ADAPTIVE ACTIVATION-BASED STRUCTURED PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying complex deep learning models on edge devices is challenging because they have substantial compute and memory resource requirements, whereas edge devices' resource budget is limited. To solve this problem, extensive pruning techniques have been proposed for compressing networks. Hand-crafted pruning requires users to explore and adjust multiple hyper-parameters, which is time consuming and often leads to sub-optimal results. It is desirable for the pruning process to automatically meet the objectives in terms of model accuracy, size, and computational complexity etc. without user intervention. This paper presents an *adaptive*, *activation-based*, *structured pruning* approach to automatically and efficiently generate small and accurate models that meet user requirements. First, it defines activation-based attention feature maps for identifying unimportant filters in iterative structured pruning. Then, it proposes adaptive pruning policies for meeting the pruning objectives of accuracy-critical, memory-constrained, and latency-critical tasks. A comprehensive evaluation shows that the proposed method can substantially outperform the state-of-arts structured pruning works. For example, on ResNet-56 with CIFAR-10, without accuracy drop, our method achieves the largest parameter reduction (63.11%), outperforming the related works by 6.81% to 50.07%, and the largest FLOPs reduction (63.61%), outperforming the related works by 7.61% to 20.01%.