Christoffel-Darboux kernels with applications in deep learning explainability

Abhik Roychoudhury (NUS, Singapore)
Jean-Bernard Lasserre (LAAS-CNRS)
Victor Magron (LAAS-CNRS)

November 15, 2021

Context: Explainability is the extent to which one can explain in "human terms" the behavior of a deep learning system. Explaining the mechanics of deep learning systems would allow to prevent from limiting the potential impact of artificial intelligence. In the context of deep learning, the last layer outputs low-dimensional representation of data [3, 2, 1, 5, 7]. Data alignment ensures that classification can be efficiently performed, for instance with linear transformation. Representation of data happen to be efficient from an empirical point of view, but they are generally considered as black boxes. Our goal is to understand the features of a given representation and explain the behavior of the related network. Prior works focused on explaining the nature of representations by studying the activation of network layers [10] or by recovering the initial data corresponding to a given representation [6]. As an alternative solution, we propose to study the geometry of the representation (dispersion, support inference, latent variety), by means of Christoffel-Darboux kernels.

Goal of the PhD thesis: Lasserre's Hierarchy is a generic tool which can be used to solve global polynomial optimization problems under polynomial positivity constraints [4]. The general idea is to reformulate the initial problem as an optimization problem over probability measures. Recent research [9], investigated the ability of Christoffel-Darboux kernels to capture information about the support of an unknown probability measure. A distinguishing feature of this approach is to allow one to infer support characteristics, based on the knowledge of finitely many moments of the underlying measure. The first investigation track will consist of analyzing the last layer of an existing classification network with the Christoffel-Darboux kernels. A more theoretical goal will be the study Christoffel-Darboux kernels to extend the approach from [8] for measures supported on specific classes of mathematical varieties. In a further step, we intend to apply this framework to deep learning network models, for which latent representation correspond to such low-dimensional varieties. Numerical experiments will be performed on several benchmark suites, including MNIST, CIFAR10 or fashion MNIST.

Potential industrial impact: Deep networks have now become methods of choice in many fields including signal processing, data driven decision systems, natural language processing and many more. Yet their black box nature and our lack of understanding of the mechanisms at stake limitate their use. Basic questions remain out of reach for practioners, including: "Why two similar input provide different outputs?", "What is the semantic information used by the network to make a decision?", "Is there adequation between a given architecture and a given dataset?", "Can a given representation obtained training on a given task be used for another task with minimal tuning?". They are often treated empirically on an ad hoc case by case basis. We will develop unsupervised learning tools with a strong geometrical ground to investigate features of the representation induced by trained networks for applications such as image recognition or natural language processing. We expect to provide new tools for a posteriori qualitative assesment of trained networks by investigation of their geometric properties. We hope to let emerge guidelines and best practices toward a middle term goal of explaining the behaviour of such systems, a crucial challenge regarding their acceptability for important applications such as vision based autonomous systems or natural language based human-computer interactions.

Requirements: A successful candidate will have a strong background in applied mathematics or computer science, having a very good knowledge of probability and statistics as well as a working knowledge

of convex optimization, real analysis and basic measure theory. The candidate is expected to have strong programming skills, be highly motivated and creative.

Funding: This PhD will be funded by DesCartes (A CREATE Programme on AI-based Decision making in Critical Urban Systems), a hybrid AI project between CNRS and Singapore. It will be co-supervised between National University of Singapore (NUS) and LAAS CNRS. The PhD candidate will be hosted in NUS, Singapore.

References

[1] Y. Bengio (2009). Learning deep architectures for AI. Foundations and trends in Machine Learning, 2(1), 1-127.

[2] G.E. Hinton (2007). Learning multiple layers of representation. Trends in cognitive sciences, 11(10):428-434.

[3] G.E. Hinton and R.R. Salakhutdinov (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786):504-507.

[4] J.B. Lasserre (2001). Global optimization with polynomials and the problem of moments. SIAM Journal on optimization, 11(3), 796-817.

[5] Y. LeCun, Y. Bengio and G. Hinton (2015). Deep learning. Nature, 521(7553), 436.

[6] A. Mahendran and A. Vedaldi (2015). Understanding deep image representations by inverting them. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5188-5196).

[7] S. Mallat, (2016). Understanding deep convolutional networks. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065).

[8] E. Pauwels, M. Putinar and J.-B. Lasserre (2019). Data analysis from empirical moments and the Christoffel function. hal-01845137

[9] E. Pauwels and J.-B. Lasserre (2019). The empirical Christoffel function with appication in data analysis. To appear in Advances in Computational Mathematics.

[10] J. Yosinski, J. Clune, T. Fuchs and H. Lipson (2014). Understanding neural networks through deep visualization. In ICML 2014 Workshop on Deep Learning.