Enabling Extreme Model Compression through Multi-Objective Optimization

Dara Varam (Advised by Dr. Mohamed I. AlHajri)

A thesis presented to the **American University of Sharjah** (College of Engineering) in partial fulfillment of the requirements for the degree of **Master of Science in Machine Learning**.

Abstract

The deployment of resource-efficient models on hardware remains a key area of research in modern-day deep learning (DL). With the continuous increase in scale of novel DL models, the challenge of compression whilst maintaining good performance is non-trivial. Several techniques, such as network pruning, quantization, low-rank factorization and knowledge distillation have been devised over the years, yet each method poses inherent trade-offs between compressibility and performance. This thesis considers model compression through a multi-objective optimization (MOO) perspective, allowing for the simultaneous optimization of both a task objective and a compression objective. By incorporating compression into the training curriculum directly, the proposed method uses loss functions such as cross-entropy and negative-log-likelihood for classification tasks, and other losses that serve as proxies for sparsity, compressibility and knowledge alignment. MOO overcomes the drawbacks of the standard weighted sum approach typically seen in the literature by ensuring that each update step for the weights is taken in the direction of common descent for all objectives. Preliminary results using pruning as the compression technique (with L_1 loss as the compression objective) show both one-shot and finetuned performances significantly outperforming existing weighted sum-based methods on standard benchmarking datasets and models. The current framework is designed to be interchangeable with any compression loss, allowing for easy extensions to quantization-based compression, knowledge distillation-based compression, and beyond, with applications in both generative and predictive DL modeling.