Geometry of deep learning: A data-driven understanding with neighborhood and graph methods

Sarath Shekkizhar

Abstract

Deep learning has shown great success in learning complex data patterns while being able to make good predictions on unseen data points. However, the limited understanding of these systems hinders further progress and application to several domains in the real world. This predicament is exemplified by the time-consuming model selection, privacy of training data, and explainability of obtained predictions in the presence of adversarial examples and model perturbations.

In this proposal, we propose to study local neighborhood and graph based tools to improve the analysis and design of deep learning. In particular, our approach is aimed at characterizing the mappings induced by a deep neural network based on the input-output geometry of the data used for learning them. Unlike previous works that focus on mathematical modeling or approximations of the components in neural networks, our goal is to develop and lay the foundation for a data-driven language for describing and comparing deep learning models.

Concretely, we will pursue the following tasks. First, we will develop metrics that take into account the local and global data geometry, rather than just considering pairwise distances between points. Second, we will study the stability of deep learning systems using proposed geometric metrics. In particular, we will characterize the connections between optimization and data geometry, and its impact on generalization. Finally, we will make use of the understandings gained to develop approaches for the adaptivity and transferability of deep learning systems.

The proposed research will focus on theoretical and practical aspects of the design and analysis of deep learning. For example, (i) on the theoretical front, we will develop statistical results linking the developed geometric properties to that of data distribution, number of parameters, and the size of training data; (ii) on the algorithmic and practical front, we propose to develop efficient methods for obtaining proposed geometric metrics such that it can be used for large scale parameter and model selection without incurring severe computational overhead.

1 Introduction

In many tasks (e.g., sensing, anomaly detection, classification, and recommendation), systems are increasingly designed by first collecting *significant* amounts of data and optimizing parameters of deep learning models using this data. Further, choices such as architecture, learning paradigm, and other components that make up these systems are based on end-to-end performance of the model on training data. As these systems become more ubiquitous in our everyday interactions, characterization and analysis of these systems are becoming a major challenge for safe and secure deployment. Firstly, to ensure that a range of sectors and professions have the capacity to use deep learning in ways that are useful for them, simple tools that are intuitive and reliable are required to help one make informed and practical decisions. Secondly, for more experienced practitioners developing these systems, we need provable techniques for understanding and designing better models.

Deep neural networks (DNN) are at the core of recent advances and transformative applications in several domains. While state-of-the-art results have been achieved in these domains by using a

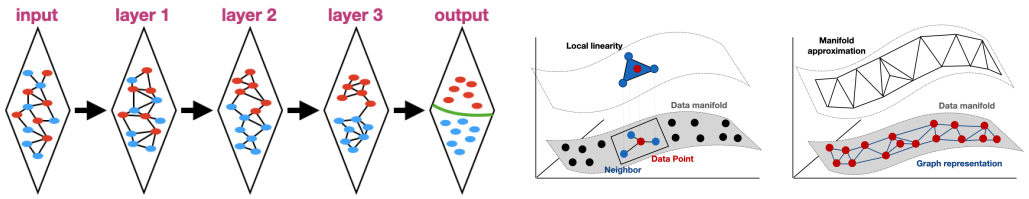

Figure 1: **Left:** Progressive transformation of input feature space over successive layers of a DNN. The samples in the dataset are the same, and thus their attributes (e.g., labels) are the same, but their position in feature space, and hence the graph and neighborhoods, changes as the model is optimized for a particular task. **Right:** Data-driven view of local and global geometry of the embedding manifold using proposed neighborhood and graph representation. Because graphs are intrinsically independent of the exact data position, we can compare observations that are fundamentally heterogeneous (e.g., representations from different dimensional spaces or different models).

performance-driven approach, there are several trends that can cause concern about the reliability of these models. First, an appropriate DNN model is closely tied to the dataset on which it is trained and is selected with significant manual engineering [1]. Second, the large number of parameters involved in DNN introduces stability issues for even small perturbations in their input. Third, it is unclear if a smaller model trained for fewer epochs or with fewer data samples could have achieved the same performance. Often, the main (and in some cases only) justification for a specific choice of model in a system is simply that it works well (in terms of accuracy or other performance metrics) with data selected for evaluation. While this is a very practical perspective that has led to significant advances, a better understanding of the systems is needed, not only for applications where safety is critical (e.g., self-driving vehicles) but also in order to understand limitations in the real world, where it can be exposed to data very different from what was available at training.

The overarching goal of this project is to develop insights into the geometry of input-output mappings learned by DNN using data neighborhood and graph constructions that can approximate the local and global geometry of the data (Figure 1). Our work is motivated by the observation that while DNN involves complex non-linear mappings, the induced transformations and the structure of the representation space can be inferred using data samples. Thus, instead of working with the high dimensional data embedding, we will focus on the relative positions of the data embeddings obtained in DNNs. This allows characterization of data geometry at any layer of the DNN and for the development of techniques and metrics that can provide a quantitative understanding of the system.

One important feature of our approach is that it enables us to compare and contrast the space surrounding a given data point in the embedded space corresponding to any layer of the network and at any stage of training. This implies that (i) it is agnostic to specific training procedures, architecture, and loss functions; (ii) it enables us to compare feature representations of the same data point using different models, even if these representations have different dimensions; (iii) it can be applied to the embeddings obtained with (out of domain or transfer) datasets not used for model training.

2 Data-driven geometrical analysis: Why now?

In the last decade, modern machine learning systems powered by deep learning have led to unprecedented success in many application domains. Three trends that have been driving progress in deep learning: (i) algorithmic innovation that has become increasingly easy to integrate owing to the modularity of components in the model, (ii) the size of trainable model parameters that have far outpaced the size of the data used to train the model, and (iii) the availability of massive amount of compute resources to training these models. In this massively overparameterized regime, deep learning models have the capacity to (over)fit arbitrary training datasets including pure noise. Further, several complex choices in the components and optimization procedures have made *training to zero loss* a feasible target. Contrary to conventional wisdom where such interpolating models (models achieving zero or near-zero training error) were considered to be poor, *some* of the deep learning models exhibit good generalization, i.e., prediction on unseen data [2, 3]. This phenomenon observed with deep learning systems has presented a number of foundational challenges, requiring researchers to revisit and propose new theories for understanding deep learning [4, 5].

Existing theoretical tools, such as model capacity [6], algorithmic stability [7], and regularization [8], predict performance on unseen test data to be close to that on training data, and are therefore unable to account for the unique challenges of overparameterized DNN models [4, 5]. More recently, researchers have developed approximation using well studied mathematical tools (e.g., scattering-networks [9], sparse-convolutional networks [10], NTK [11]) as well as simplifications of components involved in deep learning architectures (for e.g., convex relaxation of non-linear activation functions [12]) to draw similarities and better understand deep learning models. However, these understandings only provide a coarse understanding, often constrained to a single design constraint, with no direct extensions that are adaptable to the constantly evolving landscape of the deep learning systems in terms of architectures, activation functions, optimization strategy, and loss functions.

Alternatively, we propose to understand deep learning using a data-driven perspective. In particular, we explore the geometric properties of the function mapping, rather than considering mathematical models that approximate the parameter or optimization landscape during training of the model [13, 14]. Our work aims to explore the manifold of the data and signals (attributes or functions associated with the data) as observed by a complex learning model by using a graph-based representation of the input and outputs. Our approach abstracts the architecture and components of the model and provides a single framework for comparing and understanding various deep learning models.

Consider a deep learning model that is being trained on data. The dataset itself can be represented by a graph, with the labels in the case of a supervised learning problem modeled as signals on the graph. Then at each stage of the system optimization, data points in the original space are mapped to new values (in some feature space) so that we can now associate a different graph to the same dataset. This idea is illustrated in Figure 1, where we can see the same set of points evolving through different graph representations until they are separable for a classification. Our proposed method allows us to track the evolution of this mapping by measuring the properties of the graph and associated signals.

The idea of using data to understand a machine learning system is not new. In fact, accuracy and other commonly used performance measures are data-driven metrics used to benchmark and compare different machine learning models. Recently, this metric was shown to be effective for understanding the transfer performance on tasks with similar datasets as the training data, but fall short in scenarios with largely different datasets [15, 16]. Further, these empirical evaluations abstract the functional mapping of the model as well as the data used for the evaluation and thus cannot be generalized. That is, we may be able to infer that the difference in accuracy between two systems is significant on a given dataset, but the only way to say something about a model's stability is to characterize, at a much finer level, different regions of input space.

3 Proposal

The notion of manifold has been widely used in machine learning [18, 19, 20]. However, current state-of-the-art methods for deep learning, which rely on high dimensional data representations, have made it increasingly difficult to determine whether the common assumption that data belongs to a smooth manifold does in fact lead to valid insights. Three main challenges arise: (i) developing computationally efficient metrics to quantify local manifold structure, (ii) verifying that these metrics are reliable for very large datasets in high dimensional space, and (iii) incorporating knowledge about the structure of complex feature extractors (e.g., layers and channel structures in deep neural networks) into these metrics and their computation.

Our recently proposed non-negative kernel regression (NNK) graph construction [21, 22] provides key elements to address these challenges and serves as the main building block for our proposed manifold graph metrics (MGMs). NNK graphs are computed locally, with a modest increase in complexity with respect to k-nearest neighbor (kNN) graphs. In contrast with kNN, where neighbors are determined solely based on distance, NNK graphs connect neighboring points that are not geometrically redundant. As a result, the NNK neighborhood for any data point can be described by a polytope whose structure depends on the local data geometry (in particular the local dimension of the data manifold) and is invariant to other factors (data density, number of neighbors chosen to initialize kNN, etc.). While graph-based methods have been proposed to understand latent spaces in DNNs [23], NNK graphs can provide new insights into how the data is organized, for e.g., across channel outputs at a given layer of a convolutional network [24].

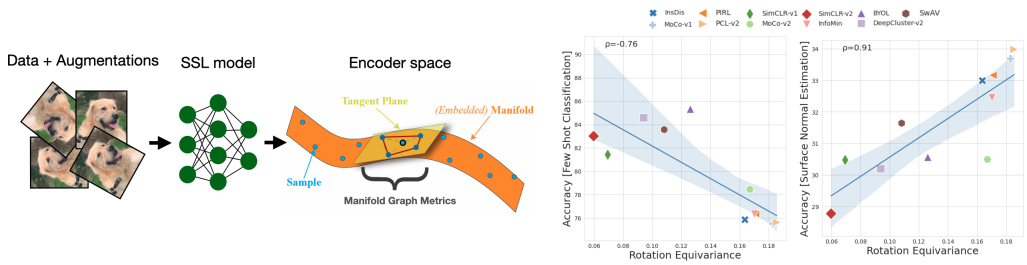

Figure 2: Quantifying rotational invariance/equivariance of various SSL models and its impact on downstream transfer. **Left** For each model, we use as input of the DNN the validation set of ImageNet as well as their augmented versions. The output of the backbone encoder is used to quantify the properties of the manifold induced by the SSL algorithm. Specifically, we develop **Manifold Graph Metrics** that capture manifold properties which are known to be crucial for transfer learning. The MGMs allow us to capture the specificity of each SSL model and to characterize their transfer learning capability. **Right** Correlation plots between measured invariance and performance of the SSL model in few-shot classification on datasets similar to ImageNet [15] and dense surface normal estimation [17]. We extract the features corresponding to various ImageNet class images and their rotated versions for a given SSL model and measure the NNK polytope diameter $\in [0, 2]$ for each input to quantify invariance (small diameter corresponds to collapse of the representations indicative of invariance). As expected, SSL models with rotation invariance perform better in classification (negative correlation) but do worse in surface normal estimation (positive correlation).

We propose to develop MGMs that can be derived from NNK graphs. Some of these metrics will be directly obtained from the set of NNK neighbors (e.g., total number of neighbors per node, diameter or estimated volume of the polytope centered around the node) while others will be properties of local tangent spaces estimated from vectors connecting data points and their NNK neighbors (e.g., dimension, orientation, curvature). Our proposed MGMs will allow us to understand quantitatively the structure of the data manifold and its variation across layers, training epochs, and channels. For example, by computing the properties of local tangent spaces and comparing them across points in the dataset, we can assess the homogeneity of the data space. Changes in these local dimensions during training have been shown to help to detect overfitting and provide insights about generalization and explainability of the model [25, 26].

We propose to use MGMs to assess the quality of a model, including its stability, and robustness to perturbations. To motivate this idea, notice that it is difficult in general to assess parameter stability for complex DNNs: even if we use the same architecture, it is not trivial to compare two DNNs that are trained with different subsets of training data. Instead, we propose to define a notion of **geometric stability** which considers the geometry of the data manifold, quantified using the MGMs under different training conditions. Thus, if $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N$ are data points, we do not require their features to be the same for all trained models. Instead, we determine that the model is stable, and hence reliable, if the local geometry of these points is consistent across all training instances.

Case study: Understanding self-supervised models using manifold properties To illustrate the potential benefits of our approach we consider a case study focused on self-supervised learning (SSL), a set of techniques that have recently empowered vision models to learn meaningful data representations from unlabeled data [27, 28]. SSL learns a representation that is aimed at being invariant to certain image augmentations (e.g., rotation, translation, color-jitter). These models are then used as general-purpose feature extractors for downstream tasks and have been shown to achieve competitive performance relative to to models trained specifically for a task [16]. However, little is understood regarding the capacity of these models: (i) How invariant is the model to different augmentations? and (ii) Which augmentations are crucial for SSL model transfer to a particular task? Our proposed framework can help us demystify the architecture of SSL by shedding light on the interplay between data augmentations, projectors, and the capacity of the feature extractor [29]. As an example, Figure 2 shows one of our MGM metrics that captures the rotation invariance of SSL models and its impact on the transfer performance on two different tasks.

References

[1] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?," in *International Conference on Machine Learning*, pp. 5389–5400, PMLR, 2019.

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations (ICLR)*, 2017.

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[4] M. Belkin, D. J. Hsu, and P. Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate," in *Advances in Neural Information Processing System*, 2018.

[5] M. Belkin, A. Rakhlin, and A. B. Tsybakov, "Does data interpolation contradict statistical optimality?," in *22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019.

[6] M. Anthony, P. L. Bartlett, P. L. Bartlett, *et al.*, *Neural network learning: Theoretical foundations*, vol. 9. cambridge university press Cambridge, 1999.

[7] O. Bousquet and A. Elisseeff, "Algorithmic stability and generalization performance," *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[8] F. Bauer, S. Pereverzev, and L. Rosasco, "On regularization algorithms in learning theory," *Journal of complexity*, vol. 23, no. 1, pp. 52–72, 2007.

[9] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[10] V. Papyan, Y. Romano, and M. Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2887–2938, 2017.

[11] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[12] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[13] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, 2018.

[14] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *International Conference on Machine Learning*, pp. 322–332, PMLR, 2019.

[15] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.

[16] L. Ericsson, H. Gouk, and T. M. Hospedales, "How well do self-supervised models transfer?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021.

[17] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 302–321, 2019.

[18] M. Belkin, *Problems of Learning on Manifolds*. PhD thesis, The University of Chicago, 2003.

[19] A. Gadde, A. Anis, and A. Ortega, "Active semi-supervised learning using sampling theory for graph signals," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[20] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[21] S. Shekkizhar and A. Ortega, "Graph construction from data by non-negative kernel regression," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3892–3896, IEEE, 2020.

[22] S. Shekkizhar and A. Ortega, "Revisiting local neighborhood methods in machine learning," in *Data Science and Learning Workshop (DSLW)*, IEEE, 2021.

[23] C. Lassance, V. Gripon, and A. Ortega, "Representing deep neural networks latent space geometries with graphs," *Algorithms*, vol. 14, no. 2, p. 39, 2021.

[24] D. Bonet, A. Ortega, J. Ruiz-Hidalgo, and S. Shekkizhar, "Channel redundancy and overlap in convolutional neural networks with channel-wise nnk graphs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022.

[25] S. Shekkizhar and A. Ortega, "Model selection and explainability in neural networks using a polytope interpolation framework," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 177–181, IEEE, 2021.

[26] D. Bonet, A. Ortega, J. Ruiz-Hidalgo, and S. Shekkizhar, "Channel-wise early stopping without a validation set via nnk polytope interpolation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

[29] R. Cosentino, A. Sengupta, S. Avestimehr, M. Soltanolkotabi, A. Ortega, T. Willke, and M. Tepper, "Toward a geometrical understanding of self-supervised contrastive learning," *arXiv preprint arXiv:2205.06926*, 2022.