Basics VAE

March 11, 2022

Useful papers - [Doe16]; [KW13]

1 Introduction

Let $x_1, ..., x_n$ - independent and identically distributed (i.i.d) random variables from distribution $P_{data}(X)$ (let‘s say this would be training dataset). Goal of any generative model is to reproduce the distribution $P_{data}(X)$ (or at least be able to make samples from it). VAE works with latent variables Z from some distribution $P(Z)$. The main idea behind latent distribution is to reproduce initial object (X variable) using hidden description and variable Z can be easy sampled from $P(Z)$. Then we need to build parameterized function $f : Z \times \Theta \to X$ to display hidden variable to initial. In this way we create samples from conditional distribution $P(X|Z, \Theta)$. We can obtain marginal distribution as

$$\begin{gathered} P(X|\Theta) = \int\limits_Z P(X, Z|\Theta) dz = [P(X, Z|\Theta) = P(Z, X|\Theta) = P(X|Z, \Theta)P(Z|\Theta), P(Z|\Theta) = P(Z)] = \\ = \int\limits_Z P(X|Z, \Theta)P(Z) dz = \mathbb{E}_{z\sim P}\, P(X|Z, \Theta) \approx \frac{1}{m}\sum_i P(X|z_i, \Theta) \end{gathered} \tag{1}$$

This marginal will be an estimate of our data distribution $P_{data}(X)$. VAE using following assumptions: $P(X|Z, \Theta) = \mathbf{N}(X|f_\Theta(Z), \sigma I)$, $P(Z) = \mathbf{N}(Z|0, I)$ .We can use maximum log likelihood estimation for approximation data distribution:

$$\begin{gathered} \Theta^* = \underset{\Theta}{\operatorname{argmax}}\, \mathbb{E}_{x\sim P_{data}}\, \mathbb{E}_{z\sim P} \log P(X|Z, \Theta) \approx \underset{\Theta}{\operatorname{argmax}} \frac{1}{m}\frac{1}{n}\sum_i\sum_j P(x_j|z_i, \Theta) \\ P_{data}(X) \approx P(X|\Theta^*) \end{gathered} \tag{2}$$

The main problem consist in integral 1. We need a lot of samples from $P(Z)$ for Monte-Carlo estimation, it‘s computational expensive. Also for most z $P(X|z, \Theta)$ will be nearly zero, because $P(Z)$ does not not carry any information about $P(X)$. The key idea of VAE is to sample values of z that are likely to produce samples from $P(X)$. For this, a new distribution is introduced - $Q(Z|X, \Phi)$. For given X we construct $Q(Z|X, \Phi)$ and sample z that a likely to produce X. Hopefully, the space of z under $Q(Z|X, \Phi)$ will be much smaller than under $P(Z)$ and we can effectively compute $\mathbb{E}_{z\sim Q}\, P(X|Z, \Theta)$. Now we need to connect $Q(Z|X, \Phi)$ and $P(X|\Theta)$.

$$\begin{gathered} \log P(X|\Theta) = \int\limits_Z Q(Z|X, \Phi) \log P(X|\Theta) dz = [P(X|\Theta) = \frac{P(X, Z|\Theta)}{P(Z|X, \Theta)}, P(Z|X, \Theta) = P(Z|X)] = \\ = \int\limits_Z Q(Z|X, \Phi) \log \frac{P(X, Z|\Theta)}{P(Z|X)} dz = \int\limits_Z Q(Z|X, \Phi) \log \frac{P(X, Z|\Theta)Q(Z|X, \Phi)}{P(Z|X)Q(Z|X, \Phi)} dz = \\ = \int\limits_Z Q(Z|X, \Phi) \log \frac{P(X, Z|\Theta)}{Q(Z|X, \Phi)} dz + \int\limits_Z Q(Z|X, \Phi) \log \frac{Q(Z|X, \Phi)}{P(Z|X)} dz = \\ = \int\limits_Z Q(Z|X, \Phi) \log P(X|Z, \Theta) dz + \int\limits_Z Q(Z|X, \Phi) \log \frac{P(Z)}{Q(Z|X, \Phi)} dz + \int\limits_Z Q(Z|X, \Phi) \log \frac{Q(Z|X, \Phi)}{P(Z|X)} dz \end{gathered} \tag{3}$$

Finally, we have

$$\log P(X|\Theta) - \mathbf{KL}(Q(Z|X,\Phi)||P(Z|X)) = \mathbb{E}_{z\sim Q}\log P(X|Z,\Theta) - \mathbf{KL}(Q(Z|X,\Phi)||P(Z)) = \mathcal{L}(\Theta,\Phi)$$

$$\log P(X|\Theta) = \mathcal{L}(\Theta,\Phi) + \mathbf{KL}(Q(Z|X,\Phi)||P(Z|X)) \tag{4}$$

$\mathcal{L}(\Theta,\Phi)$ calls evidence lower bound.

Definition 1 *Variational lower bound*

Function g(x, y(x)) calls lower bound for function f(x) if and only if
1. $\forall x, y \quad g(x, y(x)) \leq f(x)$
2. $\exists x_0 : g(x_0, y(x_0) = f(x_0)$

As can be seen $\mathcal{L}(\Theta,\Phi)$ satisfy conditionals of definition:

$$\begin{gathered} 1. \log P(X|\Theta) \leq \mathcal{L}(\Theta,\Phi); \\ \mathbf{KL}(Q(Z|X,\Phi)||P(Z|X)) \geq 0, \forall \Theta, \Phi \\ \\ 2. \ \exists \Phi^* : Q(Z|X,\Phi^*) = P(Z|X) \Rightarrow \log P(X|\Theta) = \mathcal{L}(\Theta,\Phi^*) \end{gathered} \tag{5}$$

So, the main idea of VAE is to maximize ELBO instead of maximzing likelihood directly, because $\mathbf{KL}(Q(Z|X,\Phi)||P(Z|X))$ is intractable, because we dont know $P(Z|X)$. Lets calculate $\mathbf{KL}(Q(Z|X,\Phi)||P(Z))$ with VAE assumptions: $P(Z) = \mathbf{N}(Z|0,I)$ and $Q(Z|X,\Phi) = \mathbf{N}(Z|\mu(X,\Phi),\sigma(X,\Phi))$

$$\begin{aligned} \mathbf{KL}(\mathbf{N}(Z|\mu(X,\Phi),\sigma(X,\Phi))||\mathbf{N}(Z|0,I)) &= \int_Z \mathbf{N}(Z|\mu,\sigma)\log\frac{\mathbf{N}(Z|\mu,\sigma)}{\mathbf{N}(Z|0,I)}dz = \\ = \int_Z \mathbf{N}(Z|\mu,\sigma)\log\mathbf{N}(Z|\mu,\sigma)dz &- \int_Z \mathbf{N}(Z|\mu,\sigma)\log\mathbf{N}(Z|0,I)dz = \mathbf{I}_1 - \mathbf{I}_2 \end{aligned} \tag{6}$$

Let $Z,\mu,\sigma \in \mathbb{R}^n$. Also $z_1,...,z_n-$ independent random variables, that's why $\mathbf{Cov}(Z) = diag(\sigma)$

$$\begin{aligned} \mathbf{N}(Z|\mu,\sigma) &= \frac{1}{(2\pi)^{\frac{n}{2}}\mathbf{Cov}^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(Z-\mu)^T\mathbf{Cov}^{-1}(Z-\mu)\right) = \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\prod_i\sigma_i}\exp\left(-\frac{1}{2}\sum_j(z_j-\mu_j)^2\frac{1}{\sigma_j^2}\right) \\ \mathbf{N}(Z|0,I) &= \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}Z^TZ\right) = \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}\sum_j z_j^2\right) \end{aligned} \tag{7}$$

Let's consider $\mathbf{I}_2$

$$\begin{gathered} \mathbf{I}_2 = \frac{1}{(2\pi)^{\frac{n}{2}}\prod_i\sigma_i}\int_Z \exp\left(-\frac{1}{2}\sum_j(z_j-\mu_j)^2\frac{1}{\sigma_j^2}\right)\log\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}\sum_j z_j^2\right)\right) = \\ = \frac{1}{(2\pi)^{\frac{n}{2}}\prod_i\sigma_i}\left[\int_Z \exp\left(-\frac{1}{2}\sum_j(z_j-\mu_j)^2\frac{1}{\sigma_j^2}\right)\log\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}\int_Z\sum_j z_j^2\exp\left(-\frac{1}{2}\sum_j(z_j-\mu_j)^2\frac{1}{\sigma_j^2}\right)\right] = \\ = \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2(2\pi)^{\frac{n}{2}}\prod_i\sigma_i}\mathbf{J} \end{gathered} \tag{8}$$

$$\mathbf{J} = \int\limits_Z \sum_j z_j^2 \exp\left(-\frac{1}{2}\sum_j (z_j-\mu_j)^2 \frac{1}{\sigma_j^2}\right) = \int\limits_Z z_1^2 \exp\left(-\frac{1}{2}\sum_j (z_j-\mu_j)^2 \frac{1}{\sigma_j^2}\right) + ...+$$
$$+ \int\limits_Z z_n^2 \exp\left(-\frac{1}{2}\sum_j (z_j-\mu_j)^2 \frac{1}{\sigma_j^2}\right) =$$
$$= \mathbf{J_1} + ... + \mathbf{J_n}$$
$$\mathbf{J_1} = \int\limits_{-\infty}^{\infty} z_1^2 \exp\left(-\frac{1}{2}(z_1-\mu_1)^2\frac{1}{\sigma_1^2}\right) ... \int\limits_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(z_n-\mu_n)^2\frac{1}{\sigma_n^2}\right) = \int\limits_{-\infty}^{\infty} z_1^2 \exp\left(-\frac{1}{2}(z_1-\mu_1)^2\frac{1}{\sigma_1^2}\right)(2\pi)^{\frac{n-1}{2}} \prod_{i=2}^{n} \sigma_i =$$
$$= (2\pi)^{\frac{n-1}{2}} \prod_{i=2}^{n} \sigma_i \int\limits_{-\infty}^{\infty} ((z_1-\mu_1)^2 + 2\mu_1 z_1 - \mu_1^2) \exp\left(-\frac{1}{2}(z_1-\mu_1)^2\frac{1}{\sigma_1^2}\right) =$$
$$= (2\pi)^{\frac{n-1}{2}} \prod_{i=2}^{n} \sigma_i \left[\int\limits_{-\infty}^{\infty} (z_1-\mu_1)^2 \exp\left(-\frac{1}{2}(z_1-\mu_1)^2\frac{1}{\sigma_1^2}\right) + 2\mu_1 \int\limits_{-\infty}^{\infty} z_1 \exp\left(-\frac{1}{2}(z_1-\mu_1)^2\frac{1}{\sigma_1^2}\right)\right] -$$
$$-(2\pi)^{\frac{n-1}{2}} \prod_{i=2}^{n} \sigma_i \mu_1^2 \int\limits_{-\infty}^{\infty} ((z_1-\mu_1)^2) \exp\left(-\frac{1}{2}(z_1-\mu_1)^2\frac{1}{\sigma_1^2}\right) = (2\pi)^{\frac{n}{2}} \prod_{i}^{n} \sigma_i (\sigma_1^2 + \mu_1^2)$$
$$\mathbf{J} = (2\pi)^{\frac{n}{2}} \prod_{i}^{n} \sigma_i \sum_{i}^{n} (\sigma_i^2 + \mu_i^2)$$
$$\mathbf{I}_2 = \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}\sum_i (\sigma_i^2 + \mu_i^2) \tag{9}$$

Analogue

$$\mathbf{I}_1 = \log\left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) - \frac{1}{2}\sum_i (1 + \log\sigma_i^2) \tag{10}$$

Finally

$$\mathbf{KL}(\mathbf{N}(Z|\mu(X,\Phi),\sigma(X,\Phi))||\mathbf{N}(Z|0,I)) = -\frac{1}{2}\sum_i (1 + \log\sigma_i(X,\Phi)^2 - \sigma_i(X,\Phi)^2 - \mu_i(X,\Phi)^2) \tag{11}$$

Our goal is

$$\underset{\Theta,\Phi}{\operatorname{argmax}}\, \mathbb{E}_{x\sim P_{data}} \mathcal{L}(\Theta,\Phi) = \underset{\Theta,\Phi}{\operatorname{argmax}}\, \mathbb{E}_{x\sim P_{data}} \left(\mathbb{E}_{z\sim Q} \log P(X|Z,\Theta) - \mathbf{KL}(Q(Z|X,\Phi)||P(Z))\right) \approx$$
$$\approx \underset{\Theta,\Phi}{\operatorname{argmax}} \frac{1}{N}\left(\frac{1}{m}\sum_{j=1}^{N}\sum_{i=1}^{m} \log P(x_j|z_i,\Theta) + \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{n}\left(1 + \log\sigma_k(x_j,\Phi)^2 - \sigma_k(x_j,\Phi)^2 - \mu_k(x_j,\Phi)^2\right)\right) =$$
$$= \left[\log P(x_j|z_i,\Theta) = \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{||x_j - z_i||^2}{2\sigma^2})\right)\right] =$$
$$= \underset{\Theta,\Phi}{\operatorname{argmax}} \frac{1}{N}\sum_{j=1}^{N}\left(-\frac{1}{m}\sum_{i=1}^{m}||x_j - z_i||^2 + \frac{1}{2}\sum_{k=1}^{n}\left(1 + \log\sigma_k(x_j,\Phi)^2 - \sigma_k(x_j,\Phi)^2 - \mu_k(x_j,\Phi)^2\right)\right) -? \tag{12}$$

It‘s looks like

References

[Doe16] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

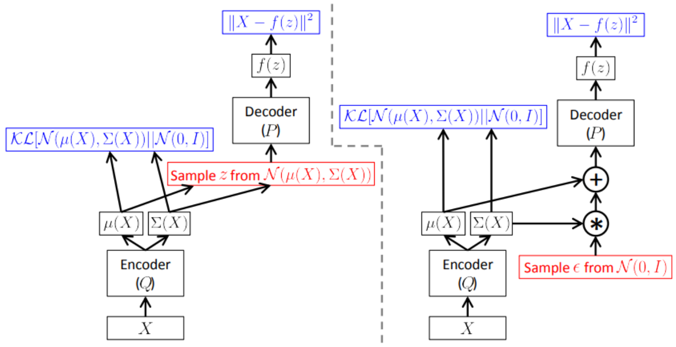

Figure 1: Comparison between the best individuals (archive population) for proposed approaches with surrogate and real model.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.