Geometric Analysis of Generative Models

Their geometric signatures, characterization and controllability through the lens of Differential Geometry

The main focus of my research is to develop a theoretical framework and practical algorithms for analyzing high-dimensional traces of intelligent systems, in particular, modern deep generative models, to investigate and improve their internal mechanisms (*e.g*. defects and signatures) in a principled way.

In recent years, we have seen a rapid development and integration of generative models into our society. Vision generative models like Stable Diffusion [9] and Sora [5] have revolutionized ways to synthesize images and videos with high realism, and Large-Language Models like GPT 3 [6] are assisting the generation of human language in practical applications. Despite this advancement, there is still a big gap in understanding what makes these models behave as they do, and what causal mechanism controls their different behaviors. For example, we still do not fully understand how the composition of training datasets influences downstream model capabilities, how to attribute model behaviors to subcomponents inside the model, and which algorithmic choices really drive performance. Underlying these limitations is the lack of a formal framework that allows us to **represent**, **estimate** and **analyze** their behaviors in their high-dimensional spaces, with both *theoretical guarantee* and *computational efficiency*.

To this end, the primary focus of my research is to develop a mathematical framework and efficient algorithms – using the well-established theory of Differential Geometry – to **(1)** represent the complex behaviors of generative models, **(2)** extract and analyze their geometric signatures, and **(3)** identify causal subcomponents inside the models that control their generative behaviors, with theoretical guarantee in both their capabilities and degenerative behaviors (*e.g*., artifacts, biases). Addressing these problems is crucial for both the scientific understanding and the real-world integration of safe and responsible AI. I approach this goal from the following three directions:

Geometric Signatures of Generative Models: One of the central questions that have surfaced with the development of generative models is, what makes their synthetic data different from natural data (*e.g*., real images taken by camera and documents written by human writers), as well as from synthetic data generated by different models. While this question has been studied as the problem of DeepFake detection [12; 13], its full extension, *i.e*., distinguishing amongst different methods of data synthesis (Model Attribution), remains mostly under-explored.

My work [11; 10] investigates this important, arising problem of model attribution from both theoretical and practical perspectives: In [11], accepted to CVPR 2024, we formalize, for the first time in the literature, the definition of artifact and fingerprint of generative models and propose an effective attribution method to study and compare vision generative models. We conduct extensive experiments on a large array of generative models from all four main families (GAN, VAE, Flow, Score-based) and show the effectiveness of our definition and attribution method, outperforming existing SoTA methods. Furthermore, our method generalizes better to unseen datasets, making it more robust and effective in real-world applications. We believe our definitions will lay an important step towards formalizing the characteristics of generative models and preparing for their integration into our society by helping to develop more effective attribution methods.

Future directions: First, I am working to generalize our current fingerprints of generative models to non-euclidean data manifold by learning Riemannian metric from data [2] and estimating the geodesics [1] to estimate our fingerprints. Our results so far look promising with improved attribution accuracy and generalization to unseen classes of data and generative models. Another direction I'm embarking on is to investigate other geometric signatures (*e.g*., wave or heat signatures [8; 7; 14]) of natural vs. model-generated data, by developing efficient algorithms for the high-dimensional spaces these generative models operate on, as an extension of our methods [11].

Mechanistic Interpretation of Causal Subnetworks in Generative Models: Understanding the inner workings of generative models is critical both for improving their capacities and for aligning their behaviors to human values and safety. Mechanistic interpretability is a prominent line of research that aims to fully specify a neural network's computation, similarly to reverse-engineering a computer program, and decompose a big block-box network into smaller subnetworks whose functions are identifiable and interpretable by humans.

By building up on the previous work of GAN Dissection [4; 3] and our definition of artifacts and fingerprints of GMs [11], our goal is to develop methods that can (1) identify *causal subnetworks* in generative models and (2) *intervene* their pathways to *control* the model behaviors for better value alignment and safety. Theoretically, such method can provide safety bounds on the model's behaviors in new environments, and algorithmically, it can help reprogram a model (without retraining from scratch) to steer their behaviors away from undesirable actions (*e.g*., LLM generating socially biased texts; StyleGAN adding checkerboard artifacts to their images). As a first step, we are working to identify artifact-generating circuits in vision generative models, and possible ways to intervene.Future investigations will include LLMs and multimodal models like Vision-Language Models.

Summary: Going forward, there are many promising and critical open questions to investigate on the behaviors of generative models and the causal mechanisms behind their behaviors. I believe the theory of differential geometry provides rich language and computational tools to develop reliable and efficient methods to represent, estimate, and analyze their behaviors, leading to insights into this new family of AI models that are rapidly changing our daily lives. I believe my approach of blending theory and practice will lead to more reliable, interpretable, and responsible deployment of generative models.

REFERENCES

[1] Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitrios Kalatzis, and Soren Hauberg. Pulling back information geometry. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 4872–4894. PMLR, May 2022.

[2] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: On the Curvature of Deep Generative Models, Dec. 2021.

[3] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a Deep Generative Model, July 2020.

[4] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks, Dec. 2018.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics*, 32(5):1–11, Sept. 2013.

[8] Alexander Grigor'yan, Jiaxin Hu, and Ka-Sing Lau. Heat Kernels on Metric Measure Spaces. In De-Jun Feng and Ka-Sing Lau, editors, *Geometry and Analysis of Fractals*, volume 88, pages 147–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[10] Hae Jin Song and Wael AbdAlmageed. Learning Robust Representations Of Generative Models Using Set-Based Artificial Fingerprints, June 2022.

[11] Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. Manifpt: Defining and analyzing fingerprints of generative models. *arXiv preprint arXiv:2402.10401*, 2024.

[12] Sheng-Yu Wang, O. Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints, Aug. 2019.

[14] Yufan Zhou, Changyou Chen, and Jinhui Xu. Learning Manifold Implicitly via Explicit Heat-Kernel Learning. *Advances in Neural Information Processing Systems*, 33:477–487, 2020.