Recreating Variational Autoencoders

Sahith Vavilala
University of California, Irvine
Irvine, CA 92617
svavilal@uci.edu

Sanjita Venkatesh Nayak
University of California, Irvine
Irvine, CA 92617
nayaksv@uci.edu

Gnana Heemmanshuu Dasari
University of California, Irvine
Irvine, CA 92617
ghdasari@uci.edu

Abstract

In this project, we re-implement the Variational Autoencoder (VAE) framework that was introduced by Kingma and Welling and evaluate its generative behavior on the MNIST handwritten digit dataset and the Frey faces dataset. We conducted data exploration, designed a multi-layer perceptron VAE, and trained models with varying latent dimensions to study how architectural choices influence reconstruction fidelity and sample quality. Our experiments show that lower-dimensional latent spaces restrict diversity, while higher-dimensional spaces increase variability but may introduce blur due to a mismatch between the prior and learned posterior. Overall, our results confirm the core behaviors reported in Kingma & Welling's work and demonstrate that VAEs learn meaningful latent representations while enabling efficient generative modeling.

1 Introduction

Deep generative models aim to learn probability distributions over high-dimensional data such as images, audio, and text. Variational Autoencoders (VAEs), introduced by Kingma and Welling [1], are a core approach to latent-variable generative modeling. They pair neural networks with variational inference, making the otherwise intractable marginal likelihood optimizable. VAEs allow sampling from learned distributions and produce latent spaces that support interpolation and clustering.

In practice, VAE performance depends strongly on design choices such as latent dimensionality, decoder likelihood, and architecture. Details like pixel-wise Gaussian assumptions or the KL term's weight can meaningfully affect reconstruction quality and sample sharpness. Reproducing reported VAE results therefore requires careful attention to these choices.

Our project re-implements a basic MLP-based VAE and trains it on MNIST and the Frey faces dataset to replicate results from the original paper and study how architectural decisions influence sample quality. During this process, we found that some of the visual results in the paper such as sharp Frey face manifolds are hard to match without closely reproducing preprocessing and decoder parameterizations. Increasing latent dimensionality reduced reconstruction error but often increased blur due to mode-averaging behavior. These observations highlight the gap between the clean theory and the practical realities of implementing VAEs. This report documents our findings from an implementation-first perspective.

Preprint. Under review.

2 Related Work

Generative models come in several forms, but for this project we focus on latent-variable approaches. Classical latent models [3] assume that observations are generated from low-dimensional hidden factors; VAEs extend this idea using neural networks to parameterize both the prior and the likelihood.

Because the true posterior $p(z \mid x)$ is intractable, VAEs use variational inference with an approximate posterior $q_\phi(z \mid x)$. The ELBO objective and the reparameterization trick make it possible to train the model with standard gradient methods.

GANs [4] offer a contrasting approach that often produces sharper samples but lacks an explicit likelihood and can be unstable to train. In comparison, VAEs provide a stable probabilistic framework and interpretable latent spaces.

In this project, we re-implement the basic VAE and study its empirical behavior to better understand how architectural choices affect reconstruction quality and generated samples.

3 Method

3.1 Datasets and Preprocessing

We use MNIST (70k 28×28 grayscale images) and the Frey Faces dataset (1,965 28×20 grayscale frames). For both datasets, we normalize pixel values to the [0,1] range and keep the original resolutions to match the setup in the VAE paper.

3.2 Model Architecture

We implement a basic MLP VAE with symmetrical encoder and decoder networks.

Encoder. The encoder is a two-layer MLP with `tanh` activations. The final hidden layer feeds into two linear heads that produce the mean and log-variance of $q_\phi(z \mid x)$. The latent dimensionality is varied across experiments (2–20).

Reparameterization. We sample latent variables using

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

Decoder. The decoder mirrors the encoder and ends with a sigmoid layer. For MNIST, we use a Bernoulli likelihood. For Frey Faces, we also test a Gaussian decoder and observe increased blur when learning pixel variances.

3.3 Training Objective

We optimize the standard ELBO:

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

The reconstruction term uses binary cross-entropy for the Bernoulli decoder.

4 Experiments and Results

4.1 Training Setup

All models were trained with:

- Adagrad optimizer
- learning rate 1e-3
- Batch size 128 for MNIST and 64 for Frey Faces
- 30–50 epochs for MNIST and 300–500 for Frey Faces

We trained separate models for each latent dimensionality and compared both reconstruction error and generative sample quality.

Figure 1: MNIST digit reconstruction. The first row shows the input images provided to the VAE, and the second provides the reconstruction.

(a) Input Frey images (b) Reconstructed images (c) After adding sigmoid

Figure 2: Adding sigmoid activation to the decoder fixes the blur in reconstructed Frey images

4.2 Results

4.2.1 Reconstruction Quality

MNIST reconstructions As shown in 1, the model produced sharp, digit-faithful reconstructions that matched typical results from the literature. Reconstruction quality was stable across seeds and hyperparameters.

Frey reconstructions Initial reconstructions for Frey Faces were noticeably blurry as seen in 2b and often collapsed to a nearly identical "average face." This was traced to the decoder output lacking a sigmoid constraint, which produced unstable pixel intensities and poor gradient signals for the Gaussian likelihood. After adding the sigmoid activation 2c, reconstruction sharpness increased substantially.

4.2.2 Random Sampling

Random samples from the MNIST latent space (4a) show good variety. The generated digits change smoothly in shape and style, suggesting that the model learned a reasonable latent structure.

For the Frey dataset (4b), random samples show very little variation. Even when sampling widely in the latent space, most generated faces look almost the same. Some modes of the dataset like the frames where the subject sticks out his tongue-never appear. We observed the same behavior in other public VAE implementations [5,6,7]. The original VAE paper [1] does not show the tongue images either, but its samples do show more variation in expressions than ours.

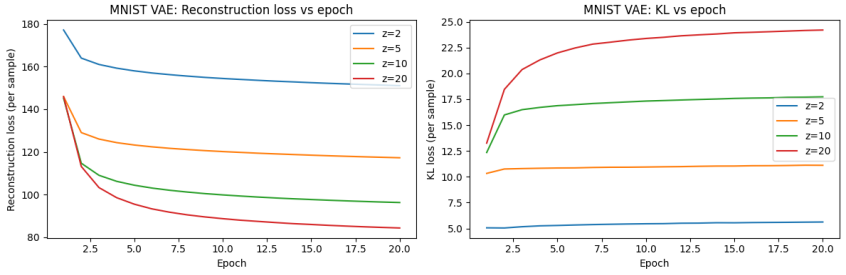
4.2.3 Posterior Collapse

We initially observed posterior collapse, especially on the Frey dataset:

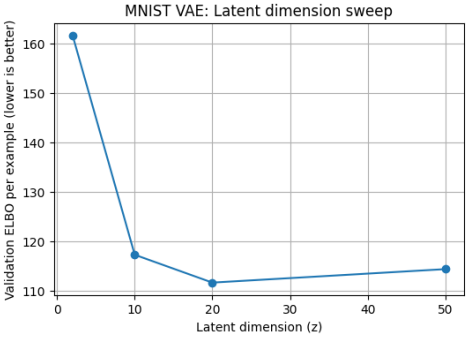
- The encoder produced $\mu \approx 0$ and $\log \sigma^2 \approx 0$ for nearly all samples.
- KL divergence decreased toward zero.
- Reconstructions converged to a single averaged face.

This was resolved through a combination of enforcing the sigmoid constraint on decoder means, reducing model capacity, and lowering the latent dimensionality.

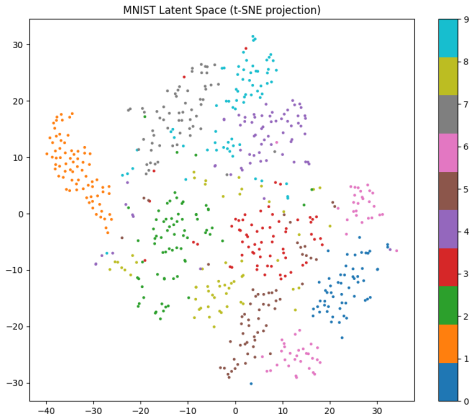
After these adjustments, KL terms stabilized at non-zero values and reconstructions recovered variety.



(a) MNIST loss with varying latent dimensions



(b) Latent dimension sweep



(c) MNIST latent space

Figure 3

4.2.4 Loss Curves

The loss curves in Figure 3a indicate stable training dynamics the MNIST dataset across varying latent dimensions. Across all settings, increases in latent dimensionality lead to consistently lower reconstruction loss but higher KL divergence, reflecting increased information capacity at the cost of weaker prior regularization. This trend persists throughout training and directly explains the improved reconstructions seen at larger latent sizes. MNIST converges faster and to a lower variational bound, reflecting its simpler and more homogeneous structure. In contrast, the Frey dataset showed slower convergence and a higher reconstruction term, consistent with our qualitative findings that its reconstructions are more challenging.

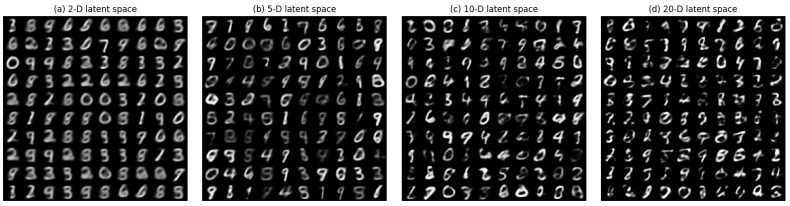
A latent-dimension sweep in 3b reveals that the ELBO is minimized at z = 20 dimensions, with the objective degrading as the dimensionality increases further. Interestingly, although the ELBO is lowest at 20 dimensions, the reconstructions are noticeably blurrier at this size (4a), suggesting that ELBO alone does not fully correlate with perceptual quality. This aligns with observations in related work that VAEs often trade off reconstruction sharpness for better latent structure.

Finally, 3c visualizes the t-SNE projection of the learned MNIST latent space, showing well-separated clusters corresponding to digit classes, indicating that the VAE successfully learns a semantically meaningful manifold.

5 Discussion

Experiments revealed clear differences between MNIST and Frey Faces with respect to VAE behavior:

1. **Dataset Complexity and Size:** Frey Faces is small (1,965 examples) and highly correlated. The limited variation encourages the model to "average" faces during training as seen in 4b, producing outputs that look nearly identical unless the model is carefully regularized.



(a) MNIST digit sampling for [2, 5, 10, 20]-D latent dimensions

(b) Frey images

Figure 4: MNIST digit sampling for increasing dimensions yields lower reconstruction loss but some amount of blurring. Sampled Frey faces seem to converge around the average face.

2. **Mode Coverage Issues:** VAEs, particularly with Gaussian decoders, tend to under-represent small or isolated modes in the data distribution. The Frey images where the subject sticks out his tongue are rare and visually distinct; the decoder fails to represent these modes and instead collapses to the majority distribution.
3. **Sensitivity to Decoder Likelihood:** For continuous data, applying the sigmoid before the Gaussian mean computation was crucial. Without this, the decoder produced unstable pixel values, leading to blurry reconstructions and collapse as in 2b.
4. **Posterior Collapse:** Frey Faces is more vulnerable to posterior collapse because a large latent dimension is unnecessary to represent such a homogeneous dataset. This causes the KL term to vanish, and the decoder to ignore the latent variables entirely. Our adjustments restored meaningful latent usage but did not solve the mode collapse entirely.

6 Conclusion

We successfully re-implemented the Variational Autoencoder of Kingma & Welling (2013) and attempted to reproduce its qualitative behavior on MNIST and Frey Faces. While MNIST produced clean reconstructions, the Frey dataset proved substantially more challenging. Its limited variability, small size, and strong correlation structure contributed to blurry reconstructions, mode averaging, and a tendency toward posterior collapse.

By enforcing the correct sigmoid constraint in the Gaussian decoder and adjusting the latent dimensionality, we alleviated posterior collapse and improved reconstruction quality. However, even with these modifications, the learned Frey manifold remained narrow and failed to capture rare poses such as the tongue-out expressions. This is consistent with results observed in other implementations we found [5].

Overall, our experiments illustrate the strengths of VAEs on well-behaved datasets like MNIST, as well as their limitations when modeling small, highly correlated real-world datasets. We plan to explore and reimplement hierarchical VAEs, normalizing flows, GANs or discrete latent models to better capture multi-modal distributions such as those present in the Frey dataset, and gain a much more hands-on understanding of these models as we have done with VAEs.

References

[1] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114, 2013.

[2] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel Recurrent Neural Networks. International Conference on Machine Learning (ICML), 2016.

[3] R. Salakhutdinov and G. Hinton. Deep Boltzmann Machines. Artificial Intelligence and Statistics (AISTATS), 2009.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In NeurIPS 2014.

[5] AntixK. PyTorch-VAE. GitHub repository, https://github.com/AntixK/PyTorch-VAE, accessed 2025.

[6] W. Goldie. VAE (PyTorch). GitHub repository, https://github.com/wgoldie/VAE, accessed 2025.

[7] NoviceStone. Variational Autoencoder (PyTorch). GitHub repository, https://github.com/NoviceStone/VAE, accessed 2025.