Deep Learning

19 Variational Autoencoder

Dr. Konda Reddy Mopuri
Dept. of AI, IIT Hyderabad

Jan-May 2025

Autoencdoers

1. Designed to reproduce input, especially reproduce the input from a learned encoding

Autoencdoers

1. Designed to reproduce input, especially reproduce the input from a learned encoding
2. We attempted to project the data into the latent space and model it via a probability distribution

Autoencdoers

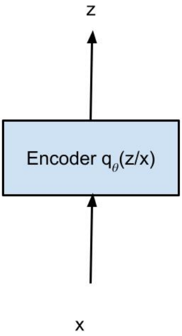
1. Designed to reproduce input, especially reproduce the input from a learned encoding
2. We attempted to project the data into the latent space and model it via a probability distribution
3. This wasn't satisfying

Variational Autoencoders

1. ‘Regularized’ autoencoder to enforce latent space ‘organization’

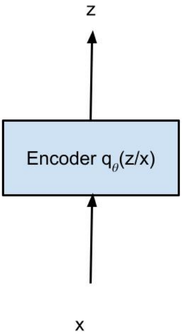
Variational Autoencoders

1. Key idea is to make both Encoder and Decoder stochastic
    - instead of encoding an i/p as a single point, we encode it as a distribution over the latent space



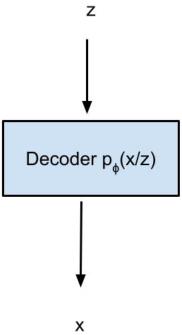
Variational Autoencoders

1. Key idea is to make both Encoder and Decoder stochastic
   - instead of encoding an i/p as a single point, we encode it as a distribution over the latent space
2. Latent variable z is drawn from a probability distribution for the given input x



Variational Autoencoders

1. Then, the reconstruction is chosen probabilistically from the sampled z



VAE Encoder

1. Takes i/p and returns the parameters of a probability density (e.g. Gaussian, mean and covariance matrix)

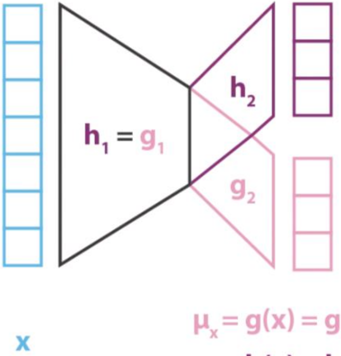
VAE Encoder

1. Takes i/p and returns the parameters of a probability density (e.g. Gaussian, mean and covariance matrix)
2. We can sample this to get random values of the latent variable z

VAE Encoder

1. Takes i/p and returns the parameters of a probability density (e.g. Gaussian, mean and covariance matrix)
2. We can sample this to get random values of the latent variable z
3. NN implementation of the encoder gives (for every input x) a vector mean and a diagonal covariance

VAE Encoder



$$\mu_x = g(x) = g_2(g_1(x))$$

$$\sigma_x = h(x) = h_2(h_1(x))$$

VAE Decoder

1. Decoder takes the latent vector z and returns the parameters for a distribution

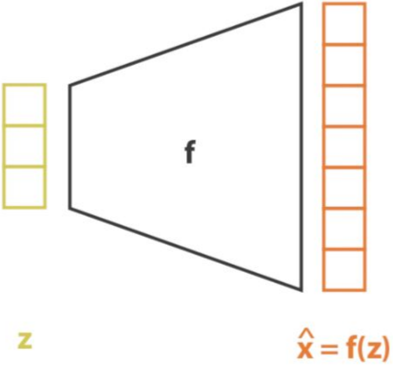
VAE Decoder

1. Decoder takes the latent vector z and returns the parameters for a distribution
2. $p_\phi(x/z)$ gives mean and variance for each pixel in the output

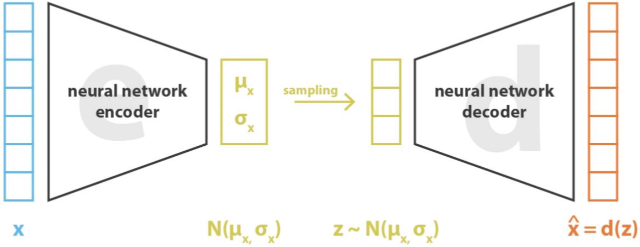
VAE Decoder

1. Decoder takes the latent vector z and returns the parameters for a distribution
2. $p_\phi(x/z)$ gives mean and variance for each pixel in the output
3. Reconstruction of x is via sampling

VAE Decoder



VAE Forward pass



VAE loss function

1. Loss for AE: l_2 distance between the input and its reconstruction

VAE loss function

1. Loss for AE: l_2 distance between the input and its reconstruction
2. In case of VAE: we need to learn parameters of two probability distributions

VAE loss function

1. Loss for AE: l_2 distance between the input and its reconstruction
2. In case of VAE: we need to learn parameters of two probability distributions
3. For each input x_i we maximize expected value of returning x_i (or, minimize the NLL)

$$-\mathbb{E}_{z \sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)]$$

VAE loss function

$$-\mathbb{E}_{z\sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)]$$

1. Problem: Input images may be memorized in the latent space

VAE loss function

$$-\mathbb{E}_{z\sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)]$$

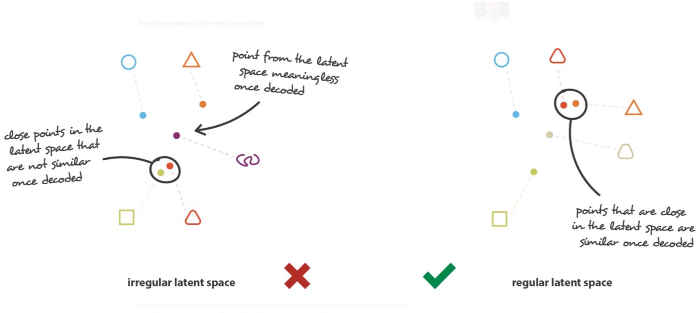
1. Problem: Input images may be memorized in the latent space
   - → similar inputs may get different representations in z space

VAE loss function

$$-\mathbb{E}_{z \sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)]$$

1. Problem: Input images may be memorized in the latent space
   - → similar inputs may get different representations in z space
   - → close points in the latent space should not give two completely different contents once decoded

VAE loss function


Intuitions about the regularisation
point from the latent space meaningless once decoded
close points in the latent space that are not similar once decoded
irregular latent space
✗
✓
points that are close in the latent space are similar once decoded
regular latent space

VAE loss function

$$-\mathbb{E}_{z \sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)]$$

1. Continuity and Completeness: We prefer continuous latent representations to give meaningful parameterization (e.g. smooth transition between i/ps)

VAE loss function

$$-\mathbb{E}_{z\sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)]$$

1. Continuity and Completeness: We prefer continuous latent representations to give meaningful parameterization (e.g. smooth transition between i/ps)
2. Solution: Force $q_\theta(z/x_i)$ to be close to a standard distribution (e.g. Gaussian)

VAE loss function

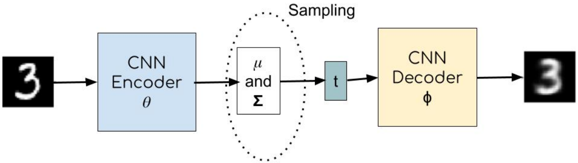
$$l_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)] + \mathbb{KL}(q_\theta(z/x_i)||p(z))$$

1. First term promotes recovery, sencond term keeps encoding continuous (beats memorization)

VAE loss function

$$l_i(\theta, \phi) = -\mathbb{E}_{z\sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)] + \mathbb{KL}(q_\theta(z/x_i)||p(z))$$

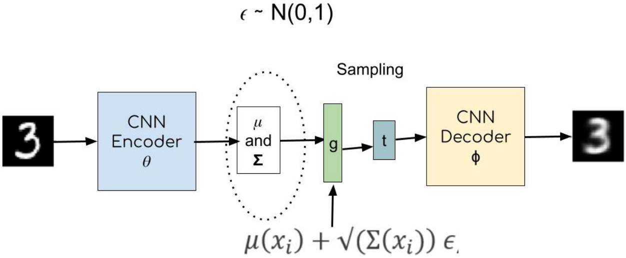
1. Problem: Differentiating over θ and ϕ



VAE loss function

$$l_i(\theta,\phi) = -\mathbb{E}_{z\sim q_\theta(z/x_i)}[log\, p_\phi(x_i/z)] + \mathbb{KL}(q_\theta(z/x_i)||p(z))$$

1. Reparameterization: Draw samples from N(0,1) → doesn't depend on parameters



Generation with VAE

- Sample z from the prior p(z)

Generation with VAE

- Sample z from the prior p(z)
- Run z through the decoder (ϕ) → distribution over data

Generation with VAE

- Sample z from the prior p(z)
- Run z through the decoder (ϕ) $\rightarrow$ distribution over data
- Sample from that distribution to generate the sample x

Generation with VAE

- Sample z from the prior p(z)
- Run z through the decoder (ϕ) $\rightarrow$ distribution over data
- Sample from that distribution to generate the sample x
- For simplicity, in practice, only the means of the pixels are inferred (deterministic)

Generation with VAE

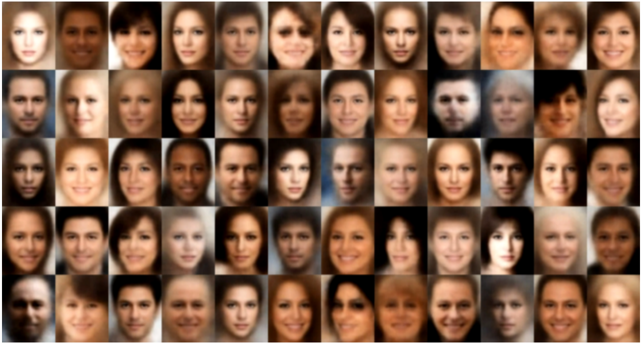


Figure credits: Wojceich

Generation with VAE

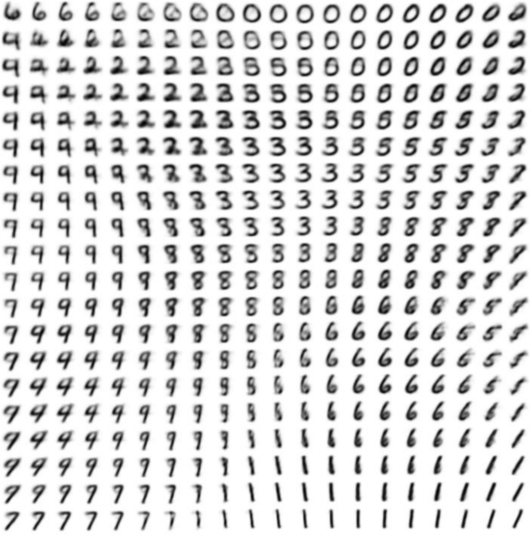
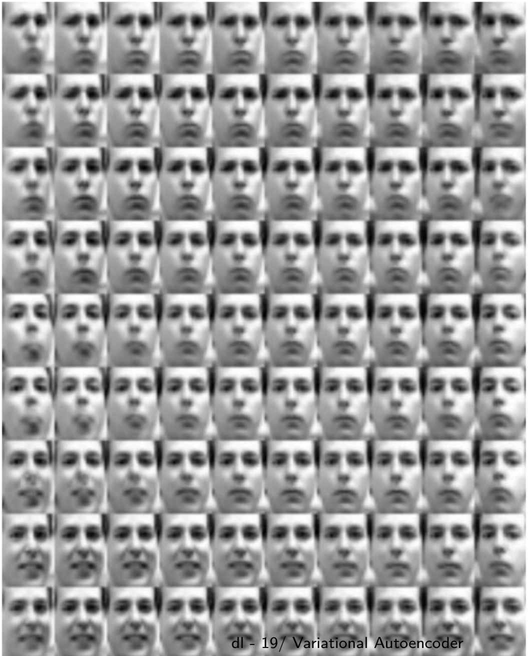


Figure credits: Kingma et al.

Edit/Manipulate samples with VAE



The Evidence Lower Bound (ELBO)

Latent Variable Models

1. Latent variable $\rightarrow$ variable which is not directly observable and is assumed to affect the response variables

Latent Variable Models

1. Latent variable $\rightarrow$ variable which is not directly observable and is assumed to affect the response variables
2. Aim

Latent Variable Models

1. Latent variable → variable which is not directly observable and is assumed to affect the response variables
2. Aim
   - representing the effect of unobservable covariates/factors

Latent Variable Models

1. Latent variable → variable which is not directly observable and is assumed to affect the response variables
2. Aim
   - representing the effect of unobservable covariates/factors
   - account for measurement errors

Latent Variable Models

1. Latent variable → variable which is not directly observable and is assumed to affect the response variables
2. Aim
   - representing the effect of unobservable covariates/factors
   - account for measurement errors
   - controlled/customized generation of the samples

Latent Variable Models

1. They model the probability distribution over latent variables

Latent Variable Models

1. They model the probability distribution over latent variables
2. Because the latent variables explain the data in a simpler way

Latent Variable Models - terminology

1. Data samples x follow a distribution $p(x)$

Latent Variable Models - terminology

1. Data samples x follow a distribution $p(x)$
2. They are mapped on to latent variable z that follow a distribution $p(z)$

Latent Variable Models - terminology

1. Data samples x follow a distribution $p(x)$
2. They are mapped on to latent variable z that follow a distribution $p(z)$
3. $p(z)$ prior distribution that models the behavior of latent variables

Latent Variable Models - terminology

1. $p(x/z)$, likelihood, defines how to map latent variables to the data points

Latent Variable Models - terminology

1. $p(x/z)$, likelihood, defines how to map latent variables to the data points
2. $p(x,z) = p(x/z)p(z)$, describes the model

Latent Variable Models - terminology

1. $p(x/z)$, likelihood, defines how to map latent variables to the data points
2. $p(x,z) = p(x/z)p(z)$, describes the model
3. Marginal distribution $p(x)$ (goal of the model) describes how likely a sample is

Latent Variable Models - terminology

1. $p(x/z)$, likelihood, defines how to map latent variables to the data points
2. $p(x, z) = p(x/z)p(z)$, describes the model
3. Marginal distribution $p(x)$ (goal of the model) describes how likely a sample is
4. $p(z/x)$, posterior, describes the latent variables that can be produced by a data sample

Latent Variable Models - terminology

1. Generation - process of computing the data point x from the latent variable z

Latent Variable Models - terminology

1. Generation - process of computing the data point x from the latent variable z
2. We move from the latent space to the actual data distribution

Latent Variable Models - terminology

1. Generation - process of computing the data point x from the latent variable z
2. We move from the latent space to the actual data distribution
3. Represented by the likelihood $p(x/z)$

Latent Variable Models - terminology

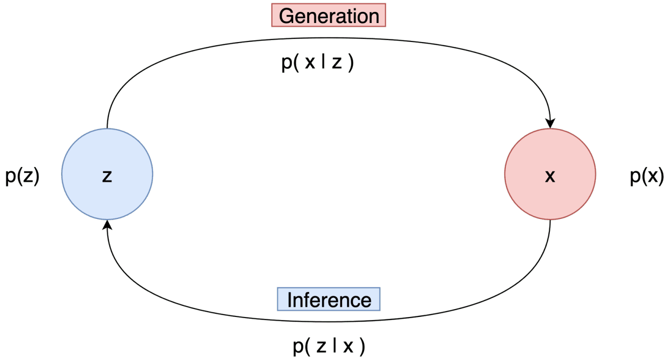
1. Inference - process of finding the latent variable z from the data point x

Latent Variable Models - terminology

1. Inference - process of finding the latent variable z from the data point x
2. Formulated by the posterior distribution $p(z/x)$

Generation-Inference

- If we assume that we (somehow) know the likelihood $p(x/z)$, the posterior $p(z/x)$, the marginal $p(x)$, and the prior $p(z)$



Aim/Question of latent Variable Models

1. How to find these distributions?

These can be connected

1. $p_\theta(z/x) = \frac{p_\theta(x/z) \cdot p_\theta(z)}{p_\theta(x)}$

These can be connected

1. $p_\theta(z/x) = \frac{p_\theta(x/z) \cdot p_\theta(z)}{p_\theta(x)}$
2. $\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{Evidence}}$

These can be connected

1. $p_\theta(z/x) = \frac{p_\theta(x/z) \cdot p_\theta(z)}{p_\theta(x)}$
2. posterior $= \frac{\text{likelihood}\cdot\text{prior}}{\text{Evidence}}$
3. But?

These can be connected

1. $p_\theta(z/x) = \frac{p_\theta(x/z) \cdot p_\theta(z)}{p_\theta(x)}$
2. posterior $= \frac{\text{likelihood} \cdot \text{prior}}{\text{Evidence}}$
3. But?
4. Evidence computation $\int p_\theta(x/z) \cdot p_\theta(z) dz$ (over all the latent space) is intractable → can't compute the LHS

These can be connected

1. $p_\theta(z/x) = \frac{p_\theta(x/z) \cdot p_\theta(z)}{p_\theta(x)}$
2. posterior $= \frac{\text{likelihood} \cdot \text{prior}}{\text{Evidence}}$
3. But?
4. Evidence computation $\int p_\theta(x/z) \cdot p_\theta(z) dz$ (over all the latent space) is intractable → can't compute the LHS
5. Variational inference suggests to use another (known) distribution $q_\phi(z/x)$ to approximate the posterior → (allows to compute the evidence and sample)

Variational Inference

1. $p_\theta(z/x) \approx q_\phi(z/x)$

Variational Inference

1. $p_\theta(z/x) \approx q_\phi(z/x)$
2. We have to learn the parameters of $q_\phi(z/x)$

Variational Inference

1. $p_\theta(z/x) \approx q_\phi(z/x)$
2. We have to learn the parameters of $q_\phi(z/x)$
3. $\rightarrow$ need to formulate an objective that captures the dissimilarity between the GT and approximation

Variational Inference

1. $p_\theta(z/x) \approx q_\phi(z/x)$
2. We have to learn the parameters of $q_\phi(z/x)$
3. $\rightarrow$ need to formulate an objective that captures the dissimilarity between the GT and approximation
4. KL Divergence

Variational Inference

1. $D_{KL} = \mathbb{E}_{q_\phi}\left[log \frac{q_\phi(z/x)}{p_\theta(z/x)}\right]$

Variational Inference

1. $D_{KL} = \mathbb{E}_{q_\phi}\left[log\frac{q_\phi(z/x)}{p_\theta(z/x)}\right]$
2. Note that we don't know the denominator (the GT)

Variational Inference

1. $D_{KL} = \mathbb{E}_{q_\phi}\left[log \frac{q_\phi(z/x)}{p_\theta(z/x)}\right]$
2. Note that we don't know the denominator (the GT)
3. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z/x)]$

Variational Inference

1. $D_{KL} = \mathbb{E}_{q_\phi}\left[log\frac{q_\phi(z/x)}{p_\theta(z/x)}\right]$
2. Note that we don't know the denominator (the GT)
3. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z/x)]$
4. $D_{KL} = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}\left[log\, \frac{p_\theta(z,x)}{p_\theta(x)}\right]$

Variational Inference

1. $D_{KL} = \mathbb{E}_{q_\phi}\left[log \frac{q_\phi(z/x)}{p_\theta(z/x)}\right]$
2. Note that we don't know the denominator (the GT)
3. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z/x)]$
4. $D_{KL} = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}\left[log\ \frac{p_\theta(z,x)}{p_\theta(x)}\right]$
5. $D_{KL} = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x)]$

ELBO

1. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z, x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x)]$

ELBO

1. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x)]$
2. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$

ELBO

1. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x)]$
2. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
3. It is the marginal log likelihood or the log evidence

ELBO

1. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x)]$
2. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
3. It is the marginal log likelihood or the log evidence
4. We can't compute because we don't have its analytical form

ELBO

1. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$

ELBO

1. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
2. $log\, p_\theta(x) = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + D_{KL}(q_\phi||p_\theta)$

ELBO

1. $D_{KL}(q_\phi || p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
2. $log\, p_\theta(x) = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + D_{KL}(q_\phi || p_\theta)$
3. Here, we know that $D_{KL} \geq 0$

ELBO

1. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
2. $log\, p_\theta(x) = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + D_{KL}(q_\phi||p_\theta)$
3. Here, we know that $D_{KL} \geq 0$
4. $log\, p_\theta(x) \geq -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$

ELBO

1. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
2. $log\, p_\theta(x) = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + D_{KL}(q_\phi||p_\theta)$
3. Here, we know that $D_{KL} \geq 0$
4. $log\, p_\theta(x) \geq -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$
5. This is the lower bound on the evidence

ELBO

1. $D_{KL}(q_\phi||p_\theta) = \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] - \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + log\, p_\theta(x)$
2. $log\, p_\theta(x) = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)] + D_{KL}(q_\phi||p_\theta)$
3. Here, we know that $D_{KL} \geq 0$
4. $log\, p_\theta(x) \geq -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$
5. This is the lower bound on the evidence
6. Now, in order to reduce the D_{KL}, we can maximize the ELBO

ELBO

1. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$

ELBO

1. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$
2. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$

ELBO

1. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z, x)]$
2. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
3. ELBO $= \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$

ELBO

1. $\text{ELBO} = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$
2. $\text{ELBO} = -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
3. $\text{ELBO} = \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
4. $\text{ELBO} = \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}\left[log\, \frac{q_\phi(z/x)}{p_\theta(z)}\right]$

ELBO

1. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$
2. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
3. ELBO $= \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
4. ELBO $= \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}\left[log\, \frac{q_\phi(z/x)}{p_\theta(z)}\right]$
5. These represent the reconstruction and KLD (approx. posterior, the prior)

ELBO

1. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z,x)]$
2. ELBO $= -\mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
3. ELBO $= \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}[log\, q_\phi(z/x)] + \mathbb{E}_{q_\phi}[log\, p_\theta(z)]$
4. ELBO $= \mathbb{E}_{q_\phi}[log\, p_\theta(x/z)] - \mathbb{E}_{q_\phi}\left[log\, \frac{q_\phi(z/x)}{p_\theta(z)}\right]$
5. These represent the reconstruction and KLD (approx. posterior, the prior)
6. VAEs model $p_\theta(x/z)$ and $q_\phi(z/x)$ as neural networks