Dense Hopfield networks in the teacher-student setting

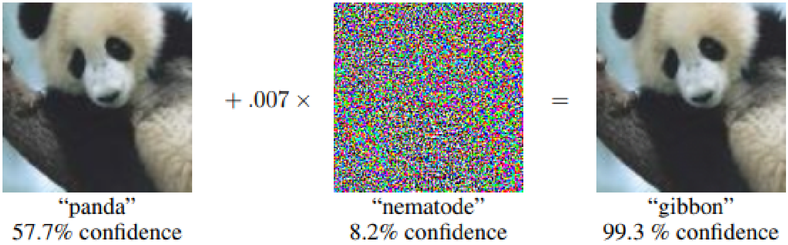
Robin Thériault
Supervised by Daniele Tantari

October 31, 2024

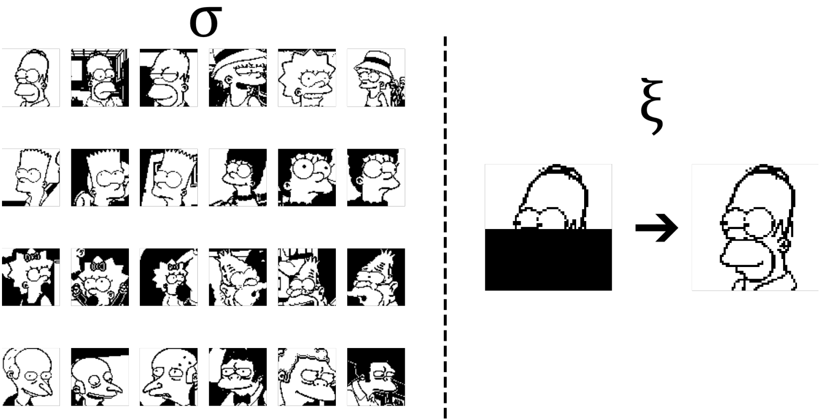
arXiv 2401.04191

Adversarial attack

Neural networks are fooled by small adversarial perturbations.



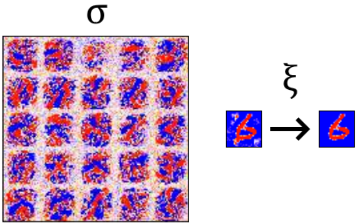
- Task: restore incomplete pattern ξ with N pixels using stored memories σ^a.
- Robust to adversarial perturbations.



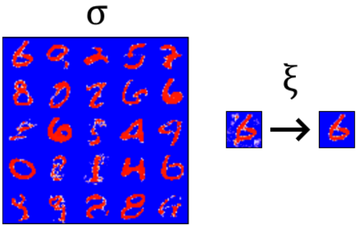
Dense Hopfield Network

Memories can be learned from data.

Feature learning



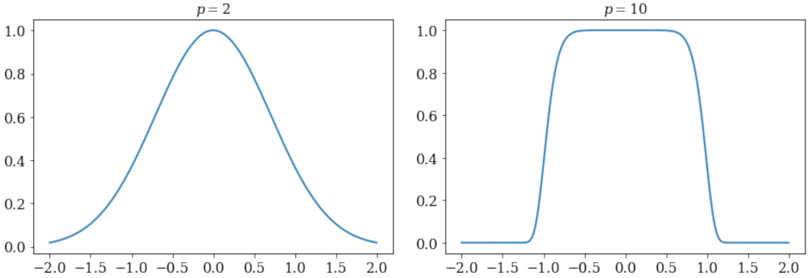
Example retrieval



Example retrieval empirically known to be robust.

Direct problem

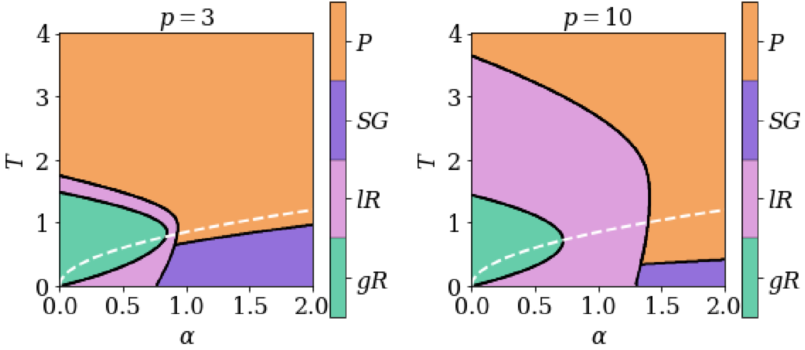
Model task: retrieve memories σ^a by sampling
$\xi \sim P(\xi|\sigma; p, T) = \mathcal{Z}^{-1} \exp(-E[\xi]/T)$ where
$E[\xi] = -\frac{p!}{N^{p-1}} \sum_a \sum_{i_1<\ldots<i_p} \xi_{i_1} \ldots \xi_{i_p} \sigma^a_{i_1} \ldots \sigma^a_{i_p}$.



- Well-studied theoretically.
- Capacity of $M \sim \mathcal{O}\left(N^{p-1}\right)$ i.i.d. random memories.

Direct problem

Parameters: temperature T and memory load $\alpha = \frac{Mp!}{N^{p-1}}$.



Caveats:

- Does not explain adversarial robustness.
- Does not capture the feature regime.
- Does not say how much data a dense net needs.
- SG transition hard to obtain exactly.

Teacher-student setting

Teacher-student setting:

- Sample examples $\sigma = \{\sigma^a\}_{a=1}^M$ from teacher with pattern ξ^*: $\sigma^a \sim P\left(\sigma^a | \xi^*; p_{\text{teacher}}, T_{\text{teacher}}\right)$.
- Sample pattern ξ from student with memories $\sigma = \{\sigma^a\}_{a=1}^M$: $\xi \sim P\left(\xi | \sigma; p_{\text{student}}, T_{\text{student}}\right)$.

Student task: recover ξ^*.

Our goal: predict student performance as a function of p_{teacher}, T_{teacher}, p_{student}, T_{student} and α (normalized M).

Teacher-student setting

Compute the order parameters m, q and q^*, where

- $m = \lim_{N\to\infty} \left\langle \frac{1}{N} \sum_i \xi_i \sigma_i^a \right\rangle_{\xi^*,\sigma,\xi}$,
- $q = \lim_{N\to\infty} \left\langle \frac{1}{N} \sum_i \xi_i^1 \xi_i^2 \right\rangle_{\xi^*,\sigma,\xi}$,
- $q^* = \lim_{N\to\infty} \left\langle \frac{1}{N} \sum_i \xi_i^* \xi_i \right\rangle_{\xi^*,\sigma,\xi}$.

m measures the proximity of the student pattern and memories.

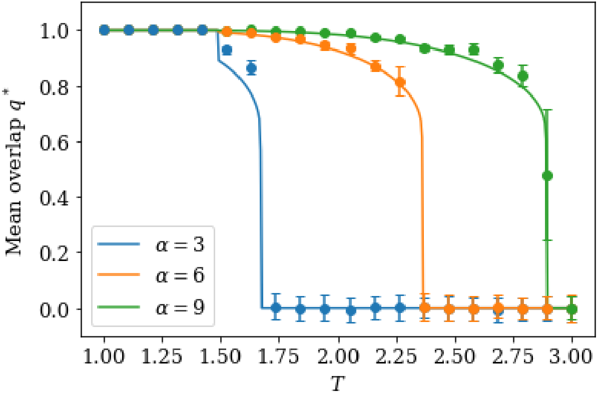
q measures the tendency of the student to stay frozen in specific configurations.

q^* measures the student performance.

Phase diagrams

Consider $p_{\text{teacher}} = p_{\text{student}}$ and $T_{\text{teacher}} = T_{\text{student}}$.

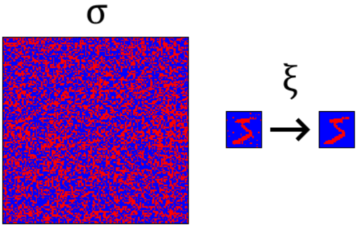
- Study the performance q^*.
- Plot: calculations (lines) consistent with simulations (dots).



Phase diagrams

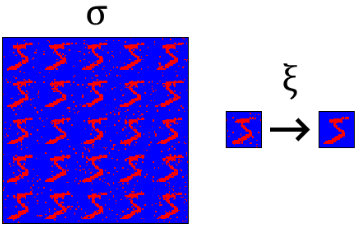
Obtain feature learning (gR/lR) and example retrieval (eR) phases.

gR/lR



$q^* = q \neq 0$ and $m = 0$

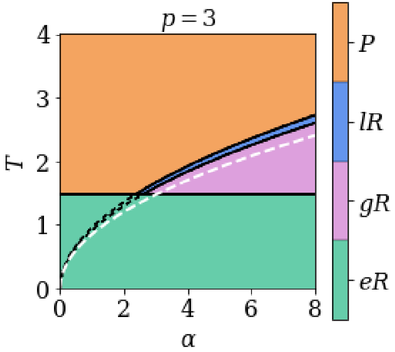
eR



$q^* = q \neq 0$ and $m \neq 0$

Phase diagrams

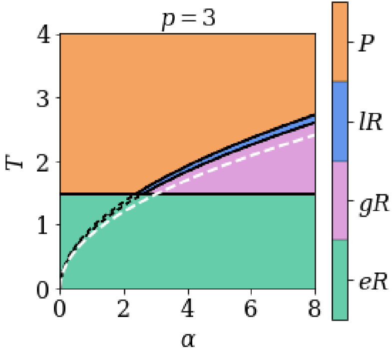
Obtain feature learning (gR/lR) and example retrieval (eR) phases.



- P: $q^* = q = 0$ and $m = 0$
- gR/lR: $q^* = q \neq 0$ and $m = 0$
- eR: $q^* = q \neq 0$ and $m \neq 0$

Phase diagrams

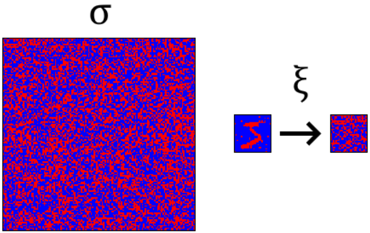
Obtain feature learning (gR/lR) and example retrieval (eR) phases.



The gR transition overlaps with the direct model SG transition.
arXiv 2401.04191

Phase diagrams

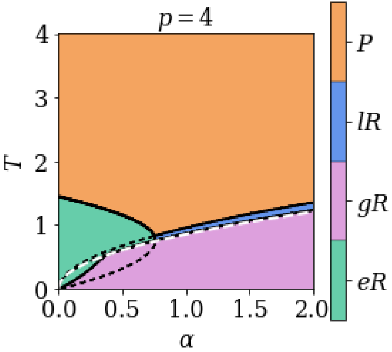
- Next: $p_{\text{teacher}} = 2$ and $p_{\text{student}} \geq 3$.
- In this regime, teacher examples are noisy, so eR is inaccurate.



Phase diagrams

Case 1:

- $M \sim \mathcal{O}\left(N^{p-1}\right)$
- $T_{\text{teacher}} \sim \mathcal{O}\left(N^{1-2/p}\right)$, i.e. can infer ξ^* despite large T_{teacher}.

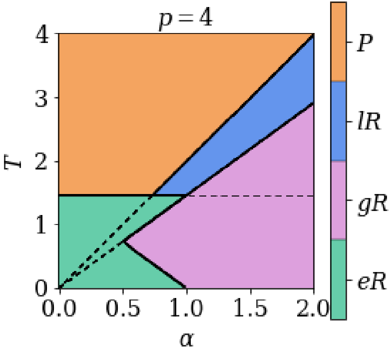


Extensive tolerance to teacher noise.

arXiv 2401.04191

Case 2:

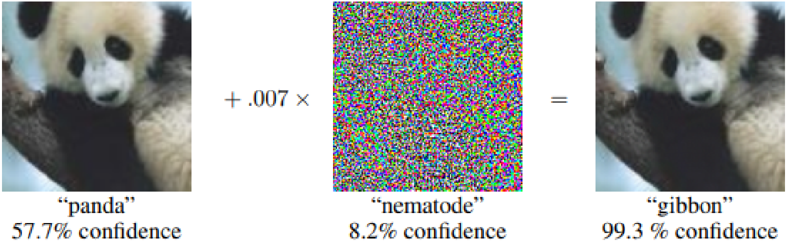
- $T_{\text{teacher}} \sim \mathcal{O}(1)$
- $M \sim \mathcal{O}\left(N^{p/2}\right)$, i.e. much smaller M than in the first case.



Avoids pattern interference.
arXiv 2401.04191

Adversarial attacks

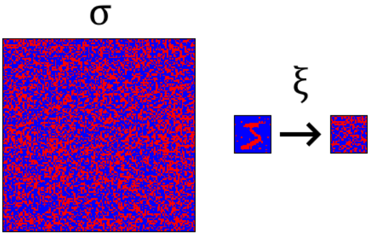
Neural networks are fooled by small adversarial perturbations.



Neural networks with more parameters are more robust.

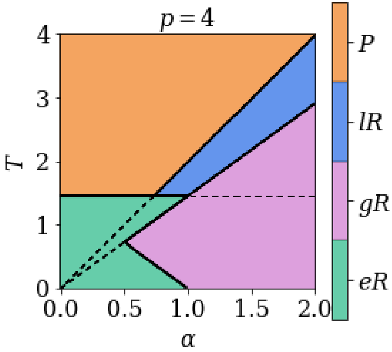
Adversarial attacks

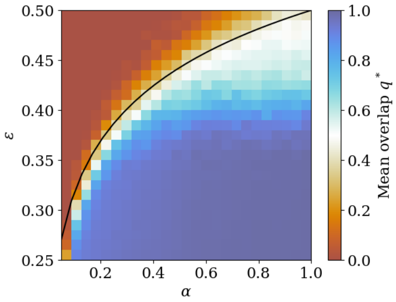
When $p_{\text{teacher}} = 2$ and $p_{\text{student}} \geq 3$, memories σ^a are noisy, so eR is inaccurate.



Adversarial attacks

Study adversarial attacks of size ε at $T_{\text{student}} = 0$.



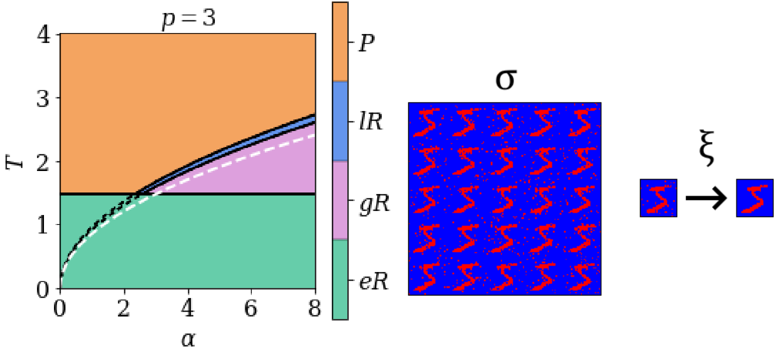


- Right plot: calculations (line) agree with simulations (image).
- Adversarial robustness increases with the memory load α, in line with empirical observations of neural networks.

arXiv 2401.04191

Adversarial attacks

Adversarial attacks fail when $p_{\text{teacher}} = p_{\text{student}}$ and $T_{\text{teacher}} = T_{\text{student}}$ because eR is accurate.



Clarifies why accurate example retrieval is adversarially robust.
arXiv 2401.04191

Summary of results

- The teacher-student feature learning phase transition overlaps the direct problem spin-glass phase transition.
- Feature learning can resist extensive teacher noise and pattern interference.
- As seen in neural networks, feature learning becomes more adversarially robust as the memory load increases.
- Our model clarifies why the example retrieval phase of dense Hopfield networks is adversarially robust.

arXiv 2401.04191

Dense Hopfield model

Sampling: $\xi \sim P\left(\xi|\sigma;p,T\right) = \mathcal{Z}^{-1}\exp\left(-E\left[\xi\right]/T\right)$ where

- $E\left[\xi\right] = -\frac{p!}{N^{p-1}}\sum_a \sum_{i_1<\ldots<i_p} \xi_{i_1}\ldots\xi_{i_p}\sigma^a_{i_1}\ldots\sigma^a_{i_p}$,
- $P\left(\xi|\sigma;p,T\right)$ is sampled with a Monte-Carlo simulation.
- $E\left[\xi\right]$ reduces to $E\left[\xi\right] = -\frac{1}{N}\sum_a \sum_{i\neq j} \xi_i\xi_j\sigma^a_i\sigma^a_j$ when $p=2$,
- MC reduces to $\xi_i = \text{sign}\left(\sum_a \sigma^a_i \sum_j \xi_j\sigma^a_j\right)$ when $T=0$.

Overlapping gR and SG lines

Sampling: $\xi \sim P\left(\xi|\sigma;p,T\right) = \mathcal{Z}^{-1}\left(\sigma\right)\exp\left(-E\left[\xi|\sigma\right]/T\right)$ where $\mathcal{Z}\left(\sigma\right)$ is a normalization constant.

The gR line overlaps with the SG line because

- $P\left(\sigma\right) = \frac{1}{2^{MN}}\frac{\mathcal{Z}(\sigma)}{\langle\mathcal{Z}\rangle}$,
- $\lim_{N\to\infty}\left\{\frac{\log\mathcal{Z}-\log\langle\mathcal{Z}\rangle}{N}\right\} = 0$ in the paramagnetic phase.

Adversarial robustness formula

Adversarial boundary: $\varepsilon^* = \frac{[\eta\alpha]^{\frac{1}{p-1}}}{[\eta\alpha]^{\frac{1}{p-1}}+1}$ where $\eta = \frac{2[\beta^*]^2}{(1-2\beta^*)^2}$ when $p = 4$.

Bibliography I

- F. Alemanno, L. Camanzi, G. Manzan, and D. Tantari, "Hopfield model with planted patterns: A teacher-student self-supervised learning model," *Applied Mathematics and Computation*, vol. 458, p. 128253, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0096300323004228
- E. Gardner, "Multiconnected neural network models," *Journal of Physics A: Mathematical and General*, vol. 20, no. 11, p. 3453, 1987.
- D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," *Advances in neural information processing systems*, vol. 29, 2016.
- D. Krotov and J. Hopfield, "Dense associative memory is robust to adversarial inputs," *Neural computation*, vol. 30, no. 12, pp. 3151–3167, 2018.

Bibliography II

I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

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

J. Puigcerver, R. Jenatton, C. Riquelme, P. Awasthi, and S. Bhojanapalli, "On the adversarial robustness of mixture of experts," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 9660–9671. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/3effb91593c4fb42b1da1528328eff49-Paper-Conference.pdf