# Understanding Variational Auto Encoder

Kaiwen Cai

## 1. Understanding Variational Auto Encoder

Variational Auto Encoder (VAE)[1] encodes images into vectors in a latent space, and then decode the latent vectors into images. We denote

- $z$: latent variable, $z \in \mathbb{R}^J$
- $x$: data (images), $x \in \mathbb{R}^{H \cdot W}$
- $p(x)$: evidence probability
- $p(z)$: prior probability
- $p(z|x)$: posterior probability
- $p(x|z)$: likelihood probability

The goal is to find $p(z|x)$ given $p(z)$ and $x$. Once $p(z|x)$ is known, for each sample in $x$, we can represent it with a low dimensional latent vector $z$ by network forward propagration: $z \leftarrow p(z|x) \leftarrow x$. However,

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x|z)p(z)dz} = \frac{p(x|z)p(z)}{\iiint \iiint \ldots p(x|z)p(z)dz}$$

$p(z|x)$ is intractable due to the intractable denominator. We resort to variational inference which approximates $p(z|x)$ with a distribution $q(z|x)$ from a tractable family (e.g., Gaussian distribution). Then the task is translated to: find $q(z|x)$ that is as close as possible to $p(z|x)$. Formally, their distributional distance to be minimized is measured by KL divergence:

$$\begin{aligned}
KL\left(q(z|x) \| p(z|x)\right) &= -\int q(z|x) \log \frac{p(z|x)}{q(z|x)} \\
&= -\int q(z|x) \log \frac{p(x|z)p(z)/p(x)}{q(z|x)} \cdots using\ Bayesian\ theorem \\
&= -\int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} + \int q(z|x) \log p(x) \\
&= -\int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} + \log p(x) \int q(z|x) \cdots note\ that \int q(z|x) = 1 \\
&= \underbrace{-\int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)}}_{EvidenceLowerBound} + \underbrace{\log p(x)}_{Evidence}
\end{aligned}$$

Since KL divergence is positive:

$$KL\left(q(z|x) \| p(z|x)\right) = \underbrace{-\int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)}}_{EvidenceLowerBound} + \underbrace{\log p(x)}_{Evidence} > 0$$

We derived what is called Evidence Lower Bound:

$$\underbrace{\log p(x)}_{Evidence} > \underbrace{\int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)}}_{EvidenceLowerBound}$$

More specifically:

$$\begin{aligned}\underbrace{\int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)}}_{EvidenceLowerBound} &= \int q(z|x) \log \frac{p(z)}{q(z|x)} + \int q(z|x) \log \frac{p(x|z)}{q(z|x)} \\ &= \underbrace{KL\left(q(z|x) \| p(z)\right)}_{Distribution} + \underbrace{\mathbb{E}_{q(z|x)} \log p(x|z)}_{Reconstruction}\end{aligned}$$

Based on the above formulation, we finally arrive at the loss function:

$$\mathcal{L} = -\underbrace{KL\left(q(z|x) \| p(z)\right)}_{Distribution} - \underbrace{\mathbb{E}_{q(z|x)} \log p(x|z)}_{Reconstruction}$$

Minimizing the loss equals minimizing the distribution distance between $q(z|x)$ and $p(z|x)$. Most importantly, each item in $\mathcal{L}$ is tractable:

- $q(z|x)$: the encoder output $\{z_j \sim \mathcal{N}(\mu_j, \sigma_j^2) | j = 1, 2, \cdots, J\}$.
- $p(z)$ is defined as Gaussian priors $\{z_j \sim \mathcal{N}(0, 1) | j = 1, 2, \cdots, J\}$.
- $\log p(x|z) = \log \mathcal{N}(x - \hat{\mu}, \hat{\sigma}^2)$

Calculating the loss[2]:

$$\begin{aligned}\mathcal{L} &= -\underbrace{KL\left(q(z|x) \| p(z)\right)}_{Distribution} - \underbrace{\mathbb{E}_{q(z|x)} \log p(x|z)}_{Reconstruction} \\ &= \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log\left((\sigma_j)^2\right) - (\mu_j)^2 - (\sigma_j)^2\right) - \sum_{m=1}^{H \cdot W} (x_m - \hat{\mu_m})^2\end{aligned}$$

## References


[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Stephen Odaibo. Tutorial: Deriving the standard variational autoencoder (vae) loss function. *arXiv preprint arXiv:1907.08956*, 2019.