Advanced Data Compression Methods (ADCM)

*A **ReNew** research suggestion white paper*

Paul Cadaret
UNICON Inc, Rancho Santa Margarita, CA, 92688
Phone: 949-742-7538; Fax: 877-243-6431; Email: alias-renew-2009[at]unicon.ws

Background

UNICON's recent participation in the *Small Business Innovation Research* (**SBIR**) programs at the DOD and DOE has inspired us to envision how very large and fast artificial neural networks could be constructed. We believe that our technology [1] allows us to employ very large neural networks in a way that enables a concept we call "***exhaustive learning***" to be applied for the benefit of certain types of complex and slow computational problems [2]. In mid 2008 we developed a paper [3] for the DOD that described how a large-scale fast/fuzzy pattern recognition system could be utilized as the basis for an unusually effective imagery data compression engine engine. This ReNeW paper proposes utilizing a variation of our method [3] to develop highly effective data compression systems appropriate for the types of massive data sets that are typically generated by fusion energy experiments.

1 The Data Management Problem

Within the last few years we have come to better understand certain needs of the DOE fusion energy community. We understand that the community has been pursuing activities that are significantly focused on high-energy tokamak reactor science. Such scientific exploration demands experimentation. Such experimentation typically requires that data be accumulated at high rates so that an accurate record of experimental behavior can be acquired. This data is then typically analyzed so that a more thorough understanding of the science regarding an experiment can be acquired. Although the massive data sets that result from such experiments are initially stored locally, they are frequently transmitted around the world to support data analysis efforts by various scientists. This presents a significant data management problem for the fusion energy community. Massive data sets are a burden on data storage capabilities, data communication capabilities, computing resources, and operating budgets. Most significantly, the management of such data sets is a drain on the precious and limited time of the scientific personnel called upon to analyze them. We believe that such problems exist throughout the fusion energy community and a likely applicable to ITER and the larger fusion energy community.

2 Technical Requirements For A Solution

Ideally, what is desired are fast, highly effective, and lossless data compression mechanisms that can be used to faithfully compress massive scale experimental data sets. Generally, we want compression methods that are capable of compressing almost arbitrary experimental data while exhibiting high compression ratios. Commonly available lossless data compression schemes such as the GNU "gzip" program can compress binary data and provide compression ratios on the order of 3 or 4 to 1. Although such compression methods are useful, when dealing with multi-gigabyte or terabyte scale data sets it is clear that the availability of much more effective data compression methods would be advantageous. Although we ideally seek lossless data compression mechanisms, we are reminded that the fusion energy community generates experimental data that inherently has some element of noise associated with it. Therefore, it may be possible to consider this fact and use methods that are fast, highly effective, and minimally lossy. If the community had access to data compression methods that exhibit very high compression ratios while introducing a minimal amount of compression error (imprecision), then such a capability might be an important option for the community to consider in its overall data management and data communication strategy.

3 Elements Of Research Thrusts Needed To Provide The Solution

3.1 A Brief Overview Of The Unusual Data Compression Method Proposed

Several years ago we noted a paper [4] that was generated by NASA and Carnegie Mellon University (CMU) that described their work exploring the utility of a fuzzy-feature-indexing method for data compression that exhibited unusually high data compression ratios. The method they described was apparently specific to the compression of LIDAR image data; however, we recognized that the underlying method that they employed could likely be applied in a wide variety of other data domains. The compression ratio that they ***demonstrated*** was astonishing (**2000:1**); however, the data compression speed they attained was very slow due to their method of implementation. Nevertheless, we recognized the fundamental importance of the approach that they explored in their work. We also recognized that the speed problem that they encountered might likely be overcome with our technology [1]. As a result, we generated a paper [3] that presented our thoughts regarding how the NASA-CMU data compression approach [4] might be enhanced and applied in a broader range of applications. The abstract that we presented in our paper [3] reads as follows:

"This slide set presents a brief overview of a compression method we call **Fuzzy Data Compression (FDCMP)**. Our idea is based upon an unusual idea that we first observed a few years ago in a 2003 NASA-CMU paper that showed how high LIDAR image compression ratios (>99%) could be achieved if one is willing to trade a small amount (<3%) of compression imprecision (fuzziness)

to achieve a high compression ratio. In some ways this capability appeared to us to be a similar idea to the "image quality" feature of JPEG image compression. The NASA-CMU paper showed how they were able to achieve outstanding compression results, but their compression speed was very limited due the tools they had available to implement their solution (now obsolete IBM ZISC chips). We believe that UNICON's CogniMax® pattern recognition technology might enable the NASA-CMU compression approach to be more effectively implemented and thus broadly applied."

In summary, believe that our ideas would allow the NASA-CMU fuzzy-feature-indexing based data compression approach to be applied in a way that allows more neurons to be utilized to learn more "features" about a data set. We then describe why we believe that the NASA-CMU compression approach might be ***greatly enhanced*** through the use of our technology [1] to:

- ***Achieve even higher compression ratios with equivalent compression-imprecision (fuzziness), or***
- ***Achieve similar compression-ratios with higher compression-precision (improved fidelity), and***
- ***Greatly enhance compression-speed (always)***

If such an approach were successfully developed, then certain "features" present within massive-scale data sets could be learned and then later recognized using advanced pattern recognition methods to exploit this knowledge as the basis for unusually effective data compression mechanisms. Massive datasets exhibiting repetitive patterns provide extensive opportunities for feature learning and exploitation.

As an example, it might be possible to develop a data compression system capable of 5,000:1 compression while exhibiting <3% compression imprecision (fuzziness). Alternatively, it might be possible to develop data compression systems capable of 500:1 compression while exhibiting <0.03% compression imprecision (fuzziness). In such a case the level of compression imprecision might be far less than the noise floor for the data being compressed. This would mean that the additional "*compression noise*" introduced might be negligible. Such additional noise in the compressed data set would then likely be inconsequential to the scientific analysis that might be performed on the compressed data set. More significantly, if a 100GB experimental data set could even be compressed by only 100:1 while retaining reasonable data fidelity this would provide great benefits in terms of data storage resources, data communication resources, and precious researcher time.

Overall, we suspect that it may be possible to develop data compression systems that can provide a measurable and adjustable trade-off between compression-ratio and compression-precision. This may provide the fusion energy community with important options for the effective management of massive-scale experimental data sets.

3.2 Research Thrusts Needed To Develop Effective Tools

Looking forward across the spectrum of potential needs for ITER and the larger DOE fusion energy community we believe that various opportunities exist for the application of advanced pattern recognition methods as the basis for unusual solutions to address important technical challenges. The following list of research thrusts is provided for consideration:

A. **Advanced Data Compression (ADC):** One research thrust suggested (the focus of this paper) is the study of advanced methods of experimental data compression. Many DOE experiments generate vast quantities of data. Such data must often be transmitted around the world to enable analysis by various scientists. If such data could be highly compressed while maintaining high fidelity this would enable the more effective use of important data storage, data communication, and computing resources. More importantly, it would also enable the more effective utilization of precious scientific personnel time. As a starting point the unusual method described in the papers [4] and [3] are offered for consideration. Using the methods offered, supporting research will likely be needed to determine the effectiveness of and appropriate utilization of the method for different types of experimental data.

Organizational issues: Given the fact that the solution enabling method described in this section [3] originates within the private sector we suggest that the DOE draw upon or partner with private sector resources in a significant way to accelerate the research and development of advanced systems to meet this need within the fusion energy community.

4 Anticipated Research Outcome

Should the DOE OFES look favorably on the research thrust topics listed above this section presents a forward looking ***and admittedly optimistic*** view of what types of tools might be developed as a result of the research and development efforts proposed.

(a) **ADC:** Given a significant and successful research and development effort we believe that fast and unusually effective experimental data compression systems can be developed. Such systems would likely be able to compress extremely large data sets at high rates and exhibit high compression ratios (hopefully, >5000:1) while introducing minimal loss during the compression process (hopefully, <<1%). Such a capability would allow the more effective

utilization of data storage resources, data communication resources, data computing resources, and precious scientific personnel time.

5 Additional References

[1] *CogniMax® pattern recognition technology;* COGNIMAX is a trademark of UNICON Inc.

[2] CADARET, P., "*What Is Exhaustive-Learning?*" (WIEL). Available within the DTIC IR&D collection (www.dtic.mil) with document accession number is 08241207

[3] CADARET, P., "*Fuzzy Data Compression*" (FDCMP). Available within the DTIC IR&D collection (www.dtic.mil) with document accession number is 08241205

[4] Cai, P.; Hu, Y.; Siegel, M.; Gollapalli, S.; Venugopal, A.; Bardak, U.; "*Onboard Feature Indexing from Satellite Lidar Images*"; Proceedings of IEEE IWADC, Perugia, Italy, 2003.