On the Information Bottleneck

Abstract

The Information Bottleneck (IB) formalizes the notion of a "good" representation in terms of the fundamental tradeoff between having a concise representation and one with good predictive power. It was introduced by Naftali Tishby 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 Information Bottleneck

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

$$p(X, Y, Z) = p(Z|X, Y)p(Y|X)p(X) = p(Z|X)p(Y|X)p(X). \tag{1}$$

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 $I(Z, Y)$ between our encoding and output

$$I(Z, Y) = \iint p(z, y) \log \frac{p(z, y)}{p(z)p(y)} dy\, dz = \iint p(y, z) \log \frac{p(y|z)}{p(y)} \tag{2}$$

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

$$p(y|z) = \int p(x, y|z)dx = \int p(y|x)p(x|z)dx = \int \frac{p(y|x)p(z|x)p(x)}{p(z)} dx. \tag{3}$$

If maximizing (2) 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:

$$\min_{P(Z|X)} I(Z, Y) \quad \text{s.t.} \quad I(X, Z) \leq I_c. \tag{4}$$

($P(Z|X)$ 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:

$$R(\theta) = I(Z, Y) - \beta I(Z, X). \tag{5}$$

(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 (5) 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 (3) is intractable.

2 Minimal Sufficient Statistics

3 Variational Formulation

(β-VAE here)

4 Information Plane

(youtube deep-NN here)

References