A Geometric View of Data Complexity: Efficient Local Intrinsic Dimension Estimation with Diffusion Models

Hamidreza Kamkari, Brendan Ross, Rasa Hosseinzadeh, Jesse Cresswell, Gabriel Loaiza-GanemLayer 6 AI, Toronto, CanadaSpotlight presentationarXiv:2406.03537





The Manifold Hypothesis

Data used in machine learning often lies on **low-dimensional** submanifolds of their ambient space [1]

(Definition) Local Intrinsic Dimension LID(x)

The dimension of the submanifold datapoint x belongs to

Local Intrinsic Dimension: Relative Complexity



Applications



Detecting Al-generated content [3, 4]

Detecting memorization [5]

Use a diffusion model trained on the data manifold, to extract the LID.

Fokker-Planck LID

The forward process:
$$t \in [0, 1]$$
, $dX_t = f(X_t, t)dt + g(t)dW_t$ The transition kernel: $p_{X_t|X_0}(x_t|x_0) = \mathcal{N}(x_t; \psi(t) x_0, \sigma^2(t))$

Diffusion learns the score [6]: $\hat{s}_{\theta}(\cdot, t) \approx s(\cdot, t) \coloneqq \nabla \log p(\cdot, t)$

$$\widehat{\text{LID}}(x; \delta) \coloneqq D + \sigma^2(t(\delta)) \left(\operatorname{tr} \nabla \widehat{s}_{\theta} \left(\psi(t(\delta))x, t(\delta) \right) + \|\widehat{s}_{\theta} \left(\psi(t(\delta))x, t(\delta) \right)\|_2^2 \right)$$
$$\operatorname{LID}(x) = \lim_{\delta \to -\infty} \widehat{\text{LID}}(x; \delta) \qquad Use \ \text{Hutchinson with a few JVPs!}$$

LID of LAION Images





Largest LIDs

A single JVP on Stable Diffusion yields data complexity on ~1 million dimensional images.

111111111/1/	312 〒2:81113 日間にある		- The share the
1/ 1///////////////////////////////////	50 60 12 E 2 304		NST Current P 1 - K
/ \ / / / / / 1 1 1 / 1	5 10 / 49 4 121 4 218 2	n n a i m n T e 🕆 n h 🕇	
1/1/1/1/1/	14 1411 1 014 8 19 241		
111111111111	2 3 4 5 HS 57 14 6 4 4 D 7 1 8 17 1	T4 ~ † -1 † ~	
/ / / / / / / / / / / / /	C2 1 719 2 151 9 510 2		
1///(//\\//	B 21, 0 7678 21 2 30 7 19 2		
////!!!!!!!!	THE T ST DATENIA 1 6 197		T - W W
11/////////	い日間 日本分 日本12月1日日	1 A A ¥ 3 I A I 7 V 1 B	C C C C C C C C C C C C C C C C C C C
/ / / / / / / / / / / / / /	10 55 24 18 18 1 1 1 4 2		
1 / / / (1 / 1 / 1 / / / / / / / / / / /	12 4 13 12 10 3f 1 1711 2		
1/1////////////////////////////////////		<u>e-tillaenilt</u>	
1711177777777			Hard Contraction
7 (7) / / 1 / \ 1	31 _ 10 4 152 2.5 0 0		
	6 71 4 4 4 1 1 2 1 2 1		
	10 5 J 10 5 J 10 5 10 10 10	<u></u>	
		▋₵₵ड़ॾड़ॖॾॵॖॿॖख़ॖॻ॒ॖॖॖॖ	
583600292600		N ~ 1 图 2 頁 2 明 ~ 2 頁 2 图 😁	
688207002802	1 6 5 0 6 1 4 1 4 5 White 1 1 1 3		
505005354085			
038548230680			
		(1)() (1)() (1)() (1)() (1)()()()()()()(
0 C 7 C 9 V 0 C 2 V 8 X		新 2. 4.14 高 梁 6 高 6 高 6 高	
240229400002			
2500385805-0			
202000000000000	11 20-20 20 40 45 33 7 4 4- 45 24		
0332802/0343		8 4 4 1 1 1 2 8 1 1 4 1 7 1 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
855005782803	7 2 2 30 13451 2029 6 28/0018		
500030580056	18/25 25 75 (31017803)0/126	前日月 一日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日	ALL NALTER STATE
509584233079	3 5244240 1774 78 1 13 20 44 44	8 # T mm 8 7 7 4 1 44 7	
235282883675	01451 52 41 62 442 869 176	**************************************	
062883025280	0 7 15 16 15 28 11525 5 80 5 28	的新常的最富贵的最贵重的	
590682889588	12 136 51 2 163 216 146 3 94 38 565		

LID for Synthetic Data

	Model-based					Model-free				
Synthetic Manifold	FLIPD		NB		LIDL		ESS		LPCA	
String within doughnut $\subseteq \mathbb{R}^3$	0.06	1.00	1.48	0.48	1.10	0.99	0.02	1.00	0.00	1.00
$\mathcal{L}_5 \subseteq \mathbb{R}^{10}$	0.17	-	1.00	-	0.10	-	0.07	-	0.00	-
$\mathcal{N}_{90} \subseteq \mathbb{R}^{100}$	0.49	-	0.18	-	0.33	-	1.67	-	21.9	-
$\mathcal{U}_{10} + \mathcal{U}_{30} + \mathcal{U}_{90} \subseteq \mathbb{R}^{100}$	1.30	1.00	61.6	0.34	8.46	0.74	21.9	0.74	20.1	0.86
$\mathcal{N}_{10} + \mathcal{N}_{25} + \mathcal{N}_{50} \subseteq \mathbb{R}^{100}$	1.81	1.00	74.2	0.34	8.87	0.74	7.71	0.88	5.72	0.91
$\mathcal{F}_{10} + \mathcal{F}_{25} + \mathcal{F}_{50} \subseteq \mathbb{R}^{100}$	3.93	1.00	74.2	0.34	18.6	0.70	9.20	0.90	6.77	1.00
$\mathcal{U}_{10} + \mathcal{U}_{80} + \mathcal{U}_{200} \subseteq \mathbb{R}^{800}$	14.3	1.00	715	0.34	120	0.70	1.39	1.00	0.01	1.00
$\mathcal{U}_{900} \subseteq \mathbb{R}^{1000}$	12.8	-	100	-	24.9	-	14.5	-	219	-

Summary

Our LID estimator



qualitatively extracts data complexity with a pre-trained model



- correlates strongly with **PNG Compression** size
- is theoretically grounded



achieves **SOTA** on synthetic data



is the first to scale to extremely high dimensions





is **differentiable** by design.

References

- [1] Bengio, Courville, Vincent. "Representation learning: A review and new perspectives." IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (8):1798–1828, 2013.
- [2] Kamkari, Ross, Cresswell, Caterini, Krishnan, Loaiza-Ganem. "A geometric explanation of the likelihood OOD detection paradox." International Conference on Machine Learning 2024
- [3] Tulchinskii, Kuznetsov, Kushnareva, Cherni- avskii, Nikolenko, Burnaev, Barannikov, Piontkovskaya. "Intrinsic dimension estimation for robust detection of Al-generated texts". Neural Information Processing Systems, 2023
- [4] Ma, Li, Wang, Erfani, Wijewickrema, Schoenebeck, Song, Houle, Bailey. "Characterizing adversarial subspaces using local intrinsic dimensionality." In International Conference on Learning Representations, 2018.
- [5] Ross, Kamkari, Wu, Hosseinzadeh, Liu, Stein, Cresswell, Loaiza-Ganem. "A geometric framework for understanding memorization in generative models." *arXiv preprint arXiv:2411.00113* (2024).
- [6] Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

Thank You!



Paper on **arXiv**



Manifolds website