International Journal of Grid and High Performance Computing
Volume 10 • Issue 1 • January-March 2018

GPU Implementation of Image Convolution Using Sparse Model with Efficient Storage Format

Saira Banu Jamal Mohammed, VIT University, Vellore, India
M. Rajasekhara Babu, VIT University, Vellore, India
Sumithra Sriram, VIT University, Vellore, India

ABSTRACT

With the growth of data parallel computing, role of GPU computing in non-graphic applications such as image processing becomes a focus in research fields. Convolution is an integral operation in filtering, smoothing and edge detection. In this article, the process of convolution is realized as a sparse linear system and is solved using Sparse Matrix Vector Multiplication (SpMV). The Compressed Sparse Row (CSR) format of SPMV shows better CPU performance compared to normal convolution. To overcome the stalling of threads for short rows in the GPU implementation of CSR SpMV, a more efficient model is proposed, which uses the Adaptive-Compressed Row Storage (A-CSR) format to implement the same. Using CSR in the convolution process achieves a 1.45x and a 1.159x increase in speed compared to the normal convolution of image smoothing and edge detection operations, respectively. An average speedup of 2.05x is achieved for image smoothing technique and 1.58x for edge detection technique in GPU platform usig adaptive CSR format.

KEYWORDS

Convolution, CSR, Edge Detection and Image Smoothening, GPU, SpMV

INTRODUCTION

Image processing involves the subjection of the source image to various operations, depending on the need of the application, to get a desired target image. The process of convolution is one of the most integral operations that are used for modifying the pixel values of the images. Various operations such as smoothing, edge-detection, sharpening, etc. can be achieved using the process of convolution. A multitude of kernels has been proposed over the years to better achieve the output image. Recently, sparse matrices have been found to play a vital role in image processing (Wang, Yan, Pan, & Xiang, 2011). Sparse matrices are matrices that are majorly populated with zeroes. The image processing operations are represented in the form of SpMV, where the sparse matrix kernel depends on the operation to be performed, and the vector is the input image represented as a one-dimensional vector. This can be easily extended to implement complex operations that involve the use of more than one of the basic processes. High efficiency gains can be obtained when all the constituent processes can be represented as a sparse matrix, and the entire image manipulating operation can be implemented as a series of SpMV operations. These complex processes however can involve multiple SpMV operations of very large matrices, depending on the size of the image and the kernel. Due to the large size of the sparse matrix kernel involved, various sparse matrix storage formats such as the

DOI: 10.4018/IJGHPC.2018010104

Coordinate (COO) format, the Compressed Row Storage (CSR) format, the Quadtree CSR (QCSR) format, etc., can be used to further increase the space and time complexity. The CSR format has been found to be the most efficient in terms of space and time complexity of the SpMV operation. With the demand for the need for greater efficiency when using large datasets, high performance computing has become a booming area of research. Much of the concentration to achieve this is on multi-threading and graphics processing unit (GPU). An improvised version of the CSR format, the Adaptive CSR format has been recently proposed for the GPU platform which gives a considerable speedup over the implementation of the normal CSR format.

In this paper, we propose a model for the implementation of convolution for images which involves the use of SpMV using the adaptive CSR format. Results are shown for various image processing applications such as edge detection and smoothing, and it is clear that the proposed model shows considerable time gain.

LITERATURE REVIEW

An image is typically represented as a two-dimensional matrix in the computer memory. There are multiple operations that can be performed on image pixels, to get a target image of a specified type, based on the application needs (Marasco, Abaza, & Cukic, 2015), (Yan, Sethi,Rangarajan, Vatsavai, & Ranka, 2017). Image editing and smoothening operation plays an important role in removing the noise from the images. This is important for any image processing task to be performed. (Alvarez, Lions, & Morel, 1992) presented a non-linear diffusion model for edge detection and selective smoothening operation. (Chen, Li, Zhang, Hsu, & Wang, 2017) proposed multifeature fusion for large scale real time duplicate image detection.Storing high resolution images can occupy a lot of space on the hard disk and performing various operations can be computationally expensive. Hence, a variety of robust compression techniques have been proposed over the years to reduce the size of the image stored, and GPU computing has been increasingly used to implement the algorithms. (Arora, & Shukla, 2014) have presented a survey on various lossless and lossy image compression techniques. JPEG compression, based on the Discrete Cosine Transformation (DCT) is the most widely used commercial algorithm. (Patel, Wong, Tatikonda, & Marczewski, 2009) in their paper have explored the improvements that can be brought to the JPEG compression by using the concept of GPU computing. They were able to obtain 61% increase in performance by doing so. (Hasan, Nur, Noor, & Shakur, 2012) have proposed a robust algorithm for lossless image compression, which can be used in the transmission of digital images over the internet. It is based on the concept of run-length coding, and can work well on today's heterogeneous network structure. Medical imaging is an important field which requires image processing techniques to extract useful information from medical images obtained from different sources. (Eklund, Dufort, Forsberg, & LaConte, 2013) in their paper gave a comprehensive overview about the various important image processing algorithms that are used in medical imaging. They have also analysed the impact of massive parallelization on some of the computationally expensive operations. (Haque, Kaisan, Saniat, & Rahman, 2014) have worked on fractal image compression techniques for medical imaging and used GPU computing to speed up the operation. The only problem with this technique is that the compression is irreversible, and hence loss of information occurs, which is critical in the case of medical imaging. The matrix used to represent an image with the majority pixel values as zeroes, is known as sparse matrix. Storing these matrices using various sparse matrix storage formats can lead to better efficiency. Image compression is one such operation where this idea can be involved. Binary images are represented as 0s and 1s. Hence, compression of binary images using various sparse matrix storage formats helps in reducing the storage space effectively and thereby increases the compression ratio. The recent research trend is to use sparse matrix storage format for compressing grayscale and colour images. (Tianxu, & Yonghui, 2006) have done extensive works to implement this concept. (Guha, & Ward, 2014) have used sparse matrices to improve compression of similar images. (He, & Chen, 2008) have analysed