Extracting Meaningful Information: from Rate Distortion Theory to the Information Bottleneck Method

30592 Research Project

Samuele Nicolò Straccialini
BAI - Bocconi University

April 16, 2025

Overview

1. Introduction
2. Quantization and Rate Distortion Theory
   - 2.1 What is quantization
   - 2.2 Distortion function
   - 2.3 Blauht-Arimoto Algorithm
3. The Information Bottleneck
4. Applications
   - 4.1 Deep Neural Network
   - 4.2 IB as a measure of optimality
   - 4.3 Generalization Bounds

Introduction

- **Extracting a meaningful representation of data** is a key challenge in machine learning and statistical inference.
- Shannon, in the original formulation of information theory concepts, focuses more on the problem of transmitting information rather than assessing its value.
- Tishby et al. (2000) see information theory as a natural approach to the question of identifying the "relevant" information, especially when dealing with lossy compression.
- Determining **"relevant" information** is strictly related to the actual **definition of relevance**.

Introduction

- The **standard approach** to lossy compression is through **rate-distortion theory**, which balances compression and fidelity.
- This tradeoff is characterized by the rate-distortion function $R(D)$, which measures the lowest possible bit rate R needed to encode a signal, while ensuring that the average distortion D remains below a given threshold level of distortion.
- Problem: **the distortion function**, which determines the relevant features of the signal, **must be specified in advance**. These features, however, are often unknown a priori.
- We will use an information-theoretic approach to extend rate-distortion theory and **derive an iterative algorithm** for finding representations of the signal that capture its essential structure.

Quantization

Quantization refers to **mapping some input signal** x from the space X with a fixed probability measure $p(x)$ to a signal $\tilde{x}$ **in a smaller space** $\tilde{X}$ characterized by the conditional p.d.f. $p(\tilde{x}|x)$. The marginal of the compressed representation is

$$p(\tilde{x}) = \sum_{x} p(\tilde{x}|x)p(x). \tag{1}$$

Quantization

- The **average volume of elements** of X which are mapped to the same codeword $\tilde{x}$ is

$$2^{H(X|\tilde{X})}$$

- $H(X|\tilde{X})$ is the **conditional entropy** $H(X|\tilde{X}) = \sum_{x \in X} p(x) \sum_{\tilde{x} \in \tilde{X}} p(\tilde{x}|x) \log p(\tilde{x}|x)$
- By the standard asymptotic arguments, the **average cardinality of partitioning** X is given by the ratio of its volume to that of the mean partition:

$$\frac{2^{H(X)}}{2^{H(X|\tilde{X})}} = 2^{I(X;\tilde{X})}$$

- $I(X;\tilde{X})$ is the **mutual information**: $I(X;\tilde{X}) = \sum_{x \in X} \sum_{\tilde{x} \in \tilde{X}} p(x,\tilde{x}) \log \frac{p(\tilde{x}|x)}{p(\tilde{x})}$

Distortion function

Mutual information alone is not enough to define a good quantization: we need another constraint that, in classical rate-distortion theory, is provided by the distortion function.

Definition (distortion function)

A **distortion function** or **distortion measure** $d : X \times \tilde{X} \to \mathbb{R}^+$ is a mapping from the set of source alphabet-reproduction alphabet pairs into the set of nonnegative real numbers.

This is the cost of compressing the source signal x with its representation $\tilde{x}$. The **expected distortion** is then:

$$\mathbb{E}_{x,\tilde{x}}[d(x,\tilde{x})] = \langle d(x,\tilde{x}) \rangle_{p(x,\tilde{x})} = \sum_{x \in X} \sum_{\tilde{x} \in \tilde{X}} p(x,\tilde{x}) d(x,\tilde{x}).$$

Rate Distortion Function

The tradeoff between the rate of quantization and the expected distortion is monotonic: the larger the rate, the smaller the achievable distortion.

Theorem (Shannon-Kolmogorov)

The rate distortion function $R(D)$ for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \tilde{x})$ is equal to the associated information rate distortion function:

$$R(D) = \min_{p(\tilde{x}|x):\mathbb{E}_{x,\tilde{x}}[d(x,\tilde{x})]<D} I(X; \tilde{X})$$

Intuitively, the **rate distortion function** $R(D)$ is the minimum number of bits per data unit needed to represent the data so that the expected distortion doesn't exceed some threshold D.

Finding $R(D)$

Finding the rate distortion function is a variational problem, which can be solved by introducing a Lagrange multiplier:

$$\mathcal{F}[p(\tilde{x}|x)] = I(X;\tilde{X}) + \beta\mathbb{E}_{x,\tilde{x}}[d(x,\tilde{x})] \tag{2}$$

Theorem

The solution of the variational problem $\frac{\delta\mathcal{F}[p(\tilde{x}|x)]}{\delta p(\tilde{x}|x)}$ *for normalized distributions* $p(\tilde{x}|x)$ *is given by:*

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} e^{-\beta d(x,\tilde{x})}, \tag{3}$$

where $Z(x,\beta)$ *is a normalization factor.*

Finding $R(D)$

- Notice that equations (1) and (3) **must hold simultaneously**. A natural approach to solve these is to **alternate between them until reaching convergence**, which is assured by the Csiszár-Tusnády Lemma.
- The alternative iteration for calculating the rate distortion function is known as **Blauht-Arimoto (BA)**.
- The BA algorithm optimizes the partitioning of X given the set $\tilde{X}$ of representatives, and not the choice of the representation $\tilde{X}$ itself. In practice, it is also crucial to find the optimal representatives that minimize the expected distortion during partitioning.

Blauht-Arimoto Algorithm

Theorem (Blauht-Arimoto Algorithm)

Equations (1) and (3) are satisfied simultaneously at the minimum of

$$\mathcal{F} = -\mathbb{E}_x[\log Z(x,\beta)] = I(X;\tilde{X}) + \beta\mathbb{E}_{x,\tilde{x}} d(x,\tilde{x}),$$

where minimization is done independently over $\{p(\tilde{x})\}$ *and* $\{p(\tilde{x}|x)\}$*:*

$$\min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \mathcal{F}[p(\tilde{x}); p(\tilde{x}|x)]$$

This corresponds to alternating iterations over the two equations. Denoting t the iteration step,

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

The Information Bottleneck

- Finding the "right" distortion measure to appropriately capture the cost of compression is a challenging task.
- **Instead, we can preserve the relevant information about another variable Y.** We assume the joint distribution $p(x, y)$ is known.
- The goal is still to compress X in $\tilde{X}$ as much as possible. This time, however, we capture as much of the information about Y as possible:

$$I(\tilde{X}; Y) = \sum_{y \in Y} \sum_{\tilde{x} \in \tilde{X}} p(y, \tilde{x}) \log \frac{p(y, \tilde{x})}{p(y)p(\tilde{x})} \leq I(X; Y),$$

The Information Bottleneck

- Similarly to the previous situation, the optimal assignment is found by minimizing the functional
$$\mathcal{L}[p(\tilde{x}|x)] = I(\tilde{X}; X) - \beta I(\tilde{X}; Y), \tag{4}$$
where again, β is the Lagrange multiplier. Notice that at $\beta = 0$, we have the most extreme quantization, with all information shrunk to a single point. For $\beta > 0$ we have increasingly detailed quantization.

- Differently to the rate distortion case, equation (4) is now nonlinear, due to the nonlinearity of the constraint. Surprisingly, it is still possible to obtain an exact solution.

The Information Bottleneck

Theorem

The optimal assignment minimizing equation (4) satisfies

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left[-\beta \sum_{y\in Y} p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})}\right] \tag{5}$$

where $p(y|\tilde{x})$ in the exponent can be found as

$$p(y|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(y|x)p(\tilde{x}|x)p(x). \tag{6}$$

The optimal "distortion measure" is revealed to be $d(x,\tilde{x}) = D_{KL}[p(y|x)||p(y|\tilde{x})]$. This **KL divergence** is not assumed anywhere else, and is therefore natural to be considered as the "correct" distortion in this setting.

The Information Bottleneck

Theorem

Equations (5) and (6) are satisfied simultaneously at the minimum of

$$\mathcal{F} = -\mathbb{E}_x[\log Z(x,\beta)] = I(X;\tilde{X}) + \beta \mathbb{E}_{x,\tilde{x}}\Big[D_{KL}[p(y|x)||p(y|\tilde{x})]\Big], \tag{7}$$

where minimization is done independently over $\{p(\tilde{x})\}$, $\{p(\tilde{x}|x)\}$, *and* $p(y|\tilde{x})$*:*

$$\min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \min_{p(y|\tilde{x})} \mathcal{F}[p(\tilde{x}); p(\tilde{x}|x); p(y|\tilde{x})] \tag{8}$$

This corresponds to alternating iterations over three equations,

$$\begin{cases} p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta D_{KL}[p(y|x)||p(y|\tilde{x})]) \\ p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_{t+1}(y|\tilde{x}) = \sum_y p(y|x) p_t(x|\tilde{x}) \end{cases} \tag{9}$$

Applications

- The information bottleneck principle provides a **unified framework** for addressing various problems, from clustering to document classification or spectral analysis.
- A particularly intriguing application is perhaps to **Deep Neural Networks (DNNs)**. Tishby and Zaslavsky (2015) analyze DNNs in the framework of the IB principle to determine the optimal information-theoretic limits of a DNN and derive finite-sample bounds for the generalization error.
- The goal of DNNs (and supervised learning technique in general) is to capture and efficiently represent the relevant information in the input variable about the output variable. This aligns precisely with the IB method's goal.

Simple DNN

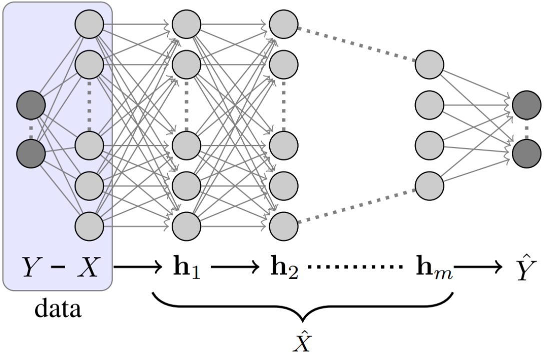

Figure: Simple Neural Network model. An n-dimensional input $x \in X$, where each x is associated to a true label $y \in Y$, is sent through m layers $\mathbf{h}_i$ $(i \in \{1, \ldots, m\})$. The network outputs a predicted label $\hat{y} \in \hat{Y}$, aiming to obtain the true label.

Information of layer

- Each layer processes only the information received from the previous one. So, for $i \geq j$, it holds
$$I(Y; X) \geq I(Y; \mathbf{h}_j) \geq I(Y; \mathbf{h}_i) \geq I(Y; \hat{Y}).$$
This means we are progressively losing information.
- Each layer should aim to **provide the most concise and relevant input representation** by maximizing $I(Y; \mathbf{h}_i)$ while minimizing $I(\mathbf{h}_{i-1}; \mathbf{h}_i)$ as much as possible. Reducing $I(\mathbf{h}_{i+1}, \mathbf{h}_i)$ can be interpreted as obtaining the minimal description length of the layer, while $I(Y; \hat{Y})$ serves as a natural quantifier for the quality of the DNN.

IB as a measure of optimality

- IB principle can be used **both as a measure of optimality for the output layer and for evaluating the optimality of each hidden layer**. Namely, each layer can be compared to the optimal IB limit for some β:

  $$I(\mathbf{h}_{i+1};\mathbf{h}_i) + \beta I(Y;\mathbf{h}_{i+1}|\mathbf{h}_i),$$

  where $\mathbf{h}_0 = X$ and $\mathbf{h}_{m+1} = \hat{Y}$.
- This framework is more advantageous than other cost functions, which are generally not applicable for evaluating the optimality of the hidden layers. The input level clearly has the least IB distortion and requires the longest description. Then, each subsequent layer can only increase the IB distortion level, but at the same time, it will compress its inputs, hopefully eliminating only irrelevant information.

Generalization Bounds

- In ML, we are interested in finding a joint distribution $p(X, Y)$, which we know only for a limited number of samples, and **we try to generalize it**. Representations that encode more information than necessary for prediction may end up fitting noise or irrelevant structure in the input data, causing overfitting.
- It has been shown that **IB principle can serve as a learning objective**, since it introduces a tradeoff between the amount of information the representation $\hat{X}$ retains about the input – $I(X; \hat{X})$ – and the amount of relevant information it retains for predicting the target – $I(\hat{X}; Y)$.

Generalization Bounds

- The objective becomes finding a stochastic encoder $p(\hat{x}|x)$ that minimizes the IB Lagrangian:

$$\mathcal{L}_{IB} = I(X;\hat{X}) - \beta I(\hat{X};Y) \tag{10}$$

- The parameter $\beta > 0$ balances compression and prediction performance. This interpretation leads to the insight that compression acts as a **form of regularization** and emerges naturally from the IB objective.
- The complexity of the representation is measured by $I(X;\hat{X})$, and the higher this value is, the higher the risk of overfitting.

Generalization Bounds

The generalization bounds found are as follows:

$$I(\hat{X};Y) \leq \hat{I}(X,Y) + \mathcal{O}\left(\frac{K|Y|}{\sqrt{n}}\right),$$

$$I(X;\hat{X}) \leq \hat{I}(X,\hat{X}) + \mathcal{O}\left(\frac{K}{\sqrt{n}}\right),$$

where $\hat{I}$ the empirical estimate of the mutual information based on the finite sample distribution $\hat{p}(x,y)$, n is the sample size, and $K := |\hat{X}|$.

- These **bounds get worse with** K. The mutual information is well estimated for learning compressed representations (small K), and is badly estimated for learning complex models (large K).
- **Compression helps generalization**, and inserting more layers or neurons without appropriate compression won't help if one has limited data.

References

Main References

- Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. 2000.
- Naftali Tishby and Noga Zaslavsky. “Deep learning and the information bottleneck principle”. In: 2015 IEEE Information Theory Workshop (ITW). Jerusalem, Israel: IEEE, Apr. 2015.

Other References

- Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. en. 1st ed. Wiley, Sept. 2005.
- W.H.R. Equitz and T.M. Cover. “Successive refinement of information”. In: IEEE Transactions on Information Theory 37.2 (Mar. 1991).
- Ohad Shamir, Sivan Sabato, and Naftali Tishby. “Learning and generalization with the information bottleneck”. en. In: Theoretical Computer Science 411.29-30 (June 2010).

Full report

Questions?