Universality of intrinsic dimension of latent representations across models

Teresa Karen Scheidt
tksc@dtu.dk

Lars Kai Hansen
lkai@dtu.dk

Section for Cognitive Systems, DTU Compute
Technical University of Denmark
2800 Kongens Lyngby, Denmark

Abstract

While state-of-the-art transformer networks use several hundreds of latent variables per layer, it has been shown that these features can actually be represented locally by relatively low dimensional manifolds. The intrinsic dimension is a geometrical property of the manifold latent representations populate, viz., a minimal number of parameters needed to describe the representations. In this work, we compare the intrinsic dimensions of three image transformer networks for classes of the cifar10 and cifar100 dataset. We find compelling evidence that the intrinsic dimensions differ among classes but are universal across networks. This universality persists across different pretraining strategies, fine-tuning and different model sizes. Our results strengthen the hypothesis that different models learn similar representations of data and show great potential that further investigation of intrinsic dimension could lead to more insights on the universality of latent representations.

1 Introduction

According to the manifold hypothesis, high-dimensional real-world datasets populate low-dimensional manifolds. This is thought to be one of the key features leading to generalization in machine learning models. A question that naturally arises from this in the context of similarity in latent representation: Do different models learn the same low-dimensional manifold when presented with the same data? And following this, does the dimensionality of that manifold depend on the data, the different classes in the data and on the model?

One tool to investigate these questions, is measuring the intrinsic dimension (ID) of the latent space. The ID is the minimum amount of variables needed to represent the data without significant information loss. It has been shown, that different models with the same task (e.g. image classification) show similar distibution of ID across layers and also it was found that a low ID in the last layer correlates with higher accuracy [1]. Similar observations were made in self-supervised networks, that exhibit similar 'evolution' of ID even across tasks [9].

In previous work, ID was reported for supervised learning cases as an average across all data [1], hence potentially losing important information about ID class differences. As the ID is a local measure and assuming that any given neighborhood is dominated by a single class (i.e., assuming high classification accuracy is possible) we suggest stratifying ID over classes. Differences in ID between classes could, for example, derive from different invariances and symmetry groups. These questions are new and we have developed a workflow to address whether ID is dependent on the classes in a dataset, i.e. whether some classes are represented in a lower or higher-dimensional manifolds than others, and if this is universal across models.

Our results show, that the ID is depended on the classes of the dataset, but not on the model. The ID of different classes compares across models and exhibits the same relative ordering and distinctive shapes. This leads to the conclusion that different networks do learn similar low-dimensional representations of data and the ID of classes is universal across models.

2 Methods

2.1 Data and models

We analyze the intrinsic dimension of three image transformer networks: vision transformer (ViT) [5], data2vec [2] and BEiT [3]. All models are retrieved from `https://huggingface.co`, see Table 1 for the exact models. While all models share the base transformer architecture, the pretraining strategy between the models differs. While ViT is pretrained by predicting the next image patch, data2vec and BEiT are trained by masking parts of the image. All models are pretrained on ImageNet-1k [8] (although not exclusively). Additionally, we investigate a fine-tuned version of each model. For ViT and BEiT we analyze models fine-tuned on cifar10 and for data2vec a model fine-tuned on ImageNet.

To test our hypothesis that the models learn similar representations, despite their different pre-training settings, we compare the intrinsic dimensions and neighbourhoods of the hidden representations of the cifar10 and cifar100 dataset [7]. The cifar10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The cifar100 dataset consists of 100 classes with 600 images each, of which each class belongs to one of 20 super-classes. There are again 50000 training images and 10000 test images. The data is retrieved from `https://huggingface.co` (cifar10, cifar100). For every model, we extract the mean latent representation after each transformer block for the training images of both datasets. From the latent representations we estimate the ID as described in the following sections.

Table 1: Analyzed models

Model	Layers	Embedding Size	Huggingface model
data2vec	12	768	facebook/data2vec-vision-base
data2vec finetuned	12	768	facebook/data2vec-vision-base-ft1k
ViT base	12	768	google/vit-base-patch16-224
ViT large	24	1024	google/vit-large-patch16-224
ViT huge	32	1280	google/vit-huge-patch14-224-in21k
ViT base finetuned	12	768	nateraw/vit-base-patch16-224-cifar10
BEiT base	12	768	microsoft/beit-base-patch16-224
BEiT large	24	1024	microsoft/beit-large-patch16-224
BEiT base finetuned	12	768	jadohu/BEiT-finetuned

2.2 Intrinsic dimension

The intrinsic dimension (ID) is a geometrical property of the manifold latent representations populate, viz., a minimal number of parameters needed to describe the representations. We use the 'TwoNN' estimator [6] to estimate the ID of the data representations, which is based on the local distance between the datapoint and its two nearest neighbours. The estimator is shown to be robust to curvature, density variations, scale and embedding dimensions [1, 6]. Due to its simplicity it's also well suited for large datasets and high embedding dimensions [6]. Estimated IDs over 20 should be interpreted as lower bounds [1]. For each datapoint, calculate the ratio between the distance to the first (r_1) and second neighbour (r_2) as $\mu = r_2/r_1$. The probability of $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_N)$ is given by [6]

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

With N being the number of datapoints and d the intrinsic dimension. An ID estimator and its posterior uncertainty can be obtained from the likelihood L:

$$\log L\left(\boldsymbol{\mu} \mid d\right) = \log \prod_i P\left(\mu_i \mid d\right) = N\log(d) - (d+1)\sum_i \log(\mu_i)$$

$$\hat{d} = \frac{N}{\sum_i \log \mu_i}, \quad \hat{\sigma}(\hat{d}) = \frac{\hat{d}}{\sqrt{N}}$$

(2)

For nearest neighbour search we use euclidean distances and employ NN-Descent, an efficient algorithm for approximate k-nearest neighbour graph construction [4].

2.3 Comparison

For comparison across models, we use Pearson's r for linear correlation and Spearman's rho for rank correlation. These are calculated for the IDs of each class for each layer and for the mean ID of each class. P-values < 0.05 are considered significant.

3 Results and Discussion

We investigate how the intrinsic dimension of the different classes behaves across different networks. The hypothesis is that the different models converge to similar latent representations, which should also be reflected in similar intrinsic dimensions.

We extracted the ID of the representations of all classes after each transformer layer and observe similar behavior across models. In Figure 1 the ID for the classes of cifar10 is shown for different layers of the three different models (ViT, BEiT and data2vec) and two things can be observed. The exact ID differs between the models, but follows a very similar pattern across models: The ID increases more in early than in later network layers, which could also seen in [1]. Additionally, the ranking of ID of classes is very similar across models. The models generally have a smaller or higher intrinsic dimension for the same classes.

The correlation between the mean ID of different classes (for cifar10) across the models is shown in Table 2. The correlation for the mean ID of each class is above 0.85 for every comparison (all with p<0.005), the correlation on a layer comparison is above 0.70 for all layers and base-size models. The same can be observed for the cifar100 dataset, when looking at the coarse and fine labels (spearman's rho > 0.84 and 0.85 respectively for all comparisons, see Appendix A for detailed results). This indicates that the complexity of learned representations is relative with respect to the classes and not dependent on the used model. Even some distinct behavior of ID evolution can be observed across models, e.g. the ID of class cat is going up after the middle of the network while the ID of most other classes go down, which can be observed for almost all models (see Figure 1).

Fine-tuning of the models did not change the relative ordering of ID of classes and had only little influence on the general development of ID (see Figure 1). For ViT and BEit fine-tuning on cifar10 reduced the ID of the cifar10 classes but the ID curve shape stayed unchanged. In contrast, an increase in ID and a change in the ID curve shape throughout layers could be observed for data2vec after fine-tuning. The fact that this model was fine-tuned on ImageNet and not cifar could contribute to these differences.

We also investigated different sizes of networks (larger latent dimension and more layers). The change in size also did not affect the relative ordering of IDs of the classes, as it can be seen in high correlations with the base model size and other models (see Table 2, last rows/columns).

4 Conclusion and Outlook

In this work, we aimed to investigate the ID of latent representations stratified by classes and compare its evolution and behavior across different models. We showed that the ID of representations of data is dependent on the class and this dependency is universal across models. This supports the hypothesis that models learn similar representations when presented with the same stimuli. The universality in ID is consistent when comparing different model-sizes and also remains after fine-tuning models.
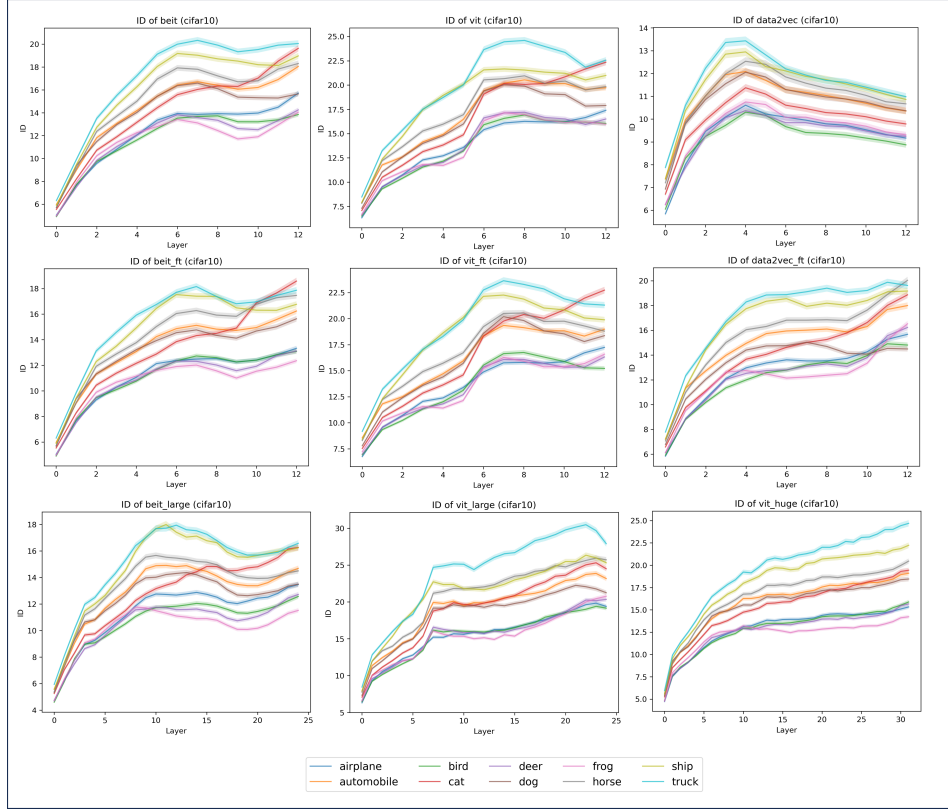
Figure 1: ID for latent representations of classes of cifar10 for BEiT, ViT and data2vec (top row) and their fine-tuned versions (middle row) and larger models BEit/ViT large and ViT huge (bottom row). Shading represents the uncertainties. Similar ordering and development of IDs for classes can be observed across models. Plots for cifar100 can be found in Appendix A.

Table 2: Correlations in between models for mean ID of classes in cifar10, p-value < 0.005 for all correlations. Upper triangle: rank correlation (Spearman's rho), lower triangle: linear correlation (Pearson's r). Correlations for cifar100 can be found in Appendix A.

	data2vec	data2vec ft	vit	vit ft	beit	beit ft	vit large	beit large	vit huge
data2vec	1.000	0.952	0.939	0.903	0.915	0.903	0.927	0.867	0.927
data2vec ft	0.948	1.000	0.988	0.939	0.988	0.964	0.903	0.891	0.927
vit	0.960	0.978	1.000	0.903	0.976	0.939	0.927	0.927	0.964
vit ft	0.951	0.967	0.995	1.000	0.952	0.976	0.891	0.867	0.867
beit	0.961	0.982	0.995	0.993	1.000	0.988	0.879	0.879	0.915
beit ft	0.964	0.968	0.988	0.991	0.994	1.000	0.867	0.855	0.891
vit large	0.876	0.875	0.928	0.908	0.889	0.887	1.000	0.954	0.976
beit large	0.908	0.950	0.985	0.980	0.968	0.961	0.954	1.000	0.964
vit huge	0.940	0.937	0.981	0.969	0.962	0.963	0.966	0.978	1.000

These findings lead to several potential research questions, which we plan to investigate in the future. How can the ID of representations be related to the data? How similar are the manifolds on which the data is represented (in terms of shape, organization, ...)? How can these insights be used to align models with each other and foster knowledge transfer between models?

Acknowledgments and Disclosure of Funding

References

[1] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[3] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[4] W. Dong, C. Moses, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586, 2011.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

[7] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[9] L. Valeriani, D. Doimo, F. Cuturello, A. Laio, A. Ansuini, and A. Cazzaniga. The geometry of hidden representations of large transformer models. *arXiv preprint arXiv:2302.00294*, 2023.

A   Additional results

Results for cifar100: Table 3 and 4 shows the correlations between mean IDs of all classes for fine and coarse labels. Figure 2 and 3 show the development of ID across layers.

Table 3: Correlations in between models for mean ID of classes in cifar100 (fine labels), p-value < 0.005 for all correlations. Upper triangle: rank correlation (Spearman's rho), lower triangle: linear correlation (Pearson's r)

	data2vec	vit	vit large	beit	beit large
data2vec	1.000	0.851	0.850	0.847	0.840
vit	0.869	1.000	0.916	0.983	0.938
vit large	0.838	0.905	1.000	0.905	0.966
beit	0.855	0.984	0.888	1.000	0.929
beit large	0.844	0.926	0.962	0.920	1.000

Table 4: Correlations in between models for mean ID of classes in cifar100 (coarse labels), p-value <
0.005 for all correlations. Upper triangle: rank correlation (Spearman's rho), lower triangle: linear
correlation (Pearson's r)

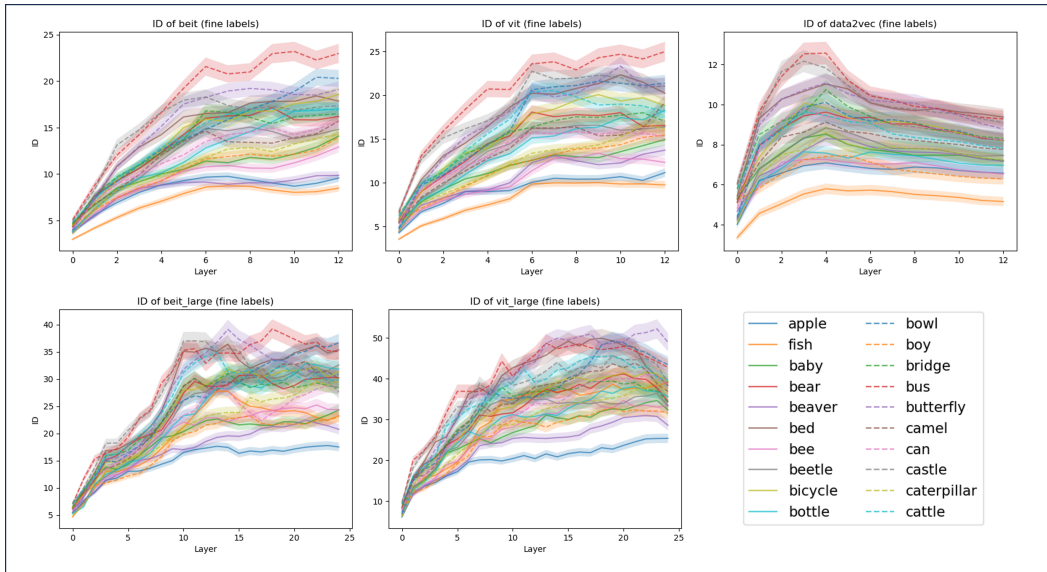|	data2vec	vit	vit large	beit	beit large
data2vec	1.000	0.856	0.798	0.874	0.832
vit	0.831	1.000	0.929	0.988	0.955
vit large	0.772	0.885	1.000	0.914	0.977
beit	0.855	0.992	0.883	1.000	0.955
beit large	0.817	0.916	0.962	0.917	1.000

Figure 2: ID for latent representations of fine classes of cifar100 for BEiT, ViT and data2vec (top
row) and larger models BEit/ViT large (bottom row). Shading represents the uncertainties. Similar
ordering and development of IDs for classes can be observed across models. Only the first 20 classes
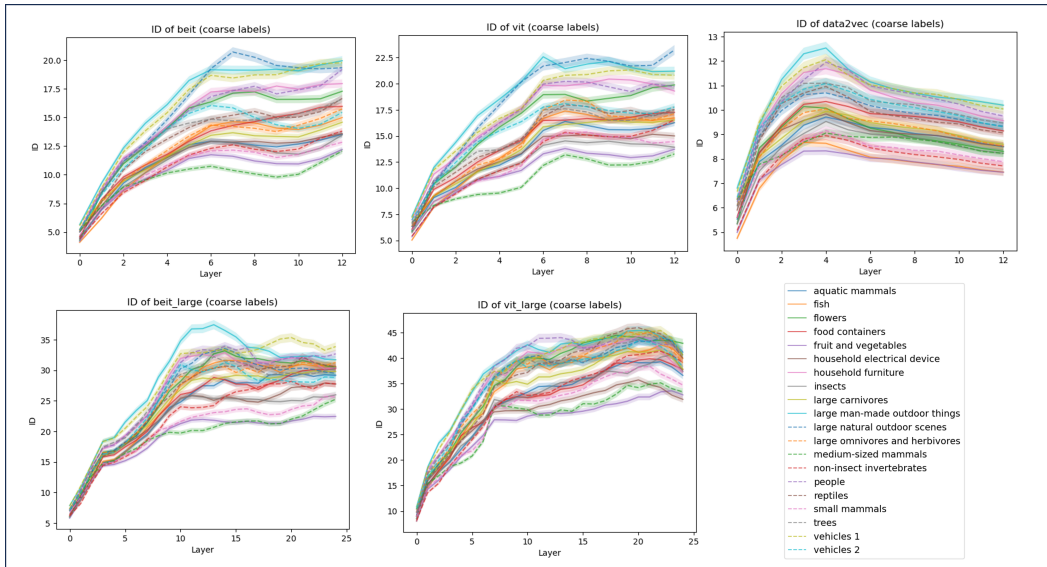are shown to make the figure comprehensible.

Figure 3: ID for latent representations of coarse classes of cifar100 for BEiT, ViT and data2vec
(top row) and BEit large (bottom row). Shading represents the uncertainties. Similar ordering and
development of IDs for classes can be observed across models..

6