Variational Autoencoder with Implicit Optimal Priors

Hiroshi Takahashi[1], Tomoharu Iwata[2], Yuki Yamanaka[3], Masanori Yamada[3], Satoshi Yagi[1]
[1]NTT Software Innovation Center, [2]NTT Communication Science Laboratories, [3]NTT Secure Platform Laboratories

1. Variational Autoencoder (VAE)

- The VAE[1] estimates the probability of a data point $\mathbf{x}$ by using latent variable $\mathbf{z}$:

$$p_\theta(\mathbf{x}) = \int \underbrace{p_\theta(\mathbf{x} \mid \mathbf{z})}_{\text{decoder}} \underbrace{p_\lambda(\mathbf{z})}_{\text{prior}} \mathrm{d}\mathbf{z}$$

- The VAE is trained to maximize the expectation of evidence lower bound (ELBO):

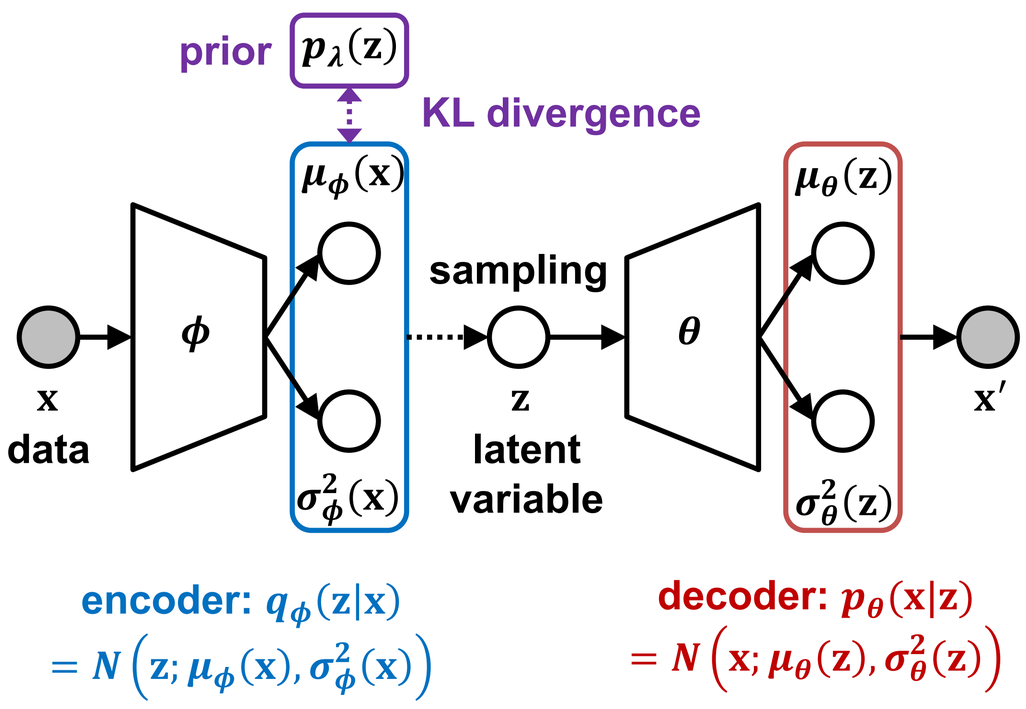
$$\max_{\theta,\phi} \int \underbrace{p_{\mathcal{D}}(\mathbf{x})}_{\text{data distribution}} \underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} \mathrm{d}\mathbf{x}$$

- ELBO can be written as the sum of reconstruction error and Kullback-Leibler (KL) divergence:

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p_\theta(\mathbf{x} \mid \mathbf{z})\right]}_{\text{negative reconstruction error}} - \underbrace{D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\lambda(\mathbf{z})\right)}_{\text{KL divergence}}$$

$q_\phi(\mathbf{z}|\mathbf{x})$: **encoder**

e.g. VAE with Gaussian encoder and decoder



encoder: $q_\phi(\mathbf{z}|\mathbf{x}) = N\left(\mathbf{z}; \mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})\right)$

decoder: $p_\theta(\mathbf{x}|\mathbf{z}) = N\left(\mathbf{x}; \mu_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{z})\right)$

2. Problem: Over-Regularization by the Prior

- The encoder is regularized by the prior using KL divergence. Although the standard Gaussian $p(\mathbf{z}) = N(\mathbf{z}; \mathbf{0}, \mathbf{I})$ is usually used for the prior, this simple prior incurs over-regularization, which is one of the causes of the poor performance of VAE.

- As a sophisticated prior, the aggregated posterior[2] has been introduced, which is the optimal prior in terms of maximizing the expectation of ELBO:

$$\arg\max_{p_\lambda(\mathbf{z})} \int p_{\mathcal{D}}(\mathbf{x})\mathcal{L}(\mathbf{x};\theta,\phi)\mathrm{d}\mathbf{x}$$

$$= \underbrace{\int p_{\mathcal{D}}(\mathbf{x}) q_\phi(\mathbf{z} \mid \mathbf{x})\mathrm{d}\mathbf{x} \equiv q_\phi(\mathbf{z})}_{\text{aggregated posterior}}$$

- However, KL divergence with the aggregated posterior $D_{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel q_{\boldsymbol{\phi}}(\mathbf{z})\right)$ cannot be calculated in a closed form, which prevents us from using this optimal prior.

- In previous work, the aggregated posterior is modeled by using the finite mixture of encoders[3]. Nevertheless, it has sensitive hyperparameters such as the number of mixture components, which are difficult to tune.

3. Our Approach: Estimating the KL Divergence

- We propose the approximation method of this KL divergence without modeling the aggregated posterior explicitly.

- This KL divergence is the expectation of the logarithm of the density ratio $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})/q_{\boldsymbol{\phi}}(\mathbf{z})$. We try to estimate this density ratio directly by the density ratio trick[4].

$$D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| q_\phi(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\underbrace{\ln \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z})}}_{\text{density ratio}}\right]$$

- Since this density ratio depends on both $\mathbf{x}$ and $\mathbf{z}$, this becomes high-dimensional with high-dimensional $\mathbf{x}$. Unfortunately, the density ratio trick works poorly in high dimensions.

- To avoid this, we rewrite the KL divergence as follows:

$$D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| q_\phi(\mathbf{z})) \qquad p(\mathbf{z})\text{: } N(\mathbf{z}; \mathbf{0}, \mathbf{I})$$

$$= \underbrace{D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z}))}_{\text{This can be calculated in a closed form.}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln \frac{q_\phi(\mathbf{z})}{p(\mathbf{z})}\right]}_{\text{low-dimensional density ratio}}$$

- We estimate this density ratio with neural net $T_{\boldsymbol{\psi}}(\mathbf{z})$ as follows:

$$T^*(\mathbf{z}) = \ln \frac{q_\phi(\mathbf{z})}{p(\mathbf{z})}$$

$$T^*(\mathbf{z}) = \max_\psi \mathbb{E}_{q_\phi(\mathbf{z})}\left[\ln(\sigma(T_\psi(\mathbf{z})))\right] + \mathbb{E}_{p(\mathbf{z})}\left[\ln(1 - \sigma(T_\psi(\mathbf{z})))\right]$$

- Therefore, we can estimate the KL divergence by

$$D_{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right) - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[T^*(\mathbf{z})\right]$$

- We alternately optimize $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$ and $T_{\boldsymbol{\psi}}(\mathbf{z})$ like GANs.

4. Experiments

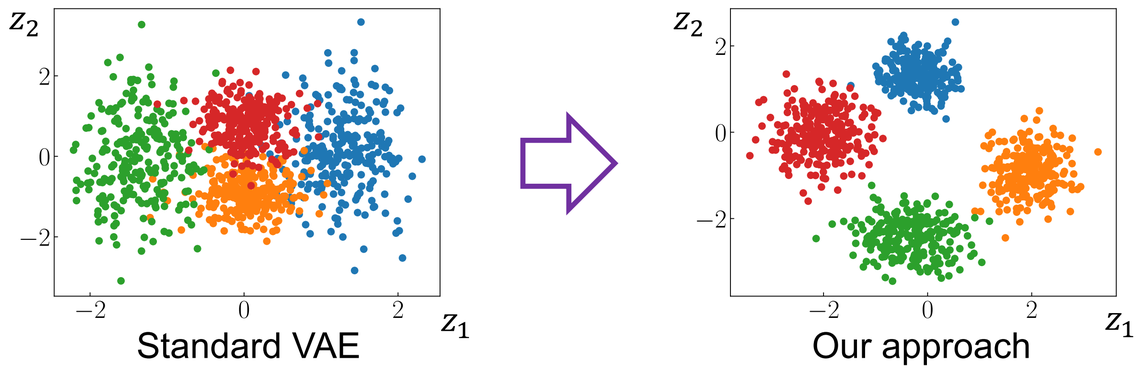
1. Comparison of test log-likelihoods on image datasets

	Standard VAE	VampPrior[3]	Our approach
MNIST	-85.84±0.07	-83.90±0.08	**≈-83.21±0.13**
OMNIGLOT	-111.39±0.11	-110.53±0.09	**≈-108.48±0.16**
FreyFaces	1382.53±3.57	**1392.62±6.25**	**≈1396.27±2.75**
Histopathology	1081.53±0.70	1083.11±2.10	**≈1087.42±0.60**

- Our approach achieved high density estimation performance.

2. Why can our approach achieve good performance?

- To explain this, we did experiment with 4-dimensional One Hot Vector dataset, and plotted the latent vectors $\mathbf{z} \in \mathbb{R}^2$.



*Samples in each color correspond to each latent representation of one hot vectors.

- Our approach makes $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ different from each other and the data point $\mathbf{x}$ is easy to reconstruct from the latent vector $\mathbf{z}$, which improves the density estimation performance.

Reference

[1] Kingma, D. P., and Welling, M. 2013. "Auto-Encoding Variational Bayes"
[2] Hoffman, M. D., and Johnson, M. J. 2016. "ELBO surgery: yet another way to carve up the variational evidence lower bound"
[3] Tomczak, J. M., and Welling, M. 2018. "VAE with a VampPrior"
[4] Sugiyama, M., Suzuki T. and Kanamori T. 2012. " Density ratio estimation in machine learning"