Variational Autoencoders: A Detailed Overview

1. Introduction

A **Variational Autoencoder (VAE)** is a type of generative model that combines ideas from variational inference and deep learning. It models a data distribution $P(X)$ using latent variables Z, and uses a neural network to approximate the posterior distribution $P(Z \mid X)$.

2. Generative Story

- Sample latent variable: $Z \sim p(z) = \mathcal{N}(0, I)$
- Generate data: $X \sim p_\theta(x \mid z)$

The marginal likelihood is:

$$P(X) = \int p_\theta(x \mid z) p(z) dz$$

which is intractable for complex likelihoods p_θ, hence the need for variational inference.

3. Variational Inference

Introduce an approximate posterior $q_\phi(z \mid x)$, and use the Evidence Lower Bound (ELBO):

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \mathrm{KL}(q_\phi(z \mid x) \| p(z))$$

Objective: Maximize the ELBO with respect to parameters θ and ϕ.

4. Encoder and Decoder Networks

- **Encoder (inference model):** $q_\phi(z \mid x) = \mathcal{N}(z \mid \mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x)))$
- **Decoder (generative model):** $p_\theta(x \mid z)$ is modeled by a neural network

5. Reparameterization Trick

To allow backpropagation through stochastic sampling, reparameterize:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

6. VAE Loss Function

The VAE is trained to minimize the negative ELBO:

$$\mathcal{L}(x) = \underbrace{-\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)]}_{\text{Reconstruction loss}} + \underbrace{\mathrm{KL}(q_\phi(z \mid x) \| p(z))}_{\text{Regularization}}$$

7. Training Procedure

1. For each data point x, compute $\mu_\phi(x)$, $\sigma_\phi(x)$
2. Sample z using the reparameterization trick
3. Compute reconstruction loss and KL divergence
4. Backpropagate and update θ, ϕ

8. Extensions and Variants

- **Beta-VAE:** Scales the KL term with a factor β to encourage disentanglement
- **Conditional VAE:** Models $p(x \mid z, y)$ for supervised generation
- **VQ-VAE:** Uses discrete latent variables via vector quantization

9. Summary

Variational Autoencoders are powerful generative models combining neural networks with probabilistic reasoning. They enable unsupervised learning of latent representations and generation of realistic data.