Balancing Accuracy and Efficiency in Budget-Aware Early-Exiting Neural Networks

Youva Addad*[1]

[1]Université de Caen Normandie - UFR Langues vivantes étrangères – Université de Caen Normandie – France

Abstract

Deep learning models have achieved remarkable success, but they often come with high computational costs, making them impractical for resource-constrained scenarios. To address this, researchers have explored E*arly Exit Neural Networks*, which allow samples to exit the network at different stages based on their complexity, reducing computation without sacrificing accuracy. In this work we present an Early Exit Neural Network architecture, which enables budgeted classification by dynamically selecting the most relevant exit point for each input sample of a dataset to achieve the best performance while adhering to a pre-defined computational budget. The key contribution of this work is a novel method that jointly learns the classifier model and the sample exiting policy, in contrast to prior approaches that treated these components separately. Specifically, we introduce a bi-level optimization framework that simultaneously optimizes the cross-entropy loss of the classifier and the probabilities of each sample exiting at different stages of the network. This joint learning approach allows the classifier parameters and the sample-dependent exiting policy to be mutually optimized, leading to improved classification accuracy under computational constraints. The proposed EENN method is evaluated on three computer vision benchmarks - CIFAR-10, CIFAR-100, and ImageNet - and demonstrates state-of-the-art results in budgeted classification compared to existing early exit strategies.

*Speaker