Memorization and consolidation in associative memory networks

Danil Tyulmankov
Columbia University
dt2586@columbia.edu

Kimberly Stachenfeld
DeepMind
Columbia University
stachenfeld@deepmind.com

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

LF Abbott
Columbia University
lfa2103@columbia.edu

Abstract

Humans, animals, and machines can store and retrieve long-term memories of individual items, while at the same time consolidating and learning general representations of categories that discard the individual examples from which the representations were constructed. Classical neural networks model only one or the other of these two regimes. In this work, we propose a biologically motivated model that can not only consolidate representations of common items but also memorize exceptional ones. Critically, we consider the unsupervised learning regime where exceptional items are not labeled as such a priori, so the signal to either memorize or consolidate items must be generated by the network itself. We propose a number of metrics for this control signal and compare them for two different algorithms inspired by traditional imbalanced data learning approaches – loss reweighting and importance sampling. Overall, our model serves not only as a framework for concurrent memorization and consolidation processes in biological systems, but also as a simple illustration of related phenomena in large-scale machine learning models, as well as a potential method for debiasing artificial intelligence algorithms.

1 Introduction

In human and animal experience, long-term memories of individual items can coexist with generalized representations of categories. Individual items are, by definition, stored verbatim through *memorization*, such that the specific details of the item can be recovered when needed. In other cases, it can be advantageous to blend multiple examples from a category to create a flexible representation, with the specifics of the individual examples being forgotten after *consolidation*. These processes are two essential but often opposing aspects of flexibly representing knowledge about the environment.

Classical models typically operate in only one of these modes. For example, Hopfield networks [1] can memorize and retrieve a predetermined set of datapoints, whereas deep networks [2] learn a compact representation of a large dataset for classification, regression, or generation. In this work, we describe a model and learning paradigm that can flexibly perform both of these processes simultaneously, memorizing examples of rare items and learning consolidated representations of common ones.

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

2 Dataset

To operationalize this scenario, we consider the problem of storing a dataset consisting of R "regular" items from a commonly occurring class, of which the learner sees many examples; and E "exceptions," which comprise a few examples from an uncommon class:

$$\mathcal{D} = \mathcal{R} \cup \mathcal{E} = \{x_1, \ldots, x_R\} \cup \{x_{R+1}, \ldots, x_{R+E}\}, \text{ with } R \gg E \tag{1}$$

As in the classical autoassociative memory task, the goal of the network is to simply store the dataset such that individual items can then be retrieved from partial cues, i.e. perturbed versions of the original items. Unlike the classical task, however, the dataset size is much larger than the memory capacity of the network, so to successfully encode the dataset it is necessary to perform some sort of compression (consolidation). Since the goal is to simply store the data as reliably as possible, we do not consider generalization performance on a held-out dataset, although we note that consolidation of stored data may indeed have advantages for generalization.

In this case, there is no binary notion of whether an item was stored. We instead evaluate performance by considering the mean-squared error between stored items (the network's fixed points) and the datapoints in $\mathcal{D}$. We consider the average errors on the regular and exceptional items separately and weight them equally:

$$\mathcal{L}_{\text{eval}} = \frac{1}{2}\left(\frac{1}{R}\sum_{i\in\mathcal{R}} \ell(x_i, g_i^*) + \frac{1}{E}\sum_{i\in\mathcal{E}} \ell(x_i, g_i^*)\right) \tag{2}$$

where g_i^* is the retrieved item corresponding to datapoint x_i, and $\ell(y, z) = ||y - z||_2^2$ for vectors y, z.

This models the scenario in which, although the exceptions are rare, it is just as important to store them as the common category. Critically, the data are not labeled as such, and the learner does not know in advance which items belong to the regular class or the exceptions. These must be inferred online over the course of learning. This furthermore allows the possibility of treating unusual examples of regular items as exceptions and storing them verbatim, which can play a role analogous to data pruning [3] or label memorization in learning long-tailed data distributions [4].

3 Associative memory network model

We consider a framework inspired by the complementary learning systems theory [5], [6] in which a fast learner (e.g. the hippocampus) rapidly accumulates data, which it then "replays" [7] to a slow learner (e.g. the cortex) for consolidation and long-term storage. We do not explicitly model the fast learner; we simply assume that there exists a mechanism for storing a dataset in a buffer from which minibatches can be assembled for replay to the slow learner. Such a buffer might be implemented biologically by a fast memory system such as the Hopfield network with Hebbian learning rules and forgetting [1], [8], or a key-value memory network with three-factor learning rules and rapid overwriting [9].

To model the slow learner, we consider a case of Dense Associative Memory, also known as the modern Hopfield network (MHN) [10]–[12], consisting of a visible layer with activity given by the vector v and a hidden layer h. The dynamics of this network are given by [11]:

$$\begin{aligned} \tau_h \frac{dh}{dt} &= -h + Wg(v) \\ \tau_v \frac{dv}{dt} &= -v + W^\top f(h) \end{aligned} \tag{3}$$

Extending the set of models described in [11], we introduce a new model in which the visible layer activity is restricted to lie on a unit sphere and the hidden layer has an attention mechanism, represented by the transfer functions $g(v) = \frac{v}{||v||}$ and $f(h) = \text{softmax}(\beta h)$. Biologically these might be implemented by divisive normalization [13], lateral inhibition [14], or intermediate neurons [15]. The weight matrix $W \triangleq [\xi_1, \ldots \xi_M]^\top$ is normalized such that $||\xi_\mu|| = 1$, ensuring that the softmax inverse-temperature parameter β has a well-defined scale.

To run the network, we generate a perturbed input $\widetilde{x}_i$, corresponding to an item x_i from the dataset with a random 50% of the pixels set to zero. This is used as the initial state of the network $g(t = 0)$.

The unperturbed pixels are clamped to their initial values, and the rest evolve according to Eq. 3, discretized with $\tau_v = 1, dt = 0.05$. Assuming the dynamics of the hidden layer are much faster than those of the visible layer (adiabatic limit [16]) we set $\tau_h = 0$, so that $h = Wg(v)$ for all t. We run the network until it reaches a fixed point, so $g(T + dt) = g(T) \triangleq g^*$, which is the network output.

4 Learning with unlabeled imbalanced data

As the main contribution of this work, we propose a biologically motivated implementation of two methods for learning imbalanced data – loss reweighting and importance sampling [17]. Critically, unlike traditional supervised machine learning approaches where class labels (and therefore relative frequencies) are known a priori, we consider autoassociative recall, an unsupervised learning setting where the class labels are not known, and the loss weights or sampling probabilities must be inferred online during learning.

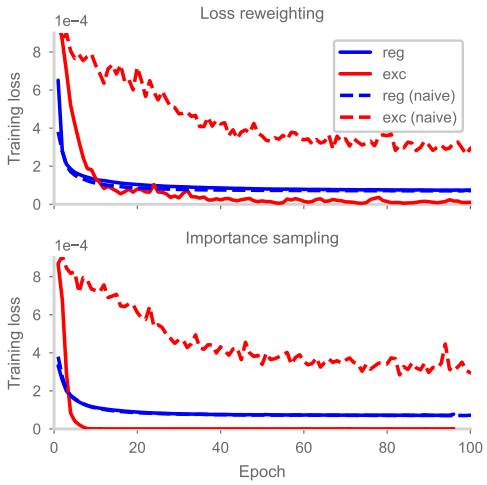

Figure 1: Comparison of ground-truth reweighted/resampled (solid) and unweighted/uniformly sampled (dashed) training. Both loss reweighting and importance sampling significantly improve memorization of exceptions without impairing performance on regular items.

Prior work has considered models that can memorize large numbers of patterns [18], those that learn consolidated representations by gradient descent [10], [19], or intermediate ones that can trade off information per pattern and number of patterns stored [20]. In contrast, we propose a training paradigm in which an associative memory model can exhibit behavior at both extremes of this continuum simultaneously – memorizing individual exceptional items while consolidating those from a commonly occurring class.

We train the network with stochastic gradient descent [21]. Although it may be possible to use a biologically plausible variant of backpropagation [22], [23] or biologically-inspired local plasticity rules [24], we compute gradients using the standard backpropagation through time (BPTT) algorithm. On the k^{th} iteration, we draw a batch $\mathcal{B}_k \subset \mathcal{D}$ from the dataset, generate a perturbed input $\widetilde{x}_b$ (see previous section), compute the corresponding output g_b^* for every item $x_b \in \mathcal{B}_k$, and update the parameters in proportion to the gradient of the loss $\nabla\mathcal{L}$ with respect to the network parameters. In a standard learning scenario, batches would be drawn uniformly without replacement so that every datapoint is seen exactly once per epoch, and the loss $\mathcal{L}$ would be the arithmetic mean of individual losses $\ell(x_b, g_b^*)$. In the case of imbalanced data, a common technique is to sample the batches *with* replacement according to a probability distribution inversely proportional to the relative frequency of the classes,

$$p(x_i \in \mathcal{B}_k) = \hat{\alpha}_i \tag{4}$$

or to compute a weighted loss for the batch, where the weights $\hat{\alpha}_b$ are taken inversely proportional to the relative frequency of the classes,

$$\mathcal{L} = \sum_{b=1}^{|\mathcal{B}_k|} \hat{\alpha}_b \, \ell(x_b, g_b^*) \tag{5}$$

In many situations for both biological and artificial agents, however, the class labels are not known a priori and relative class frequencies cannot be used for importance sampling or loss weighting. One of the contributions of our work is identifying signals that can be extracted directly from the network (rather than given class labels) and used to automatically reweight the losses or sample the training

data. We consider the following candidates for the unnormalized exception signals (see Section A.1):

$$s_i = \begin{cases} \mathbb{I}[x_i \in \mathcal{E}] & \text{ground truth (control condition/evaluation metric)} \\ \ell(x_i, g_i^*) & \text{loss per item} \\ -T & \text{(negative) time to converge to fixed point} \\ -H(f_i) & \text{(negative) entropy of hidden layer} \\ \max(f_i) & \text{max of hidden layer} \\ ||f_i|| & \text{norm of hidden layer} \end{cases} \tag{6}$$

These are re-computed for each training iteration, and vary over the course of training. To meaningfully compare the scores, we normalize them to the range $[0, 1]$ using the cumulative min/max of s_i seen so far during training:

$$\alpha_i = \frac{s_i - \min(s_{i'})}{\max(s_{i'}) - \min(s_{i'})} \tag{7}$$

To ensure they sum to 1, the final loss weights or sampling probabilities are then given by

$$\hat{\alpha}_i = \frac{\exp(\gamma \alpha_i)}{\sum_{i'=1}^{|\mathcal{D}|} \exp(\gamma \alpha_{i'})} \tag{8}$$

where the inverse-temperature parameter γ controls the relative magnitude of the weights. Note, $\gamma = 0$ reduces to an unweighted loss/uniform sampling. Furthermore, $s_i = \mathbb{I}[x_i \in \mathcal{E}]$ with $\gamma = \ln \frac{R}{E}$ corresponds to the standard methods with known labels, reweighting or sampling proportional to the relative frequency of the classes. This is equivalent to the evaluation loss in Eq. 2.

5 Results

As an example to demonstrate proof-of-principle, we use the MNIST dataset, where the "regular" subset consists of $R = 6000$ examples of zeroes, and the "exception" subset consists of $E = 3$ examples of ones. We use a network with 784 visible units (the dimensionality of MNIST images) and 100 hidden units. The hidden layer $\mathrm{softmax}(\cdot)$ inverse-temperature parameter is fixed to $\beta = 5$, determined empirically to give the overall best performance for the network of this size.

Training the network naively with uniform weighting/sampling leads to successfully learning the regular items, but failing to store the exceptions (Figure 1, dashed curves). Although the loss on exceptions decreases, this is incidental due to overlapping pixels in MNIST 1's and 0's: visually inspecting the stored representations (rows of weight matrix, Section A.2) ξ_μ in the trained network, all resemble 0's and none resemble a 1. On the other hand, a network trained with ground truth knowledge of whether an item is an exception – either upweighting or resampling exceptions – successfully learns both types of items, notably learning exceptions without degrading performance on regulars (Figure 1, solid curves).

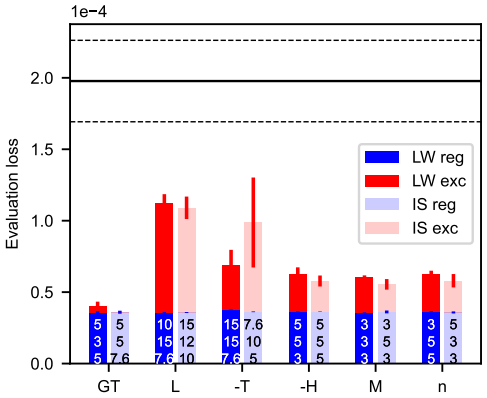

Figure 2: Evaluation loss after 150 epochs of training, split between loss on regulars (blue) and exceptions (red), trained with loss reweighting (LR, dark) or importance sampling (IS, light) using exception signals s_i (Eq. 6). Numbers at the bottom of each bar indicate the optimal γ (Eq. 8) for each of $n = 3$ runs, optimized by sweeping over $\gamma \in \{0, 1, 3, 5, 7.58, 10, 12, 15\}$ ($\gamma = \ln \frac{R}{E} \approx 7.58$ for this dataset). Black lines indicate baseline performance with $\gamma = 0$ (uniform weighting/sampling). Error bars and dashed lines indicate $\pm$ 1 standard deviation across $n = 3$ runs.

As our key result, to demonstrate the performance of reweighting/resampling using our proposed metrics, we plot the evaluation loss (Eq. 2) at the end of training (Figure 2) for each exception signal (Eq. 6). Although, as expected, using ground truth (GT) information about whether an item is regular or exceptional has the lowest evaluation loss (Eq. 2), all of the metrics proposed in Eq. 6 show significant improvement over the naive unweighted scenario (black line). A potential issue with these metrics is the variability in learning exceptions (red error

bars) compared to learning regular items (blue error bars), especially convergence time (-T). Finally, counter-intuitively, we note that even when training with ground-truth signal, the optimal γ is not equal to $\ln \frac{R}{E} \approx 7.58$ (the effective weighting in the evaluation loss).

6 Discussion

Beyond proposing a biological framework for methods to concurrently memorize individual examples of rare categories and learn consolidated representations of common ones, our model has potential applications to artificial intelligence. Large language models have been shown to output memorized sequences from their training corpus, raising significant privacy concerns [25]. Our network can be used as a simplified tractable model of this phenomenon to study this problem, particularly given its similarities to the attention mechanism of Transformers [19], enabling identification, updating, or removal of memorized items. From the perspective of fairness in AI [26], our work suggests a potential technique to help mitigate some of the common problems arising from underrepresented classes, automatically balancing training on biased data by downweighting overrepresented classes and enhancing rare ones.

7 Acknowledgments

We are grateful to John Cunningham, Ashok Litwin-Kumar, Stefano Fusi, and James Fitzgerald for helpful discussions and suggestions. Thanks to Jan Funke for suggesting importance sampling as a potential approach. Research was supported by the Gatsby Charitable Foundation GAT3708, Kavli Foundation, and NIH T32NS064929.

References

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, Apr. 1, 1982, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.79.8.2554. [Online]. Available: https://www.pnas.org/content/79/8/2554.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, Number: 7553 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature14539. [Online]. Available: https://www.nature.com/articles/nature14539.

[3] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos, *Beyond neural scaling laws: Beating power law scaling via data pruning*, Nov. 15, 2022. arXiv: 2206.14486[cs,stat]. [Online]. Available: http://arxiv.org/abs/2206.14486.

[4] V. Feldman and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 2881–2891. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1e14bfe2714193e7af5abc64ecbd6b46-Abstract.html.

[5] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," *Psychological Review*, vol. 102, pp. 419–457, 1995, Place: US Publisher: American Psychological Association, ISSN: 1939-1471. DOI: 10.1037/0033-295X.102.3.419.

[6] D. Kumaran, D. Hassabis, and J. L. McClelland, "What learning systems do intelligent agents need? complementary learning systems theory updated," *Trends in Cognitive Sciences*, vol. 20, no. 7, pp. 512–534, Jul. 1, 2016, ISSN: 1364-6613. DOI: 10.1016/j.tics.2016.05.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661316300432.

[7] L. Wittkuhn, S. Chien, S. Hall-McMaster, and N. W. Schuck, "Replay in minds and machines," *Neuroscience & Biobehavioral Reviews*, vol. 129, pp. 367–388, Oct. 1, 2021, ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2021.08.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0149763421003444.

[8] G. Parisi, "A memory which forgets," *Journal of Physics A: Mathematical and General*, vol. 19, no. 10, pp. L617–L620, Jul. 1986, ISSN: 0305-4470. DOI: 10.1088/0305-4470/19/10/011. [Online]. Available: https://doi.org/10.1088%2F0305-4470%2F19%2F10%2F011.

[9] D. Tyulmankov, C. Fang, A. Vadaparty, and G. R. Yang, "Biological learning in key-value memory networks," in *Advances in Neural Information Processing Systems*, vol. 34, Curran Associates, Inc., 2021, pp. 22 247–22 258. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/bacadc62d6e67d7897cef027fa2d416c-Abstract.html.

[10] D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," *arXiv:1606.01164 [cond-mat, q-bio, stat]*, Sep. 27, 2016. arXiv: 1606.01164. [Online]. Available: http://arxiv.org/abs/1606.01164.

[11] D. Krotov and J. Hopfield, "Large associative memory problem in neurobiology and machine learning," *arXiv:2008.06996 [cond-mat, q-bio, stat]*, Mar. 2, 2021. arXiv: 2008.06996. [Online]. Available: http://arxiv.org/abs/2008.06996.

[12] D. Krotov, "A new frontier for hopfield networks," *Nature Reviews Physics*, pp. 1–2, 2023.

[13] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," en, *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, Jan. 2012, Number: 1 Publisher: Nature Publishing Group, ISSN: 1471-0048. DOI: 10.1038/nrn3136. [Online]. Available: https://www.nature.com/articles/nrn3136.

[14] Z.-H. Mao and S. G. Massaquoi, "Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition," *IEEE Transactions on Neural Networks*, vol. 18, no. 1, pp. 55–69, Jan. 2007, Conference Name: IEEE Transactions on Neural Networks, ISSN: 1941-0093. DOI: 10.1109/TNN.2006.883724.

[15] M. A. Snow and J. Orchard, "Biological softmax: Demonstrated in modern hopfield networks," *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, no. 44, 2022. [Online]. Available: https://escholarship.org/uc/item/3jd1t2hr.

[16] D. Krotov, "Hierarchical associative memory," *arXiv:2107.06446 [cs]*, Jul. 13, 2021. arXiv: 2107.06446. [Online]. Available: http://arxiv.org/abs/2107.06446.

[17] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, Mar. 19, 2019, ISSN: 2196-1115. DOI: 10.1186/s40537-019-0192-5. [Online]. Available: https://doi.org/10.1186/s40537-019-0192-5.

[18] M. Demircigil, J. Heusel, M. LÃ¶we, S. Upgang, and F. Vermet, "On a model of associative memory with huge storage capacity," *Journal of Statistical Physics*, vol. 168, no. 2, pp. 288–299, Jul. 1, 2017, ISSN: 1572-9613. DOI: 10.1007/s10955-017-1806-y. [Online]. Available: https://doi.org/10.1007/s10955-017-1806-y.

[19] H. Ramsauer, B. Schafl, J. Lehner, *et al.*, "Hopfield networks is all you need," *arXiv:2008.02217 [cs, stat]*, Jul. 16, 2020. arXiv: 2008.02217. [Online]. Available: http://arxiv.org/abs/2008.02217.

[20] S. Sharma, S. Chandra, and I. Fiete, "Content addressable memory without catastrophic forgetting by heteroassociation with a fixed scaffold," in *Proceedings of the 39th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jun. 28, 2022, pp. 19 658–19 682. [Online]. Available: https://proceedings.mlr.press/v162/sharma22b.html.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980 [cs]*, Jan. 29, 2017. arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980.

[22] T. P. Lillicrap and A. Santoro, "Backpropagation through time and the brain," *Current Opinion in Neurobiology*, Machine Learning, Big Data, and Neuroscience, vol. 55, pp. 82–89, Apr. 1, 2019, ISSN: 0959-4388. DOI: 10.1016/j.conb.2019.01.011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438818302009.

[23] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 335–346, Jun. 2020, Number: 6 Publisher: Nature Publishing Group, ISSN: 1471-0048. DOI: 10.1038/s41583-020-0277-3. [Online]. Available: https://www.nature.com/articles/s41583-020-0277-3.

[24] D. Krotov and J. J. Hopfield, "Unsupervised learning by competing hidden units," *Proceedings of the National Academy of Sciences*, vol. 116, no. 16, pp. 7723–7731, 2019.

[25] S. Biderman, U. S. Prashanth, L. Sutawika, *et al.*, *Emergent and predictable memorization in large language models*, Apr. 21, 2023. arXiv: 2304.11158[cs]. [Online]. Available: http://arxiv.org/abs/2304.11158.

[26] S. Bird, M. Dudík, R. Edgar, *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in AI," May 18, 2020. [Online]. Available: https://www.scinapse.io/papers/3030081171.

A Appendix

A.1 Candidate signal selection

To establish candidate metrics to use as exception signals, we consider various quantities as the network is trained using ground-truth loss weighting. Several metrics can differentiate the regular items from exceptions when the network is trained with ground truth weighting (Figure A1, top), suggesting that it may be possible to use them to bootstrap the exception signal without prior knowledge of which items are exceptions. Note that some of the signals (entropy, convergence time) have the opposite sign from the ground truth signal, so we use their negation as the signal s_i. Indeed, if we train the network using (negative) entropy as the exception signal, the regulars and exceptions are again differentiated, as shown in Figure A1 (bottom).

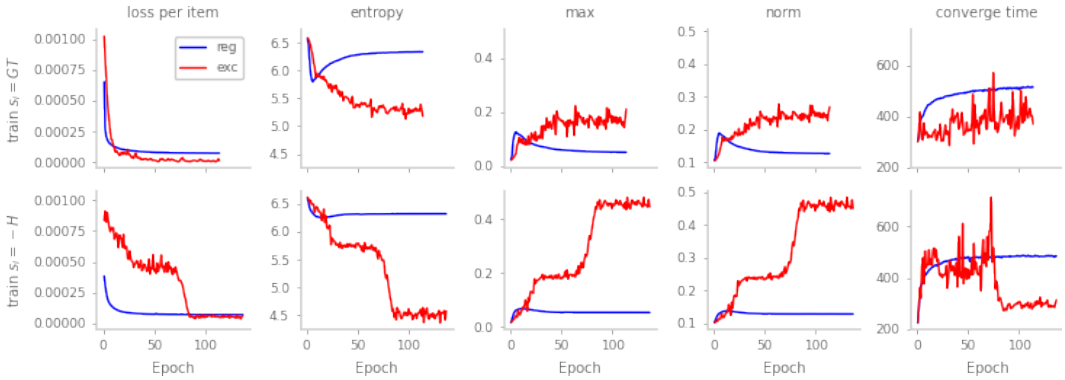

Figure A1: Signals extracted from the network can differentiate regular items from exceptions, and used to bootstrap learning for concurrent memorization and consolidation. Top: candidate signals for automatic loss weighting or importance sampling, measured as the network is trained using ground truth knowledge of whether each item is an exception. Bottom: the same signals, measured as the network is trained using negative entropy $s_i = -H(f)$ as the exception signal.

A.2 Example trained weights

Figure A2: Final trained weights $\{\xi_\mu\}, (\mu = 1, \ldots, 100)$, corresponding to Figure 1, reshaped to match the dimensions of the MNIST digits used for training (28×28).