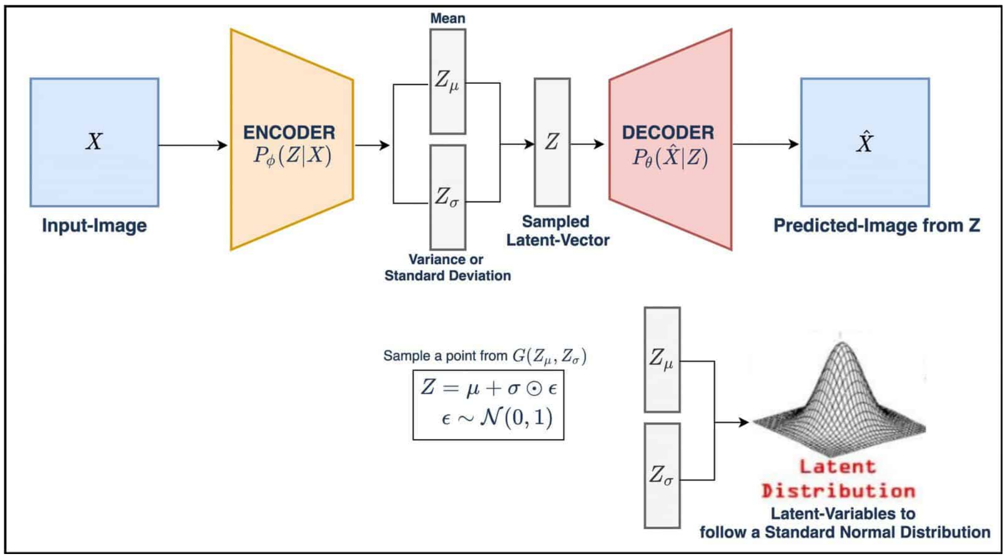
Mathematical Formulation and code explanation of Variational Autoencoders

May 20, 2025

Architecture



A Variational Autoencoder consists of:

- An encoder network $q_\phi(z|x)$ that maps input x to a latent distribution.
- A decoder network $p_\theta(x|z)$ that reconstructs data from latent variable z.

Latent variables z are sampled from a prior $p(z)$, typically a standard normal distribution:

$$p(z) = \mathcal{N}(0, I)$$

The encoder produces parameters $\mu(x)$ and $\sigma^2(x)$ to approximate a Gaussian posterior distribution over latent variable z given input x:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x)))$$

where $\mu(x)$ is the mean and $\sigma^2(x)$ is the variance of the latent variable, ϕ are the parameters of the encoder network, and the covariance matrix is diagonal.

Objective Function

The variational lower bound (ELBO) on the log-likelihood $\log p_\theta(x)$ is:

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

- Reconstruction loss: $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$
- Regularization: $\mathrm{KL}(q_\phi(z|x)\|p(z))$

KL Divergence Term

For Gaussian posterior and standard normal prior:

$$\mathrm{KL}(q_\phi(z|x)\|p(z)) = \frac{1}{2}\sum_{i=1}^{d}\left(\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1\right)$$

Reparameterization Trick

To backpropagate through the sampling process:

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

This allows rewriting the expectation:

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \approx \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)})$$

with $z^{(l)} = \mu + \sigma \odot \epsilon^{(l)}$

Final Loss Function

The total loss to minimize (negative ELBO):

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