VAE Review Notes

SUT, CE-40959

Spring 2020

1 Introduction

Variational Autoencoder is a deep generative model, working explicitly with the density function.
Suppose we have a distribution on X named $p_D(x)$. We want to model this distribution with a neural network (θ), to be able to take new samples from it.

What is an ordinary way to train the network and learn θ?
A natural training objective for learning θ is maximum likelihood:

$$\max_{\theta} \mathbb{E}_{p_D(x)}[\log p_\theta(x)] \tag{1}$$

Assuming that every data point is generated from an underlying latent representation z, we can write $p_\theta(x)$ as $p_\theta(x, z)$ marginalized on z. We can rewrite the equation 1 as follows:

$$\mathbb{E}_{p_D(x)}[\log p_\theta(x)] = \mathbb{E}_{p_D(x)}[\log \int_z p_\theta(x|z)p(z)dz] \tag{2}$$

This equation is intractable and can not be optimized in an efficient manner.

2 How to overcome intractability?

Until now we have a network (θ) that maps from z to x. This network provides us with a joint distribution on X and Z (Which is called a generative distribution):

$$p_\theta(x, z) = p_\theta(x|z)p(z) \tag{3}$$

We create another network (ϕ) to map from x to z. This network too, models a joint distribution on X and Z (Which is called an inference distribution):

$$q_\phi(x, z) = q_\phi(z|x)p_D(x) \tag{4}$$

The auxiliary $q_\phi(z|x)$ distribution can help us overcome intractability.

We can now rewrite the intractable $\log \int_z p_\theta(x|z)p(z)dz$ in equation 2 as

$$\log \int_z p_\theta(x|z)p(z)dz = \log \int_z \frac{q_\phi(z|x)}{q_\phi(z|x)} p_\theta(x|z)p(z)dz \tag{5}$$

$$= \log \int_z q_\phi(z|x) \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} dz \tag{6}$$

$$= \log \mathbb{E}_{q_\phi(z|x)}[\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}] \tag{7}$$

$$\geq \mathbb{E}_{q_\phi(z|x)}[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}] \tag{8}$$

$$= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathbb{E}_{q_\phi(z|x)}[\log \frac{p(z)}{q_\phi(z|x)}] \tag{9}$$

$$= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \mid\mid p(z)) \tag{10}$$

in which equation 8 is based on Jensen's inequality.

3 VAE Objective Function

We define the objective function for a specific x as

$$\begin{aligned} \mathcal{L}_{\mathrm{ELBO}}(x) &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \mid\mid p(z)) \\ &\leq \log \int_z p_\theta(x|z)p(z)dz = \log p_\theta(x) \end{aligned} \tag{11}$$

which, as shown in the equation, is a lower bound on log of likelihood. We also define the final objective function of the VAE as:

$$\max_{\phi,\theta} \mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{p_D(x)} \left[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \mid\mid p(z))\right] \tag{12}$$

4 How can we optimize $\mathcal{L}_{\mathrm{ELBO}}$?

We know that we can estimate $\mathbb{E}_{f(u)}[g(u)]$ as $\frac{1}{N}\sum_{n=1}^{N} g(u_n)$ which $u_1, \ldots, u_N$ are N samples taken from the $f(u)$ distribution. Since we can take samples from both $p_D(x)$ and $q_\phi(z|x)$, we can compute both of the expectations of equation 12.

We can compute $\log p_\theta(x|z)$ analytically since the distribution is set to be either Gaussian or Bernoulli. (See theoretical problems)

We can also compute $\mathrm{KL}(q_\phi(z|x) \mid\mid p(z))$ analytically, since both distributions are set to be Gaussian. (See theoretical problems)

So, all of the terms in the equation 12 can be computed efficiently, and we can optimize ϕ and θ to maximize it.

5 Equivalent Forms of the ELBO

$\mathcal{L}_{\text{ELBO}}$ has different forms, which may not be directly optimizable, but are useful in theoretical analyses. You may see these alternative forms in papers extending VAE framework.

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} &\equiv -\operatorname{KL}(q_\phi(x,z) \,||\, p_\theta(x,z)) && (13) \\ &= -\operatorname{KL}(p_D(x) \,||\, p_\theta(x)) - \mathbb{E}_{p_D(x)}[\operatorname{KL}(q_\phi(z|x) \,||\, p_\theta(z|x))] && (14) \\ &= -\operatorname{KL}(q_\phi(z) \,||\, p(z)) - \mathbb{E}_{q_\phi(z)}[\operatorname{KL}(q_\phi(x|z) \,||\, p_\theta(x|z))] && (15) \end{aligned}$$

Proof of the equation 13:

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} &= \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \operatorname{KL}(q_\phi(z|x) \,||\, p(z))\right] \\ &= \mathbb{E}_{p_D(x)}\left[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathbb{E}_{q_\phi(z|x)}[\frac{q_\phi(z|x)}{p(z)}]\right] \\ &= \mathbb{E}_{q_\phi(x,z)}\left[\log p_\theta(x|z) + \log p(z) - \log q_\phi(z|x)\right] \\ &= \mathbb{E}_{q_\phi(x,z)}\left[\log p_\theta(x,z) - \log q_\phi(z|x) - \log p_D(x) + \log p_D(x)\right] && (16) \\ &= \mathbb{E}_{q_\phi(x,z)}\left[\log p_\theta(x,z) - \log q_\phi(x,z) + \log p_D(x)\right] \\ &= \mathbb{E}_{q_\phi(x,z)}\left[\log \frac{p_\theta(x,z)}{q_\phi(x,z)} + \log p_D(x)\right] \\ &= -\operatorname{KL}(q_\phi(x,z) \,||\, p_\theta(x,z)) + \mathbb{E}_{p_D(x)}\left[\log p_D(x)\right] \end{aligned}$$

Since $\mathbb{E}_{p_D(x)}[\log p_D(x)]$ is a constant (entropy of a fixed distribution), optimizing $-\operatorname{KL}(q_\phi(x,z) \,||\, p_\theta(x,z))$ is equivalent to optimizing $\mathcal{L}_{\text{ELBO}}$.

Proofs of the equivalence of optimizing equations 14 and 15 to equation 12 are left for exercise.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[2] S. Zhao, J. Song, and S. Ermon, "Infovae: Balancing learning and inference in variational autoencoders," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5885–5892, 2019.

[3] M. Soleymani, "Deep learning course slides." University Lecture, May 2020. Computer Engineering Department of Sharif University of Technology.