Data Compression using Neural Networks

Master Thesis Task Description

by

Mahesh Sadupalli

M.Sc. Artificial Intelligence

Supervisor:	Prof. Dr.-Ing. Michael Oevermann
Mentor:	M.Sc. Abhishek Dhiman
University:	Brandenburgische Technische Universität Cottbus-Senftenberg
Faculty I:	Fachgebiet Numerische Mathematik und Wissenschaftliches Rechnen

Abstract

This master's thesis investigates the application of neural networks for concurrent and real-time data compression in streaming spatio-temporal datasets. As modern applications generate increasingly large data volumes due to higher resolutions and longer runtimes, traditional storage and post-processing approaches face significant I/O bottlenecks and scalability limitations. This work proposes an in-situ and in-transit compression framework that employs deep learning neural networks to learn compact representations of data during runtime.

The methodology integrates neural networks that approximate data patterns as continuous functions of their inputs, replacing large discrete datasets with a compact set of network parameters. This enables concurrent, real-time compression without interrupting primary workflows, reducing the need for storing full data snapshots while maintaining sufficient accuracy for downstream analysis and visualization.

The framework is validated across diverse test cases, ranging from simplified benchmarks to complex, real-world scenarios. Evaluation metrics include data reduction effectiveness, reconstruction accuracy, and computational overhead. Results demonstrate substantial reduction in storage requirements with minimal impact on runtime performance.

The in-situ and in-transit implementation leverages modern computing infrastructures, utilizing CPUs and high-performance computing environments for data generation, neural network training, and inference. This work contributes to the broader field of machine learning–driven data management by providing a practical solution for handling large-scale streaming datasets in real-time applications.

Task Description

Project Objective

Develop and implement an in-situ and in-transit data compression system using deep learning neural networks, integrated with modern computing frameworks to enable concurrent, real-time processing and visualization of large-scale spatio-temporal datasets.

Problem Statement

Modern scientific simulations generate massive amounts of streaming spatio-temporal data that overwhelm traditional storage and post-processing workflows. Current approaches require:

- Storing complete field data at every timestep
- Significant I/O overhead during simulation
- Large storage requirements for time-resolved datasets

These limitations create bottlenecks that prevent efficient utilization of computational resources and delay scientific insights from simulation data.

Proposed Approach/Methodology

An in-situ and in-transit neural network-based compression system that:

- Learns compact field representations during simulation runtime
- Eliminates traditional file I/O bottlenecks through memory-based data staging
- Provides real-time compression and decompression capabilities
- Maintains engineering accuracy for analysis and visualization

The core methodology involves training deep learning neural networks to learn mappings from input coordinates to target variables, where the network approximates complex patterns as continuous functions of the inputs, effectively replacing large discrete datasets with a compact set of network parameters.

Expected Outcomes

- Functional in-situ and in-transit compression system with data generator
- Neural network architecture optimized for high-dimensional data approximation
- Performance benchmarks against conventional approaches
- Open-source implementation with full documentation
- Best practices for integrating neural networks into large-scale data workflows
- Framework for real-time data compression

Evaluation Metrics

- **Compression performance:** assessed through data reduction effectiveness and storage efficiency compared with conventional methods.
- **Accuracy:** evaluated using reconstruction error (e.g., mean squared error thresholds) and preservation of key structural features relevant to downstream analysis.
- **Performance:** measured by runtime overhead (maintaining minimal impact on primary workflows), scalability across large datasets, and memory efficiency through in-memory processing.
- **Benchmarking and validation:** include both simplified test cases for basic validation and more complex, real-world scenarios to assess general applicability.

Conclusion

This project addresses a critical challenge in large scale data processing by developing a practical solution for concurrent and real-time compression using neural networks. The in-situ and in-transit approaches eliminate traditional I/O bottlenecks while maintaining accuracy, contributing to the emerging field of machine learning enhanced data management. The framework's modular design ensures applicability across diverse application domains and provides a foundation for future developments in data compression.