

Measuring and Understanding Sensory Representations within Deep Networks Using a Numerical Optimization Framework

Chuan-Yung Tsai^{1,2}, David D. Cox^{1,2,3,*}

1 Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

2 Center for Brain Science, Harvard University, Cambridge, MA, USA

3 Department of Computer Science, Harvard University, Cambridge, MA, USA

* E-mail: davidcox@fas.harvard.edu

Abstract

Author Summary

Introduction

The mammalian ventral visual pathway consists of a hierarchical cascade of visual areas that progressively reformat visual information from pixel-like retinotopic representations into formats suitable for high-level tasks [1, 2], such as recognizing visual objects. Recently, computer vision algorithms inspired by the hierarchical structure of the visual system—convolutional neural networks [3–5], or so-called “deep learning” approaches [6]—have become the focus of tremendous attention in the computer vision and machine learning communities, due to their success in a variety of practical domains. However, we are still far from achieving human-level capabilities for unconstrained visual tasks [7], and even for the most successful current deep learning systems which, on controlled visual tasks, perform relatively close to human subjects [8–13] and produce representations similar to mammalian visual systems’ [14, 15], inexplicable failures still occur [16]. Undeniably, we lack a comprehensive functional understanding of how these systems work at the theoretical level. Here, we take the success of artificial “deep learning” architectures as an opportunity to develop tools for studying and gaining insight about the function of deep networks of nonlinear units—tools that can in principle be used to study both artificial and real neuronal systems.

Explaining the sensory representations of neurons, one of the most important aspects of studying the underpinnings of sensory processing circuitry, has been a widely researched topic in both theoretical and experimental neuroscience, where artificial and biological neural networks are mainly used and often cross-studied for developing better explanations of sensory processing mechanisms. Brief reviews of methods used in both fields are as follows. In theoretical neuroscience studies, methods can be categorized by the artificial models being used [17]: parametric or nonparametric (which are usually analytically or only numerically analyzable). Berkes et al. [18] studied quadratic networks, $f(x) = \frac{1}{2}x^T Qx + L^T x + c$, in which the optimal stimulus uniquely exists and can be efficiently computed, and local invariance and selectivity directions can be analytically derived through eigendecomposition of the symmetric quadratic term Q (i.e. the Hessian), which are eigenvectors of the least and most negative eigenvalues. Saxe et al. [19] studied single-level convolutional networks $f(x) = \|K \otimes x\|_F$ which can be viewed as the building block of deep learning networks, and showed the optimal stimulus can be analytically approximated using a Gabor-like filter with its frequency component collocated with the peak of the Fourier spectrum of convolution kernel K , whether K itself is structural or random. Zeiler et al. [20] and Simonyan et al. [21] studied multi-layer convolutional networks (i.e. deep learning networks), numerically derived and approximated the optimal stimuli of neurons of various depths, and qualitatively showed in deeper layers, neurons are tuned to gradually more complex visual patterns. Zeiler et al. [20] also used parametric deformations (translating, rotating, and scaling) to test the invariance of neurons as in [22]. Although methods in [20, 21] didn’t explicitly require the network being analytical, the capability of the network

being able to perform backpropagation is however needed. Le et al. [23] extended the method in [18] onto multi-layer convolutional networks through numerically estimating the optimal stimulus and the Hessian. Le et al. [24] also visualized the optimal stimulus of multi-layer convolutional networks, but unlike [20, 21], treated the network as non-analytical black box. Erhan et al. [25] studied multi-layer networks and proposed to characterize invariance in a larger extent, through numerically searching for non-local (i.e. farther to the optimal stimulus) solutions, instead of only locally approximating it through Hessian decomposition. Szegedy et al. [16] also used similar approach to study the selectivity of top-layer neurons inside multi-layer convolutional networks. In experimental neuroscience studies, methods can be categorized by the stimulus used to characterize the biological—naturally nonparametric—models [17]: parametric or nonparametric. An N dimensional parametric stimulus x can be either functionally generated through a P dimensional controlling parameter p , i.e. $x(p) \in \mathbb{R}^N$ and $p \in \mathbb{R}^P$ where $P < N$, or sampled from a stimulus dictionary of limited size that mostly only spans part of the N dimensional space [26], and e.g., moving bars, sine gratings, natural images, etc. fall into this category. It is an efficient and commonly adopted stimulus for studying the sensory representation of neurons in various stages along the sensory pathways, and can be coupled with characterization procedures ranging from closed-loop methods like genetic algorithm [27, 28], to open-loop methods like spike-triggered average (or reverse correlation) [29–31] and Bayesian methods [32, 33]. Contrarily, an N dimensional nonparametric stimulus $x \in \mathbb{R}^N$ can have all its variables valued independently and thus spans the entire N dimensional space. White noise is the most commonly used modality, which can be coupled with characterization procedures ranging from closed-loop methods like hill climbing [34], to open-loop methods like spike-triggered covariance [35, 36] and clearly spike-triggered average as well [29]. Methods focusing on closed-loop characterization are also reviewed and summarized by DiMattina et al. in [37].

In this work, we propose a *unified framework* based on modern numerical optimization techniques to help us uncover the sensory representation encoded by neurons deep inside neural networks, where the most flexible and generalizable settings—nonparametric network model with nonparametric stimulus—are supported to maximize the applicability of this framework, with the inherent difficulties and inefficiency resolved by carefully constructing and constraining the numerical framework. Characterization of the “tuning landscape” of neurons to the high-dimensional nonparametric stimulus includes both first- and quasi-second-order “landscape features”, i.e. optimal stimulus and its invariance and selectivity directions—the most significant structural features of the surrounding landscape—where the inefficiency of modeling the entire Hessian can be avoided. Through incorporating multiple randomized searches, the invariance and selectivity subspaces can be efficiently estimated as well. Representation measures for analyzing of numerical searching results are also designed to perform dimensionality-insensitive characterization of deep networks. Using the proposed framework for sensory representation characterization, we directly address two important questions: (1) Why are deep networks better than shallow networks? (2) Among experimented networks of the same depth, why are certain networks better than others?

Methods

Figure 1 illustrates and defines the setup of the closed-loop numerical optimization framework for studying sensory representations encoded within the deep network of interest. The proposed framework consists of two main methods: (1) optimal stimulus search and (2) invariance/selectivity path search, or the first-order and quasi-second-order characterization methods, which equivalently correspond to the linear analysis and low-rank quadratic analysis (i.e. only eigenvectors of the largest and smallest eigenvalues considered) respectively. Representation r , in this work, is equivalent to the response(s) of artificial neuron(s), and is denoted as unit representation (or scalar representation, i.e. $r \in \mathbb{R}$) when referring to single neurons and population representation (or vector representation, i.e. $r \in \mathbb{R}^R$) when referring to groups of neurons (with group size R). Arguably, for studying both artificial and biological neural networks, especially when evaluating task-related performances, the ability of characterizing population representations is usually

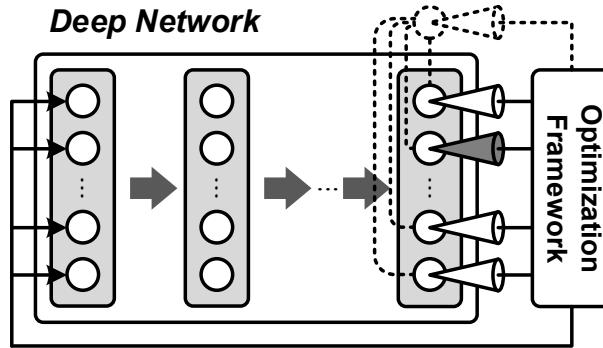


Figure 1. Schematic of numerical optimization framework. Deep network f : hierarchical cascade of layers (shaded tall rectangles) of artificial neurons (circles) transforming the stimulus x (bottom layer, leftmost) into high-level representations r (top layer, rightmost) suitable for linear classification. Optimization framework: closed-loop algorithm that repeatedly measures the response(s) of one or multiple neuron(s), calculate the fitness, and adaptively generate stimuli to characterize the target neuron(s) via optimizing the fitness functions defined for different purposes (see Eq. (1–6)); when characterizing multiple neurons for population representation of given optimal or reference stimulus, response of an imaginary neuron (dashed circle) tuned to the given stimulus is optimized (see Eq. (2, 4, 6)) equivalently.

more important than of characterizing unit representations, considering the differences in representational powers. This framework supports characterizing both unit (using Eq. (1, 3, 5), though Eq.(2, 4, 6) can be used as well) and population (using Eq.(2, 4, 6)) representations in a unified way.

As aforementioned, this framework also treats network as non-analytical black boxes and poses no assumption about networks' properties. Nevertheless, all stimuli used in this work follow the constant energy constraint $\|x\| = E$, as suggested by previous studies where simply enhancing the contrast of a fixed stimulus can as well cause increased firing rates [38,39], which in most cases leads to meaningless outcomes since the spatial “patterns” of stimulus are not changed. For simplicity of mathematical formulation, the setting of $E = 1$ is used for the rest of the paper, while in experiments E is set to the average of task-related stimulus energy before normalization. In both methods, numerical optimization are executed multiple times from randomized initial settings, for two main reasons: (1) increasing robustness of the numerical solutions of, e.g., optimal stimulus and invariance/selectivity path searches (2 runs executed), and (2) providing statistical samplings of the solution spaces of, e.g., encoding specificity and invariance/selectivity subspace searches (10 and 20 runs executed). The phase of optimal stimulus search may be omitted when studying the invariance, selectivity and encoding specificities of representations of given reference stimuli.

First-Order Characterization: Optimal Stimulus Search and Analysis

In this work, optimal stimulus is numerically derived through the iterative optimization as

$$\hat{x} = \arg \max_x f(x) = \arg \max_{x_g \in \Omega(x_{g-1}, m_{g-1})} f(x_g), \quad (1)$$

where $m_g = U(\{f(x_g)\}, m_{g-1})$ and $1 \leq g \leq G$, subject to $\|x\| = 1$. The algorithm starts from random initial point x_0 on an N dimensional sphere, as the noisy pattern shown in the bottom of Fig. 2, and samples a set of λ neighboring points (through function Ω) as candidates to be evaluated, from an initial model m_0 with null distribution (in this work, multivariate Gaussian distribution starting with covariance matrix $\Sigma = I$). Through measuring the fitnesses of sampled points $\{f(x_g)\}$, the algorithm updates its

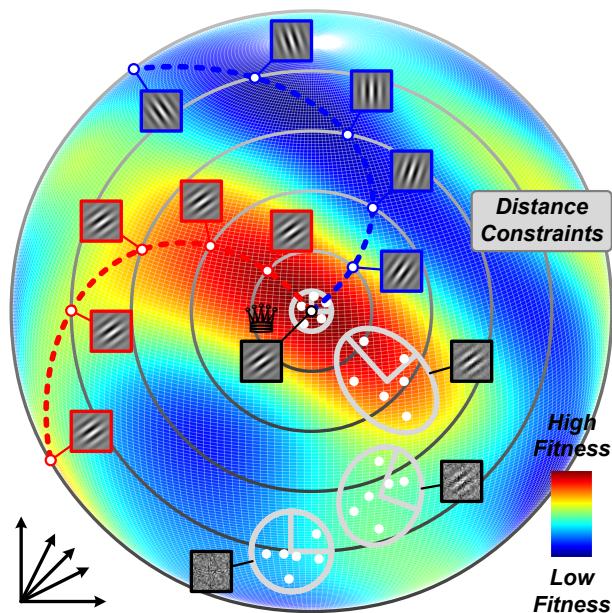


Figure 2. Visualization of numerical optimization procedures. Spherical constraint: solution space of the stimulus with energy constraint $\|x\| = 1$, an N dimensional sphere centered on the origin. Optimal stimulus search trajectory: best solutions (patterns with black frame boarders) from sequential search iterations with adaptive landscape modeling (gray eclipses with varying diameters); visualization following [40]. Distance constraints (gray circles): solution spaces of stimuli with 0.1π to 0.5π distances from the optimal stimulus. Invariance path (dashed red curve): solutions of the invariance path searches. Selectivity path (dashed blue curve): solutions of the selectivity path searches. Target neuron in this visualization is optimally tuned to a 45° Gabor pattern, invariant to phase changes, and selective to orientation changes.

model m_g (i.e. mean and covariance of the distribution, through function U) for generating samples in the following iterations, where the same operations repeat. Usually the maximally allowed iteration number G is set to prevent the algorithm from running unreasonably long; nevertheless, early termination is also practically possible when optimality is reached and the neighborhood size shrinks below threshold, as shown in the center of Fig. 2. Though the fitness landscape can be highly nonlinear, such sequential optimization procedure is usually capable of gradually accumulating knowledge of the landscape and adjusting its search directions to climb onto higher fitness areas, and the smoother the landscape is, the faster the algorithm converges. As exemplified in Fig. 2, the “most salient pattern”, or direction leading to high fitness areas usually can be rapidly extracted (see Fig. 4 for more examples). In most scenarios, fixing the maximally allowed number of fitness evaluations λG to, e.g., $100N$ (as adopted for the optimal stimulus search in this work) can lead to reasonably good convergence speed vs. quality trade-off. The aforementioned sequential optimization concept is in fact commonly adopted in modern stochastic optimization [41], and the CMA-ES (Covariance Matrix Adaptation Evolution Strategy) algorithm [42] is chosen as the back-end solver of this numerical framework, for its decent convergence speed and capability of handling rugged landscapes. Readers can refer to [42] for detailed definitions of model m and functions Ω and U . The energy constraint of stimulus is simply handled through the spherical projection before evaluation, i.e. $f(p_s(x))$ where $p_s(x) = x/\|x\|$, while the solver works unconstrainedly.

For analyzing the results of optimal stimulus searches, the following measures are adopted: (1) *Spectral complexity*, which is estimated through the L^1 norm of the (2 dimensional) Fourier power spectrum of optimal stimulus, i.e. $\|\mathcal{F}(\hat{x})\|_1$, where higher value suggests higher non-sparsity (i.e. more spectral components required to represent the signal). (2) *Explanation power*, which is estimated as the mean of the linear explainability of task-related stimuli—rectified inner-product distances between a set of n task-related stimuli $\{x^t\}$ and the optimal stimulus itself, i.e. $\frac{1}{n} \sum_{i=1}^n \max(\langle x_i^t, \hat{x} \rangle, 0)$ —and measures how well the optimal stimulus can linearly “approximate” task-related stimuli (or intuitively how much neuronal response might be elicited). (3) *Encoding specificity*, which utilizes the optimal stimulus search as an inverse function f^{-1} of a representation r (i.e. searching for the optimal stimulus of an imaginary neuron, as shown in Fig. 1, tuned to the representation of a reference stimulus) and measures the average of structural similarities (SSIM) [43] between task-related reference stimulus x^* and n reconstructed stimuli (with randomized initializations), i.e. $\frac{1}{n} \sum_{i=1}^n \text{SSIM}(x^*, \hat{x}_i^*)$ where

$$\hat{x}^* = \arg \max_x \left(e^{-\|f(x)-r\|} \right) \quad (2)$$

and $r = f(x^*)$, indicating how specific (or non-confounding) a representation r is encoded.

Quasi-Second-Order Characterization: Invariance and Selectivity Path Search and Analysis

With respect to the optimal stimulus \hat{x} , the searches of invariance and selectivity paths, which consist of sets of invariant stimuli ($\{x_\delta^+\}$, i.e. optimally excitatory) and selective stimuli ($\{x_\delta^-\}$, i.e. optimally inhibitory) respectively, are formulated as

$$x_\delta^+ = \arg \max_{x_\delta} f(x_\delta) \quad (3)$$

$$= \arg \max_{x_\delta} \left(e^{-\|f(x_\delta)-f(\hat{x})\|} \right); \quad (4)$$

$$x_\delta^- = \arg \min_{x_\delta} f(x_\delta) \quad (5)$$

$$= \arg \min_{x_\delta} \left(e^{-\|f(x_\delta)-f(\hat{x})\|} \right); \quad (6)$$

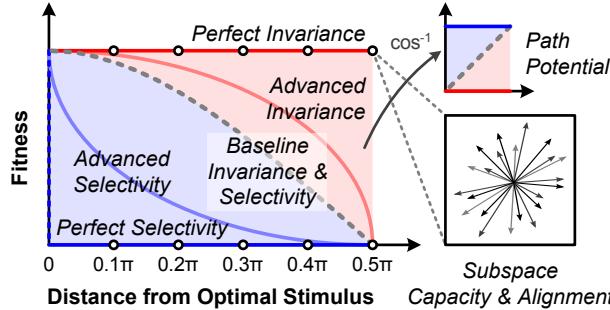


Figure 3. Fitness-distance diagram. Invariance (red) and selectivity (blue) curves: fitnesses of invariance and selectivity path search results plotted against distances from the optimal stimulus. Baseline (dashed gray) curve: graph of cosine function; invariance and selectivity curves of single inner-product neuron. Subsequent analyses, including path potential, subspace capacity and subspace alignment, can be performed. Invariance and selectivity path potentials can be defined on either the arccosine normalized diagram (shaded red and blue areas), or on the original diagram (line integrals).

where $0 < \delta \leq \frac{\pi}{2}$, subject to $\|x_\delta\| = 1$ and $\langle x_\delta, \hat{x} \rangle = \cos(\delta)$. As visualized in Fig. 2, the invariance path characterizes the N dimensional “curve” that leads toward (or ideally maintains at) fitness as high as possible while moving away from the optimal stimulus, and the selectivity path characterizes the “curve” that leads toward fitness as low as possible while moving away. The search process is implemented as multiple runs of maximization/minimization on discretized $\delta \in \{0.1\pi, 0.2\pi, 0.3\pi, 0.4\pi, 0.5\pi\}$, as the circular distance constraints shown in Fig. 2, where each run is initialized with the result from previous run (and the 0.1π run directly with optimal stimulus \hat{x}) to increase the path continuity and searching speed. Each δ of both paths has a budget of $20N$ fitness evaluations which in this work appears to be practically sufficient. The maximization and minimization simply use the same back-end solver, and the linear constraint $\langle x_\delta, \hat{x} \rangle = \cos(\delta)$ can be easily handled through an extra conic projection before evaluation, i.e. $f(p_c(p_s(x)))$ where $p_c(x) = \cos(\delta) \hat{x} + \sin(\delta) \dot{x} / \|\dot{x}\|$ and $\dot{x} = x - \langle \hat{x}, x \rangle \hat{x}$. The way the simple linear constraint is constructed to enforce the exploration of a larger extent of the fitness landscape is the one of the main differences compared to [25]. In this work, the distance constraint δ only goes up to $\frac{\pi}{2}$, since on the N dimensional sphere, stimuli fall in $\frac{\pi}{2} < \delta \leq \pi$ are simply “negatives” of those in $0 < \delta \leq \frac{\pi}{2}$, which are of less uniqueness and interest; nevertheless, going up to the full range $0 < \delta \leq \pi$ is numerically supported.

For analyzing the results of invariance/selectivity path searches, the following measures are adopted:

- (1) *Path potential*: while the search results of invariance/selectivity paths can be visualized via the *fitness-distance diagram* [44] as shown in Fig. 3 where perfect invariance and selectivity are flat lines at highest and lowest fitnesses respectively, the baselines of invariance and selectivity also can be intuitively (and analytically) defined as paths of the simplest form of neural networks—an inner-product neuron, $f(x) = w^T x$ —which precisely overlap and follow the monotonic cosine falloff from its optimal stimulus $\hat{x} = w$, as $f(x_\delta) = \cos(x_\delta)$ by definition; the invariance (and selectivity, similarly) path potential of a unit representation can thus be defined as $\int_0^{\frac{\pi}{2}} |\cos^{-1}(f(x_\delta^+)) - \delta| d\delta / \frac{\pi}{2}$, the area sandwiched between the invariance (and selectivity) and baseline curves (in \cos^{-1} domain such that both potential values fall in the range $[0, 1]$), measuring how invariant (and selective) a target neuron is compared to the baseline (i.e. zero invariance/selectivity potential); for population representation where no baseline can be easily identified, the invariance (and selectivity) path potential is alternatively defined as $\int_0^{\frac{\pi}{2}} \exp(-\|f(x_\delta^+) - f(\hat{x})\|) d\delta / \frac{\pi}{2}$.
- (2) *Subspace capacity*: compared to path potential, which is designed to characterize the best invariance/selectivity path even when only one of such exists, subspace capacity estimates the “dimensionality” (i.e. how diverse different paths can be) of the linear subspace formed by multiple path search results via

the nuclear norm of concatenation of n results $\|[x_{\delta,1}, \dots, x_{\delta,n}]\|_*$ (in this work $n = 20$ runs at $\delta = 0.1\pi$). (3) *Subspace alignment*, which measures the alignment between subspaces formed by task-related stimuli and invariance/selectivity paths (i.e. how likely the invariance/selectivity can benefit, e.g., stimulus recognition) via estimating the sparsity of n path search results projected onto the principal component vectors V of task-related stimuli $\{x^t\}$, i.e. $\frac{1}{n} \sum_{i=1}^n \|Vx_{\delta,i}\|_1$ ($n = 20$ runs at $\delta = 0.1\pi$ as well).

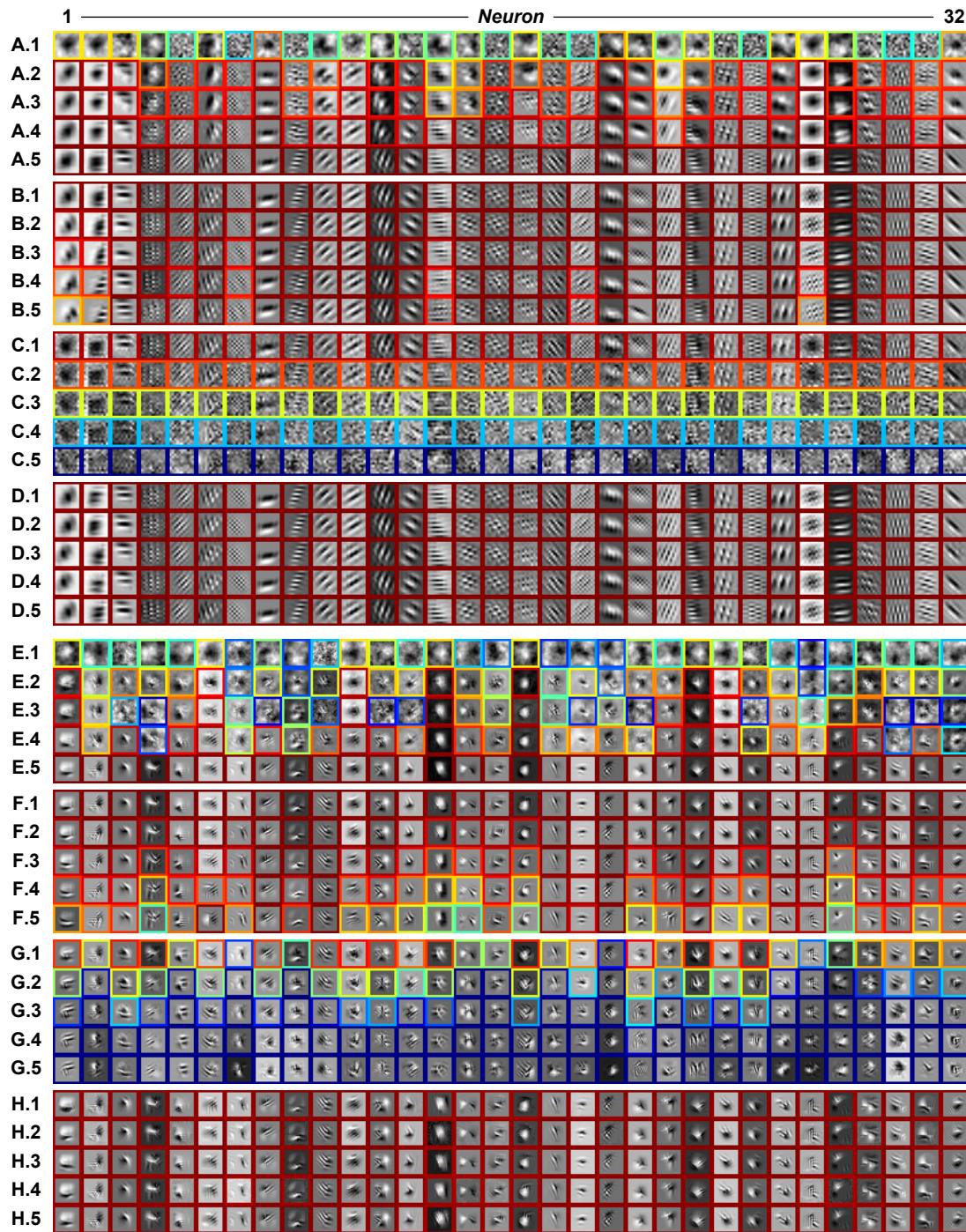
Results

The methods and measures proposed in this work are tested on the Spatio-Temporal Hierarchical Object Representation (STHOR) network [45, 46], an instance of various convolutional neural network (i.e. deep learning algorithm) implementations [3–6], which also consists of the standard cascade of levels of “convolution, nonlinear activation, pooling, and normalization” layers (which together define a single “level” in this work), for its highly parameterizable design (i.e. wide diversity of network structures can be tested) and decent performance on various visual recognition tasks [13, 45, 47] even without any convolution kernel training (i.e. fast experiments on large number of networks can be achieved, using kernels with random values). In this work, 100 shallow networks (i.e. one-level, or L1, models) and 100 deep networks (i.e. two-level, or L2, models) are randomly generated (all with 32 top-layer neurons) and tested to see how their representations vary with networks’ depths and affect networks’ performances (as shown in Fig. S1) on face pair matching tasks—accuracy of identifying pairs of different pictures from the same person and rejecting those from different persons—against the Labeled Faced in the Wild (LFW/LFW-a) dataset [48, 49]. Stimulus dimensionalities of shallow and deep neurons are $N = 121$ and $N = 441$ (i.e. 11×11 and 21×21 receptive field sizes) respectively.

Shallow vs. Deep Representations

Figure 4A and 4E exemplify the trajectories of optimal stimulus searches of 32 top-layer neurons from the best performing shallow network and 32 from the best performing deep network respectively. Worth to note, to further increase the searching speed, the initial stimulus is selected from 1000 $1/f^\alpha$ random stimuli with the best fitness, where $\alpha \in \{-4, -3, -2, -1, 0\}$, without sacrificing the nature of random initialization. From the results, two major differences can be observed. First, the fitness landscapes of deep neurons are significantly more complex than those of shallow neurons, as a large portion of searching trajectories for deep neurons actually go down then up before reaching the optima, which also suggests this numerical framework can handle non-convexity reasonably well. Second, the optimal stimuli of deep neurons are also relatively more complex (i.e. consisting of more textural components) than those of shallow neurons (which are more uni-textured or Gabor-like).

Figure 4B/4C and 4F/4G exemplify the invariance/selectivity paths on the same sets of shallow and deep neurons as in Fig. 4A and 4D. It can be observed that, while invariance paths are mostly phase changes and selectivity paths are leading toward meaningless noises (all at the same falloff rate) for shallow neurons, both types of paths consist of sophisticated shape deformations for deep neurons. Also, although most shallow neurons are selective to manually rotated optimal stimuli (i.e. Gabor filters of different orientations) as well, their fitnesses still do not drop faster than the nonparametric numerical solutions as presented in Fig. 4C. In fact, this intriguing difference generalizes across all shallow vs. deep neurons tested in this work. As the fitness-distance diagrams shown in Fig. 5, on average, though shallow neurons (Fig. 5A) have good invariance (i.e. red curve stays high), they do not have any selectivity (i.e. blue curve drops slow and completely overlaps with baseline), while deep neurons (Fig. 5B) show both good invariance and selectivity (i.e. red curve stays high; blue curve drops fast). Random walks on the fitness landscape [44], shown as the green curve in Fig. 5, can be defined as $f(\cos(\delta)\hat{x} + \sin(\delta)\tilde{x})$ where \tilde{x} is any random stimulus such that $\|\tilde{x}\| = 1$ and $\langle \hat{x}, \tilde{x} \rangle = 0$, which can be obtained through random projection of Null(\hat{x}). Comparing the gaps between selectivity and random-walk curves in Fig. 5A and 5B again



shows the selectivity of deep neurons are in fact way more “selective” than of shallow neurons. To further characterize the invariance properties, multiple runs of invariance path searches are executed and results are visualized in Fig. 4D and 4H. The diversity/dimensionality (i.e. subspace capacity) of the “invariance plateau” (e.g. the central high fitness region in Fig. 2) of deep neurons is also more apparent than that of shallow neurons. Worth to note, it can be seen in Fig. 5 that small fractions of invariance path search results in fact have higher fitness than the optimal stimuli due to nature of non-convex problems, which however does not cause noticeable differences on resultant statistics.

Comparisons of spectral complexities, invariance/selectivity path potentials, and invariance subspace capacities of shallow and deep neurons (i.e. unit representations) are summarized in Fig. 6, where all measures are normalized to have analytical upper and lower bounds being 1 and 0. These measures break down the differences between shallow and deep representations, in addition to the straightforward representation separability/classifiability as used in previous studies [20,50]. For instance, subtle differences of visual features have better chances being distinguished when the “gap” between invariance and selectivity curves (i.e. the dynamic range of neuronal responses, or amplification ratio of differences) is large. The increasing of complexity in deeper neurons’ preferred/optimal stimuli also agrees with neurobiological findings in higher visual cortical areas (e.g. V2 and V4 [51, 52]). In addition, the highly structural representations of neurons from deep networks with purely random convolution kernels found in the experiments explain why such type of random deep networks can still have good performances, on top of the theoretical analysis of shallow networks [19].

Good vs. Bad Representations

To test how well optimal stimulus may linearly explain task-related stimuli and how the explanation power correlates with network’s performance, 10,000 image patches are randomly sampled out of the predefined facial regions of pictures from the LFW-a dataset [49] and compared against the optimal stimuli of all 32 top-layer neurons from all 100 deep networks. Figure 7 shows the results of best 3 (in 7A–C) and worst 1 (in 7D) networks in terms of explanation power. For simplicity and clarity of presentation, the explanation power measure is calculated using the top 1% (i.e. 100) of the task-related stimuli with highest explainabilities, without biasing the results compared to using all task-related stimuli. It can be observed that, though none of the task-related stimuli (or arguably, of all natural stimuli) is highly similar to any optimal stimulus, there is an $R^2 = 0.36$ correlation (with details shown in Fig. S2) between network’s explanation power and performance, and when alternatively comparing the average of the top 1% of task-related stimuli and the optimal stimuli, the similarities become apparent.

It has been shown that invariance and selectivity are extremely important for recognition in various modalities and species [53–55]; however, to how these properties directly affect the recognition performances

Figure 4 (preceding page). Visualization of shallow and deep representations. (A) Optimal stimulus search trajectories of shallow neurons. Each column of (A) demonstrates the optimal stimulus search trajectory of one neuron, by showing the initial $1/f$ random stimulus in (A.1), resultant optimal stimulus in (A.5), and 3 intermediate stimuli corresponding to 3 largest second derivatives (i.e. curvatures) of the fitness history in (A.2–4). (B) Invariance paths of shallow neurons. Each column of (B) demonstrates the invariance path search results of one neuron, starting from corresponding optimal stimulus as shown in (A.5) and moving away with distance constraints from $\delta = 0.1\pi$ to 0.5π as shown in (B.1–5) accordingly. (C) Selectivity paths of shallow neurons. Definitions follow (B). (D) Invariance subspaces of shallow neurons. Each column of (D) randomly shows 5 results out of 20 runs of invariance path searches at $\delta = 0.1\pi$. (E–H) Optimal stimulus search trajectories, invariance paths, selectivity paths, and invariance subspaces, of deep neurons, respectively. Definitions follow (A–D). Color indicates fitness (i.e. response) of a neuron (definition of color map follows Fig. 2).

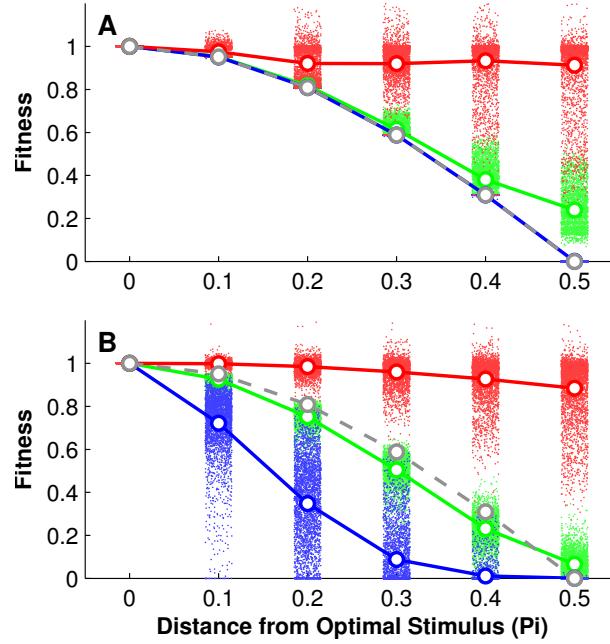


Figure 5. Comparison of shallow and deep fitness landscapes. (A) Fitness-distance diagram of shallow neurons. Red, blue and green dots indicate results of invariance and selectivity path searches, and random walks. Means of results are plotted as solid lines in corresponding colors, and the cosine baseline curve as dashed line in gray. (B) Fitness-distance diagram of deep neurons. Definitions follow (A).

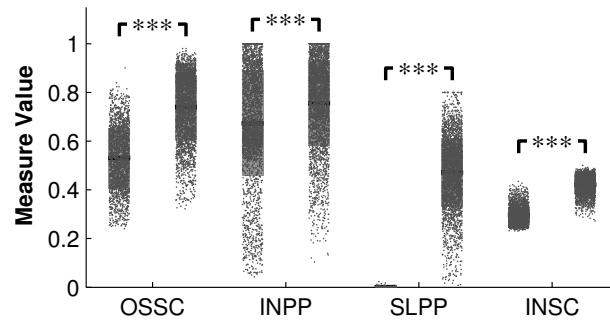


Figure 6. Comparisons of shallow and deep representation measures. Representation measures from left to right are, optimal stimulus spectral complexity (OSSC), invariance path potential (INPP), selectivity path potential (SLPP), and invariance subspace capacity (INSC), where data points from shallow networks (left distributions) and deep networks (right distributions) are shown side by side. Significances of differences of means are obtained through permutation tests, and $p < 0.001$ in all four measures. Sensitivity indexes (d') between distributions of shallow and deep representation measures from left to right are 1.63, 0.42, 4.73, and 3.67.



Figure 7. Explanation power of optimal stimulus. (A) Deep network with the best average explanation power of optimal stimulus. Each column of row (A.1) shows the optimal stimulus of the corresponding neuron. Each column of row (A.3–5) shows the task-related stimuli with the first, second, and third highest explainabilities (i.e. inner-product distances to the optimal stimulus) respectively. Each column of row (A.2) shows the average of task-related stimuli with top 1% explainabilities. (B–D) Deep networks with the second best, third best, and worst average explanation power of optimal stimulus, respectively. Definitions of rows follow (A). Color indicates explainability, and higher means better (definition of color map follows Fig. 2).

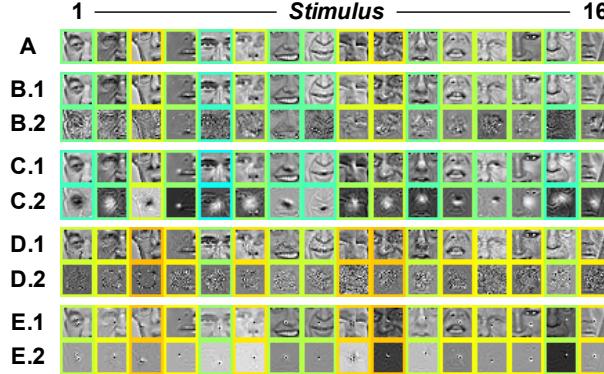


Figure 8. Subspace alignment between invariance and selectivity paths and task-related stimuli. (A) Reference stimuli used as starting points of invariance and selectivity path searches. (B) Best invariance subspace alignments. Each column of row (B.1) shows an invariance path search result with the best alignments against the principle component vector space of task-related stimuli (i.e. the eigenface vector space). Each column of row (B.2) shows the difference between result in (B.1) and corresponding reference stimulus in (A) for better visual comparison. All columns do not necessary come from the same deep network. (C–E) Best selectivity subspace alignments, worst invariance subspace alignments, and worst selectivity subspace alignments, respectively. Definitions of rows follow (B). Color indicates alignment measure, and lower means better (definition of color map follows Fig. 2).

remains unclear. To address this puzzle, starting from 16 randomly sampled reference stimuli (as shown in Fig. 8A), invariance and selectivity path searches are performed at $\delta = 0.1\pi$ using Eq. (4, 6) for population representations. Figure 8B–E demonstrates the results of invariance and selectivity subspace alignment analyses. It can be observed that, good alignments (i.e. sparser representations in principle component vector spaces) correspond to highly structural deformations or lighting changes, which are known to be important factors in visual recognition, while bad alignments correspond to mostly meaningless noises. The correlation between network’s invariance/selectivity subspace alignment measure and performance is $R^2 = 0.19/0.31$ (with details shown in Fig. S3). Network’s invariance subspace capacity (in terms of population representation), which though has an $R^2 = 0.15$ correlation as well, is in fact negatively correlated to network’s performance (with details shown in Fig. S4), mainly due to the fact that poor performing network’s invariance path search results are usually noisier (as depicted in Fig. 8D), and thus should be interpreted differently from unit representation cases. Invariance/selectivity path potentials for population representations, on the other hand, are weaker in explaining networks’ performances (with $R^2 = 0.10/0.06$), due to the fact that all deep networks have high but less ordered invariance and selectivity (as the fitness-distance diagram shown in Fig. S5).

To further understand the meaning of population representation r formed by a deep network f , the inverse function $f^{-1}(r)$ is approximated through the numerical optimization framework to reveal what kind of stimulus can drive the deep network to give output r . Instead of inverting randomly generated r , which may not have feasible or interpretable solutions, we use r from known reference stimulus x^* , and study the encoding specificity and its relationship to performance. Figure 9 shows the reference stimuli (in 9A), top 3 (in 9B–D) and worst 1 (in 9A) examples of encoding specificity, and the overall correlation between encoding specificity and performance is $R^2 = 0.41$ (with details shown in Fig. S6; SSIM measured using default settings of the standard implementation [43] and different local window sizes yield similar correlations; other similarity measures like SNR, PSNR, and inner-product distance yield similar correlations as well). It can be observed that, though the reconstructed stimuli are overall of lower spatial frequencies, SSIM is still able to capture the relative similarity (e.g. reconstructions of

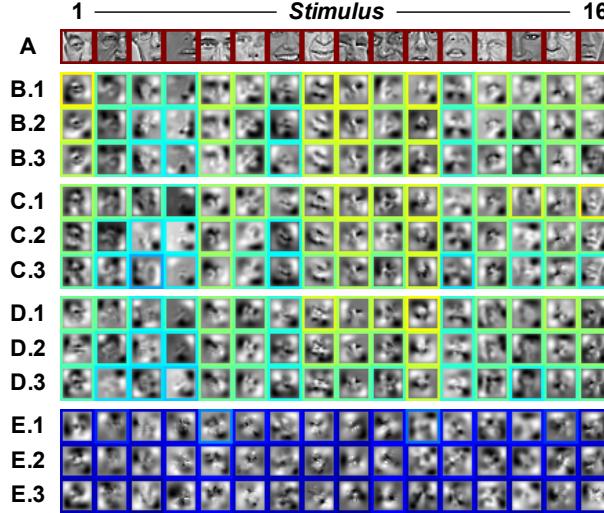


Figure 9. Encoding specificity of task-related stimulus. (A) Reference stimuli used as sources of representations. (B) Best encoding specificities. Each column of row (B.1–3) shows reconstructed stimuli of corresponding reference stimulus in (A) with top 3 SSIM measures (out of 10 reconstructions), from deep network of the best average encoding specificity. All columns do not necessarily come from the same deep network. (C–E) Second best, third best, and worst encoding specificities, respectively. Definitions of rows follow (B). Color indicates SSIM measure, and higher means better (definition of color map follows Fig. 2).

task-related stimulus 1 and 5).

Spearman's correlations between all representation measures and deep networks' performances are summarized in Fig. 10, and 71% of the variance of deep networks' performances can be explained by the proposed measures altogether. These measures further break down the differences between deep representations and suggest features like network's explanation power, invariance and selectivity subspace properties, and encoding specificity against the task-related stimuli can be extremely important for the effectiveness of representations. Encoding specificity, as a single representation measure, can best explain the deep network's performance. Though bearing strong similarity to the invariance path search in terms of mathematical formulation, encoding specificity is in fact a radically different and complementary measure, since its numerical optimization is not constrained through distances to the reference stimulus, while distance constraints can induce *a priori* similarities; it can better function as an unconstrained global characterization of the “encoding landscape” (i.e. multivariate tuning landscape of a population) and estimate if the (reconstructed) stimuli that a deep network is invariant to are “selective” (i.e. visually similar to original stimuli) at the same time. Explanation power of optimal stimulus is the second best single representation measure. Following the efficient coding theory [56, 57], one may exclude the possibility of an optimal stimulus looking highly similar to the task-related stimuli. Nevertheless, in a high-dimensional stimulus space, how to optimally construct the tuning landscape (and place the optimal stimuli) remains unclear, and whether the observed phenomenon (as shown in Fig. 7), in which optimal stimuli of well performing networks in fact are highly similar to the averages of neighboring task-related stimuli, is essential to the optimal tuning landscape still needs to be further studied. Worth to note, similar phenomenon was also reported in previous studies on well performing networks [21, 24]. Such “linearities” around the task-related stimuli out of highly nonlinear networks may also explain why natural stimuli can be more efficient in estimating the tuning of neurons [58].

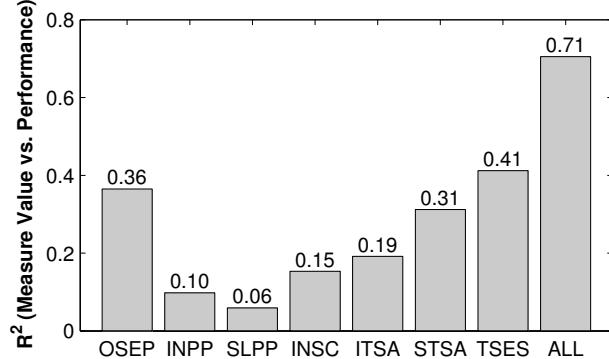


Figure 10. Spearman’s correlations between deep network’s representation measures and performance. Representation measures from left to right are, optimal stimulus explanation power (OSEP), invariance path potential (INPP), selectivity path potential (SLPP), invariance subspace capacity (INSC), invariance vs. task-related stimuli subspace alignment (ITSA), selectivity vs. task-related stimuli subspace alignment (STSA), task-related stimulus encoding specificity (TSES), and all measures together (ALL) using linear multiple correlation analysis. Pearson’s correlations following the same order are 0.37, 0.02, 0.02, 0.19, 0.17, 0.23, 0.39, and 0.69.

Discussion

Why Do We Need a New Framework?

As aforementioned, the study of sensory representations of biological neural networks has been an extremely important research topic for decades and numerous methods have been proposed; nevertheless, most methods can be viewed as implementation instances of functional frameworks of certain generalized concepts [17, 37]. We argue than existing frameworks being actively used are still mostly highly parametric, in terms of both the model assumption and stimulus generation. As moving away from peripheral or primary sensory circuitry and going deeper into higher-level sensory cortices, the modeling power of parametric models becomes inevitably insufficient, and even though parametric stimulus may facilitate the speed of representation characterization, its inherently assumptive nature can strongly bias the results. Although there also exist approaches which utilize the nonparametric artificial neural networks as models of sensory circuitry, only usage of shallow networks has been reported [59, 60], and interpreting the resultant model remains difficult and largely approximate [17]. In fact, this also directly links to the fact that attempts of characterizing/interpreting artificial neural networks themselves are also relatively difficult and scarce, and sometimes algorithms using handcrafted representations (e.g. HoG and SIFT [61]) are still preferred for their interpretability. Such problems only get worse when deeper networks are being adopted (either as system identification models of biological circuitry, or as algorithms directly applied in tasks), since understanding them also becomes even harder. Another important issue is the lack of principled and efficient approach for characterizing neuronal populations, due to the fact that, e.g., there is no straightforward generalization of the definition of optimal stimulus to neural populations, and existing methods can suffer from the exponential growth of required measurements, with respect to the population size [37].

Major benefits of the proposed framework can be categorized as follows. (1) Generality: as aforementioned, the proposed framework tackles the challenges in studying sensory representation in a highly generalizable and unbiased way via supporting both nonparametric models and nonparametric stimuli. Although not explicitly tested in this work, such setting in practice can be easily reconfigured and extended, if desired, to support both parametric models, by numerically analyzing them as nonparametric models

(especially when analytical solutions are not easily derivable), and parametric stimuli, by simply reframing the optimization in the parameter space $p \in \mathbb{R}^P$ and evaluating the fitness through the generative function $f(x(p))$ (especially when certain stimulus type or deformation is confirmed to be highly relevant). Also, the proposed framework supports characterizing both unit and population representations, as described in Methods; the fitness-distance diagram can be viewed as a “generalized tuning curve” characterization of the landscape, which is also dimensionality- and population size-insensitive. (2) Efficiency: the proposed framework is designed to be efficient using low-rank and stochastic approximations of the complex high-dimensional tuning landscape and its properties. As reported in multiple existing works [35, 36] which utilized second-order model for system identification and representation characterization, often only eigenvectors corresponding to few of the most positive and negative eigenvalues have significant and interpretable structures. Directly searching for invariance and selectivity paths eases this problem and saves potentially costly measurement requirement, which is on the order of $\mathcal{O}(N^2)$, to $\mathcal{O}(N)$, in addition to the fact that non-local characterization can be obtained in this work as well. Stochastic sampling of the solution space of the invariance subspace capacity provides an efficient approximation of the intrinsic dimension estimation approach [62, 63], and, of the encoding specificity, an efficient alternative of the maximally informative ensemble estimation approach [64], both of which can require substantially more measurements in order to yield robust and meaningful results. As aforementioned, the back-end solver with relatively simple constraints also positively contributes to the speed of this framework.

How Close Are We to Fully Understanding Deep Networks?

Via the proposed numerical optimization framework, the following important aspects of sensory representation research are enabled or facilitated: (1) Decoding unit or population representation (as optimal or reconstructed stimulus). (2) Discovering what changes in stimulus the representation is invariant or selective to. (3) Explaining how deep network’s representation may affect its task performances. Main findings of this work can be summarized as:

- Complexity of representation increases along network’s depth.
- Unlike deep representation which is both invariant and selective, shallow representation is only invariant, not selective, and its capacity of invariance is significantly lower than deep representation’s.
- How well the optimal stimulus can explain task-related stimuli, and how specific the representation encoding of task-related stimuli is, both decently explain network’s task performance.
- How well the invariance and selectivity of representation align with the actual “distribution” of task-related stimuli also partially explains network’s task performance.

We argue the importance of these findings as follows. First, the experiments are conducted in large scale and all results are statistically significant. Second, the results are obtained with relatively unbiased methods (i.e. nonparametric model plus nonparametric stimulus). Third, the representation measures match what are known to be crucial properties of performing accurate visual recognition, and explain network’s performance decently. Finally, the results also point out the practicability of the relatively less explored paradigm, where not only population representation can be examined, but extrinsic properties (i.e. measures based on task-related stimuli) also can be integrated to more effectively characterize various aspects of the deep network, in addition to only using intrinsic properties (i.e. measures based on optimal stimuli). In fact, if just using the four intrinsic properties (i.e. OSSC, INPP, SLPP and INSC as shown in Fig. 6; not included in Fig. 10) of the unit representations of the deep networks (with measures averaged from all 32 top-layer neurons within a network), we can only explain the network’s performance with $R^2 = 0.34$, which is significantly lower than the correlation based on extrinsic properties (i.e. $R^2 = 0.71$; combining both properties only marginally increase R^2 to 0.73). This also corresponds to our observation that the ranking of deep networks may differ from task to task (e.g. face pair matching in this work

vs. object recognition in [45]), and the fact that neural circuitry can (and maybe constantly) dynamically shape its tuning (i.e. representation) according to the tasks being performed [65] such that “intrinsic” properties may not mean much. Also, combining the results from machine learning and computer vision studies about deep networks’ advanced theoretical efficiency [66, 67] and practical performances over shallow networks’, we argue that network depth is possibly the most plausible way of implementing efficient sensory processing.

However, there are still undeniably plenty of unsolved and even unaddressed mysteries about deep networks. Here we list the ones we think to be of the most importance. First, even though we have successfully visualized the representations of certain deep neurons, as what is argued in Results about the optimal construction of tuning landscapes, how these representations explicitly facilitate generalization (i.e. invariance) across features that are distinct but of the same class, and discrimination (i.e. selectivity) between features that are similar but of different classes, remains ambiguous. Second, even though some smaller-scale studies using heuristic evaluation [20] or theoretical characterization [68] of representations’ evolution have been conducted, how the training algorithms (e.g. backpropagation) explicitly (and orientedly) shape the representations, and how different training settings alter the resultant representations, remain largely unclear. Third, the ability of “bottom-up attention” seems to be prominently functioning in these feedforward deep networks [20, 21] for unaligned and mildly cluttered visual object recognition; even though pooling operations are generally considered capable of providing such function, to what degree it (or other operations) contributes to this, how it explicitly (or implicitly) resolves, e.g., objects’ labels from cluttered scenes, and how much variation can be tolerated, remain unclear. Finally, some “higher-level vision” tasks also have been reported working decently using deep networks (e.g. image style recognition [69] and scene understanding [70]), in which what might the representations being used are can be even more ambiguous and undefined, and more “holistic”, instead of part-based, representations, are likely to dominate. As argued in [71], object recognition alone is a limited operational definition of high-level vision, and may not even be the right framing problem for studying high-level vision and visual representations any more. Like [72], we start this subsection by asking how close we are to solving this puzzle, yet with all the remaining inexplicable properties of deep networks, maybe asking how far we are from solving this, and how to precisely address these mysteries, would be more important and fruitful questions to be asked.

Challenges and Possible Directions

In addition to the inexplicable properties of deep networks listed above, which still lack clearer definitions and directions, here in the following, with no specific order, we also give a list of more defined challenges that we will need to solve, and possible directions to specifically tackle the challenges. Arguably, sensory representations of the most interest are at neither the bottom layer nor the top layer, but intermediate layers, since the former is relatively well understood and usually consist of Gabor (or Gabor-like) filters, and the later is highly class selective, as directly suggested by the “behavioral” (i.e. classification) results and the visualizations shown in [50, 73]. However, for most advanced deep networks [74, 75] which can have around 20 layers of operations, even intermediate layers can be very computationally expensive to measure and may have way more complex tuning landscapes (compared to the two-level deep networks experimented in this work, which consist of up to 8 layers of operations). Instead of CPU-based deep network simulation adopted in this work, GPU- or FPGA-/ASIC-based simulation may be required to tackle the potential speed problem. The back-end numerical solver used in this work may not suffice the needs for even deeper networks either; nevertheless, the proposed framework itself is neutral to the selection of solvers, thus more advanced numerical optimization algorithms can always be adopted to improve the quality of numerical solutions. Similar challenges may be encountered when moving from purely feedforward architectures into more biologically plausible ones—networks with feedback pathways, or recurrent networks—for their widely known nature of complex dynamics. Though in this work, around 70% of the variance of performance differences on face pair matching can be explained with the proposed

representation measures, yet we are not clear how high this number can be driven, and if these measures can generalize well enough onto even deeper representations or radically different types of tasks (e.g. phoneme recognition). Therefore, the validation/search of existing/new representation measures shall still be an urgent direction for future research.

Another difficult yet potentially fruitful future challenge is to apply this framework to real biological deep networks. While results are exclusively shown for artificial networks, the methods presented in this work by nature require a small budget of neuronal measurements that they could potentially be applied to real neurons in real experiments, even when spiking variability is taken into account, given that at least two effective techniques—stochastic averaging and measurement reweighting—are built in the core of CMA-ES, the back-end solver, for robustness against measurement noises. In fact, we argue that the mathematical formulation of CMA-ES can be viewed as an extension/generalization of the spike-trigger covariance method, which is known to be working effectively for biological neurons [36]. Potential inefficiency of measurement arising from neural adaptation [76] can also benefit from, e.g., Particle Swarm CMA-ES [77] or other multipopulation/multimodal optimization methods, where evaluation of stimuli from different local searches can be interwoven to minimize neural adaption and maintain search speed. Overall, as aforementioned, since the proposed framework supports efficient nonparametric search and analysis, a biological neural network can be characterized *as is* to reduce the chances of obtaining premature or biased numerical results. On the other hand, based upon recently reported results of deep networks being able to produce representations similar to mammalian visual systems' [14, 15], which though may still seem incomplete, we argue that the possibility of deep networks being a good surrogate model for understanding biologically plausible ways of representation encoding, and even guiding the research direction of ventral visual pathway studies, should not be neglected. In addition to the tools proposed in this work, neurophysiological methods like patch clamping, lesioning, "gene" (i.e. hyperparameters) knockout/knockdown, etc. may be as well used to inspire the development of new tools for characterizing artificial deep networks. For instance, the effect of surround suppression is also observed in experimented artificial neurons in this work as visualized in Fig. 4 (i.e. flat gray peripheries in the optimal stimulus patterns; responses drop when manually place, e.g., dots in peripheries) and, in our preliminary tests, can be easily shutdown via "lesioning" the prenormalization operations (i.e. early-stage lateral inhibition) of the network, which is consistent with neurophysiological findings [78].

Finally, though via the proposed numerical framework, we have demonstrated how previously unknown or unconfirmed properties can be positively identified with high statistical significance, we know this does not entirely cover what will be needed to "fully" explain observed properties and ultimately everything about sensory representations within deep networks—a theoretical framework yet to be constructed/completed. Notable recent works on bottom-up construction of theoretical frameworks for feedforward convolutional deep networks include magic theory [79], scattering transform [80], convolutional kernel network [81], etc. Saxe et al. [19] and Szegedy et al. [16] also addressed particular phenomena observed in deep networks using top-down theoretical approaches, furthermore to the theoretical studies [66, 67] on deep network's properties mentioned above. While undoubtedly all different theoretical approaches should keep being actively researched and interactively inspired by each other, we argue that the proposed numerical framework can nicely function as an independent *diagnostic tool* for examining, verifying, and facilitating the construction of theories.

Acknowledgments

References

1. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends in cognitive sciences 11: 333–341.

2. DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? *Neuron* 73: 415–434.
3. Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* 36: 193–202.
4. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86: 2278–2324.
5. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nature neuroscience* 2: 1019–1025.
6. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. pp. 1097–1105.
7. Ghodrati M, Farzamhdi A, Rajaei K, Ebrahimpour R, Khaligh-Razavi SM (2014) Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in computational neuroscience* 8.
8. Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences* 104: 6424–6429.
9. Cireşan D, Meier U, Masci J, Schmidhuber J (2012) Multi-column deep neural network for traffic sign classification. *Neural Networks* 32: 333–338.
10. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. (2014) Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:14090575*.
11. Taigman Y, Yang M, Ranzato M, Wolf L (2013) Deepface: Closing the gap to human-level performance in face verification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1701–1708.
12. Sun Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. *arXiv preprint arXiv:14064773*.
13. Vig E, Dorr M, Cox D (2014) Large-scale optimization of hierarchical features for saliency prediction in natural images. In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE.
14. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* : 201403112.
15. Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, et al. (2014) Deep neural networks rival the representation of primate it cortex for core visual object recognition. *arXiv preprint arXiv:14063284*.
16. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, et al. (2013) Intriguing properties of neural networks. *arXiv preprint arXiv:13126199*.
17. Wu MCK, David SV, Gallant JL (2006) Complete functional characterization of sensory neurons by system identification. *Annu Rev Neurosci* 29: 477–505.
18. Berkes P, Wiskott L (2006) On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural computation* 18: 1868–1895.
19. Saxe A, Koh PW, Chen Z, Bhand M, Suresh B, et al. (2011) On random weights and unsupervised feature learning. In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 1089–1096.
20. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: *Computer Vision–ECCV 2014*, Springer. pp. 818–833.
21. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:13126034*.
22. Goodfellow I, Lee H, Le QV, Saxe A, Ng AY (2009) Measuring invariances in deep networks. In: *Advances in neural information processing systems*. pp. 646–654.
23. Le QV, Ngiam J, Chen Z, Chia D, Koh PW, et al. (2010) Tiled convolutional neural networks. In: *Advances in Neural Information Processing Systems*. pp. 1279–1287.

24. Le QV, Ranzato MA, Monga R, Devin M, Chen K, et al. (2012) Building high-level features using large scale unsupervised learning. In: International Conference on Machine Learning.
25. Erhan D, Courville A, Bengio Y (2010) Understanding representations learned in deep architectures. Technical report, Technical Report 1355, Université de Montréal/DIRO.
26. Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. *JOSA A* 4: 2379–2394.
27. Bleeck S, Patterson R, Winter I (2003) Using genetic algorithms to find the most effective stimulus for sensory neurons. *Journal of neuroscience methods* 125: 73–82.
28. Yamane Y, Carlson ET, Bowman KC, Wang Z, Connor CE (2008) A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience* 11: 1352–1360.
29. Ringach D, Shapley R (2004) Reverse correlation in neurophysiology. *Cognitive Science* 28: 147–166.
30. Hansen KA, David SV, Gallant JL (2004) Parametric reverse correlation reveals spatial linearity of retinotopic human V1 BOLD response. *Neuroimage* 23: 233–241.
31. Dotsch R, Todorov A (2012) Reverse correlating social face perception. *Social Psychological and Personality Science* 3: 562–571.
32. Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* 63: 902–915.
33. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, et al. (2011) Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* 21: 1641–1646.
34. Harth E, Tzanakou E (1974) Alopex: A stochastic method for determining visual receptive fields. *Vision Research* 14: 1475–1482.
35. Touryan J, Lau B, Dan Y (2002) Isolation of relevant visual features from random stimuli for cortical complex cells. *The Journal of neuroscience* 22: 10811–10818.
36. Rust NC, Schwartz O, Movshon JA, Simoncelli E (2004) Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey V1. *Neurocomputing* 58: 793–799.
37. DiMattina C, Zhang K (2013) Adaptive stimulus optimization for sensory systems neuroscience. *Frontiers in neural circuits* 7.
38. Albrecht DG, Hamilton DB (1982) Striate cortex of monkey and cat: contrast response function. *J Neurophysiol* 48: 217–237.
39. Cheng K, Hasegawa T, Saleem KS, Tanaka K (1994) Comparison of neuronal selectivity for stimulus speed, length, and contrast in the prestriate visual cortical areas V4 and MT of the macaque monkey. *Journal of Neurophysiology* 71: 2269–2280.
40. Müller CL (2010) Black-box landscapes: characterization, optimization, sampling, and application to geometric configuration problems. Ph.D. thesis.
41. Spall JC (2005) Introduction to stochastic search and optimization: estimation, simulation, and control, volume 65. John Wiley & Sons.
42. Hansen N, Ostermeier A (2001) Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation* 9: 159–195.
43. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on* 13: 600–612.
44. Jones T, Forrest S (1995) Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Proceedings of the 6th International Conference on Genetic Algorithms. pp. 184–192.
45. Pinto N, Doukhan D, DiCarlo JJ, Cox DD (2009) A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology* 5: e1000579.
46. Pinto N, Poilvert N, Stone Z. STHOR library. <https://github.com/nsf-ri-ubicv/sthor>.

47. Cox D, Pinto N (2011) Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. IEEE, pp. 8–15.
48. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.
49. Wolf L, Hassner T, Taigman Y (2011) Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33: 1978–1990.
50. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, et al. (2014) Decaf: A deep convolutional activation feature for generic visual recognition. In: Proceedings of The 31st International Conference on Machine Learning. pp. 647–655.
51. Hegdé J, Van Essen DC (2000) Selectivity for complex shapes in primate visual area V2. *J Neurosci* 20: 61–66.
52. Pasupathy A, Connor CE (2001) Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of neurophysiology* 86: 2505–2519.
53. Desimone R (1991) Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience* 3: 1–8.
54. Ito M, Tamura H, Fujita I, Tanaka K (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology* .
55. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435: 1102–1107.
56. Barlow HB (1961) Possible principles underlying the transformation of sensory messages. *Sensory communication* : 217–234.
57. Laughlin SB, et al. (1981) A simple coding procedure enhances a neurons information capacity. *Z Naturforsch* 36: 51.
58. Talebi V, Baker CL (2012) Natural versus synthetic stimuli for estimating receptive field models: a comparison of predictive robustness. *The Journal of Neuroscience* 32: 1560–1576.
59. Lau B, Stanley GB, Dan Y (2002) Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences* 99: 8974–8979.
60. Prenger R, Wu MCK, David SV, Gallant JL (2004) Nonlinear v1 responses to natural scenes revealed by neural network analysis. *Neural Networks* 17: 663–679.
61. Szeliski R (2010) Computer vision: algorithms and applications. Springer.
62. Kégl B (2002) Intrinsic dimension estimation using packing numbers. In: Advances in neural information processing systems. pp. 681–688.
63. Levina E, Bickel PJ (2004) Maximum likelihood estimation of intrinsic dimension. In: Advances in neural information processing systems. pp. 777–784.
64. Machens CK (2002) Adaptive sampling by information maximization. *Physical Review Letters* 88: 228104.
65. Gilbert CD, Sigman M (2007) Brain states: top-down influences in sensory processing. *Neuron* 54: 677–696.
66. Delalleau O, Bengio Y (2011) Shallow vs. deep sum-product networks. In: Advances in Neural Information Processing Systems. pp. 666–674.
67. Montúfar G, Pascanu R, Cho K, Bengio Y (2014) On the number of linear regions of deep neural networks. arXiv preprint arXiv:14021869 .
68. Saxe AM, McClelland JL, Ganguli S (2013) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:13126120 .
69. Karayev S, Hertzmann A, Winnemoeller H, Agarwala A, Darrell T (2013) Recognizing image style. arXiv preprint arXiv:13113715 .

70. Khosla A, An B, Lim JJ, Torralba A (2014) Looking beyond the visible scene. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Ohio, USA.
71. Cox DD (2014) Do we understand high-level vision? *Current opinion in neurobiology* 25: 187–193.
72. Olshausen BA, Field DJ (2005) How close are we to understanding V1? *Neural computation* 17: 1665–1699.
73. Azizpour H, Razavian AS, Sullivan J, Maki A, Carlsson S (2014) From generic to specific deep representations for visual recognition. *arXiv preprint arXiv:14065774* .
74. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, et al. (2014) Going deeper with convolutions. *arXiv preprint arXiv:14094842* .
75. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556* .
76. Kohn A (2007) Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of neurophysiology* 97: 3155–3164.
77. Muller C, Baumgartner B, Sbalzarini IF (2009) Particle swarm cma evolution strategy for the optimization of multi-funnel landscapes. In: Evolutionary Computation, 2009. CEC'09. IEEE Congress on. IEEE, pp. 2685–2692.
78. Cavanova JR, Bair W, Movshon JA (2002) Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology* 88: 2547–2556.
79. Anselmi F, Leibo JZ, Rosasco L, Mutch J, Tacchetti A, et al. (2013) Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint arXiv:13114158* .
80. Mallat S (2012) Group invariant scattering. *Communications on Pure and Applied Mathematics* 65: 1331–1398.
81. Mairal J, Koniusz P, Harchaoui Z, Schmid C (2014) Convolutional kernel networks. *CoRR abs/1406.3332*.

Supporting Information

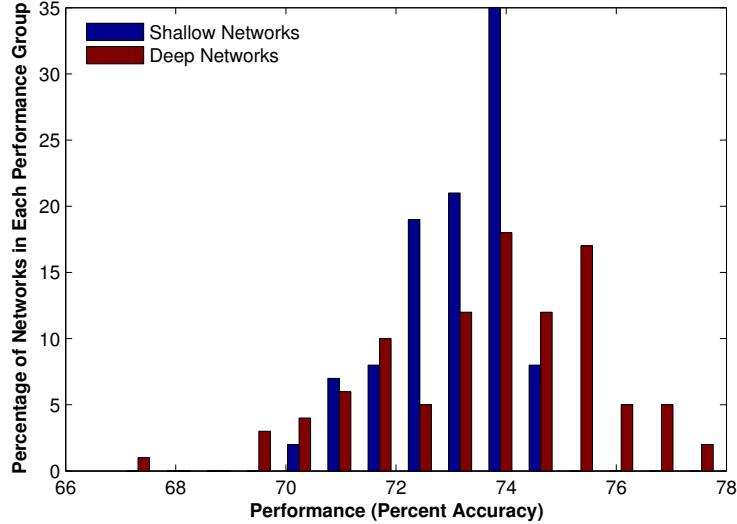


Figure S1. Distributions of shallow and deep network performances. Performances of the randomly generated 100 shallow and 100 deep networks are obtained via 10-fold cross validations of 6,000 face image pairs. Feature vectors used in the linear SVM classifier are generated via sliding the receptive fields throughout entire images and concatenating the population representations (as defined in convolutional neural networks). Maximum and mean of performances of the deep networks are both significantly higher than those of the shallow networks (with permutation tests, $p < 0.001$ and $p = 0.008$ respectively). See [47] for more details.

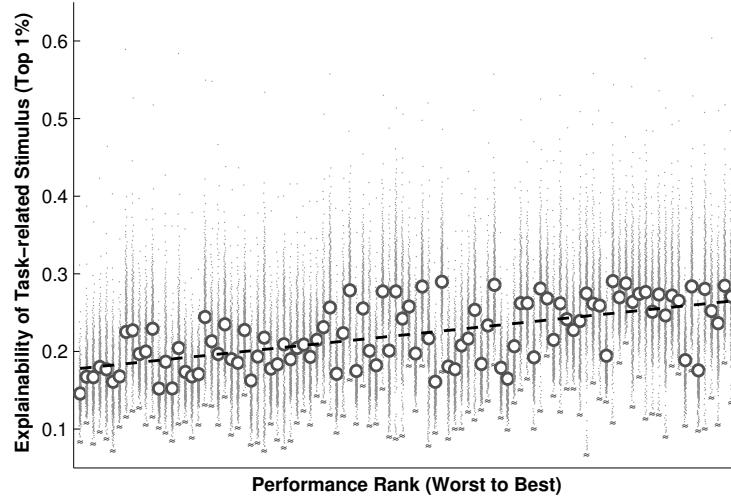


Figure S2. Explanation power vs. performances of deep networks. Each distribution from left to right shows the top 1% explainabilities of the task-related stimuli within each of the 32 top-layer unit representations of a deep network. Performance ranks, instead of performance values, are used for visualization purposes. Means of distributions are plotted as gray circles and the linear regression as dashed black line. Significance of slope using permutation test has $p < 0.001$.

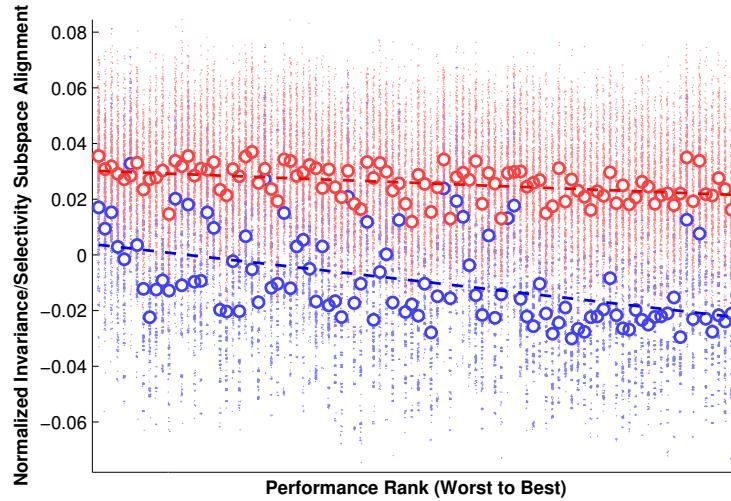


Figure S3. Invariance and selectivity subspace alignments vs. performances of deep networks. Each red distribution from left to right shows the invariance subspace alignments against the task-related stimuli of population representations of a deep network given the 16 reference stimuli. Alignment scores are normalized by subtracting the corresponding reference stimuli's own alignment scores. Performance ranks, instead of performance values, are used for visualization purposes. Means of distributions are plotted as red circles and the linear regression as dashed red line. Blue distributions for selectivity subspace alignments follow the same definitions. Significances of slopes using permutation tests both have $p < 0.001$.

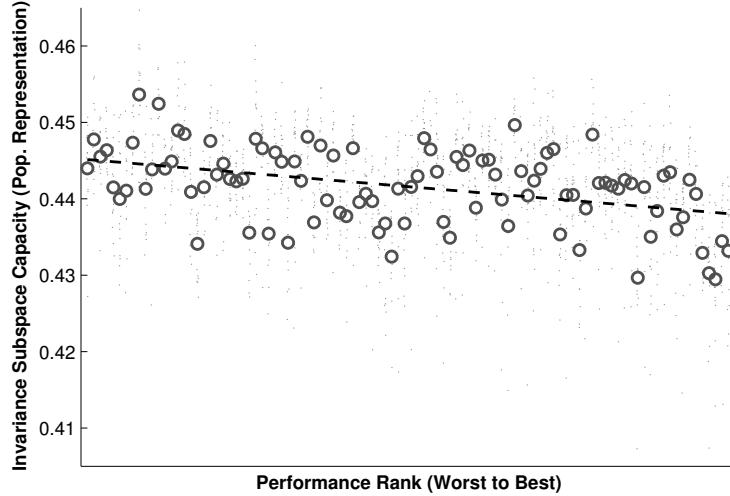


Figure S4. Invariance subspace capacities vs. performances of deep networks. Each distribution from left to right shows the invariance subspace capacities of population representations of a deep network given the 16 reference stimuli. Performance ranks, instead of performance values, are used for visualization purposes. Means of distributions are plotted as gray circles and the linear regression as dashed black line. Significance of slope using permutation test has $p < 0.001$.

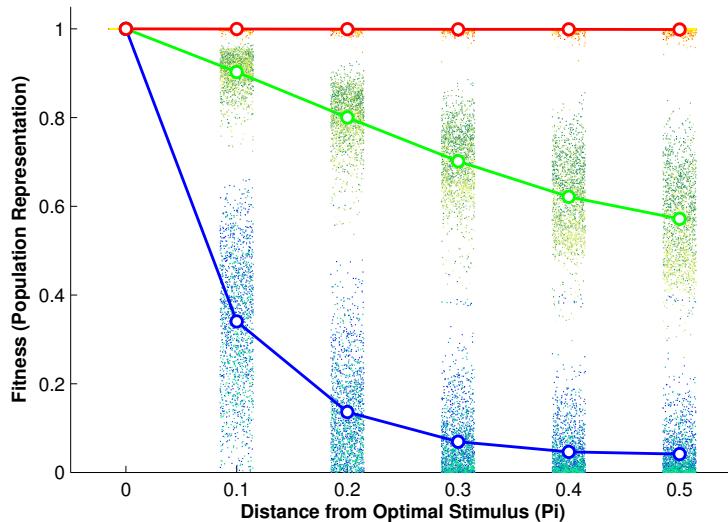


Figure S5. Fitness-distance diagram of population representations within deep networks. Red, blue, and green dots indicate results of invariance and selectivity path searches, and random walks, where brighter shades of a color are of better performing networks, and darker shades of poor performing networks. Means of results are plotted as solid lines in corresponding colors. As visualized, there is no apparent strong correlations between invariance and selectivity path potentials and performance. Random walk results vs. performance, on the other hand, has $R^2 = 0.18$ correlation which may arise from the differences in sensitivities to noises, though incorporating it into representation measures does not improve the final multiple correlation.

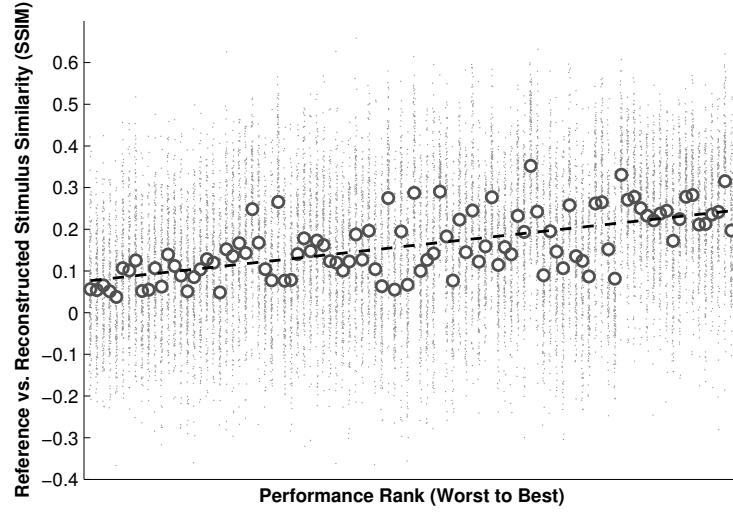


Figure S6. Encoding specificities vs. performances of deep networks. Each distribution from left to right shows the SSIM scores of the reconstructed stimuli of population representations of a deep network against the 16 reference stimuli. Performance ranks, instead of performance values, are used for visualization purposes. Means of distributions are plotted as gray circles and the linear regression as dashed black line. Significance of slope using permutation test has $p < 0.001$.