Accelerated COVID-19 CT Image Enhancement via Sparse Tensor Cores

Ayush Chaturvedi, Wu-Chun Feng

Virginia Tech

ABSTRACT

Many deep learning model architecture are an inspiration of how human brain works however their implementation in computer programming deviates in the sense that these networks over time become dense or are intentionally designed in such a way to achieve better generalization and accuracy whereas neural architecture in brain is highly sparse. In this work we target a similar deep learning model designed to enhance CT images of Covid-19 chest scans namely DD-Net (short for Dense Net and Deconvolution Network) from prior work of ComputeCovid19+. The model follows an auto encoder decoder architecture in the deep learning paradigm and has high dimensionality due to presence of stack convolution layers and deconvolution layers and thus takes many compute hours of training. We propose a set of techniques which target these two aspects of model - dimensionality and training time. We implement structured sparsity along with a hybrid training schedule. By pruning neurons, we make the model sparse and thus reduce the effective dimensionality and then retrain this sparsified model with minimal additional overhead of re-training. We also apply set of techniques tailored with respect to underlying hardware in order to better utilize the existing components of hardware (such as tensor core) and thus further reduce the overall time and associated computational cost required to train this model with the new hybrid training schedule.

INTRODUCTION

Name	Value
Architecture	DDNET
Input image size	512x512
Number of layers	45
Number of Convolution layers	25
Number of deconvolution layers	7
Number of Parameters	783,809
Cost Function	MSE + 0.2 * MS-SSIM

Table 1. Architectural Details of DDNet

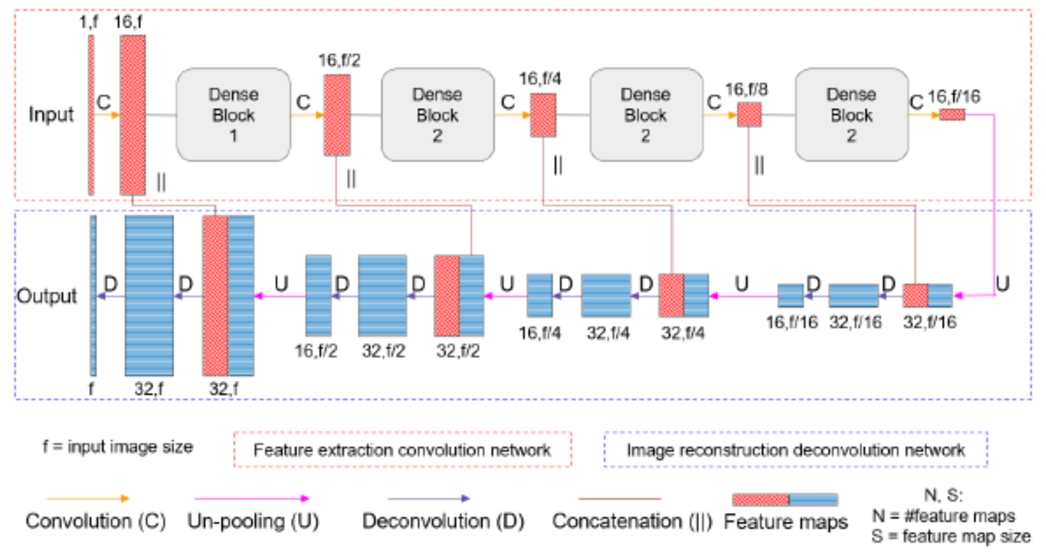

Figure 1. Architecture of Dense Net and Deconvolution Net (DDNet).

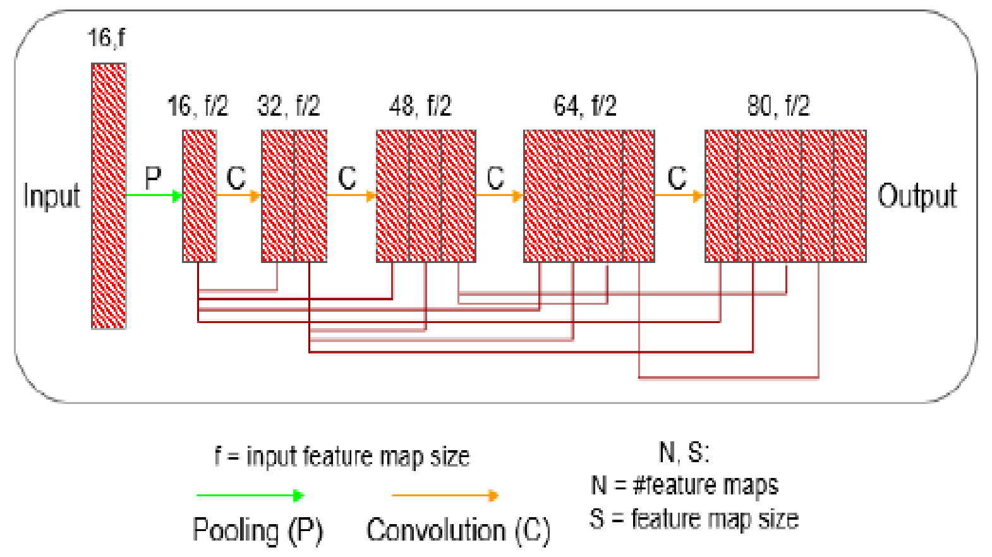

Figure 2. Architecture of a Dense block in DDNet.

A DL neural network was constructed from scratch namely DDNet[1]. To which we employ structured pruning based on L1 distribution of weights and activations which results in accuracy loss (as expected). Therefore, In order retain generalization we re-train this sparse model. Additionally, to better leverage sparse tensor cores we enabled mixed precision sparse re-training of model. The efficiency of model is evaluated in terms of enhanced image quality evaluated in terms of MSE and MS-SSIM and training overhead in terms of time/epochs required to reach the similar image quality baseline The evaluation of performance improvement is considered as the reduction in this overhead required in retraining.

METHODOLOGY

In our implementation, we first trained the DDNet model without any sparsity. This densely trained model is then used as a baseline for sparsified re-training. We then implement structured sparsity and retrain the model to achieve the same generalization and observe the difference in performance and overall training time. Additionally, we further accelerate the sparsified re-training with mixed precision which leverages tensor cores in the Nvidida GPUs. The sparsified convolution and deconvolution layers become ideal candidate to be accelerated by tensor cores. Tensor Cores and their associated data paths are custom-designed to dramatically increase floating-point compute throughput with high energy efficiency. Ampere architecture which builds on top of volta architecture and introduces **hardware optimized sparse GEMM operations** giving a significant reduction in number of clock cycles needed to compute FMA (Fused Multiply Add) operations: Fig. 3

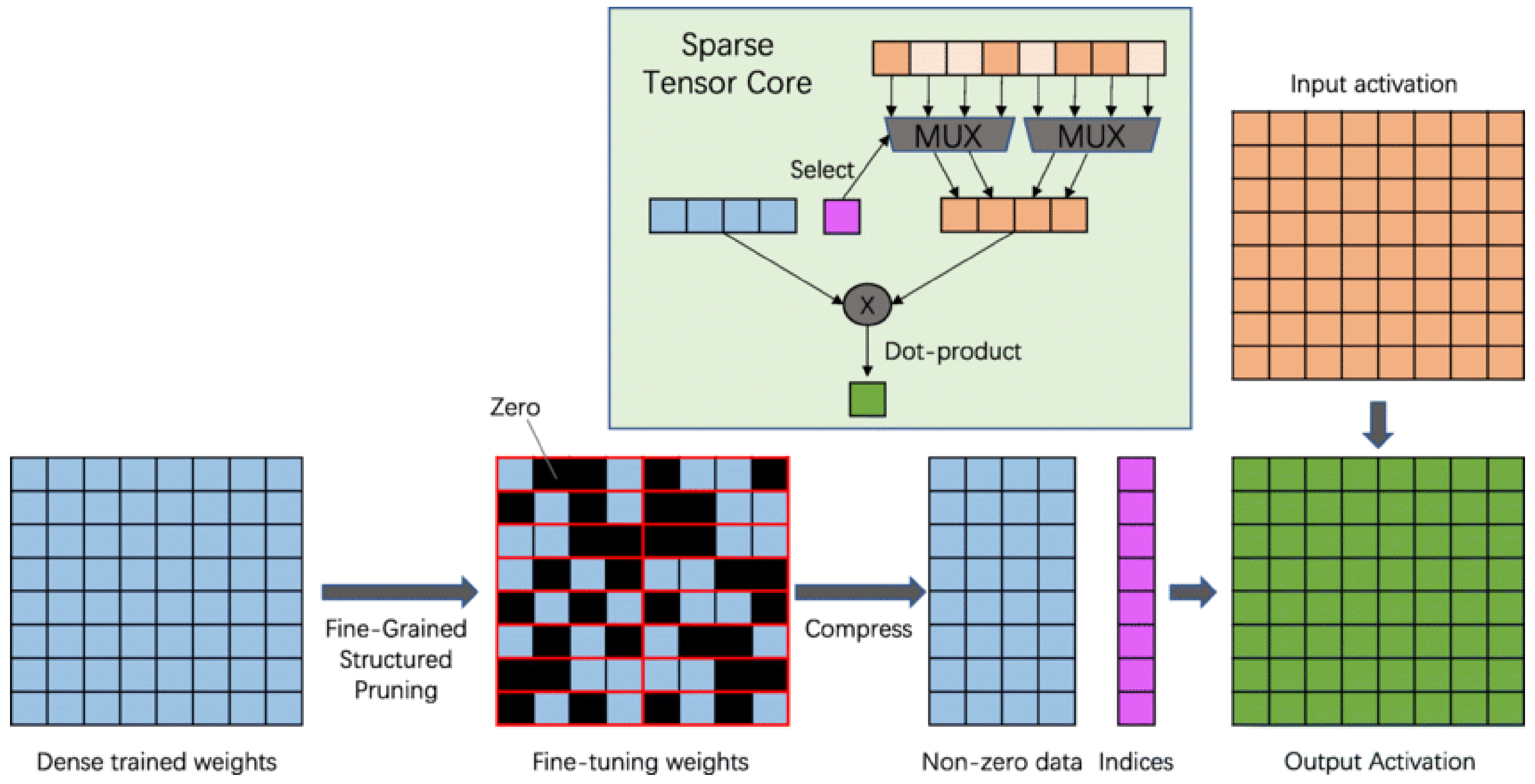

Figure 3. FMA operation with fine grain structured sparsity in Nvidia Ampere sparse tensor cores[2]

RESULTS

	MS-SSIM	MSE	Total Epochs/Time	Epochs/Time (Dense)	Epochs/Time (Sparse)	% Sparse
Baseline	**93.79±0.49**	0.0029±0.0051	50 / 73 min	50 / 73 min	0 / 0 min	0
Structured Sparsity	93.58±0.01	**0.0026±0.0067**	35 / 50 min	30 / 42 min	5 / 8 min	50
Structured Sparsity + Mixed Precision	92.38±0.05	0.0044±0.0007	**35 / 37 min**	30 / 32 min	5 / 5 min	50

Table 2. The mean values of MSE, MS-SSIM obtained using the BIMCV and MIDRC Lung test dataset. 5000 training images used along with 700 for testing.

In this section, we present both a qualitative and a quantitative comparison of the results of sparsified DDNet model using the deep neural network described above. We see that sparsified model does a better job at reducing noise (Fig.4) and retains structural similarity with a little overhead.

EVALUATION AND DISCUSSION

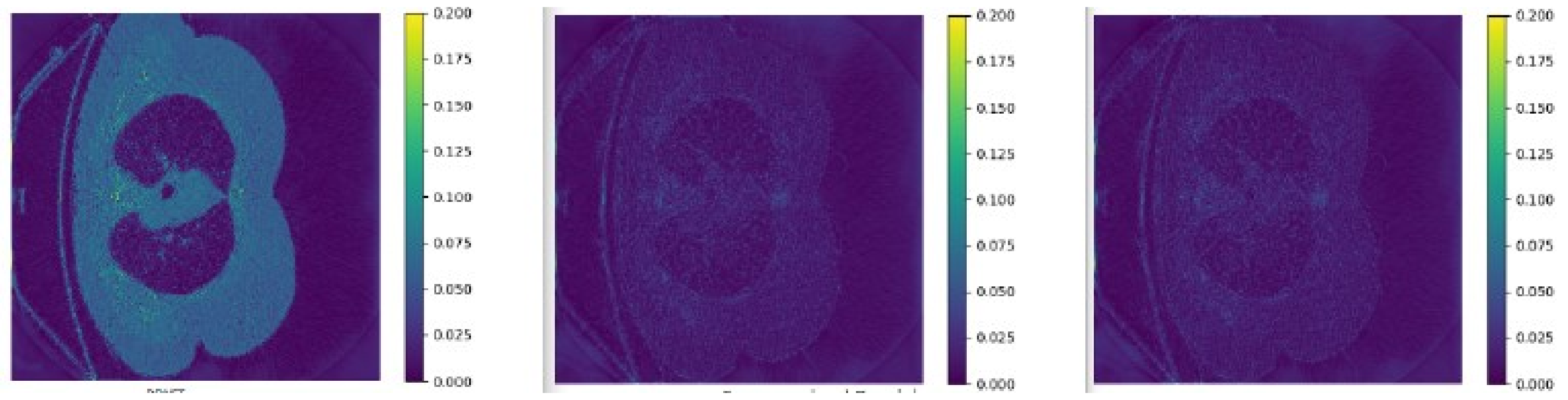

Figure 4. Left to Right Difference maps (calculated as pixel-wise MSE b/w enhanced image and target high quality CT image) of DDNet baseline, Structured sparsity and Structured Sparsity with mixed Precision [**Darker is better**]

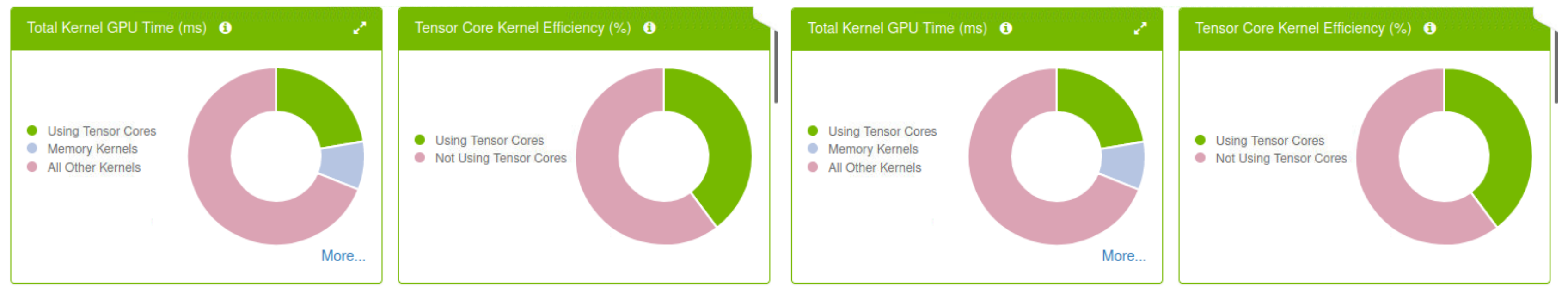

Figure 5. Tensor core efficiency. Left: without sparse training Right: With Sparse Training

We see that in order to better leverage tensor cores on Nvidia Ampere architecture; 50% sparsification of model layers (such as conv. & deconv.) is must. Doing that we can see a better tensor core utilization to 39% from previous 8% by the underlying cuda kernels and thus the performance speedup.

CONCLUSION AND WORK IN PROGRESS

We see that with the sparse re-training schedule the model leads to a 1.46x speedup in training time. Moreover, incorporating mixed precision enables operations to be offloaded to tensor core which have a higher IPC and thus a further speedup of 1.37x was achieved. A speed up of 1.9x over dense training baseline with hybrid schedule under mixed precision. We are further evaluating this work with on other modern processors with capabilities to accelerate sparse matrix vector operation such as AMD MI200/250 GPUs and Cerebras-CS2 system.

REFERENCES

[1] Zhicheng Zhang, Xiaokun Liang, Xu Dong, Yaoqin Xie, and Guohua Cao, "A Sparse-View CT Reconstruction Method Based on Combination of DenseNet and Deconvolution". IEEE Trans Med Imaging 37(6): p. 1407-1417 (2018). [2[Chen, Zhengbo & Zheng, Fang & Yu, Qi & Sun, Rujun & Guo, Feng & Chen, Zuoning. (2022). Evaluating performance of AI operators using roofline model. Applied Intelligence. 52. 10.1007/s10489-021-02794-5.

[2] Garvit Goel, Atharva Gondhalekar, Jingyuan Qi, Zhicheng Zhang, Guohua Cao, and Wu Feng. 2021. ComputeCOVID19+: Accelerating COVID-19 Diagnosis and Monitoring via High-Performance Deep Learning on CT Images. In 50th International Conference on Parallel Processing (ICPP '21), August 9–12, 2021, Lemont, IL, USA. ACM, New York, NY, USA 11 Pages. https://doi.org/10.1145/3472456.3473523.