How Information Bottleneck Helps Representation Learning: From the Rate Distortion Theory to $\beta-$VAE

Yichao Cai

yichao.cai@adelaide.edu.au

April 16, 2024

Abstract

This blog is on the topic of Information Bottleneck (IB) principle, clarifying its connection with rate-distortion theory, exploring different optimization strategies for IB, and discussing the application of the Variational Information Bottleneck (VIB) and β-VAE within the context of representation learning. Designed as both a personal reference and a resource for others, this blog aims to provide a holistic view understanding of the IB principle. Due to the breadth of concepts covered, some proofs are not included here. Readers seeking more in-depth exploration are encouraged to consult the references provided. **(The content is currently incomplete and will be updated periodically as I continue to expand my knowledge and understanding of the topic.)**

Contents

1 Preliminaries: Information Measures

Information measures are pivotal to this topic and essential for the majority, if not all, learning objectives aimed at optimizing machine learning models. Consequently, the author is compelled to outline some fundamental definitions and the intuitive implementations of these concepts. Given their foundational nature, readers may opt to proceed directly to section 2.

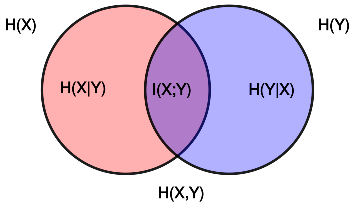

Figure 1: Venn diagram showing relations of information measures (figure taken from wikipedia).

1.1 Entropy Measures

1.1.1 Information Entropy

Given a discrete random variable X, taking values from $\mathcal{X}$ with probability mass function (pmf) $p_X : \mathcal{X} \to [0, 1]$, the information entropy is

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x) \log p_X(x) = \mathbb{E}_{p_X}[-\log p_X(x)]. \tag{1}$$

The entropy measures the average amount of information contained in the random variable X, with values ranging from $0 \leq H(X) \leq log|\mathcal{X}|$. The entropy takes the minimum value 0, when X is a deterministic event, meaning there is no uncertainty about its outcome; and it reaches its maximum value when X is uniformly distributed across its support set, indicating maximum uncertainty or randomness in its outcomes. This maximum uncertainty necessitates more information to represent X, as each outcome is equally likely and no predictions can be made based on probabilities skewed toward particular outcomes.

1.1.2 Joint Entropy

Given a pair of random variables (X, Y) with pmf $p_X : \mathcal{X} \to [0, 1]$, $p_Y : \mathcal{Y} \to [0, 1]$ and joint probability $p_{X,Y} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$, the joint entropy is given by

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y) = \mathbb{E}_{p_{X,Y}}[-\log p_{X,Y}(x, y)]. \tag{2}$$

The joint entropy measures the uncertainty with a set of (X, Y), which indicates how much information we need to represent both X and Y. Clearly, as shown in fig. 1, we know that $\max[H(X), H(Y)] \leq H(X, Y) \leq H(X)+H(Y)$, where the minimum is met when one variable is the subevent of the other while the maximum value is met when two variables are independent.

It is worth noting that the term cross entropy, $H(p, q) = \mathbb{E}_{x\sim p_X}[-\log q_X(x)]$, is a distinct concept from joint entropy, which measures the information needed for encoding two different distributions, i.e., $p_X : \mathcal{X} \to [0, 1]; q_X : \mathcal{X} \to [0, 1]$, over the same underlying set $\mathcal{X}$.

1.1.3 Conditional Entropy

The conditional entropy $H(Y|X)$ quantifies the amount of information required to describe Y when X is known. It is defined as:

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)} = \mathbb{E}_{p_{X,Y}}[-\log p_{Y|X}(y|x)]. \tag{3}$$

We have $H(Y|X) = H(X,Y) - H(X)$. If Y is completely determined by X, then $H(Y|X) = 0$, indicating no remaining uncertainty about Y once X is known. Conversely, if X and Y are independent variables, then $H(Y|X) = H(Y)$, as knowledge of X provides no information about Y. It is important to note that conditional entropy is not symmetric, meaning $H(Y|X)$ is not necessarily equal to $H(X|Y)$.

1.1.4 Kullback–Leibler Divergence

Kullback-Leibler (KL) divergence, also known as relative entropy, quantifies the difference between two probability distributions, p_X and q_X, over the same set $\mathcal{X}$. It is defined as:

$$D_{KL}(p_X\|q_X) = \sum_{x\in\mathcal{X}} p_X(x)\log\frac{p_X(x)}{q_X(x)} = \mathbb{E}_{p_X}[\log\frac{p_X(x)}{q_X(x)}]. \tag{4}$$

KL divergence is non-negative, not upper-bounded, and asymmetric. Formally, it can be expressed as the difference between the cross entropy of p_X with respect to q_X and the entropy of p_X, i.e., $D_{KL}(p_X\|q_X) = H(p_X, q_X) - H(p_X)$.

1.2 Mutual Information

The mutual information (MI) between two variables X and Y quantifies the information gained about one variable through observing the other, reflecting their mutual dependence. It is given by:

$$I(X;Y) = D_{KL}(p_{X,Y}\|p_X \otimes p_Y) = \mathbb{E}_{p_{X,Y}}[\log\frac{p_{X,Y}(x,y)}{p_X(x)\cdot p_Y(y)}]. \tag{5}$$

Mutual information is inherently non-negative and equals zero specifically when X and Y are independent variables. Intuitively, mutual information measures the amount of information shared between two variables: it quantifies how much knowing one variable reduces uncertainty about the other.

Building on these definitions, mutual information can also be related to entropy in the following ways:

$$I(X;Y) = H(Y) - H(Y|X) = H(X,Y) - H(X|Y) - H(Y|X). \tag{6}$$

2 Rate Distortion Theory

The problem of lossy compression investigates how to minimize the quantization rate—thereby maximizing compression—by encoding a signal X into a representation $\tilde{X}$ while maintaining minimal distortion. We define the distortion function[1] $d : \mathcal{X}\times\tilde{\mathcal{X}} \to \mathbb{R}^+$. The compression of X, through a mapping $p(\tilde{x}|x)$, results in an *expected distortion* given by:

$$\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} = \sum_{x\in\mathcal{X}}\sum_{\tilde{x}\in\tilde{\mathcal{X}}} p(x,\tilde{x})\cdot d(x,\tilde{x}). \tag{7}$$

The rate function is quantified by the mutual information (MI) between X and $\tilde{X}$, denoted as $I(X;\tilde{X})$. According to its definition, a lower MI corresponds to a reduced quantization rate (greater compression).

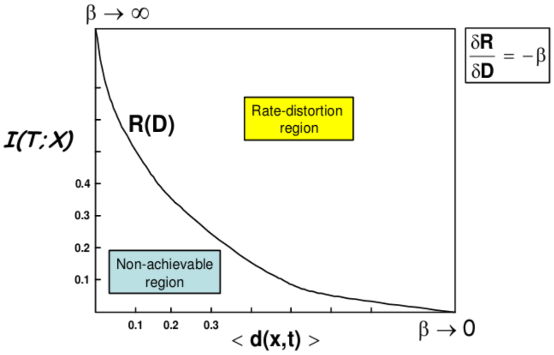

Figure 2: An example of rate distortion curve (figure taken from Slonim_PhD.pdf). The notations T and t mean $\tilde{X}$ and $\tilde{x}$ in this blog.

There exists a monotonic trade-off between these two objectives: the higher the rate, the lower the achievable distortion. Figure 2 illustrates this relationship with a representative case.

The *Rate-Distortion Theorem* elucidates the trade-off between compression rate and distortion by establishing a constraint D, through a expected distortion function. The corresponding rate-distortion (R-D) functional is defined as:

$$R(D) = \min_{p(\tilde{x}|x):\langle d(x,\tilde{x})\rangle \leq D} I(X;\tilde{X}). \tag{8}$$

To solve this functional, one can introduce a Lagrange multiplier, β, associated with the expected distortion term, transforming the objective into minimizing the following variational functional:

$$\mathcal{L}_{RD}[p(\tilde{x}|x)] = I(X;\tilde{X}) + \beta\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})}. \tag{9}$$

This formulation fundamentally seeks to determine the smallest quantization rate achievable under the constraint that the expected distortion does not exceed D. The optimal trade-off is realized at the minimum point of $\mathcal{L}_{RD}$.

2.1 Solution for the R-D functional

To minimize the rate-distortion functional $\mathcal{L}_{RD}$, the derivative is taken with respect to the conditional probability distribution $p(\tilde{x}|x)$ such that $\frac{\delta\mathcal{L}_{RD}}{\delta p(\tilde{x}|x)} = 0$, under the assumption of *normalized distributions* $p(\tilde{x}|x)$. This leads to:

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)}\exp[-\beta d(x,\tilde{x})], \tag{10}$$

where $Z(x,\beta) = \sum_{\tilde{x}} p(\tilde{x})\exp[-\beta d(x,\tilde{x})]$ functions as a normalization factor. Intuitively, $Z(x,\beta)$ determines how the input data X is divided into intervals to form the encoded $\tilde{X}$. Furthermore, the parameter β is determined by the required distortion level, such that $\frac{\delta R}{\delta D} = -\beta$, indicating a positive value due to the concavity of the R-D function, as demonstrated in Figure 2.

2.2 Optimization with Blahut-Arimoto Algorithm

To optimize the rate-distortion functional effectively, we aim to satisfy both equations of $p(\tilde{x})$ and $p(\tilde{x}|x)$ simultaneously at the minimum of the R-D functional, expressed as:

$$\min_{p(\tilde{x})}\min_{p(\tilde{x}|x)} \mathcal{L}_{RD}[p(\tilde{x}|x)]. \tag{11}$$

The Blahut-Arimoto (BA) algorithm provides a systematic method to address this optimization challenge:

[1] Note that the distortion function here is a general form, which could be various specific criteria.

1. *Initialization.* Begin with an initial guess for the marginal distribution $p(\tilde{x})$ based on the data points.
2. *R-D optimization.* Optimize the marginal and conditional distributions *independently* through iterative processing. Let t represent the time step:

$$\begin{cases} p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp[-\beta d(x,\tilde{x})] \end{cases} \tag{12}$$

3. *Iteration.* Calculate the normalization function $Z_t(x, \beta)$ and repeat the optimization until convergence is achieved.

At the convergence of the optimization, the learning object converges to a unique minimum of $\mathcal{L}_{_{RD}}$ within the convex sets of the two distributions. However, the trajectories and the final distributions are not unique, indicating that the model solution is not unique.

The BA algorithm is adept at dividing a data set X into groups or partitions based on an optimality criterion but does not address the selection of the best representatives for each group. Beyond merely partitioning the data, it is essential to select the optimal representatives for each partition to retain *relevant information* for specific tasks related to a variable Y.

3 The Information Bottleneck Principle

The application of rate-distortion theory in practical scenarios is often limited due to the challenge of defining an appropriate distortion function. To overcome this, the Information Bottleneck (IB) method offers a direct and effective approach by focusing on preserving relevant information about another variable Y. This variable Y is assumed to be statistically dependent on X, evidenced by a mutual information (MI) measure $I(X;Y) > 0$. Furthermore, it is assumed access to the joint distribution $p(x, y)$ is available, enabling supervised learning with labeled data. By definition, $I(\tilde{X};Y) \leq I(X;Y)$, indicating no new information is generated during the encoding process; equivalence occurs when the representation does not lose any relevant information about Y.

The core trade-off involves maximizing the use of the minimum bit rate to represent features (*compressing representation*), while simultaneously preserving as much relevant information related to Y as possible (*preserving relevant information*). This trade-off is addressed by minimizing the following functional:

$$\mathcal{L}_{_{IB}}[p(\tilde{x}|x)] = I(\tilde{X};X) - \beta I(\tilde{X};Y), \tag{13}$$

where β is a Lagrange multiplier that emphasizes the importance of preserving meaningful information. This multiplier also helps maintain the normalization of the mapping $p(\tilde{x}|x)$ for each x. By adjusting the hyper-parameter β, one can explore the balance between information preservation and compression.

The intuition behind the Information Bottleneck (IB) principle is clear and compelling: by imposing constraints on the preservation of relevant information, the method encourages simpler and potentially sparser representations. This simplification acts to prevent complicated relationships within the representations, aiding in the disentanglement of semantic information, which in the original data is usually deeply intertwined and complex. By minimizing extraneous information while maximizing the retention of relevant information, the IB principle enhances the clarity and utility of the resulting representations, making them more manageable and insightful for further analysis.

3.1 Solution for IB

The IB method presents a more complex solution compared to the traditional Rate-Distortion (R-D) function, primarily because the constraints weighted by β are directly tied to the encoding model.[2] However, a formal

[2] In the context of this blog, which adopts a representational perspective, the conditional distribution $p(\tilde{x}|x)$ is sometimes referred to as the encoding model or encoder.

solution for minimizing the functional can be achieved. By establishing the Markovian relation $\tilde{X} \leftrightarrow X \leftrightarrow Y$ and assuming β and $p(x, y)$ are given, the conditional $p(t|x)$ reaches a stationary point if and only if,

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x, \beta)} \exp(-\beta D_{KL}[p(y|x) \| p(y|\tilde{x})]), \forall \tilde{x} \in \tilde{\mathcal{X}}, \forall x in \mathcal{X}, \tag{14}$$

where $Z(x, \beta)$ is a normalization function ensuring that $\sum_{\tilde{x}} p(\tilde{x}|x) = 1$ for each x. *It is critical to clarify that the Markovian relation is not a modeling assumption about the latent variable model or the underlying causal structure.* Rather, it defines the problem setting—establishing a linkage between $\tilde{X}$ and Y through X so that the marginal distribution over $p(x, y, \tilde{x})$ relative to X and Y remains consistent with the input distribution $p(x, y)$, indicating that the encoding process does not modify the distribution of inputs[3].

The IB functional can be viewed as a specific, effective R-D function, where the distortion criterion $d(x, \tilde{x}) = D_{KL}[p(y|x) \| p(y|\tilde{x})]$ is guided by the target variable Y but not assumed a priori. Contrary to traditional R-D theory, where encoding is performed through partitioning alone, the IB method focuses on conditional distributions $p(y|\tilde{x})$, highlighting not only the optimal partitioning but also the selection of representatives, emphasizing how the encoded representation actively involves choices concerning the portrayal of the representatives.

3.2 Iteration algorithm for IB

The optimization target, i.e., minimizing the functional, is outlined as follows:

$$\mathcal{L}_{IB}[p(\tilde{x}|x); p(\tilde{x}); p(y|\tilde{x})] = -\langle \log Z(x, \beta) \rangle_{p(x)} = I(X; \tilde{X}) + \beta \langle D_{KL}[p(y|x) \| p(y|\tilde{x})] \rangle_{p(x,\tilde{x})}, \tag{15}$$

With the known joint distribution $p(x, y)$, trade-off parameter β, and convergence criterion ε, this minimization can be conducted independently over the convex sets of the normalized distributions $\{p(\tilde{x}|x)\}$, $\{p(\tilde{x})\}$, and $\{p(y|\tilde{x})\}$.

The iterative optimization process involves:

1. **Initialization.** Randomly initialize $p(\tilde{x}|x)$.
2. **Iteration.** Denote by t the iteration step:

$$\begin{cases} p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta D_{KL}[p(y|x) \| p_t(y|\tilde{x})]) \\ p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_{t+1}(y|\tilde{x}) = \frac{1}{p_{t+1}(\tilde{x})} \sum_x p_t(\tilde{x}|x) p(y|x) p(x) \end{cases} \tag{16}$$

3. **Evaluate.** If $\mathcal{L}_{IB} \leq \varepsilon$, terminate; otherwise, continue the iteration.

As shown in fig. 3, the solutions to these self-consistent equations correspond to a family of annealing curves, all starting from the trivial point (0, 0) in the information plane with infinite slope, parameterized by the value of β. Increasing the value of β allows exploration along convex curves in the information plane, analogous to rate-distortion curves, which exist for every choice of the cardinality of $\tilde{X}$. Interestingly, every two curves in this family separate (bifurcate) at some finite (critical) point through a second-order phase transition. These transitions create a hierarchy of relevant quantizations for different cardinalities of $\tilde{X}$. Similar to R-D theory, the trajectory to the stationary point is independently optimized for the three convex sets, thus the solution is not unique.

[3] In many representation learning scenarios, the latent variable corresponding to the representation is considered a confounder between X and Y. This complicates the model significantly because it challenges the assumption that marginal distributions remain invariant.

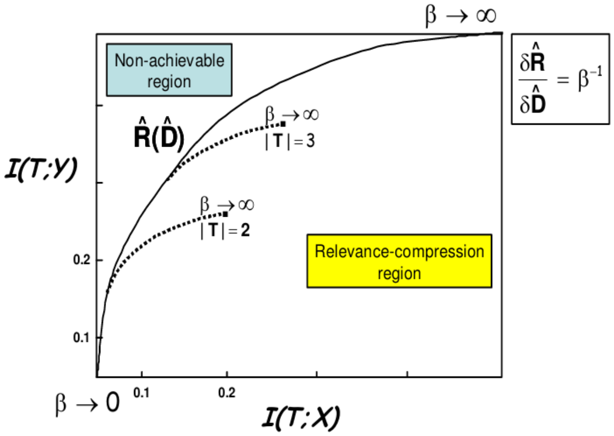

Figure 3: An example of information-compression curve of IB (figure taken from Slonim_PhD.pdf).

References

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.

Slonim, N. (2002). The information bottleneck: Theory and applications (Doctoral dissertation, Hebrew University of Jerusalem).

Tishby, N., & Zaslavsky, N. (2015, April). Deep learning and the information bottleneck principle. In 2015 ieee information theory workshop (itw) (pp. 1-5). IEEE.