**Summary of Changes**

The submitted manuscript, titled “Optimizing Depthwise Separable Convolution Operations on GPUs” presents our latest research results that have not yet been published or submitted to a journal. The preliminary results of this work were presented in the 2020 IEEE Cluster Conference (Cluster2020). This manuscript presents clear improvements over the conference version, which are described as follows:

1. **Proposing new optimization methods for pointwise convolution:** We introduce a dynamic tile size scheme to optimize pointwise convolution. Our approach first adjusts the work assigned to each GPU thread so that we have a sufficient number of tiles to be distributed to GPU threads to improve the GPU utilization.
2. **Proposing channel distribution to increase arithmetic intensity:** When partitioning the computation workload into tiles, a small tile size may be selected to generate enough tiles to saturate GPU. In order to hide global memory access latency, we distribute channels across GPU threads within a warp to increase arithmetic intensity for each GPU thread.
3. **Adding experiments on an embedded device and 8-bit integer data type:** We test old and new optimization methods on an embedded NVIDIA Jetson AGX Xavier GPU for both 32-bit floating point and 8-bit integer data types. In the Cluster2020 paper, we only conducted experiments on NVIDIA RTX 2080Ti for 32-bit floating point data type. New experiments demonstrate that our approach consistently outperforms cuDNN by delivering the overall best performance.
4. **Adding experiments for end-to-end training and inference of MobileNet**: We apply our approach to MobileNet and test it on two platforms (NVIDIA RTX 2080Ti and NVIDIA Jetson AGX Xavier) for two data types (32-bit floating point and 8-bit integer). In the Cluster2020 paper, we did not conduct end-to-end experiments. New experiments demonstrate that our approach can reduce 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.

The specific improvements are detailed as follows:

1. We update Section 1 to include a more specific description of pointwise convolution and challenges in optimizing depthwise and pointwise convolutions.
2. We update Section 2 to include description of depthwise separable convolution.
3. We add Section 4 to describe our dynamic tile size scheme for optimizing pointwise convolution.
4. We extend experiment results section to include additional evaluation results. We re-ran all experiments, all figures and tables are updated, and the corresponding discussions are included.
   1. In Section 6.1, we add new experiments on NVIDIA Jetson AGX Xavier and experiments on 8-bit integer data type.
   2. We add Section 6.2 to include experiments for pointwise convolution on two platforms and two data types.
   3. We add Section 6.3 to include experiments for end-to-end inference and training of MobileNet on two platforms and two data types.
5. We extend related work section to include more description of pointwise convolution.
6. We also update the Abstract and Conclusion with the improved methods and the new evaluation results accordingly.