Generalized IWAE for Contrastive Representation Learning

Vaden Masrani

Abstract

Mainly a rewrite of Rob's write up for now

1 Background

1.1 Mutual Information Bounds

The mutual information between two random variables $\mathbf{x}$ and $\mathbf{z}$ is defined

$$\mathbf{I}(\mathbf{x};\mathbf{z}) := \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})}\right] = \mathcal{H}(\mathbf{x}) - \mathcal{H}(\mathbf{x}\,|\,\mathbf{z}) = \mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\log p(\mathbf{x}\,|\,\mathbf{z})\right] - \mathbb{E}_{p(\mathbf{x})}\left[\log p(\mathbf{x})\right], \tag{1}$$

where $\mathcal{H}(\mathbf{x})$ and $\mathcal{H}(\mathbf{x}\,|\,\mathbf{z})$ represent the entropy and conditional entropy of $p(\mathbf{x})$ and $p(\mathbf{x}\,|\,\mathbf{z})$, respectively. The mutual information represents the reduction in uncertainty of random variable $\mathbf{x}$ given knowledge of random variable $\mathbf{z}$, with independent RV's having zero mutual information.

One can obtain variational bounds of MI by substituting into (1) variational bounds of $\log p(\mathbf{x})$. For instance, substituting in the well-known evidence lower bound (ELBO) and evidence upper bound (EUBO) bounds

$$\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right] \leq \log p(\mathbf{x}) \leq \mathbb{E}_{p_\theta(\mathbf{z}\,|\,\mathbf{x})}\left[\frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right] \tag{2}$$

one arrives at

$$\mathbb{E}_{p(\mathbf{x},\mathbf{z})}\left[\frac{q_\phi(\mathbf{z}\,|\,\mathbf{x})}{p(\mathbf{z})}\right] \leq \mathbf{I}(\mathbf{x};\mathbf{z}) \leq \mathbb{E}_{p(\mathbf{x})}\,\mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}\,|\,\mathbf{x})}{p(\mathbf{x},\mathbf{z})}\right] - \mathcal{H}(\mathbf{x}\,|\,\mathbf{z}), \tag{3}$$

where the LHS is the well-known Barber-Agakov lower bound (CITE) and the RHS is the corresponding Barber-Agakov upper bound. Note that because we subtract the $\mathcal{H}(\mathbf{x})$ entropy term in the mutual information expression, lower bounds on $\log p(\mathbf{x})$ translate to upper bounds on the mutual information and vice versa.

We can generalize (2) to generic target and proposal distributions which will allow us to construct alternate bounds on the mutual information using multiple samples, as in the case of importance weighted auto encoder (IWAE). In general, for $q_{\text{prop}} = \tilde{q}_{\text{prop}}/\mathcal{Z}_{\text{prop}}$ and $p_{\text{tgt}} = \tilde{p}_{\text{tgt}}/\mathcal{Z}_{\text{tgt}}$, with $\mathcal{Z}_{\text{tgt}}/\mathcal{Z}_{\text{prop}} = \log p(\mathbf{x})$ we have

$$\mathbb{E}_{q_{\text{prop}}}\left[\log \frac{\tilde{p}_{\text{tgt}}}{\tilde{q}_{\text{prop}}}\right] = \mathbb{E}_{q_{\text{prop}}}\left[\log \frac{p_{\text{tgt}}}{q_{\text{prop}}}\right] + \mathbb{E}_{q_{\text{prop}}}\left[\log \frac{\mathcal{Z}_{\text{tgt}}}{\mathcal{Z}_{\text{prop}}}\right] = \log p(\mathbf{x}) - \text{KL}[q_{\text{prop}}||p_{\text{tgt}}] \tag{4}$$

$$\mathbb{E}_{p_{\text{tgt}}}\left[\log \frac{\tilde{p}_{\text{tgt}}}{\tilde{q}_{\text{prop}}}\right] = \mathbb{E}_{p_{\text{tgt}}}\left[\log \frac{p_{\text{tgt}}}{q_{\text{prop}}}\right] + \mathbb{E}_{p_{\text{tgt}}}\left[\log \frac{\mathcal{Z}_{\text{tgt}}}{\mathcal{Z}_{\text{prop}}}\right] = \log p(\mathbf{x}) + \text{KL}[p_{\text{tgt}}||q_{\text{prop}}] \tag{5}$$

Because of the non-negativity of the KL divergence, we have the following bounds on the log evidence

$$\mathbb{E}_{q_{\text{prop}}}\left[\log \frac{\tilde{p}_{\text{tgt}}}{\tilde{q}_{\text{prop}}}\right] \leq \log p(\mathbf{x}) \leq \mathbb{E}_{p_{\text{tgt}}}\left[\log \frac{\tilde{p}_{\text{tgt}}}{\tilde{q}_{\text{prop}}}\right]. \tag{6}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

which can then be used to bound the mutual information in (1) In the next section we'll derive the Generalized IWAE (GIWAE) bound through careful choice of q_{prop} and p_{tgt} which incorporates in an energy function, and which can be used when the full joint density $p(\mathbf{x}, \mathbf{z})$ is unavailable, as in the case of representation learning.

1.2 Generalized IWAE

We can improve upon the Barber-Agakov bounds by considering multiple samples from the proposal distribution, as in the case of IWAE, which uses a K-sample proposal and target distributions defined as need to talk to rob to get intuition for $\tilde{p}_{\text{tgt}}$, and why summation is allowed to disappear in posterior expectation

$$q^{\text{iwae}}_{\text{prop}}(\mathbf{z}^{1:K} \mid \mathbf{x}) = \prod_{k=1}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x}) \tag{7}$$

$$\tilde{p}^{\text{iwae}}_{\text{tgt}}(\mathbf{x}, \mathbf{z}^{1:K}) = \frac{1}{K} \sum_{s=1}^{K} p_\theta(\mathbf{x}, \mathbf{z}^s) \prod_{k \neq s}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x}) \tag{8}$$

$$p^{\text{iwae}}_{\text{tgt}}(\mathbf{z}^{1:K} \mid \mathbf{x}) = \frac{\tilde{p}^{\text{iwae}}_{\text{tgt}}(\mathbf{x}, \mathbf{z}^{1:K})}{p(\mathbf{x})} = \frac{1}{K} \sum_{s=1}^{K} p_\theta(\mathbf{z}^s \mid \mathbf{x}) \prod_{k \neq s}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x}) \tag{9}$$

where the unnormalized IWAE target distribution is defined as a uniform mixture over K components, with the k'th sample coming from $p(\mathbf{x}, \mathbf{z})$ and the remainder from $q(\mathbf{z} \mid \mathbf{x})$, similarly to how sampling is performed in the wake-phi update in the reweighted wake sleep algorithm cite tuananh. The log ratio reduces to the correct IWAE expression

$$\log \frac{\tilde{p}_{\text{tgt}}}{\tilde{q}_{\text{prop}}} = \log \frac{\frac{1}{K} \sum_{s=1}^{K} p_\theta(\mathbf{x}, \mathbf{z}^s) \prod_{k \neq s}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x})}{\prod_{k=1}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x})} = \log \frac{1}{K} \sum_{s=1}^{K} \frac{p_\theta(\mathbf{x}, \mathbf{z}^s)}{q_\phi(\mathbf{z}^s \mid \mathbf{x})}, \tag{10}$$

and the corresponding MI bound can be computed as

$$\mathbf{I}(\mathbf{x}; \mathbf{z}) = \mathbb{E}_{p(\mathbf{x}, \mathbf{z})} \left[\log p(\mathbf{x} \mid \mathbf{z}) \right] - \mathbb{E}_{p(\mathbf{x})} \left[\log p(\mathbf{x}) \right] \tag{11}$$

$$\geq \mathbb{E}_{p(\mathbf{x}, \mathbf{z})} \left[\log p(\mathbf{x} \mid \mathbf{z}) \right] - \mathbb{E}_{p(\mathbf{x})} \left[\mathbb{E}_{p_{\text{tgt}}} \left[\log \frac{\tilde{p}_{\text{tgt}}}{\tilde{q}_{\text{prop}}} \right] \right] \tag{12}$$

$$\geq \mathbb{E}_{p(\mathbf{x}, \mathbf{z}^1)} \mathbb{E}_{\prod_{k=2}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x})} \left[\log p(\mathbf{x} \mid \mathbf{z}^1) \right] - \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{p(\mathbf{z}^1 \mid \mathbf{x}) \prod_{k=2}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x})} \left[\log \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(\mathbf{x}, \mathbf{z}^k)}{q_\phi(\mathbf{z}^k \mid \mathbf{x})} \right] \tag{13}$$

$$= \mathbb{E}_{p(\mathbf{x}, \mathbf{z}^1)} \mathbb{E}_{\prod_{k=2}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x})} \left[\frac{\log p(\mathbf{x} \mid \mathbf{z}^1)}{\frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(\mathbf{x}, \mathbf{z}^k)}{q_\phi(\mathbf{z}^k \mid \mathbf{x})}} \right]. \tag{14}$$

The GIWAE objective extends IWAE through use of a uniform index variable $\mathcal{U}(s) = 1/K$ for $s \in \{1, 2, ..., K\}$ that specifies which sample $\mathbf{z}^s \sim p(\mathbf{z}^s \mid \mathbf{x})$ is drawn from the true posterior and which samples $\mathbf{z}^{-s} \sim q(\mathbf{z}^{-s} \mid \mathbf{x})$ are drawn from the proposal. This leads to a joint distribution over $(\mathbf{x}, \mathbf{z}^{1:K}, s)$ and posterior over s defined as

$$\tilde{p}^{\text{giwae}}_{\text{tgt}}(\mathbf{x}, \mathbf{z}^{1:K}, s) = p(s) p(\mathbf{x}, \mathbf{z}^{1:K} \mid s) = \frac{1}{K} p_\theta(\mathbf{x}, \mathbf{z}^s) \prod_{k \neq s}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x}) \tag{15}$$

$$p^{\text{giwae}}_{\text{tgt}}(\mathbf{z}^{1:K}, s \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z}^{1:K}, s)}{p(\mathbf{x})} = \frac{1}{K} p_\theta(\mathbf{z}^s \mid \mathbf{x}) \prod_{k \neq s}^{K} q_\phi(\mathbf{z}^k \mid \mathbf{x}) \tag{16}$$

$$\tilde{p}^{\text{giwae}}_{\text{tgt}}(s \mid \mathbf{x}, \mathbf{z}^{1:K}) = \frac{\tilde{p}^{\text{giwae}}_{\text{tgt}}(\mathbf{x}, \mathbf{z}^{1:K}, s)}{\sum_{s'=1}^{K} \tilde{p}^{\text{giwae}}_{\text{tgt}}(\mathbf{x}, \mathbf{z}^{1:K}, s')} = \frac{w^s}{\sum_{s'=1}^{K} w^{s'}} \qquad \text{for } w^s = \frac{p_\theta(\mathbf{x}, \mathbf{z}^s)}{q_\phi(\mathbf{z}^s \mid \mathbf{x})}. \tag{17}$$

Inspired by the form of $\tilde{p}^{\text{giwae}}_{\text{tgt}}(s|\,\mathbf{x}, \mathbf{z}^{1:K})$ the GIWAE objective considers drawing s according to variational self-normalized importance sampling (SNIS) weights defined using energy function $T_\phi(\mathbf{x}, \mathbf{z})$

$$q^{\text{giwae}}_{\text{prop}}(\mathbf{z}^{1:K}, s|\,\mathbf{x}) := q(s|\,\mathbf{x}, \mathbf{z}^{1:K}) q(\mathbf{z}^{1:K}\,|\,\mathbf{x}) = \left(\prod_{k=1}^{K} q_\phi(\mathbf{z}^k\,|\,\mathbf{x})\right) q(s|\,\mathbf{x}, \mathbf{z}^{1:K}) \tag{18}$$

where

$$q(s|\,\mathbf{x}, \mathbf{z}^{1:K}) := \frac{\exp\left(T_\psi\left(\mathbf{x}, \mathbf{z}^s\right)\right)}{\sum_{s'=1}^{K} \exp\left(T_\psi\left(\mathbf{x}, \mathbf{z}^{s'}\right)\right)}. \tag{19}$$

Similar calculations to (14) yield the corresponding GIWAE lower bound

$$\mathbf{I}(\mathbf{x}; \mathbf{z}) - \geq \mathbf{I}_{\text{GIWAE}_L} \tag{20}$$

$$= \mathbb{E}_{p(\mathbf{x}, \mathbf{z}^1)}\left[\frac{\log q_\phi(\mathbf{z}^1\,|\,\mathbf{x})}{p(\mathbf{z}^1)}\right] + \mathbb{E}_{p(\mathbf{x}, \mathbf{z}^1)}\,\mathbb{E}_{\prod_{k=2}^{K} q_\phi(\mathbf{z}^k\,|\,\mathbf{x})}\left[\log \frac{\exp\left(T_\psi\left(\mathbf{x}, \mathbf{z}^1\right)\right)}{\sum_{k=1}^{K} \exp\left(T_\psi\left(\mathbf{x}, \mathbf{z}^k\right)\right)}\right] \tag{21}$$

Missing intuition about why optimal energy function, which equals the true log importance weights up to constant, should learn to discriminate between positive and negative samples.

note somewhere that giwae is strictly worse than iwae (cite giwae paper), but doesn't require access to the density $p(\mathbf{x}, \mathbf{z})$, only that we can sample from it. We therefore turn to the representation learning setting where these conditions are met.

1.3 Information Bottleneck method

In the information bottleneck method we are interested in the following objective

$$\max_{q_\psi, p_\theta} I(Z; Y) - \beta I(Z; X) \tag{22}$$

between random variables X, Y and Z, where Z is a stochastic encoding of the input source X, learned to be maximally informative about the target random variable Y, which could be for instance class labels. Expanding the first term and using the markov factorization $p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{z}\,|\,\mathbf{x})p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x})$, the first term can be rewritten

$$I(Z; Y) = \int p(\mathbf{y}, \mathbf{z}) \log \frac{p(\mathbf{y}, \mathbf{z})}{p(\mathbf{z})p(\mathbf{y})} d\,\mathbf{z}\, d\,\mathbf{y} \tag{23}$$

$$= \int p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \log \frac{p(\mathbf{y}\,|\,\mathbf{z})}{p(\mathbf{y})} d\,\mathbf{z}\, d\,\mathbf{y}\, d\,\mathbf{z} \tag{24}$$

$$= \mathbb{E}_{p(\mathbf{x}, \mathbf{y}) p_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{p(\mathbf{y}\,|\,\mathbf{z})}{p(\mathbf{y})}\right], \tag{25}$$

where $p_\phi(\mathbf{z}\,|\,\mathbf{x})$ is a stochastic encoder parameterized by ϕ, and $p(\mathbf{y}\,|\,\mathbf{z})$ is specified in terms of the encoder and markov chain as

$$p(\mathbf{y}\,|\,\mathbf{z}) = \int p(\mathbf{y}, \mathbf{x}\,|\,\mathbf{z}) d\,\mathbf{x} = \int p(\mathbf{y}\,|\,\mathbf{x}) p(\mathbf{x}\,|\,\mathbf{z}) d\,\mathbf{x} = \int \frac{p(\mathbf{y}\,|\,\mathbf{x}) p_\phi(\mathbf{z}\,|\,\mathbf{x}) p(\mathbf{x})}{p(\mathbf{z})} d\,\mathbf{x} \tag{26}$$

References

A Appendix