Compressing neural networks: sparsity, quantization, and low-rank approximation

Rayan Saab

UCSD

We will discuss recent advances in the compression of pre-trained neural networks using computationally efficient algorithms. Our approaches leverage sparsity, low-rank approximations of weight matrices, and weight quantization to achieve significant reductions in model size, while maintaining performance. We provide rigorous theoretical error guarantees for our methods as well as numerical experiments that demonstrate the practical effectiveness and efficiency of our techniques.

In Special Session 1 on Applied Harmonic Analysis and Operator Theory at the Mathematical Congress of the Americas 2025.

MCA 2025 internal code: SS01-T14