On the Information Bottleneck

Abstract

The Information Bottleneck (IB) formalizes the notion of an information-theoretic "optimal" representation in terms of the fundamental tradeoff between having a concise representation and one with good predictive power. It was introduced by Naftali Tishby et al. in 1999 and appears to be fundamental to a deep understanding of representations. We draw connections to (1) minimal sufficient statistics, (2) the formulation of variational auto-encoders, and, (3) the topology of and SGD dynamics deep neural networks.

1 Information Theory

Entropy

Let X be a random variable, then the entropy $H(X)$ is

$$H(X) = E[-\log X] = -\int p(x) \log p(x) dx. \tag{1}$$

Let X and Y be random variables, then conditional entropy $H(X|Y)$ is

$$H(X) = \iint p(x, y) \log p(x|y) dx\, dy. \tag{2}$$

Markov Chain

A Markov chain is a collection of random variables $\{X_i\}$ having the property that, given the present, the future is conditionally independent of the past. That is, the Markov process is a "memoryless" (also called "Markov Property") stochastic process.

$$P(X_t = x_t | X_0 = x_0, X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = P(X_t = x_t | X_{t-1} = x_{t-1}). \tag{3}$$

A simple random walk, or Brownian motion, is an example of a Markov chain.

Kullback-Leibler Divergence

Let p and q denote two probability distributions, then the Kullback-Leibler divergence is

$$\begin{aligned} D_{KL}(p \| q) &= \int p(x) \log \frac{p(x)}{q(x)} dx & (4) \\ &= -\int p(x) \log q(x) dx + \int p(x) \log p(x) dx & (5) \\ &= H(P, Q) - H(P). & (6) \end{aligned}$$

Mutual Information

Given any two random variables, X and Y, with joint distribution $p(x,y)$, their Mutual Information $I(X;Y) = I(Y;X) \geq 0$ is defined as:

$$\begin{aligned} I(X;Y) &= D_{KL}[p(x,y)\|p(x)p(y)] \\ &= \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dx\,dy \\ &= \iint p(x,y)\log\frac{p(x|y)}{p(x)}dx\,dy \\ &= H(X) - H(X|Y) \end{aligned} \tag{7}$$

where $D_{KL}[p\|q]$ denotes the Kullback-Leibler divergence of distributions p and q, and $H(X)$ and $H(X|Y)$ are the entropy and conditional entropy of X and Y, respectively. Note, if $X \perp Y$, then $p(x,y) = p(x)p(y)$, and therefore:

$$X \perp Y \;\Leftrightarrow\; \log\frac{p(x,y)}{p(x)p(y)} = \log 1 \;\Leftrightarrow\; I(X;Y) = 0. \tag{8}$$

The concept is intricately linked to that of entropy of a random variable, a fundamental notion that defined "amount of information" held in a random variable:

$$I(X;Y) = H(X) - H(X|Y) = H(X) - H(X|Z) = H(X) + H(Y) - H(X,Y). \tag{9}$$

The mutual information $I(X;Y)$ quantifies the "amount of information", the average number of relevant bits, obtained about one random variable X, through the other random variable Y. It measures the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence. Define X as some input variable and Y as the label. Then, an optimal learning problem can be cast as the construction of an *optimal encoder* of that relevant information via an efficient representation, a minimal sufficient statistic of X with respect to Y, if such can be found. A minimal sufficient statistic can enable the *decoding* of the relevant information with the smallest number of binary questions (on average), i.e. an *optimal code*.

Reparametrization Invariance & DPI

Two properties of the mutual information are fundamental in our context. First, the *Reparametrization Invariance*, that is the invariance to invertible transformations

$$I(X;Y) = I(\psi(X);\phi(Y)) \tag{10}$$

for any invertible functions $\psi(\cdot)$ and $\phi(\cdot)$. Second, the *Data Processing Inequality* (DPI), that is for any 3 random variables which form a Markov chain $X \to Y \to Z$ it holds

$$I(X;Y) \geq I(X;Z). \tag{11}$$

2 Information Bottleneck

Optimal Representation

(refine with `https://arxiv.org/abs/1703.00810`) (change Z to T)

Let random variable X denote an input source, Z a compressed representation, and Y observed output. We assume a Markov chain $X \to Y \to Z$. That is, Z cannot directly depend on Y. Then, the joint distribution $p(X, Y, Z)$ factorizes as

$$p(X, Y, Z) = p(Z|X, Y)p(Y|X)p(X) = p(Z|X)p(Y|X)p(X). \tag{12}$$

where we assume $p(Z|X, Y) = p(Z|X)$. Our goal is to learn an encoding Z that is maximally informative about our target Y. As a measure we use the mutual information $I(Z; Y) \geq 0$ between our encoding Z and output X

$$I(Z; Y) = \iint p(z, y) \log \frac{p(z, y)}{p(z)p(y)} dy\, dz = \iint p(y, z) \log \frac{p(y|z)}{p(y)} \tag{13}$$

where $p(y|z)$ is fully defined by stochastic encoder $p(Z|X)$ and Markov chain as

$$p(y|z) = \int p(x, y|z)dx = \int p(y|x)p(x|z)dx = \int \frac{p(y|x)p(z|x)p(x)}{p(z)} dx. \tag{14}$$

If maximizing (13) was our only objective, then the trivial identity encoding ($Z = X$) would always ensure a maximal informative representation. Instead, we would like to find the maximally informative representation subject to a constraint on it's complexity. Naturally, we constrain the mutual information between our encoding Z and the input data X such that $I(X; Z) \leq I_c$ where I_c denotes the information constraint. This suggests our objective:

$$\min I(Z; Y) \quad \text{s.t.} \quad I(X; Z) \leq I_c. \tag{15}$$

Equivalently, we introduce a Lagrange multiplier β and write the objective as:

$$I(Z; Y) - \beta I(Z; X). \tag{16}$$

Here, our goal is to learn an encoding Z that is maximally expressive about Y while being maximally compressive about X. Then, $\beta \geq 0$ controls the tradeoff between informativeness and compression where large β corresponds to highly compressed representations. (this is inverse to Tishbys formulation, fix) This approach is known as the Information Bottleneck (IB). Intuitively, the first term in (16) encourages Z to be "predictive" of Y; the second term encourages Z to "forget" X. Essentially, it forces Z to act like a minimal sufficient statistic of X for predicting Y.

The IB is appealing, since it defines a "optimal" representation in terms of the fundamental tradeoff between having a concise representation and one with good predictive power. The main drawback is that computing the mutual information is, in general, computationally challenging since (14) is intractable.

Relaxed Minimal Sufficient Statistic

What characterizes the optimal representation of X with respect to Y? The classical notion of minimal sufficient statistics provides good candidates for optimal representations. In our setting, sufficient statistics $S(X)$ are a partitioning on X, that captures all the information that X has on Y. That is, $I(S(X);Y) = I(X;Y)$.

Minimal sufficient statistics, $T(X)$, are the simplest sufficient statistics and induce the coarsest sufficient partition on X. Formally, they are functions of any other sufficient statistic. We can formulate this by a Markov chain:

$$Y \to X \to S(X) \to T(X), \tag{17}$$

which holds for any minimal sufficient statistic $T(X)$ with any other sufficient statistic $S(X)$. Using the DPI in (11), we cast this into an optimization problem:

$$T(X) = \underset{\{S(X):I(S(X;Y))=I(X;Y)\}}{\arg\min} I(S(X);X). \tag{18}$$

Since exact minimal sufficient statistics only exist for distributions of exponential families, Tishby relaxed this optimization problem by first, allowing the map to be stochastic, defined as an encoder $P(T|X)$, and second, by allowing the map to capute *as much as possible* of $I(X;Y)$, not necessarily all of it. This leads to the *Information Bottleneck* tradeoff, which provides a computational framework for finding approximate minimal sufficient statistics, or, the optimal tradeoff between compression of X and prediction of Y. In this sense, efficient representations are approximate minimal sufficient statistics. Define $t \in T$ as a compressed representation of $x \in X$, then the mapping $p(t|x)$ defines the representation of x. This Information Bottleneck tradeoff is formulated by the following optimization problem, carried independently for the distributions $p(t|x), p(t), p(y|t)$, with Markov chain $Y \to X \to T$,

$$\min_{p(t|x),p(t),p(y|t)} \{I(X;T) - \beta I(T;Y)\}. \tag{19}$$

The Lagrange multipler β determines the level of relevant information $I(T;Y)$ captured by the representation T, which is directly related to the error in the label prediction from this representation. The implicit solution to this problem is given by three self-consistent equations:

$$\begin{cases} p(t|x) &= \frac{p(t)}{Z_\beta(x)} \exp\left(-\beta D_{KL}\left[p(y|x)\|p(y|t)\right]\right) \\ p(t) &= \int p(t|x)p(x)dx \\ p(y|t) &= \int p(y|x)p(x|t)dx \end{cases} \tag{20}$$

where $X_\beta(x)$ denotes the normalization function. These equations are satisfied along the *information curve*, which is a monotonic concave line of optimal representations that separates achievable and unachievable regions in the information-plane. For smooth $p(X, Y)$, i.e. when Y is not a completely deterministic function of X, the information curve is strictly concave with unique slope $beta^{-1}$, at every point. In these cases, β determines a single point on the information curve with specified encoder $P_\beta(T|X)$ and decoder $P_\beta(Y|T)$.

Information Bottleneck Bound

3 Deep Neural Networks

DNN As Markov Chains

Information Plane Theorem

Any representation T, defined as a (possibly stochastic) map of input variable X, is characterized by its joint distributions with X and Y, or by its encoder and decoder distributions, $P(T|X)$ and $P(Y|T)$, respectively. Given $P(X, Y)$, T is uniquely mapped to a point in the information plane with coordinates $(I(X;T), I(T;Y))$. Given a Markov chain $Y \to X \to T_1 \to \ldots \to T_k \to \hat{Y}$ with a chain of representations $\{T_i : i = 1, \ldots, k\}$ and predicted output $\hat{Y}$, then $\{T_i\}$ are mapped to K monotonic connected points in the plane. This unique *information path* satisfies the DPI chains:

$$I(X;Y) \geq I(T_1;Y) \geq T(T2, Y) \geq \ldots \geq I(T_k;Y) \geq I(\hat{Y};Y), \tag{21}$$

$$H(X) \geq I(X;T_1) \geq I(X;T_2) \geq \ldots \geq I(X;T_k) \geq I(X;\hat{Y}). \tag{22}$$

(figure here)
(youtube deep-NN here) (generalization bound here)

4 Variational Bottleneck

Relation to β-VAE

(β-VAE here)

References