UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Retrieval of Hierarchically-Organized Concepts in a Recurrent Memory System

Permalink
https://escholarship.org/uc/item/9d6123h4

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Ralston, Robert
Sloutsky, Vladimir

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Retrieval of Hierarchically-Organized Concepts in a Recurrent Memory System

Robby Ralston (ralston.123@osu.edu)
Vladimir Sloutsky (sloutsky.1@osu.edu)
Department of Psychology, Ohio State University
1835 Neil Ave, Columbus, OH 43210 USA

Abstract

Exemplar models have been criticized for lacking mechanisms to explain key conceptual phenomena such as the hierarchical organization of concepts. Here, we offer a potential solution. We show that a broad class of exemplar models can be viewed as a special case of global matching models of memory, and that global matching models are themselves discrete-time approximations of Dense Associative Memories (DAMs), a type of recurrent network. Interpreted this way, exemplar models retrieve hierarchical prototypes by modulating competition during retrieval. We demonstrate this ability using artificial data and pretrained GLoVe and Word2Vec embeddings. Our results suggest that exemplar models remain plausible candidates for a broader theory of concepts and provide a natural algorithmic account of attractor-like retrieval in the hippocampus, highlighting their relevance in learning theory and cognitive neuroscience. **Keywords:** memory; conceptual organization; exemplar models; global matching models; attractor networks

Concepts are ubiquitous in mental life, enabling reasoners to classify items into meaningful categories, extract important commonalities, and apply this knowledge in new situations (Murphy, 2004). Researchers have proposed several ways that concepts may be represented in the brain and many mechanisms to explain their formation and use (Ashby et al., 2011; Gopnik & Wellman, 2012; Minda & Smith, 2001).

One such proposal is exemplar theory, according to which concepts are represented as memory traces of many individual items, each of which contributes to our understanding of a category (Logan, 2002; Nosofsky, 1986; Turner, 2019). For example, to classify an organism as a dog, one compares the item to all memory traces formed from encounters with dogs and other organisms throughout the lifespan, and a 'dog' response occurs because memories of dogs are most strongly activated. Exemplar models have been used extensively to explain data from experiments involving human and nonhuman animals and a substantial body of work examining the mnemonic, attentional, and decision processes underlying categorization (Murphy, 2004; Nosofsky, 2011).

Although exemplar models can explain experimental results, they have been criticized for lacking a broader theory of conceptual organization (Murphy, 2016). Here, we focus on hierarchical concepts, an important aspect of conceptual knowledge. Critics argue that an exemplar-based account may find it difficult to explain how a dog could be selectively represented as a cocker spaniel, a dog, or a mammal when useful to do so. This is a problem for a theory of concepts because, among other reasons, it may fail to explain inferences in some domains (Hayes & Heit, 2018). For example, one does not need to individually memorize that dogs, cats, bears, and other mammals possess mammary glands, but can infer this based on their common superordinate category (mammal).

In this paper, we address this critique, showing that hierarchical organization of concepts can emerge naturally from episodic retrieval in biologically plausible memory systems. Specifically, we show that exemplar models are one-update approximations of recurrent network architectures known as a Dense Associative Memory (DAM) networks (Demircigil et al., 2017; Krotov & Hopfield, 2020; Krotov & Hopfield, 2016). We then demonstrate that the relevant DAMs exhibit *hierarchical retrieval*: when memory traces form hierarchical clusters, these systems selectively retrieve prototypes of an item's category at each level of the hierarchy. In other words, exemplar models can use a cue, such as the word 'cereal,' to retrieve the prototype of cereal, breakfast foods, and foods in general, and can toggle between these representations at the time of retrieval using known mechanisms.

While hierarchical retrieval is readily achieved by humans, few cognitive models have been shown to possess this property (but see Palmeri, 1999; Saxe et al., 2019; Whittington et al., 2020). Below, we provide simulations with artificial data and precomputed word embeddings to demonstrate this ability in exemplar models, examine how knowledge is affected by item noise, and show that models infer the existence of hierarchical structure (though imperfectly) in realistic data under assumptions about attention allocation. Importantly, our models were not trained to recover the hierarchical structure. Rather, structure emerged during memory retrieval from general attentional biases. Word embeddings were chosen only as a demonstration, and our findings generalize to any data where items possess sufficient clustering[1].

We first observe that exemplar models are members of a broader class of global matching models, a popular approach to episodic memory in computational psychology. Global matching models are one-step approximations of dynamical DAM networks, with many being equivalent to a special case known as Modern Hopfield Networks (MHN) with interest-

[1]Whether actual representations in a given domain possess the required structure remains an open question.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

ing retrieval properties. We simulate hierarchical retrieval using MHNs on artificial data and word embeddings. These findings show how to derive a recurrent, continuous-time, biologically-plausible network which is well-approximated by an exemplar model, and provide a method by which exemplar models can explain hierarchical retrieval.

Background and Theory

We begin by giving an overview of our formal approach, which we call AuToassociative and HEtero Neural Attention (ATHENA). We refer to a forthcoming paper for full details (Ralston et al., 2025). We can represent the inferential process of exemplar models such as the Generalized Context Model (GCM) in a condensed vector-matrix form:

$$A(u) = f\left(k\left(u \mid K\right)\right)V, \tag{1}$$

where $A(u)$ gives the evidence in favor of each category given item u, K is a memory matrix with traces as rows, $k\left(u \mid K\right)$ applies a similarity kernel between u and each row of K, f is an optional function allowing memories to compete for retrieval, and V contains trace-category associations. For example, in GCM, $f(x) = x$ is the identity, and evidence in favor of each category $A(u)$ is a linear function of the similarity between u and each memory trace. Exemplar models use the resulting evidence values to drive a decision process; for example, by using Luce's choice rule (Nosofsky, 1986; Nosofsky, 2011). However, here we primarily focus on pre-response retrieval dynamics.

Exemplar Models and Global Matching Processes

Global matching models represent memory as a similarity-based comparison process, where cues are compared to memory traces, and item-trace similarity is the basis for recognition and/or recall (see Osth and Dennis, 2020 for a review). A key insight from the global matching approach is that similarity-based comparisons such as Equation (1) do not need to be implemented separately for category labels and other item features. Instead, a common pattern completion architecture can fill-in missing information based on any features currently available (e.g., Sloutsky and Fisher, 2004).

A typical pattern completion mechanism can be implemented by replacing K and V above with the augmented memory matrix

$$M^* = \begin{bmatrix} K & V \end{bmatrix}.$$

Any number of present/missing item features and categories can then be inferred with:

$$A(u) = f\left(k\left(u \mid M^*\right)\right)M^*, \tag{2}$$

where k ignores missing features and retrieves only relevant dimensions for the task. Many memory models are consistent with Equation (2), including exemplar models expressible with Equation (1) (Osth & Dennis, 2020). We refer to these collectively as *cognitive memory models*

Network Implementations of Memory Models

Many cognitive memory models can be expressed as the single-update limit of a two-layer recurrent architecture. To see this, we use spectral decomposition[2] (Schölkopf & Smola, 2002) to represent $k\left(u \mid M^*\right)$ as a linear operation,

$$k\left(u \mid M^*\right) = \phi(u)\begin{bmatrix}\phi(m_1^*)^\top & \dots & \phi(m_n^*)^\top\end{bmatrix} = xM^\top$$

where ϕ is known as a feature map, $m_1^*, \dots, m_n^*$ are memory traces, $x = \phi(u)$, and M is the matrix with feature-mapped memory traces as rows (rather than raw item representations). This gives a network representation of the same inferential process:

$$A(u) = f\left(xM^\top\right)M^*. \tag{3}$$

On this interpretation, cues are presented to the visible 'feature layer' x, and then progress to the hidden 'memory layer,' resulting in $xM^\top = k(u \mid M^*)$. After applying the competition function f, representations are returned to the feature layer, where target information is decoded, e.g., as a category label.

Since the retrieved representation is returned to the feature layer, this naturally raises the possibility that Equation (3) could be iterated. By remaining in feature space, i.e., substituting M for M^* in (3), this produces a time-dependent, recurrent network with the update equation:

$$x_{t+1} = g\left[f\left(x_t M^\top\right)M\right], \tag{4}$$

where g is an additional (potentially) nonlinear activation function or renormalization. Making the network recurrent in this way has surprising benefits for the recall process, allowing the system to reconstruct approximations of items which have been seen before, with MINERVA 2's 'echo' process as a paradigmatic example (Hintzman, 1986).

Cognitive Attractor Networks

The network in Equation (4) is known as a Dense Associative Memory network (DAM) (Demircigil et al., 2017; Krotov and Hopfield, 2016; see Krotov and Hopfield, 2020 for a full explanation). DAMs are attractor networks, generalizing classical autoassociative models such as Hopfield networks which are used to model pattern completion in hippocampal subfield CA3 (Hopfield, 1982; Rolls, 2010). Since cognitive memory models can be implemented with Equation (4), the networks discussed in the previous section are also special cases of this architecture.

In Equation (4), different choices of g and f give rise to networks with different properties (see Krotov and Hopfield, 2020 for a discussion). When g is the identity and f is the

[2]Traditional kernel methods only apply to symmetric and positive semidefinite similarity functions. However, for our purposes, the method can be extended to apply to functions which are not positive definite (e.g., Shiffrin and Steyvers, 1997) if retrieval cues and memory traces are encoded differently. With different encodings, the improper kernel k^* is given by $k^*\left(u \mid M\right) = \phi(u)\begin{bmatrix}\psi(m_1^*)^\top & \dots & \psi(m_n^*)^\top\end{bmatrix}$.

softmax function, this results in a Modern Hopfield Network (MHN; Ramsauer et al., 2020) with the update:

$$x_{t+1} = \text{softmax}\left(\beta x_t M^\top\right) M. \tag{5}$$

Viewed through a cognitive lens, MHNs make inferences via a typical exemplar/global matching model ($k(u \mid M^*) = xM^\top$), but where trace activation is subjected to a softmax competition process, and the amount of competition is indexed by the nonnegative β parameter. Single-step inference is identical to many exemplar models, where softmax behavior emerges from the combination of exponential similarity and response processes that divide the similarity to a category by the overall similarity (e.g., Nosofsky, 1986).

In exemplar models, β is analogous to the sensitivity parameter (often written as c) which gives the width of the similarity kernel-e.g., resulting in nearest-neighbor classification when this parameter is large (Nosofsky, 2011). The arguments of Ramsauer et al. (2020) provide insight about this parameter, showing that c (AKA β) is critical to the multi-update behavior of their network implementations. In the following section, we discuss this finidng and how it allows exemplar-based attractor networks, i.e., MHNs, to represent category prototypes and, ultimately, achieve hierarchical retrieval.

Prototypes, Competition, and Overall Attention

Modern Hopfield Networks are known to possess many interesting properties for a memory retrieval system, reminiscent of the schema-abstraction results from MINERVA 2 (Hintzman, 1986). Most importantly, Ramsauer et al. (2020) showed that the amount of competition (β) plays a crucial role in which memories are retrievable (i.e., the location of stable points in Equation (5)). Strict competition (large β) leads to tightly-tuned retrieval of individual traces, corresponding to nearest neighbor classification in exemplar models. However, when data form meaningful clusters, relaxing competition leads to widespread memory activation and stable points that average across many exemplars. This quantity therefore affects whether the system retrieves a specific trace or an average of several traces, i.e., a prototype.

An interesting possibility arises if the level of competition can be modulated during retrieval: variable competition would allow a memory system the flexibility to retrieve either individuals or prototypes. In exemplar models, this would correspond to varying the sensitivity parameter (c), dynamically affecting the spatial scale of memory retrieval. If data possesses hierarchical clusters-i.e., clusters that occur across spatial scales, with smaller clusters making up larger clusters-this raises the possibility that exemplar models and MHNs could extract prototypes at different levels of the conceptual hierarchy on demand. But is it plausible that competition could be modulated in this way?

Interpreted as a biological network, several neural mechanisms could vary the amount of competition in a memory system. For example, feedforward inhibition provides a plausible mechanism that could allow external control to influence the overall amount of competition during retrieval (McKenzie, 2018). Alternatively, selective attention could modulate the effective level of competition by acting on the feature layer representations of items (Galdo et al., 2022; Nosofsky, 1986; Turner, 2019). To see this, consider (5) and let α represent a vector of non-negative attention weights. We can implement a typical multiplicative attention process:

$$x_{t+1} = \text{softmax}\left(\beta(\alpha \odot x_t)M^\top\right) M \tag{6}$$

$$= \text{softmax}\left(\beta\|\alpha\|(\hat{\alpha} \odot x_t)M^\top\right) M, \tag{7}$$

where $\odot$ is the element-wise (Hadamard) product and $\hat{\alpha}$ is the unit vector in the direction of α. Note that the magnitude of the attention vector $\|\alpha\|$ has the same role as β; if the overall amount of attention can be influenced during retrieval, this would have the same effect as varying the amount of competition on demand. Models that vary the amount of attention have been discussed previously (Galdo et al., 2022; Kruschke, 2001), though not directly on the feature-space representation of an item, and the empirical effects of varying attention in this way is not yet understood.

We have shown that cognitive memory models can be implemented as dynamical attractor networks, and that exemplar models with exponential similarity and a relative response rule are closely related to Modern Hopfield Networks. In MHNs, varying the amount of competition changes the memories retrieved by the system, potentially toggling between individuals and prototypes. Below, we assess whether varying the amount of competition can achieve recall across the conceptual hierarchy through a simulation study.

Computational Methods

Our goal was to determine if the memory system described by Equation (6) exhibits hierarchical retrieval. To do this, we considered the stable points of Modern Hopfield Networks, i.e., exemplar models implemented as attractor networks, and examine how these points change when competition β is altered. We did not attempt to model the learning process, instead storing noisy examples of each item. This allowed us to assess the effects of varying competition on pre-established memory traces during retrieval and avoids needing to model a complex and modality-dependent learning process.

Datasets

Artificial Hierarchical Clusters. To assess model behavior in ideal conditions, we generated items with hierarchical clusters according to a pre-defined similarity matrix. There were three higher-level clusters, each with two lower-level clusters of ten items each (60 true items total). We first created a pre-similarity matrix where items sharing only a higher-level category were assigned a similarity of 15, items sharing a lower-level category were assigned a similarity of 20, and items were assigned a self-similarity of 25. We then

Table 1: Category terms used to generate word embeddings.

High-Level	Lower-Level	Items
Emotions	Positive	Joy, Love, Pride, Happiness, Gratitude, Excitement, Contentment, Hope, Serenity, Compassion
	Negative	Anger, Fear, Sadness, Guilt, Shame, Jealousy, Frustration, Loneliness, Anxiety, Disgust
Countries	Europe	France, Germany, Italy, Spain, Netherlands, Sweden, Norway, Greece, Portugal, Switzerland
	Asia	China, India, Japan, Thailand, Vietnam, Indonesia, Malaysia, Philippines, Singapore, Pakistan
Foods	Breakfast	Pancakes, Oatmeal, Cereal, Toast, Bagel, Yogurt, Muffin, Granola, Waffles, Porridge
	Dinner	Steak, Pasta, Tacos, Curry, Pizza, Sushi, Burger, Salad, Lasagna, Chili

took its eigendecomposition and obtained item representations by multiplying eigenvectors by the diagonal matrix containing the square root of eigenvalues. This resulted in 60-dimensional vectors for each item where the dot product between items was equal to their pre-similarity. We then normalized the vectors to unit length. These vectors served as seeds for generating noisy memory traces and were used to cue the network when assessing stable points.

To obtain noisy memory traces, we stored each true item ten times with noise. Since cosine similarity measures the angle between two vectors, we obtained noisy traces by adding angular noise to each true item. Specifically, we sampled a von Mises-Fisher distribution centered at the true item with several values for the precision parameter κ.

Noisy GLoVe Embeddings To examine the ability of the system to extract prototypes across hierarchical levels under more realistic item noise, we utilized pre-computed word embeddings. We first chose a list of three ontologically-distinct superordinate category terms (Emotions, Countries, Foods) which possess the hierarchical structure shown in Table 1. These categories were chosen to exhibit the performance of the model in a typical case, and investigation of additional categories yielded similar results. We leave a systematic investigation of other categories to future work.

After selecting terms, we obtained pretrained GLoVe embeddings of the items from a model trained on 6 billion tokens from a 2014 Wikipedia dump and the Gigaword 5 database (Pennington et al., 2014). This resulted in 300-dimensional vectors for each term. After unit normalization, we treated vectors identically to the true items above, and used them as retrieval cues when assessing the model's stable points. To obtain memory traces, we obtained ten samples from a von Mises-Fisher distribution centered on each word vector with $\kappa = 1000$. Cosine similarity matrices showing the amount of noise relative to signal can be seen in Figure 2.

Noisy Word2Vec Embeddings To avoid conclusions based on the idiosyncrasies of GLoVe embeddings and the selected datasets, we also obtained 300-dimensional, pretrained Word2Vec embeddings for each term using using the Python package gensim (Mikolov et al., 2013; Řehůřek & Sojka, 2010). These embeddings were obtained using the Google News corpus containing 3 billion tokens. Otherwise, embeddings were treated identically to GLoVe embeddings.

Model Initialization

To initialize the model, we first obtained attention weights (α) for the target dataset. For artificial clusters, we set attention weights equal to 1 because similarity was pre-defined. For distributional embeddings, raw vectors contained very little hierarchical structure and substantial cross-category similarity. This is not surprising, as, even under ideal conditions, word vectors represent many aspects of word usage. However, by attending to different subsets of dimensions, different aspects of structure can emerge from the same vectors. We chose to use attention weights that maximized similarity within lower-level categories and minimized similarity between items that did not share a lower-level category, consistent with general attentional biases that distinguish between categories, a important explanatory principle in exemplar models (Nosofsky, 1986). Notably, this attentional set does not, by itself, help the model to extract hierarchical structure, only to distinguish lower-level categories. The resulting cosine similarity matrices can be seen in Figure 2 below.

To obtain weights, we initialized the attention vector to all 1s. We used mean squared error loss between a) the cosine similarity matrix of attention-weighted item vectors and b) the matrix where within-lower-level categories have similarity equal to 1 and other pairs have similarity -1. We fit the attention vector over 1000 epochs using the Adam optimizer implemented in PyTorch (Paszke et al., 2019). After obtaining weights, we renormalized the attention-weighted memory vectors to unit length, to ensure that the competition parameter (β) was comparable across simulations.

Simulated Retrieval

To obtain retrievable memories, we estimated the stable points obtained after initialization with each retrieval cue across 50 values of β. For each cue, we iterated (4) either one thousand times or until an update moved the vector less than a distance of 10^{-7}. This procedure should reflect the stable points of (4) given the convergence properties discussed by Ramsauer et al. (2020). We considered two retrieval cues to share the same stable point if the resulting vectors were within a distance of 10^{-6}. This resulted in a sequence of stable points for each retrieval cue representing the memory retrieved by the model under different levels of competition.

Evaluation

We considered hierarchical retrieval to be achieved when:

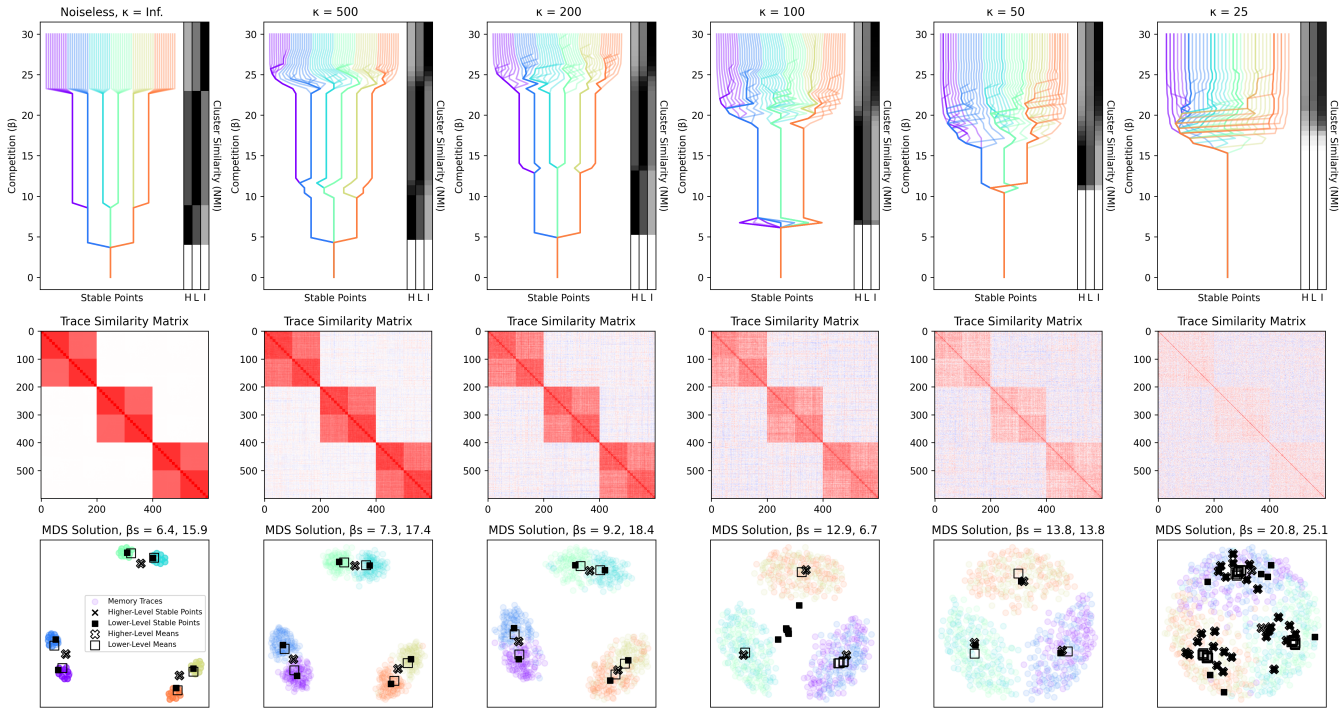

Figure 1: Results with artificial clusters. Columns vary in angular noise during storage (lower κ = more noise). Top row: stable points as a function of competition (β). Each colored line tracks a retrieval cue; several lines collapsing to one indicate a shared stable point. Colors reflect true lower-level categories; horizontal position is arbitrary. Shaded bars on the right show NMI scores for higher-level (H), lower-level (L), and individual item (I) clustering, with darker shades indicating higher NMI (better match to ground truth). Middle row: cosine similarity between stored traces (red = high similarity, blue = low). Bottom row: MDS solutions visualizing stored items, stable points, and category means at selected β values minimizing NMI for higher- and lower-level categories (Xs = higher-level means, squares = lower-level means).

1. There exists a value of β where items in the same category share a stable point which excludes other items. This should occur at a different β for each hierarchical level.

2. The stable point representing a group of items is approximately its prototype, i.e., near the mean of those items.

To assess (1), we first obtained visual representations of each stable point. Since retrieval cues often share stable points, the model's representation at each β could be represented as a cluster solution, with items sharing a stable point in the same cluster. To assess clustering accuracy we calculated the Normalized Mutual Information (NMI) for each β between model representations and the ground-truth higher- (H) or lower- (L) level category structure, where all members of a category are in the same set. NMI is a useful similarity measure between two clustering solutions, equal to 1 when they are identical and 0 when cluster assignments are independent. For both category levels, we found the level of competition where NMI was maximized. Maximum NMI gives an assessment the model's ability to recover each category structure at some level of competition. We also obtained this value for an 'individual' clustering (I), which assigned each retrieval cue to its own cluster. To achieve (1), the model should have NMI near 1 for both H and L categories.

For criterion (2), we obtained the prototype (directional mean) of each category. We considered item representations during runs where β maximize NMI from one of the category structures (see above). We then computed a Multidimensional Scaling (MDS) Solution using the scikit-learn package in Python to visualize how the model's stable points correspond to the prototype and distribution of memory traces (Pedregosa et al., 2011). To obtain dissimilaritiesfor MDS, we used one minus the cosine similarity.

Results

Artificial Hierarchical Clusters

Figure 1 shows our results with artificial clusters under different amounts of noise during storage. With low competition, the model considers only global structure, with every item collapsing to the same stable point. Then, as competition increases, models with $\kappa \geq 50$ possess three stable points corresponding to the mean of higher-level categories ($\max \text{NMI} = 1$ for each model). This can be seen in the existence of regions in the top row of Figure 1 with three stable points as well as the close overlap of item means and stable points in the MDS solutions. The remaining simulation fails

to capture the higher-level categories (max NMI $= 0.53$).

As β increases, three simulations ($\kappa = \infty, 500, 200$) retrieve representations of the lower-level categories (max NMI $= 1$), corresponding visually to the same simulations where the 2D MDS solution still possesses separable clusters. In each, stable points closely correspond to cluster means. The remaining simulations fail to capture the lower-level categories, max NMI $= 0.79, 0.24, 0.31$ for $\kappa = 100, 50, 25$ respectively.

Word Embeddings

Results with word embeddings can be seen in Figure 2. Qualitatively, results are similar to those with artificial data. At low β, models possess a single stable point. Then, as β increases, stable points develop near the means of higher-level categories (emotions, countries, foods). For these values of β, each model achieved perfect classification (max NMI $= 1$) on the higher-level category. This is especially noteworthy because attention weights were selected specifically to maximize lower-level category distinctiveness, yet the model was able to recover the hierarchical structure.

In contrast, no model achieved perfect classification of lower-level categories. The model with stored GLoVe embeddings recovered separate stable points for Countries (European vs. Asian) and Foods (Breakfast vs. Dinner), but not for emotions, (max NMI = 0.93). In contrast, the Word2Vec model achieved different stable points for Emotions (Positive vs. Negative) and Foods, but not for countries, (max NMI = 0.85). While imperfect, these scores are near ceiling and reflect cluster assignments far above chance. Trace similarity matrices and the MDS solutions show that clusters that are not recovered are those with the most confusable lower-level categories. In each case, as β increases, individual terms 'break off' from the main stable point one by one instead of forming a stable sub-cluster. However, when there is a stable point corresponding to a lower-level category, it is typically near the category's prototype.

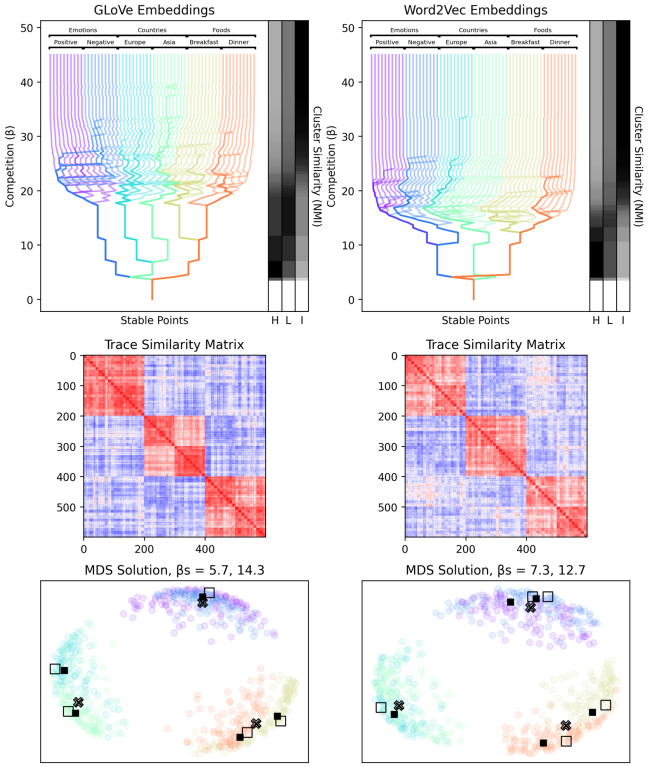

Figure 2: Results with distributional word embeddings. Plot conventions are the same as in Figure 1. In the top plot, the order of categories from left to right is: positive and negative emotions, European and Asian countries, and finally breakfast and dinner foods.

Discussion

In this study, we showed how exemplar models of categorization and global matching models of memory can be implemented as the update of an attractor network with the ability to extract hierarchical structure from stored traces. Specifically, we examined retrieval in Modern Hopfield Networks, an attractor architecture closely related to exemplar models with exponential similarity (Nosofsky, 1986; Nosofsky, 2011). With ideal data, our results show that exemplar models can flexibly access prototypes along a conceptual hierarchy. This shows that competitive, recurrent retrieval is a plausible mechanism for representing and accessing hierarchies in these approaches.

Additionally, we found the models to be fairly robust to light- and moderate noise during storage under an attentional set that maximizes differences in lower-level categories. However, even as noise increased further, models lost the ability to retrieve small-scale features of the data, but retained higher-level distinctions. This finding was observed with word embeddings from the GLoVe and Word2Vec datasets, where both models could retrieve higher-level prototypes, but not lower-level prototypes when clusters were undifferentiated. Therefore, while noise and overlapping traces can lead to insurmountable interference, large-scale structure persists. This parallels the acquisition of everyday categories, which are learned through experience but where one is unable to remember the specific experiences where this knowledge was acquired.

Finally, we have shown that attractor networks arise naturally as implementations of existing cognitive models and offer additional explanatory power. Attractor networks are suspected to enable pattern completion in the hippocampal region, enabling episodic retrieval (Rolls, 2010). Therefore, our results provide a theoretical bridge between recurrent models of the hippocampal region (Schapiro et al., 2018; Whittington et al., 2020) and exemplar models of categorization (Nosofsky, 1986; Turner, 2019; see also Kumaran and McClelland, 2012 for similar observations). In the future, we hope to use the approach developed here to better understand connections between network models of the hippocampus and cognitive memory models, and to translate findings across disciplines.

References

Ashby, G. F., Paul, E. J., & Maddox, T. W. (2011). Covis. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization*. Cambridge University Press.

Demircigil, M., Heusel, J., Löwe, M., Upgang, S., & Vermet, F. (2017). On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, *168*, 288–299.

Galdo, M., Weichart, E. R., Sloutsky, V. M., & Turner, B. M. (2022). The quest for simplicity in human learning: Identifying the constraints on attention. *Cognitive Psychology*, *138*, 101508.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085.

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(3), e1459.

Hintzman, D. L. (1986). " schema abstraction" in a multiple-trace memory model. *Psychological review*, *93*(4), 411.

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

Krotov, D., & Hopfield, J. (2020). Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996*.

Krotov, D., & Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *Advances in neural information processing systems*, *29*.

Kruschke, J. (2001). The inverse base-rate effect is not explained by eliminative inference. *Journal of experimental psychology: Learning, Memory, and Cognition*, *27*(6), 1385.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological review*, *119*(3), 573.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological review*, *109*(2), 376.

McKenzie, S. (2018). Inhibition shapes the organization of hippocampal representations. *Hippocampus*, *28*(9), 659–671.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Minda, J. P., & Smith, D. J. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 775–799.

Murphy, G. (2016). Is there an exemplar theory of concepts? *Psychonomic bulletin & review*, *23*, 1035–1042.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal approaches in categorization*, 18–39.

Osth, A. F., & Dennis, S. (2020). Global matching models of recognition memory. *PsyArXiv*.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, *6*(3), 495–503.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Ralston, R., Sloutsky, V. M., & M., T. B. (2025). Autoassociative and hetero neural attention (athena) [Manuscript in preparation].

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*.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rolls, E. T. (2010). Attractor networks. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 119–134.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature communications*, *9*(1), 3920.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem—retrieving effectively from memory. *Psychonomic bulletin & review*, *4*, 145–166.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166.

Turner, B. M. (2019). Toward a common representational framework for adaptation. *Psychological Review*, *126*, 660–692.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.