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 deep neural networks.

1 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:

$$\begin{aligned} I(X,Y) &= D_{KL}[p(x,y) \| p(x)p(y)] \\ &= \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx\, dy \\ &= \iint p(x,y) \log \frac{p(x|y)}{p(x)} dx\, dy \\ &= H(X) - H(X|Y) \end{aligned} \tag{1}$$

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:

$$X \perp Y \;\Leftrightarrow\; \log \frac{p(x,y)}{p(x)p(y)} = \log 1 \;\Leftrightarrow\; I(X,Y) = 0. \tag{2}$$

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:

$$I(X,Y) = H(X) - H(X|Y) = H(X) - H(X|Z) = H(X) + H(Y) - H(X,Y). \tag{3}$$

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

$$I(X,Y) = I(\psi(X), \phi(Y)) \tag{4}$$

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

$$I(X,Y) \geq I(X,Z). \tag{5}$$

2 Information Bottleneck

(refine with `https://arxiv.org/abs/1703.00810`)

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

$$p(X,Y,Z) = p(Z|X,Y)p(Y|X)p(X) = p(Z|X)p(Y|X)p(X). \tag{6}$$

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

$$I(Z,Y) = \iint p(z,y) \log \frac{p(z,y)}{p(z)p(y)} dy\, dz = \iint p(y,z) \log \frac{p(y|z)}{p(y)} \tag{7}$$

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

$$p(y|z) = \int p(x,y|z)dx = \int p(y|x)p(x|z)dx = \int \frac{p(y|x)p(z|x)p(x)}{p(z)} dx. \tag{8}$$

If maximizing (7) 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:

$$\min_{P(Z|X)} I(Z,Y) \quad \text{s.t.} \quad I(X,Z) \leq I_c. \tag{9}$$

($P(Z|X)$ correct?) (doesn't match with `https://en.wikipedia.org/wiki/Information_bottleneck_method`, see comment 3 page 1) Equivalently, we introduce a Lagrange multiplier β and write the objective as:

$$R(\theta) = I(Z, Y) - \beta I(Z, X). \tag{10}$$

(θ?) 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 approach is known as the Information Bottleneck (IB). Intuitively, the first term in (10) 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 "good" 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 (8) is intractable.

3 Minimal Sufficient Statistics

4 Variational Formulation

(β-VAE here)

5 Information Plane

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

References