Supplementary Materials of Auto-NBA: Efficient and Effective Search Over The Joint Space of Networks, Bitwidths, and Accelerators

Yonggan Fu [1] Yongan Zhang [1] Yang Zhang [2] David Cox [2] Yingyan Lin [1]

1. Ablation studies about the accelerator search engine

As mentioned, the proposed accelerator search engine is one of the key enabler of our Auto-NBA framework. To evaluate its efficacy, we compare the acceleration efficiency of Auto-NBA generated accelerators with SOTA accelerators under the same datasets, models, and hardware resources. For FPGA-based accelerators, we consider three representative SOTA accelerators including (Qiu et al., 2016; Xiao et al., 2017; Zhang et al., 2018) for two DNN models (AlexNet and VGG16) on ImageNet. For a fair comparison, when using our own engine to generate optimal accelerators, we adopt the same precision and FPGA resource as the baselines. The results in Tab. 1 show that the Auto-NBA generated accelerators outperform **both SOTA expert-designed and tool-generated** accelerators under the same dataset, DNNs, and FPGA resources. For example, the Auto-NBA generated accelerators achieve up to 2.16× improvement in throughput on VGG16. The consistent better performance of Auto-NBA's automatically generated accelerators validates the effectiveness of our accelerator search engine in being able to navigate over the *large* and *discrete* design space of accelerators to efficiently identify/locate the optimal accelerators.

[1] Department of Electrical and Computer Engineering, Rice University [2] MIT-IBM Watson AI Lab. Correspondence to: Yingyan Lin <yingyan.lin@rice.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

References

Qiu, J., Wang, J., Yao, S., Guo, K., Li, B., Zhou, E., Yu, J., Tang, T., Xu, N., Song, S., et al. Going deeper with embedded fpga platform for convolutional neural network. In *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 26–35. ACM, 2016.

Xiao, Q., Liang, Y., Lu, L., Yan, S., and Tai, Y.-W. Exploring heterogeneous algorithms for accelerating deep convolutional neural networks on fpgas. In *Proceedings of the 54th Annual Design Automation Conference 2017*, DAC '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349277. doi: 10.1145/3061639.3062244. URL `https://doi.org/10.1145/3061639.3062244`.

Zhang, X., Wang, J., Zhu, C., Lin, Y., Xiong, J., Hwu, W.-m., and Chen, D. Dnnbuilder: An automated tool for building high-performance dnn hardware accelerators for fpgas. In *Proceedings of the International Conference on Computer-Aided Design*, ICCAD '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359504. doi: 10.1145/3240765.3240801. URL `https://doi.org/10.1145/3240765.3240801`.

Table 1. Auto-NBA generated FPGA accelerators vs. SOTA FPGA accelerators built on top of an SOTA FPGA board, Zynq XC70Z45, adopting a frequency of 200 Mhz for different networks with a fixed precisio of 16-bit on ImageNet.

	(Zhang et al., 2018)	(Xiao et al., 2017)	(Qiu et al., 2016)	**Auto-NBA generated**	(Zhang et al., 2018)	**Auto-NBA generated**
Network	VGG16	VGG16	VGG16	VGG16	AlexNet	AlexNet
Resource Utilization	680/900 DSP	824/900 DSP	780/900 DSP	723/900 DSP	808/900 DSP	704/900 DSP
Performance (GOP/s)	262	230	137	**291**	247	**272**