Variational Autoencoders on MNIST

ELBO Derivation, KL Annealing, and CVAE Extension (v2)

Alexis Le Trung, Khatir Youyou, Tidjani Adam Kandine,
Yahya Ahachim, Aniss Outaleb, Quentin Wurtlin

Generative AI and Diffusion Models

January 20, 2026

Project Objectives

1. **Derive the ELBO** from first principles with closed-form KL for Gaussians
2. **Implement a Convolutional VAE** on MNIST with 2D latent space
3. **Track reconstruction vs KL** separately, implement KL annealing
4. **Visualize**: latent space, latent traversals, interpolations
5. **Extend to CVAE** for controlled digit generation

ELBO Derivation

Starting from intractable marginal likelihood, we derived:

Evidence Lower Bound

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

Closed-form KL for Gaussian encoder $q = \mathcal{N}(\mu, \sigma^2)$ and prior $p = \mathcal{N}(0, I)$:

$$\mathrm{KL}(q \| p) = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1 \right)$$

ELBO — Proof of Lower Bound

Start from the marginal log-likelihood:

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

Introduce an arbitrary approximate posterior $q_\phi(z|x)$ and rewrite:

$$\begin{aligned}\log p_\theta(x) &= \log \int q_\phi(z|x)\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\,dz \\ &= \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right]\end{aligned}$$

Apply Jensen's inequality (log is concave):

$$\begin{aligned}\log p_\theta(x) &\geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}\right] \\ &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(z)}{q_\phi(z|x)}\right] \\ &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x)\|p(z))\end{aligned}$$

This establishes $\log p_\theta(x) \geq \mathcal{L}(\theta,\phi;x)$.

ELBO — Gradients / Derivative

We optimize parameters θ (decoder) and ϕ (encoder) by maximizing the ELBO. The ELBO for a single datum is:

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

Gradient w.r.t. decoder parameters θ (decoder appears only in $p_\theta(x|z)$):

$$\begin{aligned} \nabla_\theta \mathcal{L} &= \nabla_\theta \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\nabla_\theta \log p_\theta(x|z)\right]. \end{aligned}$$

In practice estimate by Monte Carlo: sample $z \sim q_\phi(z|x)$ via reparameterization $z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon$. Gradient w.r.t. encoder parameters ϕ uses reparameterization to push gradient through samples:

$$\nabla_\phi \mathcal{L} = \nabla_\phi \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}\left[\log p_\theta(x|\mu_\phi + \sigma_\phi \odot \epsilon)\right] - \nabla_\phi \mathrm{KL}(q_\phi \| p)$$

where the KL term has closed-form gradients: $\partial_{\mu_j} \mathrm{KL} = \mu_j$, $\partial_{\sigma_j} \mathrm{KL} = \sigma_j - 1/\sigma_j$.

Architecture & Training Setup

Convolutional VAE:

- Encoder: Conv layers $\rightarrow$ 2D latent $(\mu, \log \sigma^2)$
- Decoder: Linear $\rightarrow$ ConvTranspose layers
- Reparameterization: $z = \mu + \sigma \odot \epsilon$

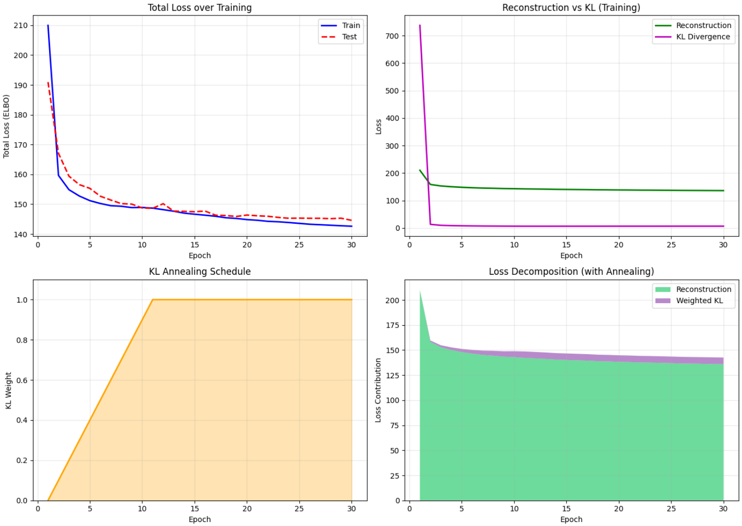
KL Annealing:

$$\beta(t) = \min\left(1, \frac{t}{10}\right)$$

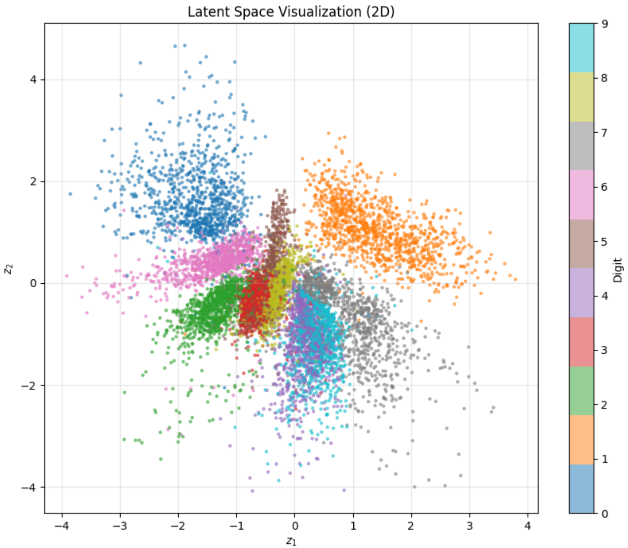
Prevents posterior collapse

Parameter	**Value**
Latent dim	2
Batch size	128
Learning rate	10^{-3}
Optimizer	Adam
Epochs	30
KL warmup	10 epochs

VAE Training Curves



Latent Space Visualization

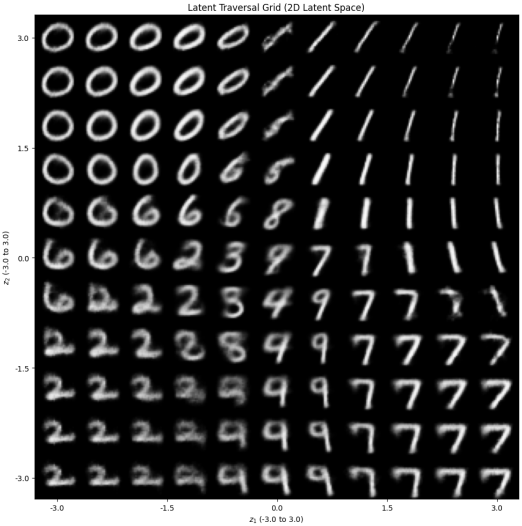


Observations:

- Clear clustering by digit class
- Similar digits nearby (4/9, 3/8, 1/7)
- Smooth, continuous manifold
- Matches $\mathcal{N}(0, I)$ prior

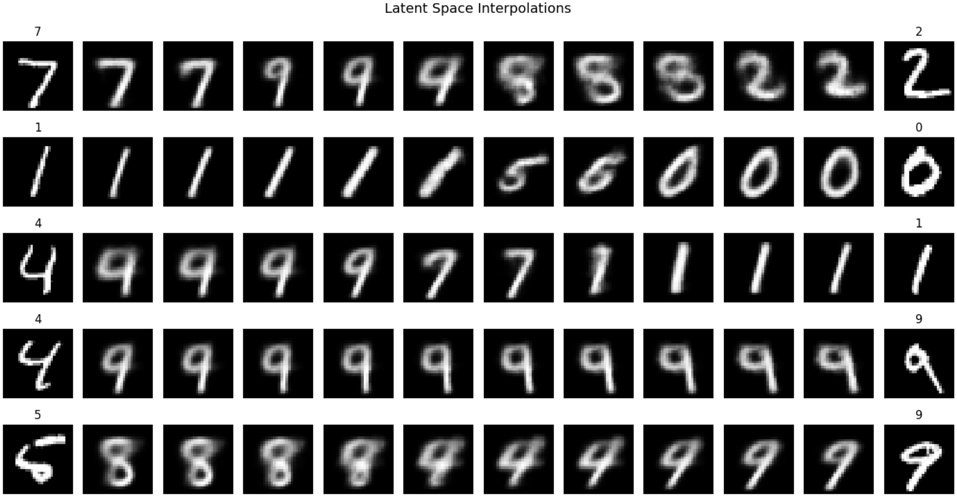
Key insight:
Organization emerges from reconstruction alone — no labels used!

Latent Traversal Grid



Traversing z_1 vs z_2: z_1 (horizontal) controls slant/rotation; z_2 (vertical) controls thickness/scale.

Latent Interpolations



Linear interpolation between encoded digit pairs. Smooth transitions validate the continuous, well-structured latent space.

CVAE Extension: Conditional Generation

Modification: Condition encoder and decoder on class label c

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

Implementation:

- Embedding the label to $\times 28 \times 28$
- Concatenate with image as extra channels
- Class-independent prior: $p(z) = \mathcal{N}(0, I)$

Result: Latent space encodes *style*, label provides *class identity*

CVAE ELBO

Condition on label c; objective per datum (x, c):

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

Proof of lower bound follows identical steps with all densities conditioned on c:

$$\begin{aligned} \log p_\theta(x|c) &= \log \int p_\theta(x|z, c)p(z)\, dz \\ &= \log \mathbb{E}_{q_\phi(z|x,c)} \left[\frac{p_\theta(x|z, c)p(z)}{q_\phi(z|x, c)} \right] \\ &\geq \mathbb{E}_{q_\phi(z|x,c)} \left[\log \frac{p_\theta(x|z, c)p(z)}{q_\phi(z|x, c)} \right] = \mathcal{L}_{\text{CVAE}}. \end{aligned}$$

Gradients: similar decomposition as VAE; decoder gradient

$$\nabla_\theta \mathcal{L}_{\text{CVAE}} = \mathbb{E}_{q_\phi(z|x,c)} \left[\nabla_\theta \log p_\theta(x|z, c) \right]$$

Encoder gradients use reparameterization with the conditional encoder $\mu_\phi(x, c), \sigma_\phi(x, c)$ and the closed-form KL.

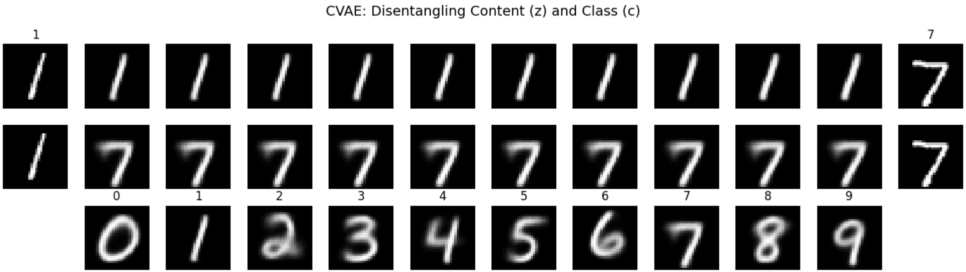
VAE vs CVAE: Quantitative Comparison

Model	Recon. Loss	KL	Total Loss
VAE	138.00	6.59	144.59
CVAE	124.36	4.80	129.16
Improvement	**9.9%**	**27.2%**	**10.7%**

Why CVAE performs better:

- Decoder doesn't need to infer class from $z \Rightarrow$ simpler task
- Latent space focuses purely on style variations
- More compact representation (lower KL)

CVAE: Disentanglement Demonstration



Row 1-2: Same z interpolation with different class labels (1 vs 7). **Row 3:** Fixed z, varying class 0-9 $\Rightarrow$ style transfer across all digits.

Key Findings

1. **ELBO provides principled training objective**
   - Reconstruction + regularization trade-off
   - Closed-form KL enables efficient optimization
2. **KL annealing is essential**
   - Without it: posterior collapse, KL $\rightarrow 0$
   - 10-epoch warmup gave balanced training
3. **VAE learns meaningful structure unsupervised**
   - Class clustering, smooth interpolations
4. **CVAE achieves explicit disentanglement**
   - 9.9% better reconstruction, controlled generation

Thank You!

Questions?

Code available in accompanying Jupyter notebook