Data Compression

Ikjun Yeom
Self-Information and Entropy

- Self-information
  - The amount of information contained in an event
  - defined by $i(A) = - \log P(A)$
  - Rare events reveal more information.

- Entropy
  - The amount of information revealed by a sequence of events
  - defined by the weighted average of the self-information of each events
  - $H = \sum P(A_i)i(A_i) = - \sum P(A_i) \log P(A_i)$
Entropy - Example

- Consider the following sequence

  \[ 12323454567898910 \]

- \( P(1) = P(6) = P(7) = P(10) = \frac{1}{16} \)

- \( P(2) = P(3) = P(4) = P(5) = P(8) = P(9) = \frac{2}{16} \)

- \( H = - \sum_{i=1}^{10} P(i) \log_2 P(i) = 3.25 \text{ bits/sample} \)
Compression Types

- Lossless compression
- Lossy compression
Lossless Compression Algorithms

- Huffman coding
- Run-length coding
  - aaabbcdddd $\rightarrow$ 3a2b1c4d
- Dictionary-based methods
Huffman Coding

Based on the following observations:

• In an optimum code, symbols that occur more frequently will have shorter codewords than symbols that occur less frequently.

• In an optimal code, the two symbols that occur least frequently will have the same length.
Huffman Coding - Example

<table>
<thead>
<tr>
<th>Letter</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>a2</td>
<td>0.4</td>
</tr>
<tr>
<td>a1</td>
<td>0.2</td>
</tr>
<tr>
<td>a3</td>
<td>0.2</td>
</tr>
<tr>
<td>a4</td>
<td>0.1</td>
</tr>
<tr>
<td>a5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

\[
a2(0.4) \quad a2(0.4) \quad a2(0.4) \quad a3''(0.6)
\]

\[
a1(0.2) \quad a1(0.2) \quad a3'(0.4) \quad 0 \quad a2(0.4)
\]

\[
a3(0.2) \quad a3(0.2) \quad 0 \quad a1(0.2) \quad 1
\]

\[
a4(0.1) \quad 0 \quad a4'(0.2) \quad 1
\]

\[
a5(0.1) \quad 1
\]
Disctionary-Based Methods

- Static disctionary
  - useful when considerable prior knowledge about the source is available

- Adaptive dictionary
LZ77 Compression

<Offset, match length, letter after the match>

Here

Source: ...cabracadabrarrarrad...

...cabracadabrarrarrad... <0,0,d>

Search buf, look-ahead buf.

...cabracadabrarrarrad... <7,4,r>

...cabracadabrarrarrad... <3,5,d>
Image Compression (JPEG)

- Joint Photographic Experts Group
- Used to compress still images
DCT Phase

- Discrete Cosine Transform
- transforming an 8x8 matrix of pixel values to 8x8 matrix of frequency coefficients
- no loss here
DCT Example
DCT Example
### DCT Example

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>-4</td>
<td>-4</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-4</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>149</td>
<td>113</td>
<td>115</td>
<td>101</td>
<td>107</td>
<td>61</td>
<td>66</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>17</td>
<td>24</td>
<td>-32</td>
<td>-30</td>
<td>12</td>
<td>-37</td>
<td>-72</td>
<td></td>
</tr>
<tr>
<td>-62</td>
<td>-47</td>
<td>-51</td>
<td>4</td>
<td>11</td>
<td>-25</td>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-6</td>
<td>-11</td>
<td>0</td>
<td>17</td>
<td>31</td>
<td>35</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td>8</td>
<td>0</td>
<td>-25</td>
<td>-27</td>
<td>-5</td>
<td>-2</td>
<td>-19</td>
<td></td>
</tr>
<tr>
<td>-33</td>
<td>-9</td>
<td>14</td>
<td>21</td>
<td>8</td>
<td>-5</td>
<td>-1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>
Quantization Phase

- dropping the insignificant bits of the frequency coefficients.

\[
\begin{array}{cccccccc}
3 & 5 & 7 & 9 & 11 & 13 & 15 & 17 \\
5 & 7 & 9 & 11 & 13 & 15 & 17 & 19 \\
7 & 9 & 11 & 13 & 15 & 17 & 19 & 21 \\
9 & 11 & 13 & 15 & 17 & 19 & 21 & 23 \\
11 & 13 & 15 & 17 & 19 & 21 & 23 & 25 \\
13 & 15 & 17 & 19 & 21 & 23 & 25 & 27 \\
15 & 17 & 19 & 21 & 23 & 25 & 27 & 29 \\
17 & 19 & 21 & 23 & 25 & 27 & 29 & 31 \\
\end{array}
\]
Encoding Phase

- RLE for 0 coefficients
- Huffman coding for others
JPEG Images

The first element is 1, and others are 100.

Elements in the first row are 1, and others are 100.
Video Compression (MPEG)

- Moving (Motion) Picture Experts Group

- Frame types
  - I (Intrapicture) frame
    * do not use any temporal correlation
  - P (Predicted picture) frame
    * coded using motion-compensated prediction from the last I or P frames whichever happens to be closest.
  - B (Bidirectional predicted picture) frame
    * coded using motion-compensated prediction from both the most recent and the closest future I or P frames.
Sequence of I, P and B Frames

MPEG stream

Bidirectional prediction

Forward prediction

I frame  P frame  B frame