Dear Editor:

Please find enclosed our manuscript entitled “Optimizing Depthwise Separable Convolution Operations on GPUs”, which we request you to consider for publication as regular paper in *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS*.

This manuscript has not been published elsewhere and is not under consideration by another journal. The preliminary results of this work were presented in the 2020 IEEE Cluster Conference. We have uploaded the original work and a summary of differences.

The depthwise separable convolution, including depthwise and pointwise convolutions, has been widely used in convolutional neural networks to reduce the computation overhead of a standard multi-channel 2D convolution. Existing implementations of depthwise separable convolutions target accelerating model training with large batch size, which are inadequate for small-batch-sized (<128) model training and the typical scenario of model inference. In this manuscript, we aim to bridge the gap of optimizing depthwise separable convolutions by targeting the GPU architecture. For depthwise convolution, we employ column and row reuse to improve its memory performance because depthwise convolution is memory bound. For pointwise convolution, we found that low GPU utilization is often the performance-limiting factor. Therefore, we propose a dynamic tile size scheme to adaptively partition the computation workload and generate enough tiles to saturate GPU as well as maximize data reuse. Additionally, in order to hide global memory access latency for each GPU thread, we design a channel distribution method that increases the arithmetic intensity for each GPU thread. Experimental results conducted on two platforms: an NVIDIA RTX 2080Ti and an embedded NVIDIA Jetson AGX Xavier, and two data types: 32-bit floating point and 8-bit integer demonstrate that our approach achieves an average speedup of 2x (up to 3x) over cuDNN. We show that our approach reduces the end-to-end training and inference time of MobileNet when using a moderate batch size by 11.5% and 9.7% on average, respectively.

We believe that the findings of this study are relevant to the scope of your journal and will be of interest to its readership. We have approved the manuscript and agree with submission to *IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS*. There are no conflicts of interest to declare.

We look forward to hearing from you at your earliest convenience.

Sincerely,

Zhang, Weizhe wzzhang@hit.edu.cn