Pre-processing and Compression: Understanding Hidden Representation Refinement Across Imaging Domains via Intrinsic Dimension

Nicholas Konz[1], **Maciej A. Mazurowski**[1,2,3,4]
[1] Department of Electrical and Computer Engineering, [2] Department of Radiology,
[3] Department of Computer Science, [4] Department of Biostatistics & Bioinformatics
Duke University, NC, USA
{nicholas.konz, maciej.mazurowski}@duke.edu

Abstract

In recent years, there has been interest in how geometric properties such as *intrinsic dimension* (ID) of a neural network's hidden representations change through its layers, and how such properties are predictive of important model behavior such as generalization ability. However, evidence has begun to emerge that such behavior can change significantly depending on the domain of the network's training data, such as natural versus medical images. Here, we further this inquiry by exploring how the ID of a network's learned representations changes through its layers, in essence, characterizing how the network successively refines the information content of input data to be used for predictions. Analyzing eleven natural and medical image datasets across six network architectures, we find that how ID changes through the network differs noticeably between natural and medical image models. Specifically, medical image models peak in representation ID earlier in the network, implying a difference in the image features and their abstractness that are typically used for downstream tasks in these domains. Additionally, we discover a strong correlation of this peak representation ID with the ID of the data in its input space, implying that the intrinsic information content of a model's learned representations is guided by that of the data it was trained on. Overall, our findings emphasize notable discrepancies in network behavior between natural and non-natural imaging domains regarding hidden representation information content, and provide further insights into how a network's learned features are shaped by its training data.

Introduction

In science, it is common to use relatively simple models to approximate the behavior of complex systems. The parameters of such models for real systems can often be estimated using observed data, which can then be used to extrapolate to future behavior. This approach has been increasingly applied to deep learning, such as where relatively simple measurable geometrical properties (intrinsic dimension, curvature, etc.) of a network's manifold of learned representations [2] or training data [27] can predict important behavior such as generalization, adversarial robustness, or prompt perplexity [19]. However, it has been recently found that such behavior can vary significantly between data domains—in particular natural versus medical images—with many open questions remaining about what other behavior can change due to such domain shift [16, 17, 35]. Here, we further this inquiry by exploring how the *intrinsic dimension* (ID) of a neural network's learned representations—which

38th Conference on Neural Information Processing Systems (NeurIPS 2024): Workshop on Scientific Methods for Understanding Deep Learning (SciForDL).

describes their intrinsic information content/minimum number of degrees of freedom—changes through its layers, and how this behavior changes depending on the dataset or data domain.

It has been found that for ImageNet [4] classification networks (both convolutional and transformer-based), the ID of hidden representations typically first steadily increases through successive network layers, then past a certain point, decreases until the output [2, 35]. It is hypothesized that this first stage of "dimensionality expansion" is the network *pre-processing* the data to disentangle certain correlated, nuisance features (such as luminance or contrast) which are irrelevant to the final prediction, which may have parallels in human visual processing [32, 33]. Once this is completed and the ID has peaked, the second stage begins where the network successively *compresses* the representations to lower- and lower-dimensional manifolds that are better suited for generalizing to unseen data [2, 22]. Importantly, prior results with ImageNet models have shown the ID of the final hidden layer representations to be closely correlated with generalization error [2], pointing it to being a useful property for describing network behavior. However, the generality of such findings in other datasets or data domains is unknown.

In this work, we first explore how layer-by-layer representation ID progression differs for models trained on datasets beyond ImageNet, including both three other natural image datasets and seven medical image datasets for a wide range of diagnostic tasks. We find that these curves have typically different shapes depending on if the model was trained on natural or medical images. In particular, the depth at which the ID peaks at some $d_{\text{repr}}^{\max}$ is noticeably deeper into the model on average for natural image models, which we attribute to the task-relevant features of medical images typically requiring less abstract representations to capture, and so can be "pre-processed" for compression earlier in the network. Next, we explore the relationship between $d_{\text{repr}}^{\max}$ and the intrinsic dimension of the raw data in its native space, d_{data} [27], and discover a new result: there is a strong correlation between $d_{\text{repr}}^{\max}$ and d_{data} across all datasets and models, implying that the maximum intrinsic complexity of learned representations in a model is guided by the intrinsic complexity of the model's training data.

1 Experimental Methods

The manifold hypothesis and intrinsic dimension. Neural networks work by mapping high-dimensional input data (like images) to significantly lower-dimensional manifolds of learned representations [10, 2] which describe features that generalize to new data, made possible by learning the lower-dimensional manifold structure of such datasets in their native space [7]. The intrinsic dimension of such a manifold can be computed via a relatively simple estimator that utilizes maximum likelihood estimation (MLE) [20, 23] given the high dimensional dataset (in our case, of representations) which we use here following prior work [17, 27] via the implementation of `https://github.com/mazurowski-lab/intrinsic-properties`. We set the nearest-neighbor hyperparameter to $k = 20$ following [17, 16, 27].

Consider some C-class image classification neural network $f : \mathbb{R}^n \to \mathbb{R}^C$ with L layers f_i indexed by $i = 1, \ldots, L$, *i.e.*, $f = f_L \circ \ldots \circ f_2 \circ f_1$, and $f_{\leq i}(x) := (f_i \circ \ldots \circ f_1)(x)$ denotes the ith hidden layer output given input data $x \in \mathbb{R}^n$. Given some *unlabeled* input dataset $\mathcal{X}$, the intrinsic dimension d_{repr}^i of an ith layer's manifold of representations $\mathcal{H} := \{f_{\leq i}(x) : x \sim \mathcal{X}\}$ can be computed by applying one of the aforementioned estimators to $\mathcal{H}$. In this work, we will mainly analyze the way that representation ID changes from layer to layer, *i.e.*, the sequence $d_{\text{repr}}^1, \ldots, d_{\text{repr}}^i, \ldots, d_{\text{repr}}^L$.

Datasets. We experiment on binary classification datasets from both natural imaging and medical imaging. The former includes **ImageNet** [4], **CIFAR10** [18], **SVHN** [26], and **MNIST** [5], where the two predictive classes are randomly chosen for any given experiment (*i.e.*, model training and ID estimation). The latter utilizes the same datasets and predictive tasks as [17]; including (1) brain MRI glioma detection (**BraTS**, [24]); (2) breast MRI cancer detection (**DBC**, [29]); (3) prostate MRI cancer risk scoring (**Prostate MRI**, [31]); (4) brain CT hemorrhage detection (**RSNA-IH-CT**, [8]); (5) chest X-ray pleural effusion detection (**CheXpert**, [14]); (6) musculoskeletal X-ray abnormality detection (**MURA**, [28]); and (7) knee X-ray osteoarthritis detection (**OAI**, [34]).

We follow the same procedure for dataset creation as in [17], creating randomly-sampled, even class-balanced training sets of size $N \in \{500, 750, 1000, 1250, 1500, 1750\}$ and test sets of size 750; natural image model results are averaged over 5 runs of different randomly selected class pairings. Representation ID estimates for a given model are made using its training set, following [2]. Our

results will be presented for $N = 1750$ unless otherwise stated, but we provide all results for the other N in Appendix A.1. All images are resized to 224×224 and normalized to $[0, 1]$.

Models. We evaluate six common convolutional network architectures: VGG-13, -16, -19 [30] and ResNet-18, -34, and -50 [12]. Each network is trained via Adam [15] until maximally fitting to the given training set. We analyze the representations outputted by all pooling layers, convolutional blocks (or residual blocks for the ResNet models), and fully-connected layers.

2 Experiments and Results

Progression of hidden representation intrinsic dimension through network depth. We will first analyze how representation intrinsic dimension (ID) d^i_{repr} changes with respect to relative network depth i/L for models trained on either natural or medical image datasets or, in other words, the way that intrinsic representation information content changes through the depth of the network. We will characterize these ID curves by the depth

$$i^* := \text{argmax}_i(d^i_{\text{repr}}) \tag{1}$$

at which the ID peaks at some $d^{\max}_{\text{repr}} := d^{i^*}_{\text{repr}}$, transitioning from the aforementioned "pre-processing" stage to the "compression" stage of the model. This $d^{\max}_{\text{repr}}$ is hypothesized to be the minimum number of degrees of freedom needed to represent the data such that task-irrelevant features are decorrelated with task-related features before the compression phase can begin. We show these curves for all datasets in Fig. 1, along with peak ID values $d^{\max}_{\text{repr}}$ and their depths i^*/L for each dataset in Table 1, averaged over all models.

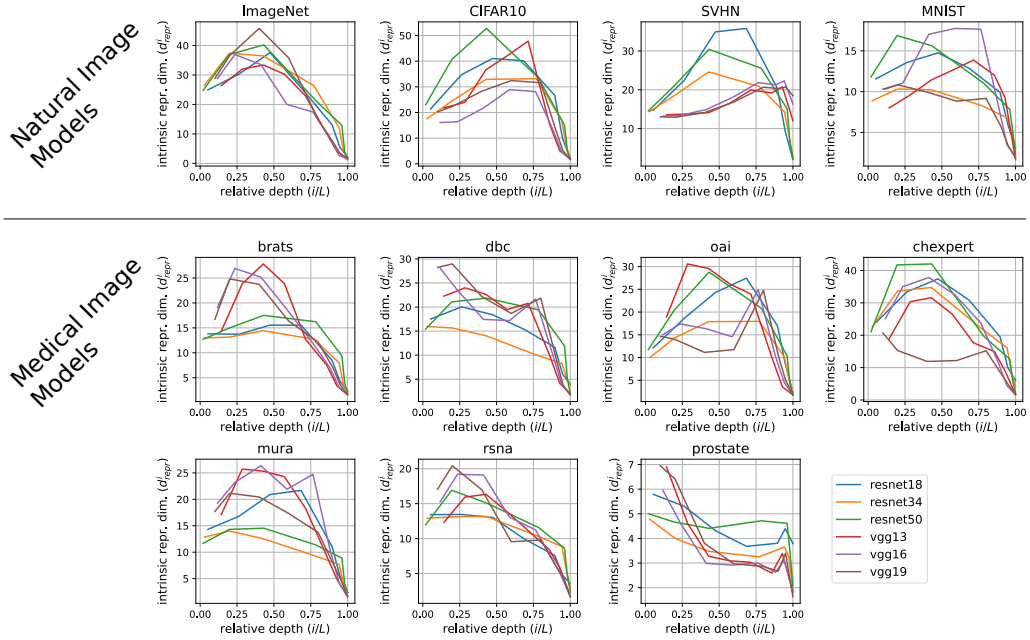

Figure 1: Progression of representation intrinsic dimension d^i_{repr} with respect to relative network depth i/L, for all models trained on each dataset.

We see that the typical shapes of these ID curves differ depending on whether the model was trained on a natural image or medical image dataset; in particular, the average $d^{\max}_{\text{repr}}$ for natural image models is 0.57, while for medical image models it is noticeably earlier in the networks at 0.35. This implies that for medical image datasets, features needed for diagnostic tasks are typically less abstract/hierarchical than those for natural image tasks and so can be pre-processed by the network for prediction usage at earlier layers. We find the same for other training set sizes (Appendix A.1.1).

	ImageNet	CIFAR-10	SVHN	MNIST	**Avg.**
$d_{\text{repr}}^{\max}$	39 ± 4	39 ± 9	26 ± 6	14 ± 3	$\mathbf{29 \pm 10}$
i^*/L	0.40 ± 0.10	0.65 ± 0.14	0.77 ± 0.24	0.47 ± 0.24	$\mathbf{0.57 \pm 0.15}$

BraTS	DBC	OAI	CheXpert	MURA	RSNA-IH-CT	Prostate-MRI	**Avg.**
21 ± 5	23 ± 5	26 ± 4	34 ± 7	21 ± 5	17 ± 3	6 ± 1	$\mathbf{21 \pm 8}$
0.40 ± 0.10	0.23 ± 0.14	0.65 ± 0.22	0.40 ± 0.16	0.40 ± 0.16	0.33 ± 0.10	0.05 ± 0.00	$\mathbf{0.35 \pm 0.17}$

Table 1: Hidden representation peak intrinsic dimension $d_{\text{repr}}^{\max}$ and the relative network depth i^*/L at which it was attained for each dataset (see Fig. 1 for reference), averaged over all architectures. **Top table:** natural image datasets. **Bottom table:** medical image datasets. Confidence intervals provided as standard deviations over all model architectures.

The relationship between hidden representation intrinsic dimension and dataset intrinsic dimension. We hypothesize that the learned representation manifolds of a network may be shaped by the properties and complexity of the dataset manifold in its native (pixel) space. Prior work [17] studied the relationship between the intrinsic dimension of final hidden layer representations for classification networks, d_{repr}^{L-1}, with the intrinsic dimension of the dataset itself, d_{data}, and theoretically and empirically showed that $d_{\text{repr}}^{L-1} \lesssim d_{\text{data}}$. Intuitively, this describes that the network learns to convert the data to a final hidden representation constructed mainly of intrinsic features that are *task-relevant*, which will be a subset of a larger number of intrinsic features that can be used to describe the data in its original space. Here we will explore this relationship between the intrinsic information content of input data and representations further.

When a network's representations reach a peak intrinsic dimension $d_{\text{repr}}^{\max}$ at some depth (Fig. 1), it implies that task-irrelevant features have fully been decorrelated from task-relevant features in the post-processing stage, such that the former can be removed from the representation in the remaining layers (the compression stage) without affecting the latter. We therefore hypothesize that $d_{\text{repr}}^{\max}$ will be proportional to the dataset's d_{data} because it describes the input data features in a decorrelated manner (as degrees of freedom) before any task-irrelevant features are discarded in later layers. We evaluated this empirically as shown in Fig. 2, averaging $d_{\text{repr}}^{\max}$ and d_{data} estimates for each dataset over all models, and intriguingly found this correlation to be quite high ($r = 0.89$), with $d_{\text{repr}}^{\max}$ being roughly double d_{data}. This relationship and correlation is quite consistent for different training set sizes (Appendix A.1.2).

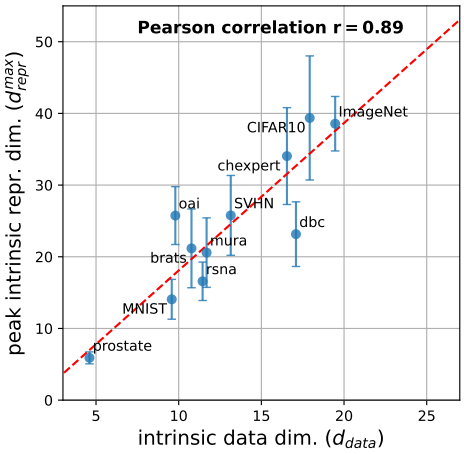

Figure 2: Correlation of peak hidden representation intrinsic dimension $d_{\text{repr}}^{\max}$ with intrinsic dataset dimension d_{data}, averaged (with standard deviation error bars) over all models.

Why is it not the case that this constant of proportionality is closer to 1, *i.e.*, $d_{\text{repr}}^{\max} \approx d_{\text{data}}$? We hypothesize that while the networks learn to compress input data to a dimensionality much lower than the *extrinsic* dimension/pixel count (on the order of 10^5) while still well-representing the data's information content, it is not quite the perfectly optimal, truly minimum number of degrees of freedom that can describe the data. While this could be due to limited network capacity, in our over-parameterized regime it is more likely that compressing the data further simply has a marginal effect on downstream task performance, as the representation dimensionality is already quite low and usable for downstream task prediction in the compression stage of the network, so the network has no need to learn to do so.

Related Work

Beyond intrinsic dimension (ID), other metrics of the intrinsic complexity, dimensionality, capacity, or related characteristics of a hidden layer's representations include manifold capacity [3], linear ID [13], local ID [1, 21], and others. We choose to analyze MLE estimator-based ID rather than manifold capacity due to it not requiring specifying labels/a task for the dataset, as manifold capacity focuses on how class representation manifolds are modified through the network, so that our findings

may more easily generalize to models for tasks beyond classification. We focus on global nonlinear ID rather than local or linear ID due to its established relationship with important phenomena such as network generalization ability [2]. Other general nonlinear manifold ID estimators beyond MLE include TwoNN [6], GeoMLE [9], and k-NN Graph Distance [11].

Conclusions

In this work, we explored how the intrinsic dimension (ID) of hidden representations changes through a neural network's depth, in particular studying how the shape of this ID curve changes depending on the training dataset and data domain. We found that these curves peak at noticeably earlier layers when the model is trained with medical imaging data compared to natural images, implying a discrepancy between the two data domains in the abstractness of prediction-relevant features. We also studied the relationship between this peak representation ID and the ID of the dataset in its input space, between which we found a strong correlation, pointing to a close connection between the information content of a model's representations and the content of its training data.

We expect that our findings will extend to other tasks beyond binary classification, not just because the ID estimator does not use labels, but also because similar "pre-processing then compression" representation ID progression curves (Fig. 1) have been seen in models for both multi-class classification [2] and other non-supervised, non-vision (as well as non-CNN) models [35]. Future research directions could include (1) further probing our proposed connections between learned feature "abstractness" and peak representation ID, such as by seeing if training for classification tasks with increasingly abstract super-classes (*e.g.*, as in WordNet [25]) results in an increased peak ID depth in the network; as well as (2) attempting to disentangle the effects of dataset ID and feature "abstractness" on peak representation ID using experiments that can control the precise ID of training data, such as in [27].

References

[1] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38, 2015.

[2] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[6] 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, 2017.

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

[8] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

[9] Marina Gomtsyan, Nikita Mokrov, Maxim Panov, and Yury Yanovich. Geometry-aware maximum likelihood estimation of intrinsic dimension. In *Asian Conference on Machine Learning*, pages 1126–1141. PMLR, 2019.

[10] Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019.

[11] Daniele Granata and Vincenzo Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific reports*, 6(1):31377, 2016.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Haiping Huang. Mechanisms of dimensionality reduction and decorrelation in deep neural networks. *Physical Review E*, 98(6):062313, 2018.

[14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[16] Nicholas Konz, Hanxue Gu, Haoyu Dong, and Maciej A Mazurowski. The intrinsic manifolds of radiological images and their role in deep learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 684–694. Springer, 2022.

[17] Nicholas Konz and Maciej A Mazurowski. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Henry Kvinge, Davis Brown, and Charles Godfrey. Exploring the representation manifolds of stable diffusion through the lens of intrinsic dimension. *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[20] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[21] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[22] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pages 3355–3364. PMLR, 2018.

[23] David JC MacKay and Zoubin Ghahramani. Comments on'maximum likelihood estimation of intrinsic dimension'by e. levina and p. bickel (2004). *The Inference Group Website, Cavendish Laboratory, Cambridge University*, 2005.

[24] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[25] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[27] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

[28] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.

[29] Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Connie E Kim, Sujata V Ghate, Ruth Walsh, and Maciej A Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British journal of cancer*, 119(4):508–516, 2018.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[31] Geoffrey A Sonn, Shyam Natarajan, Daniel JA Margolis, Malu MacAiran, Patricia Lieu, Jiaoti Huang, Frederick J Dorey, and Leonard S Marks. Targeted biopsy in the detection of prostate cancer using an office based magnetic resonance ultrasound fusion device. *The Journal of urology*, 189(1):86–92, 2013.

[32] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

[33] Sina Tafazoli, Houman Safaai, Gioia De Franceschi, Federica Bianca Rosselli, Walter Vanzella, Margherita Riggi, Federica Buffolo, Stefano Panzeri, and Davide Zoccolan. Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex. *Elife*, 6:e22794, 2017.

[34] Aleksei Tiulpin, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Scientific Reports*, 8(1):1727, 2018.

[35] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36, 2024.

A Supplementary Material

A.1 Additional Results

A.1.1 Progression of hidden representation intrinsic dimension through network depth.

We provide Table 2 to show the same results of average peak intrinsic dimension $d_{\text{repr}}^{\max}$ and the relative network depth i^*/L at which it was attained for natural vs. medical images, to supplement Table 1 for all training set sizes N. We also show the actual representation ID curve plots for each dataset for these other N in Figs. 3, 4, 5, 6, and 7 to supplement Fig. 1.

A.1.2 The relationship between hidden representation intrinsic dimension and dataset intrinsic dimension.

We provide Fig. 8 to show the same results of Fig. 2 for all other training set sizes.

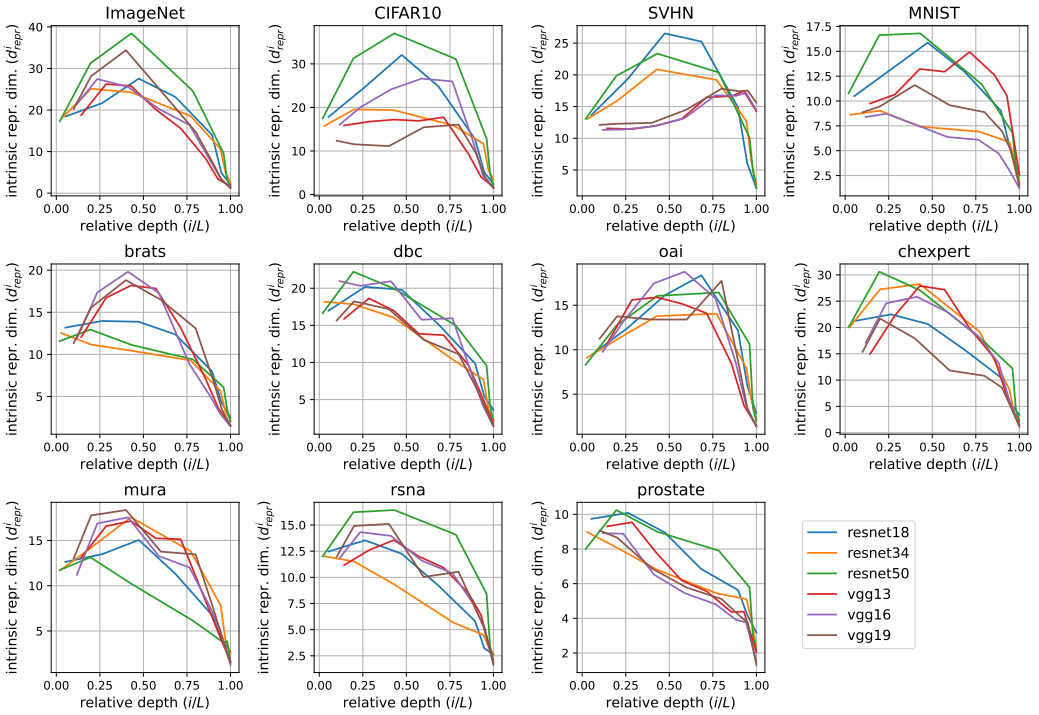

Figure 3: Progression of hidden representation intrinsic dimension d^i_{repr} with respect to relative network depth i/L, for all models trained on each dataset, for $N = 500$. **Top row:** natural image models. **Lower rows:** medical image models.

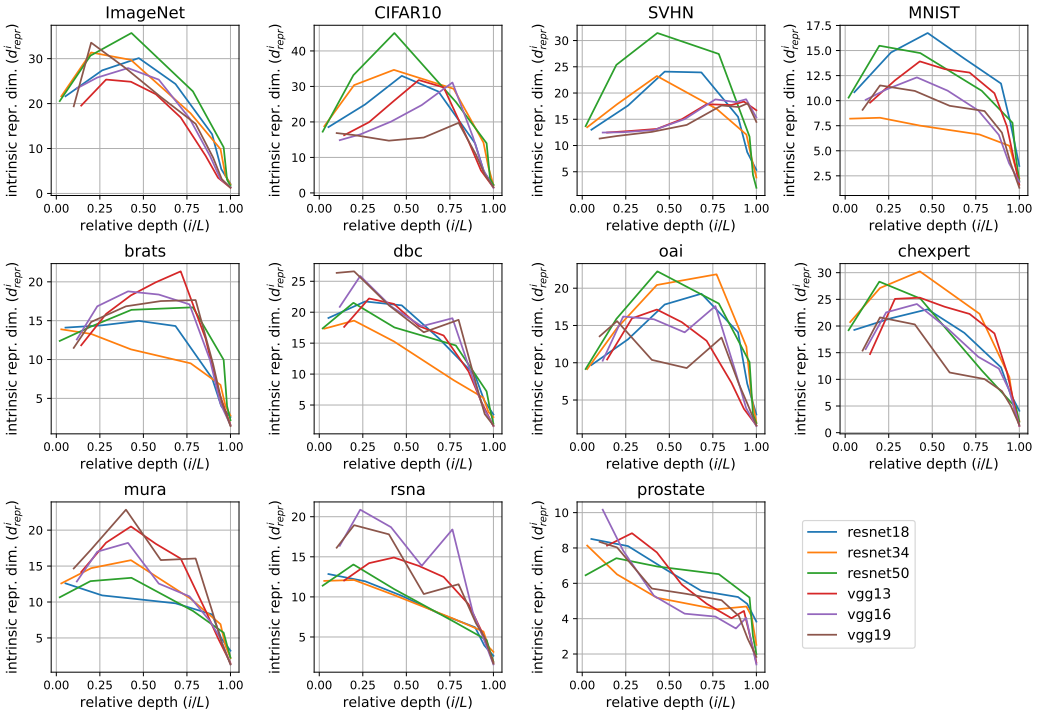

Figure 4: Progression of hidden representation intrinsic dimension d^i_{repr} with respect to relative network depth i/L, for all models trained on each dataset, for $N = 750$. **Top row:** natural image models. **Lower rows:** medical image models.

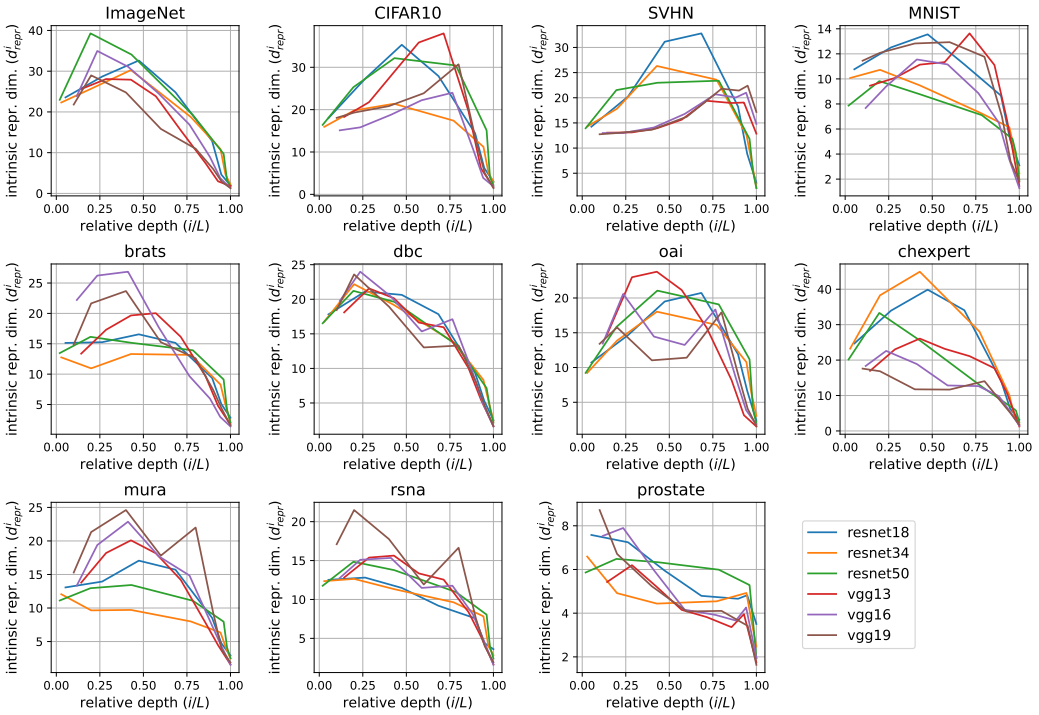

Figure 5: Progression of hidden representation intrinsic dimension d^i_{repr} with respect to relative network depth i/L, for all models trained on each dataset, for $N = 1000$. **Top row:** natural image models. **Lower rows:** medical image models.

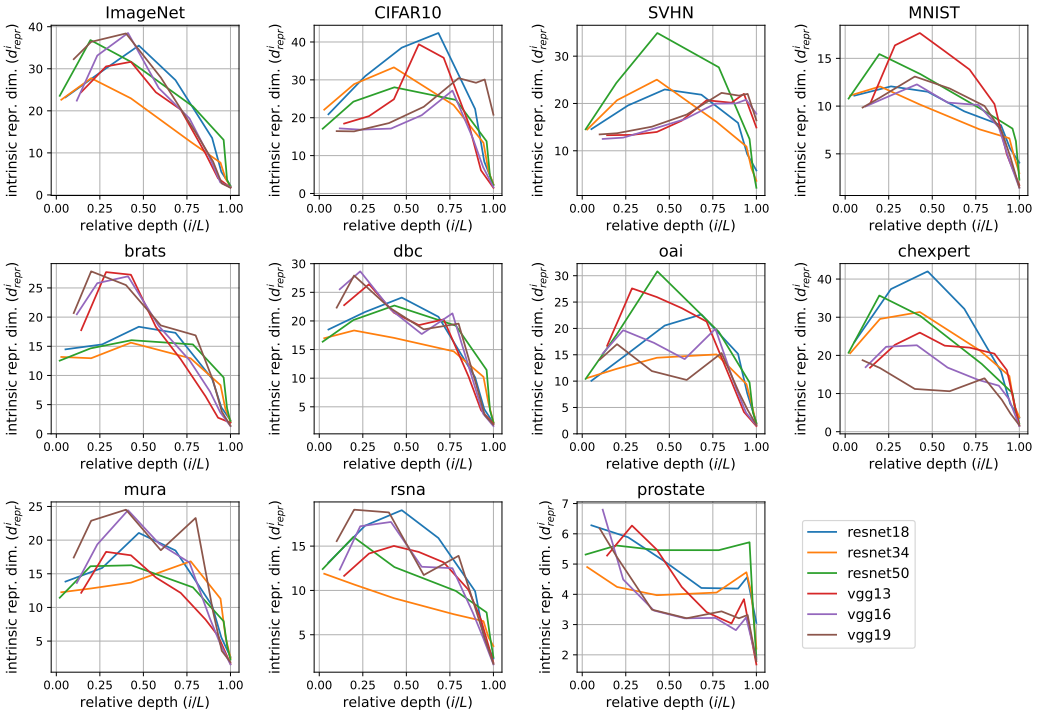

Figure 6: Progression of hidden representation intrinsic dimension d^i_{repr} with respect to relative network depth i/L, for all models trained on each dataset, for $N = 1250$. **Top row:** natural image models. **Lower rows:** medical image models.

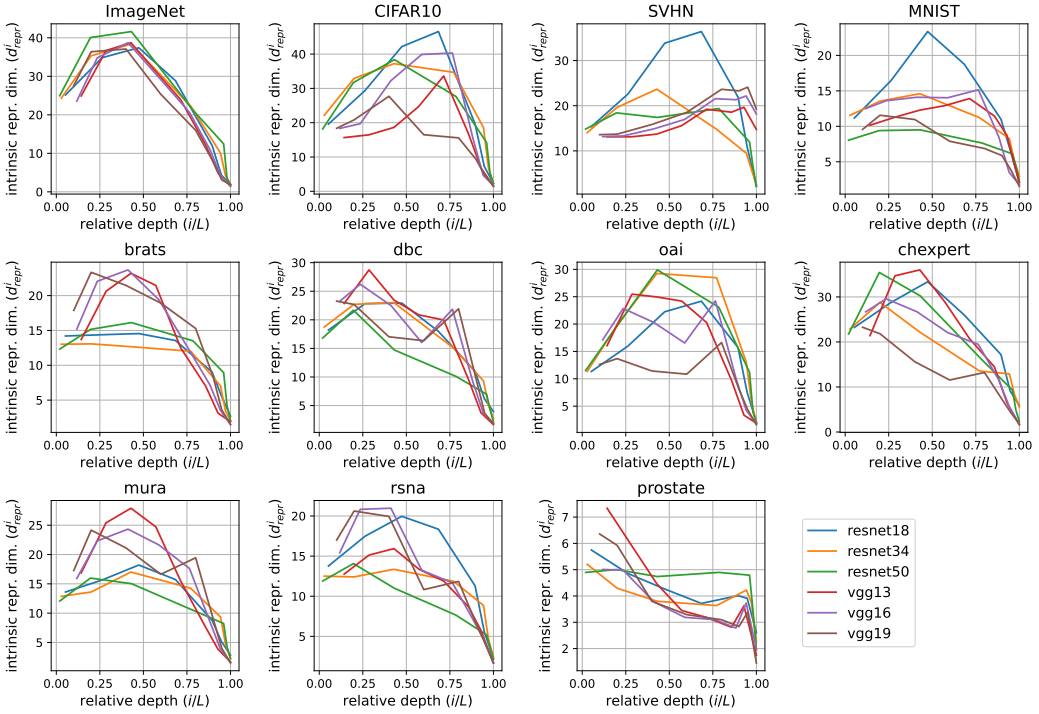

Figure 7: Progression of hidden representation intrinsic dimension d^i_{repr} with respect to relative network depth i/L, for all models trained on each dataset, for $N = 1500$. **Top row:** natural image models. **Lower rows:** medical image models.

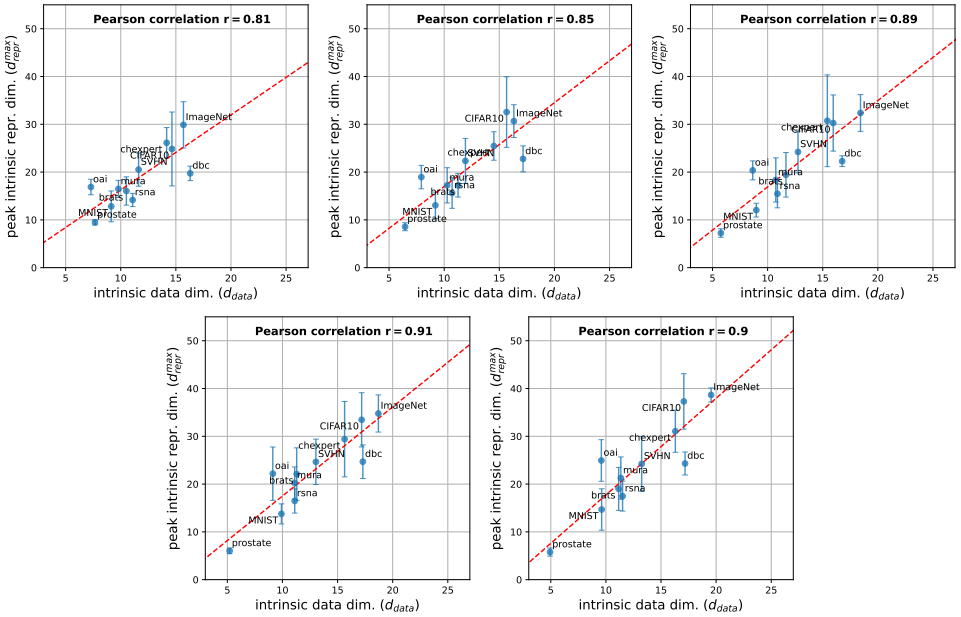

Figure 8: Correlation of peak hidden representation intrinsic dimension $d^{\max}_{\text{repr}}$ with intrinsic dataset dimension d_{data}, averaged over all models, for additional training site sizes of $N = 500, 750, 1000, 1250, 1500$, ordered left-to-right, to supplement the $N = 1750$ result of Fig. 2.

Natural Image Models

	$N = 500$	$N = 750$	$N = 1000$	$N = 1250$	$N = 1500$	$N = 1750$
$d_{\text{repr}}^{\max}$	22 ± 6	25 ± 8	25 ± 8	27 ± 8	29 ± 10	29 ± 10
i^*/L	0.54 ± 0.13	0.53 ± 0.17	0.58 ± 0.17	0.54 ± 0.16	0.63 ± 0.12	0.57 ± 0.15

Medical Image Models

	$N = 500$	$N = 750$	$N = 1000$	$N = 1250$	$N = 1500$	$N = 1750$
$d_{\text{repr}}^{\max}$	17 ± 5	18 ± 5	19 ± 7	20 ± 7	21 ± 7	21 ± 8
i^*/L	0.36 ± 0.16	0.37 ± 0.16	0.36 ± 0.12	0.38 ± 0.10	0.36 ± 0.15	0.35 ± 0.17

Table 2: Hidden representation peak intrinsic dimension $d_{\text{repr}}^{\max}$ and the relative network depth i^*/L at which it was attained, on average over all natural image models vs. all medical image models, for each training set size N (extending the results of Table 1, averaged over all architectures.