Comparing Effectiveness of Lossy and Lossless Reduction Techniques

COLEMAN NICHOLS and JON C. CALHOUN (ADVISOR)*, Holcombe Department of Electrical and Computer Engineering - Clemson University, USA

Large data sets tend to be very common in many areas of high-performance computing. Often times, the size of these data sets are so extreme that they far exceed the storage capabilities of their system. This highlights an opportunity to employ compression methods in order to reduce the data set down to a manageable size. Given that reduction methods operate on data in different ways, it is important to compare these methods with the goal of determining the optimal approach for any given data set. This poster compares the effectiveness of different data reduction methods on image data from Los Alamos National Labs based on three major parameters: PSNR, compression ratio, and compression rate. Our analysis indicated the SZ lossy compressor was the most effective for this data set, given that it offered the highest PSNR along with a very reasonable compression ratio.

Additional Key Words and Phrases: Data Reduction, Compression, Hybrid Data Sampling, Image Compression

ACM Reference Format:
Coleman Nichols and Jon C. Calhoun (Advisor). 2022. Comparing Effectiveness of Lossy and Lossless Reduction Techniques . 1, 1 (September 2022), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Data reduction is becoming more and more prevalent as the size of data sets continues to increase. One of the most common uses for these very large data sets is in the field of machine learning. This is the exact process that researchers at Los Alamos National Labs (LANL) are employing. These researchers are analyzing an additive manufacturing operation in which they have noticed structural volatility in its product. They are doing so by taking thousands of pictures of the process and feeding this image data into a machine learning algorithm. The dataset From Los Alamos National Labs can be found under the title LA-UR-21-32202

The newly emerging problem is that these images are taking up an extreme amount of storage space. On any given day, upwards of a terabyte of images are taken of the process. The clear solution to this problem is to utilize data reduction tools. These tools each operate on data in different ways. Depending on the type of data, certain reduction methods are faster, introduce less error, or are able to reduce the data down in a more efficient manner than others. Therefore it is important to compare these reduction techniques with the goal of finding which one is most suitable for the problem at hand.

This is exactly what this poster has done with regards to a segment of LANL's image data. By focusing on key metrics such as PSNR, compression rate, and compression ratio, conclusions can be drawn on which reduction tool is most efficient for this specific application. The specific reduction tools being compared in this poster include hybrid data sampling [3], SZ error-bounded lossy compressor [2], and BLOSC lossless compressor [1].

Authors' address: Coleman Nichols, cnicho5@clemson.edu; Jon C. Calhoun (Advisor), jonccal@clemson.edu, Holcombe Department of Electrical and Computer Engineering - Clemson University, 433 Calhoun Dr, Clemson, South Carolina, USA, 29634.

2022. Manuscript submitted to ACM

2 BACKGROUND AND METHODS

Many terms such as PSNR, compression rate, and compression ratio have been used thus far without proper explanation. PSNR or Peak Signal to Noise Ratio is a metric typically used to quantify the error or noise of an image relative to another in units of decibels (dB). A formula for PSNR is shown below.

$$PSNR = 10 \cdot log_{10} \frac{MAX_I^2}{MSE} \tag{1}$$

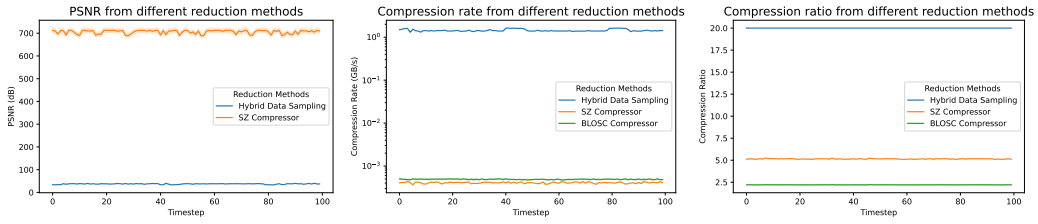
Max_I represents the range of possible pixel value and MSE represents the Mean Squared Error between the two images. The accuracy of a reduction tool can be quantified by comparing how similar the reduced data is to the original via the PSNR metric.

Compression ratio refers to the ratio of the uncompressed size of the data to the reduced size of the data set. For example, a compression ratio of 2 would mean that the reduction method was able to reduce the data to half of its original size. Compression rate refers to the speed at which the reduction tool can compress the data. This is typically measured in units of bytes per second.

The three reduction methods compared throughout this poster include hybrid data sampling [3], the SZ compressor [2], and the BLOSC compressor [1]. The hybrid data sampling tool samples segments of the original data set with full resolution, and then can reconstruct the data from these original samples. The SZ compressor is an error-bounded lossy compressor. This means that the user selects an acceptable amount of the original data to be lost in order to achieve a better compression ratio. As for the BLOSC lossless compressor, it is able to reduce the size of the data without losing any information. This ensures that the decompressed data will be identical to the original data.

3 EXPERIMENT AND RESULTS

In order to compare all of the reduction methods listed above, all of the necessary software was downloaded onto the Palmetto cluster where the experiments were conducted. This software includes Libpressio 0.79.0 [4], Python 3.9.2, OpenMPI 4.0.3, and CUDA 11.5.0. Each reduction method was used to compress and decompress the data individually, allowing us to obtain and compare the key metrics previously mentioned. The graphs shown below demonstrate how each of these reduction methods performed on the data with respect to our key metrics.



(a) PSNR (b) Compression Rate (c) Compression Ratio

Fig. 1. Comparing Key Metrics Across Reduction Methods

From these graphs, we can see that the hybrid data sampling method had the highest compression rate and compression ratio. The major downside to this method is its low PSNR value. This method excelled in reducing the size of the data in a quick manner at the expense of the data set's accuracy. These graphs also demonstrate how the compressors

operated on the data set. Since BLOSC is a lossless compressor, it will always have an infinite PSNR value which is why it is not represented in the PSNR graph. BLOSC was able to reduce the image data to around one-half of its original size with perfect accuracy; however, it struggled in terms of compression rate. Finally, we can see that the SZ compressor was able to reduce the image data down to one-fifth of its original size with an extremely high PSNR value. The only drawback to this method is its compression rate, since it was the slowest of the three reduction methods.

4 CONCLUSION

As the size of data sets continues to increase, it becomes imperative that we start to look into ways to efficiently store this data. This poster focused on comparing three different data reduction methods specifically on LANL's set of additive manufacturing image data. From this process, we have concluded that the SZ error-bounded lossy compressor offers the greatest upside for this data set. With high accuracy and very reasonable compression ratio, it is the most efficient reduction method of the three methods discussed. Future research in this area could focus on combining these methods in a way which encapsulates the strengths of each individual method.

ACKNOWLEDGMENTS

Clemson University is acknowledged for generous allotment of compute time on the Palmetto cluster. This material is based upon work supported by the National Science Foundation under Grant No. SHF-1910197 and SHF-1943114.

REFERENCES

[1] Francesc Alted. 2010. Why modern CPUs are starving and what can be done about it. *Computing in Science & Engineering* 12, 2 (2010), 68–71.

[2] Sheng Di and Franck Cappello. 2016. Fast error-bounded lossy HPC data compression with SZ. In *2016 ieee international parallel and distributed processing symposium (ipdps)*. IEEE, 730–739.

[3] Megan Hickman Fulp, Ayan Biswas, and Jon C Calhoun. 2020. Combining Spatial and Temporal Properties for Improvements in Data Reduction. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2654–2663.

[4] Robert Underwood, Victoriana Malvoso, Jon C Calhoun, Sheng Di, and Franck Cappello. 2021. Productive and Performant Generic Lossy Data Compression with LibPressio. In *2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7)*. IEEE, 1–10.