Dear Editors-in-Chief and Editors.

We are pleased to submit a novel research article entitled “Optimal Model Partitioning with Low-Overhead Profiling on the Heterogeneous Platform for Deep Learning Inference” for publication in ACM TODAES.

The heterogeneous platform, accommodating various computing devices, such as CPUs, GPUs, ASICs, FPGAs, and so on, has been widely used for energy-efficient and high-performance computation of a broad class of applications. However, the recent DNN models’ high complexity in computation and the increasing computing heterogeneity make the cost profiling and the model partitioning harder. In this work, we propose two novel algorithms to resolve the difficulties: one for profiling to recognize the minimum number of execution paths to measure all the costs and the other for partitioning to achieve the best execution performance with the profiled costs. Applying two algorithms to three state-of-the-art transformer-based models, we could profile them in polynomial time and achieve optimal performance. Our approach is relevant to the focus of ACM TODAES since we are confident that the work will significantly contribute this work to the community.

All authors have read and approved the final manuscript and agree with its submission to ACM TODAES. We confirm that this manuscript has not been published elsewhere and is not under consideration by any other journal or conference.

Best regards,

Jaewook Lee, Seok Young Kim, Yoonah Paik, Chang Hyun Kim, Won Jun Lee, and Seon Wook Kim