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: October 5, 2017)

Deep neural networks as powerful tools are widely used in various domains. However, the nature of computations in each layer of the deep networks is far from being 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 paves the way for understanding how a sensory hierarchy works in general.

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. This kind of learning is called unsupervised learning, which has long been thought of as a fundamental function of the sensory cortex [1]. In particular, it was conjectured that the ventral visual pathway from primary visual cortex to inferotemporal cortex uses its hierarchical structure to gradually untangle neuronal manifolds at lower layers [2], such that the object identity these manifolds encode can be easily read out by a simple linear decoder at a high layer [3] [4]. A neuronal manifold can be thought of as a low-dimensional population code, whose geometry defines the nature of hierarchical sensory representations [5]. 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 [6], 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.

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 complexity of the representation varies along the hierarchy. Both of these two features are fundamental properties of deep neural networks, and even information processing in vision [7]. 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 in the intermediate representation across layers is described by our mean-field theory. Then, we apply a statistical mechanical approach to study a deep belief network (DBN) composed of multiple restricted Boltzmann machines (RBMs) stacked on top of each other [8]. In this probabilistic generative model, we reveal the same computational principles for input dimensionality reduction, and our theory captures statistical properties of covariance for intermediate representations during layer-wise unsupervised learning.

A deterministic deep network.—A deep network is a many-layered neural network performing hierarchical non-linear transformations of sensory inputs. As shown in Fig. 1 the number of layers is specified by $d+1$, with $d$ being the depth (the number of hidden layers) of the network. Weights between $l-1$-th and $l$-th layers are specified by a matrix $w^l$, in which the $i$-th row corresponds to incoming connections to the neuron $i$ at the higher layer. Biases of neurons at $l$-th layer are denoted by $b^l$. For simplicity, we assume each layer has an equal number (defined by $N$) of neurons being the width. The input data vector is denoted by $v$, and $h^l$ ($l = 1, \cdots , d$) denotes a hidden representation of $l$-th layer, in which each entry $h^l_i$ defines a non-linear transformation of its weighted-sum inputs $\hat{g}^l_i \equiv [w^l h^{l-1}]_i$, as $h^l_i = \phi(\hat{g}^l_i + b^l_i)$. In this random model, without loss of generality, we choose the non-linear transfer function as $\phi(x) = \tanh(x)$, and assume the weight follows a normal distribution with zero mean and variance $\frac{1}{2}$, and the bias follows the same distribution but with a size-independent variance $\sigma_b$.

We consider a Gaussian input with zero mean, and 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 the following derivations, we define the weighted-sum 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 $a_i^l$ can be expressed as $\Delta_{ij}^l = \langle a_i^l a_j^l \rangle = [w^l C^{l-1}(w^l)^T]_{ij}$, where $C^{l-1}$ defines the co-
where the hidden representation, we define an intrinsic dimensional vector \( v \). During the transformation, a cascade of internal representations \( \{h^1, \ldots, h^d\} \) are created. Here, \( d = 3 \) denotes the depth of the deep network.

variance 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 \( m^l \) and connected correlation \( C^l \) are given separately by

\[
m^l_i = \langle h^l_i \rangle = \int D\phi \left( \sqrt{\Delta^l_{ii}} t + [w^l m^{l-1}]_i + b^l_i \right), \tag{1a}
\]

\[
C^l_{ij} = \int Dx Dy \phi \left( \Omega_{l}x + \Xi y + b^l_i + [w^l m^{l-1}]_j \right) \times \phi \left( \Omega_{j}x + \Xi y + b^l_j + [w^l m^{l-1}]_j \right) - m^l_i m^l_j, \tag{1b}
\]

where \( D = e^{-x^2/2}dx/\sqrt{2\pi}, \Omega_{l} = \frac{\Delta_{ii} - \Delta_{jj}}{\sqrt{\Delta_{ii} + \Delta_{jj} - 2\Delta_{ij}}}, \Omega_{j} = \frac{\Delta_{jj} - \Delta_{ii}}{\sqrt{\Delta_{ii} + \Delta_{jj} - 2\Delta_{ij}}}, \Xi = \sqrt{\frac{\Delta_{ii} - \Delta_{jj}}{\Delta_{ii} + \Delta_{jj} - 2\Delta_{ij}}}. \) The superscript \( l \) is omitted for the covariance of \( a^l \) in \( \Omega_{l} \) and \( \Omega_{j} \) and \( \Xi \). 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 (weight matrix and biases). Although the network parameters are chosen from a random ensemble, the network still exhibits a non-trivial collective behavior (Results), likely thanks to the non-linearity of the transformation and the depth of the network. This kind of random ensemble was confirmed to be useful to capture the expressive power of deep networks [9]. In this work, we focus on the covariance structure of the internal representation created by each intermediate layer.

To characterize the collective property of the entire hidden representation, we define an intrinsic dimensionality of the representation as [10]

\[
D = \frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{N} \lambda_i^2}, \tag{2}
\]

where \( \{\lambda_i\} \) is the eigen-spectrum of the covariance matrix \( C^l \). 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 representation will result in \( D < N \) in general. In this work, we derive the mean-field equation to address how this dimensionality changes along the depth.

A stochastic deep network.—Although the deterministic deep neural network shows non-trivial properties even with random weights, we are also interested in whether a deep generative model trained in a completely unsupervised way has the similar collective behavior. 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 \( h^1 \) at \( l \)-th layer, the neural representation at a higher \((l + 1\)-th) layer is determined by a conditional probability

\[
P(h^{l+1} \mid h^l) = \prod_i \frac{e^{b^l_i + 1} (\lambda_i + b_i^{l+1})}{2 \cosh(\lambda_i + b_i^{l+1})}. \tag{3}
\]

In addition, \( P(h^l \mid h^{l+1}) \) can be similarly derived.

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 log-likelihood of input data [3]. More precisely, RBMs are trained in a feedforward fashion using the contrast divergence algorithm [8], where Gibbs samplings of the model starting from each point in the data space 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. 3. 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 [11].

Given a trained RBM, the hidden neural activity at a higher layer (e.g., \( h^{l+1} \)) can be marginalized over using the conditional independence (Eq. 3), thus the
distribution of the representation at a lower layer (e.g., \( \mathbf{h}' \)) can be expressed as

\[
P(\mathbf{h}') = \frac{1}{Z_l} \prod_f (2 \cosh [(\mathbf{w}^{l+1} \mathbf{h}')_f + b'_{f} + 1]) \prod_i e^{\mathbf{h}'_i b'_i},
\]

where \( Z_l \) is the partition function, which is intractable for an exact computation in large-size networks. To study the statistics of the internal representation created by each hidden layer, we need to compute the free energy function of Eq. (4) defined as \( F = -\ln Z_l \), 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 \( F_{\text{bethe}} \). In physics, the Bethe approximation assumes \( P(\mathbf{h}) \approx \prod_f P_f(\mathbf{h}_f) \prod_i P_i(h_i)^{1-N} \) \[12\], where \( f \) (\( \partial f \)) indicates a factor node (its neighbors) representing the contribution of one hidden node to the joint probability (Eq. (3)) \[13\]. \( P_f \) and \( P_i \) can be obtained from a variational principle of free energy optimization \[13\]. Note that this approximation takes into account the correlations induced by nearest neighbors of each neuron in a factor graph representation, which thus improves the naive mean-field approximation where variables are assumed independent. The Bethe approximation is exact on tree graphs and asymptotically correct for sparse graphs or graphs with sufficiently weak interactions.

Covariance of neural activity under Eq. (4) 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 \( \tilde{F}_{\text{bethe}} = F_{\text{bethe}} - \frac{1}{2} \sum_i \Lambda_i (1 - m_i^2) \) \[14\] \[15\]. Following the similar procedure in our previous work \[13\], we obtain the following mean-field iterative equations:

\[
\begin{align*}
\chi_{i \rightarrow f, k} &= \tanh \left( b_i - \Lambda_i m_i + \sum_{f' \in \partial_i \setminus f} u_{f' \rightarrow i} \right), \\
u_{f' \rightarrow i} &= \frac{1}{2} \ln \frac{\cosh (b_{f'} + G_{f' \rightarrow i} + w_{f'j})}{\cosh (b_{f'} + G_{f' \rightarrow i} - w_{f'j})},
\end{align*}
\]

where the factor node index \( f \) or \( f' \) is exactly the hidden neuron’s index, and \( G_{f' \rightarrow i} = \sum_{j \in \partial_{i,j} \neq f'} u_{f' \rightarrow f} \). 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. (5) is not closed. \( \{\Lambda_i\} \) must be computed based on correlations. Therefore, we define a cavity susceptibility

\[
\chi_{i \rightarrow f, k} = \frac{\partial \ln Z_l}{\partial b_{ik}^+} \] \[17\]. According to this definition and the linear response theory, we close Eq. (4) by obtaining the following susceptibility propagation equations:

\[
\begin{align*}
\chi_{i \rightarrow f, k} &= (1 - m_i^2 - f) \sum_{f' \in \partial_i \setminus f} \Gamma_{f' \rightarrow i} P_{f' \rightarrow i, k} \\
&\quad + \delta_{ik} (1 - m_i^2 - f) - \Lambda_i C_{ik}, \\
C_{ik} &= \frac{1 - m_i^2}{1 + (1 - m_i^2) F_{ik}}, \\
\Lambda_i &= \frac{F_{ii} - 1}{1 - m_i^2},
\end{align*}
\]

where the full magnetization \( m_i \) is defined as \( m_i = \tanh (b_i - \Lambda_i m_i + \sum_{f' \in \partial_i} u_{f' \rightarrow i}) \), \( \Gamma_{f' \rightarrow i} \equiv \tanh (w_{fj}) (1 - \tanh^2 (b_f + G_{f' \rightarrow i})) \) \[1\] \[10\]. It is easy to verify that Eq. (6) leads to the consistency for the diagonal terms. Adding the diagonal constraint through Lagrange multiplier \( \Lambda_i \) can not only solve the diagonal inconsistency problem but also improve the accuracy of estimating off-diagonals. After the parameters (weights and biases) for one RBM are learned, we run the above iterative equations (Eq. (5) and Eq. (6)) from a random initialization of the messages, and estimate the covariance from the fixed point (damping techniques may be used to avoid oscillation).

In the following, we generate \( M = 60000 \) training examples (each example is an \( N \)-dimensional vector) from a random RBM whose parameters follow the normal distribution \( N(0, g) \) for weights and \( N(0, \sigma_b) \) for biases. Here we fix \( g = 0.8 \) and \( \sigma_b = 0.1 \). Then these examples are learned by RBMs in a DBN. We divide the entire dataset into mini-batches of size \( B = 400 \). One epoch corresponds to a sweep of the full 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 \( [t/10] \) at \( t \)-th epoch, a damping coefficient of 0.2 or 0.02, and an \( \ell_2 \) weight decay parameter of 0.0025.

Results.—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 reduces across network depth (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. It was argued that deeper layers capture more abstract information, which is easily read out by simple decoders \[2\] \[3\]. 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 [18, 19].

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 $\Sigma = \frac{1}{N(N-1)} \sum_{i<j} C_{ij}^2$, which is related to eigenvalues of the correlation matrix by the equality $\text{Tr}(C^2) = \sum_{i=1}^{N} \lambda_i^2$. We find that to support an effective 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 indeed indicate strong statistical dependence, and thus redundancy. But an efficient representation must not be highly redundant [20], because a highly redundant representation can not be easily disentangled and is thus not useful for computation, e.g., co-adaptation of neural activities are harmful for feature extraction [21]. The deep deterministic random network can implement decorrelation of redundant inputs, although a first boost of the correlation strength is observed when $g \geq 0.63$.

Then, 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 [8].

One typical learning trajectory for each layer is shown in the inset of Fig. 3 (a), where the reconstruction error decreases with the learning epoch. Compared to the initial input dimensionality, the representation dimensionality the successive layers create becomes lower. Different from the random deterministic deep network, the trained generative deep network seems to maintain a compact representation whose dimensionality does not keep decreasing after the first drop. This feature does not change when more neurons are used in each layer. To obtain more insights about the internal representations, we measure the correlation strength as a function of the network depth in a trained deep generative network ($N = 150$). As shown in Fig. 3 (b), the learning increases the correlation strength at the second layer (this first boost is also observed in a random deep network (Fig. 2)), after which the learning seems to preserve that level of correlations, in spite of a slight increase observed. This level of corre-
lations may be important for the effective representation of a weakly correlated input. In general, it is impossible to fully represent the structure of high-dimensional raw data in two or three dimensional space, whereas, the deep network is able to create a compact representation preserving a certain level of correlations, explaining the non-trivial correlations in the input of higher dimensionality.

Discussion.—Brain computation can be thought of as a transformation of internal representations along different stages of a hierarchy \([5, 19]\). 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 a 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 \([15]\). 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 \([2, 18, 19]\). How the compact representation shown in this study of deep unsupervised networks helps generalization or discrimination remains an open challenging problem \([22, 23]\).

Furthermore, each layer learns a non-linear representation of the data from the layer below, maintaining a relatively small level of correlations, whose symmetric property can be broken (although a symmetric operation can be carried out as \(\frac{1}{2}(C_{ij} + C_{ji})\)), 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. A further study of this special property is required to have a deep understanding of the collective representations in deep networks.

In this work, we use the training dataset generated from some well-designed probabilities, rather than real dataset where we found the convergence of the susceptibility propagation is a main problem, which will inspire further studies of advanced mean-field theory to handle the irregularity in the learned weights from real dataset. Overall, our work paves the way towards understanding how hierarchical sensory systems work in general.

This work was supported by the program for Brain Mapping by Integrated Neurotechnologies for Disease Studies (Brain/MINDS) from Japan Agency for Medical Research and development, AMED.