EXPLOITING LOW-RANK GEOMETRY IN DEEP LEARNING

FRANCESCO TUDISCO

As model and data sizes continue to expand, modern AI faces pressing questions about timing, costs, energy consumption, and accessibility. In response, there has been a surge of interest in network compression techniques aimed at mitigating computational costs while preserving model performance. While many existing methods focus on post-training pruning to reduce inference costs, a subset addresses the challenge of diminishing training overhead, with layer factorization emerging as one of the prominent approaches. In fact, a variety of empirical and theoretical evidence has recently shown that deep networks exhibit a form of low-rank bias, hinting at the existence of highly performing low-rank subnetworks.

This talk will focus on our recent work on analyzing and leveraging implicit low-rank bias for efficient model compression in deep learning. Taking advantage of the Riemannian geometry of the low-rank format, we devise a geometry-aware variation of SGD to train small, factorized network layers while simultaneously adjusting their rank. We provide theoretical guarantees of convergence and approximation capabilities together with experimental evaluation showing competitive performance across various moderate-size network architectures.