Toward an Over-parameterized Direct-Fit Model of Visual Perception

Xin Li

xin.li@ieee.org

Abstract

In this paper, we revisit the problem of computational modeling of simple and complex cells for an over-parameterized and direct-fit model of visual perception. Unlike conventional wisdom, we highlight the difference in parallel and sequential binding mechanisms between simple and complex cells. A new proposal for abstracting them into space partitioning and composition is developed as the foundation of our new hierarchical construction. Our construction can be interpreted as a product topology-based generalization of the existing k-d tree, making it suitable for brute-force direct-fit in a high-dimensional space. The constructed model has been applied to several classical experiments in neuroscience and psychology. We provide an anti-sparse coding interpretation of the constructed vision model and show how it leads to a dynamic programming (DP)-like approximate nearest-neighbor search based on ℓ_∞-optimization. We also briefly discuss two possible implementations based on asymmetrical (decoder matters more) auto-encoder and spiking neural networks (SNN), respectively.

I. Introduction

How do we learn to see in the first six months after birth? To answer this question, David Hubel and Torsten Wiesel conducted pioneering experiments in the 1950s, leading to the discovery of simple and complex cells [45]. Inspired by their discovery, David Marr developed a theory of the neocortex [64] in 1970 and a theory of the hippocampus [65] in 1971. His computational investigation of vision [63] was published in 1982 after his death. The construction of neocognitron by Fukushima [33] and connectionist models by LeCun [53] in the 1980s represented the continuing effort to construct biologically plausible computational models for visual perception. Wavelet theory [59], [25] and sparse coding [71], [72] in the 1990s further supplied mathematical formulations of multi-resolution analysis for scale-invariant representation of images. Rapid advances in deep learning [35], [54], especially the class of over-parameterized models [7], [6] have expedited both the theory and practice of data-driven/learning-based visual processing.

Despite the great progress of today, the gap between biological and artificial vision remains significant in the following aspects. First, the network architecture of the convolutional neural network (CNN) is characterized by the pooling of layers, which reduces the dimensionality of the input data. This is in sharp contrast to the increase in the number of neurons and synapses as we move from the lower layer (e.g. V1) to the higher layer (e.g. V4) of neocortex (refer to Fig. 1). This anatomical finding has inspired H. Barlow to revise his redundancy reduction hypothesis into the redundancy exploitation hypothesis [10] in 2001. Second, although recurrent neural networks including long short-term memory (LSTM) [44] take into account temporal dynamics and have found successful applications in 1D signal analysis, the role of memory in visual perception has remained explored. In the human vision system, the hippocampus is known to play a critical role in various cognition tasks, including memory consolidation and novelty detection [52]. Finally, it remains a mystery how the human brain can manage to achieve the objectives of learning and memory with more than 100-1000 trillion synapses and a power budget of less than 20W. The challenge of breaking the conventional von Neumann architecture built upon the Turing machine remains a holy grail in neuromorphic computing.

The motivation behind this paper is two-fold. On the one hand, both the human brain and CNN are characterized by the ability to optimize an astronomical number of synaptic weights. The class of over-parameterized models [7], [6] has shown some counterintuitive properties, such as double descent [69]. Analytical tools such as neural tangent kernel (NTK) offer an approach to understanding over-parameterization in the Hilbert space, but, like all kernel methods, they are not compatible with the

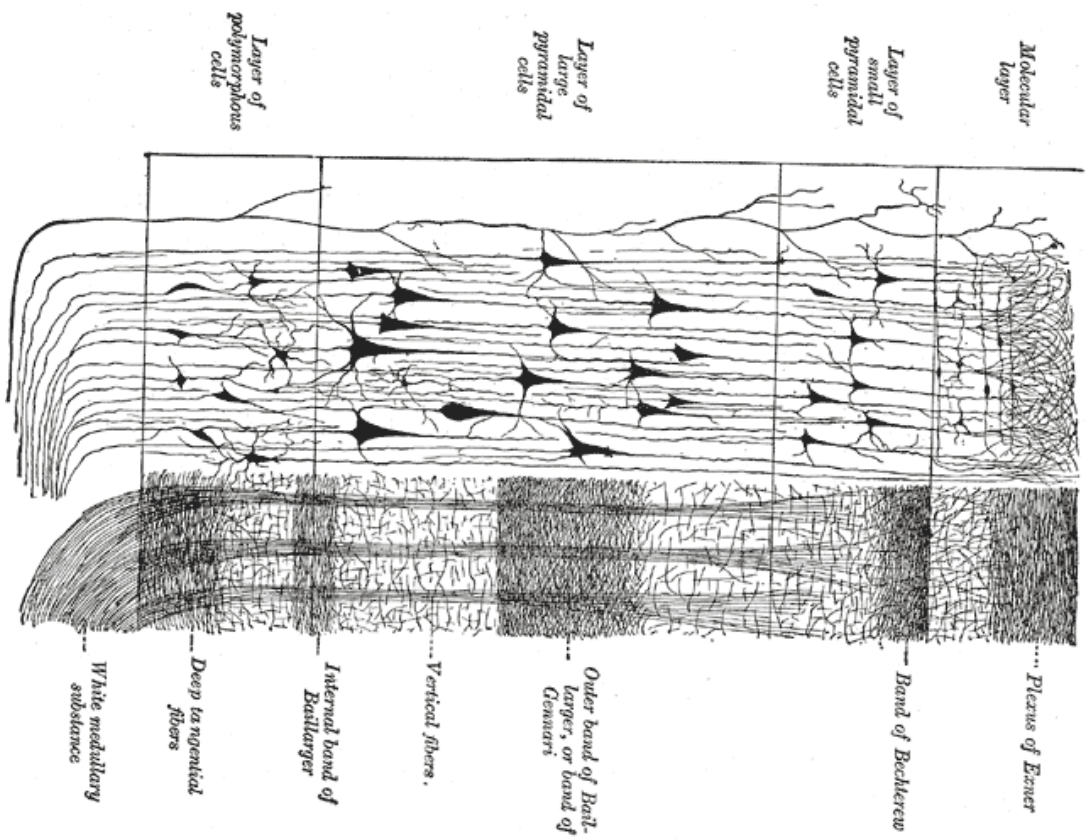


Fig. 1: The divergent architecture in neocortex - there are increasingly number of neurons and synapses as we move from low to high levels (i.e., from left to right) [36].

recursion strategy (e.g. dynamic programming that builds upon the optimality of substructures [14]). We seek to understand overparameterization in the framework of optimizing hierarchical representations [87]. On the other hand, an evolutionary perspective on biological and artificial neural networks [37] offers a direct-fit approach in brute force. Such a deceivingly simple model, when combined with over-parameterized optimization, offers an appealing solution to increase the generalization (predictive) power without explicitly modeling the unknown generative structure underlying sensory inputs.

In this paper, we construct an over-parameterized direct-fit model for visual perception. Unlike the conventional wisdom of abstracting simple and complex cells, we use space partitioning and composition as the building block of our hierarchical construction. In addition to biological plausibility, we offer a geometric analysis of our construction in topological space (i.e., topological manifolds without the definition of a distance metric or an inner product). Our construction can be interpreted as a product-topology-based generalization of the existing k-d tree [15], making it suitable for brute-force direct-fit in a high-dimensional space. In the presence of novelty/anomaly, a surrogate model that mimics the escape mechanism of the hippocampus can be activated for unsupervised continual learning [99]. The constructed model has been applied to several classical experiments in neuroscience and psychology. We also provide an anti-sparse coding interpretation [48] of the constructed vision model and a dynamic programming (DP)-like solution to approximate nearest neighbor (ANN) in high-dimensional space. Finally, we briefly discuss two possible implementations based on asymmetric autoencoder [70] and spiking neural networks (SNN) [47], respectively.

II. Neuroscience Foundation

A. Dichotomy: Excitatory and Inhibitory Neurons

During the work of Wilson and Cowan in the 1970s [93], [94], they made the crucial assumption that "all nervous processes of any complexity depend on the interaction of excitatory and inhibitory cells." Using phase plane methods, they have shown simple and multiple hysteresis phenomena and limit cycle activity with localized populations of model neurons. Their results, more or less, offer the primitive basis for memory storage, namely stimulus intensity, which can be coded in both the average spike frequency and the frequency of periodic variations in the average spike frequency [79]. However, such ad hoc sensory encoding cannot explain the sophistication of learning, memory, and recognition associated with higher functions.

B. Hebbian Learning and Anti-Hebbian Learning

Hebbian learning [40] is a dogma that claims that an increase in synaptic efficacy arises from repeated and persistent stimulation of a presynaptic cell by a postsynaptic cell. Hebbian learning rule is often summarized as "cells that fire together wire together". The physical implementation of the Hebbian learning rule has been well studied in the literature, for example, through spike timing-dependent plasticity (STDP) [18]. The mechanism of STDP is to adjust the connection strengths based on the relative timing of some neuron's input and output action potentials. STDP as a Hebbian synaptic learning rule has been demonstrated in various neural circuits, from insects to humans.

By analogy to excitatory and inhibitory neurons, it has been suggested that a reversal of Hebb's postulate, named anti-Hebbian learning, dictates the reduction (rather than increase) of the synaptic connectivity strength between neurons following a firing scenario. Synaptic plasticity that operates under the control of an anti-Hebbian learning rule has been found to occur in the cerebellum [13]. More importantly, local anti-Hebbian learning has been shown to be the foundation for forming sparse representations [29]. By connecting a layer of simple Hebbian units with modifiable anti-Hebbian feedback connections, one can learn to encode a set of patterns into a sparse representation in which statistical dependency between the elements is reduced while preserving the information. However, the sparse coding represents only a local approximation of the sensory processing machinery. To extend it to global (nonlocal) integration, we have to assume an additional postulate, called the "hierarchical organization principle ", which we will introduce in the next section.

C. Simple and Complex Cells in V1

These two classes of cells were discovered by Torsten Wiesel and David Hubel in the early 1960s [45]. Simple cells respond primarily to oriented edges and gratings, which can be mathematically characterized by Gabor filters [26]. Complex cells also respond to oriented structures; unlike simple cells, they have a degree of spatial invariance. The difference between receptive fields and the characteristics of simple and complex cells has inspired the invention of the neocognitron by Fukushima in 1979 [33], which foresaw the subsequent convolutional neural network. The hierarchical convergent nature of visual processing has also inspired the construction of the HMAX model in 1999 [82].

An important observation with the difference between simple and complex cells, as described in [45], is their neural circuits and the corresponding temporal dynamics. Simple cells are built from center-surrounding cells, which require *simultaneous summation*. On the contrary, activation of the complex cell by a moving stimulus requires *successive activation* of many simple cells. Therefore, the spatial invariance of complex cells is achieved by summation and integration of the receptive fields of simple cells. Mathematical modeling of complex cells has been extensively studied in the literature (e.g., energy model [3]). However, the abstraction strategies taken in this paper will be different from those in the open literature.

D. Mountcastle's Universal Principle

In 1978 Mountcastle suggested a universal processing principle that has been acclaimed as the Rosetta Stone of neuroscience. According to Mountcastle, all parts of the neocortex operate according to a common principle, the cortical column being the unit of computation [68]. If Mountcastle were correct, the "simple discovery" made by Hubel and Wiesel might have deeper implications in the mechanism of visual processing beyond V1. Along this line of reasoning, the striking difference between simultaneous and successive activation of simple and complex cells might illustrate a fundamental contrast between two classes of binding mechanism among neurons.

In visual perception, it has been hypothesized that the characteristics of individual objects are bound / segregated by Hebbian / anti-Hebbian learning of different groups of neurons [67]. We conjecture that there exist two types of binding mechanism (parallel vs. sequential) that are analogous to combinatorial and sequential logic in digital circuits. The former plays the role of integrating spatially overlapped parts into a whole (e.g., a horizontal edge and a vertical edge form a letter "T") or multiple features of the same object into a coherent perception (e.g., the age and gender of a face) in object recognition. Parallel binding can be interpreted as an extension of von der Malsburg's correlation theory [60]. This is the mechanism adopted by the dorsal stream to support the task of object vision. The latter is at the core of integrating spatially non-overlapped parts into a whole (e.g., the concatenation of letters into a word) in spatial vision, which belongs to the ventral stream/pathway. Sequential binding is closely related to the formation of short-term memory (e.g., Miller's law [66]) and long-term memory (e.g., Atkinson–Shiffrin model [8]) in the brain. The fundamental difference between parallel and sequential binding is that the former is invariant to permutation (the ordering of parts does not affect the perception of the whole), while the latter is sensitive to the ordering of the neuronal groups.

Binding by neuronal synchrony has been widely recognized in the literature; however, the binding problem is often thought to suffer from the so-called "superposition catastrophe" [92]. The combination coding argument often faces the dilemma of the curse of dimensionality, and it has been suggested that a representation with hierarchical structure can at least partially overcome this barrier of impracticality with combination coding. More importantly, we argue that our intuition about the capacity of the cortex might be misleading and our understanding of the power of hierarchical structures is inadequate [38]. If the curse of dimensionality can become a blessing [20], the combination coding can be made compatible with the hypothesis of redundancy exploitation [10].

III. Construction of an Over-parameterized Direct-Fit Model

In neuroscience, the principle of hierarchical organization can be roughly stated as follows: The nested structure of the physical world is mirrored by the hierarchical organization of the neocortex [38]. Unlike the mathematical construction of wavelets [25], [59], we envision that nature has discovered an elegant dynamic programming (DP)-like solution to manage the complexity of sensory stimuli in the physical world. We propose to study the following problem as the fruit-fly problem in visual perception [41].

Problem Formulation of Visual Perception

Given a visual stimulus (e.g., a sequence of images) as input, group/cluster them into different classes in an unsupervised manner.

The solution, as manifested by an infant's development of the visual cortex (primarily for ventral stream for object vision) during the first six months after birth, lies in a novel construction of a hierarchical direct-fit model based on simple and complex cells. The geometric intuition behind our construction of the new hierarchical model is best illustrated by the duality between space partitioning (i.e., subspace projection [16]) and composition (i.e., product topology [51]).

A. Preliminary on Topological Space

To facilitate our abstraction of simple and complex cells by subspace and product topology, we briefly review the basics of topological space as follows. We will follow the axiomatization of Felix Hausdorff to

construct the topological space using neighborhood as the building block. Let $\mathcal{N}$ denote the neighborhood function assigning to each point $x \in \mathbf{X}$ a non-empty subset $\mathcal{N}(x) \in \mathbf{X}$. Then the following axioms must be satisfied for $\mathbf{X}$ with $\mathcal{N}$ to be called a topological space.

1) If $\mathbf{N}$ is a neighborhood of x (i.e., $\mathbf{N} \in \mathcal{N}(x)$), then $x \in \mathbf{N}$;
2) If $\mathbf{N}$ is a subset of $\mathbf{X}$ and includes a neighborhood of x, then $\mathbf{N}$ is a neighborhood of x;
3) The intersection of two neighborhoods of x is a neighborhood of x;
4) Any neighborhood $\mathbf{N}$ of x includes a neighborhood $\mathbf{M}$ of x such that $\mathbf{N}$ is a neighborhood of each point in $\mathbf{M}$.

Note that the fourth axiom plays the role of linking the neighborhoods of different points in $\mathbf{X}$ together. Since no distance metric is defined, which is consistent with Piaget's theory of cognitive development [77], we need to define the basis as the starting point for defining a topology.

Basis of a Topology: Let $\mathbf{X}$ be a set, and suppose that $\mathcal{B}$ is a collection of subsets of $\mathbf{X}$. Then $\mathcal{B}$ is a basis for some topology in $\mathbf{X}$ if and only if the following two conditions are satisfied: a) $\cup_{\mathbf{B}\in\mathcal{B}}\mathbf{B} = \mathbf{X}$; b) If $\mathbf{B}_1, \mathbf{B}_2 \in \mathcal{B}$ and $x \in \mathbf{B}_1 \cap \mathbf{B}_2$, there exists an element $\mathbf{B}_3 \in \mathcal{B}$ such that $x \in \mathbf{B}_3 \subseteq \mathbf{B}_1 \cap \mathbf{B}_2$.

In this work, we will only consider the neighborhood basis, which is defined by

Neighborhood Basis: A neighborhood basis is a subset $\mathcal{B} \subseteq \mathcal{N}(x)$ such that for all $\mathbf{V} \in \mathcal{N}(x)$, there exists some $B \in \mathcal{B}$ such that $\mathbf{B} \subseteq \mathbf{V}$. In other words, for any neighborhood $\mathbf{V}$ we can find a neighborhood $\mathbf{B}$ on the basis of the neighborhood contained in $\mathbf{V}$.

With the above setup, the objective is to construct a hierarchical direct-fit model in the Hausdorff space (a.k.a. topological manifold), which generalizes the existing multi-resolution analysis in the Hilbert space [59]. Following our intuition above, simple and complex cells will be abstracted into subspace and product topology [57], respectively. Formally, we have the following.

Subspace Topology: Let $\mathbf{X}$ be a topological space and let $\mathbf{S} \subseteq \mathbf{X}$ be any subset. Then $\mathcal{T}_S = \{\mathbf{U} \subseteq \mathbf{S} : \mathbf{U} = \mathbf{S} \cap \mathbf{V}\ for\ some\ open\ subset\ \mathbf{V} \subseteq \mathbf{X}\}$ is the subspace topology.

Product Topology: Suppose that $\mathbf{X}_1, ..., \mathbf{X}_n$ are arbitrary topological spaces. In its Cartesian product $\mathbf{X}_1 \times ... \times \mathbf{X}_n$, the product topology is generated on the following basis: $\mathcal{B} = \{\mathbf{U}_1 \times ... \times \mathbf{U}_n : \mathbf{U}_i\ is\ an\ open\ subset\ of\ \mathbf{X}_i, i = 1, ..., n\}$.

Both subspace and product topologies have their uniqueness in terms of satisfying the characteristic property.

B. Computational Modeling of Neocortex

Simple Cells play the role of space partitioning, which can be abstracted as subspace projection [16]. A good proxy model to study the concept of space partitioning is the k-dimensional tree (k-d tree) [15], a space partitioning data structure to organize points in a k-dimensional space. The k-d tree structure can be interpreted as a class of binary space partitioning trees that extends the binary search tree (BST) for sorted arrays. It directly fits the data using hyperplanes to recursively partition the k-dimensional space. A simple variant of the k-d tree, named the random projection tree (rp tree) [24], is capable of automatically adapting to the low-dimensional structure of the data without explicit manifold learning.

The limitations of linear separability (e.g., single-layer perceptron [84]) are well known. Conventional wisdom of addressing these limitations is to introduce a nonlinear activation unit or a hidden layer (e.g., multi-layer perceptron). However, we note that some linearly non-separable set (e.g., the well-known XOR as shown in Fig. 2) can be decomposed into the linear difference between two linearly separable sets ($X \oplus Y = (X|Y) - (X \wedge Y)$). Such an observation, when combined with the data structure such as k-d/rp trees, offers a refreshing perspective of adapting to the intrinsic low-dimensional structure in high-dimensional data. That is, instead of nonlinear dimensionality reduction (e.g., the well-known Johnson-Lindenstrauss lemma [90]), we can use only a linear combination of space partitioning and differencing to approximate an arbitrary nonconvex object based on the following lemma.

Lemma 1: Convex Decomposition Lemma for Simple Cells

Any nonconvex object X can be decomposed into a finite set of differences between convex objects.

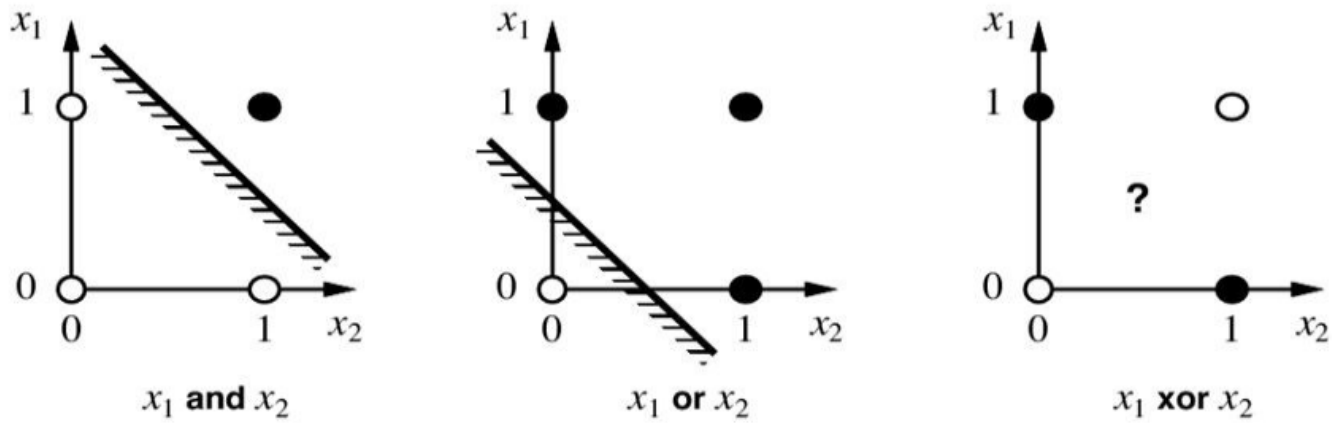


Fig. 2: XOR is not linearly separable but can be decomposed into the linear difference between linearly separable sets (the base case for proving the convex decomposition lemma).

Sketch of the Proof Prove by induction with n, which is the minimum number of convex objects $\{X_1, ..., X_n\}$ for the generation of X. Starting with $n = 2$, it is easy to construct the solution for $X = X_1 \pm X_2$. Suppose that the statement is true for the case of $n = k$, then the solution to $n = k + 1$ can be constructed similarly to $n = 2$.

Note that k-d trees have the nice property of facilitating NN/kNN search in metric space [15], [31]; but its performance degrades in high-dimensional space due to the curse of dimensionality. More importantly, one of the surprising behavior of distance metrics in high-dimensional spaces [4] is that ℓ_p ($p \to 0$) behaves much better than the popular Euclidean distance metric. Counterintuitively, in a high-dimensional space, proximity-based queries such as the NN search are meaningless and unstable because the discrimination between the nearest and farthest neighbor becomes poor. As rigorously shown in [4], the relative contrast provided by a norm with a smaller parameter p is more likely to dominate another norm with a larger parameter p as the dimensionality increases. The fractional distance concentration [30] dictates that we either work with topological space (instead of metric space) or use ℓ_0 as a pseudo-norm that recognizes the limitations of distance metrics. Interestingly, the problem of ℓ_∞-optimization has shown to be useful for an approximate NN search with anti-sparse coding [48].

Complex Cells play the role of the space composition, which can be abstracted by product topology [51]. Note that our intuition is consistent with the max/sum-pooling operation in HMAX model [82] because the objective is to achieve spatial invariance within an increased field of view. The difference lies in the way of abstraction - from simple to complex cells, we ask what will be its dual operation of space partitioning. Along this line of reasoning, if simple cells are responsible for linear separability without the change of dimensionality; complex cell must be able to increase the dimensionality for handling more sophisticated objects.

A common criticism of increasing dimensionality is concern about the so-called *curse of dimensionality* [14]. There are several ways to address this problem. First, recently proposed overparameterized models [7] or direct fit models [37] suggest that dimensionality can be a blessing when a large amount of training data is available due to a counterintuitive phenomenon called "concentration of measure" [55]. More importantly, the low-dimensional manifold structure is still preserved after space composition because of the following lemma.

Lemma 2: Product-Manifold Lemma [57] for Complex Cells

In topology, a topological manifold is a topological space that locally resembles the real n-dimensional Euclidean space. Let M be a topological m-manifold and N be a topological n-manifold, then $M \times N$ (Cartesian product of M and N) is a topological $(m + n)$-manifold.

This lemma explains the blessing of dimensionality in that the manifold structure is easier to discover in a high-dimensional space. Note that manifold learning in a higher-dimensional space requires more training data. For example, the Cartesian product of a horizontal edge and a vertical edge will produce several combinations including "T", "+", "⊢", "⊣", and "⊥". Through combination coding, our direct-fit model stores different patterns like k-d tree (i.e., the centroids of vector quantization or dictionary atoms of sparse coding) but the combination of those patterns will be further enumerated through product topology to support the direct-fit at the higher hierarchy. This enumeration perspective differs from our approach from the HMAX model [82] in which no discrimination was considered for the combination of basic patterns, as we will elaborate next.

C. Hierarchical Direct-Fit Model with Manifold-based Novelty Detection

Hierarchical Model Construction combines layers of simple and complex cells such as HMAX [82] or neocognitron [33], but with an important distinction. The network architecture of our model is not convergent, but *divergent* - one way of generalizing is to still use pooling layers, but we will consider many parallel pooling layers simultaneously. This divergent architecture is inspired by the hypothesis of redundancy exploitation [10] advocated by H. Barlow, since the number of neurons does not decrease, but increases significantly as we move to the higher level of the neocortex. It is natural to ask why higher functions in visual recognition need more neurons. In addition to the argument with combination coding, we note that achieving spatial invariance by max-pooling loses the resolution. To compensate for this sacrifice, context aggregation by dilated convolutions [98] has been developed for semantic segmentation. Mathematically, a dilated convolution is equivalent to a convolution without the follow-up max-pooling operation. Alternatively, we can still use max-pooling, but consider the generalization of dilated convolutions to deformable convolutions [23]. To achieve invariance to local geometric transformations, we can consider a pool of deformable models instead of a single one. Such a combination of space composition with deformable kernels allows us to generate invariant representations with product topology in a bottom-up fashion.

Based on the constructed hierarchical model, we note that the combination of space partitioning (by simple cells) and space composition (by complex cells) can be recursively applied to obtain an over-parameterized model in higher-dimensional space. This recursion is conceptually similar to dynamic programming (DP) [14] - at each level, the concatenation of simple and complex cells will map the visual stimuli onto a sequence of invariant representations with increasing dimensionality (field of view). Due to the product-manifold lemma, the number of training data required by our direct-fit model will not exponentially increase (thus, avoiding the curse of dimensionality). Instead, a hierarchical model allows us to automatically adapt to low-dimensional structures in high-dimensional space by extending the k-d tree as follows.

Product Manifold Tree/Forest.

The product manifold tree (pm tree) can be defined as the dual operation of the classic k-d/rp tree. Instead of space partitioning, we recursively merge low-dimensional manifolds in subspaces into higher-dimensional manifolds through product topology. A pm forest uses a pm tree as its building block and consists of an ensemble of pm trees.

How to directly fit the data to the pm tree? Such a problem has been formulated as unsupervised learning in the literature of ML [43], [11]. Unfortunately, all existing work on unsupervised learning and clustering analysis assumes a fixed dimensionality with the input data and focuses on the learning of a distance metric. Our construction of the pm tree attempts to overcome such a barrier by generalizing k-d tree-based clustering with product topology. The performance of nearest-neighbor (NN)-search is known to degrade in high-dimensional space partially due to the surprising behavior of distance metric as dimensionality increases [4]. However, such limitation can be overcome by using the following lemma (note that we do not assume any definition of a distance metric in a topological manifold).

Lemma 3: Unsupervised Clustering Lemma on PM Tree.

Let $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ denote the Cartesian product of two topological manifolds. For a vector $\boldsymbol{z} = [\boldsymbol{xy}]$, its neighborhood search can be carried out by taking the intersection of neighborhoods of $\mathbf{X}$ in $\mathbf{X}$ and $\mathbf{Y}$, in $\mathbf{Y}$ respectively.

Sketch of the Proof. It is known that the subspaces and products of the Hausdorff spaces are Hausdorff. One property of product topology is that if $\mathcal{B}_i$ is a basis for the topology of $\mathbf{X}_i$, then the set $\{\mathbf{B}_1 \times ... \times \mathbf{B}_n : \mathbf{B}_i \in \mathcal{B}_i\}$ is a basis for the product topology on $\mathbf{X}_1 \times ... \times \mathbf{X}_n$ [57]. The result follows directly from the definition of a neighborhood basis in the topological manifold.

Such hierarchical construction with interlaced simple and complex cell layers allows an organism to directly fit the visual stimuli into the multiscale model in a bottom-up fashion. However, such a feedforward process alone is insufficient; it requires a control mechanism (negative feedback) for stability and an escape mechanism (novelty detection) for adaptation. The hippocampus plays an important role in memory consolidation (both old and new) and novelty detection [52]. One way of abstracting the functionality of the hippocampus in visual perception is that it serves the dual role: 1) a control mechanism for implementing negative feedback control loops; 2) an escape mechanism for out-of-distribution (OOD) samples. The first role has been extensively studied in the literature on deep learning (for example, the backpropagation algorithm [85]). Computational modeling of novelty detection includes statistical approaches [61] and neural network-based approaches [62]. Under the framework of our direct-fit hierarchical model, we can revisit these two mechanisms as follows.

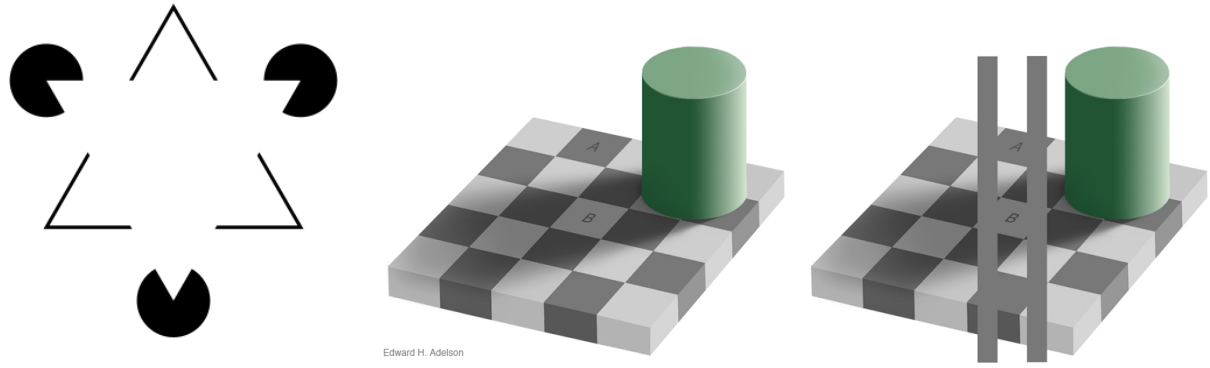
Hippocampus plays an important role in memory consolidation (both old and new) and novelty detection [52]. One way of abstracting the functionality of the hippocampus in visual perception is that it serves the dual role: 1) a control mechanism for implementing negative feedback control loops; 2) an escape mechanism for out-of-distribution (OOD) samples. The first role has been extensively studied in the literature on deep learning (for example, the backpropagation algorithm [85]). Computational modeling of novelty detection includes statistical approaches [61] and neural network-based approaches [62]. Under the framework of our direct-fit hierarchical model, we can revisit the second mechanism as follows.

Manifold-based Novelty Detection. In previous work [78], novelty detection was formulated as a twist on the manifold learning problem. First, it linearizes the parameterized manifold, capturing the underlying structure of the inlier distribution. The novelty score is then obtained by factoring in the probability and calculated with respect to the local coordinates of the tangent space of the manifold. However, the task of manifold learning is undertaken by an adversarial auto-encoder, which makes strong assumptions about the input images (e.g., fixed dimensions and inlier/outlier categories). By replacing the adversarial autoencoder with our newly constructed hierarchical model, we can achieve an improved generalization property in the following aspects. First, our construction of a product-manifold tree belongs to unsupervised learning because it directly fits the input data to a pre-constructed data structure. Second, similar to the k-d tree but working toward the opposite direction (increasing dimensionality), the product manifold tree can dynamically accommodate novel events through continual learning. The escape mechanism allows us to incorporate a novel object into the tree so that it is treated as normal in the future.

D. Supervised Learning: Perception connected to Language

In Piaget's theory of cognitive development [77], he presented a meticulous but empirical study of how children develop cognitive skills. His experimental study suggests that cognitive development consists of four stages: sensorimotor (0-2 years), preoperational (2-7 years), concrete operational (7-12 years), and formal operational (12 years and on). Recent developments in social cognitive theory [9] view people as both producers and products of social systems. Despite the wide range of topics in cognitive science, language seems to play a fundamental role in most existing cognitive models. How to endow the neocortex with the capacity of language learning is the question we attempt to connect with the proposed hierarchical direct-fit model here.

The hierarchical organization of natural languages, such as the Chomsky hierarchy [22], is a theory of formal grammars consisting of several deterministic production rules. Helen Keller's story of learning



a) Kanisza's illusion b) checkerboard illusion modified checkerboard illusion

Fig. 3: Optical illusion of an illusory contour. The White triangle is perceived as an illusory contour (the illusion will disappear as the size of the triangle increases).

the word "water" from her teacher Mrs. Sullivan [50] offers a different theory of modeling cognitive development- i.e., from unsupervised to supervised learning. The experience of Helen Keller suggests that cognitive skills, such as language, can be obtained independently of sensory modalities. Using only tactile input, Helen Keller was able to mentally reconstruct a physical world as meaningful as any other person.

The breakthrough achieved by Mrs. Sullivan is to establish the association between sensory stimuli and abstract concepts in Helen's young brain. Both input and output are driven by tactile messages (i.e., the feeling of water on the one hand and the feeling of finger dance on the other). Such an experiment suggests that the synchronization of two groups of neurons serves as the foundation of associative learning. Therefore, if the physical world can boil down to the hierarchy of associations, an overparameterized neocortex with the capacity of an astronomical number of neurons and synapses becomes a good candidate for directly fitting the formation of hierarchical associations both bottom-up and top-down.

IV. Applications to Visual Perception

A. Grandmother Cell Hypothesis and Face Identity Encoding

There is a long history of debate between localist and distributed representations in psychology and neuroscience [73]. In localist coding, a neuron codes for one familiar thing and does not directly contribute to the representation of anything else [17]. In distributed representations, knowledge is coded in a distributed manner in the mind and brain. The discovery of grandmother cells (a.k.a. Jennifer Aniston cells) [80] can be interpreted as an extreme case of sparseness. The hypothesis of a grandmother cell is also closely related to the binding problem [92] that deals with the integration of individual features into a holistic experience. Our model is in support of the hypothesis regarding the grandmother cell. Following our analysis of the ML-CSC model, the sparsity is expected to increase as the depth of the network increases. This matches our intuition that more abstraction is obtained at higher levels, implying fewer nonzero coefficients required for a global sparse representation. The extreme case will be a single coefficient or neuron. However, it should be noted that there might be massive redundancy in coding any item, such as the grandmother's face. The locations of these redundant single-neuron encodings can be distributed in the cortex.

Meanwhile, we note that familiarity is an important constraint for the grandmother cell hypothesis. Familiarity and identity of faces are two related but different concepts encoded by the medial temporal lobe (MTL) and the fusiform face area [49]. Face familiarity is related to the bias of visual stimuli ; within the framework of sparse coding, faces within the social group (e.g., grandmother's face) represent the redundancy of training data. This redundancy corresponds to a denser sampling of a local region in the face space, which implies improved sparsity. Therefore, it is easier to observe the firing of single

neurons in the class of familiar faces. On the contrary, unfamiliar faces (for example, the cross-race effect [97]) are often more difficult to recognize. This bias is associated with a degraded sparsity due to a poor sampling density. Therefore, it is more difficult to record the response of individual neurons, which is consistent with a recent finding on facial identity coding [19].

B. Optical Illusion and Context Adaptation

Optical illusions have been widely used by Gestalt psychologists as an experimental tool for studying perceptual organization in visual perception. For example, consider the famous Kanisza's illusion - an example of an illusory contour [91]. What has been less studied is the perturbed version of this illusion - i.e., as one increases the size of the black triangle and the distance among three pacmacs, it will become more and more difficult to perceive the illusory "white triangle" at the center. Such experimental findings cannot be explained by Gestalt theory because there is no prediction of the critical boundary condition for perceptual organization to fall apart. Our hierarchical direct-fit model can predict that such a threshold is determined by the size of the fovea centralis. When the distance is above this threshold, there is no previous experience (training data) to support the perceptual grouping of white triangles.

Another celebrated illusion related to context adaptation is Ted Adelson's checkerboard illusion [2] (see Fig. 3b). Two blocks with identical color values can appear to have opposite perceived luminance (brightness). Such an inconsistency is the consequence of perceptual organization; namely, the global constraint of perceiving a checkerboard dictates the interpretation at lower levels. By contrast, if two additional vertical strips having identical color are added, they interfere with the result of perceptual organization at the global level. Consequently, two blocks of the same color as two strips tend to be grouped together.

a) face image without and with eyes inverted b) upside-down face without and with eyes inverted

Fig. 4: Thatcher effect. It is much easier to detect inverted eyes from an upright face than from an upside-down face.

C. Novelty Detection of Human Vision System

Our final experiment deals with novelty detection in visual perception. In the so-called Thatcher effect [1], one can observe that it is relatively easy to detect inverted eyes and mouth from an upright face (see Fig. 4a). This detection of an anomaly (novelty) can be explained away because local patches around the eye and mouth regions are not compatible with the global impression of a human face. However, such novelty detection tasks become much more challenging with the inverted face (see Fig. 4b). Note that an inverted face is a rare event for HVS (e.g., face recognition becomes almost impossible [88]), and the hippocampus can only tell it is a novelty. Accordingly, top-down feedback in a hierarchical model will not detect the inconsistency between the inverted face and the eyes/mouth in normal position.

V. Sparse Coding Interpretation: DP-like Solution to ℓ_∞-Optimization and Approximate Nearest Neighbor Search

A. Hierarchical Convolutional Sparse Coding

It is well known that sparse coding offers a powerful analysis of the mechanism of V1 [71], [72]. Meanwhile, sparse representations have also found promising applications in unsupervised learning, such as K-SVD dictionary learning [5] and non-negative matrix factorization [56]. Unlike predetermined dictionaries (e.g., wavelet [59]), data-adaptive dictionary learning such as K-SVD based sparseland [5] has led to a multi-layer formulation of convolutional sparse coding (ML-CSC) [74], [75], which provided an attractive new theoretical framework for analyzing CNN.

Multi-Layer Convolutional Sparse Coding (ML-CSC).

The new insight brought about by ML-CSC [74], [75] is to generalize the original sparse coding in a hierarchical (multi-layer) manner. Specifically, a multilayer convolutional sparse coding (ML-CSC) model can be constructed as follows. Suppose that $\mathbf{X}$ is the input signal and a set of dictionaries is given by $\{\mathbf{D}_k\}_{k=1}^K$ where $\mathbf{D}_k$ denotes the dictionary at the level k. Then an ML-CSC model can be written as: $\mathbf{X} = \mathbf{D}_1\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2$, ..., $\mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K$ where $\mathbf{\Gamma}_i = [\boldsymbol{w}_1, ..., \boldsymbol{w}_k]$ denotes the sparse coefficients at the level i. Following the convex approximation of ℓ_0-optimization in [74], a layered thresholding algorithm runs recursively as follows.

$$\mathbf{\Gamma}_k = \mathcal{P}_{\beta_k}(\mathbf{D}_k^T\mathbf{\Gamma}_{k-1}), (k = 1, 2, ..., K) \tag{1}$$

where $\mathcal{P}_{\beta_k}$ is the standard thresholding operator and $\{\beta_k\}_{k=1}^K$ is the set of thresholds at level k.

As shown in [75], the ML-CSC model manages to decompose a signal $\mathbf{X} \in R^N$ into the superposition of multiple dictionaries $\mathbf{D} = \mathbf{D}_1\mathbf{D}_2...\mathbf{D}_K$ but the concatenation of these atoms, although it is overcomplete, remains in the space of the same dimension R^N. Unfortunately, both K-SVD and ML-CSC are still constructed within the Hilbert space without a change of dimensionality. To the best of our knowledge, no previous work exists in the open literature that extends data-adaptive sparse representation into varying dimensions. Note that wavelet theory achieves the objective of multi-resolution analysis by varying the support of basis functions. Our intuition is that a similar objective can be achieved for sparse data-adaptive representations if we can automatically adapt dictionary learning to the intrinsic low-dimensional structure of the data, such as rp trees [24]. A different way of generalizing the CSC is to construct a dictionary hierarchy with varying dimensionality.

Parallel to the product-manifold lemma, we are interested in developing a recursive strategy to decompose a high-dimensional sparse coding problem into the "product" of multiple lower-dimensional ones. Note that unlike existing work on wavelet decomposition dealing with basis construction in a Hilbert space, we opt to work with directly fitting the data, which is closer to the sparse PCA [100] and K-SVD [5]. Instead of formulating the dictionary learning in the Hilbert space with a fixed dimension (i.e., only vary dictionary size), we imagine a dynamic programming (DP)-like approach to dictionary learning from optimal substructures. The theoretical foundation of such a DP-like solution is the product manifold lemma; and sparse coding offers a well-established framework for exploiting the manifold constraint. We start from the following construction.

Hierarchical Convolutional Sparse Coding.

Let $\mathbf{X} = \mathbf{D}_x\mathbf{\Gamma}_x$ and $\mathbf{Y} = \mathbf{D}_y\mathbf{\Gamma}_y$ denote two dictionary coding schemes with $\mathbf{D}_x, \mathbf{D}_y \in R^{n\times m}, (m > n)$. Then we start with a coding scheme in the direct-sum space R^{2n} by $\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_x & \boldsymbol{z} \\ \boldsymbol{z} & \mathbf{D}_y \end{bmatrix} \begin{bmatrix} \mathbf{\Gamma}_x \\ \mathbf{\Gamma}_y \end{bmatrix}$ and improve the sparsity by basis pursuit algorithm [21] using the composite dictionary $[\mathbf{D}_{x,i}\ \mathbf{D}_{y,j}], (1 \le i, j \le m)$ (total m^2 atoms). Such process of basis composition can be recursively applied to obtain sparse bases in higher-dimensional spaces (i.e., $R^n \to R^{2n} \to R^{4n}$...).

Our hierarchical extension of the CSC shares a spirit similar to that of dynamic programming in that optimal substructures (atoms) contribute to the (nearly) optimal solution (molecule). Mathematically, ℓ_0-optimization is an NP-hard problem; but in biology, evolution does not have foresight [83] and therefore

the hierarchical systems generated by evolution might not have a globally optimal structure. What matters more appears to be the nearly decomposability of complex systems [87], as we will elaborate next.

B. Anti-sparse coding for Approximate Nearest Neighbor (ANN) Search

Instead of ℓ_0-optimization, we conjecture that ℓ_∞-optimization is a more appropriate framework for analyzing the high-level visual cortex for the following reasons. First, the strategy employed by V1 [71] follows Barlow's redundancy reduction principle [12]; however, Barlow has revised this principle to exploit redundancy in [10]. How to exploit the blessing of dimensionality by our hierarchical CSC calls for a fresh look at the definition of sparsity. Second, redundancy has been extensively exploited in information theory for reliable communication [86]. The neocortex faces a similar challenge of robustness to errors (e.g., sensory deprivation and lesions), especially for the high-level layers responsible for important decisions related to behavior. In the literature, it has been shown in [58], [32] that ℓ_∞-optimization leads to the so-called Kashin representation (a.k.a. spread representation [32]) where all coefficients are of the same order of magnitude. Such a class of representations is known to robustly withstand errors in their coefficients. Third, the anti-sparse coding scheme based on ℓ_∞-optimization is known to facilitate the search for the approximate nearest neighbor (ANN) [48]. Such an interesting connection implies that it is easier to construct a DP-like solution by decomposing the high-dimensional ANN search problem into several subproblems in projected subspaces.

Spread Representation For a given signal $\boldsymbol{x}$ in n-dimensional space with $\mathbf{D}_x \in R^{n\times m}, (m > n)$ representing the dictionary, we consider the following ℓ_∞-optimization problem.

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \parallel \boldsymbol{\alpha} \parallel_\infty, s.t.\ \mathbf{D}_x\boldsymbol{\alpha} = \boldsymbol{x} \tag{2}$$

where $\parallel \boldsymbol{\alpha} \parallel_\infty = max|\alpha_i|, i \in \{1, ..., m\}$. The solution to the above ℓ_∞-optimization problem [32] boils down to a binarization scheme with components $m-n+1$ reaching limits $\parallel \boldsymbol{\alpha} \parallel_\infty$ and the remaining $n-1$ between these two extreme values. This spread representation has led to an anti-sparse coding scheme for the ANN search [48].

Anti-sparse coding for ANN search. Given an inquiry vector $\boldsymbol{y}$, we can find its binarization $e(\boldsymbol{y}) = sign(\boldsymbol{\alpha}/ \parallel \boldsymbol{\alpha} \parallel_\infty)$ by solving the anti-sparse coding problem.

$$J_\lambda(\boldsymbol{\alpha}) = \frac{1}{2} \parallel \mathbf{D}_x\boldsymbol{\alpha} - \boldsymbol{y} \parallel_2^2 + \lambda \parallel \boldsymbol{\alpha} \parallel_\infty, \tag{3}$$

where λ is the Lagrangian multiplier. Then the problem of finding an ANN inquiry $\boldsymbol{z}$ boils down to

$$\mathbf{NN}(e(\boldsymbol{z})) = \underset{\boldsymbol{y}\in\mathcal{Y}}{\operatorname{argmax}}\, e(\boldsymbol{z})^T e(\boldsymbol{y}) \tag{4}$$

Combining anti-sparse coding with our hierarchical CSC, we can envision a DP-like solution to the ANN search in high-dimensional space. Thanks to product sparse coding lemma, we can recursively construct a dictionary in a high-dimensional space from the direct product of dictionaries in low-dimensional spaces. It follows that an anti-sparse coding scheme in high-dimensional space can be obtained by decomposing it into subproblems in low-dimensional spaces (conceptually similar to the principle of pf dynamic programming). It should be noted that space partitioning, such as the k-d tree and random projection [96] also supports the NN search [31]. Therefore, the NN/ANN search might serve as a common strategy to unify top-down and bottom-up processing mechanisms in visual perception.

C. Theoretical Analysis

Relative Invariance to Transformation and Illumination Variation. Unlike wavelet theory [59] in which scale/shift invariance is a property of basis functions, our direct-fit model achieves *relative* invariance to geometric transformations and changes in illumination by learning from the samples. Note that translational invariance is relatively easy to achieve with a maximum-pooling operator with an increased field of view. The invariance in the composition of geometric transformations is beyond the reach of conventional CNN. Our hierarchical model achieves such invariance through densely sampling the data space (guided by ego-motion). In other words, invariant representation arises from the interaction between sensory and motor system - as JJ Gibson comments on visual perception [34], "we see because we move; we move because we see." Note that such invariance is relative - for rotation that is beyond a critical threshold, the performance of visual recognition degrades rapidly (escape mechanism will be activated).

Robustness to Occlusion and Corruption. The robustness of sparse coding recognition [95] is mainly attributed to the use of the ℓ_1 norm in convex optimization. The hierarchical extension further improves the robustness by exploiting the redundancy of the representation. As shown in [27], when an overcomplete system is *incoherent*, the optimally sparse approximation to noisy and noiseless data differs from at most a constant multiple of the noise level. Our intuition is that incoherence becomes easier to satisfy in a higher-dimensional space due to the phenomenon of concentration of measures [55]. In addition to the stabilibility results proven in [74], additional robustness to missing and noisy observations comes from the ℓ_∞-optimization as rigorously demonstrated by the uncertainty principle in [58].

VI. Two Possible Implementations Using Artificial Neural Networks

A. Asymmetrical Autoencoder: Decoder Matters More

The autoencoder represents a popular network architecture for unsupervised learning. A straightforward application of the sparse coding principle to the autoencoder is possible [70]. The hierarchical extension of the CSC inspires us to consider the design of an asymmetrical autoencoder, as shown in Fig. 5. Its asymmetrical design is also biologically inspired by H. Barlow's redundancy exploitation hypothesis [10]. The hierarchical organization of the neocortex is believed to reflect the nested structure of the physical world [38], indicating that the decoder (responsible for reconstruction of internal representations) plays a more important role than the encoder. More specifically, we can implement a prototype by combining an over-parameterized autoencoder [81] with manifold-based novelty detection [78]. In the literature, autoencoder-based representation learning [89] is known to be capable of learning disentangled and hierarchical representations. Instead of storing training samples as attractors [81], we envision that a hierarchy of increasingly abstract concepts can become attractors of an over-parameterized asymmetrical autoencoder. In the presence of a new category (novelty), the escape mechanism will be activated and the decoder will be further expanded to accommodate the novelty class (i.e., consolidation of new memory).

Such unsupervised and continual learning can lead to monotonically increased memory capacity for associative memory implemented by overparameterized autoencoder. Unlike [81] treating sequence encoding as composition maps and limit cycles, we argue that a biologically more plausible mechanisms for memory storage is based on the rich interaction between sensory and motor systems. From the direct-fit perspective, motion dictates the regime within which the organism achieves the invariance to the composition of geometric transformations. The rich interaction between sensory and motor systems contributes to the formulation of reconstruction problems at multiple scales from parts to the whole [42] as an unsupervised mechanism of learning invariant representations. When the sensory motion goes out of the normal range (e.g., rotating a book continuously), the asymmetrical autoencoder will fail to recognize the object (i.e., it will be treated as a novelty).

B. Polychronous Neuronal Group (PNG) and Spiking Neural Networks

In [47], polychronization was conceived as a basic mechanism for computing with spikes. It is built upon Hebb's postulate, but extends it by relaxing the synchronous firing into polychronous time-locked

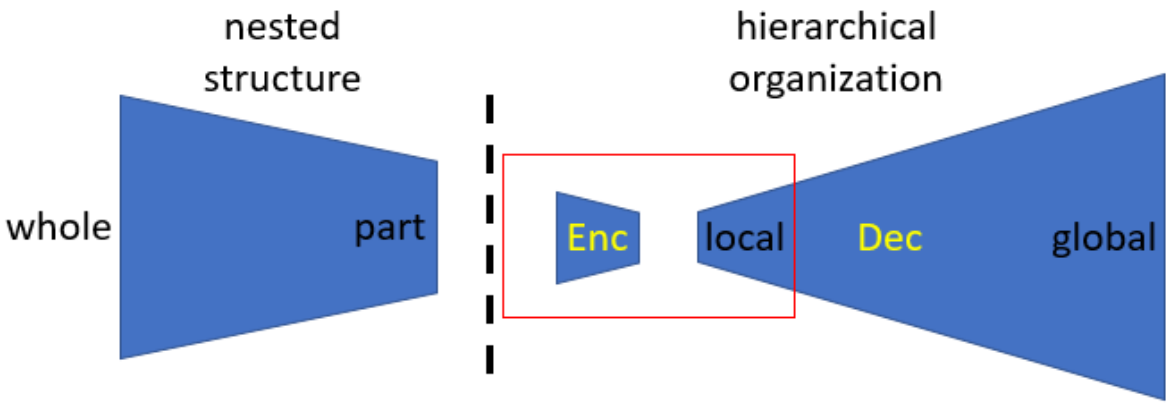
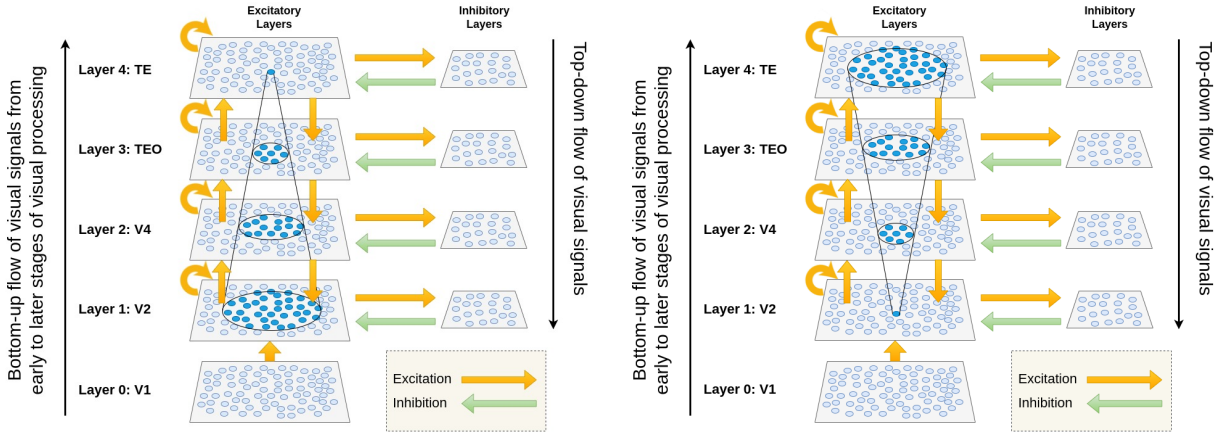


Fig. 5: Architecture of an asymmetrical autoencoder implementing the principle that nested structure in the physical world is reflected by the hierarchical organization of the neocortex (the dashed line marks the boundary between environment and organism; red box corresponds to the standard symmetric autoencoder).

patterns. Therefore, the group of neurons that are spontaneously organized by the fundamental process of spike-timing-dependent plasticity (STDP) is called Polychronous Neuronal Group (PNG). The mechanism of polychronization has recently been studied in [28], [46] as a plausible solution to the problem of feature binding. Using a spiking neural network (SNN), input training images (sensory stimuli) can be mapped to a hierarchy of PNGs by the emergence of polychronization. A mathematical abstraction of PNG is that it maps some partitioned space to a binary output (firing or no firing).



a) Convergent SNN b) Divergent SNN

Fig. 6: Current implementation of SNN [28], [46] vs. the proposed implementation (there are more cells and synapses at higher levels than at lower levels).

Unlike previous studies [28], [46], our over-parameterized direct fit model can be implemented on SNN with a divergent architecture, as shown in Fig. 6. Such a divergent architecture directly matches our intuition of abstracting complex cells by space composition. It is easy to see that the number of PNGs can grow exponentially as the number of neurons increases. Why do we need more PNGs at the higher level of the neocortex? The combinational coding argument (the end of Sect. III-B) suggests that PNGs could represent a biological plausible storage mechanism. They are physical implementations of attractors in artificial neural networks. One way to experiment with our divergent SNN is to focus on its memory capacity; the other promising direction is to build a dorsal-ventral SNN to study the binding mechanism between "what" and "where" according to recent work of SNN for object vision [39] and spatial vision [76], respectively.

VII. Conclusions

We challenge the conventional way of modeling simple/complex cells by constructing space partitioning/composition in the topological space. Without imposing additional structures such as distance metric, we can construct product-manifold trees as a novel data structure suitable for the task of direct-fit visual perception. We demonstrate the biological plausibility of our construction and offer a sparse coding interpretation. It is possible to test the developed theory by constructing an asymmetrical autoencoder or a divergent SNN. For asymmetrical autoencoder, it will be interesting to study how a divergent architecture can maximize the capacity of associative memory [81]. For divergent SNN, the objective is to experimentally verify the parallel and sequential binding mechanisms and their associated combination coding arguments, which might offer supporting evidence for the grandmother cell hypothesis.

References

[1] I. Adachi, D. P. Chou, and R. R. Hampton, "Thatcher effect in monkeys demonstrates conservation of face perception across primates," *Current Biology*, vol. 19, no. 15, pp. 1270–1273, 2009.

[2] E. H. Adelson, "Perceptual organization and the judgment of brightness," *Science*, vol. 262, no. 5142, pp. 2042–2044, 1993.

[3] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *Josa a*, vol. 2, no. 2, pp. 284–299, 1985.

[4] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*. Springer, 2001, pp. 420–434.

[5] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[6] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 242–252.

[7] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," in *International Conference on Machine Learning*. PMLR, 2018, pp. 244–253.

[8] R. C. Atkinson and R. M. Shiffrin, "Human memory: A proposed system and its control processes," in *Psychology of learning and motivation*. Elsevier, 1968, vol. 2, pp. 89–195.

[9] A. Bandura, "Social cognitive theory: An agentic perspective," *Annual review of psychology*, vol. 52, no. 1, pp. 1–26, 2001.

[10] H. Barlow, "Redundancy reduction revisited," *Network: computation in neural systems*, vol. 12, no. 3, p. 241, 2001.

[11] H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[12] H. B. Barlow *et al.*, "Possible principles underlying the transformation of sensory messages," *Sensory communication*, vol. 1, no. 01, 1961.

[13] C. C. Bell, V. Z. Han, Y. Sugawara, and K. Grant, "Synaptic plasticity in a cerebellum-like structure depends on temporal order," *Nature*, vol. 387, no. 6630, pp. 278–281, 1997.

[14] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[15] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[16] A. Blum, "Random projection, margins, kernels, and feature-selection," in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. Springer, 2005, pp. 52–68.

[17] J. S. Bowers, "On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience." *Psychological review*, vol. 116, no. 1, p. 220, 2009.

[18] N. Caporale, Y. Dan *et al.*, "Spike timing-dependent plasticity: a hebbian learning rule," *Annual review of neuroscience*, vol. 31, no. 1, pp. 25–46, 2008.

[19] L. Chang and D. Y. Tsao, "The code for facial identity in the primate brain," *Cell*, vol. 169, no. 6, pp. 1013–1028, 2017.

[20] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3025–3032.

[21] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[22] N. Chomsky, "Three models for the description of language," *IRE Transactions on information theory*, vol. 2, no. 3, pp. 113–124, 1956.

[23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[24] S. Dasgupta and Y. Freund, "Random projection trees and low dimensional manifolds," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 2008, pp. 537–546.

[25] I. Daubechies, *Ten lectures on wavelets*. SIAM, 1992.

[26] J. G. Daugman, "Complete discrete 2-d gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 7, pp. 1169–1179, 1988.

[27] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on information theory*, vol. 52, no. 1, pp. 6–18, 2005.

[28] A. Eguchi, J. B. Isbister, N. Ahmad, and S. Stringer, "The emergence of polychronization and feature binding in a spiking neural network model of the primate ventral visual system." *Psychological review*, vol. 125, no. 4, p. 545, 2018.

[29] P. Földiak, "Forming sparse representations by local anti-hebbian learning," *Biological cybernetics*, vol. 64, no. 2, pp. 165–170, 1990.
[30] D. François, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.
[31] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 209–226, 1977.
[32] J.-J. Fuchs, "Spread representations," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, 2011, pp. 814–817.
[33] K. Fukushima, "A hierarchical neural network model for associative memory," *Biological cybernetics*, vol. 50, no. 2, pp. 105–113, 1984.
[34] J. J. Gibson, "The perception of the visual world." 1950.
[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[36] H. Gray, *Anatomy of the human body*. Lea & Febiger, 1878, vol. 8.
[37] U. Hasson, S. A. Nastase, and A. Goldstein, "Direct fit to nature: an evolutionary perspective on biological and artificial neural networks," *Neuron*, vol. 105, no. 3, pp. 416–434, 2020.
[38] J. Hawkins and S. Blakeslee, *On intelligence*. Macmillan, 2004.
[39] H. Hazan, D. Saunders, D. T. Sanghavi, H. Siegelmann, and R. Kozma, "Unsupervised learning with self-organizing spiking neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–6.
[40] D. Hebb, "The organization of behavior. a neuropsychological theory," 1949.
[41] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the theory of neural computation*. CRC Press, 2018.
[42] G. Hinton, "How to represent part-whole hierarchies in a neural network," *arXiv preprint arXiv:2102.12627*, 2021.
[43] G. Hinton and T. J. Sejnowski, *Unsupervised learning: foundations of neural computation*. MIT press, 1999.
[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[45] D. H. Hubel, *Eye, brain, and vision.* Scientific American Library/Scientific American Books, 1995.
[46] J. B. Isbister, A. Eguchi, N. Ahmad, J. M. Galeazzi, M. J. Buckley, and S. Stringer, "A new approach to solving the feature-binding problem in primate vision," *Interface focus*, vol. 8, no. 4, p. 20180021, 2018.
[47] E. M. Izhikevich, "Polychronization: computation with spikes," *Neural computation*, vol. 18, no. 2, pp. 245–282, 2006.
[48] H. Jégou, T. Furon, and J.-J. Fuchs, "Anti-sparse coding for approximate nearest neighbor search," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2029–2032.
[49] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *Journal of neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.
[50] H. Keller, *The story of my life*. Library of Alexandria, 2004, vol. 1.
[51] J. L. Kelley, *General topology*. Courier Dover Publications, 2017.
[52] R. T. Knight, "Contribution of human hippocampal region to novelty detection," *Nature*, vol. 383, no. 6597, pp. 256–259, 1996.
[53] Y. Le Cun and F. Fogelman-Soulié, "Modèlearning connectionists," vol. 2, no. 1, pp. 114–143, 1987.
[54] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[55] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2001, no. 89.
[56] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[57] J. M. Lee, B. Chow, S.-C. Chu, D. Glickenstein, C. Guenther, J. Isenberg, T. Ivey, D. Knopf, P. Lu, F. Luo *et al.*, "Manifolds and differential geometry," *Topology*, vol. 643, p. 658, 2009.
[58] Y. Lyubarskii and R. Vershynin, "Uncertainty principles and vector quantization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3491–3501, 2010.
[59] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 7, pp. 674–693, 1989.
[60] C. v. d. Malsburg, "The correlation theory of brain function," in *Models of neural networks*. Springer, 1994, pp. 95–119.
[61] M. Markou and S. Singh, "Novelty detection: a review—part 1: statistical approaches," *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.
[62] ——, "Novelty detection: a review—part 2:: neural network based approaches," *Signal processing*, vol. 83, no. 12, pp. 2499–2521, 2003.
[63] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.
[64] D. Marr and W. T. Thach, "A theory of cerebellar cortex," in *From the Retina to the Neocortex*. Springer, 1991, pp. 11–50.
[65] D. Marr, D. Willshaw, and B. McNaughton, "Simple memory: a theory for archicortex," in *From the Retina to the Neocortex*. Springer, 1991, pp. 59–128.
[66] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological review*, vol. 63, pp. 81–97, 1956.
[67] P. M. Milner, "A model for visual shape recognition." *Psychological review*, vol. 81, no. 6, p. 521, 1974.
[68] V. Mountcastle, "An organizing principle for cerebral function: the unit module and the distributed system," *The mindful brain*, 1978.
[69] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.
[70] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.
[71] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.
[72] ——, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[73] M. Page, "Connectionist modelling in psychology: A localist manifesto," *Behavioral and Brain Sciences*, vol. 23, no. 4, pp. 443–467, 2000.

[74] V. Papyan, Y. Romano, and M. Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2887–2938, 2017.

[75] V. Papyan, Y. Romano, J. Sulam, and M. Elad, "Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 72–89, 2018.

[76] F. Paredes-Vallés, K. Y. Scheper, and G. C. De Croon, "Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 8, pp. 2051–2064, 2019.

[77] J. Piaget and M. T. Cook, "The origins of intelligence in children." 1952.

[78] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," *Advances in neural information processing systems*, vol. 31, 2018.

[79] G. F. Poggio and L. J. Viernstein, "Time series analysis of impulse sequences of thalamic somatic sensory neurons," *Journal of Neurophysiology*, vol. 27, no. 4, pp. 517–545, 1964.

[80] R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, "Invariant visual representation by single neurons in the human brain," *Nature*, vol. 435, no. 7045, pp. 1102–1107, 2005.

[81] A. Radhakrishnan, M. Belkin, and C. Uhler, "Overparameterized neural networks implement associative memory," *arXiv preprint arXiv:1909.12362*, 2019.

[82] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[83] S. Rose, *The future of the brain: The promise and perils of tomorrow's neuroscience*. Oxford University Press, 2005.

[84] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[85] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[86] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[87] H. A. Simon, "The architecture of complexity," *Proceedings of the American Philosophical Society*, vol. 106, no. 6, pp. 467–482, 1962.

[88] J. W. Tanaka and M. J. Farah, "Parts and wholes in face recognition," *The Quarterly journal of experimental psychology*, vol. 46, no. 2, pp. 225–245, 1993.

[89] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances in autoencoder-based representation learning," *arXiv preprint arXiv:1812.05069*, 2018.

[90] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[91] R. Von der Heydt, E. Peterhans, and G. Baumgartner, "Illusory contours and cortical neuron responses," *Science*, vol. 224, no. 4654, pp. 1260–1262, 1984.

[92] C. Von der Malsburg, "The what and why of binding: the modeler's perspective," *Neuron*, vol. 24, no. 1, pp. 95–104, 1999.

[93] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical journal*, vol. 12, no. 1, pp. 1–24, 1972.

[94] ——, "A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue," *Kybernetik*, vol. 13, no. 2, pp. 55–80, 1973.

[95] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[96] D. Yan, Y. Wang, J. Wang, H. Wang, and Z. Li, "K-nearest neighbor search by random projection forests," *IEEE Transactions on Big Data*, vol. 7, no. 1, pp. 147–157, 2019.

[97] S. G. Young, K. Hugenberg, M. J. Bernstein, and D. F. Sacco, "Perception and motivation in face recognition: A critical review of theories of the cross-race effect," *Personality and Social Psychology Review*, vol. 16, no. 2, pp. 116–142, 2012.

[98] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[99] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3987–3995.

[100] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.