Overview

- **Title:** Understanding and exploiting the loss surface of Deep Neural Networks
- **Principal Investigator:** Joan Bruna Estrach. Courant Institute, NYU, Computer Science Department, Mathematics Department and Center for Data Science. Contact: bruna@cims.nyu.edu; cell: 917-767-0617.
- **Google Sponsors:** Jascha Sohl-Dickstein, Koray Kavukcuoglu.
- **Google Contacts:** Oriol Vinyals, Vincent Vanhoucke, Jonathan Tomson.

Proposal Body

Abstract: Training Deep Neural networks is a prime example of a high-dimensional non-convex optimization problem. Despite its massive practical interest, there exist little theoretical guarantees that can inform the choice of the training hyperparameters as well as model arquitectural choices. In this proposal we will study conditions on the data distribution and model architecture that prevent the existence of poor local minima, by first focusing on statistical properties of natural images and specific convolutional architectures.

The conditioning of gradient descent is the next challenge we shall address. We study this question by estimating the geometry of the level sets of any given loss. For that purpose, we introduce an algorithm to estimate the regularity of such sets on large-scale networks. Our initial results suggest that these sets become exponentially more curvy as the energy level decays, in accordance to what is observed in practice. A byproduct of our analysis is the development of learning algorithms that can adapt to the geometry of these sets and provide an extra tool to streamline the training of large models.

Problem Statement: Optimization is a critical component in Deep Learning, governing its success in different areas of computer vision, speech processing and natural language processing. The prevalent optimization strategy is Stochastic Gradient Descent, invented by Robbins and Munro in the 50s. On the one hand, the generalization of SGD is better than one could expect in generic, non-convex loss surfaces. On the other hand, one may wonder if this generic optimization strategy can be adapted to operate in the class of loss surfaces defined by deep neural networks. Some reasons to believe that such adaptation is possible come from the variety of small modifications of SGD algorithms yielding significant speedups [1, 2, 3]. This raises a number of theoretical questions as to why neural network optimization does not suffer in practice from poor local minima. Likewise, it introduces opportunities to leverage training data into adapted optimization.

The loss surface of deep neural networks has recently attracted interest in the optimization and machine learning communities as a paradigmatic example of a hard, high-dimensional, non-convex problem. Recent work has explored models from statistical physics [4], such as spin glasses [5], in order to understand the macroscopic properties of the system, but at the expense of strongly simplifying the nonlinear nature of the model. In this proposal, we do not make any such assumption and study conditions on the data distribution and model architecture that prevent the existence of bad local minima. Together with recent results that rigorously establish that gradient descent does not get stuck on saddle points [6], our first objective will yield guarantees that gradient descent converges to a global optimum in deep rectified networks.

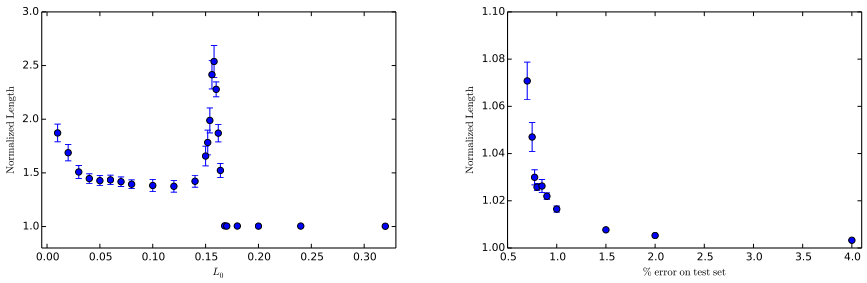

Figure 1: Normalized geodesic length as a function of the energy level for two different models. Left: a "low-dimensional" neural network attempting to fit a cubic polynomial. Right: a convolutional neural network on MNIST. Whereas the cubic fitting displays a heavily non-convex structure at mid-energy values, the MNIST example qualitatively behaves as a convex loss, in which geodesics approach straight lines.

The loss surface $F(\theta)$ of a given model can be expressed in terms of its level sets Ω_λ, which contain all parameters yielding a loss smaller or equal than the energy level λ: $\Omega_\lambda = \{\theta \; ; \; F(\theta) \leq \lambda\}$. A first question we address concerns the topology of these level sets, i.e. under which conditions they are connected. Connected level sets imply that one can always find a descent direction at each energy level, and therefore that no poor local minima can exist. One can verify that linear neural networks (cascaded linear operators without point-wise nonlinearities), despite defining a non-convex loss $F(\theta)$, have connected level sets and therefore can be optimized globally using gradient descent strategies. It is also easy to verify that the non-linear case is not true in general [7], and our first objective is thus to characterize the interplay between data distribution, overparametrization and model architecture that is needed to obtain connected level sets.

Beyond the question of whether the loss contains poor local minima, the immediate follow-up question that determines the convergence of algorithms in practice is the local conditioning of the loss surface. It is thus related not to the topology but to the shape or geometry of the level sets. As the energy level decays, one expects the level sets to reveal more complex irregular structures, which correspond to regions where $F(\theta)$ has small curvature.

In order to construct meaningful theoretical models, we have developed an algorithm that efficiently estimates these topological and geometrical properties on large deep architectures and datasets. Inspired by [8], given two parameter values θ_1 and θ_2 lying in the border of the same level set Ω_λ, our algorithm constructs a geodesic $\gamma(t)$ in Ω_λ relying θ_1 to θ_2 using a dynamic programming procedure. An example of the resulting characteristics is illustrated in Figure 1.

Besides improving our understanding of stochastic optimization in the class of neural networks, our second main objective is to leverage that understanding by developing optimization strategies adapted to the class of loss surfaces at hand, i.e. learning how to learn [9]. In particular, we shall construct targets for optimization based on the geodesic paths described above: given two parameter values θ_a, θ_b with $F(\theta_b) \leq F(\theta_a) = \lambda$ (for example obtained during the course of a vanilla optimization method at iterations t and $t+T$ respectively), a natural target to update θ_a is to follow the geodesic $\gamma(t)$ in Ω_λ relying θ_a to θ_b. The extent by which this geodesic is smooth will determine how easy it is to predict from past and present information available at θ_a. We will explore predictive models based on recurrent neural networks, inspired from [9].

Data Policy

This proposal requires mostly unlabeled data in the form of images, speech and audio signals, and text corpus. The initial plan is to use publicly available datasets. The research outcomes will be both conference and journal papers, and also publicly available code.

Budget

- Student salary: $50000 per year.
- Student conference/workshop travel: $6000 per year (three conferences a year).

References

[1] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[2] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a overview of mini–batch gradient descent."

[3] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[4] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Advances in neural information processing systems*, 2014, pp. 2933–2941.

[5] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proc. AISTATS*, 2015.

[6] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent converges to minimizers," *University of California, Berkeley*, vol. 1050, p. 16, 2016.

[7] O. Shamir, "Distribution-specific hardness of learning neural networks," 2016.

[8] I. J. Goodfellow, O. Vinyals, and A. M. Saxe, "Qualitatively characterizing neural network optimization problems," *arXiv preprint arXiv:1412.6544*, 2014.

[9] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, and N. de Freitas, "Learning to learn by gradient descent by gradient descent," *arXiv preprint arXiv:1606.04474*, 2016.