Learning Dynamics of Hierarchical Category Structure in Deep Non-Linear Networks

Anonymous

Abstract

Understanding how structured, hierarchical representations are learnt by biological or artificial neural networks is a fundamental goal in cognitive science, neuroscience, and artificial intelligence. Prior work has established that hierarchical representations in artificial networks are acquired in a stage-like fashion with general categories being learnt first, in agreement with human developmental studies. However, the scope of these studies has been limited by small datasets with hand-designed features or strictly linear network architectures. In this work, we take the first step to address both of these shortcomings by using a graph-based Gaussian generative process to create arbitrarily large datasets with control over the degree of hierarchical association between categories, and investigate the learning trajectory of non-linear deep neural networks in relation to linear ones. Our aim is to understand learning dynamics in the non-linear setting given that the human brain plausibly performs non-linear computations. We utilize a suite of tools to do so: i) Traditional metrics for investigating hierarchies; test accuracy and pairwise distances between representations, ii) Relative movement in the feature space constructed by network gradients, iii) Geometry of category centers, and iv) Unsupervised K-nearest neighbor methods. We replicate stage-like learning in non-linear networks by demonstrating how network gradients drive representations of categories across hierarchies in distinct directions. We show that hierarchies are differentially expressed across network layers as evidenced by geometrical properties of the categories over the course of learning. In early layers, learning of hierarchical structures involves starts by separation of general hierarchies and in exchange for compression for specific ones, followed by a second phase where separation of specific categories is achieved at the expense of compression of general categories. Finally, we recover stage-like transitions in a fully unsupervised manner without any knowledge about categories, by only using neighborhood statistics between categories. To the best of our knowledge, this is the first systematic analysis of how structures in large datasets, specifically hierarchies, are being learnt by non-linear deep neural networks. Our work has implications within cognitive science as we strengthen existing hypotheses about how hierarchical structures develop, as well as within artificial intelligence as our proposed tools provide increased interpretability of neural network representations.

1 Introduction

An overarching goal of cognitive science and neuroscience is to understand how humans learn and encode abstract, structured knowledge. Acquisition of hierarchical semantic knowledge is a prominent example of such learning and has been shown to follow a general to specific progression; In infants and children, broad categorical distinctions are acquired before finer ones (Keil, 1979; Mandler and McDonough, 1993; Quinn and Johnson, 1997; Gelman and Coley, 1990). To provide an analogy: Imagine an agent – a child or an artificial system – learns to perfectly distinguish among different categories of dogs and flowers. Next, you ask your agent whether an instance within a novel category,

Preprint. Under review.

a cat, is a plant or an animal. Has the agent learnt that a cat falls under the super-category animals, and not plants? By simply examining the (category) accuracy of the agent, it is unknown whether the true statistical structure of the data has been acquired. Early work on this question focused on the domain of human semantic knowledge, which led to proposals of hierarchically organized concepts as nodes in localist networks (Collins and Quillian, 1969; Collins and Loftus, 1975) to distributed representations in connectionist networks (Rumelhart and McClelland, 1985; Rumelhart et al., 1993; Hinton, 1981; Hinton et al., 1986; Schyns, 1991; Plunkett et al., 1992; Rogers and McClelland, 2004).

Early connectionist modeling work approached the problem by investigating networks containing units with predefined operations on hand-designed datasets, and used tools such as representational vector distances, network weight magnitudes, clustering algorithms and multidimensional scaling to assess learning (e.g. Rumelhart et al. (1993); Rogers and McClelland (2004)). More recently, Saxe et al. (2019) provided the first mathematical description for learning dynamics of structured semantic knowledge in linear networks. Saxe et al. traded network complexity to obtain a theoretical groundwork and showed how findings from prior work connect to progressive learning of the covariance structure of the input data. Regardless, building on these earlier approaches has remained challenging due to (at least) two reasons: i) Hand-crafted datasets used in the context of these studies are not scalable to larger neural networks, even though they aid interpretability of findings, and ii) Pre-defined network architectures are not flexible enough to be adopted for datasets with varying size and complexity. We aim to address these two challenges by relaxing assumptions about the size of the dataset as well as the size and type of network. To still be able to interrogate the learning trajectory of these larger-scale networks, we present a framework for elucidating learning dynamics that extend beyond tools previously used to investigate neural network learning of such structured data.

Our aim is to understand learning dynamics in deep non-linear networks given that neural computations are plausibly non-linear (e.g. (Gidon et al., 2020; Pagan et al., 2016; Ghazanfar and Nicolelis, 1997)). As a first step, we ask whether hierarchical structures for arbitrarily large datasets are learnt and encoded in these non-linear networks. Next, we ask how these dynamics compare to the ones obtained in the linear network regime. We define structured data as instances where levels that define super-categories are not directly evident from the data and aim to investigate the learning dynamics of these latent categories. We achieve this by generating datasets using graph grammars similar to Zhu et al. (2003); Kemp and Tenenbaum (2008) and present a metric to quantify the degree of latent structure directly from the covariance of the dataset. We develop a set of metrics to establish how such latent structures are expressed across different layers during learning. Finally, we use this setting to establish an unsupervised approach to quantify structured learning in such networks, as an attempt to scale up the study of learning to real world datasets. Within neuroscience, our work will enable hypotheses about the neural implementation; within artificial intelligence, it will enable understanding of how neural networks can obtain representations that capture the underlying data structures which will yield interpretability and potentially more robust and human-like generalization.

The paper is structured as follows: We first reproduce findings from the linear network regime using traditional metrics such as accuracy and distances among vector representations. Next, we ask how representations emerge by analyzing the network gradients. Then, we take a geometric approach to examine how category centers expand and compress across learning. Lastly, we propose an unsupervised metric to uncover progressive stage-like learning dynamics.

2 Methods

2.1 Dataset

Consider a dataset of N entities with P features belonging to C categories. Each feature f specifies how entities are related to one another, and can be defined over an undirected entity graph S_{ent} that maps the relationship between the entities. S_{ent} has $N + L$ nodes, where L represents latent nodes, and is parameterized by a weight matrix $\mathbf{W}$. The weight matrix defines the strength of connection between each pair of nodes. If a pair of nodes i and j are connected with an edge length e_{ij}, the weight between them is $w_{ij} = \frac{1}{e_{ij}}$, and $w_{ij} = 0$ otherwise. For our analyses, we constructed S_{ent} as a 6-level containing $C = 64$ categories with 1000 examples in each leaf category ($N = 64000$). From this tree we constructed a set of features such that they reflect the association among entities. Zhu et al. (2003), Kemp and Tenenbaum (2008), and Saxe et al. (2019) showed that feature values

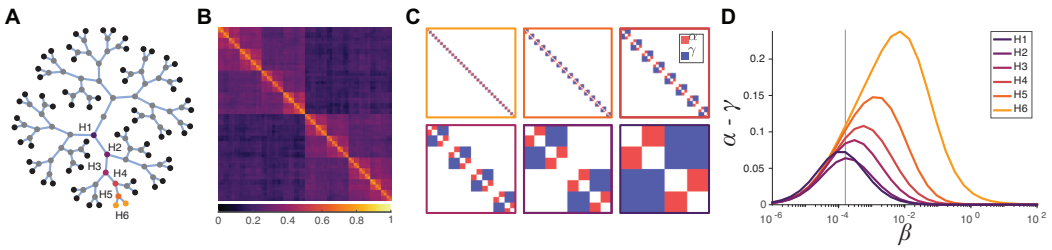

Figure 1: **Dataset. A.** Tree structure with six hierarchies. Leaf category nodes are shown as black points, and latent nodes as gray points. A single branch illustrates the color coding and hierarchy naming used throughout the paper. **B.** Covariance between entities for the hierarchical structure depicted in A. **C.** Masks for separating within-category (α) from between-category (γ) values in the covariance matrix for each hierarchy. The frame color corresponds to the hierarchy level. **D.** The difference $\alpha - \gamma$ (within - between-category) values for a range of β values (determining the strength between nodes, Eq.4). The main dataset used throughout the paper was selected to maximize distinction among all six hierarchies (gray vertical line, corresponding to the covariance matrix in B.).

can be constructed from a Gaussian distribution over S_{ent} based on $\mathbf{W}$:

$$P(f|\mathbf{W}) \propto \exp\left(-\frac{1}{4}\sum_{i,j} w_{ij}\left(f_i - f_j\right)^2\right) \tag{1}$$

Defining a degree matrix $\mathbf{D}$ and the graph Laplacian $\Delta = \mathbf{D} - \mathbf{W}$, we can rewrite Eq.1 as:

$$P(f|\mathbf{W}) \propto \exp\left(-\frac{1}{2}f^T \Delta f\right) \tag{2}$$

Eq.2 defines a Gaussian prior over each feature f with mean 0 and covariance Δ^{-1}. Following same methodology as Kemp and Tenenbaum (2008), in order to create a proper prior we fix the variance of each entity by adding a diagonal matrix $\mathbf{V}$ with values $\frac{1}{\sigma^2}$ to the covariance and define the modified covariance as $\tilde{\Delta} = \Delta + \mathbf{V}$.

$$f|\mathbf{W} \sim \mathcal{N}\left(0, \tilde{\Delta}^{-1}\right) \tag{3}$$

Next, to construct P features, we sample e_{ij} from an exponential distribution with hyperparameter β:

$$e_{ij}|\, S_{ent},\ \beta \sim \text{Exponential}(\beta) \tag{4}$$

Thus, the β parameter determines the structure-dependent covariance for each feature, while the σ^2 parameter determines its variance. This yielded a dataset $\mathbf{X}$ with size $N \times P$. Lastly, we normalized $\mathbf{X}$ by subtracting the overall mean, and dividing the square root if the maximum value of its covariance (allowing for comparisons across datasets with different parameters). See suppl. A.1 for further information on dataset selection.

2.2 Neural Network Architecture and Training Procedure

The network architecture consisted of three fully connected linear layers of size $\texttt{fc}_1\texttt{=936}$, $\texttt{fc}_2\texttt{=624}$, $\texttt{fc}_3\texttt{=208}$, $\texttt{fc}_{\texttt{softmax}}\texttt{=64}$. The hidden state in each layer was passed through a rectified linear unit (ReLU) activation function, and the hidden state in last is passed through a softmax. The network was trained with negative log likelihood loss using stochastic gradient descent with a learning rate of 0.01 and momentum 0.5, for a total of 10 epochs with batch size 32. Weights were initialized from $\mathcal{N}\left(0, 1e^{-6}\right)$ and zero bias. Every 15^{th} training batch, we evaluated model performance on an independent test batch (also size 32, 1060 test batches across 10 epochs). The linear network counterpart was identical to the network described above, except with no ReLU activations.

2.3 Gradient Distance

The aim of the gradient-based analysis was to uncover the mechanism underlying the change in the hierarchical category separation as reflected by changes in the network weights. To do so, we

first replaced the network weights with their corresponding gradient values, $\Delta\mathbf{W}$. To investigate differences among hierarchies, we randomly extracted pairs of samples from different categories conditioned on the hierarchy level, H (1500 pairs of samples for each level). Then, we computed the vector difference between sample pairs and calculated the mean thereof: $\text{diff}_{H_i}^{\Delta\mathbf{W}}$ for $i \in [1, 6]$. Next, we computed the L2 norm based on the meaned vector differences: $\left\| \sum \text{diff}_{H}^{\Delta\mathbf{W}} \right\|_2$. Finally, to obtain an aggregated metric of the gradient contribution for each hierarchy level, we computed the mean across these norm values. For each layer with P features, we normalized these averaged feature norms by $\sqrt{(\text{len}(P))}$, which we denote as the gradient distance. The gradient distance is a quantification of the gradient contribution for separating out between-category representations at a given hierarchy level. This yields a measure of the coherence of changes per feature for each hierarchy as manifested in the network weight space.

2.4 Geometric Center Covariance

The aim of the center covariance analyses was to reveal the geometric organization of the representations and how they change during the course of learning. To do so, we computed the relative cosine distance between the category centers at different levels of hierarchy. We first averaged samples that belong to the same category conditioned on each hierarchy level. Next, we computed the cosine similarity ($1 - \text{uv}/||u||_2||v||_2$) between these geometric centers. For each level of hierarchy, this yielded a square center correlation matrix with the same size at the number of categories in that hierarchy level, $P \times P$. We then created the same mask over the matrices as done previously, namely within-category α_{cos} and between-category pairs γ_{cos}. Finally, we calculated $\alpha_{cos} - \gamma_{cos}$ as a measure of separation between geometric centers.

2.5 Unsupervised K-Nearest Neighborhood

The aim of the K-Nearest Neighbor (KNN) analysis was to identify temporal dynamics of learning in an unsupervised manner without any knowledge about the category labels (inspired by Kollmorgen et al. (2020)). First, we computed a nearest neighbor graph for the representations at all time points, $\mathbf{T}$, throughout the epochs. In each of the 1060 test batches throughout learning, we randomly drew 100 subsamples, and computed the neighbors ($K_{neighbors} = 100$) based on the Euclidean distance for each sample (using the cuML library (Raschka et al., 2020)). Next, we tagged each sample with its corresponding batch number (indicating the time point in learning), and constructed a neighborhood matrix $\mathbf{K} = [\mathbf{T} \times \mathbf{S}]$ based on these tags. Each row t in $\mathbf{K}$ represents one time point of learning, and the columns for row t denote at which other time points the representations are close to the ones at time point t. We then calculated the $1, 5, 10, 15, 20, 25, 50, 75, 80, 85, 90, 95, 99$ percentiles for each t, revealing the distribution of neighbors. This was done for each time point in learning 100 samples $\times$ 1060 batches, yielding a set of vectors revealing the temporal evolution of neighbors across learning. Finally, we averaged across every 100 subsamples and normalized these vectors by $\mathbf{T}$ to obtain a relative neighborhood time for each point in training.

3 Results

We aim to investigate the learning dynamics of hierarchical structure category structure in non-linear neural networks. Thus, there are two main aspects to settle on: 1) The dataset, and 2) The neural network model. For point 1), we generate a dataset that we know contain several levels of embedded hierarchical latent structure. We wanted to avoid a bias in certain hierarchy levels being less or more learnable, and thus defined a metric to quantify the degree of separation within each hierarchy (Fig. 1, metric further explained in suppl. A.1). We selected a dataset where different levels of hierarchy have as uniform expression as possible (Fig. 1). As a control, we generated a dataset with entities that are not associated to each other beyond their target category labels (partition dataset) (suppl. A.1), used throughout the paper for comparison. For point 2), we selected a non-linear neural network (section 2.2) that is deeper than recent work on the linear counterpart thereof (Saxe et al., 2019). Additionally, as a control, we provide comparisons to the linear network throughout the paper.

3.1 Stage-like Transitions in Learning

A stage-like transition is an abrupt reorganizations of knowledge, and has been claimed to be a phenomenon of infant concept learning (Keil, 1979) and has been shown in artificial neural networks (Rogers and McClelland, 2004; Saxe et al., 2019; Cao et al., 2020). In Fig. 1, we demonstrate emergence of hierarchies in a stage-like manner in non-linear neural networks based on both test accuracy (Fig. 1A) and Euclidean distances among pairs of categories (Fig. 1B). Based on the Euclidean distances, the first difference to emerge is between the top hierarchy. This is succeeded by differentiation of intermediate and then lower hierarchies. Moreover, we show that differentiation among categories within hierarchy level commences in layer 1 and propagates through the remaining layers of the network. This is evident from the steep rise in Euclidean distance separation around learning onset in layer 1, which gradually flattens out in the other layers (Fig. 1B). As a control, we trained a linear counterpart of the network on the same hierarchical dataset, and reproduced occurrence of stage-like learning from prior work on linear network dynamics (suppl. A.2). These stage-like patterns with differences across hierarchy levels were not evident for the partition dataset (suppl. A.2), suggesting that these learning dynamics only occur in the presence of latent hierarchical structure. Thus, we conclude that non-linear networks, in addition to linear networks, have the substrate to progressively learn hierarchies in a stage-like manner, akin to human learning (e.g. Quinn and Johnson (1997); Gelman and Coley (1990)).

3.2 Hierarchical Learning Manifested in the Network Gradients

We aim to go beyond simply examining how the samples are are represented across learning (as e.g. in Fig. 2) by unraveling the computational emergence of representations. The goal is to tap into into the mechanistic underpinnings of the learning dynamics as reflected in changes of the network weights; the gradients. Traditional approaches to gradient-based analyses are inconclusive (see suppl. A.3 for an analysis on overall gradient changes across learning). Thus, we propose a metric, the gradient distance, to parse out weight changes related to learning of hierarchies (section 2.3). The gradient distance is a measure of the gradient contribution for separating out between-category representations at a given hierarchy level. Intuitively, a large gradient distance means that the categories within that hierarchy are well differentiated (Fig. 2A, section 2.3).

Based on the gradient analysis in Fig. 2, we demonstrate three findings: i) When the network accuracy starts rising, the major gradient contribution stems from the upper hierarchies. This suggests that if the network allocates gradient updates to separate out top hierarchies, learning the leaf node accuracy requires less gradient contribution. In this manner, it may be "beneficial" to learn the true statistical structure of super-categories in terms learning efficiency. ii) Based on Fig. 2B, it is evident that layer 1 most evenly separates out all hierarchies. In later layers, some of the between-category gradient distances are very similar for some hierarchy levels, indicating that the gradients no longer need to differentiate among these hierarchies. We do not evidence this differentiation across hierarchies in our control (partition) dataset (suppl. A.3). iii) In layer 1, the top hierarchy has the largest between-category gradient difference. However, in later layers: the next hierarchy takes over. This implies that the earliest layer learns to differentiate the most general (top hierarchy) features, and then more finer-grained features (from lower hierarchies) are represented later in the neural network. For comparison, the linear network counterpart shows distinct learning dynamics from the non-linear network; it does not seem to contain the capacity to differentiate categories in an ordered fashion.

3.3 Geometric Metrics Reveal Biphasic Expansion and Compression Dynamics During Learning

One of the crucial components that is not captured by our analyses so far is how the changes in representations prior to time 0.3 (epoch 3) in learning lead to subsequent increase in accuracy across hierarchies. We took a geometric approach and investigated whether the geometry of category centers might reveal the observed learning dynamics. To do so, we computed the vector mean of samples conditioned on their hierarchical categories. We aimed to investigate how geometric centers change relative to one-another as a proxy for general organization of the representations. Therefore, we computed the cosine distance between centers and used covariance masks (Fig. 4A) to study the relative organization of centers belonging to same (i.e. within-category) versus different categories (i.e. between-category). For quantification across layers and hierarchies we computed the between-category cosine distance subtracted from within-category cosine distance (section 2.4). To provide an

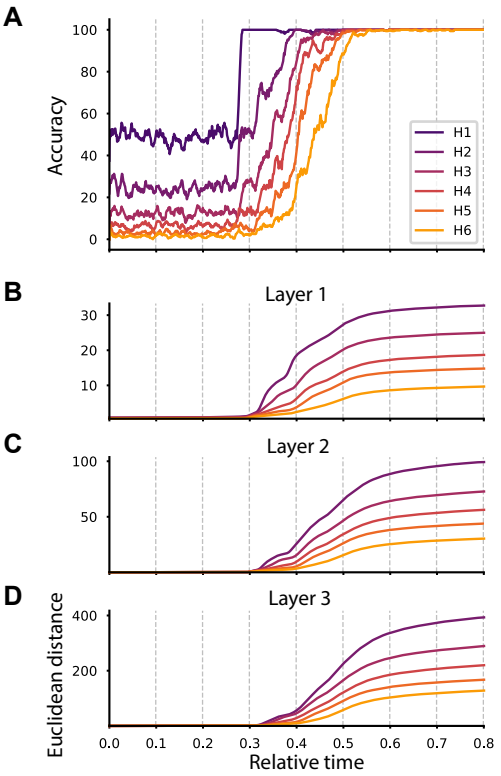

Figure 2. **Hierarchies emerge in a stage-like manner in non-linear deep networks**. **A.** Hierarchical accuracy. Test accuracies were calculated by summing up the category probabilities conditioned on each hierarchy level. The network first learns to classify among top hierarchies, and progressively learns to differentiate lower ones. **B-D.** Euclidean distances among pairs of categories from each hierarchy level as in Rogers and McClelland (2004). The average Euclidean distance between 300 sample pairs from different categories was computed at each level conditioned on membership to the same super-category. Differentiation commences in layer 1 and propagates through the subsequent layers.

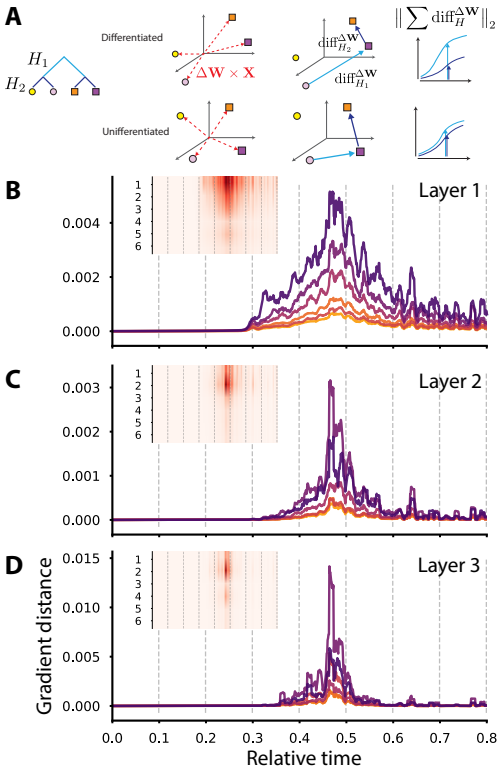

Figure 3. **Earliest layer learns the most general features A.** Schematic showing how the gradient distance metric was derived. First gradient representations were extracted, and then the distance between them were calculated based on the hierarchy level. Top panel shows an example of differentiation among hierarchies, while the bottom panel shows an undifferentiated instance. Right panel shows the expected behavior for the gradient distance given these two scenarios. **B-D.** Between-category gradient difference conditioned on hierarchies and layers. Inset depicts the same as in the main plot with less interference from overlapping lines. In layer 1, the main gradient contribution is from the top hierarchy, while in later layers, the next hierarchy takes over. moreover, early layer shows better separation between gradient distances, suggesting better differentiation of hierarchies as compared to subsequent layers.

intuition, a positive value indicates that centers belonging to the same hierarchy are more similar to each other compared to the ones belonging to different hierarchies.

Focusing on layer 1 in Fig. 4B, our analysis reveals a clear progression from general to specific categories in the representation. This occurs across layers, and more importantly, in an orderly manner from top to bottom hierarchies. During early epochs, the distance only increases for the top-most hierarchy (H_1), resulting in a separation between by time 0.3. Interestingly, at the same time the distance between categories in lower hierarchies decreases, suggesting that they are compressed to allow for separation of the top-most hierarchy. This initial phase (leading to increase in network accuracy, Fig. 2A) is followed by a second phase in which the distance for the top-most hierarchy starts to decrease, while the distance for other hierarchies increase, essentially flipping the expansion and compression behavior. Evidently, H_2 follows the same biphasic behavior; by the time 0.4 (epoch 4) the maximum separation has been reached, followed by a compression phase. Prior work in the linear network domain established a similar progression via learning of singular values of the input-output correlation matrix (Saxe et al., 2013, 2019). However, here we additionally show that in a non-linear network there is a biphasic expansion and compression in the geometry of representations that subserve stage-like learning behavior. In other words, even though in early learning the representations are organized to best learn top hierachy distinctions, they need to be distorted (unleared) to some degree afterwards to allow for learning of the lower hierarchy distinctions. This is a general phenomena that we evidence across layers and across hierarchies (Fig. 4B-D). Similarly, in earlier work by McClelland and Rogers (2003) there is also some indirect evidence of this phenomenon; the Euclidean distances among representations decrease for top hierarchies after an initial increase. Other studies on deep learning dynamics (Shwartz-Ziv and Tishby, 2017; Recanatesi et al., 2019) have also pointed to similar biphasic dynamics as we demonstrate in the current work, where expansion and compression are crucial to obtain network accuracy. Lastly, this might align with recent work on the neural basis of human concept learning, where the prefrontal cortex compresses task-irrelevant information.

3.4 Unsupervised Neighborhood Statistics Reveal Stage-Like Transitions

In all of our prior analyses we relied on knowledge about labels associated with data to reveal how learning unfolds. However, in most settings, such information is not available a priori and what a network is learning about the data is compromised by the lack of full knowledge about the data itself. Here, we take the first step in addressing such limitation by using an unsupervised neighborhood-based metric to relate geometrical changes in the representations to a temporal evaluation of neighborhood distances, inspired by a methodology recently suggested by Kollmorgen et al. (2020) to identify components in bird vocal learning. First, to gain an intuition about the overall organization of representation, we performed a tSNE analysis (Fig. 5A). We focus on layer 1 dynamics (see suppl. A.4 for other layers). Apart from few epochs, most tSNE results were hard to interpret. This might due to the fact that tSNE distorts information about relative distances, which is crucial in cases of structured, hierarchical datasets. Moreover, we relied on the category labels to interpret the tSNE visualizations. The KNN metric (section 2.5), however, reveals aspects of stage-like learning from only inspecting the neighborhood statistics with no category knowledge. During early epochs in Fig. 5B, the neighbors for each time point are from both past and future time points (large percentile values, gray lines). However, by the end of time 0.3 (epoch 3), there is a compression of neighborhood times suggesting a qualitatively different structure of representations at that stage compared to proceeding and succeeding epochs. This observation aligns well with the learning of top-most hierarchy as evidenced from geometric center covariances (Fig. 4B). Additionally, we observe a replication of such compression behavior between time 0.3 and 0.6 (epochs 3 and 6) for other hierarchy levels. Lastly, this is followed by a steady state with no additional compression or expansion. From Fig. 2A, we know that by time 0.6, the accuracy reaches its state state as well suggesting that learning in the network is stabilized by that time point. It is intriguing that such a simple metric can reveal time points of the learning that corresponds to major shifts in the representations (we do not see this behavior for our control (partition) dataset, section A.5, which suggests that the KNN metric can be further exploited under settings where the true structure of the dataset is unknown to pinpoint different stages of learning.

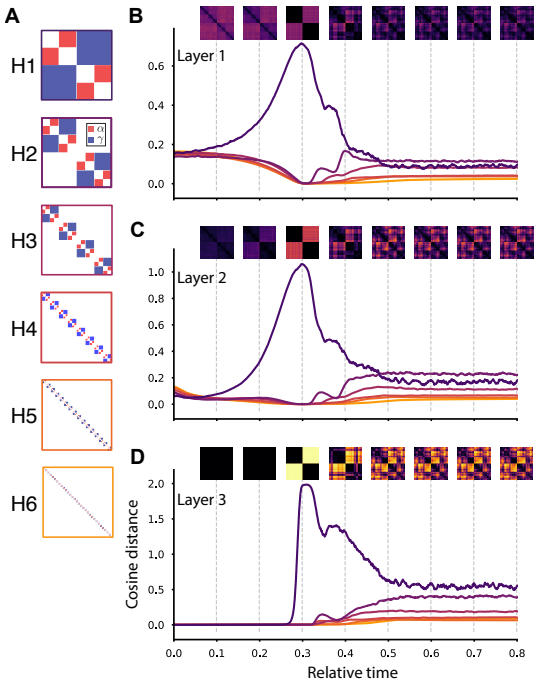

Figure 4: **Geometric center covariance reveals a progressive global to local reorganization in layer representations. A.** Covariance masks used to separate within- from between-category clusters. **B-D.** Evolution of center separation ($\alpha_{cos} - \gamma_{cos}$) across all layers. Inset shows the geometric center covariance for all 64 categories. The first major reorganization happens towards the end of time 0.3 (epoch 3) in which categories in H_1 start to separate while compressing the bottom hierarchies, $(\alpha_{cos} - \gamma_{cos}) \longrightarrow 0$, which is followed by a reversal: H_1 starts to compress allowing the progressive expansion and separation of subsequent hierarchies over the next three epochs until stabilization by the end of learning.

4 Conclusion

In this work, we have demonstrated how non-linear networks learn hierarchical categories for large datasets in a general-to-specific manner with stage-like transitions, akin to human learning (Keil, 1979; Mandler and McDonough, 1993; Quinn and Johnson, 1997; Gelman and Coley, 1990). This extends prior work on learning dynamics in the linear regime (Saxe et al., 2013, 2019; Cao et al., 2020) and network size of previous non-linear models (Rogers and McClelland, 2004; Hinton et al., 1986; Rumelhart et al., 1993). Moreover, we extend the dataset regime: Prior work has focused on one-hot vectors with associated binary properties as input (Cao et al., 2020; Rogers and McClelland, 2004; Hinton et al., 1986; Rumelhart et al., 1993; Rumelhart and McClelland, 1985), cf. (Schyns, 1991), while here we analyze large datasets without hand-crafted features or constraints besides a hierarchical covariance structure (Fig. 1, section 2.1).

To reveal the features of learning, we used established methods as well as developed supervised and unsupervised approaches to dissect how the structure of dataset gets embedded in the layer representations throughout learning. We showed a biphasic expansion and compression in the geometry of category centers that starts with general (top hierarchy) categories and propagates to specific ones. Interestingly, learning of specific categories comes at a cost of reduction in the separation of general ones, a phenomenon that to the best of our knowledge have not been observed in non-linear neural network learning of hierarchical datasets. There has been some suggestive evidence for these biphasic learning dynamics for structured data (McClelland and Rogers, 2003) as well as general studies of learning dynamics of network not specific to hierarchies (Shwartz-Ziv and Tishby, 2017; Recanatesi et al., 2019). Finally, we proposed a K-nearest neighbor method to recover stage-like transitions in an fully unsupervised manner without any knowledge about categories. We do not claim that type of dataset, the neural network, or the learning rule we used in our work is a

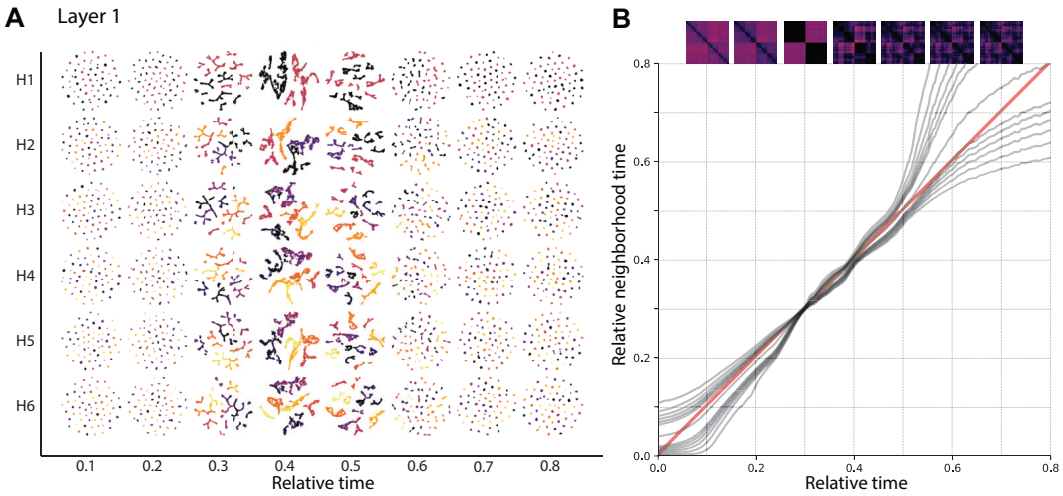

Figure 5: **Unsupervised methods recover stage-like learning of structure across hierarchies. A.** Representations across hierarchies visualized by the tSNE technique (Van der Maaten and Hinton, 2008) (learning rate=10, otherwise default parameters from scikit-learn (Pedregosa et al., 2011)). Rows denote hierarchies (H1 through H6). Early in learning, the top hierarchies get separated, however their representations get distorted subsequently. tSNE cannot uncover the behavior in the bottom hierarchies as well as top ones. **B.** Fully unsupervised KNN method for identifying time points in learning corresponding to transitions in the structure of representations. The gray lines denote different percentiles for relative neighborhood time, while the red line shows unity. The reorganization of representations corresponds to compression of relative neighbors. There are multiple phases of compression and expansions in the relative neighbor statistics corresponding to the different hierarchies being learnt.

realistic model of the brain or a region thereof. Yet, given that the brain has been suggested to perform non-linear computations (e.g. (Gidon et al., 2020; Pagan et al., 2016; Ghazanfar and Nicolelis, 1997)) and is capable of accommodating the complexity of natural sensory input, our work is a step closer to building models that can recapitulate aspects of human learning, and ultimately generate novel hypothesis about how the brain constructs structured representations of world for perception and action.

Broader Impact

Our work aims to elucidate the nature of how hierarchically structured concepts are learned by distributed neural networks, and we hope that this work will be of interest to researchers in neuroscience and cognitive sciences who use deep neural networks to understand the structure and function of the biological brain. We acknowledge that the application of our work is far in the future. Yet, if the methods presented in this work lead to applications, there can be downstream implications in education (e.g., better curriculum design for human development) and medical settings (e.g., rehabilitation from brain injury).

We provide a framework for identifying and interpreting the latent structures that are learnt and represented in deep neural networks. Interpretability is crucial for the ethical dimension of artificial intelligence to increase algorithmic fairness (Piano, 2020). On the other hand, development of methods for interpreting these deep neural network models can have negative consequences in privacy and security (e.g., adversarial probing of personal information from embedded representations within models).

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or

[N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]** [Yes] Section 1.
   (b) Did you describe the limitations of your work? **[TODO]** [Yes] Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]** [Yes] Yes, in Broader Impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]** [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]** [N/A]
   (b) Did you include complete proofs of all theoretical results? **[TODO]** [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]** [Yes] Not added in anonymous version, but available.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]** [Yes] Section 2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]** [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]** [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]** [N/A]
   (b) Did you mention the license of the assets? **[TODO]** [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]** [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]** [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]** [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]** [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]** [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]** [N/A]

Acknowledgments and Disclosure of Funding

References

Cao, Y., Summerfield, C., and Saxe, A. (2020). Characterizing emergent representations in a space of candidate learning rules for deep networks. *Advances in Neural Information Processing Systems*, 33.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247.

Gelman, S. A. and Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental psychology*, 26(5):796.

Ghazanfar, A. A. and Nicolelis, M. A. (1997). Nonlinear processing of tactile information in the thalamocortical loop. *Journal of neurophysiology*, 78(1):506–510.

Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., Holtkamp, M., Vida, I., and Larkum, M. E. (2020). Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473):83–87.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In Hinton, G. E. and Anderson, J. A., editors, *Parallel Models of Associative Memory*, pages 161–187. Erlbaum, Hillsdale, NJ.

Hinton, G. E. et al. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

Keil, F. C. (1979). Semantic and conceptual development: An ontological perspective.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.*, 105(31):10687–10692.

Kollmorgen, S., Hahnloser, R. H., and Mante, V. (2020). Nearest neighbours reveal fast and slow components of motor learning. *Nature*, 577(7791):526–530.

Mandler, J. M. and McDonough, L. (1993). Concept formation in infancy. *Cognitive development*, 8(3):291–318.

McClelland, J. L. and Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, 4(4):310–322.

Pagan, M., Simoncelli, E. P., and Rust, N. C. (2016). Neural quadratic discriminant analysis: Nonlinear decoding with v1-like computation. *Neural computation*, 28(11):2291–2319.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7(1):1–7.

Plunkett, K., Sinha, C., Møller, M. F., and Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4(3-4):293–312.

Quinn, P. C. and Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. *Journal of experimental child psychology*, 66(2):236–263.

Raschka, S., Patterson, J., and Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*.

Recanatesi, S., Farrell, M., Advani, M., Moore, T., Lajoie, G., and Shea-Brown, E. (2019). Dimensionality compression and expansion in deep neural networks. *arXiv preprint arXiv:1906.00443*.

Rogers, T. T. and McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rumelhart, D. E. and McClelland, J. L. (1985). On learning the past tenses of english verbs. Technical report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE.

Rumelhart, D. E., Todd, P. M., et al. (1993). Learning and connectionist representations. *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, 2:3–30.

Saxe, A. M., Mcclelland, J. L., and Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci. U. S. A.*

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, 15(4):461–508.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Zhu, X., Lafferty, J., and Ghahramani, Z. (2003). *Semi-supervised learning: From Gaussian fields to Gaussian processes*. School of Computer Science, Carnegie Mellon University.

A Appendix

A.1 Dataset Selection

A.2 Accuracy and Euclidean Distance Analysis

A.3 Gradient Distance

A.4 tSNE

A.5 KNN