VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CSI3019- ADVANCED DATA COMPRESSION TECHNIQUES

HYBRID COMPRESSION USING ARITHMETIC CODING AND RECURRENT NEURAL NETWORK

TEAM MEMBERS

PRAVEENA S – 22MID0106

PRARTHANA S – 22MID0118

BOOMIKA S – 22MID0126

ABSTRACT

The rapid growth of digital information has created a strong need for efficient data compression techniques that can minimize storage requirements while ensuring accurate and lossless recovery of the original data. Traditional entropy-based compression algorithms such as Shannon–Fano and Huffman coding use static probability models that assign fixed code lengths based solely on symbol frequencies. Although these methods are simple and fast, their compression performance is limited because they do not consider the contextual relationship between consecutive symbols. As a result, these techniques become inefficient for complex, patterned, or highly structured data.

This project proposes a **hybrid compression model that integrates Recurrent Neural Networks (RNN) with Arithmetic Coding** to significantly improve compression efficiency. In the proposed system, an RNN is trained to predict the probability of the next byte based on the sequence of previously encountered bytes. Unlike static probability models, the RNN generates **adaptive, context-aware probability distributions** that capture long-term dependencies and repeating patterns in the data. These dynamic probabilities are then used by an arithmetic encoder, which converts highly accurate predictions into compact fractional-length codes, achieving near-optimal compression according to information theory.

The system architecture includes four major components: file preprocessing, neural probability modelling using an LSTM-based RNN, arithmetic encoding using predicted probabilities, and lossless reconstruction through arithmetic decoding with the same neural model. The implementation is carried out in Python using PyTorch, and the model is tested on general binary data. The results demonstrate a **significant improvement in compression ratio**, where the proposed RNN + Arithmetic Coding method achieves **up to 85.6% reduction in file size**, compared to **43% reduction using Shannon–Fano coding** on the same dataset. The SHA-256 hash verification confirms that the decoded output perfectly matches the original input, ensuring complete loss lessness.

This hybrid approach effectively resolves the major limitations of classical compression algorithms. While arithmetic coding alone depends heavily on static or inaccurate probability models, the RNN provides highly accurate and adaptive predictions. This allows the encoder to produce much smaller code lengths. The system also avoids the zero-frequency problem, handles long-range context, and adapts to complex data patterns automatically through training.

Overall, this project demonstrates that combining machine learning–based probability modelling with arithmetic coding significantly enhances compression performance. The findings highlight the potential of neural-assisted compression systems in modern applications where data volume is high and storage efficiency is critical. This hybrid design can be extended to text, images, audio, and other structured data, paving the way for advanced neural compression solutions.