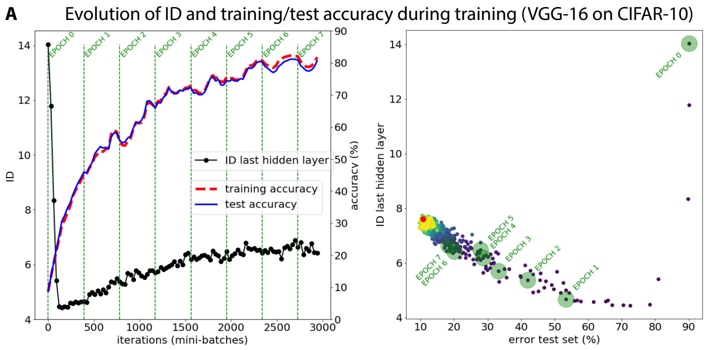
We thank the reviewers for their constructive feedback. Our study explores how the intrinsic dimension (ID) of data manifolds varies across the layers of state-of-the-art deep neural networks (DNNs). Our findings can be summarized as follows: **1)** the ID follows a "hunchback" shape across the layers of a trained DNN (Fig. 3); **2)** data representations live on low-dimensional, but curved manifolds (Fig. 5); **3)** in an untrained network, the ID in all layers is very similar to the ID of the input (Fig. 5C), which can be made arbitrarily low by adding low-level features (Fig. 6); and **4)** the classification accuracy of a network depends on how low the ID of data manifolds is in its last hidden layer (Fig. 4).

As recognized by reviewers R2 and R3, these findings are new. Therefore, we are surprised that reviewer R1 found our conclusions not to be novel enough. We agree that we unfortunately overlooked one very pertinent reference [Ma et al, 2018] and we thank the reviewer for pointing it out – we will cite and discuss this paper in our revision. However, **our results are complementary with [Ma], rather than overlapping.** In fact, [Ma] studied how the ID of a specific layer (the second last) in a DNN varies during optimization, and not how the ID varies across the layers of fully trained DNNs. Only one analysis in our study is partially redundant with [Ma] (and we will acknowledge this accordingly) – the increase of ID produced by random labels reported in Sec. 3.5 (red curve in Fig. 6B). But **none of the major findings summarized above can be found in, or easily inferred from, [Ma] or any other study we are aware of**.

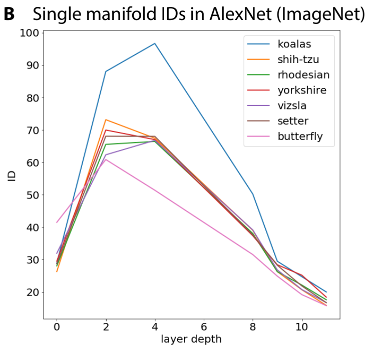
We are particularly surprised by R1 conclusion that "the findings of 3.2 ... have limited novelty", since "[Ma] has already observed that the intrinsic dimension decreases during training (and the accuracy increases ...)". First, [Ma] analysis is limited to the last hidden layer. Our study shows that, for early and middle layers, the trend is exactly the opposite - as a result of training, the ID increases (Fig. 5C). Second, [Ma] does not show that the ID in the last hidden layer predicts the accuracy achieved by different state-of-the-art DNNs, as demonstrated in our Fig. 4. Finally, **following R2 suggestion, we run new analyses to monitor the ID evolution during training**. We observed a monotonic decrease of the ID only in some layers of a NN trained with MNIST, but not in the last hidden layer of VGG-16 trained with CIFAR-10 (see New Fig A). Here, after an initial drop, the ID slowly increased, but, differently from [Ma], without producing any overfitting. This shows that our results not only are largely novel with respect to those of [Ma], but also challenge their generality. Specifically, this new analysis suggests that the evolution of the ID during training may depend on the specific architecture and dataset. This, of course, calls for further investigation.

Further expanding on **R2 suggestion**, we also estimated the variability of the ID in the last hidden layer of a VGG-16 adapted for CIFAR-10 across 50 different trainings, finding no correlation with accuracy (r=-0.003), likely because of the little variation in accuracy produced by different random weight initializations. This suggests that differences in accuracy across well-trained networks (Fig. 4) are mostly due to differences in the architecture. In **response to R2**, we have also verified that the ID variation across layers is generally consistent across object classes (see New Fig. B).

In our revision, following **R1**, we will underline that our ID estimator is a global, not a local one. We will also address **R3 questions** about the reliability of our ID estimates. **First**, we will expand the discussion on the robustness of the ID estimator with respect to the dimension of the embedding space. We already performed a test on artificial data of known ID, embedded in a 100,000 dimensional space. The test did not reveal any significant degradation of accuracy. **Second**, we will comment on the relevance of the independence assumption we made before eq 1. 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 in practice this does not significantly affect the estimate. **Third**, we will stress that our ID estimate is not severely affected by the presence of "hubs". Indeed, the ID is not strongly affected by aggressive decimation of the dataset (Fig. 2B), a procedure which would kill the hubs. **Fourth**, we will provide an interpretation of the value of the ID in the output layer. This value cannot be 1, since all the networks we considered perform classification among N_c categories. Its value should be of the order of $\log(N_c)/\log(2)$. We will also extend the discussion on the implications of the reported low ID. Indeed, we believe that the coordinates on the manifolds should capture the invariance of the representation of objects, and are therefore "meaningful", even if their explicit representation cannot be extracted easily. We will also cite and discuss ref [Achille et al], which we agree is relevant.



A Evolution of ID and training/test accuracy during training (VGG-16 on CIFAR-10)



B Single manifold IDs in AlexNet (ImageNet)