**Compression Outline**

**Introduction**: Lossy vs. Lossless, Benchmarks, ...

**Information Theory**: Entropy, etc.

**Probability Coding**: Huffman + Arithmetic Coding

**Applications of Probability Coding**: PPM + others

**Lempel-Ziv Algorithms**: LZ77, gzip, compress, ...

**Other Lossless Algorithms**: Burrows-Wheeler

- **Lossy algorithms for images**: JPEG, MPEG, ...
  - Scalar and vector quantization
  - JPEG and MPEG

**Compressing graphs and meshes**: BBK

---

**Scalar Quantization**

Quantize regions of values into a single value:

- **output**
  - uniform
  - non uniform

Can be used to reduce # of bits for a pixel

---

**Vector Quantization**

- **Generate Vector**
  - Find closest code vector
- **Encode**
  - Codebook Index
  - Index
  - Codebook
- **Generate Output**
  - Decode
**Vector Quantization**

What do we use as vectors?
- Color (Red, Green, Blue)
  - Can be used, for example to reduce 24 bits/pixel to 8 bits/pixel
  - Used in some terminals to reduce data rate from the CPU (colormaps)
- K consecutive samples in audio
- Block of K pixels in an image

How do we decide on a codebook?
- Typically done with clustering

---

**Linear Transform Coding**

Want to encode values over a region of time or space
- Typically used for images or audio

Select a set of linear basis functions \( \phi_i \) that span the space
- sin, cos, spherical harmonics, wavelets, ...
- Defined at discrete points

**Vector Quantization: Example**

---

**Linear Transform Coding**

Coefficients:

\[
\Theta_i = \sum_j x_j \phi_i(j) = \sum_j x_j a_{ij}
\]

- \( \Theta_i \) = \( i^{th} \) resulting coefficient
- \( x_j \) = \( j^{th} \) input value
- \( a_{ij} \) = \( ij^{th} \) transform coefficient = \( \phi_i(j) \)

In matrix notation:

\[
\begin{align*}
\Theta &= Ax \\
x &= A^{-1}\Theta
\end{align*}
\]

Where \( A \) is an \( n \times n \) matrix, and each row defines a basis function
**Example: Cosine Transform**

\[ \phi_0(j) \quad \phi_1(j) \quad \phi_2(j) \quad \ldots \]
\[ \Theta_i = \sum_j x_j \phi_i(j) \]

**How to Pick a Transform**

**Goals:**
- Decorrelate
- Low coefficients for many terms
- Basis functions that can be ignored by perception

Why is using a Cosine of Fourier transform across a whole image bad?

How might we fix this?

**Other Transforms**

**Polynomial:**
\[ 1 \quad x \quad x^2 \]

**Wavelet (Haar):**

**Usefulness of Transform**

Typically transforms \( A \) are orthonormal: \( A^{-1} = A^T \)

**Properties of orthonormal transforms:**
- \( \sum x^2 = \sum \Theta^2 \) (energy conservation)

Would like to compact energy into as few coefficients as possible

\[ G_{TC} = \frac{1}{n} \sum_{i=1}^{n} \frac{\sigma_i^2}{\prod \sigma_i^2} \] (the **transform coding gain**)

\[ \bar{\sigma}_i = (\Theta_i - \Theta_{\text{ave}}) \]

The higher the gain, the better the compression
Case Study: JPEG

A nice example since it uses many techniques:
- Transform coding (Cosine transform)
- Scalar quantization
- Difference coding
- Run-length coding
- Huffman or arithmetic coding

JPEG (Joint Photographic Experts Group) was designed in 1991 for lossy and lossless compression of color or grayscale images. The lossless version is rarely used.

Can be adjusted for compression ratio (typically 10:1)

JPEG in a Nutshell

JPEG: Quantization Table

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>11</td>
<td>10</td>
<td>16</td>
<td>24</td>
<td>40</td>
<td>51</td>
<td>61</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>14</td>
<td>19</td>
<td>26</td>
<td>58</td>
<td>60</td>
<td>55</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>16</td>
<td>24</td>
<td>40</td>
<td>57</td>
<td>69</td>
<td>56</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>22</td>
<td>29</td>
<td>51</td>
<td>87</td>
<td>80</td>
<td>62</td>
</tr>
<tr>
<td>18</td>
<td>22</td>
<td>37</td>
<td>56</td>
<td>68</td>
<td>109</td>
<td>103</td>
<td>77</td>
</tr>
<tr>
<td>24</td>
<td>35</td>
<td>55</td>
<td>64</td>
<td>81</td>
<td>104</td>
<td>113</td>
<td>92</td>
</tr>
<tr>
<td>49</td>
<td>64</td>
<td>78</td>
<td>87</td>
<td>103</td>
<td>121</td>
<td>120</td>
<td>101</td>
</tr>
<tr>
<td>72</td>
<td>92</td>
<td>95</td>
<td>98</td>
<td>112</td>
<td>100</td>
<td>103</td>
<td>99</td>
</tr>
</tbody>
</table>

Also divided through uniformly by a quality factor which is under control.

JPEG: Block scanning order

Uses run-length coding for sequences of zeros
Case Study: MPEG

Pretty much JPEG with interframe coding

Three types of frames
- \( I \) = intra frame (aprox. JPEG) anchors
- \( P \) = predictive coded frames
- \( B \) = bidirectionally predictive coded frames

Example:
Type: \( I \) B \( B \) P \( B \) \( B \) P \( B \) \( B \) \( B \) I
Order: 1 3 4 2 6 7 5 9 10 8 12 13 11

\( I \) frames are used for random access.

MPEG: Compression Ratio

356 x 240 image

<table>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
<th>Compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I )</td>
<td>18KB</td>
<td>7/1</td>
</tr>
<tr>
<td>( P )</td>
<td>6KB</td>
<td>20/1</td>
</tr>
<tr>
<td>( B )</td>
<td>2.5KB</td>
<td>50/1</td>
</tr>
<tr>
<td>Average</td>
<td>4.8KB</td>
<td>27/1</td>
</tr>
</tbody>
</table>

30 frames/sec x 4.8KB/frame x 8 bits/byte = 1.2 Mbits/sec + .25 Mbits/sec (stereo audio)

HDTV has 15x more pixels
= 18 Mbits/sec
MPEG in the “real world”

- DVDs
  - Adds "encryption" and error correcting codes
- Direct broadcast satellite
- HDTV standard
  - Adds error correcting code on top
- Storage Tech “Media Vault”
  - Stores 25,000 movies

Encoding is much more expensive than encoding. Still requires special purpose hardware for high resolution and good compression.

Wavelet Compression

- A set of localized basis functions
- Avoids the need to block

**mother function** \( \phi(x) \)

\[ \phi_{s,l}(x) = \phi(2^s x - l) \]

\( s = \) scale \quad \( l = \) location

**Requirements**

\[ \int_{-\infty}^{\infty} \phi(x) dx = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} |\phi(x)|^2 dx < \infty \]

Many mother functions have been suggested.

Haar Wavelets

Most described, least used.

\[ \phi(x) = \begin{cases} 
1 & 0 \leq x < 1/2 \\
-1 & 1/2 \leq x < 1 \\
0 & \text{otherwise}
\end{cases} \]

Many mother functions have been suggested.

Haar Wavelet in 2d

+ DC component = \( 2^{k+1} \) components
Discrete Haar Wavelet Transform

How do we convert this to the wavelet coefficients?

for (j = n/2; j >= 1; j = j/2) {
    for (i = 1; i < j; i++) {
        b[i] = (a[2i-1] + a[2i])/2;
        b[j+i] = (a[2i-1] – a[2i])/2;
    }
    a[1..2*j] = b[1..2*j];
}

Linear time!

Haar Wavelet Transform: example

a = 2 1 2 -1 -2 0 2 -2
   = 1.5 .5 -1 0 5 1.5 -1 2
   = 1 -.5 .5 -.5
   .25 .75
a = .25 .75 .5 .5 .5 1.5 -1 2

Wavelet decomposition
Morlet Wavelet

\[ \phi(x) = \text{Gaussian} \times \cos(x) = e^{-x^2/2} \cos(5x) \]

Corresponds to wavepackets in physics.

Daubechies Wavelet

JPEG2000

**Overall Goals:**
- High compression efficiency with good quality at compression ratios of .25bpp
- Handle large images (up to \(2^{32} \times 2^{32}\))
- Progressive image transmission
  - Quality, resolution or region of interest
- Fast access to various points in compressed stream
- Pan and Zoom while only decompressing parts
- Error resilience

JPEG2000: Outline

Main similarities with JPEG
- Separates into Y, I, Q color planes, and can downsample the I and Q planes
- Transform coding

Main differences with JPEG
- Wavelet transform
  - Daubechies 9-tap/7-tap (irreversible)
  - Daubechies 5-tap/3-tap (reversible)
- Many levels of hierarchy (resolution and spatial)
- Only arithmetic coding
JPEG2000: 5-tap/3-tap

\[ h[i] = a[2i-1] - (a[2i] + a[2i-2])/2; \]
\[ l[i] = a[2i] + (h[i-1] + h[i] + 2)/2; \]

- \( h[i] \): is the "high pass" filter, ie, the **differences**
  - it depends on 3 values from a (3-tap)
- \( l[i] \): is the "low pass" filter, ie, the **averages**
  - it depends on 5 values from a (5-tap)

Need to deal with boundary effects.
This is reversible: assignment

---

JPEG 2000: Outline

- **A spatial and resolution hierarchy**
  - **Tiles:** Makes it easy to decode sections of an image. For our purposes we can imagine the whole image as one tile.
  - **Resolution Levels:** These are based on the wavelet transform. High-detail vs. Low detail.
  - **Precinct Partitions:** Used within each resolution level to represent a region of space.
  - **Code Blocks:** blocks within a precinct
  - **Bit Planes:** ordering of significance of the bits

---

JPEG2000: Precincts

---

JPEG vs. JPEG2000

- **JPEG:** .125bpp
- **JPEG2000:** .125bpp
Compression Outline
Introduction: Lossy vs. Lossless, Benchmarks, ...
Information Theory: Entropy, etc.
Probability Coding: Huffman + Arithmetic Coding
Applications of Probability Coding: PPM + others
Lempel-Ziv Algorithms: LZ77, gzip, compress, ...
Other Lossless Algorithms: Burrows-Wheeler
Lossy algorithms for images: JPEG, MPEG, ...
Compressing graphs and meshes: BBK

Compressing Structured Data
So far we have concentrated on Text and Images, compressing sound is also well understood.
What about various forms of "structured" data?
- Web indexes
- Triangulated meshes used in graphics
- Maps (mapquest on a palm)
- XML
- Databases

Compressing Graphs
Goal: To represent large graphs compactly while supporting queries efficiently
  - e.g., adjacency and neighbor queries
  - want to do significantly better than adjacency lists (e.g. a factor of 10 less space, about the same time)
Applications:
  - Large web graphs
  - Large meshes
  - Phone call graphs

How to start?
Lower bound for n vertices and m edges?
1. If there are N possible graphs then we will need log N bits to distinguish them
2. in a directed graph there are n² possible edges (allowing self edges)
3. we can choose any m of them so N = (n² choose m)
4. We will need log (n² choose m) = O(m log (n²/m)) bits in general
For sparse graphs (m = kn) this is hardly any better than adjacency lists (perhaps factor of 2 or 3).
**What now?**

Are all graphs equally likely?
Are there properties that are common across “real world” graphs?
Consider
- link graphs of the web pages
- map graphs
- router graphs of the internet
- meshes used in simulations
- circuit graphs

**LOCAL CONNECTIONS / SMALL SEPARATORS**

---

**Edge Separators**

An edge separator for \((V,E)\) is a set of edges \(E' \subseteq E\) whose removal partitions \(V\) into two components \(V_1\) and \(V_2\).

**Goals:**
- balanced \((|V_1| \approx |V_2|)\)
- small \(|E'|\) is small

A class of graphs \(S\) satisfies a \(f(n)\)-edge separator theorem if
\[
\exists \alpha < 1, \beta > 0 \quad \forall (V,E) \in S, \exists \text{ separator } E',

|E'| < \beta f(|V|),

|V_i| < \alpha |V|, i = 1,2
\]

Can also define vertex separators.

---

**Separable Classes of Graphs**

Planar graphs: \(O(n^{1/2})\) separators
Well-shaped meshes in \(\mathbb{R}^d\): \(O(n^{1-1/d})\) [Miller et al.]
Near-neighbor graphs
In practice, good separators from circuit graphs, street graphs, web connectivity graphs, router connectivity graphs

Note: All separable classes of graphs have bounded density (\(m\) is \(O(n)\))

---

**Main Ideas**

- Number vertices so adjacent vertices have similar numbers
  - Use separators to do this
- Use difference coding on adjacency lists
- Use efficient data structure for indexing
Compressed Adjacency Tables

Log-sized Codes

Log-sized code: Any prefix code that takes $O(\log(d))$ bits to represent an integer $d$.
Gamma code, delta code, skewed Bernoulli code

Example: Gamma code
Prefix: unary code for $\lfloor \log d \rfloor$
Suffix: binary code for $d-2^{\lfloor \log d \rfloor}$
(binary code for $d$, except leading 1 is implied)

Difference Coding
For each vertex, encode:
- Degree
- Sign of first entry
- Differences in adjacency list
Concatenate vertex encodings to encode the graph
Theorem (edge separators)

Any class of graphs that allows $O(n^e)$ edge separators can be compressed to $O(n)$ bits with $O(1)$ access time using:
- Difference coded adjacency lists
- $O(n)$-bit indexing structure

Performance: Adjacency Table

<table>
<thead>
<tr>
<th>Graph</th>
<th>Array time</th>
<th>Array space</th>
<th>List time</th>
<th>List space</th>
<th>bu-cf/semi time</th>
<th>bu-cf/semi space</th>
</tr>
</thead>
<tbody>
<tr>
<td>auto</td>
<td>0.24</td>
<td>34.2</td>
<td>0.61</td>
<td>66.2</td>
<td>0.51</td>
<td>7.17</td>
</tr>
<tr>
<td>feocean</td>
<td>0.04</td>
<td>37.6</td>
<td>0.08</td>
<td>69.6</td>
<td>0.09</td>
<td>11.75</td>
</tr>
<tr>
<td>m14b</td>
<td>0.11</td>
<td>34.1</td>
<td>0.29</td>
<td>66.1</td>
<td>0.24</td>
<td>6.70</td>
</tr>
<tr>
<td>ibm17</td>
<td>0.15</td>
<td>33.3</td>
<td>0.40</td>
<td>65.3</td>
<td>0.34</td>
<td>7.72</td>
</tr>
<tr>
<td>ibm18</td>
<td>0.14</td>
<td>33.5</td>
<td>0.38</td>
<td>65.5</td>
<td>0.32</td>
<td>7.33</td>
</tr>
<tr>
<td>CA</td>
<td>0.34</td>
<td>43.4</td>
<td>0.56</td>
<td>75.4</td>
<td>0.58</td>
<td>11.66</td>
</tr>
<tr>
<td>PA</td>
<td>0.19</td>
<td>43.3</td>
<td>0.31</td>
<td>75.3</td>
<td>0.32</td>
<td>11.68</td>
</tr>
<tr>
<td>google</td>
<td>0.24</td>
<td>37.7</td>
<td>0.49</td>
<td>69.7</td>
<td>0.45</td>
<td>7.86</td>
</tr>
<tr>
<td>googleO</td>
<td>0.24</td>
<td>37.7</td>
<td>0.50</td>
<td>69.7</td>
<td>0.51</td>
<td>9.90</td>
</tr>
<tr>
<td>lucent</td>
<td>0.02</td>
<td>42.0</td>
<td>0.04</td>
<td>74.0</td>
<td>0.05</td>
<td>11.87</td>
</tr>
<tr>
<td>scan</td>
<td>0.04</td>
<td>43.4</td>
<td>0.06</td>
<td>75.4</td>
<td>0.08</td>
<td>12.85</td>
</tr>
</tbody>
</table>

Time is to create the structure, normalized to time for DFS

Performance: Overall

Conclusions

$O(n)$-bit representation of separable graphs with $O(1)$-time queries
Space efficient and fast in practice for a wide variety of graphs.
Compression Summary

Compression is all about probabilities

\[ \text{Model} \rightarrow \text{Coder} \]

\[ \text{Message} s \in \mathbb{N} \]

We want the model to skew the probabilities as much as possible (i.e., decrease the entropy)

\[ \text{Compress} \]

\[ |w| = 1_{\text{sh}}(s) = -\log p(s) \]

How do we figure out the probabilities

- Transformations that skew them
  - Guess value and code difference
  - Move to front for temporal locality
  - Run-length
  - Linear transforms (Cosine, Wavelet)
  - Renumber (graph compression)

- Conditional probabilities
  - Neighboring context

In practice one almost always uses a combination of techniques