**Summary of Changes**

A preliminary version of this article entitled "Optimizing GPU Memory Transactions for Convolution Operations" by Gangzhao Lu, Weizhe Zhang, Zheng Wang appeared in The 22nd IEEE International Conference on Cluster Computing (Cluster) , 2020.

The extended version makes several additional contributions over the conference paper:

* It proposes a new dynamic tile size scheme to optimize pointwise convolution (Section 4), which is a key component of depthwise separable convolution.
* It includes new experiments performed on embedded devices and using 8-int integers for the neural networks. The new experiments demonstrate that robustness of the proposed approach, showing that it consistently outperforms cuDNN by delivering the overall best performance (Sections 6.1 and 6.2)
* It uses MobileNet as a new case study to show the impact of our optimization on the end-to-end model training and inference time (Section 6.3).
* It extends the related work section (Section xx) to cover in-depth discussions of pointwise convolution
* We have greatly rewritten and extended the paper to provide insights and the motivation of our approach.