Fixing a broken ELBO

Alemi, et al. 2018
Motivation

- Progress in VAEs have led to good generative performance.
- Training via ELBO does not necessarily lead to good representation performance.
- This work focuses on developing losses that give good representations.
Mutual information based loss

- “Good” representation is analogous to higher mutual information between the observations and the latents

\[
I_e(X; Z) = \int \int dx \, dz \, p_e(x, z) \log \frac{p_e(x, z)}{p^*(x)p_e(z)}.
\]  (1)

- Computing the mutual information can be intractable, lower and upper bounds can be developed:

\[
H - D \leq I_e(X; Z) \leq R
\]  (2)
Mutual information based loss

Here,

Constant for a given data distribution

\[ H \equiv - \int dx \, p^*(x) \log p^*(x) \]  
(3)

Distortion

\[ D \equiv - \int dx \, p^*(x) \int dz \, e(z|x) \log d(x|z) \]  
(4)

Rate

\[ R \equiv \int dx \, p^*(x) \int dz \, e(z|x) \log \frac{e(z|x)}{m(z)} \]  
(5)

“Decoder”, approximation to \( p(x|z) \)

“Marginal”, approximation to \( p(z) \)
Distortion-Rate phase maps

- Assuming that obtaining tighter bounds to MI is good, it is of interest to get closer to the $D + R = H$ line.
- It can be shown that a “perfect” $m(z)$ and a “perfect” $d(x|z)$ will take us to the $D + R = H$ line.

\[
H \equiv - \int dx \, p^*(x) \log p^*(x) \tag{3}
\]

\[
D \equiv - \int dx \, p^*(x) \int dz \, e(z|x) \log d(x|z) \tag{4}
\]

\[
R \equiv \int dx \, p^*(x) \int dz \, e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}
\]
Distortion-Rate based loss

- Low distortion implies better encoding and decoding of data.
- Therefore, a good loss function must minimize $R$ while also keeping $D$ as low as possible.

\[
H \equiv -\int dx\ p^*(x) \log p^*(x) \tag{3}
\]
\[
D \equiv -\int dx\ p^*(x) \int dz\ e(z|x) \log d(x|z) \tag{4}
\]
\[
R \equiv \int dx\ p^*(x) \int dz\ e(z|x) \log \frac{e(z|x)}{m(z)} \tag{5}
\]
Distortion-Rate based loss

Consider the following optimization problem and its equivalent form:

\[
\begin{align*}
\min_{e,d,m} & \quad D \\
\text{s.t.} & \quad R \leq \epsilon \\
\min_{e,d,m} & \quad D + \beta R \\
\min_{e(z|x), m(z), d(x|z)} & \quad \int dx \, p^*(x) \int dz \, e(z|x) \\
& \quad \left[ -\log d(x|z) + \beta \log \frac{e(z|x)}{m(z)} \right].
\end{align*}
\]

\(\beta = 1\) Corresponds to the standard VAE loss.
Experiments: Toy problem

Figure 2. Toy Model illustrating the difference between fitting a model by maximizing ELBO (b) vs minimizing distortion for a fixed rate (c).

**Top (i):** Three distributions in data space: the true data distribution, $p^*(x)$, the model’s generative distribution, $g(x) = \sum_z m(z)d(x|z)$, and the empirical data reconstruction distribution, $d(x) = \sum_{x'} \sum_z \tilde{p}(x')e(z|x')d(x|z)$. **Middle (ii):** Four distributions in latent space: the learned (or computed) marginal $m(z)$, the empirical induced marginal $e(z) = \sum_x \tilde{p}(x)e(z|x)$, the empirical distribution over $z$ values for data vectors in the set $X_0 = \{x_n : z_n = 0\}$, which we denote by $e(z_0)$ in purple, and the empirical distribution over $z$ values for data vectors in the set $X_1 = \{x_n : z_n = 1\}$, which we denote by $e(z_1)$ in yellow. **Bottom:** Three $K \times K$ distributions: (iii) $e(z|x)$, (iv) $d(x|z)$ and (v) $p(x'|x) = \sum_z e(z|x)d(x'|z)$. 
Experiments: Binary MNIST

(a) Reconstructions from $-v$ with $\beta = 0.1 - 1.1$.

(b) Generations from $-v$ with $\beta = 0.1 - 1.1$
References