Resource-Aware and Resilient Learning in Mobile Networks

Giuseppe Di Giacomo

In the last decade, Machine Learning (ML) and, more specifically, Deep Learning have gained significant prominence, driving innovation across a variety of domains. ML models are commonly trained in a centralized manner, requiring extensive computational resources and large volumes of data to be aggregated on a central server, which also raises privacy concerns. To tackle these issues, the Cooperative Learning (CL) paradigm has emerged as a promising solution: it leverages the computational capabilities of a plethora of heterogeneous nodes, enabling local training while exchanging the models' parameters, without sharing sensitive raw data with third parties.

Many factors characterize the cooperative training, such as the choice of the model, the data to use and the nodes to employ; making decisions about these aspects is challenging, as they affect one another. The first part of this work presents two distinct CL scenarios. In the first one, the layers of a Deep Neural Network (DNN), or multiple instances of them, are run at different devices of the mobile-edge-cloud continuum. The second scenario focuses on sequential learning across sets of nodes and capitalizes on pruning, a well-established technique to compress DNNs, which requires additional decisions about when and how much to prune the model during the training. Importantly, each node trains all the layers of the model, whether full or pruned. For each scenario, we design an algorithmic framework that, given the interdependencies among the abovementioned factors, makes joint decisions about them to optimize the training energy consumption while meeting time and quality constraints. The proposed frameworks have polynomial time complexity and are proven to make near-optimal decisions, outperforming alternative methods, as validated through our extensive performance evaluation. Another key challenge in CL lies in the lack of incentive for nodes to participate in the learning, as they will not allocate their computational and communication resources for training unless it is beneficial to them. Thus, to foster cooperation among nodes, we develop a game-theoretic approach based on the Generous Tit-for-Tat strategy. The designed method, which accounts for

the heterogeneity of both nodes' resources and models to be trained, constitutes a Nash equilibrium and converges to the Pareto optimal operating point.

In the second part of this thesis, we focus on Generative Artificial Intelligence models owing to their ability to generate high-quality synthetic samples, which have been demonstrated to improve the performance of DNNs. In line with this, we first design a diffusion model for Novel Views Synthesis, endowed with the capability of accommodating a flexible number of input views of an object to generate multiple novel views thereof. The model outperforms the alternative state-of-the-art methods by up to 15% in terms of perceptual similarity between the generated and ground truth views. Finally, we employ a smaller-scale model with a similar architecture to generate synthetic novel views that are used to augment the training set in the context of image classification, resulting in higher accuracy. Additionally, we propose a method to combine synthetic and real views also at inference time, achieving further improvements. Overall, augmenting both the training and inference pipelines leads to an accuracy gain of up to 20%. Importantly, our inference scheme can also be applied to enhance the resilience in edge-assisted classification systems where the inference task is offloaded to an edge node.