DUL: Variational Autoencoders

Naeemullah Khan

naeemullah.khan@kaust.edu.sa

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

KAUST Academy
King Abdullah University of Science and Technology

May 27, 2025

Table of Contents

Variational Autoencoders: Motivation

Why VAEs?

- Traditional autoencoders are deterministic and lack generative capabilities.
- Need for models that can generate new, diverse data samples.

Applications

- Image generation (e.g., generating new faces or handwritten digits).
- Data compression and denoising.
- Anomaly detection in various domains.

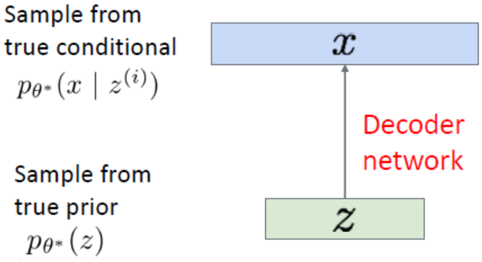
Variational Autoencoders: AE Recap

Autoencoders:

- **Architecture**:
  - **Encoder**: Compresses input x into a latent representation z.
  - **Decoder**: Reconstructs the input x from the latent representation z.
- **Objective**: Minimize reconstruction error $\|x - \hat{x}\|^2$.
- **Limitation**: Cannot generate new data; lacks a probabilistic foundation.

- Autoencoders do not work as generative models because the latent space Z is too discrete.
- **Solution**: Let us rectify that.
- **Variational Autoencoders**: Probabilistic spin on autoencders will let us sample from the model to generate data.

- We want a generative model, which given a prior **z** outputs a new sample from data.
- To make the latent space Z continuous, let's choose it to be Gaussian
- So now, we want to estimate the true parameters θ^* of this generative model.



Variational Autoencoders: Introduction (cont.)

- **Probabilistic Encoder:** Instead of encoding an input x to a fixed point z, the encoder learns a distribution over the latent variable z conditioned on the input:

  $$q(z \mid x)$$

  Typically, this is a multivariate Gaussian with parameters $\mu(x)$ and $\sigma(x)$ learned by a neural network.

- **Probabilistic Decoder:** Given a sample z from the latent distribution, the decoder reconstructs the input by modeling:

  $$p(x \mid z)$$

  This allows for generating new data by sampling from the latent space.

- **Latent Space:** The model learns a continuous latent space where similar data points are close together. This latent space captures the underlying structure or factors of variation in the data.

Objectives and Training: The training objective of a VAE is to maximize the **Evidence Lower Bound (ELBO)**:

$$\log p(x) \geq \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{\mathrm{KL}}(q(z|x) \| p(z))$$

- The first term, $\mathbb{E}_{q(z|x)}[\log p(x|z)]$, encourages accurate reconstruction.
- The second term, $D_{\mathrm{KL}}(q(z|x) \| p(z))$, regularizes the latent space by making the approximate posterior $q(z|x)$ close to the prior $p(z)$, usually $\mathcal{N}(0, I)$.

Why Use VAEs?:

- Enable generative modeling: generate new samples by sampling from the latent distribution.
- Learn smooth, structured, and interpretable latent representations.
- Provide a principled probabilistic framework for inference and generation.

Latent Variable Models:

- **Definition**: Models that assume the data is generated from some unobserved (latent) variables.
- **Goal**: Learn a mapping from observed data x to latent variables z and vice versa.
- **Generative Process**:
    - Sample latent variable z from a prior distribution $p(z)$.
    - Generate data x from the latent variable using a likelihood function $p(x|z)$.
- **Inference Problem**: Given observed data x, infer the posterior distribution $p(z|x)$.
- **Challenge**: Direct computation of posterior $p(z|x)$ is often intractable, leading to the need for approximations.

VAE: Variational Inference

- Now, how to train this model?
- How about following the same strategy as in FVSBNs? Learn model parameters to maximize the likelihood of training data.

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z) dz$$

- But there is a problem here. It is Intractible to compute $p(x|z)$ for every z!
- Intuitively, need to figure out which z corresponds to each x in the dataset, but such mapping is unknown.
- This also makes posterior density $p(z|x)$ intractable because it depends on $p_\theta(x)$

$$p(z|x) = \frac{p_\theta(x|z) p_\theta(z)}{p_\theta(x)}$$

Solution

- Let's approximate $p(z|x)$ with another distribution by another distribution $q(z)$
- If $q(z)$ is a tractable distribution e.g. Gaussian distribution
- **Approach**: We can adjust parameters of $q(z)$ and make it as close to $p(z|x)$, i.e. $q(z) \approx p(z|x)$.
- **Goal**: Minimize the Kullback-Leibler (KL) divergence $KL(q||p)$.

VAE: Variational Inference (cont.)

$$
\begin{aligned}
KL(q(z)||p(z|x) &= -\sum q(z) log \frac{p(z|x)}{q(z)} \\
&= -\sum q(z) log \frac{\frac{p(x,z)}{p(x)}}{q(z)} \\
&= -\sum q(z) log \left(\frac{p(x,z)}{q(z)} \frac{1}{p(x)} \right) \\
&= -\sum q(z) \left[log \frac{p(x,z)}{q(z)} + log \frac{1}{p(x)} \right] \\
&= -\sum q(z) \left[log \frac{p(x,z)}{q(z)} - log\, p(x) \right] \\
&= -\sum_z q(z) log \frac{p(x,z)}{q(z)} + log\, p(x) \sum_z q(z) \\
&= -\sum_z q(z) log \frac{p(x,z)}{q(z)} + log\, p(x) \quad \because \sum_z q(z) = 1
\end{aligned}
$$

VAE: Variational Inference (cont.)

- $$KL(q(z)||p(z|x) = -\sum_z q(z) log \frac{p(x,z)}{q(z)} + log\ p(x)$$

- We can also write above equation as:

$$log\ p(x) = KL(q(z)||p(z|x) + \sum_z q(z) log \frac{p(x,z)}{q(z)}$$

- Given x , $log\, p(x)$ is a constant
- $KL(q(z)||p(z|x))$ is the quantity we wanted to minimize
- Assume $L = \sum_z q(z) log \frac{p(x,z)}{q(z)}$, then

$$\text{constant} = KL + L$$

$$L \leq log\, p(x) \quad \because kl \geq 0$$

- **Instead of minimizing KL we can maximise L**

What is ELBO?:

- Allows us to optimize the model.
- It provides a lower bound on the log-likelihood of the data, which we aim to maximize during training.
- The ELBO consists of two main components:
  - **Reconstruction term**: Measures how well the model can reconstruct the input data from the latent representation.
  - **Regularization term**: Encourages the learned latent distribution to be close to a prior distribution (usually Gaussian).

- Mathematically, the ELBO can be expressed as:

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

where:

  - $q_\phi(z|x)$ is the approximate posterior distribution (encoder).
  - $p_\theta(x|z)$ is the likelihood of the data given the latent variable (decoder).
  - D_{KL} is the Kullback-Leibler divergence between the approximate posterior and the prior distribution $p(z)$.

- The goal is to maximize the ELBO with respect to the model parameters θ and ϕ.
- By maximizing the ELBO, we ensure that the model learns a meaningful latent representation while also being able to generate new samples.

VAE: Evidence Lower Bound (ELBO) (cont.)

Looking at Lower bound L

$$
\begin{aligned}
L &= \sum_z q(z) log \frac{p(x,z)}{q(z)} \\
&= \sum_z q(z) log \frac{p(x|z)p(z)}{q(z)} \\
&= \sum_z q(z) \left[log\ p(x|z) + log \frac{p(z)}{q(z)} \right] \\
&= \underbrace{\sum_z q(z) log\ p(x|z)}_{\text{Expectation } E_{q(z)}(log\ p(x|z))} + \underbrace{\sum_z q(z) log \frac{p(z)}{q(z)}}_{-KL(q(z)||p(z)}
\end{aligned}
$$

So,

$$
L = E_{q(z)}(log\ p(x|z)) - KL(q(z)||p(z))
$$

- $E_{q(z)}(log\ p(x|z))$ is conceptually reconstruction
- We can assume z to be Standard Normal Distribution

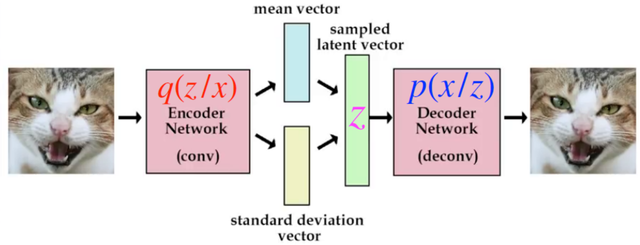
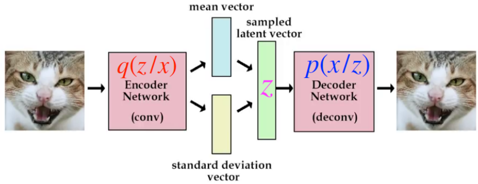


Figure 2: Variational Autoencoder Architecture

VAE: Evidence Lower Bound (ELBO) (cont.)



$$p(x|\hat{x}) = e^{-|x-\hat{x}|^2}$$

$$log\ e^{-|x-\hat{x}|^2} = -|x-\hat{x}|^2$$

$$L = E_{q(z)}(-|x-\hat{x}|^2) - KL(q(z)||p(z))$$

$$\min |x-\hat{x}| + KL(q(z|x)||\mathcal{N}(0,1))$$

$$\min |x-\hat{x}| - 0.5 * (1 + log\,\sigma^2 - \sigma^2 - \mu^2)$$

For full derivation of KL Loss, read here

VAE: Reparameterization Trick

Problem: Cannot backpropagate through stochastic sampling.

Solution: Reparameterize z as:

$$z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Benefit: Enables gradient-based optimization by making the sampling operation differentiable.

VAE: Reparameterization Trick (cont.)

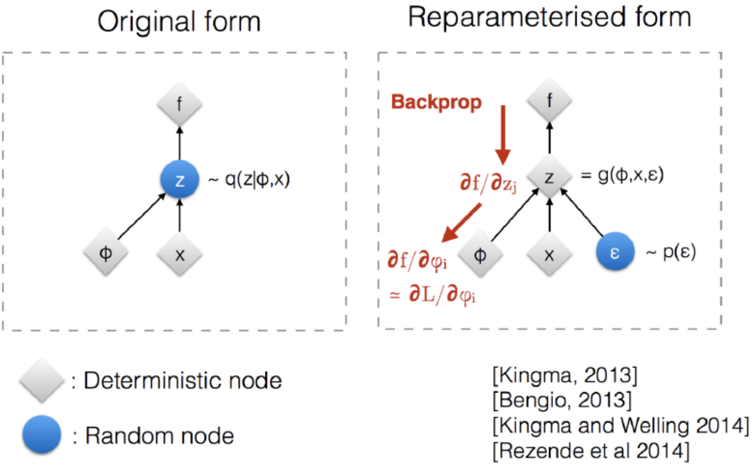


Figure 3: Reparameterization trick to make back propagation possible

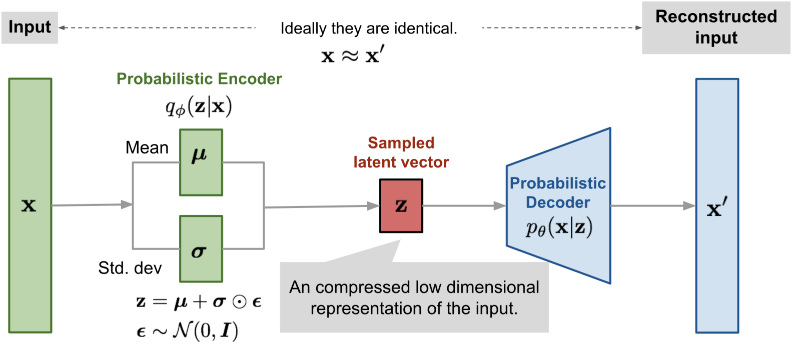


Figure 4: Variational Autoencoder with reparameterization trick

VAE: Loss Function Breakdown

Total Loss:

$$L = \text{Reconstruction Loss} + \text{KL Divergence}$$

Reconstruction Loss:

- Measures how well $\hat{x}$ matches x.
- Common choices: Mean Squared Error (MSE) or Binary Cross-Entropy.

KL Divergence:

- Measures how much $q(z \mid x)$ diverges from the prior $p(z)$.
- Encourages the latent space to follow a standard normal distribution.

VAE: Results

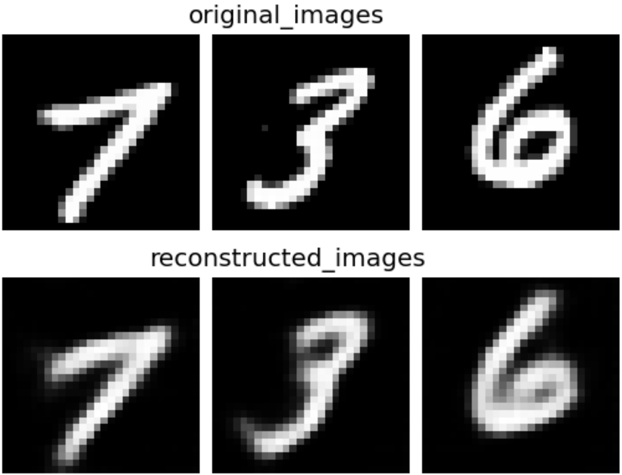


Figure 5: Image reconstruction with variational autoencoders on MNIST digits dataset

VAE: Results (cont.)

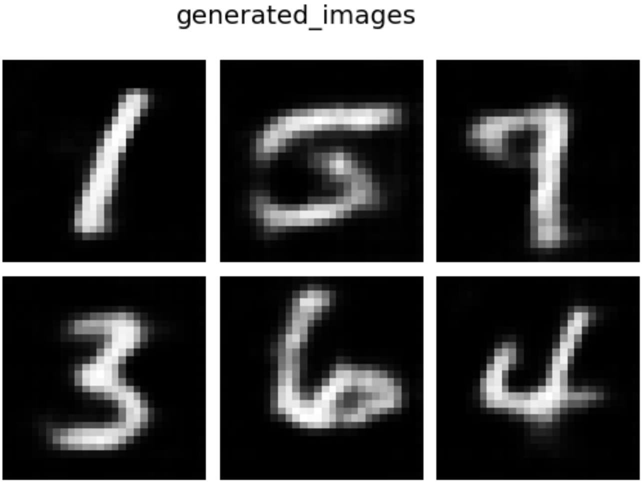


Figure 6: Image generation with variational autoencoders on MNIST digits dataset. Sample an encoding vector from $\mathcal{N}(0, 1)$ and passed it through decoder

VAE: Results (cont.)

Figure 7: Image generation with variational autoencoders on CIFAR-10 32x32 dataset

VAE: Variants and Extensions

β-VAE:

- Introduces a hyperparameter β to control the trade-off between reconstruction and regularization.
- Encourages disentangled representations.

Conditional VAE (CVAE):

- Incorporates additional information y (e.g., class labels) into the encoder and decoder.
- Enables generation of data conditioned on y.

Discrete VAE:

- Utilizes discrete latent variables.
- Techniques like Gumbel-Softmax are used for differentiable sampling.

Disentangled Variational Autoencoders (β-VAEs)

- **Basic Idea**: Different neurons in latent space should be uncorrelated, i.e. they all try to learn something different about input data.
- **Implementation**:

$$\mathcal{L}(\theta, \phi; x, z, \beta) = E_{q_\phi(z|x)}(log\ p_\theta(x|z)) - \beta\ KL(q_\phi(z|x)||p(z))$$

- Increasing the β is forcing variational autoencoder to encode the information in only few latent variables

Figure 8: Azimuthal rotation in β-VAEs and simple VAEs. β-VAEs produce more disentangled rotation, whereas some other features also change in simple VAEs.

VAE: Limitations and Challenges

Limitations:

- **Gaussian Assumption:** The assumption that the latent variables follow a Gaussian distribution may not hold for all datasets.
- **Over-smoothing:** The model may produce overly smooth reconstructions, losing fine details in the data.
- **Sensitivity to Hyperparameters:** The performance of VAEs can be sensitive to the choice of hyperparameters, such as the weight of the KL divergence term.
- **Computational Complexity:** Training VAEs can be computationally expensive, especially for large datasets or complex models.
- **Evaluation Metrics:** Evaluating the quality of generated samples can be subjective and challenging, as traditional metrics may not capture the nuances of the data.

Challenges:

- **Training Instability:** VAEs can be difficult to train, especially with complex datasets.
- **Mode Collapse:** The model may generate samples from only a subset of the latent space.
- **Balancing Reconstruction and Regularization:** Finding the right balance between reconstruction loss and KL divergence can be tricky.
- **Posterior Collapse:** In some cases, the model may ignore the latent variables, leading to poor representations.
- **Limited Expressiveness:** The Gaussian assumption for the latent space may not capture complex data distributions.
- **Disentanglement:** Achieving disentangled representations can be challenging, especially in high-dimensional spaces.

VAE: Limitations and Challenges (cont.)

Future Directions:

- **Improved Training Techniques:** Developing better optimization methods to stabilize training.
- **Advanced Architectures:** Exploring more complex latent variable models, such as Normalizing Flows or Hierarchical VAEs.
- **Better Regularization Techniques:** Investigating alternative regularization methods to improve disentanglement and representation quality.
- **Hybrid Models:** Combining VAEs with other generative models (e.g., GANs) to leverage their strengths.
- **Application-Specific Variants:** Tailoring VAEs for specific applications, such as text or video generation.

VAE: Summary

- Add a probabilistic spin to Autoencoders to make them generative models
- Assume Z to be from Gaussian Distribution.
- But $p(z|x)$ is intractable.
- **Solution**: Approximate $p(z|x)$ with Gaussian distribution $q(z)$.
- To minimize the KL divergence between them maximize the Evidence lower bound

$$L = \sum_{z} q(z) log \frac{p(x,z)}{q(z))}$$

- But image produced by variational autoencoders are blurry.

Reference Slides

- Fei-Fei Li "Generative Deep Learning" CS231
- Hao Dong "Deep Generative Models"
- Hung-Yi Lee "Machine Learning"
- Murtaza Taj "Deep Learning" CS437
- Aykut Erdem, COMP547: Deep Unsupervised Learning, Koc University

Credits

Dr. Prashant Aparajeya

Computer Vision Scientist — Director(AISimply Ltd)

p.aparajeya@aisimply.uk

This project benefited from external collaboration, and we acknowledge their contribution with gratitude.