Variational Inference for Deep Learning

November 27, 2021

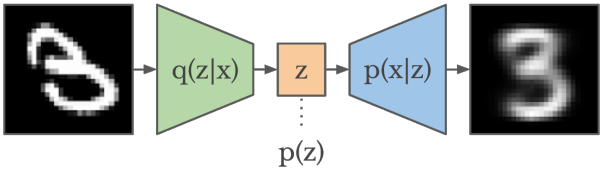

Figure 1: Variational AutoEncoder

1 ELBO for VAE

Let: $p(z) \sim$ prior on latent, $q_\phi(z|x) \sim$ inference model, and $p_\theta(x|z) \sim$ generator model.

1. $\log p_\theta(x) = \int dz\, q_\phi(z|x) \log p_\theta(x)$
2. $= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x)]$
3. $= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x)]$
4. $= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z,x)}{p_\theta(z|x)}\right]$
5. $= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z,x) q_\phi(z|x)}{p_\theta(z|x) q_\phi(z|x)}\right]$
6. $= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z,x)}{q_\phi(z|x)}\right] + \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)}\right]$
7. $= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z,x)}{q_\phi(z|x)}\right] + D_{KL}(q_\phi(z|x)|p_\theta(z|x))$
8. Let $ELBO(x) = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z,x)}{q_\phi(z|x)}\right]$

$$= \log p_\theta(x) - D_{KL}(q_\phi(z|x)|p_\theta(z|x) = \mathcal{L}_{\theta,\phi}(x)$$

$$\leq \log p_\theta(x)$$

2 SGD on ELBO for VAE

SGD is not an unbiased estimator of the gradients of ϕ due to backpropagation through the latent r.v. z.

$$\nabla ELBO \; w.r.t. \; \phi : \nabla_\phi \mathcal{L}_{\theta,\phi}(x) = \nabla_\phi \, \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x,z) - \log q_\phi(z|x) \right]$$
$$\neq \mathbb{E}_{q_\phi(z|x)} \left[\nabla_\phi (\log p_\theta(x,z) - \log q_\phi(z|x)) \right]$$

To get around this issue, we use the reparameterization trick.

1. Let $z = g(\epsilon, \phi, x)$
2. $\mathbb{E}_{q_\phi(z|x)}[f(g(\epsilon,\phi,x))] = \mathbb{E}_{q_\phi(z|x)}[f(z)] = \mathbb{E}_{p(\epsilon)}[f(z)]$
3. Thus, $\nabla_\phi \, \mathbb{E}_{q_\phi(z|x)}[f(z)] = \nabla_\phi \, \mathbb{E}_{p(\epsilon)}[f(z)] = \mathbb{E}_{p(\epsilon)} \, \nabla_\phi [f(z)] \simeq \nabla_\phi f(z)$

Thus, we can form a MC-Estimator $\mathcal{L}^{\sim}_{\theta,\phi}(x)$ from a single noise sample ϵ from $p(\epsilon)$:

$$\mathcal{L}^{\sim}_{\theta,\phi}(x) = \log p_\theta(x,z) - \log q_\phi(z|x), \quad \epsilon \sim p(\epsilon), \quad z = g(\epsilon,\phi,x)$$

Averaged over $\epsilon \sim p(\epsilon)$, the gradient equals the single-datapoint ELBO gradient:

1. $\mathbb{E}_{p(\epsilon)}[\nabla_{\theta,\phi} \mathcal{L}^{\sim}_{\theta,\phi}(x;\epsilon)] = \mathbb{E}_{p(\epsilon)}[\nabla_{\theta,\phi}(\log p_\theta(x,z) - \log q_\phi(z|x))]$
2. $= \nabla_{\theta,\phi}[\mathbb{E}_{p(\epsilon)}[(\log p_\theta(x,z) - \log q_\phi(z|x))]]$
3. $= \nabla_{\theta,\phi} \mathcal{L}^{\sim}_{\theta,\phi}(x)$

Thus, with the MC-Estimator $\mathcal{L}^{\sim}_{\theta,\phi}(x)$, we can maximize ELBO using SGD.

3 ELBO in practice for VAE

As shown in the previous section:

$$\begin{aligned}\mathcal{L}^{\sim}_{\theta,\phi}(x) &= \log p_\theta(x,z) - \log q_\phi(z|x),\ \epsilon \sim p(\epsilon),\ z = g(\epsilon,\phi,x) \\ &= \log p_\theta(x) - D_{KL}(q_\phi(z|x)|p_\theta(z|x))\end{aligned}$$

In practice, $\log p_\theta(x)$, is calculated by taking the cross-entropy of the batch of data against the reconstructed output from the decoder (recall, cross-entropy is equivalent to log-likelihood with i.i.d assumption).

$$\text{Mini-batch}\mathcal{M} \subset \mathcal{D} \text{ of size } \mathcal{N}_\mathcal{M};\ \frac{1}{\mathcal{N}_\mathcal{M}} \log p_\theta(\mathcal{D}) \simeq \frac{1}{\mathcal{N}_\mathcal{M}} \log p_\theta(\mathcal{M}) = \frac{1}{\mathcal{N}_\mathcal{M}} \sum_{x \in \mathcal{M}} \log p_\theta(x)$$

Yet, calculating $\log q_\phi(z|x)$ from $D_{KL}(q_\phi(z|x)|p_\theta(z|x)) = \int dz\ q_\phi(z|x) \log(\frac{q_\phi(z|x)}{p_\theta(z|x)})$ may be "cumbersome" depending on our choice for the prior $p(z)$. A common choice for $p(z)$ is a normal.

4 Closed form ELBO for VAE with Gaussian Latent

1. Say $p(z) \to \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp(\frac{-(x-\mu_p)^2}{2\sigma_p^2})$ and $q_\phi(z|x) \to \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp(\frac{-(x-\mu_q)^2}{2\sigma_q^2})$
2. $D_{KL}(q_\phi(z|x)|p(z)) = \int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp(\frac{-(x-\mu_q)^2}{2\sigma_q^2}) \log(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp(\frac{-(x-\mu_p)^2}{2\sigma_p^2})}{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp(\frac{-(x-\mu_q)^2}{2\sigma_q^2})})\ dz$
3. $= \frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp(\frac{-(x-\mu_q)^2}{2\sigma_q^2}) \big\{ \log(\frac{\sigma_q}{\sigma_p}) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2} \big\} dz$
4. $-D_{KL}(q_\phi(z|x)|p(z)) = \mathbb{E}_{q_\phi(z|x)} \big[\log(\frac{\sigma_q}{\sigma_p}) - \frac{(x-\mu_p)^2)}{2\sigma_p^2} + \frac{(x-\mu_q)^2)}{2\sigma_q^2} \big]$

   $= \log(\frac{\sigma_q}{\sigma_p}) - \frac{1}{2\sigma_p^2} \mathbb{E}_{q_\phi(z|x)}[(x-\mu_p)^2] + \frac{1}{2\sigma_q^2} \mathbb{E}_{q_\phi(z|x)}[(x-\mu_q)^2]$
5. Notice, $\sigma_q^2 = \mathbb{E}_{q_\phi(z|x)}[(x-\mu_q)^2]$

   $\Rightarrow -D_{KL}(q_\phi(z|x)|p(z)) = \log(\frac{\sigma_q}{\sigma_p}) - \frac{1}{2\sigma_p^2} \mathbb{E}_{q_\phi(z|x)}[(x-\mu_p)^2] + \frac{\sigma_q^2}{2\sigma_q^2}$

   $= \log(\frac{\sigma_q}{\sigma_p}) - \frac{1}{2\sigma_p^2} \mathbb{E}_{q_\phi(z|x)}[(x-\mu_q+\mu_q-\mu_p)^2] + \frac{1}{2}$
6. Recall, $(a+b)^2 = a^2 + 2ab + b^2$

   $\Rightarrow D_{KL}(q_\phi(z|x)|p(z)) = \log(\frac{\sigma_q}{\sigma_p}) - \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \frac{1}{2}$

7. Yet, for $z \sim \text{Normal}, \sigma_p = 1, \mu_p = 0$
   Thus, $-D_{KL}(q_\phi(z|x)|p(z)) = \frac{1}{2}[1 + \log(\sigma_q^2) - \sigma_q^2 - \mu_q^2]$

8. Thus, plugging into ELBO we have the following objective function for the latent space with dimension J,

$$\mathcal{L} = -\sum_{j=1}^{J} \frac{1}{2}[1 + \log(\sigma_j^2) - \sigma_j^2 - \mu_j^2] - \frac{1}{L}\sum_l \mathbb{E}_{q_\phi(z|x)}[\log p(x|z^l)]$$

5 Variational Information Bottleneck (VIB)

Fortunately, the loss function for VIB, is almost identical to the ELBO loss for the VAE. Below is the objective for VIB when the prior is normal.

$$\mathcal{L} = -\sum_{j=1}^{J} \frac{1}{2}[1 + \log(\sigma_j^2) - \sigma_j^2 - \mu_j^2] - \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y|z)$$

6 VIB with Normalizing Flow

$$\mathcal{L}(\theta, \phi; x, y) = \mathbb{E}_{q\phi}[\log q_\phi(z|x) - \log p_\theta(y, z)]$$

$$= \mathbb{E}_{q_0}[\log q_0(z_0|x) - \log p_\theta(y, z)] - \mathbb{E}_{q_0}[\sum_{k=1}^{K} \log|\det \frac{\partial f_k(z_{k-1}; \lambda_k(x))}{\partial z_{k-1}}|)]$$

$$= \mathbb{E}_{q_0}[\log q_0(z_0|x) - \log p_\theta(y|z) - \log p_\theta(z)] - \mathbb{E}_{q_0}[\sum_{k=1}^{K} \log|\det \frac{\partial f_k(z_{k-1}; \lambda_k(x))}{\partial z_{k-1}}|)]$$

- $\log q_0(z_0|x) = -\frac{1}{2}\sum_d[\log(2\pi) + \log(\sigma_d^2) + (\frac{(z_0 - \mu_d)}{\sigma_d^2})^2]$, log prob of normal
- $\log p_\theta(z) = -\frac{1}{2}\sum_d[\log(2\pi) + z_d^2]$, log prob of standard normal
- $\log p_\theta(y|z) = CrossEntropy(y, \hat{y})$
- $\sum_{k=1}^{K} \log|\det \frac{\partial f_k(z_{k-1}; \lambda_k(x))}{\partial z_{k-1}}|)]$ depends on chosen flow

```
import tensorflow_probability as tfp
import tensorflow as tf

def gaussian_log_pdf(z, mu, var):
    """
    Log probability from a diagonal covariance normal distribution.
    """
    return tfp.distributions.MultivariateNormalDiag(
                loc = mu, scale_diag = tf.maximum(tf.sqrt(var),
                  1e-4)).log_prob(z + 1e-4)

bce_loss = tf.keras.losses.BinaryCrossentropy(from_logits=True,
    reduction=tf.keras.losses.Reduction.SUM)

# Foward Pass
z0, mu, var = Encoder(x)
zk, sum_logdet_jacobian = FLOWS(z0)
yh = Generator(zk)

# Calculate loss
log_q0_z0 = gaussian_log_pdf(z0, mu, var)
log_qk_zk = log_q0_z0 - sum_logdet_jacobian
log_p_zk =  gaussian_log_pdf(zk, tf.zeros_like(mu),
            tf.ones_like(mu))
log_p_x_given_zk = log_q0_z0 - sum_logdet_jacobian
kl_loss = log_qk_zk - log_p_zk
recons_loss = bce_loss(y,yh)

# Final loss
elbo_loss = tf.reduce_mean(kl_loss + recons_loss)
```

Listing 1: VIB with Flows