Abstract der fertigen Arbeit

The use of artificial intelligence (AI) models on edge devices with limited resources faces obstacles due to insufficient computational capacity and excessive energy consumption. Split computing addresses this issue by dividing neural networks (NNs) so that some computations occur on edge devices and some in the cloud. This approach manages energy efficiency and latency demands. However, finding the best layer to split and correct hardware setups is complex. This difficulty is due to a vast array of configurations, non-linear interactions between software and hardware, diverse hardware features, and fluctuating workload scenarios.

In response to these hurdles, we introduce DynaSplit, an extensive framework to optimize hardware and software in two distinct phases. DynaSplit dynamically adjusts software components, such as the split layer, along with hardware configurations such as accelerator usage and CPU frequency, to enhance performance. During the Offline Phase, we tackle the optimization issue by employing a multi-objective approach with a meta-heuristic algorithm to find Pareto-optimal setups. Meanwhile, the Online Phase employs a scheduling algorithm to select the optimal settings for every incoming inference task, thereby minimizing energy usage while adhering to the latency thresholds imposed by the application's quality of service (QoS) constraints.

By deploying DynaSplit on a real-world prototype with widely used pre-trained AI models, we achieved notable energy efficiency while meeting application requirements. Our experimental data indicate that DynaSplit can reduce energy usage by as much as 72% in contrast to cloud-only solutions and can meet around 90% of user-defined latency targets, thus greatly exceeding baselines.