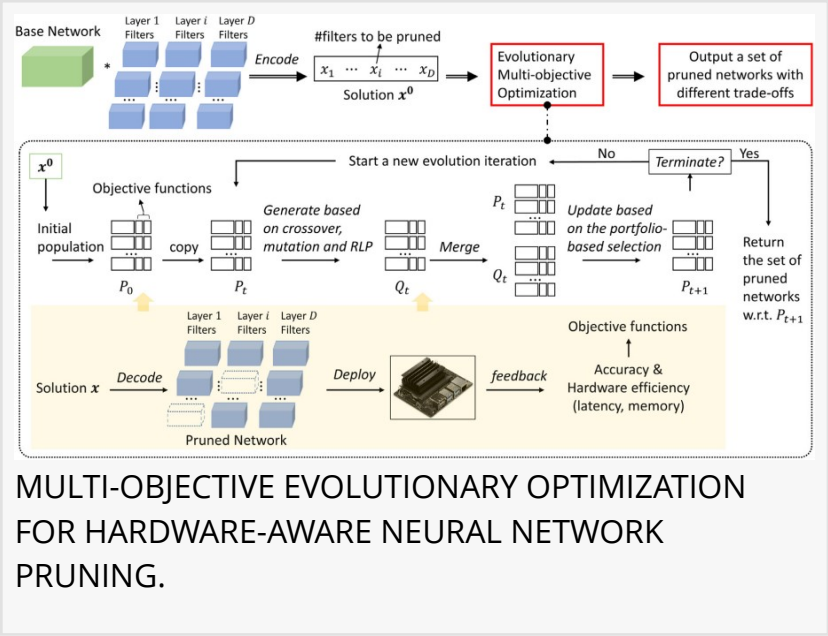
Multi-objective evolutionary optimization for hardware-aware neural network pruning

GA, UNITED STATES, September 25, 2024 /EINPresswire.com/ -- This paper frames hardware-aware neural network pruning as a multi-objective optimization problem and introduces HAMP, a memetic Multi-Objective Evolutionary Algorithm (MOEA) that optimizes both accuracy and hardware efficiency through portfolio-based selection and surrogate-assisted local search. Experiments on the NVIDIA Jetson Nano, an advanced edge device, show that HAMP outperforms existing methods by effectively managing the trade-off between accuracy, latency, and memory, while also providing a diverse set of solutions for flexible user selection.



MULTI-OBJECTIVE EVOLUTIONARY OPTIMIZATION FOR HARDWARE-AWARE NEURAL NETWORK PRUNING.

Neural network pruning is a key technique for deploying artificial intelligence (AI) models based on deep neural networks (DNNs) on resource-constrained platforms, such as mobile devices. However, hardware conditions and resource availability vary greatly across different platforms, making it essential to design pruned models optimally suited to specific hardware configurations. Hardware-aware neural network pruning offers an effective way to automate this process, but it requires balancing multiple conflicting objectives, such as network accuracy, inference latency, and memory usage, that traditional mathematical methods struggle to solve.

In a study (doi: https://doi.org/10.1016/j.fmre.2022.07.013) published in the journal Fundamental Research, a group of researchers from Shenzhen, China, present a novel hardware-aware neural network pruning approach based on multi-objective evolutionary optimization.

“We propose to employ Multi-Objective Evolutionary Algorithms (MOEAs) to solve the hardware neural network pruning problem,” shares Ke Tang, senior and corresponding author of the study.

Compared with conventional optimization algorithms, MOEAs have two advantages in tackling

this problem. One is that MOEAs do not require particular assumptions like differentiability or continuity and possess strong capacity for black-box optimization. The other is their ability to find multiple Pareto-optimal solutions in a single simulation run, which is very useful in practice because it offers flexibility to meet different user requirements.

Specifically, once such a set of solutions has been found. End users can easily choose their preferred configurations of DNN compression, such as latency first or memory consumption first, with just one click on the corresponding solutions," adds Tang.

The team's findings further revealed that, while multi-objective evolutionary algorithms hold significant potential, they still struggle with low search efficiency. To that end, the researchers developed a new MOEA, namely Hardware-Aware Multi-objective evolutionary network Pruning (HAMP), to address this issue.

"It is a memetic MOEA that combines an efficient portfolio-based selection and a surrogate-assist local search operator. HAMP is currently the only network pruning approach that can effectively handle multiple hardware direct feedback and accuracy simultaneously." explains first author Wenjing Hong. "Experimental studies on the mobile NVIDIA Jetson Nano demonstrate the effectiveness of HAMP over the state-of-the-art and the potential of MOEAs for hardware-aware network pruning."

The team's results show that HAMP not only manages to achieve solutions that are better on all objectives, but also delivers simultaneously a set of alternative solutions.

"These solutions present different trade-offs between latency, memory consumption, and accuracy, and hence can facilitate a rapid deployment of DNNs in practice," concludes Hong.

DOI
10.1016/j.fmre.2022.07.013

Original Source URL
https://doi.org/10.1016/j.fmre.2022.07.013

Funding information
This work was supported by grants from the National Natural Science Foundation of China (62106098), the Stable Support Plan Program of Shenzhen Natural Science Fund (20200925154942002), and the MOE University Scientific-Technological Innovation Plan Program.

Lucy Wang
BioDesign Research
email us here