Optimization algorithms for sparse deep networks

Keywords. Deep neural networks; sparse regularization; proximal optimization; matrix factorization

Supervisor. Rémi Gribonval (remi.gribonval@inria.fr)

Location. DANTE Team, LIP, ENS de Lyon

1 Context

Learning with neural networks empirically leads to state of the art results for many tasks in computer vision (image segmentation, image classification, etc.), as well as in many research domains (signal processing, robotics, etc) under the umbrella of what is now called "artificial intelligence",

Yet, today's best deep networks can have up to billions of parameters and their empirical success typically requires two ingredients: 1) very large training datasets and 2) substantial computing power. This practically limits the applicability of these techniques in fields where data is scarce, or in scenarios where energy consumption or training time become crucial resources.

To circumvent these bottlenecks, a possible alternative is to consider *sparse* deep networks, where only few network parameters are nonzero. Indeed, sparsity plays the role of a proxy both for the memory required to store such networks and for the complexity of computing the network's output given its input.

As an example, recent work [1] has empirically shown that fast linear transforms such as the Hadamard transform can be reverse-engineered by optimizing *linear* sparse deep networks using so-called proximal algorithms, with guaranteed convergence to stationary points of the objective function.

2 Goals

The goal of this internship is twofold: to adapt these proximal algorithms to the complex-valued case; to explore these approaches to train *nonlinear* sparse deep networks, with a focus on the ReLU nonlinearity.

Proximal algorithms for multilayer sparse matrix factorizations yield approximations of a given matrix $\mathbf{A}$ as a product of L sparse factors, $\mathbf{A} \approx \prod_{\ell=1}^{L} \mathbf{S}_\ell$. Optimizing one factor, the other ones being fixed, is an instance of a sparsity constrained linear inverse problem, which has been widely studied [2] and successfully addressed with iterative proximal methods [3] either based on convex or nonconvex sparsity-promoting penalties. Proximal factorization approaches have been empirically explored with success to reverse-engineer real-valued linear transforms [1], yet many important transforms such as the Fourier transform are indeed complex-valued.

A primary objective is to extend these algorithms (and possibly their convergence analysis) to the complex case, before exploring their adaptation to train sparse deep networks with the ReLU nonlinearity.

In a first step, the intern will get familiar with the concepts and tools of proximal optimization [3], sparse regularization for linear inverse problems [2], and multilayer sparse factorization [1, 4] via a bibliographic study. In order to establish a testbed and baseline for experiments, getting acquainted with the FAμST library (Python and Matlab interfaces, `https://faust.inria.fr`) developed in the team will be needed, as well as with a standard deep learning tool such as pytorch. Then, the intern will propose, implement, and test modified optimization algorithms, starting with the toy problem of reverse-engineering the Fast Fourier Transform. Standard datasets for large-scale learning [5][1] [6] will serve as a testbed for the resulting algorithms. Successful work is expected to lead to a paper submission.

Further information: Please contact Rémi Gribonval for more information. The intern can receive a "gratification" if needed and continuation as a PhD is possible.

[1] `https://www.cs.toronto.edu/~kriz/cifar.html`

References

[1] Luc Le Magoarou and Rémi Gribonval. Flexible multi-layer sparse approximations of matrices and applications. *IEEE J. Selected Topics in Signal Processing*, 2016.

[2] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, May 2012.

[3] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, May 2010.

[4] Luc Le Magoarou. *Matrices efficientes pour le traitement du signal et l'apprentissage automatique*. PhD thesis, INSA de Rennes, November 2016.

[5] A Krizhevsky and G Hinton. Learning multiple layers of features from tiny images. 2009.

[6] Wei Dong, Richard Socher, Li Li-Jia, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

When are sparse deep networks unique and optimal ?

Keywords. Mathematical optimization; sparse regularization; matrix factorization; neural networks

Supervisor. Rémi Gribonval (remi.gribonval@inria.fr)

Location. DANTE Team, LIP, ENS de Lyon

1 Context

Learning with neural networks empirically leads to state of the art results for many tasks in computer vision (image segmentation, image classification, etc.), as well as in many research domains (signal processing, robotics, etc) under the umbrella of what is now called "artificial intelligence",

Yet, today's best deep networks can have up to billions of parameters and their empirical success typically requires two ingredients: 1) very large training datasets and 2) substantial computing power. This practically limits the applicability of these techniques in fields where data is scarce, or in scenarios where energy consumption or training time become crucial resources.

To circumvent these bottlenecks, a possible alternative is to consider *sparse* deep networks, where only few network parameters are nonzero. Indeed, sparsity plays the role of a proxy both for the memory required to store such networks and for the complexity of computing the network's output given its input.

As an example, recent work [1] has empirically shown that fast linear transforms such as the Hadamard transform can be reverse-engineered by optimizing *linear* sparse deep networks using so-called proximal algorithms, with guaranteed convergence to stationary points of the objective function.

2 Goals

The goal of this internship is to explore from a mathematical perspective the optimization problems and algorithms involved in the training of sparse deep networks. This is expected to yield a better understanding of their conditions of success, and to possible adaptations with improved performance.

Proximal algorithms for multilayer sparse matrix factorizations yield approximations of a given matrix $\mathbf{A}$ as a product of L sparse factors, $\mathbf{A} \approx \prod_{\ell=1}^{L} \mathbf{S}_\ell$. Optimizing one factor, the other ones being fixed, is an instance of sparsity constrained linear inverse problems which have been widely studied [2] and successfully addressed with iterative proximal methods [3]. These methods are based on convex (ℓ^1 norm) or nonconvex (ℓ^0 pseudo-norm) sparsity-promoting penalties, and are endowed with a solid mathematical understanding of their conditions of success. In contrast, despite empirical successes [1], little is known about the conditions of success of sparse factorization approaches even for two-layer sparse factorization.

A primary objective is to study conditions under which a matrix admits a *unique* two-layer sparse factorization $\mathbf{A} = \mathbf{S}_1\mathbf{S}_2$, up to natural equivalence classes corresponding to permutation and scaling of the rows (resp. columns) of $\mathbf{S}_1$ (resp. of $\mathbf{S}_2$). The extension of uniqueness to more layers, in the spirit of [4, Chapter 7], will then be investigated before characterizing properties of the local and global optima of the underlying cost functions in order to shed light on the optimization landscape of sparse deep networks, either linear or with the ReLU nonlinearity.

In a first step, the intern will get familiar with the concepts and tools of sparse regularization for linear inverse problems [2], multilayer sparse factorization [1, 4], and on the optimization landscape of deep networks [5, 6, 7] via a bibliographic study. To support the mathematical exploration with empirical experiments and illustrations, the intern is encouraged to get acquainted with software tools for deep learning such as pytorch, and with the FAμST library (Python and Matlab interfaces, `https://faust.inria.fr`) developed in the team for multilayer sparse factorization. Successful work is expected to lead to a paper submission.

Further information: Please contact Rémi Gribonval for more information. The intern can receive a "gratification" if needed and continuation as a PhD is possible.

References

[1] Luc Le Magoarou and Rémi Gribonval. Flexible multi-layer sparse approximations of matrices and applications. *IEEE J. Selected Topics in Signal Processing*, 2016.

[2] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, May 2012.

[3] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, May 2010.

[4] Luc Le Magoarou. *Matrices efficientes pour le traitement du signal et l'apprentissage automatique*. PhD thesis, INSA de Rennes, November 2016.

[5] Alexis Benichoux, Emmanuel Vincent, and Remi Gribonval. A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors. In *ICASSP 2013*, pages 6108–6112, Vancouver, Canada, March 2013. IEEE.

[6] Luca Venturi, Afonso S Bandeira, and Joan Bruna. Neural Networks with Finite Intrinsic Dimension have no Spurious Valleys. *CoRR*, 2018.

[7] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask. May 2019.