Frugal Models and Algorithms for Machine Learning

PhD position 2021–2024

LIS and I2M, Aix-Marseille University, France

Context and objectives

The data deluge and the recent trends in machine learning result in the explosion in the size of the models, with possibly hundreds of billions of parameters [Brown et al., 2020]. Consequences of this phenomenon include major concerns: the difficulty to control those models in terms of design, training, interpretation, security; the need for large computational resources, for training but also for making predictions; the environmental impact that reaches unsustainable levels [Strubell et al., 2019].

The objective of this PhD project is to propose frugal models that are able to handle large volumes of data with efficiency while being structured to provide a reduced time and space complexity. As opposed to distillation techniques [Hinton et al., 2015] that are applied after training, the target structures are intrinsic to the proposed models.

Either in deep neural networks or in other machine learning models, the space and time complexity is mainly due to the linear part of the models, involving large matrices or tensors of data and parameters. A key challenge is to reduce this particular aspect.

Appart from the well-known low-rank approaches [Mishra et al., 2013], one promising strategy is to decompose a large $N \times N$ matrix into a product of sparse matrices. Such models, named Flexible Multilayer Sparse Approximations [Le Magoarou and Gribonval, 2016] or butterfly factorizations [Dao et al., 2019, Vahid et al., 2020], inherit the structures of the fast transforms like the fast Fourier or Hadamard transforms. Consequently, their typical complexity is in $\mathcal{O}(N \log N)$ in space and in time for matrix-vector multiplications. Preliminary works have shown how to leverage such models to revisit K-means in a frugal way [Giffon et al., 2021] and to use them for compressing neural networks [Giffon, 2020]. However, such models have not been well controlled so far, both in their ability to model arbitrary data and in their training procedure [Cheukam Ngouonou, 2020, Le Quoc, 2020, Zheng, 2020].

The expected works should contribute to the following research directions:

- proposing new frugal models, e.g., based on sparse/butterfly factorizations;
- studying the properties of such models (expressivity, frugality);
- developing learning algorithms for those models, with techniques including convex, non-convex and combinatorial optimization;
- deploying such models in machine learning models (neural networks, kernel machines, and so on), tasks and use cases.

Supervision and research group description

The PhD project is located in Marseille, France. It will be mainly supervised by Valentin Emiya within QARMA team at LIS lab, and cosupervised by Caroline Chaux at I2M lab.

Main supervisor: Valentin Emiya, LIS, Aix-Marseille University. V. Emiya has developed an expertise in machine learning and signal processing. He has been a member of the Équipe d'Apprentissage de Marseille (QARMA) at Laboratoire d'Informatique et Systèmes (LIS) since 2011. In the last five years, Valentin Emiya has been the supervisor of 3 postdocs, 3 PhD candidates and 16 interns. He was the principal investigator of 4 projects including ANR JCJC MAD and had about 25 collaborators. He is currently preparing his Habilitation à Diriger les Recherches. He also dedicates a large effort on science popularization, as a leader of the Treize Minutes Marseille project since 2013. V. Emiya is currently co-supervising Marina Kreme's PhD on time-frequency inpainting (2017-2021) and Raphael Sturgis' PhD on the optimization of vessel trajectories by machine learning techniques (2019-2022, with company Searoutes).

Co-supervisor: Caroline Chaux, I2M, Aix-Marseille University. C. Chaux has developed an expertise in convex optimization, sparse representations and signal processing. She is a member of the Équipe Signal et Image (SI) at Institut de Mathématiques de Marseille (I2M) since 2012. In the last five years, Caroline Chaux has been the supervisor of 1 postdoc, 3 PhD candidates and 4 interns. She defended her Habilitation à Diriger les Recherche (HdR) in January 2019. She is the principal investigator of the Amidex project Bifrost. For three years, she has been a member of the ANR scientific evaluation comittee on signal processing. C. Chaux is co-supervising Marina Kreme's PhD with V. Emiya and B. Torrésani.

Research team. QARMA is composed of about 10 permanent and 15 non-permanent members with complementary skills conducting research works in machine learning, including theory, algorithms and applications. It is part of the Data Science department of Laboratoire d'Informatique et Systèmes (LIS), which hosts about 375 members in 21 teams.

Application and important dates

Applicants should have excellent general skills in maths and computer science, ideally with some expertise in machine learning, optimization and signal processing. Some science popularization tasks will also be attached to this position.

Applications must be sent to Valentin Emiya and Caroline Chaux (firstname.lastname@univ-amu.fr). They should include a CV, a motivation letter, detailed grades at least for Bachelor and Master degrees, recommendation letters and/or contact of referees, possible scientific reports from past projects and any additional useful document.

The timeline is tight, with the following expected dates (subject to changes).

- May 2nd: application deadline (interviews may be conducted before this date, on the fly when application are received)
- May 10th: end of interviews, notification of results
- May 17th: final candidates' answer
- October 1st: start of the PhD

Candidates are invited to contact V. Emiya and C. Chaux at any time to obtain more details.

References

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165 [cs].

[Cheukam Ngouonou, 2020] Cheukam Ngouonou, J. (2020). Apprentissage de transformées rapides par décomposition en produit de facteurs 2-uniformes. Master's thesis, Aix-Marseille Univ.

[Dao et al., 2019] Dao, T., Gu, A., Eichhorn, M., Rudra, A., and Ré, C. (2019). Learning fast algorithms for linear transforms using butterfly factorizations. In *Proc. of Int. Conf. on Machine Learning (ICML)*.

[Giffon, 2020] Giffon, L. (2020). *Approximations parcimonieuses et méthodes à noyau pour la compression de modèles d'apprentissage*. PhD thesis, Aix-Marseille Université.

[Giffon et al., 2021] Giffon, L., Emiya, V., Ralaivola, L., and Kadri, H. (2021). QuicK-means: Acceleration of K-means by learning a fast transform. *Machine Learning*, to appear.

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *Proc. of NIPS Workshop on Deep Learning*.

[Le Magoarou and Gribonval, 2016] Le Magoarou, L. and Gribonval, R. (2016). Flexible Multilayer Sparse Approximations of Matrices and Applications. *IEEE Journal of Selected Topics in Signal Processing*, 10(4).

[Le Quoc, 2020] Le Quoc, T. (2020). Multilayer Sparse Matrix Factorization. Master's thesis, ENS de Paris-Saclay.

[Mishra et al., 2013] Mishra, B., Meyer, G., Bach, F., and Sepulchre, R. (2013). Low-Rank Optimization with Trace Norm Penalty. *SIAM Journal on Optimization*, 23(4):2124–2149.

[Strubell et al., 2019] Strubell, E., Ganesh, A., and Mccallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In *Proc. of Annual Meet. of the Association for Computational Linguistics (ACL)*, pages 3645–3650, Florence, Italy.

[Vahid et al., 2020] Vahid, K. A., Prabhu, A., Farhadi, A., and Rastegari, M. (2020). Butterfly Transform: An Efficient FFT Based Neural Architecture Design. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[Zheng, 2020] Zheng, L. (2020). Identifiability in Matrix Sparse Factorization. Master's thesis, ENS de Paris-Saclay.