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, Yoram Singer, 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 the model arquitecture in such non-convex scenarios. In this proposal we will study conditions on the data distribution and model architecture that prevent or limit 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 performance of SGD is better than one could expect in generic, arbitrary 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 modifications of SGD algorithms yielding significant speedups [4, 7, 8, 9]. 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 [3], such as spin glasses [2], 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 [10], our first objective will yield guarantees that gradient descent converges to a global optimum in deep rectified networks.

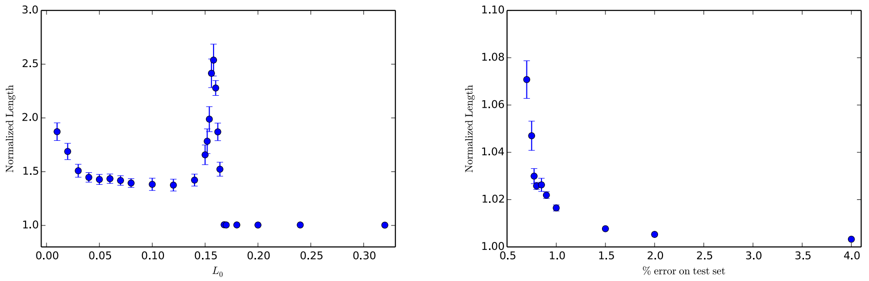

Figure 1: Normalized geodesic length as a function of the energy level for two different models. Left: a "low-dimensional" neural network attempting to fit a cubic polynomial. Right: a convolutional neural network on MNIST. Whereas the cubic fitting displays a heavily non-convex structure at mid-energy 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 θ 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 [5]. It is also easy to verify that the non-linear case is not true in general [11], 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 [5].

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 [5] that efficiently estimates these topological and geometrical properties on large deep architectures and datasets. Inspired by [6], 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. This tool may help us to construct good mathematical models that explain the behavior of SGD on such large convolutional neural networks, especially the interplay between overparametrization and regularity of the sets Ω_λ.

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 [1]. 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 it from past and present information available at θ_a. We will explore predictive models based on recurrent neural networks, inspired from [1].

Data Policy

This proposal requires mostly labeled datasets in the form of images, text corpus, as well as synthetically generated examples. We plan to use publicly available datasets and to release our code implementation for reproducibility.

The research outcome will be both conference and journal papers, and also publicly available code on `github.com`. Our current implementation is in TensorFlow, running on both CPU and GPU machines at UC Berkeley. Our algorithm is prone to parallelization, and thus we would be interested in hardware infrastructures that can leverage it.

Budget

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

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. *arXiv preprint arXiv:1606.04474*, 2016.

[2] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Proc. AISTATS*, 2015.

[3] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

[4] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[5] Daniel Freeman and Joan Bruna. Topology and geometry of deep rectified network optimization landscapes. *manuscript in preparation*, 2016.

[6] Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

[7] Geoffrey Hinton, N Srivastava, and Kevin Swersky. Lecture 6a overview of mini–batch gradient descent.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[9] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.

[11] O. Shamir. Distribution-specific hardness of learning neural networks. 2016.

Joan Bruna

Contact Information

+1 917 767 0617
bruna@cims.nyu.edu
http://cims.nyu.edu/~bruna/

Citizenship

Spanish

Research Interests

Invariant Signal Representations, Pattern recognition, Harmonic analysis, Stochastic Processes, Machine Learning, Deep Learning.

Education

Ecole Polytechnique, Palaiseau, France

Ph.D. in Applied Mathematics, "Scattering Representations for Recognition", 2008-2013.

- Thesis Topics: *Invariant signal representations, classification, pattern recognition, stochastic processes, invariance learning, differential geometry.*
- Adviser: Professor Stéphane Mallat.

Ecole Normale Superieure, Cachan, France

MSc "*Mathématiques, Vision, Apprentissage*" in Applied Mathematics, 2004-2005.

- Image, audio and video processing, Harmonic analysis, Machine Learning, Wavelet theory.
- *Mention Très bien.*

Universitat Politècnica de Catalunya, Barcelona, Spain

BS, MSc Telecommunications, 1999-2004,

- Master Thesis developed in Nokia Denmark (Aalborg), 2003-2004.
- Title: "New Modeling techniques for HSDPA/WCDMA" *(A with honors).*

BS Mathematics, 1998-2002.

- *With Honors, Ranked 2nd.*

Professional Experience

Courant Institute, NYU, New York, NY

Assistant Professor, Department of Computer Science,
Department of Mathematics (affiliated) and Center for Data Science, **Fall 2016**

UC Berkeley, Berkeley, CA

Assistant Professor, Dept of Statistics, **Jan 2015 (on leave)**

Facebook AI Research, New York, NY

Post Doctoral Fellow, AI Research, **Oct-Dec 2014**

Courant Institute, NYU, New York, NY

Postdoctoral Researcher **Oct 2012 - Sep 2014**

- Invariance Learning, Deep learning architectures, scattering representations of stochastic processes
- Applications to Speech recognition, Pattern recognition, Texture synthesis, Bioacustics.
- Supervisor: Professor Yann LeCun.

CSR (through acquisition), Malakoff, France

Senior Reseach Consultant **Oct 2010 - July 2012**

- Conception and supervision of real-time algorithms for video processing applications: frame-rate conversion, 2d to 3d conversion.
- Machine Learning, Statistical Model Selection, Non-supervised learning for robust motion estimation.

Zoran (through acquisition) , Malakoff, France

Senior Reseach Engineer **June 2008 - Sep 2010**

- Conception, development and implementation of real-time algorithms for several video processing applications: frame-rate conversion, 2d to 3d conversion, denoising.
- Geometric Spatio-temporal representations, Sparse evolutive models, multiscale representations, real-time algorithm design.

Let it Wave, Palaiseau and Malakoff, France.

Research Engineer **April 2005 to June 2008**

- Algorithmic research on Video Super-Resolution, Denoising, and Deinterlacing
- Conception and development of real-time video up-converter based on spatio-temporal geometric processing.

Preprint Papers

Freeman, D., Bruna, J "Topology and Geometry of Deep Rectified network optimization landscapes", *preprint*, 2016.

Bruna, J., Mallat, S. "Max-Entropy Gaussianization by Multiscale Scattering", *Preprint*, 2016.

Dokmanic, I. , Bruna, J. Mallat, S. De Hoop, M. "Inverse Problems with Invariant Multiscale Statistics", *Submitted*, 2016.

Bronstein, M., Bruna, J., Szlam, A. LeCun, Y, Vandergyst, P. "Geometric Deep Learning: going beyond Euclidean data", *Submitted*, 2016.

Bruna, J., Moreau, Th, "Adaptive acceleration of Sparse Coding via Matrix Factorization", *Preprint, arxiv 1609.00285*, 2016.

Published Papers

See https://scholar.google.com/citations?user=L4bNmsMAAAAJ&hl=en.

Patents

See https://cims.nyu.edu/~bruna

Awards

- ENS Cachan Scolarship, 2004-2005
- Honors Mention in Spanish Physics Olympiad, 1998
- 2nd national place in Cangur Mathematical Contest, 1996

Collegial

- JMLR, ACHA, NIPS, EECV, IEEE TPAMI, Annals of Statistics, IEEE TIT, ICLR, CVPR, PNAS reviewer (2013-present).
- Co-organizer of Machine Learning Summer School, Cadiz, 2016.
- Co-organizer of "Mathematics of Deep Learning" tutorial: ICCV, 2015, CVPR 2016.

Language Skills

Catalan: Mother tongue **Spanish**: Bilingual
French: Bilingual **English**: Bilingual **Japanese**: Beginner level (2 years)

Miscellaneous

Travelling, Food, Photography, Literature, New Wave music.