Variational Auto-Encoder

Explainations & Experiments

Tianwei Yue

March 1, 2018

Latent Variable Model

Latent Variable Model

Problem scenario

Assumption: dataset $X = \{x_i\}_{i=1}^{N},\ i.i.d.$ are generated by some random process consist of two steps:

1. $z \sim p_{\theta^*}(z)$
2. $x \sim p_{\theta^*}(x \mid z)$

Latent Variable Model

Problem scenario

Assumption: dataset $X = \{x_i\}_{i=1}^N,\ i.i.d.$ are generated by some random process consist of two steps:

1. $z \sim p_{\theta^*}(z)$
2. $x \sim p_{\theta^*}(x \mid z)$

$$\max_\theta p_\theta(x) = \int_z p(z) p_\theta(x|z)$$

This involves a possibly intractable integral over z.

Variational Inference

Assumption: dataset $X = \{x_i\}_{i=1}^N,\ i.i.d.$ are generated by some random process consist of two steps:

1. $z \sim p_{\theta^*}(z)$
2. $x \sim p_{\theta^*}(x \mid z)$

Variational Auto-Encoder

Encoder: recognition/inference model $q_\phi(z \mid x)$

Decoder: generative model $p_\theta(x \mid z)$

Variational Inference v.s. MCMC

Variational Inference v.s. MCMC

Variational Inference (VI) and Markov chain Monte Carlo (MCMC) sampling are both widely used to approximate posterior densities for Bayesian models.

They resolve the same problem from different perspectives:

- MCMC: **sampling**
- Variational Inference: **optimization**

Variational Inference v.s. MCMC

MCMC: sampling

- Construct an ergodic Markov chain on z whose stationary distribution is the posterior $p(Z \mid X)$
- Sample from the chain
- Approximate the posterior with empirical samples.

Variational Inference: optimization

- Posit a family of densities $\mathcal{Q}$ over Z
- Find $q \in \mathcal{Q}$ such that

$$q_{\phi^*}(Z) = \arg\min_{\phi} \mathcal{KL}(q(Z) \parallel p(Z \mid X))$$

Variational Inference

Goal: to approximate a conditional density of latent variables Z given observed variables X.
Idea: to solve this problem with optimization.

- Posit a family of densities $\mathcal{Q}$ over z
- Find $q \in \mathcal{Q}$ such that

$$q_{\phi^*}(Z) = \arg\min_{\phi} \mathcal{KL}(q(Z) \parallel p(Z \mid X))$$

The members of $\mathcal{Q}$ are parameterized by free "variational" parameters, denote as $q_\phi(z)$.

Variational Inference v.s. MCMC

- VI is faster than MCMC
- VI is easier than MCMC to scale to large data
- VI (2000s) has been studied less rigorously than MCMC (1970s)
- VI's statistical properties are less well understood than MCMC

Variational Inference: Preliminaries

Variational Inference: Preliminaries

- Bayes's Theorem
- Jensen's Inequality
- Kullback-Leibler (KL) divergence

Preliminary: Bayes' Theorem

- Posterior

$$\boxed{p(Z \mid X)} = \frac{\boxed{p(X \mid Z)}\,\boxed{p(Z)}}{\boxed{p(X)}}$$

Preliminary: Bayes' Theorem

- Posterior
- Prior

$$p(Z \mid X) = \frac{p(X \mid Z)\, p(Z)}{p(X)}$$

Preliminary: Bayes' Theorem

- Posterior
- Prior
- Likelihood

$$p(Z \mid X) = \frac{p(X \mid Z)\, p(Z)}{p(X)}$$

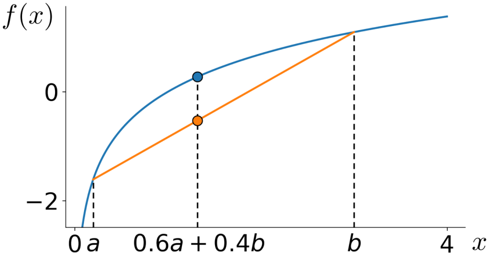
Preliminary: Bayes' Theorem

- Posterior
- Prior

$$p(Z \mid X) = \frac{p(X \mid Z)\, p(Z)}{p(X)}$$

- Likelihood
- Evidence

Preliminary: Jensen's Inequality



Concave Function

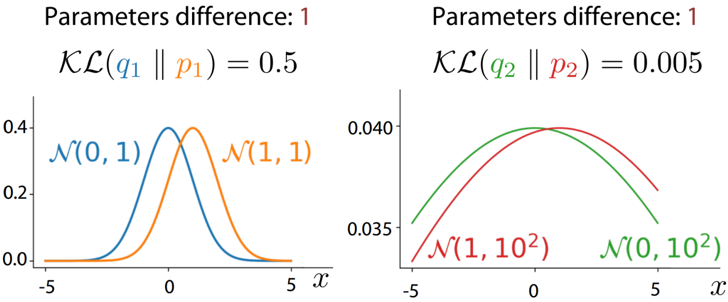
Ref: Polykovskiy and Novikov

Jensen's Inequality: $f\left(\mathbb{E}_{p(t)}(t)\right) \geq \mathbb{E}_{p(t)} f(t)$

Preliminary: Kullback-Leibler (KL) divergence

$$\mathcal{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

- A way to compare distributions
- Not a proper distance



Ref: Polykovskiy and Novikov

Variational Lower Bound

a.k.a. Evidence Lower Bound (ELBO)

Variational Lower Bound

$$\log p_\theta(X) = \log \int_Z p_\theta(X, Z)$$

Variational Lower Bound

$$
\begin{aligned}
\log p_\theta(X) &= \log \int_Z p_\theta(X, Z) \\
&= \log \int_Z \frac{q_\phi(Z)}{q_\phi(Z)} p_\theta(X, Z) = \log \int_Z q_\phi(Z) \frac{p_\theta(X, Z)}{q_\phi(Z)}
\end{aligned}
$$

Variational Lower Bound

$$
\begin{aligned}
\log p_\theta(X) &= \log \int_Z p_\theta(X, Z) \\
&= \log \int_Z \frac{q_\phi(Z)}{q_\phi(Z)} p_\theta(X, Z) = \log \int_Z q_\phi(Z) \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \log \mathbb{E}_{q_\phi(Z)} \frac{p_\theta(X, Z)}{q_\phi(Z)}
\end{aligned}
$$

Variational Lower Bound

$$\begin{aligned}\log p_\theta(X) &= \log \int_Z p_\theta(X,Z)\\ &= \log \int_Z \frac{q_\phi(Z)}{q_\phi(Z)} p_\theta(X,Z) = \log \int_Z q_\phi(Z) \frac{p_\theta(X,Z)}{q_\phi(Z)}\\ &= \log \mathbb{E}_{q_\phi(Z)} \frac{p_\theta(X,Z)}{q_\phi(Z)} \quad \textit{then by Jensen's Inequality}\end{aligned}$$

Variational Lower Bound

$$
\begin{aligned}
\log p_\theta(X) &= \log \int_Z p_\theta(X, Z) \\
&= \log \int_Z \frac{q_\phi(Z)}{q_\phi(Z)} p_\theta(X, Z) = \log \int_Z q_\phi(Z) \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \log \mathbb{E}_{q_\phi(Z)} \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&\geq \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)}
\end{aligned}
$$

Variational Lower Bound

$$
\begin{aligned}
\log p_\theta(X) &= \log \int_Z p_\theta(X, Z) \\
&= \log \int_Z \frac{q_\phi(Z)}{q_\phi(Z)} p_\theta(X, Z) = \log \int_Z q_\phi(Z) \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \log \mathbb{E}_{q_\phi(Z)} \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&\geq \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&\triangleq \mathcal{L}(\theta, q)
\end{aligned}
$$

Variational Lower Bound

$$\begin{aligned}
\log p_\theta(X) &= \log \int_Z p_\theta(X, Z) \\
&= \log \int_Z \frac{q_\phi(Z)}{q_\phi(Z)} p_\theta(X, Z) = \log \int_Z q_\phi(Z) \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \log \mathbb{E}_{q_\phi(Z)} \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&\geq \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&\triangleq \mathcal{L}(\theta, q) \quad \textbf{Variational Lower Bound, } \textit{a.k.a.} \textbf{ ELBO}
\end{aligned}$$

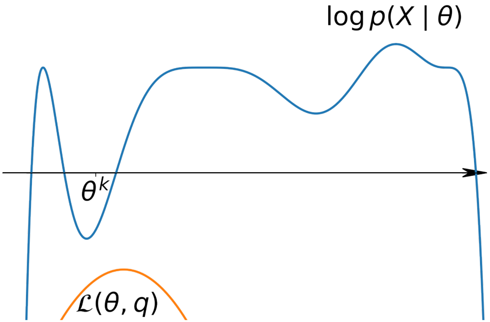
Variational Lower Bound

Then $\mathcal{L}(\theta, q)$ is referred as variational lower bound, *a.k.a.* evidence lower bound (ELBO), and it always holds true that

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q), \text{ for any } q.$$

Understanding Variational Lower Bound

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

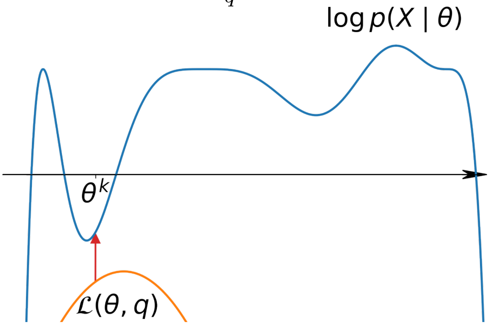


Ref: Polykovskiy and Novikov

Understanding Variational Lower Bound

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

$$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$$

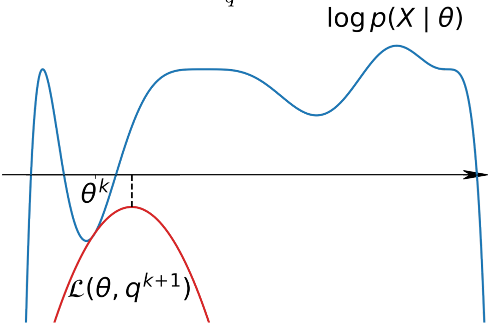


Ref: Polykovskiy and Novikov

Understanding Variational Lower Bound

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

$$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$$

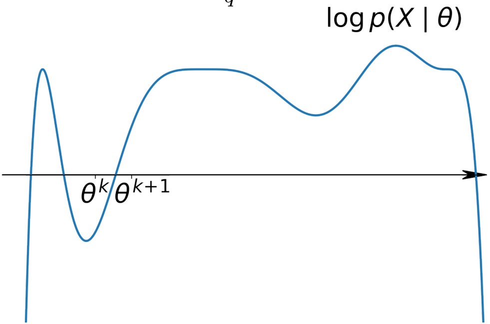


Ref: Polykovskiy and Novikov

Understanding Variational Lower Bound

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

$$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$$

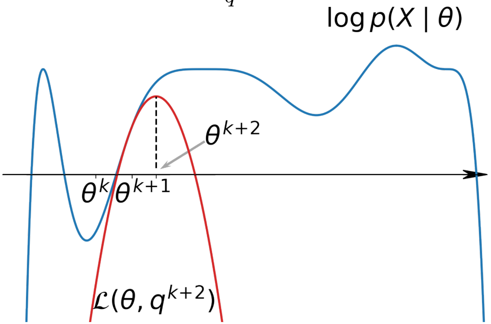


Ref: Polykovskiy and Novikov

Understanding Variational Lower Bound

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

$$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$$



Ref: Polykovskiy and Novikov

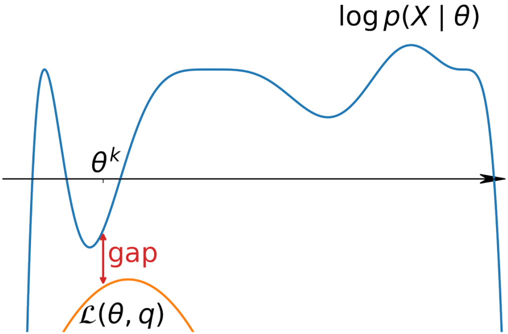
Understanding Variational Lower Bound

To calculate the GAP between $\log p_\theta(X)$ and $\mathcal{L}(\theta, q)$,

$$
\begin{aligned}
&\log p_\theta(X) - \mathcal{L}(\theta, q) \\
=&\log p_\theta(X) - \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
=&\mathbb{E}_{q_\phi(Z)} \left[\log p_\theta(X) - \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \right] \\
=&\mathbb{E}_{q_\phi(Z)} \left[\log \frac{p_\theta(X)}{p_\theta(X, Z)} \cdot q_\phi(Z) \right] \\
=&\mathbb{E}_{q_\phi(Z)} \left[\log \frac{q_\phi(Z)}{p_\theta(Z \mid X)} \right] \\
=&\mathcal{KL}(q_\phi(Z) \parallel p_\theta(Z \mid X))
\end{aligned}
$$

Understanding Variational Lower Bound

$$\log p_\theta(X) - \mathcal{L}(\theta, q) = \mathcal{KL}(q_\phi(Z) \parallel p_\theta(Z \mid X))$$



Ref: Polykovskiy and Novikov

Revisiting Variational Lower Bound

a.k.a. Evidence Lower Bound (ELBO)

Revisiting Variational Lower Bound

Note: $q_\phi(Z \mid X)$ *is usually written as* $q_\phi(Z)$ *for simplicity*

$$\mathcal{L}(\theta, q) = \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)}$$

Revisiting Variational Lower Bound

Note: $q_\phi(Z \mid X)$ is usually written as $q_\phi(Z)$ for simplicity

$$\begin{aligned}
\mathcal{L}(\theta, q) &= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X \mid Z) p_\theta(Z)}{q_\phi(Z)}
\end{aligned}$$

Revisiting Variational Lower Bound

Note: $q_\phi(Z \mid X)$ is usually written as $q_\phi(Z)$ for simplicity

$$\begin{aligned}
\mathcal{L}(\theta, q) &= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X \mid Z) p_\theta(Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(Z)}{q_\phi(Z)} + \mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z)
\end{aligned}$$

Revisiting Variational Lower Bound

Note: $q_\phi(Z \mid X)$ is usually written as $q_\phi(Z)$ for simplicity

$$
\begin{aligned}
\mathcal{L}(\theta, q) &= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X \mid Z) p_\theta(Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(Z)}{q_\phi(Z)} + \mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z) \\
&= -\mathcal{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z \mid X)) + \mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z)
\end{aligned}
$$

Revisiting Variational Lower Bound

Note: $q_\phi(Z \mid X)$ *is usually written as* $q_\phi(Z)$ *for simplicity*

$$
\begin{aligned}
\mathcal{L}(\theta, q) &= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X, Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(X \mid Z) p_\theta(Z)}{q_\phi(Z)} \\
&= \mathbb{E}_{q_\phi(Z)} \log \frac{p_\theta(Z)}{q_\phi(Z)} + \mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z) \\
&= -\mathcal{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z \mid X)) + \mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z) \\
&= \text{Regularization} + \text{Reconstruction}
\end{aligned}
$$

Revisiting Variational Lower Bound

Now we have rewritten $\mathcal{L}(\theta, q)$ into two parts as follows:

$$
\begin{aligned}
\mathcal{L}(\theta, q) &= -\mathcal{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z \mid X)) + \mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z) \\
&= \text{Regularization} + \text{Reconstruction}
\end{aligned}
$$

Given the evidence X of N samples $x^{(1)}, \ldots, x^{(N)}$, $\mathcal{L}(\theta, q)$ is calculated by taking the average.

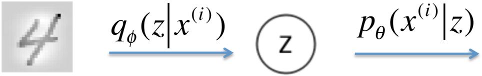
$$
\mathcal{L}(\theta, q) = \sum_{i=1}^{N} -\mathcal{KL}(q_\phi(z|x^{(i)}) \parallel p_\theta(z|x^{(i)})) + \mathbb{E}_{q_\phi(z|x^i)} \log p_\theta(x^{(i)}|z)
$$

Revisiting Variational Lower Bound

$$\mathcal{L}(\theta, q) = \sum_{i=1}^{N} -\mathcal{KL}(q_\phi(z|x^{(i)}) \parallel p_\theta(z|x^{(i)})) + \mathbb{E}_{q_\phi(z|x^i)} \log p_\theta(x^{(i)}|z)$$

Take $x^{(i)}, i \in 1, \ldots, N$ for example:

Example x(i)



Ref: https://tensorchiefs.github.io/bbs/files/vae.pdf

Regularization: $p(z)$ is usually a simple prior $\mathcal{N}(0, 1)$
Reconstruction quality: $\log(1)$ if $x^{(i)}$ gets always reconstructed perfectly (z produces $x^{(i)}$)

Calculation of the Regularization

Common assumptions

- use $\mathcal{N}(0,1)$ as prior for $p(Z)$
- $q(Z \mid X)$ is Gaussian with parameters $(\mu(X), \sigma(X))$ determined by Neural Networks

The Regularization term can be calculated analytically:

$$
\begin{aligned}
&-\mathcal{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z \mid X)) \\
&= \frac{1}{2} \sum_{i=1}^{N} \sum_{s=1}^{D} \left(1 + \log(\sigma(z_s^{(i)^2})) - \mu_s^{(i)^2} - \sigma_s^{(i)^2})\right)
\end{aligned}
$$

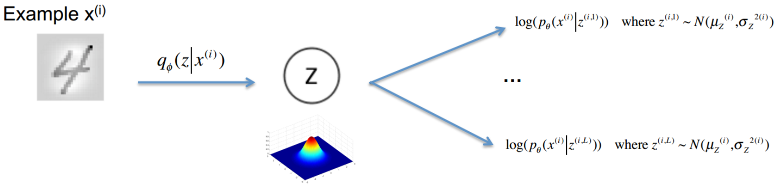
N is the number of samples; D is the number of latent variables.

Sampling to calculate the Reconstruction

Approximating $\mathbb{E}_{q_\phi}$ with sampling from the distribution $q(Z \mid X)$

$$\mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X \mid Z) = \frac{1}{N \cdot J} \sum_{i=1}^{N} \sum_{j=1}^{J} \log p_\theta(x^{(i)} \mid z^{(i,j)})$$

- $-\log p_\theta(x_{(i)} \mid z^{(i,j)})$ can use L2 loss & cross entropy loss, *etc.*
- J is the number of z sampled from $q(z \mid x^{(i)})$



Ref: https://tensorchiefs.github.io/bbs/files/vae.pdf

Revisiting ELBO: Putting it altogether

We have now arrived at

$$\mathcal{L}(\theta, q) = \frac{1}{2}\sum_{i=1}^{N}\sum_{s=1}^{D}\left(1 + \log(\sigma({z_s^{(i)}}^2)) - {\mu_s^{(i)}}^2 - {\sigma_s^{(i)}}^2)\right)$$
$$+ \frac{1}{N \cdot J}\sum_{i=1}^{N}\sum_{j=1}^{J} \log p_\theta(x^{(i)} \mid z^{(i,j)})$$

Recall that

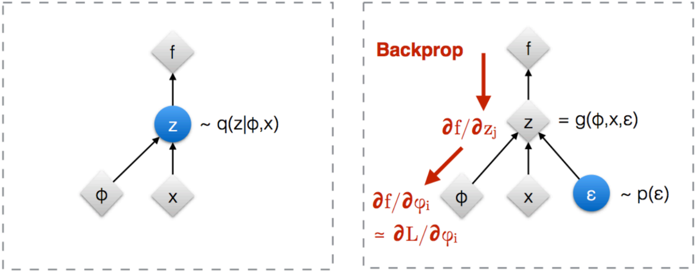
$$\log p_\theta(X) - \mathcal{L}(\theta, q) = \mathcal{KL}(q_\phi(Z) \parallel p_\theta(X \mid Z))$$

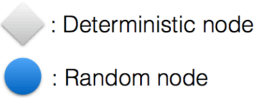
Use mini batch gradient decent to perform optimization.

maximize $\mathcal{L}(\theta, q)$ (ELBO) $\Longleftrightarrow$ minimize $\mathcal{KL}(q_\phi(Z) \parallel p_\theta(X \mid Z))$

Reparameterization Trick

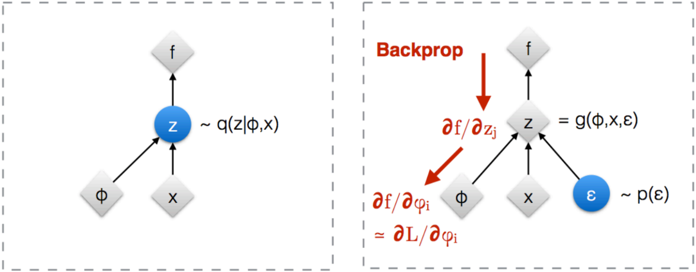
Reparameterization Trick





[Kingma, 2013]
[Bengio, 2013]
[Kingma and Welling 2014]
[Rezende et al 2014]

Reparameterization Trick



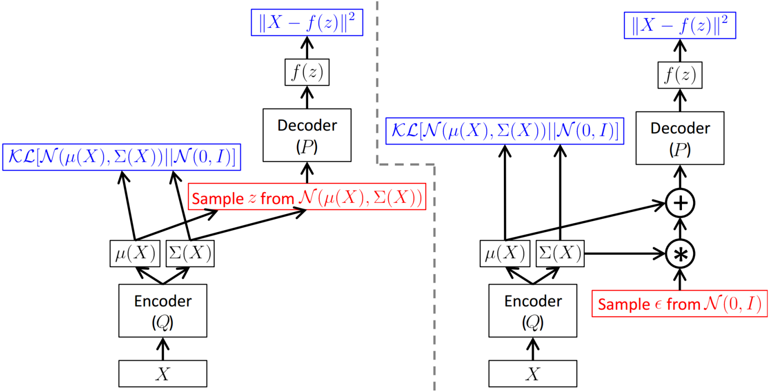
$$\mathcal{L}(\theta, q) = \frac{1}{2}\sum_{i=1}^{N}\sum_{s=1}^{D}\left(1 + \log(\sigma(z_{\phi,s}^{(i)^2})) - \mu_{\phi,s}^{(i)^2} - \sigma_{\phi,s}^{(i)^2}\right)$$

$$+ \frac{1}{N \cdot J}\sum_{i=1}^{N}\sum_{j=1}^{J}\log p_\theta(x^{(i)} \mid z^{(i,j)})$$

where $z^{(l)} = \mu_\phi(x) + \epsilon \odot \sigma_\phi(x)$ and $\epsilon^{(l)} \sim \mathcal{N}(\epsilon; 0, \mathbf{I})$

VAE Structure

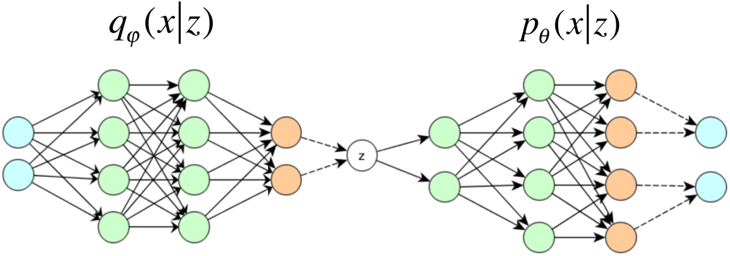
VAE Structure



Ref: Doersch [2016, Tutorial on Variational Autoencoders]

VAE Structure

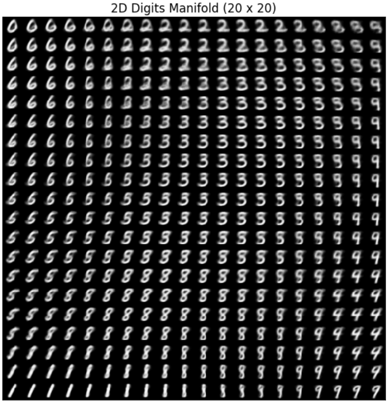
e.g. use a multi-layer perception (MLP) as en/decoder



Ref: https://tensorchiefs.github.io/bbs/files/vae.pdf

VAE Experiments on MINST

2D Digits Manifold

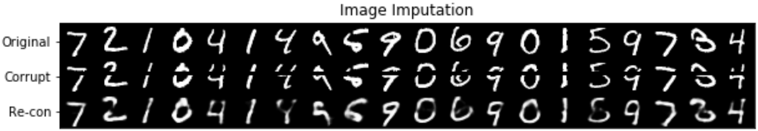


VAE (MLP with 2D latent space)

Linearly spaced coordinates on the unit square were transformed through the inverse CDF of the Gaussian to produce values of the latent variables z, since the prior of the latent space is Gaussian.

code: https://github.com/ThitherShore/VAE-toy/blob/master/vae.py

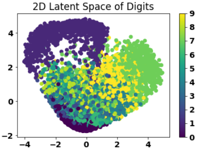
Image Reconstrution and Denoising



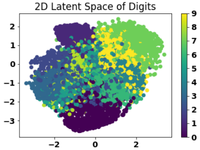
Latent $=$ 20, Epochs $=$ 70

code: https://github.com/ThitherShore/VAE-toy/blob/master/vae.py

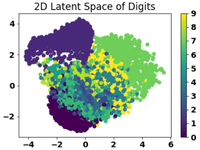
2D Latent Space



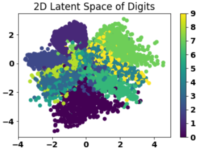
(a) epoch = 3



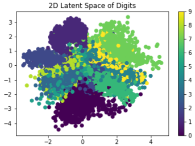
(b) epoch = 5



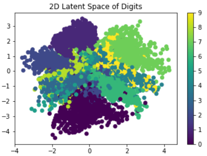
(c) epoch = 10



(d) epoch = 30



(e) epoch = 50

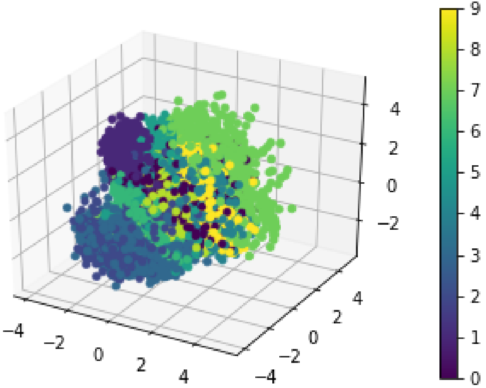


(f) epoch = 60

2D Latent Space over Epochs

code: https://github.com/ThitherShore/VAE-toy/blob/master/vae.py

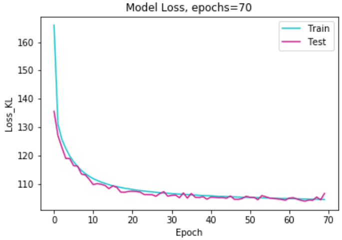
3D Latent Space



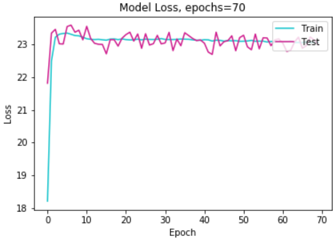
Latent = 3, Epochs = 70

code: https://github.com/ThitherShore/VAE-toy/blob/master/vae.py

Loss over Epochs



(a) Loss = KL loss + cross entropy

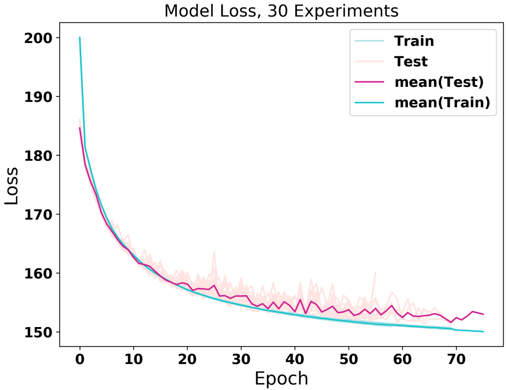


(b) KL Loss

Latent = 20, Epochs = 70

code: https://github.com/ThitherShore/VAE-toy/blob/master/vae.py

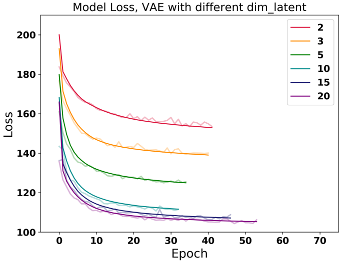
Compare Multiple Experiments



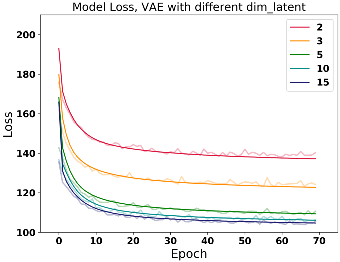
Latent = 2, 30 Experiments

code: https://github.com/ThitherShore/VAE-toy/blob/master/exps_multiple.py

Different Latent Dimensions



(a) EarlyStooping: patience = 5



(b) over 70 epochs

Dim Latent = 2,3,5,10,15,20

code: https://github.com/ThitherShore/VAE-toy/blob/master/exps_diff_latent.py

patience: number of epochs with no improvement after which training will be stopped.

Conditional Variational Auto-Encoder

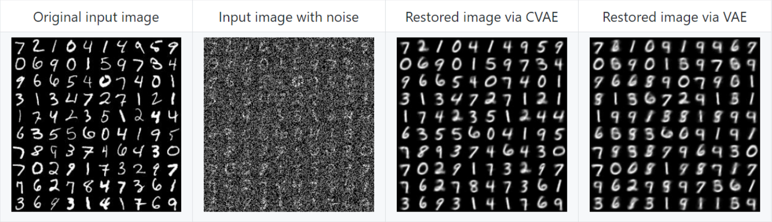
Conditional Variational Auto-Encoder (CVAE)

	Input image	CVAE	VAE
Epoch 1			
Epoch 20			

Ref (paper): Kingma et al., Semi-S L with Deep Generative Models

Ref (codes): https://zhuanlan.zhihu.com/p/25401928

Conditional Variational Auto-Encoder (CVAE)



Ref: https://zhuanlan.zhihu.com/p/25401928

Conditional Variational Auto-Encoder (CVAE)

Learned MNIST manifold with a condition of label 2

Learned MNIST manifold with a condition of label 3

Learned MNIST manifold with a condition of label 4

Ref: https://zhuanlan.zhihu.com/p/25401928

C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. 2014. *arXiv preprint arXiv:1406.5298*.

D. Polykovskiy and A. Novikov. Bayesian methods for machine learning. Coursera.

Thank You!