Information-theory

Shalev Habany

September 2024

1 Introduction

We will introduce here the compression algorithms that we have implemented in our project and also compare it to the Lempel-Ziv algorithm that is implemented in the windows file system when using the zip files.

2 Methods

2.1 Arithmetic Coding

Arithmetic coding is a form of entropy encoding used in lossless data compression. Arithmetic coding signs a sequence of symbols as a single number in the interval [0, 1). This is achieved by recursively subdividing the interval based on the probabilities of the symbols. The more probable a symbol, the larger the subinterval it occupies. As more symbols are processed, the interval narrows, and the final number uniquely represents the entire sequence. This method is highly efficient and can approach the theoretical limit of compression defined by the entropy of the source.

2.2 Huffman Coding

Huffman coding is another form of entropy encoding used in lossless data compression. It assigns variable-length codes to input characters, with shorter codes assigned to more frequent characters. The process involves building a binary tree where each leaf node represents a character from the input. The tree is constructed by repeatedly merging the two nodes with the lowest frequencies until

only one node remains, which becomes the root of the tree. The path from the root to each leaf node determines the code for that character.

3 Our Experiment

We compared the two algorithms that were explained above and implemented it in python. Then we used it to compress (and decompress just for a check that the compression works) the "dickens.txt" file and checked what gives us the best performance in compression to zip which uses Lampel-Ziv.

4 Results

Algorithm	**Original Size (bytes)**	**Compressed Size (bytes)**
Arithmetic Coding	30721	16385
Huffman Coding	30721	17418
Lempel-Ziv (zip)	30721	11833

Table 1: Comparison of Compression Algorithms

In The bottom line the LZ algorithm compressed the file better than the algorithms we implemented, however we also saw that the arithmetic coding performed the compression on this file better then the huffman codding.