Table 1: Comparing with channel pruning. Speedups for HRank[11] are inferred from its theoretical FLOPs reduction.

Network	Sparsity ratio	Accuracy	Speedup
Baseline	Dense	76.12	1×
HRank[11]	37%	74.98	1.77×
OVW	50%	75.76	1.86×
HRank[11]	46%	71.98	2.63×
OVW	70%	73.35	2.79×

Table 2: Comparing speedup with different vector lengths.

Resnet50 layer		64	32	16
Conv0	50%	**1.51**	1.27	1.12
	90%	2.14	**2.76**	1.93
Conv11	50%	**2.23**	1.73	0.33
	90%	**7.75**	6.62	1.38
Conv41	50%	**1.62**	0.76	0.31
	90%	**7.93**	3.72	1.53

A Appendix

A.1 Ablation Study

Table 1 shows the accuracy of our method compared to advanced channel pruning methods, HRank [11]. Channel pruning requires specialized training from scratch methods to recover its enormous accuracy drop. The OVW pattern demonstrates a better accuracy-speed tradeoff against it.

Table 2 supports our arguments for V that only vector lengths as large as 32 or 64 can minimize convolution kernel runtime.

Table 2 justified our system is robust on different types of GPU as long as its architecture supports optimization for the implicit GEMM convolution algorithm.

References

[1] MegEngine:A fast, scalable and easy-to-use deep learning framework. 2020.

[2] *Cao Shijie, Zhang Chen, Yao Zhuliang, Xiao Wencong, Nie Lanshun, Zhan Dechen, Liu Yunxin, Wu Ming, Zhang Lintao*. Efficient and effective sparse LSTM on FPGA with bank-balanced sparsity // Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. 2019. 63–72.

Table 3: Comparing speedup on RTX 3070.

Resnet50 layer		V100		RTX 3070	
		V=64	V=32	V=64	V=32
Conv11	50%	2.23	1.73	1.71	1.55
	90%	7.75	6.62	6.71	6.39
Conv41	50%	1.62	0.76	1.30	0.68
	90%	7.93	3.72	6.30	3.31

[3] *Gale Trevor, Zaharia Matei, Young Cliff, Elsen Erich*. Sparse gpu kernels for deep learning // SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. 2020. 1–14.

[4] *Gray Scott, Radford Alec, Kingma Diederik P*. Gpu kernels for block-sparse weights // arXiv preprint arXiv:1711.09224. 2017. 3. 2.

[5] *Guo Cong, Hsueh Bo Yang, Leng Jingwen, Qiu Yuxian, Guan Yue, Wang Zehuan, Jia Xiaoying, Li Xipeng, Guo Minyi, Zhu Yuhao*. Accelerating sparse dnn models without hardware-support via tile-wise sparsity // SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. 2020. 1–15.

[6] *Han Song, Mao Huizi, Dally William J*. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding // arXiv preprint arXiv:1510.00149. 2015.

[7] *Han Song, Pool Jeff, Tran John, Dally William*. Learning both weights and connections for efficient neural network // Advances in neural information processing systems. 2015. 28.

[8] *He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian*. Deep residual learning for image recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. 770–778.

[9] *Huang Guyue, Li Haoran, Qin Minghai, Sun Fei, Din Yufei, Xie Yuan*. Shfl-BW: Accelerating Deep Neural Network Inference with Tensor-Core Aware Weight Pruning // arXiv preprint arXiv:2203.05016. 2022.

[10] *Li Hao, Kadav Asim, Durdanovic Igor, Samet Hanan, Graf Hans Peter*. Pruning filters for efficient convnets // arXiv preprint arXiv:1608.08710. 2016.

[11] *Lin Mingbao, Ji Rongrong, Wang Yan, Zhang Yichen, Zhang Baochang, Tian Yonghong, Shao Ling*. Hrank: Filter pruning using high-rank feature map // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. 1529–1538.

[12] *Liu Zhuang, Sun Mingjie, Zhou Tinghui, Huang Gao, Darrell Trevor*. Rethinking the value of network pruning // arXiv preprint arXiv:1810.05270. 2018.

[13] *Ma Xiaolong, Qin Minghai, Sun Fei, Hou Zejiang, Yuan Kun, Xu Yi, Wang Yanzhi, Chen Yen-Kuang, Jin Rong, Xie Yuan*. Effective Model Sparsification by Scheduled Grow-and-Prune Methods // arXiv preprint arXiv:2106.09857. 2021.

[14] *Malinen Mikko I, Fränti Pasi*. Balanced k-means for clustering // Joint iapr international workshops on statistical techniques in pattern recognition (spr) and structural and syntactic pattern recognition (sspr). 2014. 32–41.

[15] *Mishra Asit, Latorre Jorge Albericio, Pool Jeff, Stosic Darko, Stosic Dusan, Venkatesh Ganesh, Yu Chong, Micikevicius Paulius*. Accelerating sparse deep neural networks // arXiv preprint arXiv:2104.08378. 2021.

[16] *Pool Jeff, Yu Chong*. Channel Permutations for N: M Sparsity // Advances in Neural Information Processing Systems. 2021. 34.

[17] *Sui Yang, Yin Miao, Xie Yi, Phan Huy, Aliari Zonouz Saman, Yuan Bo*. CHIP: CHannel Independence-based Pruning for Compact Neural Networks // Advances in Neural Information Processing Systems. 2021. 34.

[18] *Wen Wei, Wu Chunpeng, Wang Yandan, Chen Yiran, Li Hai*. Learning structured sparsity in deep neural networks // Advances in neural information processing systems. 2016. 29.

[19] *Zhang Xiang, Zhao Junbo, LeCun Yann*. Character-level convolutional networks for text classification // Advances in neural information processing systems. 2015. 28.

[20] *Zhou Aojun, Ma Yukun, Zhu Junnan, Liu Jianbo, Zhang Zhijie, Yuan Kun, Sun Wenxiu, Li Hongsheng*. Learning n: m fine-grained structured sparse neural networks from scratch // arXiv preprint arXiv:2102.04010. 2021.

[21] *Zhou Yangjie, Yang Mengtian, Guo Cong, Leng Jingwen, Liang Yun, Chen Quan, Guo Minyi, Zhu Yuhao*. Characterizing and Demystifying the Implicit Convolution Algorithm on Commercial Matrix-Multiplication Accelerators // 2021 IEEE International Symposium on Workload Characterization (IISWC). 2021. 214–225.

[22] *Zhu Maohua, Zhang Tao, Gu Zhenyu, Xie Yuan*. Sparse tensor core: Algorithm and hardware co-design for vector-wise sparse neural networks on modern gpus // Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture. 2019. 359–371.