Data Compression

Michael J. Watts

http://mike.watts.net.nz

Lecture Outline

- What is compression?
- Huffman encoding
- Lempel-Ziv encoding
- Fraunhofer compression
- JPEG compression

What is Compression?

- Reduction of the length of symbols used to represent data
  - Symbols can be anything
- Based on redundancy
  - Information theory
- Lossless vs lossy
- Degree of compression depends on amount of redundancy
  - Random data compression?

What is Compression?

- Represent symbols with codewords
  - One symbol == one codeword
  - Length of codeword < length of symbol
- Many algorithms exist
  - Huffman
  - Lempel-Ziv
  - Fraunhofer

Huffman Encoding

- Based on probability (frequency) of symbols
- Assign codewords as bits
- Builds a binary tree based on probability
  - Frequent / high probability symbols at the top
  - Infrequent / low probability symbols lower down
  - Branches labelled with 0 and 1
- Frequent symbols get short bit strings
- Infrequent symbols get longer bit strings

Huffman Encoding

- Algorithm
  - Arrange symbols in order of decreasing probability
  - Combine two symbols with lowest probability
    - Assign a new name, add their probabilities
  - Rebuild the list
  - Combine the next two lowest symbols
  - Repeat until there is one symbol, with probability of one
  - Build a binary tree
  - Form codewords by tracing down the tree to the symbol

Huffman Encoding

- Decompression is the opposite of compression
  - Follow bit sequence to a terminal node
- Needs to see the entire data set
  - Must know symbol probabilities
- Also combined with other techniques
  - MP3 encoding

Lempel-Ziv

- Dictionary based encoder
- Lossless
- Replaces phrases with tokens
  - Tokens smaller than the phrases
- Converts input stream to compressed output stream
- Doesn't need to see entire data set
  - Dictionary is built as it moves through the stream

Lempel-Ziv

- Separates input stream into tokens
- Each token represents the shortest phrase that has not been seen
- Tokens are numbered
- Tokens contain other tokens
- Compression is inefficient at the start
- Better later
  - Larger dictionary

Fraunhofer / MP3 Encoding

- Lossy algorithm
- Compression of audio data
- Perceptual encoding
  - Psycho-acoustic model
  - Throws out what can't be heard
    - Frequency bands
    - Masking effects
- Output is further compressed
  - Uses Huffman encoding

MP3 Encoding

- Only usable on audio
- Part of the MPEG standard
  - DVDs
- Rather widely used by itself

JPEG Compression

- Joint Photographic Experts Group
- Used to compress images
- Compresses groups of pixels
- Uses a Discrete Cosine Transformation
  - DCT
- Quantises the results of the DCT
  - Lossy!
- Further compression is applied

Summary

- Compression is an application of information theory
- Encodes long strings as shorter strings
- All schemes depend on redundant / repetitive data
- Common gotchas
  - A compression algorithm cannot compress its own output
  - Cannot easily compress completely random data
    - No / little redundancy