Dimensionality compression and expansion in Deep Neural Networks

Stefano Recanatesi[*]
Center for Computational Neuroscience
University of Washington
Seattle, WA
stefanor@uw.edu

Matthew Farrell[*]
Center for Computational Neuroscience
University of Washington
Seattle, WA
msf9@uw.edu

Madhu Advani
Center for Brain Science
Harvard University
Cambridge, MA
madvani@fas.harvard.edu

Timothy Moore
Center for Computational Neuroscience
University of Washington
Seattle, WA
tjm36@uw.edu

Guillaume Lajoie
Dept. of Mathematics and Statistics
Université de Montréal
Montréal, Québec, Canada
lajoie@dms.umontreal.ca

Eric Shea-Brown
Center for Computational Neuroscience
University of Washington
Seattle, WA
etsb@uw.edu

Abstract

Datasets such as images, text, or movies are embedded in high-dimensional spaces. However, in important cases such as images of objects, the statistical structure in the data constrains samples to a manifold of dramatically lower dimensionality. Learning to identify and extract task-relevant variables from this embedded manifold is crucial when dealing with high-dimensional problems. We find that neural networks are often very effective at solving this task and investigate why. To this end, we apply state-of-the-art techniques for intrinsic dimensionality estimation to show that neural networks learn low-dimensional manifolds in two phases: first, dimensionality expansion driven by feature generation in initial layers, and second, dimensionality compression driven by the selection of task-relevant features in later layers. Our mathematical analysis shows how Stochastic Gradient Decent balances the dimensionality of neural representations by inducing an effective regularization term in the loss. We highlight the important relationship between low-dimensional compressed representations and generalization properties of the network. Our work contributes by shedding light on the success of deep neural networks in disentangling data in high-dimensional space while achieving good generalization. Furthermore, it invites new learning strategies focused on optimizing measurable geometric properties of learned representations, beginning with their intrinsic dimensionality.

1 Introduction

Deep neural networks are able to accurately classify high-dimensional data, not only achieving high training accuracy but also generalizing well to held-out samples. This is in spite of the myriad

[*]These authors contributed equally.

Preprint. Under review.

challenges associated with high-dimensional spaces, often referred to collectively as the *curse of dimensionality*. This is also in spite of deep networks typically existing in a highly over-parameterized regime where the number of parameters greatly exceeds the number of data samples. What is the reason for this unreasonable effectiveness? Here, we find new answers by probing networks with nonlinear metrics for dimensionality and developing theory that shows how deep networks naturally learn to compress the representation dimensionality of their inputs, sidestepping its apparent curse.

For concreteness we consider image classification datasets, but the observations and arguments we make are more general. While a dataset of images is naturally embedded in a high-dimensional space – the rgb space of 32x32-pixel images has dimension 3072 – the statistics of the dataset generally constrain the images to lie on a lower-dimensional structure, a nonlinear *manifold* [15]. What is the shape and dimension of this manifold, and how does learning influence these attributes? Recent developments in data science have yielded techniques for estimating the intrinsic dimensionality of manifolds which are robust to the high dimensionality of the embedding space as long as the manifold itself is low-dimensional [26, 42, 5, 4]. Here we deploy these state-of-the-art tools to analyze the dimensionality of image datasets (Fashion-MNIST and CIFAR-10) and of their deep manifold representations.

We train two deep neural network models to classify images from these datasets, and we use local and global metrics for dimensionality – the first time they have been applied to deep neural networks to the best of our knowledge – to analyze the geometry of the resulting manifold representations at each layer through the network architecture. Throughout training, each layer develops a specific representation in its high-dimensional neural space with properties determined by both task demands and by learning mechanisms. We find:

1. The dimensionality of representation manifolds is very low when compared to the number of neurons in each layer.
2. The dimensionality of representation manifolds evolves through trained networks in two distinct phases: initial layers expand dimensionality, and final layers compress it [24, 44, 46, 39].
3. Dimensionality expansion and compression are automatically balanced by SGD, and can be understood through an effective loss with two competing terms: one enforcing task demands on training data, and one compressing the manifold dimension.

Our results on the low dimensionality of learned deep representation manifolds helps explain why deep networks show good generalization properties despite using massive numbers of parameters: low-dimensional or otherwise minimal representations are thought to support good generalization [20, 49, 39]. We close the paper by discussing how probing and controlling dimensionality suggests new avenues to improve both AI and our understanding of neural coding strategies in brain circuits.

2 Methodology

Intrinsic dimensionality and its estimation Estimating the dimensionality of data manifolds is crucial for understanding how and why neural networks learn – but it is difficult, because linear analyses often fail to capture the effect of nonlinearities in embedding low-dimensional structures in high-dimensional spaces. By employing novel techniques [22, 16] we overcome the limits of previous analyses based on linear methods [50], uncovering new important phenomena. This builds on a rich literature on the estimation of intrinsic dimensionality of manifolds [23, 38, 12, 27, 42, 5, 4]. Here we provide a brief treatment of the essential concepts of intrinsic dimensionality estimation.

By means of example consider a simple sheet of paper. On a *local scale* a paper has three dimensions on which its molecules are organized, although on a more *global scale* we could say that it has only two dimensions on which we may draw or print. Importantly, independently of how it is folded or crumpled, these properties are persistent: locally it is three dimensional, globally it is two dimensional. To formalize these ideas we consider our sheet of paper folded to resemble a Swiss-roll [35], Fig. 1a, this is a curved manifold with local dimensionality of 3 and global dimensionality of 2 embedded in 3d. We remark that embedding the paper in a four or higher dimensional space would leave its local and global dimensionality unaffected; this point is very important in the following where we consider the local and global properties of manifolds embedded in high-dimensional spaces.

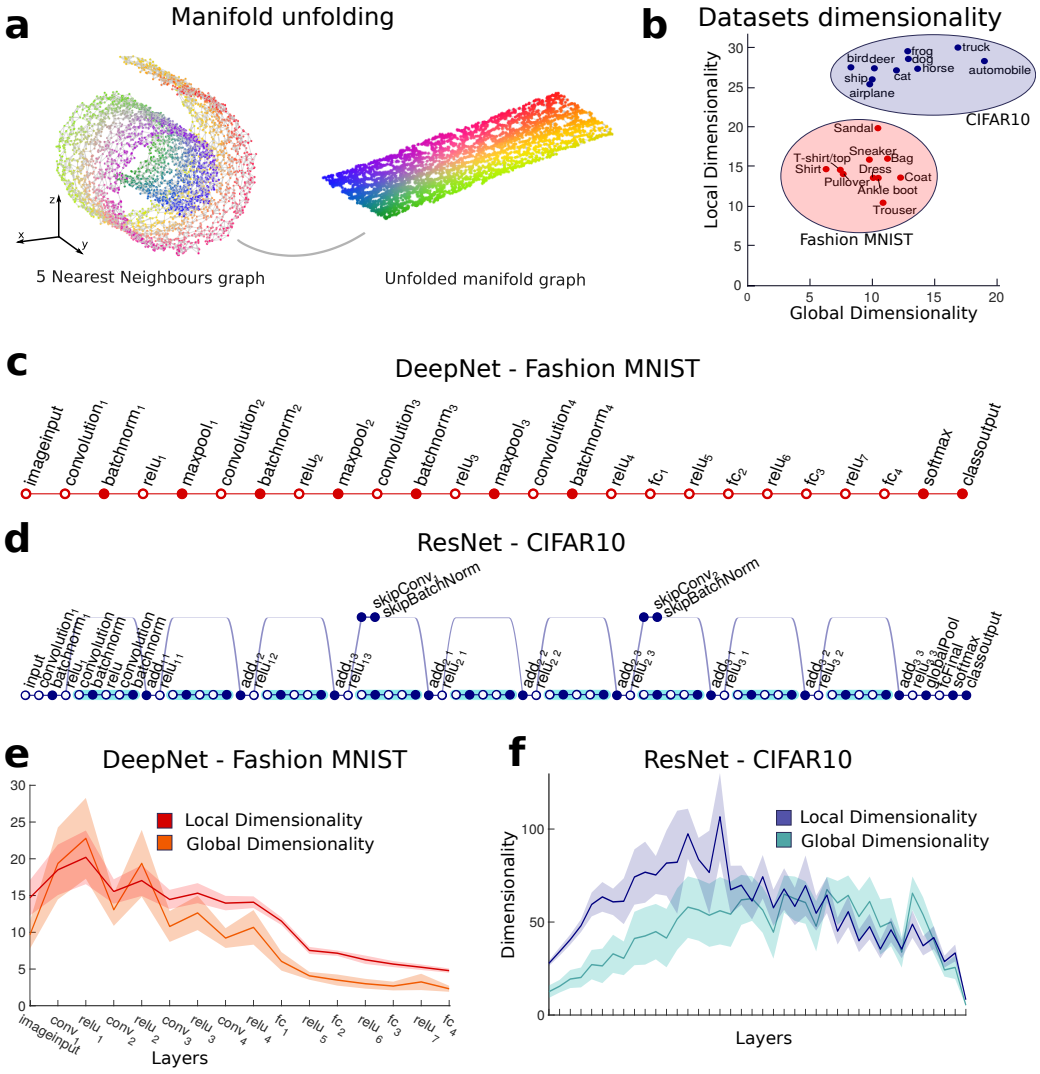

Figure 1: Dimensionality analysis of representation manifolds. a) Example of unfolding a Swiss-roll manifold. b) Local and Global dimensionality of data manifolds for Fashion-MNIST and CIFAR-10. c) Structure of DeepNet for classifying Fashion-MNIST. d) Structure of ResNet for classifying CIFAR-10. e) Local and global dimensionality of manifold representations through the layers of DeepNet and ResNet (f). Error bars indicate 95% confidence intervals across classes. The displayed layers are the ones colored white in panels (c) and (d).

In the example above the dimensionality changes as a function of the radius – this property is called multiscaling, a common intrinsic property of statistical manifolds [35, 4, 29]. For this reason we study two different measures of dimensionality, at a local scale ($r \to 0$) [16] and a global scale (for values of r around the mode value) [22]. In the local case the dimensionality is computed from the scaling of the probability distribution of nearest neighbor distances $\mathcal{P}_{NN}(r)$. A linear fit of the log probability of this distribution is proportional to the intrinsic dimensionality d. The 95% confidence interval on the fit is used to report uncertainty (cf. [16] for details). In the global case the k-nearest-neighbor graph ($k = 20$) is built and geodesic distances are computed. The resulting probability distribution of global distances $\mathcal{P}_G(r)$ is then analyzed around its mode value r_m. Specifically the portion of the distribution falling in between $r_m - r_\sigma$ and $r_m + \frac{1}{2}r_\sigma$, where r_σ is the standard deviation of the distribution, is compared to the same portion of the distribution of distances between points drawn from a hypersphere of varying dimensionality d. The value of d which minimizes the least square error between the two portions of distributions in the considered interval is the estimated global

intrinsic dimensionality (cf. [22] for details). These two methods have been chosen on a criterion of robustness and minimality.

Related intrinsic dimensionality estimation methods [11, 27, 6] yield consistent results with the metrics here selected whenever a robust convergence is achieved. In Fig. 1b we visualize the local and global dimensionality for the training set of the ten classes of CIFAR-10 and Fashion-MNIST, where the dimensionality of each class is measured individually. Different classes have slightly different dimensionalities but are overall consistent in their value of global and local dimensionality. This suggests that the methods we use are able to extract consistent information from datasets with similar statistics. Linear techniques based on singular value decomposition or principal component analysis are not able to provide such an accurate dimensionality estimation, largely overestimating the dataset dimensionality (data not shown). An alternative approach to measuring dimensionality of representation manifolds was developed in [8, 9] and recently applied to deep neural networks in [10]. This measure captures the arrangement of class manifolds in space from the perspective of a maximum margin linear classifier.

Deep network representation spaces Next we turn to the intrinsic dimensionality of the representations developed in feedforward neural networks. To assess the dimensionality of deep representations [3] we considered two benchmarks: a deep neural network trained to classify Fashion-MNIST and a ResNet [37] trained to classify both CIFAR-10 and CIFAR-100. The architectures of the two networks are reported respectively in Fig. 1c and Fig. 1d. The two networks were trained with SGD with a starting learning rate of 0.01 decreasing linearly by 0.0001 per epoch (fixed policy). Both networks were trained for 100 epochs and then the epoch where the validation accuracy was first minimized within a 0.1% accuracy was selected. The two networks achieved 90.96% and 87.6% testing accuracy, respectively. Importantly network architectures were chosen that keep the layer width constant as much as possible. DeepNet layer sizes decreased only via max-pooling layers. The 784 input variables passed through the architecture in Fig. 1c, where the total layer size decreases at each occurrence of a max-pooling or fully connected layer according to the sequence (25088, 6272, 1568, 288, 64, 10). Similarly in ResNet the initial 3072 variables followed the sequence of layer sizes (16384, 8192, 4096, 64, 10) throughout the network, decreasing only in the the case of Skip Convolution or fully connected layers. This helps disentangle the effect of layer sizes on representation dimensionality. Each layer induces a set of representation manifolds over the ensemble of inputs, one manifold corresponding to each class. To measure intrinsic dimensionality, we compute the dimensionality of the representation manifold for each class individually and average the results, reporting the 95% confidence intervals of the deviations.

3 Intrinsic dimensionality of learned representations

We computed the intrinsic dimensionality of deep representations for two deep neural networks, DeepNet and ResNet, trained on the Fashion-MNIST and CIFAR-10 datasets, respectively. Initial layers expanded the dimensionality of the input dataset while final ones carried out a dimensionality reduction (cf. Fig. 1e and Fig. 1f).

Figs. 1e and 1f indicate roles for specific layer types in increasing and decreasing dimensionality. In particular, ReLU nonlinearities consistently increased the dimensionality of their inputs by a factor that we measure to be 1.16 ± 0.19 across all network instances and classes (data not shown). Dimensionality compression was driven primarily by the application of the weight matrix before the ReLU nonlinearity was applied. Early convolutional layers tended to increase dimensionality. This highlights the tools that the network can use to create higher-dimensional as well as lower-dimensional feature representations.

Note that dimensionality expansion is related to feature generation [31, 2], akin to support vector machine kernel space expansion, and dimensionality reduction can be viewed in terms of feature selection [26, 39]. In the context of image classification, while earlier layers are often thought of as generating spatial features, final layers are thought to select and combine features that are relevant to classifying images according to their labels [46].

We emphasize that estimation properties make computing intrinsic dimension in high-dimensional spaces a serious challenge. A common approach in this setting is to consider multiple metrics. Here, both dimensionality metrics plotted above – and others not shown – are in agreement with the

trends we describe. This strengthens our confidence in identifying robust trends of the dimension of representation manifolds.

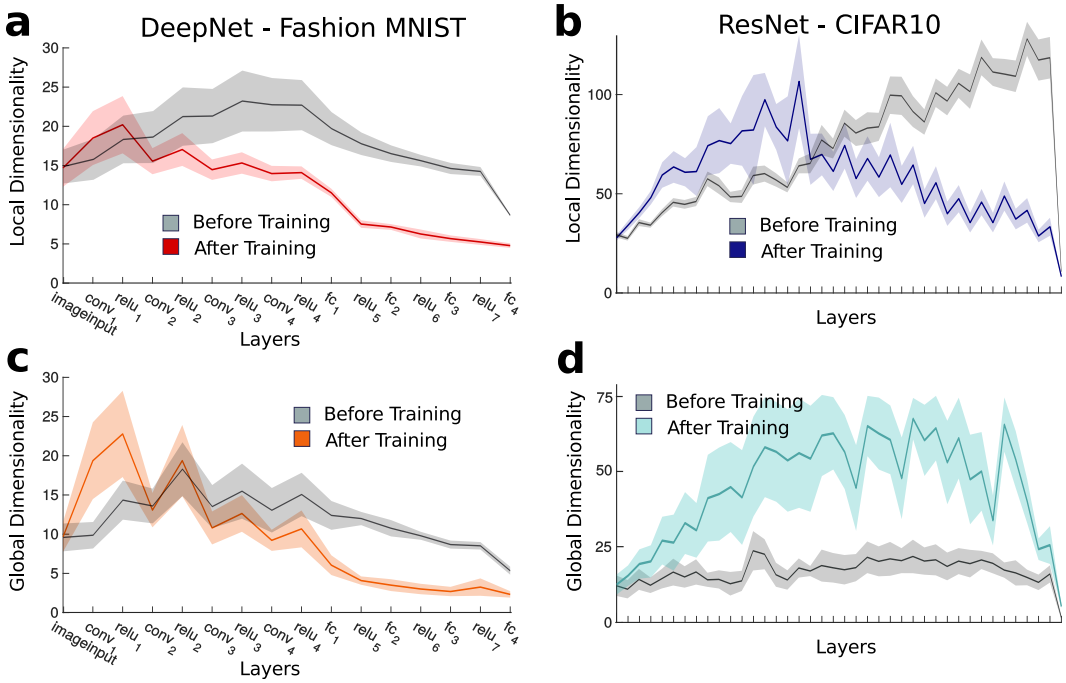

Figure 2: Dimensionality of representation manifolds before and after training. a) Local dimensionality of manifolds through the layers of DeepNet before and after training. Global dimensionality analysis as in (c). b) Local dimensionality for ResNet before and after training. Global dimensionality analysis as in (d). Error bars indicate 95% confidence intervals across class manifolds.

4 The role of learning in shaping the dimensionality of representations

How does training shape the dimensionality of network representations? To address this question we compare local and global dimensionality before and after training, for DeepNet (Fig. 2c) and ResNet (Figs. 2a and 2b).

Compared with the untrained network, training slightly increased local dimensionality in initial layers, and significantly decreased it in final ones (Figs. 2a to 2b). We note that before training, both DeepNet and ResNet showed the same layer-specific effects for local dimensionality: convolutional layers tended to expand local dimensionality while fully connected layers tended to decrease it (Fig. 2a and Fig. 2b). ResNet – a network that has convolutional layers throughout its depth – exhibited this phenomenon most clearly, with a nearly monotonic increase in dimension before training. For DeepNet, training had the same effects on global dimensionality as for local dimensionality (fig. 2c), increasing these dimensionalities in early layers and decreasing them in later ones. However, for the ResNet trained on CIFAR-10 (fig. 2d), learning increased the global dimensionality of all the layers, while the increasing-decreasing trend across layers was preserved after learning. We hypothesize that this occurs because ResNet primarily extracts local features before learning, expressing them globally only after learning. This is evidenced by the significant difference between local and global dimensionalities for the untrained networks (Fig. 2b vs. Fig. 2d). Training serves to express local information useful for solving the task globally; once that has occurred, local and global dimensions become nearly equal (Fig. 1f).

Overall the results shown here (Figs. 2a to 2d) are consistent with the interpretation that trained neural networks generate high-dimensional collections of features in early layers and select out a low-dimensional combination of these features in later layers. This suggests that the network is driven both by the need to expand and by the need to compress the dimensionality of its representation, and that the learned behavior of the network constitutes a balancing of these two demands. Which

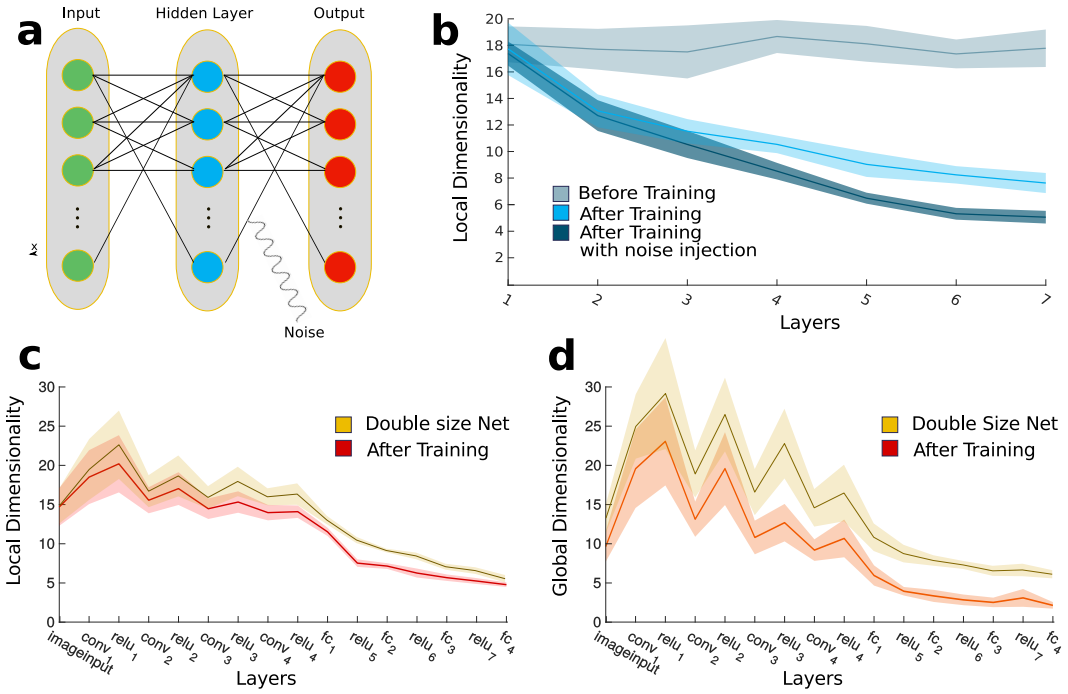

Figure 3: Theoretical analysis of SGD dimensionality compression. a) Schematics of a 2 layer networks with noise injected in the second layer. b). Dimensionality of deep network with constant layer size of 200 trained to classify Fashion-MNIST. The three lines display the dimensionality before and after training with and without the injection of noise in all the feedforward connections ($\sigma = 0.005$). Input weights are initialized to be random while all other weights are initialized to the identity matrix. c) Local dimensionality of DeepNet representation manifolds after training for the original network and the same network with intermediate layers doubled in size. Same for global dimensionality in panel (d). Error bars indicate two standard deviations.

mechanisms induce the network to strike this optimal balance? In the next section we show how SGD itself naturally produces these two complementary effects.

5 SGD balances task demands with dimensionality compression

We analyze a two-layer neural network trained to classify inputs according to c classes (see Fig. 3a). The equations for the network are:

$$\begin{aligned} h_i &= \sum_j \phi\left(w_{ij}^{(1)} x_j\right) \\ \hat{y}_i &= \sum_j w_{ij}^{(2)} h_j, \end{aligned} \tag{1}$$

where x, h and $\hat{y}$ are the input, the hidden representation, and the output of the network, respectively. Additionally, $w^{(1)}, w^{(2)}, \phi$ are respectively the input weights, readout weights, and a (possibly nonlinear) activation function. We consider the cost function $L(\Theta) = \sum_{\mu=1}^{P} \|y^\mu - \hat{y}^\mu\|_2^2$ where $\Theta = [w^{(1)}, w^{(2)}]$ and μ indexes the training set of size P. Here the output $\hat{y}^\mu$ is a length c vector, and the targets y^μ one-hot encode the labels.

Over training, SGD generates effective noise in the parameter updates (see e.g. [51, 36]) due to the fact that each update is performed on a subset of the training data. In general SGD leads to noisy gradient updates of the form:

$$\Delta\Theta = -\eta \nabla_\Theta L(\Theta) + z_t \tag{2}$$

where η is the learning rate and z_t is the noise generated from each mini-batch, or the difference between the gradient of the full batch and mini-batch t. This noise in the gradient updates is correlated.

Here we simplify the analysis by modeling this noise in parameter updates by adding noise of variance σ^2 directly to the output weights $w^{(2)}$, and by assuming this noise is isotropic Gaussian with zero mean. This is the simplest setting sufficient to provide intuition for the phenomena highlighted in this work: namely expansion and compression of the dimension of representations. Our analysis leads to the effective cost function:

$$L(\Theta, \sigma) = \sum_{\mu=1}^{P} \sum_{i} (y_i^\mu - \sum_j (w_{ij}^{(2)} + \sigma \xi_{ij}^\mu) h_j^\mu)^2, \tag{3}$$

where ξ is Gaussian noise with unit variance independently sampled across μ. Taking an average over ξ, for a learning rate small enough, we can rewrite Eq. 3 in the form:

$$L(\Theta, \sigma) \approx \langle L(\Theta, \sigma) \rangle_\xi = L(\Theta, 0) + R(\sigma) = L(\Theta, 0) + \sigma^2 \text{Tr}(C), \tag{4}$$

with $C_{jj'} = \sum_\mu h_j^\mu h_{j'}^\mu$ and where we view $R(\sigma) = \sigma^2 \text{Tr}(C)$ as a regularization term. Similar effective regularization terms have been shown to arise from dropout [45, 21], and the effects of regularization on generalization have been heavily studied [21]. Here we focus on the effects of such regularization in shaping the dimensionality of the representation. While the compressive effects of an initial step of SGD have been previously noted [17], here we consider the limit of many steps.

Geometrically, all the directions of the representation h^μ in the span of the readouts $w^{(2)}$ contribute to the cost both in $L(\Theta, 0)$ and $R(\sigma)$, while the directions orthogonal to this span contribute only to the regularization penalty $R(\sigma)$. By penalizing the norm of the representation $\|h^\mu\|_2^2$, the regularizer $R(\sigma)$ encourages the reduction of all h^μ components, including the orthogonal components $h_\perp^\mu$ of the representation with respect to the readout weights $w^{(2)}$. We refer to this action as compression of task-irrelevant directions (directions orthogonal to the readout $w^{(2)}$). For $\sigma > 0$ there is indeed a unique solution $(h^\mu)^*$ with null orthogonal components $(h_\perp^\mu)^* = 0$ that minimizes the cost eq. 4: $(h^\mu)^* = ((w^{(2)})^T w^{(2)} + \sigma^2 I)^\dagger (w^{(2)})^T y^\mu$. Here $\dagger$ denotes the Moore-Penrose pseudo-inverse. The uniqueness is a consequence of the strict convexity of $R(\sigma)$ and the convexity of $L(\Theta, 0)$ as functions of h^μ when h^μ is unconstrained. The cost increase of straying from this solution is quadratic.

A network that learns the task balances minimizing the task cost $L(\Theta, 0)$ with encouraging the compression of task-irrelevant directions due to $R(\sigma)$. For instance, learning a task that is not linearly separable with large amounts of training data requires a higher-dimensional hidden representation, which may come at the expense of increasing $R(\sigma)$, since $R(\sigma)$ increases isotropically as h^μ diverges from $(h^\mu)^*$. In other words, balancing the two terms $L(\Theta, 0)$ and $R(\sigma)$ in the loss shapes the representation h^μ so that the manifold dimension of h^μ is expanded only when aiding the reduction of the task loss $L(\Theta, 0)$ by separating the classes (see [13] for formal connections between dimensionality expansion and class separation).

To provide further intuition regarding this balance, we consider the case of linear activations ($\phi(z) = z$). In this setting we can write a closed form expression for the first layer weights. When $\sigma > 0$ the cost Eq. 4 is strictly convex with respect to the weights, and the unique minimizer $\hat{w}^{(1)}$ is:

$$\hat{w}^{(1)} = ((w^{(2)})^T w^{(2)} + \sigma^2 I)^\dagger (w^{(2)})^T Y^T X (X^T X)^\dagger, \tag{5}$$

where X is a $P \times d$ matrix of input samples and Y is a $P \times c$ matrix of labels. Here d is the embedding space dimension for the inputs. This equation reveals that the range of $\hat{w}^{(1)}$ lies within the span of the output weights, which implies that $\hat{h}^\mu = \hat{w}^{(1)} x^\mu$ lies in the span of the output weights as well. Equivalently, $\hat{h}_\perp^\mu = 0$ for all μ. Strict convexity assures that for an appropriate learning rate scheme, SGD will converge to this solution. The linear network is nearly always able to achieve $(h^\mu)^*$ when $P \leq d$, but not necessarily for larger numbers of samples. However, in either setting it will still remove all task-irrelevant directions from the representation.

We can consider how well these results generalize to the commonly applied ReLU nonlinearity ($\phi(x) = \max\{x, 0\}$). In the limit as $\sigma \to 0$, the hidden activity minimizing the loss will have the form $h^* + h_\perp$, where $w^{(2)} h_\perp = 0$. The addition of $h_\perp$ will not impact the ability to fit the training data, but is required to satisfy the nonnegativity constraint imposed by the ReLU activation. Thus, the dimensionality of the representation is larger in this nonlinear setting, and as the regularization strength is increased it leads to a trade-off between reducing dimensionality and fitting the training data.

Our theory suggests that the effective regularizer $R(\sigma)$ induced by fluctuations in weight updates encourages the compression of task-irrelevant directions and that dimensionality expansion of the activity in the hidden layer can occur when task complexity requires it. Although these results have been developed in the context of a two-layer network we hypothesize that they apply more broadly to deep networks. In this case the noise in the "output" weights of each layer generated by SGD encourages compression of the hidden representation of each layer. To support this conjecture we track the dimensionality of the ten Fashion-MNIST classes through a 7 layer feedforward fully connected network with 200 units per layer and a ReLU nonlinearity in Fig. 3b. Upon training with SGD on a mean squared error loss, we see that the dimensionality reduction through layers is more pronounced when noise is injected into the weights during the training process (dark blue line). This is what the theory above predicts, as higher effective noise induces a larger regularization $R(\sigma)$ and in turn stronger compression and lower dimensionality. As also predicted, we found a similar trend for another manipulation that increases the effective noise, decreasing the batch size, with smaller batches inducing stronger compression (data not shown).

Finally, we note that the learning of a low-dimensional representation is thought to prevent overfitting and aid generalization. Formal bounds on generalization are typically written in terms of the complexity of the class of functions used to fit data or of the parameters learned by the network (such as in [28, 43]). To the best of our knowledge, formal theoretical links between dimensionality of deep neural network representations and generalization have not been explicitly established. However, it is intuitive that representation geometry [33, 34] is connected to these ideas. Lower-dimensional distributions require fewer samples be drawn before the structure of the distribution can be inferred [7, 20]. This means that weights trained to transform a low-dimensional representation will require fewer training samples before the true distribution is learned [18, 19], i.e. before the weights generalize. See [20, 32, 49, 39] for more in-depth discussions of this topic.

Our simulations and theoretical analysis suggest that the dimensionality of the representations in deep networks is driven by balancing training data task demands with compression of task-irrelevant directions, as opposed to being driven by the number of neurons in the layers. To provide evidence for this argument, we double the number of units in each layer of DeepNet before training and compare the dimensionality of its representations (Fig. 3d). The accuracy (which increased by 2.1%, data not shown) and the dimensionality of layer representations are only slightly affected, despite the major increase in model complexity. Similarly, training ResNet on CIFAR-100 in place of CIFAR-10 doesn't lead to a significant increase in the layers' manifold dimensionality (data not shown). This supports the hypothesis that the SGD learning rule discovers a minimal-dimensional manifold that solves the task, and does so in a way that is insensitive to network size – helping to make the generalization power of deep architectures robust to over-parametrization.

6 Conclusion

In this paper we deploy state-of-the-art nonlinear dimensionality estimation techniques to measure the intrinsic dimensionality of deep neural networks' activations during a classification task. Our results show that the representation manifolds are very low-dimensional when compared to the network architecture, on the order of ten dimensions compared to the thousands of neurons per layer. We identify two distinct phases of dimensionality expansion and compression through the network's layers. A natural interpretation is that the expansion phase generates features that aid in solving the task [2, 20], while the compression phase selects out the key task-relevant features from among those generated [46]. As an example, feature selection allows for the building of invariances to class-irrelevant transformations (such as rotations and translations). Both simulation and theoretical analysis of SGD demonstrate that these phases emerge as learning balances task demands with a tendency to compress the manifold dimension. See [10] for recent work that finds similar trends for a different measure of dimensionality.

Our advances in measuring and understanding trends in the dimensionality of representation manifolds have a number of applications. The minimal number of neurons needed in the bottleneck of an autoencoder to reproduce the dataset can be viewed as an alternative measure of intrinsic dimensionality of the dataset (by this measure MNIST is believed to has dimensionality $\sim$13) [26, 48, 4]. Here we move beyond considering the intrinsic dimensionality of datasets to study the dimension of deep representations of these datasets that networks use to classify them. Following the autoencoder example, we posit that our results may provide a foundation for future work to determine the most

efficient sizes of networks that learn classification tasks [25, 40, 41]. For instance, if the maximum dimensionality achieved by a network is 50 in a middle layer, we conjecture that this will inform the layer size of a deep neural network that solves the task with high performance, either via standard training procedures or those that add pruning or compression steps.

Furthermore, our analysis suggests that representations learned by deep architectures tend to be low-dimensional due to the intrinsic regularization properties of SGD. Analyzing terms in an approximate loss shows that SGD encourages the representation manifold to be as low-dimensional as possible without compromising task accuracy, in effect removing task-irrelevant dimensions. We note that [39] and preceding works advanced similar ideas in terms of task-irrelevant information. Overall, low-dimensional representations add perspective to the community's efforts to link deep neural networks to generalization, [20, 49, 39, 1, 47], and will likely be critical in furthering understanding of why deep networks generalize well [30]. Moreover, the dimensionality-oriented perspective outlined in this work opens new possibilities in explicitly regularizing networks to improve performance on unobserved data. Examples include explicitly encouraging dimensionality compression through addition of tailored noise during training, adding dimension regularizing norms explicitly, or encouraging dimensionality expansion through other means including the choice of architecture in early layers. In addition, the connections drawn between noise, dimensionality compression, and generalization provide hints for understanding the representations formed in biological learning circuits [20], which are themselves noisy and often low-dimensional [14].

Acknowledgements

This work was supported by NSF DMS Grants #1514743 and #1256082, an NSERC Discovery Grant (RGPIN-2018-04821), an FRQNT Young Investigator Startup Program (2019-NC-253251), and an FRQS Research Scholar Award, Junior 1 (LAJGU0401-253188), as well as the Boeing Endowment at University of Washington Applied Mathematics.

References

[1] Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv:1710.03667 [physics, q-bio, stat]*, October 2017. arXiv: 1710.03667.

[2] Baktash Babadi and Haim Sompolinsky. Sparseness and expansion in sensory representations. *Neuron*, 83(5):1213–1226, 2014.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[4] Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, January 2016.

[5] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework, 2015.

[6] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. DANCo: Dimensionality from Angle and Norm Concentration. *arXiv:1206.3881 [cs, stat]*, June 2012. arXiv: 1206.3881.

[7] Siu-Wing Cheng, Tamal K. Dey, and Edgar A. Ramos. Manifold reconstruction from point samples. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1018–1027. ACM, New York, 2005.

[8] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Linear readout of object manifolds. *Phys. Rev. E*, 93:060301, Jun 2016.

[9] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and Geometry of General Perceptual Manifolds. *Physical Review X*, 8(3):031003, July 2018.

[10] Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and Geometry of Object Manifolds in Deep Neural Networks. *bioRxiv*, page 644658, May 2019.

[11] J. A. Costa and A. O. Hero. Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. In *2004 12th European Signal Processing Conference*, pages 369–372, September 2004.

[12] Jose Costa and Alfred Hero. Manifold Learning with Geodesic Minimal Spanning Trees. *arXiv:cs/0307038*, July 2003. arXiv: cs/0307038.

[13] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, June 1965.

[14] John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, nov 2014.

[15] Marzieh Edraki and Guo-Jun Qi. Generalized loss-sensitive adversarial learning with manifold margins. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[16] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, September 2017.

[17] Matthew S Farrell, Stefano Recanatesi, Guillaume Lajoie, and Eric Shea-Brown. Dynamic compression and expansion in a classifying recurrent network. *bioRxiv*, 2019.

[18] Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan. Reconstruction and interpolation of manifolds i: The geometric whitney problem. *arXiv preprint arXiv:1508.00674*, 2015.

[19] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[20] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, April 2016.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[22] Daniele Granata and Vincenzo Carnevale. Accurate Estimation of the Intrinsic Dimension Using Graph Distances: Unraveling the Geometric Complexity of Datasets. *Scientific Reports*, 6:31377, August 2016.

[23] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1):189–208, October 1983.

[24] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[25] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[26] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313(5786):504–507, July 2006.

[27] Elizaveta Levina and Peter J. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, 2005.

[28] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-Rao Metric, Geometry, and Complexity of Neural Networks. November 2017.

[29] Anna V. Little, Mauro Maggioni, and Lorenzo Rosasco. Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature. *Applied and Computational Harmonic Analysis*, 43(3):504–567, November 2017.

[30] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and Generalization in Neural Networks: an Empirical Study. *arXiv:1802.08760 [cs, stat]*, February 2018. arXiv: 1802.08760.

[31] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, December 1997.

[32] Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[33] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

[34] Ankita Shukla, Shagun Uppal, Sarthak Bhagat, Saket Anand, and Pavan Turaga. Geometry of deep generative models for disentangled representations. *arXiv preprint arXiv:1902.06964*, 2019.

[35] Vin D Silva and Joshua B Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in neural information processing systems*, pages 721–728, 2003.

[36] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

[37] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[38] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[39] Naftali Tishby. The Information Bottleneck Theory of Deep Neural Networks. *Bulletin of the American Physical Society*, 2018.

[40] Ming Tu, Visar Berisha, Yu Cao, and Jae-sun Seo. Reducing the model order of deep neural networks using information theory. In *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 93–98. IEEE, 2016.

[41] Frederick Tung and Greg Mori. Deep neural network compression by in-parallel pruning-quantization. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[42] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.

[43] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[44] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.

[45] Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pages 351–359, 2013.

[46] Weiran Wang and Miguel Á. Carreira-Perpiñán. The role of dimensionality reduction in linear classification, 2014.

[47] Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

[48] Shujian Yu, Kristoffer Wickstrøm, Robert Jenssen, and Jose C Principe. Understanding convolutional neural networks with information theory: An initial exploration. *arXiv preprint arXiv:1804.06537*, 2018.

[49] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[50] Mengxiao Zhang, Wangquan Wu, Yanren Zhang, Kun He, Tao Yu, Huan Long, and John E. Hopcroft. The Local Dimension of Deep Manifold. *arXiv:1711.01573 [cs]*, November 2017. arXiv: 1711.01573.

[51] Yao Zhang, Andrew M. Saxe, Madhu S. Advani, and Alpha A. Lee. Energy-entropy competition and the effectiveness of stochastic gradient descent in machine learning. *Molecular Physics*, 116(21-22):3214–3223, November 2018.