Statistics-guided Associative Memories

Hongzhi Wang, Satyananda Kashyap, Niharika Shimona D'Souza, Tanveer Syeda-mahmood
IBM Almaden Research Center, San Jose, CA 95120

Abstract

Content-associative memories such as Hopfield networks have been studied as a good mathematical model of the auto-associative features in the CA3 region of the hippocampal memory system. Modern Hopfield networks (MHN) are generalizations of the classical Hopfield networks with revised energy functions and update rules to expand storage to exponential capacity. However, they are not yet practical due to spurious metastable states leading to recovery errors during memory recall. In this work, we present a fresh perspective on associative memories using joint co-occurrence statistics, and show that accurate recovery of patterns is possible from a partially-specified query using the maximum likelihood principle. In our formulation, memory retrieval is addressed via estimating the joint conditional probability of the retrieved information given the observed associative information. Unlike previous models that have considered independence of features, we do recovery under the maximal dependency assumption to obtain an upper bound on the joint probability of occurrence of features. We show that this new approximation substantially improves associative memory retrieval accuracy on popular benchmark datasets.

1 Introduction

The hippocampal system in the brain is responsible for long-term declarative memory, which involves remembering facts and events [WC21]. For any incoming stimulus, the representations forming the memories are created within the trisynaptic circuit of the hippocampus consisting of distinct regions, which includes the CA3 region of cells. The existence of recurrent synaptic connections in CA3 led to the hypothesis that CA3 is an auto-associative network similar to the Hopfield network formulation of [Hop82]. Several key ideas have emerged from the theoretical analysis of Hopfield networks, and these have strongly influenced how neuroscientists analyze memory networks [AIL07].

Classical Hopfield networks can store a collection of multidimensional vectors, or memories as fixed point attractor states of a recurrent dynamical system through Hebbian learning [Hop82]. Despite their biological plausibility, Hopfield networks have not seen great adoption in content storage systems due to their limited storage capacity.

The Modern Hopfield Network was introduced as a continuous relaxation of the original Hopfield network from the 1980s and has been shown theoretically to have exponential storage capacity [KH16, DHL+17]. However, they are not yet practical because they have a predilection to enter spurious metastable states, leading to memorizing bogus patterns even while storing a small number of inputs.

In this paper, we revisit associative memories to offer new insights based on joint statistical modeling to mitigate some of the limitations of Hopfield networks. Specifically, we show that by storing joint statistics of neuronal activation as an aggregation of information contained in the patterns, we can later recover accurately when partial evidence for the pattern is provided as a query.

Associative Memory & Hopfield Networks in 2023. NeurIPS 2023 workshop.

2 Associative memory modeling using Hopfield network

Associative memory [Pal80, AM88] models the relationship between different entities. It allows information retrieval for one entity given the information of the associated entity. The Hebbian rule is a simple yet effective method for neural network learning that is commonly used for associative memory [Heb05]. It is based on an intuitive concept that neurons firing together should be connected together. Given a binary dataset $X = \{x\}$, the Hebbian rule can be formulated as:

$$w^{ij} = \sum_{x \in X} x^i x^j \tag{1}$$

, where $x^i \in [-1, 1]$ is the i^{th} feature for sample x. The Hebbian rule was applied in several early associative memory models such as Associatron [Nak72] besides the Hopfield network [Hop82]. Once the learning is complete, given a neural input x, the Hopfield network updates the neural states in an iterative fashion. At each iteration, the state of the i^{th} neuron is updated by:

$$x^i = \begin{cases} 1 & \text{if } \sum_{j \neq i} w^{ij} x^j \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

When patterns are correctly stored, the converged state corresponds to one stored pattern. In a recent development, the modern Hopfield network (MHN) significantly improves the theoretical memory capacity by replacing the quadratic energy function used by the classical Hopfield network with higher order energy functions [KH16].

2.1 The metastable states in Hopfield Network

Despite the large theoretical memory capacity, MHN is highly sensitive to the choice of parameters, which may vary widely across different data representations [BR90, RSL$^+$20, BBF18]. Furthermore, the Hopfield networks are known to enter spurious attractors states due to a combination of similar looking patterns [RSL$^+$20]. Since Hopfield networks rely on evaluated similarity between queries and stored memories for accurate associative memory retrieval, poor direct image similarity assessments tend to produce weak basins of attraction in the Hopfield energy landscape space, leading to metastable formulations. This happens even when storing a small number of patterns as shown in Fig. 2.

3 Statistics-guided associative memories

In this work, we revisit the classical Hopfield network by reformulating the Hebbian rule in a statistical setting. Basically, the Hebbian rule captures the joint statistics of the input features, which can be expressed through joint feature histograms.

$$w^{ij} = \sum_{x \in X} x^i x^j = \sum_{v^i \in \{-1,1\}, v^j \in \{-1,1\}} v^i v^j h_{i,j}(x^i = v^i, x^j = v^j) \tag{3}$$

, where $h_{i,j}$ is the joint feature histogram of i^{th} and j^{th} feature over X.

Given a non-binary dataset $X = \{x\}$ we can extend the Hebbian rule using joint feature statistics. Let h_{ij} be the first order joint histogram between the i^{th} and the j^{th} feature over X.

$$h_{ij}(v^i, v^j) = h_{ji}(v^j, v^i) = \sum_{x \in X} I(x^i = v^i) I(x^j = v^j) \tag{4}$$

I is an indicator function, which equals 1 when the statement is true and 0 otherwise. For an input vector x, let x^O and x^U be the visible and hidden features, respectively. A hidden feature $u \in U$ can be inferred from a visible feature based on the first order joint histograms by:

$$p(x^u | x^i, i \in O) = \frac{h_{iu}(x^i, x^u)}{\sum_v h_{ih}(x^i, x^u = v)} \propto h_{iu}(x^i, x^u) \tag{5}$$

For more reliable estimation, the hidden feature should be inferred based on all visible features $p(x^u|x^O)$. Although this high order conditional probability is difficult to calculate, it can be approximated using the first order conditional probability (5). For example, under the independence assumption, the product rule can be applied:

$$p(x^u | x^O) \sim \prod_{i \in O} p(x^u | x^i) \tag{6}$$

The Hopfield network can be interpreted as applying a summation rule for the approximation so that to infer a hidden feature from one visible feature we get

$$\bar{x}^{u|i} = w^{iu}x^i = \sum_{x^u=-1}^{1} x^u p(x^u|x^i) \tag{7}$$

To infer a hidden feature using all visible features, a simple summation rule is applied as:

$$\bar{x}^{u|O} = \sum_{i \in O} \bar{x}^{u|i} \tag{8}$$

This is equivalent to applying a summation rule to estimate the joint conditional probability with:

$$p(x^u|x^O) \propto \sum_{i \in O} p(x^u|x^i) \tag{9}$$

With the estimated conditional probability, the hidden feature can be determined by:

$$x^u = \operatorname*{argmax}_v p(x^u = v|x^O) \tag{10}$$

3.1 Maximal dependency for high order conditional probability approximation

For data with strong inter-feature dependencies, the independent assumption may lead to less accurate approximation. To address this problem, we propose the approximation derived from the maximal inter-feature dependency assumption for associative memory applications. The following equations show the derived upper bound using first order joint histograms.

$$p(x^u|x^O) = \frac{p(x^u, x^O)}{p(x^O)} \leq \frac{\min_{i \in O} p(x^u, x^i)}{p(x^O)} \propto \min_{i \in O} h_{ui}(x^u, x^i) \tag{11}$$

Unlike the product and summation rules that assume inter-feature independence, the upper bound approach assumes maximal dependency among visible features, which may better fit imaging data.

4 Experiments

We evaluated the proposed method for associated memory retrieval on images.

In the first experiment, we focused on comparing the effectiveness of the above mentioned higher order probability approximation methods. Hence, retrieval was done directly through the maximum likelihood principle in (10), without iterative updates. The CIFAR-100 dataset (50000 colored training images of size 32×32) was used. The images were converted to gray scale with integer value in the range $[0, 100]$. To simulate performance at various memory capacities, a various number of images (i.e. 1000, 2500, 5000, 10000, 20000, 50000) were stored, respectively. For retrieval test, every pixel within the center region of size 16×16 was hidden and the rest of the image were visible.

For each size of stored images, all pairwise first order joint histograms were calculated using the stored images, which resulted in $1024 \times 1023/2$ unique first order histograms in total. For evaluation, two metrics were used: 1) the number of images with at least one hidden pixel incorrectly inferred, and 2) the average percentage of incorrectly inferred pixels in each image with retrieval errors.

# of stored images	1000	2500	5000	10000	20000	50000
Summation rule (9)	30/7%	210/25%	2928/12%	9866/30%	20000/58%	50000/80%
Product rule (6)	0/0	2/1%	7/5%	292/4%	14223/16%	49993/68%
Upper Bound (11)	0/0	0/0	0/0	0/0	1969/4%	45355/70%

Table 1: Retrieval performance by the three joint conditional probability estimation methods. The number of images used is shown in the first row. The results are shown in total number of images with retrieval errors / percentage of incorrectly retrieved pixels within each incorrectly retrieved image.

The results are summarized in Table 1. The summation rule for joint conditional probability estimation performed the worst, with over 8% images incorrectly retrieved when 2500 images were stored. The

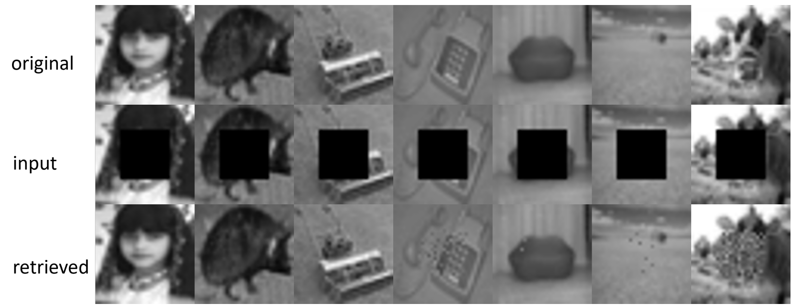

Figure 1: Example retrieval results produced by our method. The first six examples were tested with 20000 images. The three examples on the left were correctly retrieved, while the next three images have retrieval errors. The last column shows retrieving errors when tested with 50000 images.

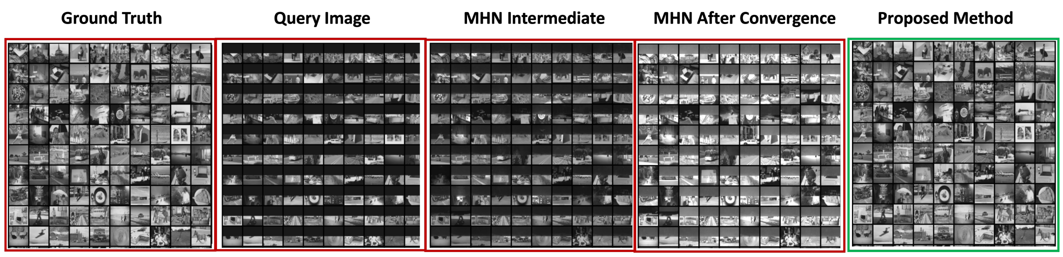

Figure 2: The following sequence showcases image retrieval results by the proposed method and the Modern Hopfield Network (MHN). Left to right - original image, the query image, an intermediate update step, the final MHN reconstruction, and the proposed method. For most images, MHN incorrectly converged to metastatble states, while the proposed method made no retrieval mistake.

error kept increasing as the number of images used in the test increased. The product rule substantially outperformed the summation rule. Our upper bound approximation consistently performed the best. Fig. 1 shows a few retrieval results produced by our method.

To compare with the modern Hopfield network (MHN) [KH16], we conducted a second retrieval experiment using 6000 images from the MS-COCO dataset [LMB+14]. Again, these images were converted to grayscale and resized to 32×32. For this test, each image's upper half was hidden. For MHN, a dot product-based energy formulation proposed in [KH20] was used. We pre-processed the images in the same way as previously described and set our β temperature to 150. Recurrent updates were repeated for 300 iterations, and we compared the final set of recovered images with the originals in the memory banks. If there were no meta-stable states, the mean squared error (MSE) and the 1-structural similarity index (SSIM) would be zero. However, we observed that the MHN energy updates had metastable states (as shown in Figure 2), resulting in 1-SSIM and MSE values of 0.441 and 0.058, respectively. In contrast, the proposed method displayed no meta-stable states with a 1-SSIM and MSE of 0.0, confirming the advantage of the proposed method in image associative memory retrieval.

5 Conclusions

We revisited the Hebbian rule and extended it with joint feature statistics for associative memory. Higher order joint statistics is critical for accurate memory retrieval. Both the product rule and the summation rule were applied to approximate higher order joint statistics using lower order statistics under the often oversimplified independent assumption. We proposed to use the upper bound approximation derived under the maximal dependency assumption as a more accurate alternative. For testing associative memory retrieval on natural images, we showed that the proposed approximation significantly improved the retrieval accuracy.

References

[AIL07] Licurgo De Almeida, Marco Idiart, and John E. Lisman. Memory retrieval time and memory capacity of the ca3 network: Role of gamma frequency oscillations. *Learning Memory*, 14:795, 11 2007.

[AM88] Shun-Ichi Amari and Kenjiro Maginu. Statistical neurodynamics of associative memory. *Neural Networks*, 1(1):63–73, 1988.

[BBF18] Adriano Barra, Matteo Beccaria, and Alberto Fachechi. A new mechanical approach to handle generalized hopfield neural networks. *Neural Networks*, 106:205–222, 2018.

[BR90] Jehoshua Bruck and Vwani P Roychowdhury. On the number of spurious memories in the hopfield model (neural network). *IEEE Transactions on Information Theory*, 36(2):393–397, 1990.

[DHL+17] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

[Heb05] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

[Hop82] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[KH16] Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

[KH20] Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*, 2020.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Nak72] Kaoru Nakano. Associatron-a model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):380–388, 1972.

[Pal80] Günther Palm. On associative memory. *Biological cybernetics*, 36(1):19–31, 1980.

[RSL+20] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

[WC21] Christopher D Wickens and C Melody Carswell. Information processing. *Handbook of human factors and ergonomics*, pages 114–158, 2021.