Exploiting Low-Rank structures in Deep Neural Networks

Emanuele Zangrando, Francesco Tudisco

Abstract

Neural networks have revolutionized numerous fields, demonstrating extraordinary success across a wide range of applications. However, their substantial memory footprint and high computational demands can make them impractical for deployment in resource-constrained environments, where hardware and energy limitations pose significant challenges.

In recent years, a growing body of empirical evidence, beginning with works such as [1, 2], has shown that modern neural networks contain a striking degree of redundancy in their parameters. Despite this, contemporary experience suggests that sparse and low-rank architectures derived from pruning are difficult to train from scratch, limiting the ability to fully leverage their computational advantages during early training stages.

While sparsity often needs to be enforced explicitly within neural network parameters, low-rankness interestingly frequently emerges naturally in deep networks [6], suggesting it as a more natural structure to impose. Understanding the conditions under which this low-rank implicit bias manifests is of crucial importance, as it enables us to predict a priori when a compressed model can perform as well as a much larger one. Although considerable progress has been made in this direction, a comprehensive theoretical characterization remains missing. In this direction, in [7] the authors establish a connection between the rank of matrices at stationary points and the total cluster variation of intermediate embeddings, linking low-rank bias to a phenomenon known as deep neural collapse [8, 9, 10], suggesting a tendency towards small rank matrices.

In this talk, we will first discuss the implicit bias of deep neural networks toward low-rank structures and then highlight their practical significance in large-scale applications such as model compression, fine-tuning, and stability.
Regarding applications, we will touch on different methods for constructing highly efficient low-rank representations of neural networks, such as methods leveraging variational principles [3, 4], and methods more rooted in optimization.

Lastly, we will introduce a novel perspective on low-rank compression, reformulating it as a bilevel optimization problem [5]:

$$\begin{aligned} &\min_{s\in\mathbb{R}^r:\|s\|\leq\tau} f_1(s) = L_1(U\,\mathrm{diag}(s)V^\top), \\ &\text{s.t.} \quad (U,V) \in \arg\min_{U\in\mathbb{R}^{n\times r},V\in\mathbb{R}^{m\times r}} f_2(U,V) = L_2(U\,\mathrm{diag}(s)V^\top), \end{aligned}$$

where f_1, f_2 are two loss functions, typically calculated on subsets of the original dataset. One of the central challenges in approximating solutions of this optimization problem is computing ∇f_1, since it depends on the solution of a linear system involving second-order information $\nabla^2 f_2$, effectively rendering first-order iterative methods as already computationally prohibitive.
To address this, we propose an algorithm based on conditional gradient methods, where an approximation of $\nabla f_1(s)$ is derived in closed form as a function of the minimizer of the lower-level problem.

Along with this, a series of numerical results will be presented, comparing different approaches for large-scale applications.

References

[1] J. Frankle, M. Carbin. *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks.* ICLR 2019.

[2] V. Lebedev, Y. Ganin, M. Rakhuba, I. Oseledets, V. Lempitsky. *Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition.* ICLR 2015.

[3] O. Koch, C. Lubich. *Dynamical low-rank approximation.* SIMAX 2007.

[4] S. Schotthöfer, E. Zangrando, J. Kusch, G. Ceruti, F. Tudisco. *Low-rank lottery tickets: finding efficient low-rank neural networks via matrix differential equations.* NeurIPS 2022.

[5] E. Zangrando, S. Venturini, F. Rinaldi, F. Tudisco. *dEBORA: Efficient Bilevel Optimization-based low-Rank Adaptation.* ICLR 2025.

[6] T. Galanti, Z. S. Siegel, A. Gupte and T. A. Poggio. *SGD and Weight Decay Secretly Minimize the Rank of Your Neural Network.* NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning.

[7] E. Zangrando, P. Deidda, S. Brugiapaglia, N. Guglielmi, F. Tudisco. *Neural Rank Collapse: Weight Decay and Small Within-Class Variability Yield Low-Rank Bias.* arXiv preprint, 2024.

[8] A. Rangamani, M. Lindegaard, T. Galanti, T. A. Poggio. *Feature learning in deep classifiers through Intermediate Neural Collapse.* ICML 2023.

[9] P. Súkeník, C. H. Lampert, M. Mondelli. *Neural collapse vs. low-rank bias: Is deep neural collapse really optimal?.* NeurIPS 2024.

[10] P. Súkeník, M. Mondelli, C. H. Lampert. *Deep Neural Collapse Is Provably Optimal for the Deep Unconstrained Features Model.* NeurIPS 2023.