Denoising Critertion for Variational Auto-encoding Framework

Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, Yoshua Bengio
University de Montreal

Université de Montréal

Motivations and Contributions

- Recent advance in variational inference is to use the inference network as the approximate posterior distribution.
- Obtaining a class of variational distributions which is flexible enough to accurately model the true posterior distribution is a challenge.
- The denoising criterion - input is corrupted by adding some noise and the model is asked to recover the original input.
- We show that injecting noise both in input and in the stochastic hidden layer can be advantageous.

Background

Variational Inference

Variational inference is an approximate inference method where the goal is to approximate the intractable posterior distribution $p(z|x)$, by a tractable approximate distribution $q_\phi(z)$.

$$\log p(x) = \mathbb{E}_{q_\phi(z)}\left[\log \frac{p(x,z)}{q_\phi(z)}\right] + \mathbb{KL}(q_\phi(z)||p(z|x)).$$

Variational auto-encoder

Variational auto-encoder (VAE) is that the approximate distribution q is conditioned on the observation x, resulting in a form $q_\phi(z|x)$

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \quad (1)$$

$$= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathbb{KL}(q_\phi(z|x)||p(z)).$$

- The *inference network* represents $q_\phi(z|x)$. The variational parameter ϕ is the weights of the neural network.
- The *generative network* represents $p_\theta(x|z)$. The θ is the weights of the neural network.

Training Procedure

A simple way of training VAE with the denoising criterion:

1. sample a corrupted input $\tilde{x}^{(m)} \sim p(\tilde{x}|x)$,
2. sample $z^{(l)} \sim q(z|\tilde{x}^{(m)})$
3. sample reconstructed images from the generative network $p_\theta(x|z^{(l)})$.

The above procedure can be seen as a special case of optimizing the following objective.

$$\mathcal{L}_{dvae} \simeq \frac{1}{MK}\sum_m^M \sum_k^K \log \frac{p_\phi(x, z^{(k|m)})}{q_\phi(z^{(k|m)}|\tilde{x}^{(m)})} \quad (4)$$

where $\tilde{x}^{(m)} \sim p(\tilde{x}|x)$ and $z^{(k|m)} \sim q_\phi(z|\tilde{x}^{(m)})$.

Noise Injection to Inference Network

Example 1. *Let $\mathbf{x} \in \{0,1\}^D$ be a D-dimension observation, and consider a Bernoulli corruption distribution $p_{\boldsymbol{\pi}}(\tilde{\mathbf{x}}|\mathbf{x}) = Ber(\boldsymbol{\pi})$ around the input $\mathbf{x}$. Then,*

$$\mathbb{E}_{p_{pi}(\tilde{x}|x)}\left[q_\phi(z|\tilde{x})\right] = \sum_{i=1}^{K} q_\phi(z|\tilde{x}_i)p_{\boldsymbol{\pi}}(\tilde{x}_i|x) \quad (2)$$

has the form of a finite mixture of Gaussian and the number of mixture component K is 2^D.

Example 2. *Consider a Gaussian corruption model $p(\tilde{x}|x) = N(x|0,\sigma I)$. Let $q_\phi(z|\tilde{x})$ be a Gaussian inference network. Then,*

$$\mathbb{E}_{p(\tilde{x}|x)}\left[q_\phi(z|\tilde{x})\right] = \int_{\tilde{x}} q_\phi(z|\tilde{x})p(\tilde{x}|x)d\tilde{x}. \quad (3)$$

1. *If $q_\phi(z|\phi^T\tilde{x}) = \mathcal{N}(z|\mu = \phi^T\tilde{x}, \sigma = \sigma^2 I)$ such that the mean parameter is a linear model of weight vector ϕ and input $\tilde{x}$, then the Equation 3 is a Gaussian distribution.*
2. *If $q_\phi(z|\tilde{x}) = \mathcal{N}(z|\mu(\tilde{x}),\sigma(\tilde{x}))$ where $\mu(\tilde{x})$ and $\sigma(\tilde{x})$ are non-linear functions of $\tilde{x}$, then the Equation 3 is an infinite mixture of Gaussian.*

Denoising Variational Lower Bound

Lemma 1. *Consider an approximate posterior distribution of the following form:*

$$q_\Phi(z|x) = \int_{z'} q_\varphi(z|z')q_\psi(z'|x)dz',$$

here, we use $\Phi = \{\varphi, \psi\}$. Then, given $p_\theta(x,z) = p_\theta(x|z)p(z)$, we obtain the following inequality:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\Phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\varphi(z|z')}\right] \geq \mathbb{E}_{q_\Phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(x,z)}{q_\Phi(z|x)}\right].$$

The *denoising variational lower bound*

For the approximate distribution $\tilde{q}_\phi(z|x) = \int q_\phi(z|\tilde{x})p(\tilde{x}|x)d\tilde{x}$, we can write the standard variational lower bound as follows:

$$\log p_\theta(x) \geq E_{\tilde{q}_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{\tilde{q}_\phi(z|x)}\right] \stackrel{\text{def}}{=} \mathcal{L}_{cvae}. \quad (5)$$

$$\mathcal{L}_{dvae} \stackrel{\text{def}}{=} E_{\tilde{q}_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|\tilde{x})}\right]. \quad (6)$$

Applying Lemma 1 to Equation 5, we get:

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{dvae} \geq \mathcal{L}_{cvae}. \quad (7)$$

Note that the above does not necessarily mean that $\mathcal{L}_{dvae} \geq \mathcal{L}_{vae}$ where $\mathcal{L}_{vae}$ is the lower bound of VAE with Gaussian distribution in the inference network.

Proposition 1. *Maximizing $\mathcal{L}_{dvae}$ is equivalent to minimizing the following objective*

$$\mathbb{E}_{p(\tilde{x}|x)}[\mathbb{KL}(\tilde{q}_\phi(z|\tilde{x})||p(z|x))]. \quad (8)$$

Equivalently, $\log p_\theta(x) = \mathcal{L}_{dvae} + \mathbb{E}_{p(\tilde{x}|x)}[\mathbb{KL}(\tilde{q}_\phi(z|\tilde{x})||p(z|x))]$.

Results

Classification Performance

Negative variational lower bounds using different corruption levels on MNIST (the lower, the better). The salt-and-pepper noises are injected to data x during the training.

Model	# Hidden Layers	Noise Level 0	5	10	15
DVAE (K=1)	1	96.14 ± 0.09	**95.52 ± 0.12***	**96.12 ± 0.06**	96.83 ± 0.17
DVAE (K=1)	2	95.90 ± 0.23	**95.34 ± 0.17***	**95.65 ± 0.14**	96.17 ± 0.17
DVAE (K=5)	1	95.20 ± 0.07	**95.01 ± 0.04***	95.55 ± 0.07	96.41 ± 0.11
DVAE (K=5)	2	95.01 ± 0.07	**94.71 ± 0.13***	**94.90 ± 0.22**	96.41 ± 0.11
DIWAE (K=5)	1	94.36 ± 0.07	**93.67 ± 0.10***	**93.97 ± 0.07**	**94.35 ± 0.08**
DIWAE (K=5)	2	94.31 ± 0.07	**93.08 ± 0.08***	**93.35 ± 0.13**	**93.71 ± 0.07**

Negative variational lower bound using different corruption levels on the Frey Face dataset. Gaussian noises are injected to data x during the training.

Model	# Hid. Layers	Noise Level 0	2.5	5	7.5
DVAE (K=1)	1	1304.79 ± 5.71	**1313.74 ± 3.64***	**1314.48 ± 5.85**	1293.07 ± 5.03
DVAE (K=1)	2	1317.53 ± 3.93	**1322.40 ± 3.11***	**1319.60 ± 3.30**	1306.07 ± 3.35
DVAE (K=5)	1	1306.45 ± 6.13	**1320.39 ± 4.17***	**1313.14 ± 5.80**	1298.40 ± 4.74
DVAE (K=5)	2	1317.51 ± 3.81	**1324.13 ± 2.62***	**1320.99 ± 3.49**	**1317.56 ± 3.94**
DIWAE (K=5)	1	1318.04 ± 2.83	**1320.18 ± 3.43**	**1333.44 ± 2.74***	1305.38 ± 2.97
DIWAE (K=5)	2	1320.03 ± 1.67	**1334.77 ± 2.69***	**1323.97 ± 4.15**	1309.30 ± 2.95

Negative variational lower bounds using different corruption levels on MNIST (the lower, the better) with recurrent neural network as a inference network. The salt-and-pepper noises are injected to data x during the training.

Model	# Hidden Layers	Noise Level 0	5	10	15
DVAE (GRU)	1	96.07 ± 0.17	**94.30 ± 0.09***	**94.32 ± 0.12**	**94.88 ± 0.11**
DIWAE (GRU)	1	93.94 ± 0.06	**93.13 ± 0.11**	**92.84 ± 0.07***	**93.03 ± 0.04**

- All of the methods with denoising criterion surpassed the performance of vanilla VAE and vanilla IWAE as shown in Table 1 and Table 2.
- DVAE and DIWAE, both of the models are not very sensitive with respect to the two types of noises: Gaussian and salt and pepper. They are more sensitive to the *level* of the noise rather than the *type*.
- We notice that when VAE combined with GRU tend to severely overfit on the training data and it actually performed worse than having a neural network at the inference network. However, denoising criterion redeems the overfitting behaviour and produce much better results
- We have used a simple corruption distribution using a global corruption rate (the parameter of the Bernoulli distribution or the variance of the Gaussian distribution) to all pixels in the images. To see if a more sensible corruption can lead to an improvement, one may propose a more sensible noise distribution that depends on data in the future.