Intrinsic dimension of data representations in deep neural networks

Alessio Ansuini
International School for Advanced Studies
alessioansuini@gmail.com

Alessandro Laio
International School for Advanced Studies
laio@sissa.it

Jakob H. Macke
Technical University of Munich
macke@tum.de

Davide Zoccolan
International School for Advanced Studies
zoccolan@sissa.it

Abstract

Deep neural networks progressively transform their inputs across multiple processing layers. What are the geometrical properties of the representations learned by these networks? Here we study the intrinsic dimensionality (ID) of data-representations, i.e. the minimal number of parameters needed to describe a representation. We find that, in a trained network, the ID is orders of magnitude smaller than the number of units in each layer. Across layers, the ID first increases and then progressively decreases in the final layers. Remarkably, the ID of the last hidden layer predicts classification accuracy on the test set. These results can neither be found by linear dimensionality estimates (e.g., with principal component analysis), nor in representations that had been artificially linearized. They are neither found in untrained networks, nor in networks that are trained on randomized labels. This suggests that neural networks that can generalize are those that transform the data into low-dimensional, but not necessarily flat manifolds.

1 Introduction

Deep neural networks (DNNs), including convolutional neural networks (CNNs) for image data, are among the most powerful tools for supervised data classification. In DNNs, inputs are sequentially processed across multiple layers, each performing a nonlinear transformation from a high-dimensional vector to another high-dimensional vector. Despite the empirical success and widespread use of DNNs, we still have an incomplete understanding about why and when they work so well – in particular, it is not clear yet why they are able to generalize well to unseen data, not withstanding their massive overparametrization (1). While progress has been made recently [e.g. (2; 3)], guidelines for selecting architectures and training procedures are still largely based on heuristics and domain knowledge.

A fundamental geometric property of a data representation in a neural network is its *intrinsic dimension* (ID), i.e., the minimal number of coordinates which are necessary to describe its points without significant information loss. It is widely appreciated that deep neural networks are over-parametrized, and that there is substantial redundancy amongst the weights and activations of deep nets – e.g., several studies in network compression have shown that many weights in deep neural networks can be pruned without significant loss in classification performance (4; 5). Linear estimates of the ID in DNNs have been computed theoretically and numerically in simplified models (6), and local estimates of the ID developed in (7) have been related to robustness properties of deep networks to adversarial attacks (8; 9), showing that a low local intrinsic dimension correlates positively with robustness. Local ID of object manifolds can also be estimated at several locations on the tangent

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

space, and was found to decrease along the last hidden layers of AlexNet (10; 11). Both linear and nonlinear dimensionality reduction techniques have been used extensively to visualize computations in deep networks (12; 13; 14). Estimates of the local ID (7) can assume very low values in the last hidden layers (15), and can be used to signal the onset of overfitting in the presence of noisy labels. Thus, local ID can be used to drive learning towards high generalization solutions, also in conditions of severe noise contamination, showing a connection between intrinsic dimension and generalization during training of specific models (15).

However, there has not been a direct and systematic characterization of how the intrinsic dimension of data manifolds varies across the layers of CNNs and how it relates to generalization across a wide variety of architectures. We here leverage TwoNN (16), a recently developed estimator for *global* ID that exploits the fact that nearest-neighbour statistics depend on the ID (17) (see Fig. 1 for an illustration). TwoNN can be applied even if the manifold containing the data is curved, topologically complex, and sampled non-uniformly. This procedure is not only accurate, but also computationally efficient. In a few seconds on a desktop PC it provides the estimate of the ID of a data set with $O(10^4)$ data, each with $O(10^5)$ coordinates (for example the activations in an intermediate layer of a CNN), thus making it possible to map out ID across multiple layers and networks. Using this estimator, we investigated the variation of the ID along the layers of a wide range of deep neural networks trained for image classification. Specifically, we addressed the following questions:

- How does the ID change along the layers of CNNs? Do CNNs compress representations into low-dimensional manifolds, or, conversely, seek to expand the dimensionality?
- How different is the actual ID from the 'linear' dimensionality of a network, i.e., the dimensionality of the linear subspace containing the data-manifold? A substantial mismatch would indicate that the underlying manifolds are curved rather than flat.
- How is the ID of a network related to its generalization performance? Can we find empirical signatures of generalization performance in the geometrical structure of the representations?

Our analyses show that data representations in CNNs are embedded in manifolds of low dimensionality, which is typically several orders of magnitude lower than the dimensionality of the embedding space (the number of units in a layer). In addition, we found that the variation of the ID along the hidden layers of CNNs follows a similar trend across different architectures – the early layers expand the dimensionality of the representations, followed by a monotonic decrease that brings the ID to reach low values in the final layers.

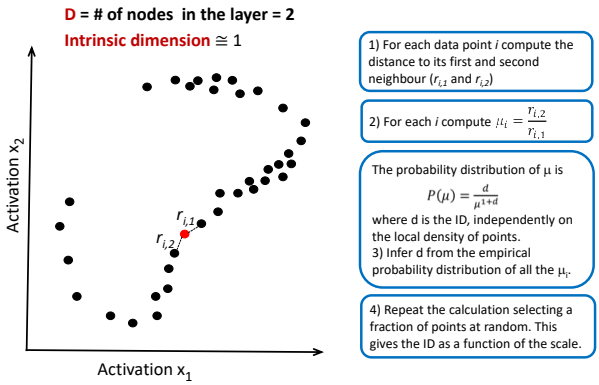

Figure 1: The TwoNN estimator derives an estimate of intrinsic dimensionality from the statistics of nearest-neighbour distances.

Moreover, we observed that, in networks trained to classify images, the ID of the training set in the last hidden layer is an accurate predictor of the network's classification accuracy on the test set – i.e, the lower the ID in this layer, the better the network capability of correctly classifying the image categories in a test set. Conversely, in the last hidden layer, the ID remains high for a network trained on non predictable data (i.e., with permuted labels), on which the network is forced to memorize rather than generalize. These geometrical properties of representations in trained neural networks were empirically conserved across multiple architectures, and might point to an operating principle of deep neural networks.

2 Estimating the intrinsic dimension of data representations

Inferring the intrinsic dimension of high-dimensional and sparsely sampled data representations is a challenging statistical problem. To estimate the ID of data-representations in deep networks, we leverage a recently developed global ID-estimator ('TwoNN') that is based on computing the ratio

between the distances to the second and first nearest neighbors (NN) of each data point (16) (see Fig. 1). This allows overcoming the problems related to the curvature of the embedding manifold and to the local variations in the density of the data points, under the weak assumption that the density is constant on the scale of the distance between each point and its second nearest neighbor.

Formally, let points x_i be uniformly sampled on a manifold with intrinsic dimension d and let N be the total number of points. Let $r_i^{(1)}$ and $r_i^{(2)}$ be the distances of the first and second neighbor of i respectively. Then $\mu_i \doteq r_i^{(2)}/r_i^{(1)}, i = 1, 2, ..., N$ follows a Pareto distribution with parameter $d + 1$ on $[1, +\infty)$, $f(\mu_i|d) = d\mu_i^{-(d+1)}$. Taking advantage of this observation, we can formulate the likelihood of vector $\boldsymbol{\mu} \doteq (\mu_1, \mu_2, ..., \mu_N)$ as

$$P(\boldsymbol{\mu}|d) = d^N \prod_{i=1}^{N} \mu_i^{-(d+1)}. \tag{1}$$

At this point d can be easily computed, for instance by maximizing the likelihood, or, following (16), by employing the empirical cumulate of the distribution of the μ values to reduce the ID estimation task to a linear regression problem. Indeed, the ID can also be estimated by restricting the product in eq. 1 to non-intersecting triplets of points, for which independence is strictly satisfied, but, as shown in ref. (16), in practice this does not significantly affect the estimate.

The ID estimated by this approach is asymptotically correct even for samples harvested from highly non-uniform probability distributions. For a finite number of data points, the estimated values remain very close to the ground truth ID, when this is smaller than ~ 20. For larger IDs and finite sample size, the approach moderately underestimates the correct value, especially if the density of data is non-uniform. Therefore, the values reported in the following figures, when larger ~ 20, should be considered as lower bounds.

For real-world data, the intrinsic dimension always depends on the scale of distances on which the analysis is performed. This implies that the reliability of the dimensionality estimate needs to be assessed by measuring the intrinsic dimension at different scales and by checking whether it is, at least approximately, scale invariant (16). In our analyses, this test was performed by systematically decimating the dataset, thus gradually reducing its size. The ID was then estimated on the reduced samples, in which the average distance between data points had become progressively larger. This allowed estimating the dependence of the ID on the scale. As explained in (16), if the ID is well-defined, its estimated value will only depend weakly on the number of data points N; in particular it will be not severely affected by the presence of "hubs", since the decimation procedure would kill them (see Fig. 2B).

To test the reliability of our ID estimator on embedding spaces with a dimension comparable to that found in the layers of a deep network, we performed tests on artificial data of known ID, embedded in a 100,000 dimensional space. The test did not reveal any significant degradation of the accuracy. Indeed, the ID estimator is sensitive only to the value of the distances between pair of points, and this distance does not depend on the embedding dimension.

For computational efficiency, we analyzed the representations of a subset of layers. We extracted representations at pooling layers after a convolution or a block of consecutive convolutions, and at fully connected layers. In the experiments with ResNets, we extracted the representations after each ResNet block (18) and the average pooling before the output. See A.1 for details.

The code to compute the ID estimates with the TwoNN method and to reproduce our experiments is available at this repository.

3 Results

3.1 The intrinsic dimension exhibits a characteristic shape across several networks

Our first goal was to empirically characterize the ID of data representations in different layers of deep neural networks. Given a layer l of a DNN, an individual data point (e.g., an image) is mapped onto the set of activations of all the n_l units of the layer, which define a point in a n_l-dimensional space. We refer to n_l as the embedding dimension (ED) of the representation in layer l. A set of N input

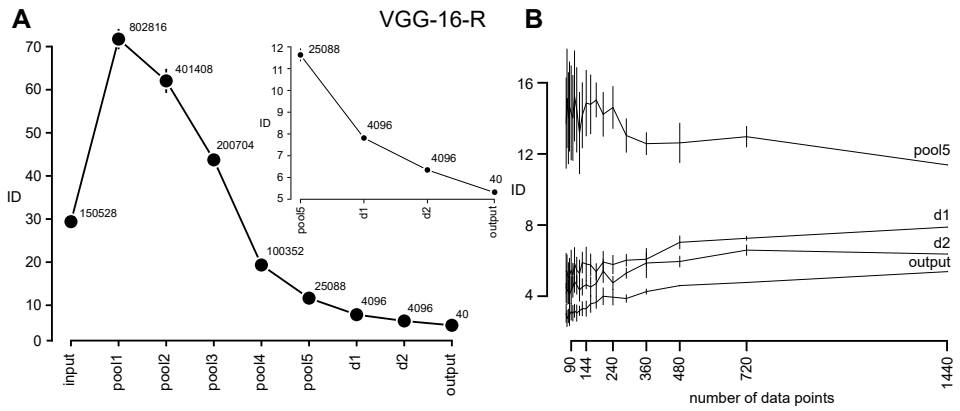

Figure 2: **Modulation of ID across hidden layers of deep convolutional networks A)** ID across layers of VGG-16-R, error bars are the standard deviation of the ID (see A.1). Numbers in plot indicate embedding dimensionality of each layer. **B** Subsampling analysis on VGG-16-R experiment, reported for the same layers as in the inset in **A** (see A.1 for details).

samples (e.g., N images) generate, in each layer l, a set of N n_l-dimensional points. We estimated the dimension of the manifold containing these points using TwoNN.

We first investigated the variation of the ID across the layers of a VGG-16 network (19), pre-trained on ImageNet (11), and fine-tuned and evaluated on a synthetic data-set of 1440 images (20). The dataset consisted of 40 3D objects, each rendered in 36 different views (we left out 6 images for each object as a test set) – it thus spanned a spectrum of different appearances, but of a small number of underlying geometrical objects. When estimating the ID of data representations on this network (referred to as 'VGG-16-R'), we found that the ID first increased in the first pooling layer, before successively and monotonically decreasing across the following layers, reaching very low values in the final hidden layers (Fig. 2A). For instance, in the fourth layer of pooling (pool4) of VGG-16-R, ID ≈ 19 and ED $\approx 10^5$, with $\frac{ID}{ED} \approx 2 \times 10^{-4}$.

One potential concern is whether the number of stimuli is sufficient for the ID-estimate to be robust. To investigate this, we repeated the analysis on subsamples randomly chosen on the data manifold, finding that the estimated IDs were indeed stable across a wide range of sample sizes (Fig. 2B). We note that, for the early/intermediate layers, the reported values of the ID are likely a lower bound to the real ID (see discussion in (16)).

Are the 'hunchback' shape of the ID variation across the layers (i.e., the initial steep increase followed by a gradual monotonic decrease), and the overall low values of the ID, specific to this particular network architecture and dataset? To investigate this question, we repeated these analyses on several standard architectures (AlexNet, VGG and ResNet) pre-trained on ImageNet (21). Specifically, we computed the average ID of the object manifolds corresponding to the 7 biggest ImageNet categories, using 500 images per category (see section A.1). We found both the hunchback-shape and the low IDs to be preserved across all networks (Fig. 3A): the ID initially grew, then reached a peak or a plateau and, finally, progressively decreased towards its final value. As shown in Fig. 8 for AlexNet, such profile of ID variation across layers was generally consistent across object classes.

The ID in the output layer was the smallest, often assuming a value of the order of ten. Such a low value is to be expected, given that the ID of the output layer of a network capable of recognizing N_c categories is bound by the condition $N_c \leq 2^{ID}$, which implies that each category is associated with a binary representation, and that the output layer optimally encodes this representation. For the ~ 1000 categories of ImageNet, this bound becomes $ID \gtrsim 10$, a value consistent with those observed in all the networks we considered.

Is the relative (rather than the absolute) depth of a layer indicative of the ID? To investigate this, we plotted ID against relative depth (defined as the absolute depth of the layer divided by the total number of layers, not counting batch normalization layers (12)) of the 14 models belonging to the three classes of networks (Fig. 3B). Remarkably, the ID profiles approximately collapsed onto a

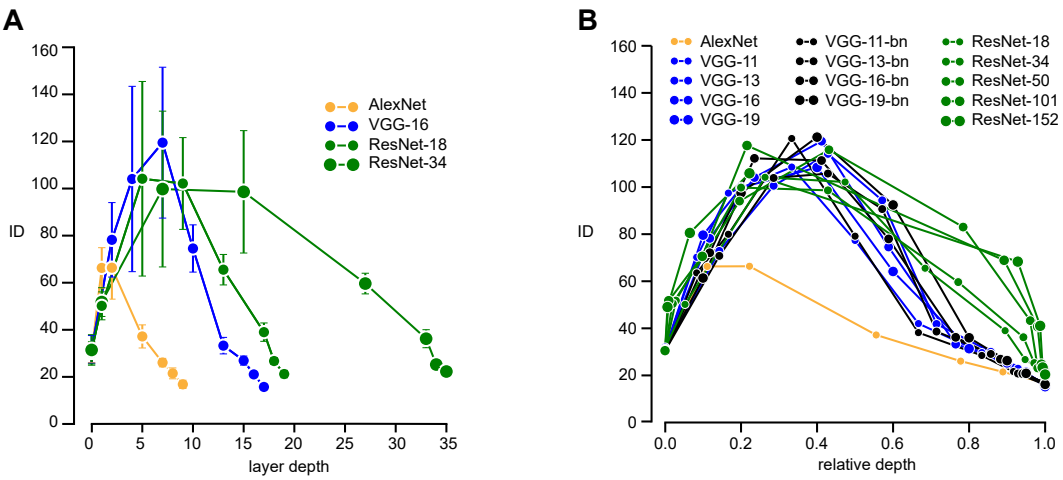

Figure 3: **ID of object manifolds across networks. A)** IDs of data representations for 4 networks: each point is the average of the IDs of 7 object manifolds. The error bars are the standard deviations of the ID across the single object's estimates (see A.1). **B)** The ID as a function of the relative depth in 14 deep convolutional networks spanning different sizes, architectures and training techniques. Despite the wide diversity of these models, the ID profile follows a typical hunchback shape (error bars not shown).

common hunchback shape [1], despite considerable variations in the architecture, number of layers, and optimization algorithms. For networks belonging to the VGG and ResNet families, the rising portions of the ID profiles substantially overlapped, with the ID reaching similar large peak values (between 100 and 120) in the relative depth range 0.2-0.4. The dependence on relative depth is consistent with the results of (12), where it was observed that similarity between layers depended on relative depth.

Notably, in all networks the ID eventually converged to small values in the last hidden layer. These results suggest that state-of-the-art deep neural networks – after an initial increase in ID – perform a progressive dimensionality reduction of the input feature vectors, a result with is consistent with the information-theoretical analysis in (22). Based on previous findings about the evolution of the ID in the last hidden layer during training (15), one could speculate that this progressive, gradual reduction of dimensionality of data-manifolds is a feature of deep neural networks which allows them to generalize well. In the following, we will investigate this idea further by showing that the ID of the last hidden layers predicts generalization performance, and by showing that these properties cannot be found in networks with random weights or trained on non predictable data.

3.2 The intrinsic dimension of the last hidden layer predicts classification performance

Although the hunchback shape was preserved across networks, the IDs in the last hidden layers were not exactly the same for all the networks. To better resolve such differences, we computed the ID in the last hidden layer of each network using a much larger pool of images of the training set ($\sim 2,000$), sampled from all ImageNet categories (see section A.1). This revealed a spread of ID values, ranging between ≈ 12 (for ResNet152) and ≈ 25 (for AlexNet, Fig. 4). These differences may appear small, compared to the much larger size of the embedding space in the last hidden layer (where the ED was between 1 and 2 orders of magnitude larger than the ID (range $= [512 - 4096]$). However, the ID in the last hidden layer on the training set was indeed a strong predictor of the performance of the network on the test set, as measured by top 5-score (Fig. 4, Pearson correlation coefficient $r = 0.94$). A tight correlation was found not only across the full set of networks, but also within each class of architectures, when such comparison was possible – i.e., in the classes of the VGG with and without batch normalization and ResNets ($r = 0.99$ in the latter case, see inset in Fig. 4).

[1] with the exception of AlexNet, and a small network trained on MNIST in a separate analysis, see section 3.4 for details and analysis

Overall, this analysis suggests that the ID in the last hidden layer can be used as a proxy for the generalization ability of a network. Importantly, this proxy can be measured without estimating the performance on an external validation set.

3.3 Data representations lie on curved manifolds

The strength of the TwoNN method lies in its ability to infer the ID of data representations, even if they lie on curved manifolds. This raises the question of whether our observations (low IDs, hunchback shapes, correlation with test-error) reflect the fact that data points live on low-dimensional, yet highly curved manifolds, or, simply, in low-dimensional, but largely flat (linear) subspaces.

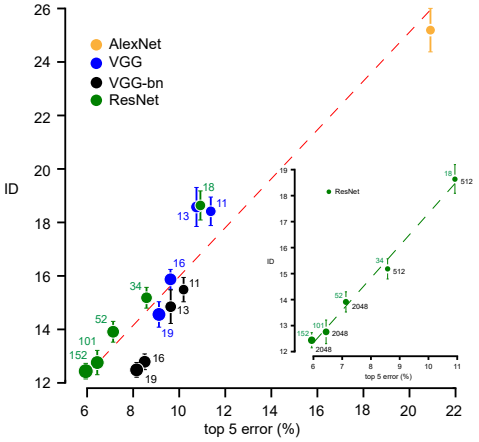

Figure 4: **ID of the last hidden layer predicts performance.** The ID of data representations (training set) predicts the top 5-score performance on the test set. **Inset** Detail for the ResNet class.

To test this, we performed linear dimensionality reduction (principal component analysis, PCA) on the normalized covariance matrix (i.e., the matrix of correlation coefficients – using the raw covariance resulted in qualitatively similar results) for each layer and network. We did not find a clear gap in the eigenvalue spectrum (Fig. 5A), a result that is qualitatively consistent with that obtained for stimulus-representations in primary visual cortex (23). The absence of a gap in the spectrum, with the magnitude of the eigenvalues smoothly decreasing as a function of their rank, is, by itself, an indication that the data manifolds are not linear. Nevertheless, we defined an 'ad-hoc' estimate of dimensionality by computing the number of components that should be included to describe 90% of the variance in the data. In what follows, we call this number PC-ID. We found PC-ID to be about one or two orders of magnitude larger than the value of the ID computed with TwoNN. For example, the PC-ID in the last hidden layer of VGG-16 was ≈ 200 (Fig. 5C, solid red line), while the ID estimated with TwoNN was ≈ 18 (solid black line).

The discrepancy between the ID estimated with TwoNN and with PCA points to the existence of strong non-linearities in the correlations between the data, which are not captured by the covariance matrix. To verify that this was indeed the case (and, e.g., not a consequence of estimation bias), we used TwoNN to compare the ID of the last hidden layer of VGG-16 with the ID of a synthetic Gaussian dataset with the same second-order moments. The ID of the original dataset was low and stable as a function of the size N of the data sample used to estimate it (Fig. 5B, black curve; similar subsampling analysis as previously shown in Fig. 2B). In contrast, the ID of the synthetic dataset was two orders of magnitude larger, and grew with N (Fig. 5B, red curve), as expected in the case of an ill-defined estimator (16).

We also computed the PC-ID of the object manifolds across the layers of VGG-16 on randomly initialized networks, and we found that its profile was qualitatively the same as in trained networks (compare solid and dashed red curves in Fig. 5C). By contrast, when the same comparison was performed on the ID (as computed using TwoNN), the trends obtained on random weights (dashed black curve) and after training the network (solid black curve) were very different. While the latter showed the hunchback profile (same as in Fig. 3), the former was remarkably flat. This behaviour can be explained by observing that the ID of the input is very low (see section 3.4 for a discussion of this point). For random weights, each layer effectively performs an orthogonal transformation, thus preserving such low ID across layers. Importantly, the hunchback profile observed for the ID in trained networks (Figs 2A, 3A,B) is a genuine result of training, which does not merely reflect the initial expansion of the ED from the input to the first hidden layers, as shown by the fact that, in VGG-16, the ID kept growing after that the ED had already started to substantially decline (compare the solid black and blue curves in Fig. 5C).

The analysis shown in Fig. 5C also indicates that intermediate layers and the last hidden layer undergo opposite trends, as the result of training (compare the solid and dashed black curves): while the ID of the last hidden layer is reduced with respect to its initial value [consistently with what reported in (15)], the ID of intermediate layers increases by a large amount. This prompted us to run an exploratory analysis to monitor the ID evolution during training in a VGG-16 network trained with CIFAR-10. We observed a behavior that was consistent with that already reported in Fig. 5C: a substantial increase of the ID in the initial/intermediate layers, and a decrease in the last layers (Fig. 9A, black vs. orange curve). Interestingly, a closer inspection of the dynamics in the last hidden layer revealed a non-monotonic variation of the ID (see Fig. 9B,C). Here, after an initial drop, the ID slowly increased, but, differently from (15), without resulting in substantial overfitting. Thus, the evolution of the ID during learning appears to be not strictly monotonic and its trend likely depends on the specific architecture and dataset, calling for further investigation.

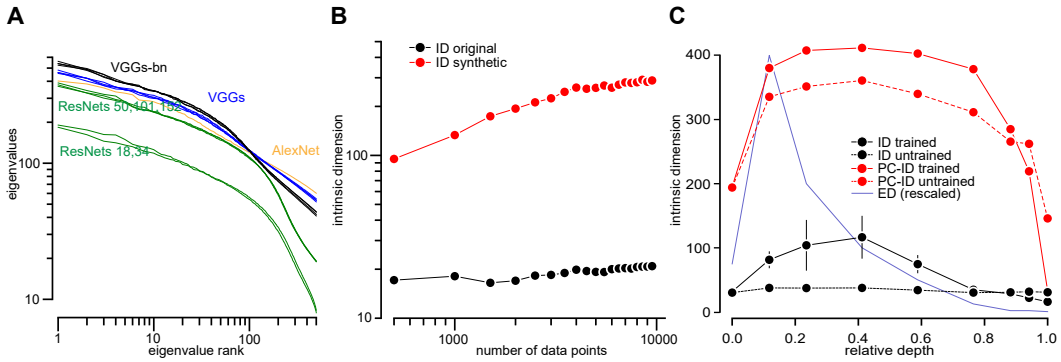

Figure 5: **Evidence that data-representations are on curved manifolds A)** Variance spectra of last hidden layer do not show a clear gap. **B)** ID in the last hidden layer of VGG-16 (black), compared with the ID of a synthetic Gaussian dataset with the same size and second-order correlations structure (red). **C)** The ID and the PC-ID along the layers of VGG-16 for a trained network and an untrained, randomly initialized network. The ED, rescaled to reach the maximum at 400, is shown in blue.

3.4 The initial increase in intrinsic dimension can arise from irrelevant features

We generally found the ID to increase in the initial layers. However, this was not observed for a small network trained on the MNIST data-set (Fig. 6B, black curve) and was also less pronounced for AlexNet (Fig. 3A, orange curve). A mechanism underlying the initial ID rise could be the fact that the input is dominated by features that are irrelevant for predicting the output, but are highly correlated between each other. To validate this hypothesis, we generated a modified MNIST dataset (referred to as MNIST*) by adding a luminance perturbation that was constant for all pixels within an image, but random across the various images (Fig. 6A). Given an image i with pixels of $x_i \in \mathcal{R}^N$ (where N is the number of pixels), we added shared random perturbations, $x_i \rightarrow x_i^\star = x_i + \lambda\xi_i$ where λ is a positive parameter and ξ_i are i.i.d. uniformly distributed random variables in the range $[0, 1]$. This perturbation has the effect of stretching the dataset along a specific direction in the input space (the vector $[1, 1, \ldots, 1]$) thus reducing the ID of the data manifold in the input layer. Indeed, with $\lambda = 100$, the ID of the input representation dropped from ≈ 13 (its original value) to ≈ 3.

The network trained on MNIST* was still able to generalize (accuracy $\approx 98\%$). However, the variation of the ID (blue curve in Fig. 6B) now showed a hunchback shape reminiscent of that already observed in Figs 2A and 3A,B for large architectures. This suggests that the growth of the ID in the first hidden layers of a deep network is determined by the presence in the input data of low-level features that carry no information about the correct labeling – for instance, in the case of images, gradients of luminance or contrast. One can speculate that, in a trained deep network, the first layers prune the irrelevant features, formatting the representation for the more advanced processing carried out by the last layers (22). The initial increase of the dimensionality of the data manifold could be the signature of such pruning. This notion is consistent with recent evidence gathered in the field of visual neuroscience, where the pruning of low-level confounding features, such as luminance, has

been demonstrated along the progression of visual cortical areas that, in the rat brain, are thought to support shape processing and object recognition (24).

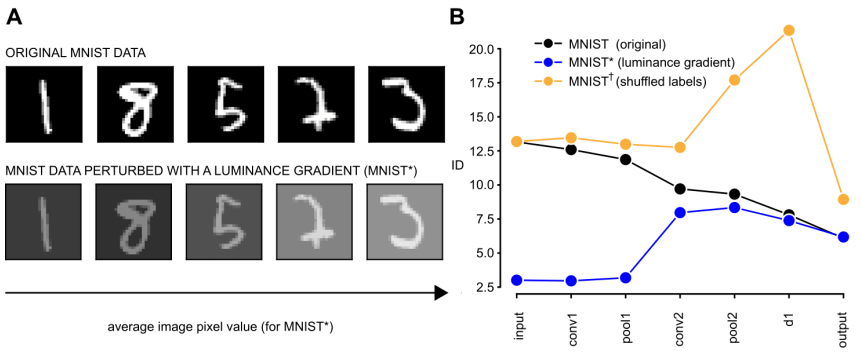

Figure 6: **A** The addition of a luminance gradient across the images of the MNIST dataset results in a stretching of the image manifold along a straight line in the input space of the pixel representation. **B** Change of the ID along all the layers of the MNIST network, as obtained in three different experiments: 1) with the original MNIST dataset (black curve) 2) with the luminance-perturbed MNIST* dataset (blue curve) and 3) with the MNIST†, in which the label of the MNIST images where randomly shuffled (red curve).

3.5 A network trained on random labels does not show the characteristic hunchback profile of ID variation

In untrained networks the ID profile is largely flat (Fig. 5C). Are there other circumstances in which the ID profile deviates from the typical hunchback shape of Figs 2A and 3A,B, with IDs that do not decrease progressively towards the output? It turns out that this is the case when generalization is impossible by construction, as we verified by randomly shuffling the labels on MNIST (we refer to the shuffled data as MNIST†).

It has been shown (1) that deep networks can perfectly fit the training set on randomly labelled data, while necessarily achieving chance level performance on the test set. As a result, when we trained the same network as in section 3.4 on MNIST†, we achieved a training error of zero. However, we found that the network had an ID profile which did not decrease monotonically (orange curve in Fig. 6B) – in contrast to the same network trained with the original dataset (black curve). Instead, it grew considerably in the second half of the network, almost saturating the upper bound, which is set by the ED, in the output layer. This suggests that the reduction of the dimensionality of data manifolds in the last layers of a trained network reflects the process of learning on a generalizable dataset. By contrast, overfitting noisy labels leads to an expansion of the dimensionality, as already reported in (15). As suggested in that study, this indicates that a network trained on inconsistent data can be recognized *without* estimating its performance on a test set, but by simply looking at whether the ID increases substantially across its final layers.

4 Conclusions and Discussion

Convolutional neural networks, as well as their biological counterparts, such as the visual system of primates (25) and other species (24; 26), transform the input images across a progression of processing stages, eventually providing an explicit (i.e. transformation-tolerant) representation of visual objects in the output layer. Leading theories in the field of visual neuroscience postulate that such re-formatting gradually untangles and flattens the manifolds produced by the different images within the representational space defined by the activity of all the neurons (or units) in a layer (25; 27; 28). This suggests that the dimensionality of the object manifolds may progressively decrease along the layers of a deep network, and that such a decrease may be at the root of the high classification accuracy achieved by deep networks. Although previous theoretical and empirical studies have provided support to this hypothesis using small/simple network architectures or focusing on single layers of large networks (6; 10; 15; 29; 30; 31), our study is the first to investigate systematically how

the dimensionality of individual object manifolds - or mixtures of object manifolds - vary in large, state-of-the-art CNNs used for image classification.

Our results can be summarized by making reference to the cartoon shown in Fig. 7. We found that the ID in the initial layer of a network is low. As shown in Fig. 6, this can be explained by the existence of gradients of correlated low-level visual features (e.g., luminance, contrast, etc.) across the image set (32), resulting in a stretching of image representations along a few, main variation axes within the input space (see Fig. 7A). Early layers of DNNs appear to get rid of these correlations, which are irrelevant for the classification task, thus leading to an increase of the ID of the object manifolds (Fig. 2A and 3A,B). As illustrated in Fig. 7B, this can be thought as a sort of whitening of the input data. Such initial dimensionality-expansion is also thought to be performed in the visual system (28; 32), and is consistent with a recent characterization of the dimensionality of image representations in primary visual cortex (23) and with the pruning of low-level information performed by high-order visual cortical areas (24).

After this initial expansion, the representation is squeezed into manifolds of progressively lower ID (Figs 2, 3A,B), as graphically illustrated in Fig. 7C,D). This phenomenon has been already observed by (29) and (6) on simplified datasets and architectures, by (10) in the final, fully connected layers of AlexNet, and by (15) in the last hidden layers of two different DNNs, where the ID evolution was tracked during training. We here demonstrate that this progressive reduction of the dimension of data manifolds is a general behavior and a key signature of every CNN we tested – both small toy models (Fig. 6B) and large state-of-the-art networks (Fig. 3A,B). More importantly, our experiments show that the extent to which a deep network is able to compress the dimensionality of data representations in the last hidden layer is a key predictor of its ability to generalize well to unseen data (Fig. 4) – a finding that is consistent with the inverse relationship between ID and accuracy reported by (15), although our pilot tests suggest that the ID, after a large, initial drop, can slightly increase during training without producing overfitting (Fig. 9). From a theoretical standpoint, this result is broadly consistent with recent studies linking the classification capacity of data manifolds by perceptrons to their geometrical properties (30; 33). Our findings also resonate with the compression of the information about the input data during the final phase of training of deep networks (34), which is progressively larger as a function of the layer's depth, thus displaying a trend that is reminiscent of the one observed for the ID in our study.

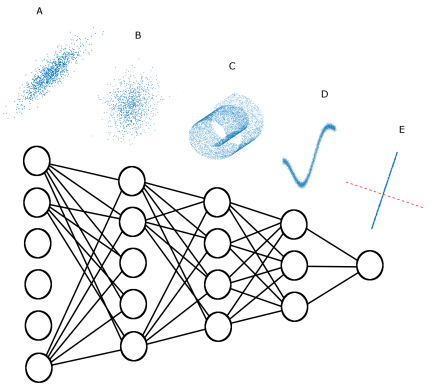

Figure 7: **A**. Input layer. The intrinsic dimensionality of the data can assume low values due to the presence of irrelevant features uncorrelated with the ground truth. **B**. The first hidden layers pre-process the data raising its intrinsic dimension. **C,D**. The representation is squeezed onto manifolds of progressively lower intrinsic dimension. These manifolds are typically not hyperplanes. **D**. In the last hidden layer (**D**) the ID shows a remarkable correlation with the performance in trained networks. **E**. The output layer.

Finally, our experiments also show that the ID values are lower than those identified using PCA, or on 'linearized' data, which is an indication that the data lies on curved manifolds. In addition, ID measures from PCA did not qualitatively distinguish between trained and randomly initialized networks (Fig. 5C). This conclusion is at odds with the unfolding of data manifolds reported by (31) across the layers of a small network tested with simple datasets. It also suggests a slight twist on theories about transformations in the visual system (25; 27) – it indicates that a flattening of data manifolds may not be a general computational goal that deep networks strive to achieve: progressive reduction of the ID, rather than gradual flattening, seems to be the key to achieving linearly separable representations.

To conclude, we hope that data-driven, empirical approaches to investigate deep neural networks, like the one implemented in our study, will provide intuitions and constraints, which will ultimately inspire and enable the development of theoretical explanations of their computational capabilities.

Acknowledgments

We thank Eis Annavini for providing the custom dataset described in A.1.1; Artur Speiser and Jan-Matthis Lückmann for a careful proofreading of the manuscript. We warmly thank Elena Facco for her valuable help in the early phases of this project. We also thank Naftali Tishby, Riccardo Zecchina, Matteo Marsili, Tim Kietzmann, Florent Krzkala, Lenka Zdeborova, Fabio Anselmi, Luca Bortolussi, Jim DiCarlo and SueYeon Chung for useful discussions and suggestions, and the anonymous referees for their useful and constructive comments.

This work was supported by a European Research Council (ERC) Consolidator Grant, 616803-LEARN2SEE (D.Z.). JHM is funded by the German Research Foundation (DFG) through SFB 1233 (276693517), SFB 1089 and SPP 2041, the German Federal Ministry of Education and Research (BMBF, project 'ADMIMEM', FKZ 01IS18052 A-D), and the Human Frontier Science Program (RGY0076/2018).

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, vol. abs/1611.03530, 2016.

[2] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, "Towards understanding the role of over-parametrization in generalization of neural networks," *arXiv preprint arXiv:1805.12076*, 2018.

[3] A. K. Lampinen and S. Ganguli, "An analytic theory of generalization dynamics and transfer learning in deep linear networks," *arXiv preprint arXiv:1809.10374*, 2018.

[4] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. De Freitas, "Predicting parameters in deep learning," in *Advances in Neural Information Processing Systems*, pp. 2148–2156, 2013.

[5] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in neural information processing systems*, pp. 598–605, 1990.

[6] H. Huang, "Mechanisms of dimensionality reduction and decorrelation in deep neural networks," *Physical Review E*, vol. 98, no. 6, p. 062313, 2018.

[7] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K.-i. Kawarabayashi, and M. Nett, "Estimating local intrinsic dimensionality," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 29–38, ACM, 2015.

[8] L. Amsaleg, J. Bailey, D. Barbe, S. Erfani, M. E. Houle, V. Nguyen, and M. Radovanović, "The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality," in *Information Forensics and Security (WIFS), 2017 IEEE Workshop on*, pp. 1–6, IEEE, 2017.

[9] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, M. E. Houle, G. Schoenebeck, D. Song, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," *CoRR*, vol. abs/1801.02613, 2018.

[10] T. Yu, H. Long, and J. E. Hopcroft, "Curvature-based comparison of two neural networks," *arXiv preprint arXiv:1801.06801*, 2018.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[12] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Advances in Neural Information Processing Systems*, pp. 6078–6087, 2017.

[13] A. Morcos, M. Raghu, and S. Bengio, "Insights on representational similarity in neural networks with canonical correlation," in *Advances in Neural Information Processing Systems*, pp. 5727–5736, 2018.

[14] D. G. Barrett, A. S. Morcos, and J. H. Macke, "Analyzing biological and artificial neural networks: challenges with opportunities for synergy?," *Current opinion in neurobiology*, vol. 55, pp. 55–64, 2019.

[15] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. Erfani, S. Xia, S. Wijewickrema, and J. Bailey, "Dimensionality-driven learning with noisy labels," vol. 80, pp. 3355–3364, 10–15 Jul 2018.

[16] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," *Scientific reports*, vol. 7, no. 1, p. 12140, 2017.

[17] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in neural information processing systems*, pp. 777–784, 2005.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] S. Vascon, Y. Parin, E. Annavini, M. D'Andola, D. Zoccolan, and M. Pelillo, "Characterization of visual object representations in rat primary visual cortex," in *European Conference on Computer Vision*, pp. 577–586, Springer, 2018.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[22] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1947–1980, 2018.

[23] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, "High-dimensional geometry of population responses in visual cortex," *bioRxiv*, p. 374090, 2018.

[24] S. Tafazoli, H. Safaai, G. De Franceschi, F. B. Rosselli, W. Vanzella, M. Riggi, F. Buffolo, S. Panzeri, and D. Zoccolan, "Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex," *Elife*, vol. 6, p. e22794, 2017.

[25] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[26] G. Matteucci, R. B. Marotti, M. Riggi, F. B. Rosselli, and D. Zoccolan, "Nonlinear processing of shape information in rat lateral extrastriate cortex," *Journal of Neuroscience*, pp. 1938–18, 2019.

[27] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in cognitive sciences*, vol. 11, no. 8, pp. 333–341, 2007.

[28] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[29] R. Basri and D. Jacobs, "Efficient representation of low-dimensional manifolds using deep networks," *arXiv preprint arXiv:1602.04723*, 2016.

[30] S. Chung, D. D. Lee, and H. Sompolinsky, "Classification and geometry of general perceptual manifolds," *Physical Review X*, vol. 8, no. 3, p. 031003, 2018.

[31] P. P. Brahma, D. Wu, and Y. She, "Why deep learning works: A manifold disentanglement perspective.," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 27, no. 10, pp. 1997–2008, 2016.

[32] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[33] S. Chung, D. D. Lee, and H. Sompolinsky, "Linear readout of object manifolds," *Physical Review E*, vol. 93, no. 6, p. 060301, 2016.

[34] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[36] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.