RESPERFNET: DEEP RESIDUAL LEARNING FOR REGRESSIONAL PERFORMANCE MODELING OF DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancements of computing technology facilitate the development of diverse deep learning applications. Unfortunately, the efficiency of parallel computing infrastructures varies widely with neural network models, which hinders the exploration of the design space to find high-performance neural network architectures on specific computing platforms for a given application. To address such a challenge, we propose a deep learning-based method, *ResPerfNet*, which trains a residual neural network with representative datasets obtained on the target platform to predict the performance for a deep neural network. Our experimental results show that ResPerfNet can accurately predict the execution time of individual neural network layers and full network models on a variety of platforms. In particular, ResPerfNet achieves 8.4% of mean absolute percentage error for LeNet, AlexNet and VGG16 on the NVIDIA GTX 1080Ti, which is substantially lower than the previously published works.