

Mean-field theory of input dimensionality reduction in unsupervised deep neural networks

Haiping Huang

RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan

(Dated: January 15, 2018)

Deep neural networks as powerful tools are widely used in various domains. However, the nature of computations at each layer of the deep networks is far from being well understood. Increasing the interpretability of deep neural networks is thus important. Here, we construct a mean-field framework to understand how compact representations are developed across layers, not only in deterministic random deep networks but also in generative deep networks where network parameters are learned from input data. Our theory shows that the deep computation implements a dimensionality reduction while maintaining a finite level of weak correlations between neurons for possible feature extraction. This work may pave the way for understanding how a sensory hierarchy works in general.

PACS numbers: 02.50.Tt, 87.19.L-, 75.10.Nr

Introduction.—The sensory cortex in the brain encodes the structure of the environment in an efficient way. This is achieved by creating progressively better representations of sensory inputs, and these representations finally become easily-decoded without any reward or supervision signals [1–3]. This kind of learning is called unsupervised learning, which has long been thought of as a fundamental function of the sensory cortex [4]. Based on the similar computational principle, many layers of artificial neural networks were designed to perform a non-linear dimensionality reduction of high dimensional data [5], which later triggered resurgence of deep neural networks. By stacking unsupervised modules on top of each other, one can produce a deep feature hierarchy, in which high-level features can be constructed from less abstract ones along the hierarchy. However, it remains rarely explored how this kind of effective representation is transformed along stages of processing. Understanding what each layer exactly computes may shed light on how sensory systems work in general.

Recent theoretical efforts focused on the layer-wise propagation of one input vector length, correlations between two inputs [6], and clustered noisy inputs [7]. Covariance of neural activity as one important feature of neural data modeling [8] is directly related to the dimensionality and complexity of hierarchical representations, yet unexplored in both deterministic and generative deep networks. Here, we propose a mean-field theory of input dimensionality reduction in deep neural networks. In this theory, we capture how a deep non-linear transformation reduces the dimensionality of a data representation, and moreover, how the representation complexity varies along the hierarchy. Both of these two features are fundamental properties of deep neural networks, and even information processing in vision [9]. Specifically, we first build a random multi-layered neural network, which implements deterministically a non-linear transformation of random Gaussian inputs with specified statistics. In

this random model, the intrinsic dimensionality of the intermediate representation at each layer is analytically computed, and the evolution of the associated covariance structure of neural activities across layers is described by our mean-field theory. Then, we apply the same framework to study a deep belief network (DBN) composed of multiple restricted Boltzmann machines (RBMs) stacked on top of each other [10]. In this probabilistic generative model, our theory captures the same computational principles of input dimensionality reduction, and furthermore statistical properties of neural covariance during layer-wise unsupervised learning. This mean-field framework thus provides a theoretical underpinning of input dimensionality reduction empirically observed in previous studies [1, 2, 5], offering a new perspective on how compact representations are transformed across layers in deep networks.

A deterministic deep network.—A deep network is a multi-layered neural network performing hierarchical non-linear transformations of sensory inputs (Fig. 1). The number of hidden layers is defined as the depth of the network, and the number of neurons at each layer is called the width of that layer. For simplicity, we assume an equal width (N). Weights between $l-1$ and l -th layers are specified by a matrix \mathbf{w}^l , in which the i -th row corresponds to incoming connections to the neuron i at the higher layer. Biases of neurons at the l -th layer are denoted by \mathbf{b}^l . The input data vector is denoted by \mathbf{v} , and \mathbf{h}^l ($l = 1, \dots, d$) denotes a hidden representation of the l -th layer, in which each entry h_i^l defines a non-linear transformation of its pre-activation $\tilde{a}_i^l \equiv [\mathbf{w}^l \mathbf{h}^{l-1}]_i$, as $h_i^l = \phi(\tilde{a}_i^l + b_i^l)$. Without loss of generality, we choose the non-linear transfer function as $\phi(x) = \tanh(x)$, and assume that the weight follows a normal distribution $\mathcal{N}(0, g/N)$, and the bias follows $\mathcal{N}(0, \sigma_b)$.

We consider a Gaussian input with zero mean, covariance $\langle v_i v_j \rangle = \rho r_{ij}$ for all $i \neq j$ (r_{ij} is a random variable taken uniformly from $[-1, 1]$), and variance $\langle v_i^2 \rangle = 1$. In

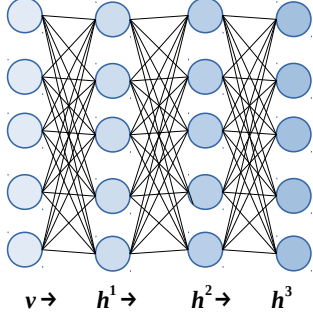


FIG. 1: (Color online) Schematic illustration of a deep neural network. The deep neural network performs a layer-by-layer non-linear transformation of the original input data (a high dimensional vector \mathbf{v}). During the transformation, a cascade of internal representations ($\mathbf{h}^1, \dots, \mathbf{h}^d$) are created. Here, $d = 3$ denotes the depth of the deep network.

the following derivations, we define the weighted-sum \tilde{a}_i^l subtracted by its mean as $a_i^l = \sum_j w_{ij}^l (h_j^{l-1} - \langle h_j^{l-1} \rangle)$, thus a_i^l has zero mean. As a result, the covariance of \mathbf{a}^l can be expressed as $\Delta_{ij}^l = \langle a_i^l a_j^l \rangle = [\mathbf{w}^l \mathbf{C}^{l-1} (\mathbf{w}^l)^T]_{ij}$, where \mathbf{C}^{l-1} defines the covariance matrix of neural activity at $l-1$ -th layer (also called connected correlation matrix in physics). Because the deep network defined in Fig. 1 is a fully-connected feedforward network, where each neuron at an intermediate layer receives a large number of inputs, the central limit theorem implies that the mean of hidden neural activity \mathbf{m}^l and connected correlation \mathbf{C}^l are given separately by

$$m_i^l = \langle h_i^l \rangle = \int Dt \phi \left(\sqrt{\Delta_{ii}^l} t + [\mathbf{w}^l \mathbf{m}^{l-1}]_i + b_i^l \right), \quad (1a)$$

$$C_{ij}^l = \int Dx Dy \phi \left(\Omega_1 x + \Xi y + b_i^l + [\mathbf{w}^l \mathbf{m}^{l-1}]_i \right) \times \phi \left(\Omega_2 x + \Xi y + b_j^l + [\mathbf{w}^l \mathbf{m}^{l-1}]_j \right) - m_i^l m_j^l, \quad (1b)$$

where $Dx = e^{-x^2/2} dx / \sqrt{2\pi}$, $\Omega_1 = \frac{\Delta_{ii} - \Delta_{jj}}{\sqrt{\Delta_{ii} + \Delta_{jj} - 2\Delta_{ij}}}$, $\Omega_2 = \frac{\Delta_{ij} - \Delta_{jj}}{\sqrt{\Delta_{ii} + \Delta_{jj} - 2\Delta_{ij}}}$, and $\Xi = \sqrt{\frac{\Delta_{ii} \Delta_{jj} - \Delta_{ij}^2}{\Delta_{ii} + \Delta_{jj} - 2\Delta_{ij}}}$. The superscript l is omitted for the covariance of \mathbf{a}^l in Ω_1 , Ω_2 and Ξ . Eq. (1) forms an iterative mean-field equation across layers to describe the transformation of the activity statistics in deep networks. The solution of this equation depends on the distribution of the network parameters (weights and biases).

To characterize the collective property of the entire hidden representation, we define an intrinsic dimensionality of the representation as $D = \frac{(\sum_{i=1}^N \lambda_i)^2}{\sum_{i=1}^N \lambda_i^2}$ [11], where $\{\lambda_i\}$ is the eigen-spectrum of the covariance matrix \mathbf{C}^l . It is easy to see that $D = N$ if each component of the representation is generated independently with the same variance. However, non-trivial correlations in the repre-

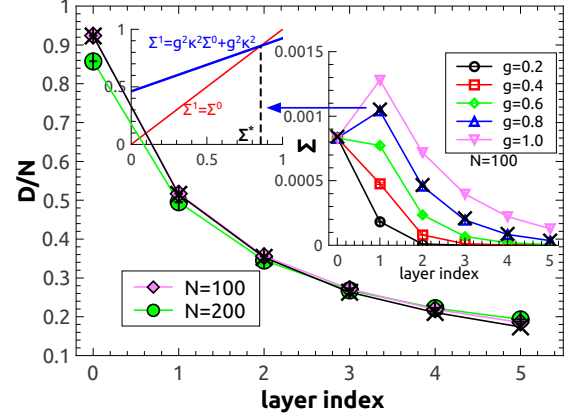


FIG. 2: (Color online) Representation dimensionality versus depth in deterministic deep networks. Ten network realizations are considered for each network width. $\rho = 0.05$, $g = 0.8$, and $\sigma_b = 0.1$. The right inset shows how the overall strength of connected correlations changes with depth and connection strength (g), and the left inset is a mechanism illustration ($\Sigma^{0,1}$ has been scaled by N). The crosses show simulation results ($g = 0.8$) obtained from 10^5 sampled configurations at each layer.

sensation will result in $D < N$ in general. The above mean-field equation addresses how this dimensionality changes along the depth.

Using the above mean-field theory, we first study the deterministic deep neural networks, where weights and biases are randomly chosen from respective normal distributions. Regardless of whichever network width used, we find that the representation dimensionality progressively decreases across layers (Fig. 2). The theoretical results agree very well with numerical simulations (indicated by crosses in Fig. 2). This shows that, even in a random multi-layered neural network, a more compact representation of the correlated input is gradually computed as the network becomes deeper. This compact representation may remove some irrelevant factors in the input, which facilitates formation of easily-decoded population-based representations at deeper layers. Although the network parameters are randomly constructed, the network depth and the non-linearity of transformations yield progressively compact representations of lower dimensionality, which is also one of basic properties in biological hierarchical computations [2, 12].

To get deeper insights about the representations, we study how the overall strength of connected correlations at each layer changes with the network depth and the connection strength (g). The overall strength of connected correlations is measured by $\Sigma = \frac{2}{N(N-1)} \sum_{i < j} C_{ij}^2$, which is related to the dimensionality by the equality $\Sigma = \frac{1}{N-1} \left[\left(\frac{1}{N} \sum_i C_{ii} \right)^2 / \tilde{D} - \frac{1}{N} \sum_i C_{ii}^2 \right]$ where $\tilde{D} = D/N$. We find that, to support an effec-

tive representation where neurons are not completely independent, the connection strength must be sufficiently strong (inset of Fig. 2), such that weakly-correlated neural activities are still maintained at later stages of processing. Otherwise, the information will be blocked from passing through that layer where the neural activity becomes completely independent. High correlations imply strong statistical dependence, and thus redundancy. An efficient representation must not be highly redundant [13], because a highly redundant representation can not be easily disentangled and is thus not useful for computation, e.g., co-adaptation of neural activities is harmful for feature extraction [14].

For a mechanistic explanation, assuming a large- N limit, we expand $C_{ij}^l = K_{ij}\Delta_{ij} + O(\Delta_{ij}^2)$ where $K_{ij} = \phi'(x_i^0)\phi'(x_j^0)$, and $x_{i,j}^0 = b_{i,j}^l + [\mathbf{w}^l \mathbf{m}^{l-1}]_{i,j}$. Then $\Sigma^l \simeq g^2 \kappa^2 \Sigma^{l-1} + \frac{g^2 \kappa^2}{N^2} \sum_i (C_{ii}^{l-1})^2$ where $\kappa = \int Dt \int Du (\phi'(\sqrt{\sigma_b}u + \sqrt{gQ^{l-1}}t))^2$, and $Q^l = \int Dt \int Du \phi^2(\sqrt{\sigma_b}u + \sqrt{gQ^{l-1}}t)$. For the random model, $N\Sigma^1 = g^2 \kappa^2 (N\Sigma^0 + 1)$ which determines a critical $N\Sigma^* = \frac{g^2 \kappa^2}{1 - g^2 \kappa^2}$, such that a first boost of the correlation strength is observed when $\Sigma^0 < \Sigma^*$ (Fig. 2); otherwise, decorrelation is achieved. Analogously, $\tilde{D}^1 = \frac{1}{(N-1)\Sigma^0 + 1 + \Upsilon}$ where $\Upsilon > 0$ and its specific value depends on the layer, and thus the dimensionality reduction is explained as $\tilde{D}^1 < \tilde{D}^0 = \frac{1}{(N-1)\Sigma^0 + 1}$, which also holds at deeper layers. Therefore, the deep deterministic random network can implement decorrelation of redundant inputs together with dimensionality reduction.

A stochastic deep network.—It is of practical interest whether a deep generative model trained in a completely unsupervised way has the similar collective behavior a random deterministic deep neural network has. We consider a DBN as a typical example of stochastic deep networks, in which at each hidden layer, each neuron's activity taking a binary value (± 1) depends on a stochastic function of its weighted-sum input shifted by a bias. Specifically, the DBN is composed of multiple RBMs stacked on top of each other (Fig. 1). RBM is a two-layered neural network, where there are no lateral connections within each layer, and the bottom (top) layer is also named the visible (hidden) layer. Therefore, given the input \mathbf{h}^l at l -th layer, the neural representation at a higher ($l+1$ -th) layer is determined by a conditional probability

$$P(\mathbf{h}^{l+1}|\mathbf{h}^l) = \prod_i \frac{e^{h_i^{l+1}([\mathbf{w}^{l+1}\mathbf{h}^l]_i + b_i^{l+1})}}{2 \cosh([\mathbf{w}^{l+1}\mathbf{h}^l]_i + b_i^{l+1})}. \quad (2)$$

Similarly, $P(\mathbf{h}^l|\mathbf{h}^{l+1})$ is also factorized.

The DBN learns a data distribution generated by a random RBM whose parameters follow the normal distribution $\mathcal{N}(0, g/N)$ for weights and $\mathcal{N}(0, \sigma_b)$ for biases. $g = 0.8$ and $\sigma_b = 0.1$ unless otherwise specified. Using the RBM as a data generator allows us to control

the complexity of the input data. In addition, characterizing the RBM representation may provide insights towards deep representations, since RBM is a building block for deep models and moreover a universal approximator of discrete distributions [15]. Given the RBM, the hidden neural activity at a higher layer (e.g., $\{h_f^{l+1}\}$) can be marginalized over using the conditional independence (Eq. (2)), thus the distribution of the representation at a lower layer (e.g., \mathbf{h}^l) can be expressed as

$$P(\mathbf{h}^l) = \frac{1}{Z_l} \prod_f \left[2 \cosh([\mathbf{w}^{l+1}\mathbf{h}^l]_f + b_f^{l+1}) \right] \prod_i e^{h_i^l b_i^l}, \quad (3)$$

where Z_l is the partition function intractable for large N . To study the statistics of the RBM representation, we need to compute the free energy function of Eq. (3) defined as $F = -\ln Z$, where l is dropped hereafter for simplicity, and unit inverse temperature is assumed. We use the Bethe approximation to compute an approximate free energy defined by F_{bethe} . In physics, the Bethe approximation assumes $P(\mathbf{h}) \approx \prod_f P_f(\mathbf{h}_{\partial f}) \prod_i P_i(h_i)^{1-N}$ [16], where f (∂f) indicates a factor node (its neighbors) representing the contribution of one hidden neuron to the joint probability (Eq. (3)) in a factor graph representation [17]. P_f and P_i can be obtained from a variational principle of free energy optimization [17]. This approximation takes into account the correlations induced by nearest neighbors of each neuron in the factor graph, which thus improves the naive mean-field approximation where neurons are assumed independent.

Covariance of neural activity under Eq. (3) can be computed from the approximate free energy using the linear response theory. However, due to the approximation, there exists a statistical inconsistency for diagonal terms computed under the Bethe approximation, i.e., $C_{ii} \neq 1 - m_i^2$. Therefore, we impose the statistical consistency of diagonal terms on a corrected free energy as $\tilde{F}_{\text{bethe}} = F_{\text{bethe}} - \frac{1}{2} \sum_i \Lambda_i (1 - m_i^2)$ [18–20]. Following the similar procedure in our previous work [17], we obtain the following mean-field iterative equation:

$$m_{i \rightarrow f} = \tanh \left(b_i - \Lambda_i m_i + \sum_{f' \in \partial i \setminus f} u_{f' \rightarrow i} \right), \quad (4a)$$

$$u_{f' \rightarrow i} = \frac{1}{2} \ln \frac{\cosh(b_{f'} + G_{f' \rightarrow i} + w_{f'i})}{\cosh(b_{f'} + G_{f' \rightarrow i} - w_{f'i})}, \quad (4b)$$

where $G_{f' \rightarrow i} \equiv \sum_{j \in \partial f' \setminus i} w_{f'j} m_{j \rightarrow f'}$, and the correction introduces an Onsager term ($-\Lambda_i m_i$). The cavity magnetization $m_{i \rightarrow f}$ can be understood as the message passing from visible node i to factor node f , while the cavity bias $u_{f' \rightarrow i}$ is interpreted as the message passing from factor node f' to visible node i . In fact, Eq. (4) is not closed. $\{\Lambda_i\}$ must be computed based on correlations. Therefore, we define a cavity susceptibility $\chi_{i \rightarrow f, k} \equiv \frac{\partial m_{i \rightarrow f}}{\partial b_k}$ [21]. According to this definition and the linear response the-

ory, we close Eq. (4) by obtaining the following susceptibility propagation equations:

$$\chi_{i \rightarrow f, k} = (1 - m_{i \rightarrow f}^2) \sum_{f' \in \partial i \setminus f} \Gamma_{f' \rightarrow i} \mathcal{P}_{f' \rightarrow i, k} \quad (5a)$$

$$+ \delta_{ik}(1 - m_{i \rightarrow f}^2) - \Lambda_i C_{ik},$$

$$C_{ik} = \frac{1 - m_i^2}{1 + (1 - m_i^2)\Lambda_i} \mathcal{F}_{ik}, \quad (5b)$$

$$\Lambda_i = \frac{\mathcal{F}_{ii} - 1}{1 - m_i^2}, \quad (5c)$$

where the full magnetization $m_i = \tanh(b_i - \Lambda_i m_i + \sum_{f' \in \partial i} u_{f' \rightarrow i})$, $\Gamma_{f \rightarrow i} \equiv \frac{\tanh(w_{fi})(1 - \tanh^2(b_f + G_{f \rightarrow i}))}{1 - \tanh^2(b_f + G_{f \rightarrow i}) \tanh^2(w_{fi})}$, $\mathcal{P}_{f \rightarrow i, k} \equiv \sum_{j \in \partial f \setminus i} \chi_{j \rightarrow f, k} w_{fj}$, and $\mathcal{F}_{ik} \equiv \sum_{f \in \partial i} \Gamma_{f \rightarrow i} \mathcal{P}_{f \rightarrow i, k} + \delta_{ik}$. It is easy to verify that Eq. (5) leads to the consistency for the diagonal terms. Adding the diagonal constraint through Lagrange multiplier Λ can not only solve the diagonal inconsistency problem but also improve the accuracy of estimating off-diagonals. After the RBM parameters (weights and biases) are specified, we run the above iterative equations (Eq. (4) and Eq. (5)) from a random initialization of the messages, and estimate the covariance and associated representation dimensionality from the fixed point.

Learning in a deep belief network can be achieved by layer-wise training of each RBM in a bottom-up pass, which was justified to improve a variational lower bound on the data log-likelihood [10]. More precisely, RBMs are trained in a feedforward fashion using the contrast divergence algorithm [10], where Gibbs samplings of the model starting from each data point are truncated to a few steps, and then used to compute model-dependent statistics for learning. The upper layer is trained with the lower layer's parameters being frozen. During the training of each RBM, the visible inputs are set to the mean activity of hidden neurons at the lower layer, while hidden neurons of the upper layer adopt stochastic binary values according to Eq. (2). With this layer-wise training, each layer learns a non-linear transformation of the data, and upper layers are conjectured to learn more abstract (complex) concepts, which is a key step in object and speech recognition problems [22].

Finally, we study the generative deep network where network parameters are learned in a bottom-up pass from the representations at lower layers. The network parameters for each stacked RBM in the DBN are updated by contrast divergence procedure truncated to one step [10]. One typical learning trajectory for each layer is shown in the top inset of Fig. 3 (a), where the reconstruction error decreases with the learning epoch. The input data complexity is captured very well by Eqs. (4) and (5) (Fig. 3 (b)). However, for a trained RBM, some components of the correlation matrix may lose the symmetry property

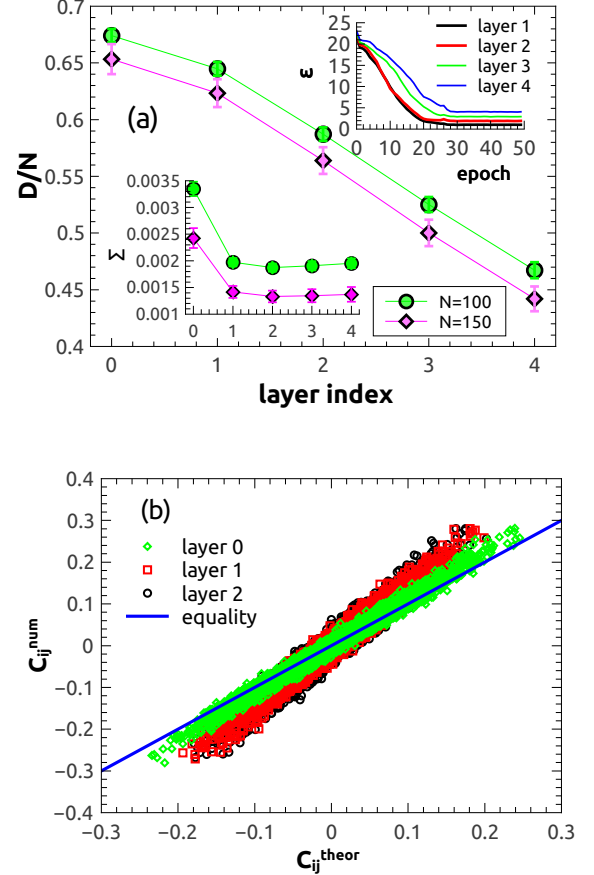


FIG. 3: (Color online) (a) Representation behavior as a function of depth in generative deep networks. Ten network realizations are considered for each network width. The top inset shows an example ($N = 150$) of reconstruction errors ε as a function of layers during learning. The bottom inset shows the overall strength of connected correlations as a function of depth. (b) Numerically estimated off-diagonal correlation versus its theoretical prediction ($N = 150$). In these plots, we generate $M = 60000$ training examples (each example is an N -dimensional vector) from the random RBM. Then these examples are learned by RBMs in a DBN. We divide the entire dataset into mini-batches of size $B = 150$. One epoch corresponds to a sweep of the entire dataset. Each RBM is trained for tens of epochs until the reconstruction error ($\varepsilon \equiv \|\mathbf{h}' - \mathbf{h}\|_2^2$) between input \mathbf{h} and reconstructed one \mathbf{h}' does not decrease. We use an initial learning rate of 0.12 divided by $\lceil t/10 \rceil$ at t -th epoch, and an ℓ_2 weight decay parameter of 0.0025.

($C_{ij} \neq C_{ji}$), likely because of the mean-field approximation incapable of dealing with an irregular distribution of learned connection weights. The irregularity means that the distribution is divided into two parts: the bulk part is around zero, while the other part is dominated by a few large values of weights (as also observed recently in spectral dynamics of learning in RBM [23]). Our mean-field formula offers a basis to be further improved to address

this interesting special property (although one can enforce the symmetry by $[C_{ij} + C_{ji}]/2$). Alternatively, we use the mean-field framework derived for deterministic networks to study the complexity propagation (starting from the statistics of the input data), which is reasonable, because to suppress the noise due to sampling, the mean activities at the intermediate layer are used as the input data when learning the next layer [10].

Compared to the initial input dimensionality, the representation dimensionality the successive layers create becomes lower (Fig. 3 (a)), which coincides with observations in the random deterministic deep network. This feature does not change when more neurons are used in each layer. Moreover, the learning decorrelates the correlated input, whereas, after the first drop, the learning seems to preserve a finite level of correlations (the bottom inset of Fig. 3 (a)). These properties can be explained qualitatively as in the random model.

Summary.—Brain computation can be thought of as a transformation of internal representations along different stages of a hierarchy [3, 12]. Deep artificial neural networks can also be interpreted as a way of creating progressively better representations of input sensory data. Our work provides a mean-field evidence about this picture that compact representations of relatively low dimensionality are progressively created by deep computation, while a small level of correlations is still maintained to make feature extraction possible, in accord with the redundancy reduction hypothesis [13]. In the deep computation, more abstract concepts captured at higher layers along the hierarchy are typically built upon less abstract ones at lower layers, and high level representations are generally invariant to local changes of the input [2], which thereby coincides with our theory that demonstrates a compact (compressed) representation formed by a series of dimensionality reduction. It was hypothesized that neuronal manifolds at lower layers are strongly entangled with each other, while at later stages, manifolds are flattened to facilitate that relevant information can be easily decoded by downstream areas [1, 2, 12], which connects to the small level of correlations preserved in the network for a representation that may be maximally disentangled [24]. Our work thus provides a theoretical underpinning of the hierarchical representations, which encourages further studies (e.g., how the compact representation shown in this study helps generalization (invariance) or discrimination (selectivity) [25, 26]) to interpret the nature of deep computation and its performance.

I thank Hai-Jun Zhou and Chang-Song Zhou for their insightful comments. This research was supported by AMED under Grant Number JP15km0908001.

- [2] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, *Neuron* **73**, 415 (2012).
- [3] N. Kriegeskorte and R. A. Kievit, *Trends in Cognitive Sciences* **17**, 401 (2013).
- [4] D. Marr, *Proceedings of the Royal Society of London B: Biological Sciences* **176**, 161 (1970).
- [5] G. E. Hinton and R. R. Salakhutdinov, *Science* **313**, 504 (2006).
- [6] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 3360–3368.
- [7] J. Kadmon and H. Sompolinsky, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016), pp. 4781–4789.
- [8] Elsayed Gamaleldin F and Cunningham John P, *Nature Neuroscience* **20**, 1310 (2017).
- [9] U. Guclu and M. A. J. van Gerven, *Journal of Neuroscience* **35**, 10005 (2015).
- [10] G. Hinton, S. Osindero, and Y. Teh, *Neural Computation* **18**, 1527 (2006).
- [11] K. Rajan, L. Abbott, and H. Sompolinsky, in *Advances in Neural Information Processing Systems 23*, edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Curran Associates, Inc., 2010), pp. 1975–1983.
- [12] Yamins Daniel L K and DiCarlo James J, *Nat Neurosci* **19**, 356 (2016).
- [13] H. Barlow, in *Sensory Communication*, edited by W. Rosenblith (Cambridge, Massachusetts: MIT Press, 1961), pp. 217–234.
- [14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Journal of Machine Learning Research* **15**, 1929 (2014).
- [15] N. Le Roux and Y. Bengio, *Neural Comput.* **20**, 1631 (2008).
- [16] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
- [17] H. Huang and T. Toyozumi, *Phys. Rev. E* **91**, 050101 (2015).
- [18] H. Huang and Y. Kabashima, *Phys. Rev. E* **87**, 062129 (2013).
- [19] J. Raymond and F. Ricci-Tersenghi, *Phys. Rev. E* **87**, 052111 (2013).
- [20] M. Yasuda and K. Tanaka, *Phys. Rev. E* **87**, 012134 (2013).
- [21] S. Higuchi and M. Mezard, *Journal of Physics: Conference Series* **233**, 012003 (2010).
- [22] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, *Commun. ACM* **54**, 95 (2011).
- [23] A. Decelle, G. Fissore, and C. Furtlehner, *EPL (Europhysics Letters)* **119**, 60001 (2017).
- [24] A. Achille and S. Soatto, arXiv: 1706.01350 (2017).
- [25] Hong Ha, Yamins Daniel L K, Majaj Najib J, and DiCarlo James J, *Nat Neurosci* **19**, 613 (2016).
- [26] Pagan Marino, Urban Luke S, Wohl Margot P, and Rust Nicole C, *Nat Neurosci* **16**, 1132 (2013).