Variational Autoencoder

Proposed by Kingma & Welling in 2014.

Input Encoder → Distribution over latent variables → Decoder → Output x → z

Encoder and Decoder implemented by two neural networks.

The latent variable z encapsulates the information used to describe the input.

At the same time, z, is the "seed" used by the decoder to generate x.

Two different z's should give us two different generated objects.

Definition: Latent Variables

In a probabilistic model, latent variables (or unobserved variables) are those not directly available in the learning dataset but are included as part of the model's structure to capture hidden or underlying factors influencing the observed data.

A bunch of new distributions.

$p(x)$: The true distribution of x. Unknown. If we know $p(x)$ in order to generate we could draw a sample x to maximize $\log p(x)$.

Self information: $I(x) = -\log p(x)$

Low probability → High information. Rare events are more surprising. High probability → Low information. Common events are predictable, unsurprising.

Entropy: $H(p) = E_{x\sim p(x)}[-\log p(x)]$ = $-\sum p(x)\log p(x)$. Is the average amount of information (aka the overall uncertainty) of $p(x)$

$p(z)$: Distribution of the latent variable. Usually $N(0,1)$. Any distribution can be generated by mapping a Gaussian through a sufficiently complicated function (e.g. neural network).

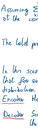
$p(x|z)$: Conditional distribution associated to the decoder. We do not have access to it. It is not the decoder itself. Likelihood of x when given z.

$p(z|x)$: Conditional distribution associated to the encoder. Posterior probability of getting z given x.

Deep Latent Variables:

$p(x) = \int p(x|z)\,p(z)\,dz$, parameterized by neural networks.

Toy example



X random variable distributed according to a gaussian mixture model. Pick 1 of 2 centers $k \in \{1,2\}$ according to the latent variable $p(k)$. $P(k=1)=\pi$, $P(k=2)=1-\pi$.

Assuming $\sum_k \pi_k = 1$ if we need to pick only in decoder 1 we look at the conditional distribution

$$P(x|z\,(X)) = N(x|\mu_1,\sigma_1^2)$$

The total probability is:

$$P_\theta(x) = \sum_{k=1}^{2} P_\theta(x|k)\,p_\theta(k) = \sum_{k=1}^{2} \pi_k\, N(x|\mu_k,\sigma_k^2)$$

In this scenario, given $p(x|z)$ we want to learn to encoder such that from any sample $x \sim p(x)$ the latent code $z \in \{1,2\}$ with a distribution $z \sim p(z)$.

Encoder: How do we obtain z from x? $z = \begin{cases} 1 & P_\theta(1|x) \geq P_\theta(2|x) \\ 2 & \text{otherwise} \end{cases}$

Decoder: Since we know means and std we just need to sample from the conditional distribution

$$P_\theta(x|z=k) = N(x|\mu_k,\sigma_k^2)$$

Alternatively, it gives one of the 2 mixture components.

We use an encoder parameterized by θ as a proxy for $p(z|x)$

$(\mu,\sigma^2) = \text{Encoding}_\theta(x) \rightarrow q_\theta(z|x) = N(z|\mu_\theta(x),\sigma_\theta^2(x))$

We use a decoder parameterized by θ as a proxy for $p(x|z)$

$p_\theta(x) = \text{Decoder}_\theta(z) \rightarrow q_\theta(x|z) = N(x|f_\theta(z),\sigma_\theta^2)$ ← hyperparameter

Trivial and known cases

$x \in \mathbb{R}^d$

$\alpha \rightarrow p(x) = N(x|\mu,\sigma^2)$

$\alpha \rightarrow p(x) = N(x|0,1)$

with $p(x)$ known. $\varepsilon = \frac{(x-\mu)}{\sigma}$ $\hat{x} = \mu + \sigma\varepsilon$

we can use the sampling function to compute the confounds:

$p(x|z) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$

$p(x|z) = \frac{1}{\sigma}\phi\left(\varepsilon - \frac{(x-\mu)}{\sigma}\right)$

If $p(x)$ is unknown we could use a linear layer for the mean and a softplus for the variance.

$q_\theta(z|x) = N(z|\,x\,w_1 b_1,\ L^2)$

$p_\theta(x|z) = N(x|\,z\,w_2 + b_2^2)$

But how do we obtain the best parameters?

Our goal is to maximize $\log p(x)$ which is unknown. We need to maximize it in tractable cases:

$\log p(x) = \log p(x) \int q_\theta(z|x)\,dz$ Multiply by 1

$= \int q_\theta(z|x)\,\log p(x)\,dz$

constant wrt z Distribution wrt z $\int q_\theta(z|x) = E_{q(z|x)}$

$= E_{q(z|x)}[\log p(x)]$ we introduced the encoder!

Now we can begin to introduce $P(x|z)$ $P(z|x)$, $P(x)$ $P(z)$

$= E_{q(z|x)}\left[\log \frac{p(x|z)}{p(z|x)}\right]$ Let's multiply by 1 again

$= E_{q(z|x)}\left[\log \frac{p(x|z)}{p(z|x)} \cdot \frac{q_\theta(z|x)}{q_\theta(z|x)}\right]$

$= E_{q(z|x)}\left[\log \frac{p(x|z)}{q_\theta(z|x)}\right] + E_{q(z|x)}\left[\log \frac{q_\theta(z|x)}{p(z|x)}\right]$

ELBO

$D_{KL}(q_\theta(z|x)\,\|\,p(z|x))$ ← Unknown

The Evidence Lower Bound will be our Loss function

$\log p(x) = \underbrace{D_{KL}(q_\theta(z|x)\,\|\,p(z|x))}_{\text{Unknown}} + \underbrace{ELBO(x)}_{\text{Known}}$

$ELBO(x) = E_{q(z|x)}\left[\log \frac{p(x,z)}{q_\theta(z|x)}\right]$ Let's use Bayes again

$= E_{q(z|x)}\left[\log \frac{p(x|z)\,p(z)}{q_\theta(z|x)}\right]$

$= E_{q(z|x)}\left[\log p(x|z)\right] + E_{q(z|x)}\left[\log \frac{p(z)}{q_\theta(z|x)}\right]$

we need to replace the intractable $p(x|z)$ → $P_\theta(x|z)$ and decoder

$= E_{q(z|x)}[\log P_\theta(x|z)] - D_{KL}(q_\theta(z|x)\,\|\,p(z))$

Reconstruction Decoder (the part we care about) Approx to the likelihood. The expectation tells you that it doesn't be as important over order but not be complete from $q_\theta(z|x)$ the closest latent space

Prior Matching (Regularizer) Encoder (too good at it) The encoder tries to cheat and fake a pattern on target distribution $N(0,\sigma^2 I)$

Limitation of ELBO

Let's use our by example $x \in \mathbb{R}^d$ $z \in \mathbb{R}^d$. Let's assume that we know the unknown order but not the variance. The mean is correlated with the latent function of $P(x|z)$ and $p(z)$

$p_\theta(x|z) = N(x|\frac{z}{\sigma^2}, I^2)$

$p_\theta(x|z) = N(x|w_2 z + b, I^2)$

To Maximize the ELBO we compute the derivatives respect to θ and φ. The derivative of the $D_{KL}(q_\theta(z|x)\,\|\,p(z))$ has closed form solution, however we have two unknowns expectation.

We just that $\sigma = \frac{1}{\sigma^2}$

Similarly taking the derivative with respect to φ and setting it to zero gives us $\varphi = \frac{z}{\sigma^2}$

Therefore $q_\varphi(z|x) = N(x|\frac{z}{\sigma^2}, \frac{1}{\sigma^2})$

$p_\theta(x|z) = N(x|w_2 z + b, I^2)$

The variance of latent had our idea. This leaving the ELBO is not entirely tractable.

These funk variances will contribute to the gradient. Interestingly, it scales linearly with $\frac{1}{\sigma^2}$ when σ is the dimension of the input. Therefore when σ is large $\frac{1}{\sigma^2}$ will be exponentially against the best distribution.

Optimization in VAE

Goal: Approximate the true distribution $p(x)$.

How: For a VAE parameterized by (θ,φ) we optimize the lower bound of $\log p(x)$

$$(\hat{\theta},\hat{\varphi}) = \underset{\theta,\varphi}{\operatorname{argmax}} \sum_{x\in X} ELBO(x)$$

→ Learning Dataset

Harness the gradient ∇_θ is intractable ↯ Let's see why

$$\nabla_\theta ELBO(x) = \nabla_\theta \int q_{\theta(z|x)}\left[\log \frac{p(x,z)}{q_\theta(z|x)}\right] \qquad \nabla_\theta \int q \cdot \log \cdot dz$$

$$= \int \nabla_\theta \left[q_\theta(z|x)\cdot \log p(x,z) - \log q_\theta(z|x)\right]$$

Let's derive further

$$\nabla_\theta \int q_\theta(z|x)\cdot \log p(x,z) - \int \nabla_\theta q_\theta(z|x)\cdot \log q_\theta(z|x)$$

Problem $\int q_\theta(z|x)\cdot E_{q(z|x)}$

Let's take again

$$= \int q_\theta(z|x)\cdot \nabla_\theta \log p(x,z) - \int q_\theta(z|x)\cdot \nabla_\theta \log q_\theta(z|x)$$

does not depend on θ

$$= E_{q(z|x)}[\nabla_\theta \log p(x,z)]$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\theta \log p(x,z^l) \quad \text{where } z^l \sim q_\theta(z|x)$$

If P_θ is a neural network we can backpropagate this gradient without issues

Now we know that the issue is with φ

$$\nabla_\varphi ELBO(x) = \nabla_\varphi \int q_{\theta(z|x)}\left[\log p(x,z) - \log q_\theta(z|x)\right] \qquad \text{Let's expand the gradient of the expectation}$$

$$= \int \nabla_\varphi \left[q_\theta(z|x)\cdot \log p(x,z) - q_\theta(z|x)\cdot \log q_\theta(z|x)\right]$$

$$= E_{q(z|x)}\left[\nabla_\varphi \log p(x,z) - \nabla_\varphi \log q_\theta(z|x)\right]$$

$$= E_{q(z|x)}\left[\nabla_\varphi \log p(x,z) - \log q_\theta(z|x)\cdot \nabla_\varphi \log q_\theta(z|x)\right] \approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\varphi \left[\log p(x,z^l) - \log q_\theta(z^l|x)\right]$$

where $z^l \sim q_\theta(z|x)$

Non differentiability of operators

$E_{q(z|x)}[f(x)] = \int f(x)\,q_\theta(z|x)\,dz$

only if f is the gradient of $f(x)$ wrt $q_\theta(z|x)$ not of q_θ and continuous

$\int f(x)\,\nabla_\theta q_\theta(z|x)\,dz$ → depends on θ

now since $q_\theta(z|x)$ depends on θ and what $f(x)$ depends on it. The gradient acts on the function inside the expectation and the expectation. Switching the continuity of the gradient

Reparameterization Trick

The intractability of the gradients. Solution from the first that we know complex. $z \sim q_\theta(z|x)$ which itself a function of θ.

The reparameterization trick expand a sec. backpropagation of random variable is independent from θ and φ

$$z = g(\varepsilon,\theta,x) \quad \varepsilon \sim p \quad \text{then } \varepsilon \sim N(0,I)$$

$$\varepsilon \sim N(\mu,\sigma^2 I)$$

which leads to $q_\theta(z|x) = \left|\frac{\partial g(\varepsilon)}{\partial \varepsilon}\right| \cdot q(\varepsilon)$

Let's rewrite the expectation in terms of $p(\varepsilon)$

$$E_{q(z|x)}[\log q_\theta(z|x)] = \int \log q_\theta(z|x)\cdot q_\theta(z|x)\,dz$$

$$= \int \log q_\theta(g(\varepsilon)|x)\cdot q_\theta(g(\varepsilon)|x)\,\left|\frac{\partial g(\varepsilon)}{\partial \varepsilon}\right|\,p(\varepsilon)\,d\varepsilon$$

Jacobian matrix by the Inverse of variance

$$= \int \log q_\theta(g(\varepsilon)|x)\cdot q_\theta(g(\varepsilon)|x)\,\left|\frac{\partial g(\varepsilon)}{\partial \varepsilon}\right|\,p(\varepsilon)\,d\varepsilon$$

$$= \int \log q_\theta(g(\varepsilon)|x)\,p(\varepsilon)\,d\varepsilon$$

$$= E_{p(\varepsilon)}[\log q_\theta(g(\varepsilon)|x)]$$

Now we can compute the gradient wrt φ

$$\nabla_\varphi E_{q(z|x)}[\log q_\theta(z|x)] = \nabla_\varphi E_{p(\varepsilon)}[\log q_\theta(g(\varepsilon)|x)]$$

$$= \int \nabla_\varphi \log q_\theta(g(\varepsilon)|x)\,p(\varepsilon)\,d\varepsilon$$

$$= \int \nabla_\varphi \log q_\theta(g(\varepsilon)|x)\,p(\varepsilon)\,\frac{q_\theta(z|x)}{q_\theta(z|x)}\,d\varepsilon$$

$$= \int \nabla_\varphi \log q_\theta(g(\varepsilon)|x)\,\frac{p(\varepsilon)}{q_\theta(z|x)}\,q_\theta(z|x)\,d\varepsilon$$

$$= E_{q(z|x)}\left[\nabla_\varphi \log q_\theta(g(\varepsilon)|x)\,\frac{p(\varepsilon)}{q_\theta(z|x)}\right]$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\varphi \log q_\theta(g(\varepsilon^l)|x)\,\frac{p(\varepsilon^l)}{q_\theta(z^l|x)} \quad \text{where } \varepsilon^l \sim p(\varepsilon)$$

The network parameter (in encoder and decoder) affect the input, image and latent is change → will $\varepsilon \sim N(0,I)$

$z \rightarrow \begin{cases} \mu_\theta \\ \sigma_\theta \end{cases}$

Let's simplify that

$$\nabla_\varphi \log q_\theta(g(\varepsilon)|x) = \nabla_\varphi \left[\log \left|\frac{\partial g(\varepsilon)}{\partial \varepsilon}\right| + \log q_\theta(z|x)\right]$$

$$= \nabla_\varphi \left[\log \left|\frac{\partial g(\varepsilon)}{\partial \varepsilon}\right| + \log \frac{1}{\sigma}\right]$$

$$= \nabla_\varphi \left[\log \left|\frac{\partial g(\varepsilon)}{\partial \varepsilon}\right| - \log \sigma