The Information Bottleneck Method

Naftali Tishby, Fernando C. Pereira, William Bialek
The Information Bottleneck Method

Naftali Tishby
The Hebrew University
Jerusalem 91904, Israel
tishby@cs.huji.ac.il

Fernando C. Pereira
AT&T Labs – Research
Florham Park, NJ 07932
pereira@research.att.com

William Bialek
NEC Research Institute
Princeton, NJ 08540
bialek@research.nj.nec.com
What is information bottleneck?

It is a technique for finding the best tradeoff between accuracy and complexity.
Speech compression: A transcript of spoken words has low entropy $\implies$ It can be compressed without losing the information about the words.
Problem Definition

- **Input** signals $x \in X$, and $y \in Y$
  - mapping function $f: X \rightarrow Y$
  - $P(X = x)$, $P(Y = y, X = x)$

- **Output**
  - $X \rightarrow \tilde{X}$
  - $\tilde{X} \rightarrow Y$
Example

1. $X = \text{Speech signal}$  $Y = \text{Transcription signal}$

2. $X = \text{Speech signal}$  $Y = \text{Speakers identity}$
Relevant quantization

Mapping $X \rightarrow \tilde{X}$

- $P(\tilde{x}|x)$ \leftarrow \begin{cases} \text{Soft Partitioning} \\ \text{Hard Partitioning} \end{cases}$
- $P(\tilde{x}) = \sum_x p(x)p(\tilde{x}|x)$
What is a good quantization?

The first factor is the rate, or the average number of bits per message needed to specify an element in the codebook without confusion. This number *per element* of $X$ is bounded from below by the mutual information

$$I(X; \tilde{X}) = \sum_{x \in X} \sum_{\tilde{x} \in \tilde{X}} p(x, \tilde{x}) \log \left[ \frac{p(\tilde{x}|x)}{p(\tilde{x})} \right]$$
\(H(X), I(X, \tilde{X}), H(X|\tilde{X})\)
The average volume of the elements of $X$ that are mapped to the same codeword is $2^{H(X|\tilde{X})}$

$$H(X|\tilde{X}) = -\sum_{x\in X} p(x) \sum_{\tilde{x}\in \tilde{X}} p(\tilde{x}|x) \log p(\tilde{x}|x)$$
Information rate alone is not enough to characterize good quantization since the rate can always be reduced by throwing away details of the original signal $x$. We need therefore some additional constraints.
\[ I(\tilde{X}; Y) = \sum_y \sum_{\tilde{x}} p(y, \tilde{x}) \log \frac{p(y, \tilde{x})}{p(y)p(\tilde{x})} \leq I(X; Y). \]
The information bottleneck

\[ \mathcal{L}[p(\tilde{x}|x)] = I(\tilde{X}; X) - \beta I(\tilde{X}; Y) \]
The optimal assignment, that minimizes previous equation, satisfies the equation

\[ p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x, \beta)} \exp \left[ -\beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})} \right] \]

\( p(y|\tilde{x}) \) can be computed by Bayes’ rule and Markov chain condition \( \tilde{X} \leftarrow X \leftarrow Y \)

\[ p(y|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(y|x)p(\tilde{x}|x)p(x). \]
The information bottleneck iterative algorithm

\[
\begin{aligned}
& p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x, \beta)} \exp(-\beta d(x, \tilde{x})) \\
& p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\
& p_{t+1}(y|\tilde{x}) = \sum_y p(y|x) p_t(x|\tilde{x})
\end{aligned}
\]
The formal solution of the self consistent equations, described above, still requires a specification of the structure and cardinality of $\tilde{X}$, as in rate distortion theory.
Document Clustering using Word Clusters via the Information Bottleneck Method

Noam Slonim and Naftali Tishby

School of Computer Science and Engineering and The Interdisciplinary Center for Neural Computation
The Hebrew University, Jerusalem 91904, Israel
email: {noamm,tishby}@cs.huji.ac.il
a novel implementation of the information bottleneck method for unsupervised document clustering.

- **Input:**  \( X = \) Documents, \( Y = \) Words
  \( P(X) \) and \( P(X, Y) \)
\[
\begin{align*}
p(\tilde{x}|x) &= \frac{p(\tilde{x})}{Z(\beta, x)} \exp(-\beta D_{KL}[p(y|x)\|p(y|\tilde{x})]) \\
p(y|\tilde{x}) &= \frac{1}{p(\tilde{x})} \sum_x p(\tilde{x}|x)p(x)p(y|x) \\
p(\tilde{x}) &= \sum_x p(\tilde{x}|x)p(x),
\end{align*}
\]
Hard Clustering

\[ \beta \to \infty \]

\[
\begin{align*}
p(\tilde{x}|x) &= \begin{cases} 
1 & \text{if } x \in \tilde{x} \\
0 & \text{otherwise}
\end{cases} \\
p(y|\tilde{x}) &= \frac{1}{p(\tilde{x})} \sum_{x \in \tilde{x}} p(x)p(y|x) \\
p(\tilde{x}) &= \sum_{x \in \tilde{x}} p(x).
\end{align*}
\]
**Input:** Joint probability distribution $p(x, y)$

**Output:** A partition of $X$ into $m$ clusters, $\forall m \in \{1...|X|\}$

**Initialization:**
- Construct $\tilde{X} \equiv X$
- $\forall i, j = 1...|X|$, $i < j$, calculate $d_{i,j} = (p(\tilde{x}_i) + p(\tilde{x}_j)) \text{D}_{JS}[p(y|x_i), p(y|x_j)]$

**Loop:**
- For $m = |X| - 1 ... 1$
  - Find the indices $\{i, j\}$ for which $d_{i,j}$ is minimized
  - Merge $\{\tilde{x}_i, \tilde{x}_j\} \Rightarrow \tilde{x}_*$
  - Update $\tilde{X} = \{\tilde{X} - \{\tilde{x}_i, \tilde{x}_j\}\} \cup \{\tilde{x}_*\}$
  - Update $d_{i,j}$ costs w.r.t. $\tilde{x}_*$
- End For
\[
p(\tilde{x}_* | x) = \begin{cases} 
1 & \text{if } x \in \tilde{x}_i \text{ or } x \in \tilde{x}_j \\
0 & \text{otherwise}
\end{cases}
\]

\[
p(y | \tilde{x}_*) = \frac{p(\tilde{x}_i)}{p(\tilde{x}_*)} p(y | \tilde{x}_i) + \frac{p(\tilde{x}_j)}{p(\tilde{x}_*)} p(y | \tilde{x}_j)
\]

\[
p(\tilde{x}_*) = p(\tilde{x}_i) + p(\tilde{x}_j).
\]
\[
\delta I(\tilde{x}_i, \tilde{x}_j) \equiv I(\tilde{X}_{before}; Y) - I(\tilde{X}_{after}; Y)
\]

\[
\delta I(\tilde{x}_i, \tilde{x}_j) = (p(\tilde{x}_i) + p(\tilde{x}_j)) \cdot D_{JS}[p(y|\tilde{x}_i), p(y|\tilde{x}_j)]
\]

\[
D_{JS}[p_i, p_j] = \pi_i D_{KL}[p_i \parallel \bar{p}] + \pi_j D_{KL}[p_j \parallel \bar{p}]
\]

\[
\{p_i, p_j\} \equiv \{p(y|\tilde{x}_i), p(y|\tilde{x}_j)\}
\]

\[
\{\pi_i, \pi_j\} \equiv \left\{ \frac{p(\tilde{x}_i)}{p(\tilde{x}_*)}, \frac{p(\tilde{x}_j)}{p(\tilde{x}_*)} \right\}
\]

\[
\bar{p} = \pi_i p(y|\tilde{x}_i) + \pi_j p(y|\tilde{x}_j).
\]
Input: Joint probability distribution \( p(x, y) \)

First Stage: Find clusters \( \tilde{Y} \) using \( \{p(x|y)\} \)

Second Stage:

- For every \( x \in X \), replace \( p(y|x) \) by the more compact representation \( p(\tilde{y}|x) \)
- Find clusters \( \tilde{X} \) using \( \{p(\tilde{y}|x)\} \)

\[
I(\tilde{X}; \tilde{Y}) \leq I(X; \tilde{Y}) \leq I(X; Y)
\]
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Newsgroups included</th>
<th>#documents per group</th>
<th>Total #documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>sci.crypt, sci.electronics, sci.med, sci.space.</td>
<td>500</td>
<td>2000</td>
</tr>
<tr>
<td>Binary1,2,3</td>
<td>talk.politics.mideast, talk.politics.misc.</td>
<td>250</td>
<td>500</td>
</tr>
<tr>
<td>Multi51,2,3</td>
<td>comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space talk.politics.mideast.</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>Multi101,2,3</td>
<td>alt.atheism, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.sport.hockey</td>
<td>50</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.gun.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Datasets details. For example, for each of the three Binary datasets we randomly chose 500 documents, evenly distributed between the news groups talk.politics.mideast and talk.politics.misc. This resulted in three document collections, Binary1, Binary2 and Binary3, each of which consisted of 500 documents.
<table>
<thead>
<tr>
<th>$\tilde{x}_1$</th>
<th>graphics</th>
<th>motorcycles</th>
<th>baseball</th>
<th>space</th>
<th>mideast</th>
</tr>
</thead>
<tbody>
<tr>
<td>78</td>
<td></td>
<td>3</td>
<td>11</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>$\tilde{x}_2$</td>
<td>3</td>
<td>68</td>
<td>7</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>$\tilde{x}_3$</td>
<td>4</td>
<td>5</td>
<td>59</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>$\tilde{x}_4$</td>
<td>6</td>
<td>14</td>
<td>13</td>
<td>68</td>
<td>13</td>
</tr>
<tr>
<td>$\tilde{x}_5$</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>13</td>
<td>63</td>
</tr>
</tbody>
</table>

Table 2: Contingency table for the $IB_{double}$ algorithm over the $Multi52$ dataset using 10 word-clusters. The accuracy is 0.67.
<table>
<thead>
<tr>
<th>Data/algorithim</th>
<th>$IB_{double}$</th>
<th>$IB_{single}$</th>
<th>$L1_{double}$</th>
<th>$L1_{single}$</th>
<th>$Ward_{double}$</th>
<th>$Ward_{single}$</th>
<th>Complete_{tf–idf}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>0.59</td>
<td>0.49</td>
<td>0.41</td>
<td>0.34</td>
<td>0.33</td>
<td>0.29</td>
<td>0.47</td>
</tr>
<tr>
<td>Binary$_1$</td>
<td>0.70</td>
<td>0.71</td>
<td>0.61</td>
<td>0.62</td>
<td>0.59</td>
<td>0.56</td>
<td>0.67</td>
</tr>
<tr>
<td>Binary$_2$</td>
<td>0.68</td>
<td>0.60</td>
<td>0.60</td>
<td>0.57</td>
<td>0.56</td>
<td>0.51</td>
<td>0.61</td>
</tr>
<tr>
<td>Binary$_3$</td>
<td>0.75</td>
<td>0.70</td>
<td>0.66</td>
<td>0.65</td>
<td>0.60</td>
<td>0.60</td>
<td>0.52</td>
</tr>
<tr>
<td>Multi5$_1$</td>
<td>0.59</td>
<td>0.42</td>
<td>0.43</td>
<td>0.38</td>
<td>0.34</td>
<td>0.27</td>
<td>0.51</td>
</tr>
<tr>
<td>Multi5$_2$</td>
<td>0.58</td>
<td>0.40</td>
<td>0.43</td>
<td>0.36</td>
<td>0.29</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>Multi5$_3$</td>
<td>0.53</td>
<td>0.50</td>
<td>0.46</td>
<td>0.34</td>
<td>0.34</td>
<td>0.29</td>
<td>0.63</td>
</tr>
<tr>
<td>Multi10$_1$</td>
<td>0.35</td>
<td>0.24</td>
<td>0.31</td>
<td>0.24</td>
<td>0.20</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>Multi10$_2$</td>
<td>0.35</td>
<td>0.26</td>
<td>0.28</td>
<td>0.27</td>
<td>0.21</td>
<td>0.20</td>
<td>0.33</td>
</tr>
<tr>
<td>Multi10$_3$</td>
<td>0.35</td>
<td>0.29</td>
<td>0.27</td>
<td>0.28</td>
<td>0.20</td>
<td>0.17</td>
<td>0.34</td>
</tr>
<tr>
<td>Average</td>
<td>0.55</td>
<td>0.46</td>
<td>0.45</td>
<td>0.40</td>
<td>0.37</td>
<td>0.34</td>
<td>0.47</td>
</tr>
</tbody>
</table>