EMQ: Evolving Training-free Proxies for Automated Mixed Precision Quantization

Peijie Dong[1†] Lujun Li[2†] Zimian Wei[1] Xin Niu[1*] Zhiliang Tian[1] Hengyue Pan[1]
[1] National University of Defense Technology, [2] HKUST
[1]{dongpeijie, weizimian16, niuxin, tianzhiliang, hengyuepan}@nudt.edu.cn, [2]lilujunai@gmail.com

Abstract

Mixed-Precision Quantization (MQ) can achieve a competitive accuracy-complexity trade-off for models. Conventional training-based search methods require time-consuming candidate training to search optimized per-layer bit-width configurations in MQ. Recently, some training-free approaches have presented various MQ proxies and significantly improve search efficiency. However, the correlation between these proxies and quantization accuracy is poorly understood. To address the gap, we first build the MQ-Bench-101, which involves different bit configurations and quantization results. Then, we observe that the existing training-free proxies perform weak correlations on the MQ-Bench-101. To efficiently seek superior proxies, we develop an automatic search of proxies framework for MQ via evolving algorithms. In particular, we devise an elaborate search space involving the existing proxies and perform an evolution search to discover the best correlated MQ proxy. We proposed a diversity-prompting selection strategy and compatibility screening protocol to avoid premature convergence and improve search efficiency. In this way, our Evolving proxies for Mixed-precision Quantization (EMQ) framework allows the auto-generation of proxies without heavy tuning and expert knowledge. Extensive experiments on ImageNet with various ResNet and MobileNet families demonstrate that our EMQ obtains superior performance than state-of-the-art mixed-precision methods at a significantly reduced cost. The code is available at https://github.com/lilujunai/EMQ-series.

1. Introduction

Deep Neural Networks (DNNs) have demonstrated outstanding performance on various vision tasks [24, 33]. However, their deployment on edge devices is challenging due to high memory consumption and computation cost [18]. Quantization techniques [23, 7, 11] have emerged as a promising solution to address this challenge by performing computation and storing tensors at lower bit-widths than floating point precision, and thus speed up inference and reduce the memory footprint.

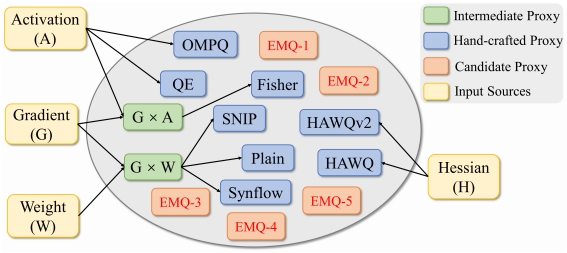

Figure 1. Illustration of the search space for EMQ. Our proposed search space encompasses the handcrafted proxies in mixed-precision quantization, whose input sources are activation(A), gradient (G), weight(W), Hessian(H), as well as their combinations (e.g., $G \times W$). The proposed search space highlights the extensive range of possible combinations, emphasizing the significant effort required to discover new MQ proxies.

Mixed-precision quantization (MQ) [56, 22, 13, 16, 11, 17] is a technique that assigns different bit-widths to the layers of a neural network to achieve a better accuracy-complexity trade-off and allows for the full exploitation of the redundancy and representative capacity of each layer. MQ methods can be categorized into training-based and training-free approaches. **Training-based methods** for MQ present it as a combinatorial search problem and adopt time-consuming Reinforcement Learning (RL) [56], Evolution Algorithm (EA) [57], one-shot [20], or gradient-based [58] methods to find the optimal bit-precision setting. However, these methods can be computationally intensive and require several GPU days on ImageNet [56, 3], limiting their applicability in scenarios with limited computing resources or high real-time requirements. Recently, **training-free approaches** [50, 41, 52, 11, 10, 25] have emerged for mixed-precision quantization, which starkly reduces the heavy computation burden. These approaches aim to reduce the computational burden by building alternative proxies to rank candidate bit-width configurations. For exam-

*Corresponding author, † equal contribution.