specifically, the maximal number of memory patterns that can be successfully retrieved from the network scales linearly with the number of feature neurons[1,3] N_f. If one attempts to store more patterns than this upper limit, the shape of the energy landscape changes in such a way that it acquires many additional local minima that have nothing to do with the memory patterns that we are trying to store (Fig. 1a). These additional local minima are closely related to spin glass states, commonly studied in statistical physics of disordered systems. Their presence is undesirable for a proper function of the Hopfield network as an associative memory system. This linear scaling relationship between the dimension of the feature space and the memory storage capacity presents a problem from the perspective of machine learning applications, in which the dimension of the feature space is given by the specific data science task of interest. At the same time, the amount of patterns that the network needs to operate with is large, possibly much larger than the dimension of the feature space.

Modern Hopfield networks

In 2016, we realized that it is possible to overcome the linear scaling problem between the number of features and memory storage capacity by introducing a rapidly growing activation function[4]. The model, dubbed 'dense associative memory', is characterized by an energy function that includes higher than quadratic interactions between the features (Fig. 1b). The key difference between dense associative memory and the traditional Hopfield network is the presence of the rapidly growing activation function $F(\cdot)$. When this function is quadratic, dense associative memory reduces to the traditional Hopfield network. However, when the activation function grows more rapidly, as the state vector approaches one of the memory patterns, the network can achieve a super-linear memory storage capacity. This makes it an attractive tool for machine learning applications.

This idea has been further extended in 2017[5] by showing that a careful choice of the activation function can even lead to an exponential memory storage capacity. Importantly, the study[5] also demonstrated that dense associative memory, like the traditional Hopfield network, has large basins of attraction of size $O(N_\mathrm{f})$. This means that the new model continues to benefit from strong associative properties despite the dense packing of memories inside the feature space. Following Refs. 4 and 5, it became clear that the Hopfield network is not just one specific model, but rather a family of many models that can be classified into universality classes based on the asymptotic behaviour of the activation function $F(\cdot)$. Most of these new Hopfield networks have super-linear memory storage capacity and exhibit strong associative properties. Additionally, dense associative memories are not restricted to binary variables and can work with continuous variables too. For this reason, they can be smoothly integrated into deep learning architectures and trained with the back-propagation algorithm in an end-to-end fashion.

Hopfield networks meet transformers

In 2020, it was noticed[6] that if one picks the activation function $F(\cdot)$ so that its derivative is equal to the softmax function (known as Boltzmann distribution among physicists), the update rule for dense associative memory reduces to the attention mechanism, commonly used in transformers[7]. Transformers are a special kind of neural network architecture, which is responsible for many of the recent exciting results in large language models, foundation models, chat generative pre-trained transformer (ChatGPT), and others. Despite being one of the most popular deep learning models, transformers are typically designed through trial and error, and the theoretical principles behind their computational strategies remain mysterious. In contrast, Hopfield networks have a well-established record of theoretical methods, but have yet to demonstrate truly impressive empirical results in large-scale machine learning systems. The correspondence between dense associative memories and transformer's attention is interesting for two reasons. First, it enables investigation of pre-trained transformer models from the perspective of energy, basins of attraction, and other theoretical concepts commonly used in statistical physics. Second, it opens up the possibility of searching for new transformer-like architectures that are fundamentally designed around the idea of associative memory.

A study reported in a recent preprint[8] pursued this second idea and proposed replacing a sequence of conventional transformer layers

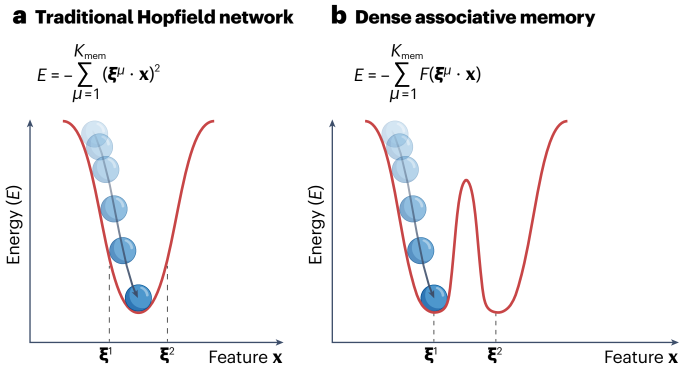

Fig. 1 | The comparative energy landscapes of the traditional and modern Hopfield networks. a, When the number of stored memories significantly exceeds the number of feature neurons, the traditional Hopfield network acquires spin glass local minima that are uncorrelated with the memory vectors (one such minimum is shown in panel **a**). **b**, In the dense associative memory model this spin glass transition happens at a much larger number of memories. Thus, even in situations when the number of memories is significantly larger than the number of feature neurons, each memory has a large basin of attraction around it, and there are no spin glass local minima. The memory patterns (vectors $\boldsymbol{\xi}^{\mu}$) are indexed by μ (going from 1 to the number of memory patterns K_{mem}) and each pattern is an N_f-dimensional vector. For continuous variables the feature vector **x** needs to additionally pass through a bounded activation function, such as sigmoid or layer normalization, to ensure that the energy E is bounded from below.

with a single dense associative memory. Transformers operate on a set of tokens, where each token represents a word in a sentence or a patch of an image, depending on the data domain. Since words in language don't appear in random order, there are certain rules that dictate how different words should be put together to form a meaningful sentence. The same applies to images. For example, if one is given an image patch that represents a left portion of the face, there needs to be another patch that represents the right portion of that face, and these two patches need to be properly located in the image plane. The idea of the new architecture[8], called energy transformer, is to treat these unwritten rules as Hopfield memories in a sophisticated energy landscape. When the tokens are arranged in a coherent pattern, the network assigns low energy to that state. In contrast, when the tokens are arranged in an incorrect way, the network assigns a high energy to that state. The memories are not hard-coded in advance, but rather are learned in a self-supervised way from the data. The network can handle many possible ways the tokens can be arranged together because it uses a dense associative memory network to memorize plausible arrangements of tokens. This approach provides a principled way to design new transformer-like architectures for language and image processing tasks that are grounded in the idea of associative memory.

The broader landscape

What insights can we gain from these recent theoretical advances? First, the longstanding problem of the coupling between memory capacity and feature space dimension in Hopfield networks has been resolved. In the practically relevant regime, the memory capacity of new Hopfield networks is restricted by the number of hidden neurons, rather than the feature space dimension[4,9]. Second, these networks can function in both discrete and continuous spaces. The continuous versions of dense associative memories[4,6,9] are particularly appropriate for use as part of other end-to-end deep learning architectures. Third, it is possible to include commonly used inductive biases such as convolutions, attention, and pooling in the general Hopfield-like framework. Lastly, large-memory-capacity Hopfield networks may be viewed as part of a broader class of energy-based models frequently discussed in the AI community[10] (see also this interview) and can act as a source of inspiration for new energy-based architectures rooted in associative memory ideas.

Dmitry Krotov
MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA, USA.
e-mail: krotov@ibm.com

Published online: 18 May 2023

References

1. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**, 2554–2558 (1982).
2. Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *PNAS* **81**, 3088–3092 (1984).
3. Amit, D. J., Gutfreund, H. & Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.* **55**, 1530 (1985).
4. Krotov, D. & Hopfield, J. J. Dense associative memory for pattern recognition. In *Advances in Neural Information Processing Systems 29* (NIPS, 2016).
5. Demircigil, M., Heusel, J., Löwe, M., Upgang, S. & Vermet, F. On a model of associative memory with huge storage capacity. *J. Stat. Phys.* **168**, 288–299 (2017).
6. Ramsauer, H. et al. Hopfield networks is all you need. In *International Conference on Learning Representations* (ICLR, 2021).
7. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30* (NIPS, 2017).
8. Hoover, B. et al. Energy transformer. Preprint at https://arxiv.org/abs/2302.07253 (2023).
9. Krotov, D. & Hopfield, J. J. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations* (ICLR, 2021).
10. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. & Huang, F. A tutorial on energy-based learning. In *Predicting Structured Data* (MIT Press, 2007).

Acknowledgements

I thank B. Hoover and J. Hopfield for numerous discussions over several years of the ideas described in this Comment.

Competing interests

The author declares no competing interests.