Modern Hopfield Networks with Continuous-Time Memories

Saul Santos[1] António Farinhas[1]
Daniel McNamee[2] André F.T. Martins[1,3]

[1]Instituto de Telecomunicações, IST, Lisbon ELLIS Unit, Portugal
[2]Champalimaud Research, Lisbon, Portugal [3]Unbabel, Lisbon, Portugal

Outline

Inspired by **continuous neural resource allocation** (Ma et al., 2014), we propose a **continuous-time memory** mechanism for **modern Hopfield networks** (MHNs), replacing discrete memories with compressed continuous representations.

- Modifies MHNs' energy function (Ramsauer et al., 2020) for **compressed memory storage**.
- Leverages a *probability density function* (PDF), recovering Martins et al. (2022)'s ∞-memory attention.
- Maintains retrieval performance with smaller memory.
- Validated on synthetic and video datasets for efficient associative retrieval.

Modern Hopfield Networks (Ramsauer et al., 2020)

Memory patterns $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, forming a matrix $\boldsymbol{X} \in \mathbb{R}^{L \times D}$, query $\boldsymbol{q}$.
Energy function:

$$E(\boldsymbol{q}) = -\frac{1}{\beta}\log \sum_{i=1}^{L} \exp(\beta \boldsymbol{x}_i^\top \boldsymbol{q}) + \frac{1}{2}\|\boldsymbol{q}\|^2 + \text{const.}$$

Update rule (via CCCP): $\boldsymbol{q}_{t+1} = \boldsymbol{X}^\top \mathrm{softmax}(\beta \boldsymbol{X}\boldsymbol{q}_t)$.

Continuous Attention (Martins et al., 2020)

- Traditional attention: operates on discrete data (e.g., words, pixels).
- Continuous attention: handles inherently continuous signals (e.g., speech, video).
- Attention defined over continuous-time signal $\bar{x}(t)$, evolving smoothly:

$$\boldsymbol{c} = \mathbb{E}_p[\boldsymbol{v}(t)] = \int p(t)\boldsymbol{v}(t)\,dt.$$

- $\boldsymbol{v}(t)$: continuous value function.
- $p(t)$: PDF instead of probability mass function .
- For $p(t)$, Martins et al. (2022): $\mathcal{N}(t;\mu,\sigma^2)$, where μ, σ^2 are input-dependent.

Ours: Continuous-Time Memory Hopfield Networks

We model sequences $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \dots, \boldsymbol{x}_L^\top] \in \mathbb{R}^{L \times D}$ as samples from a smooth function $\boldsymbol{x}(t)$, reconstructed using basis functions $\boldsymbol{\psi}(t) \in \mathbb{R}^N$ and learned coefficients $\boldsymbol{B} \in \mathbb{R}^{N \times D}$:

$$\bar{x}(t) = \boldsymbol{B}^\top \boldsymbol{\psi}(t).$$

Coefficients $\boldsymbol{B}$ are obtained via ridge regression:

$$\boldsymbol{B}^\top = \boldsymbol{X}^\top \boldsymbol{F}^\top (\boldsymbol{F}\boldsymbol{F}^\top + \lambda \boldsymbol{I})^{-1},$$

where $\boldsymbol{F} = [\boldsymbol{\psi}(t_1), \dots, \boldsymbol{\psi}(t_L)]$ evaluates basis functions at assigned time points $t_1, \dots, t_L \in [0, 1]$.

This provides a compressed representation with $N \ll L$.

Ours: Continuous Hopfield Energy and Update

Define the energy over queries $\boldsymbol{q} \in \mathbb{R}^D$:

$$E(\boldsymbol{q}) = -\frac{1}{\beta}\log \int_0^1 \exp(\beta \bar{\boldsymbol{x}}(t)^\top \boldsymbol{q})dt + \frac{1}{2}\|\boldsymbol{q}\|^2 + \text{const.}$$

Minimizing this energy yields the update with Gibbs PDF:

$$\boldsymbol{q}^{(i+1)} = \boldsymbol{B}^\top \int_0^1 p(t)\boldsymbol{\psi}(t)\,dt \quad \text{with} \quad p(t) = \frac{e^{\beta \bar{\boldsymbol{x}}(t)^\top \boldsymbol{q}^{(i)}}}{\int_0^1 e^{\beta \bar{\boldsymbol{x}}(t')^\top \boldsymbol{q}^{(i)}} dt'}.$$

Experiments

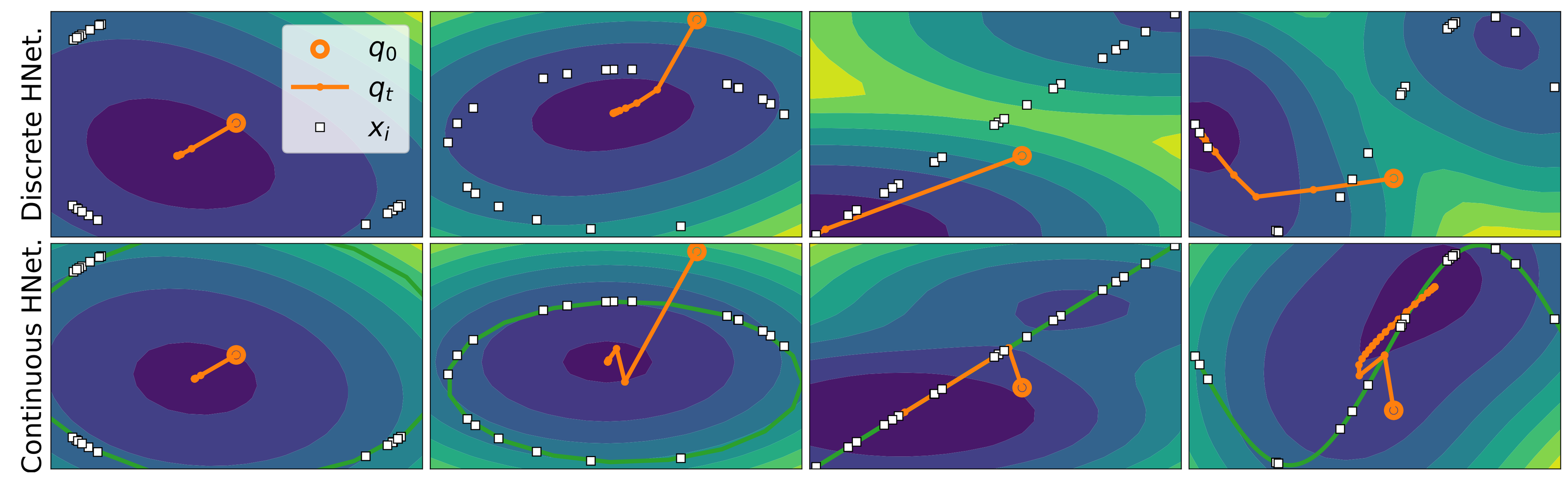

Fig. 1: Optimization trajectories and energy contours for Hopfield networks with discrete **(top)** and continuous memories **(bottom)**. Green illustrates the continuous function shaped by discrete memory points, while darker shades of blue indicate lower energy regions.

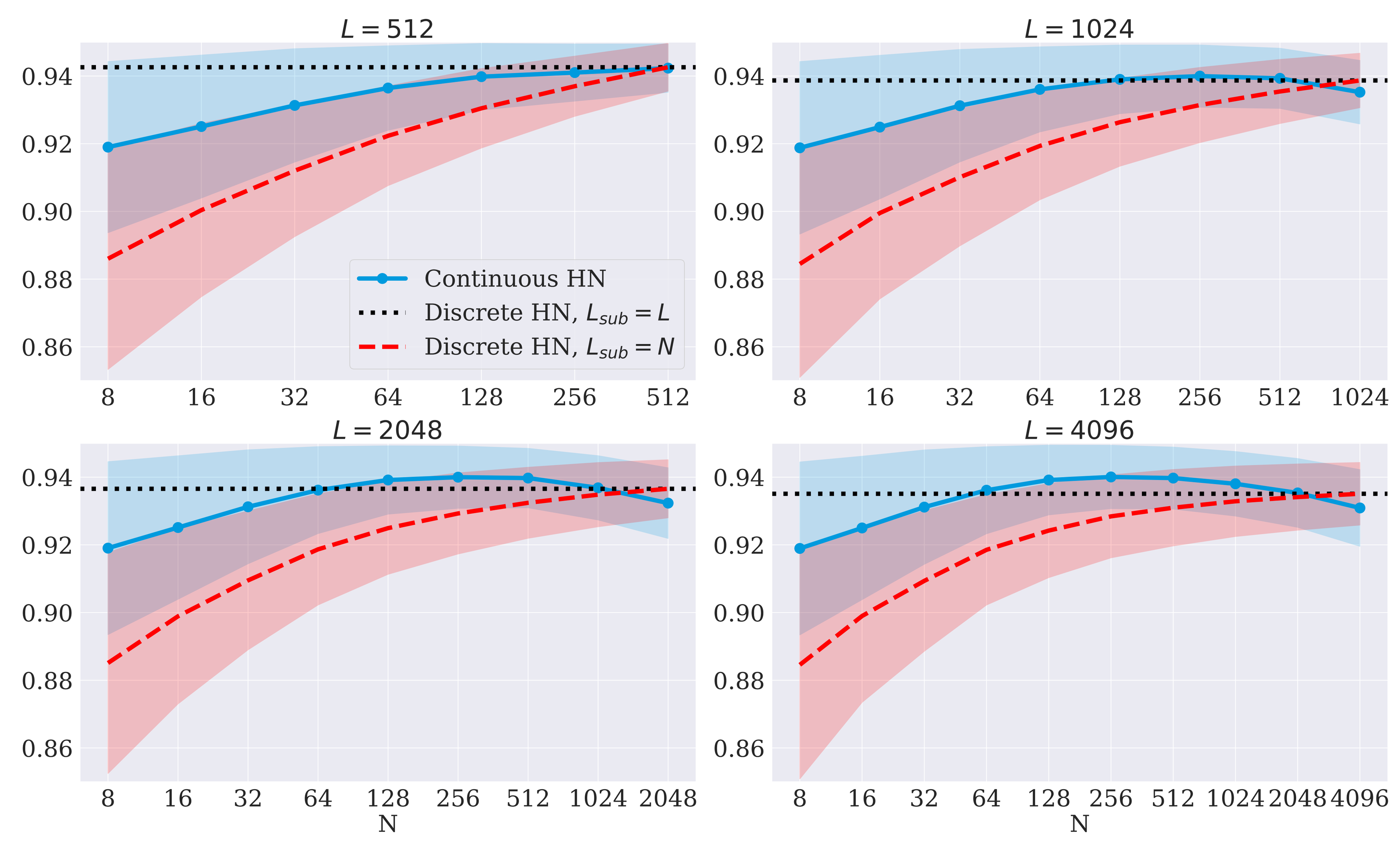

Fig. 2: Video embedding retrieval performance across different numbers of basis functions. Plotted are the cosine similarity means and standard deviations across videos.

Our method achieves competitive retrieval performance while using smaller compressed memories!

References

Ma, W. J., Husain, M., and Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356.

Martins, A., Farinhas, A., Treviso, M., Niculae, V., Aguiar, P., and Figueiredo, M. (2020). Sparse and continuous attention mechanisms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20989–21001. Curran Associates, Inc.

Martins, P. H., Marinho, Z., and Martins, A. F. (2022). ∞-former: Infinite memory transformer. In *Proc. ACL*.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., et al. (2020). Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Acknowledgments: ERC (DECOLLAGE, ERC-2022-CoG 101088763), Portuguese PRR (Center for Responsible AI, C645008882- 00000055), FCT (contract UIDB/50008/2020).