Derivations

Kelvin Hong

27th Apr 2023

1 Introduction

This document records some derivations related to VAE.

2 Compute losses in VAE objective

Given a data point x and a latent representation z, we assume they have dimensions D, d respectively. For example, in a VAE of CIFAR10 with latent dimension $d = 50$, since one image is of shape $(3, 32, 32)$, we have $D = 3 \times 32 \times 32 = 3072$.

Let $\tilde{x}$ be a reconstruction of x, then the MSE loss between them is

$$\text{MSE}(x, \tilde{x}) = \frac{1}{D} \sum_{i=1}^{D} (x_i - \tilde{x}_i)^2$$

2.1 KL Divergence & Reconstruction Losses

Let q_ϕ and p_θ be the encoder and decoder network in a VAE, which is common in many literatures. Specifically, q_ϕ takes a data point input x and output its latent variable z, either directly (Auto encoder) or parametrizing the mean and variance of z. We note this relation as a PDF $z \sim q_\phi(z|x)$. Similarly, by taking a latent z as input, p_θ can produce a reconstructed data point $\tilde{x}$. We write this as another PDF $\tilde{x} \sim p_\theta(x|z)$. We will often ignore the subscript and instead write $q(z|x), p(x|z)$ in the following section.

In a standard VAE, given a data point x we let z be a sample from the posterior, characterized by $z \sim N(z; \mu_\phi(x), \sigma_\phi(x)^2)$. Both $\mu_\phi(x)$ and $\sigma_\phi(x)$ are d-dimension vectors, which means $(\sigma_\phi(x)^2)^T I_d$ acts as a diagonal covariance matrix of the data distribution.

The prior $p(z)$ is assumed to be standard multivariate gaussian, $N(z; 0, I_d)$. We then have

$$\frac{q(z|x)}{p(z)} = \frac{1}{\prod_i \sigma_i} \exp\left[-\frac{1}{2}\sum_i \left(\frac{(z_i - \mu_i)^2}{\sigma_i^2} - z_i^2\right)\right]$$

where $\mu_\phi(x) = (\mu_1, \ldots, \mu_d)$ and $\sigma_\phi(x) = (\sigma_1, \ldots, \sigma_d)$.

The KL divergence of $q(z|x)$ and $p(z)$ is thus

$$\begin{aligned} -D_{KL}(q(z|x)||p(z)) &= \int_{\mathbb{R}^d} q(z|x)\left[\sum_i \log \sigma_i + \frac{1}{2}\sum_i \left(\frac{(z_i - \mu_i)^2}{\sigma_i^2} - z_i^2\right)\right] dz \\ &= \sum_{i=1}^{d}\left[\log \sigma_i + \frac{1}{2}(\mu_i^2 + \sigma_i^2 - 1)\right] \end{aligned}$$

Note that the resulting metric is summed instead of averaged over the latent dimension, which is an important feature to be kept in mind when using manually tuned KL coefficient in a VAE, especially beta-VAE. We suspect this is also a reason for different literatures to claim different choices of beta.

Given a latent variable z which comes from x, we let $\tilde{x}$ be the reconstruction from z, following the distribution $p(x|z) \sim N(x; \mu_\theta(z), \sigma^2)$ which the $\sigma = \sigma_\theta$ is shared between all pixels in x. The variance σ^2 might be manually chosen, learned, or analytically calculated. The negative log-likelihood of getting $\tilde{x}$ from the distribution $p(x|z)$ is evaluated as

$$-\ln p(\tilde{x}|z) = D \ln \sigma + \frac{1}{2} D \ln(2\pi) + \frac{D}{2\sigma^2}\text{MSE}(x, \tilde{x})$$

the proof can be found from Simple and Effective VAE Training with Calibrated Decoders

Combining the two losses, we can construct the original ELBO (evidence lower bound) of the vanilla VAE

$$\begin{aligned} -\mathcal{L}_{\phi,\theta}(x) &= \frac{1}{2} D \ln(2\pi) + D \ln(\sigma_\theta) + \frac{D}{2\sigma_\theta^2}\text{MSE}(x, \tilde{x}) + \sum_{i=1}^{d}\left[\log \sigma_i + \frac{1}{2}(\mu_i^2 + \sigma_i^2 - 1)\right] \\ &\equiv D \ln(\sigma_\theta) + \frac{D}{2\sigma_\theta^2}\text{MSE}(x, \tilde{x}) + \sum_{i=1}^{d}\left[\log \sigma_i + \frac{1}{2}(\mu_i^2 + \sigma_i^2 - 1)\right] \end{aligned}$$

Should put minus sign here :(