FlexiBO: A Decoupled Cost-Aware Multi-objective Optimization Approach for Deep Neural Networks (Abstract Reprint)

Md Shahriar Iqbal[1], **Jianhai Su**[1], **Lars Kotthoff**[2], **Pooyan Jamshidi**[1]

[1]University of South Carolina, Columbia, SC, USA
[2]University of Wyoming, Laramie, WY, USA

Abstract Reprint. This is an abstract reprint of a journal article by Iqbal, Su, Kotthoff, and Jamshidi (2023).

Abstract

The design of machine learning systems often requires trading off different objectives, for example, prediction error and energy consumption for deep neural networks (DNNs). Typically, no single design performs well in all objectives; therefore, finding Pareto-optimal designs is of interest. The search for Pareto-optimal designs involves evaluating designs in an iterative process, and the measurements are used to evaluate an acquisition function that guides the search process. However, measuring different objectives incurs different costs. For example, the cost of measuring the prediction error of DNNs is orders of magnitude higher than that of measuring the energy consumption of a pre-trained DNN as it requires re-training the DNN. Current state-of-the-art methods do not consider this difference in objective evaluation cost, potentially incurring expensive evaluations of objective functions in the optimization process. In this paper, we develop a novel decoupled and cost-aware multi-objective optimization algorithm, which we call Flexible Multi-Objective Bayesian Optimization (FlexiBO) to address this issue. For evaluating each design, FlexiBO selects the objective with higher relative gain by weighting the improvement of the hypervolume of the Pareto region with the measurement cost of each objective. This strategy, therefore, balances the expense of collecting new information with the knowledge gained through objective evaluations, preventing FlexiBO from performing expensive measurements for little to no gain. We evaluate FlexiBO on seven state-of-the-art DNNs for image recognition, natural language processing (NLP), and speech-to-text translation. Our results indicate that, given the same total experimental budget, FlexiBO discovers designs with 4.8% to 12.4% lower hypervolume error than the best method in state-of-the-art multi-objective optimization.

References

Iqbal, M. S.; Su, J.; Kotthoff, L.; and Jamshidi, P. 2023. FlexiBO: A Decoupled Cost-Aware Multi-Objective Optimization Approach for Deep Neural Networks. *Journal of Artificial Intelligence Research*, 77: 645–682.