VQ-VAE a posteriori with Geodesic Quantization

August 28, 2025

Alessio Lani

Abstract

This project investigates the challenge of learning robust discrete representations for generative modeling, comparing a standard end-to-end trained Vector-Quantized Variational Autoencoder (VQ-VAE) with an innovative a posteriori quantization method. The proposed approach first trains a continuous VAE to learn a latent manifold, then applies KMedoids clustering using a geodesic distance metric derived from a k-NN graph to form a discrete codebook. Our experiments on the MNIST dataset reveal a critical trade-off between the two methods. While the standard VQ-VAE achieves superior reconstruction quality, it suffers from severe codebook collapse, where only a small fraction of the discrete codes are utilized. This collapse critically impairs the performance of a subsequent autoregressive prior, leading to low-fidelity generated samples. In contrast, our a posteriori geodesic method produces a fully-utilized, geometrically-aware codebook, resulting in visually more coherent generated samples. This work demonstrates that a posteriori quantization, while computationally intensive, offers a more stable and robust alternative to the instability of end-to-end learned discrete representations.

1. Introduction

Learning discrete representations is a cornerstone of modern generative modeling, with the Vector-Quantized Variational Autoencoder (VQ-VAE) being a prominent approach. While effective, the standard end-to-end training of VQ-VAEs can suffer from issues such as codebook collapse, where only a fraction of the discrete codes are utilized, negatively impacting generative performance.

This project investigates an alternative, two-stage pipeline for learning discrete representations, which constitutes the main scientific contribution of this work. The core idea is an **a posteriori quantization** method where a standard continuous VAE is first trained to learn a smooth latent manifold. Subsequently, vector quantization is performed on this pre-trained space using a **non-Euclidean, geodesic distance metric**. This metric is computed via shortest-path distances on a k-Nearest Neighbors graph constructed from latent vector samples. The hypothesis is that this geometrically-aware clustering can produce a more meaningful and robust discrete codebook than the standard, learned approach.

To validate this method, a comprehensive comparison is performed against a standard end-to-end trained VQ-VAE baseline. Both pipelines are evaluated on reconstruction quality, codebook utilization, and the sample fidelity of a subsequent autoregressive model trained on the discrete codes.

2. Related Work

Generative modeling with deep neural networks has explored both continuous and discrete latent variable models.

Continuous Latent Variable Models. Variational Autoencoders (VAEs) (Kingma & Welling, 2013) are a cornerstone of generative modeling, learning a continuous, often Gaussian, latent space. They are trained by maximizing the Evidence Lower Bound (ELBO), which combines a reconstruction term with a Kullback-Leibler divergence regularizer. While powerful, standard VAEs can sometimes suffer from an issue known as ”posterior collapse,” where the latent variables are ignored by the decoder.

Discrete Latent Variable Models. To overcome some limitations of continuous VAEs, the Vector-Quantized Variational Autoencoder (VQ-VAE) was introduced (Van Den Oord et al., 2017). The VQ-VAE learns a discrete codebook of latent representations through a vector quantization step. This approach avoids posterior collapse and has been shown to be effective for generating high-fidelity images, particularly when paired with a powerful autoregressive prior to model the distribution of the discrete codes. However, VQ-VAEs can be difficult to train, often

Email: Alessio Lani <lani.1857003@studenti.uniroma1.it>.

Deep Learning and Applied AI 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

suffering from ”codebook collapse,” a phenomenon where only a small subset of the codebook is utilized, which was a central challenge investigated in this work.

Latent Space Geometry. The geometry of the latent spaces learned by autoencoders is often complex and non-Euclidean. The manifold hypothesis suggests that high-dimensional data, such as images, often lies on a low-dimensional, non-linear manifold. This motivates the use of non-Euclidean distance metrics, such as geodesic distance, which measures the shortest path along the surface of the manifold. Such metrics, often approximated via shortest-path algorithms on neighborhood graphs (Tenenbaum et al., 2000), can provide a more meaningful measure of similarity for clustering and analysis, forming the core hypothesis of our proposed a posteriori quantization method.

3. Method

This work investigates two distinct pipelines for learning discrete representations for image generation. The first is a proposed method based on a posteriori geodesic quantization, and the second is a standard end-to-end VQ-VAE that serves as a baseline.

3.1. Approach A: A Posteriori Geodesic Quantization

This approach is a multi-stage pipeline designed to create a discrete codebook from a pre-trained continuous latent space, respecting its underlying geometry.

1. Continuous Latent Space Learning: First, a standard continuous Variational Autoencoder, termed a `GridVAE`, is trained. The model is composed of a convolutional encoder that maps a 28 × 28 input image to a latent grid of distributions, $\mathcal{Z} \in \mathbb{R}^{H \times W \times D}$, and a corresponding convolutional decoder. The model is trained to minimize the standard Evidence Lower Bound (ELBO), which consists of a reconstruction term and a Kullback-Leibler (KL) divergence regularizer.

2. Geodesic Clustering: After training, the mean latent vectors, μ, are extracted for all images. To create a geometrically-aware codebook, a non-Euclidean clustering algorithm is applied. A key challenge is that computing the full geodesic distance matrix for all 3 million latent vectors is computationally infeasible.

To overcome this, an approximation is employed using a subset of N vectors, called landmarks. A k-Nearest Neighbors (k-NN) graph is constructed on these landmarks, and the all-pairs shortest-path distance is then computed to yield a geodesic distance matrix, D_{geo}. K-Medoids clustering is performed on this matrix to find the K centroids that form the final a posteriori codebook, $\mathcal{C} = \{c_1, \ldots, c_K\}$. The quantization of the full dataset is then achieved by assigning each vector to the cluster of its nearest landmark. This landmark-based approach provides a computationally tractable yet principled approximation of the true geodesic quantization.

3. Autoregressive Prior: To generate new samples, a generative prior is trained over the discrete codes. The 2D grid of continuous vectors for each image is quantized by assigning each vector to the nearest centroid in $\mathcal{C}$. This creates a dataset of discrete code grids. A decoder-only Transformer is then trained on these grids to learn the autoregressive probability distribution $p(z_i|z_{<i})$.

3.2. Approach B: End-to-End VQ-VAE (Baseline)

As a baseline, a standard VQ-VAE is implemented, which learns its discrete codebook end-to-end. The model consists of an encoder, a decoder, and a ‘VectorQuantizer‘ layer containing a learnable codebook. The model is trained to minimize a combination of a reconstruction loss and the VQ loss, which includes a codebook loss and a commitment loss term (Van Den Oord et al., 2017). In this implementation, a moderately high `commitment_cost` is used as the primary mechanism to encourage codebook utilization and mitigate the codebook collapse observed during preliminary experiments. For generation, a separate Transformer is trained on the discrete codes produced by this VQ-VAE.

4. Experimental Results

The proposed a posteriori geodesic quantization method (Approach A) is evaluated against a standard end-to-end VQ-VAE baseline (Approach B) on the MNIST dataset. The comparison is structured to assess both the quality of the learned discrete representations and the final performance of the full generative pipelines.

4.1. Analysis of A Posteriori Clustering

The core hypothesis of this work is that a geodesic distance metric can produce a more meaningful clustering of the VAE’s latent space than a standard Euclidean metric. To test this, a `GridVAE` was first trained, and then a sample of 10,000 of its latent vectors were clustered using both the proposed Geodesic KMedoids and a standard K-Means algorithm.

The qualitative results, visualized using t-SNE in Figure 1, reveal distinct clustering behaviors. Standard K-Means imposes a spherical bias due to its assumption of equal variance and spherical cluster shapes, whereas the geodesic method generates more organic clusters that align with the winding topology of the MNIST data manifold. This alignment with the manifold's intrinsic geometry suggests that the geodesic approach better captures the data's structure.

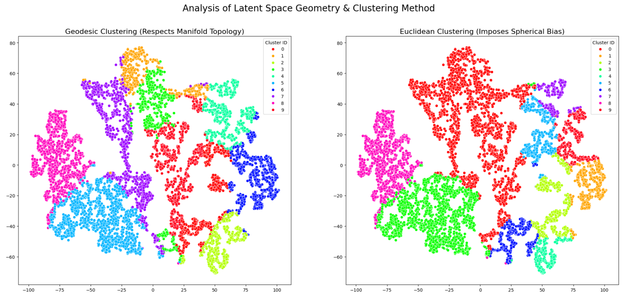

Figure 1. A side-by-side t-SNE visualization of the clustering results on the `GridVAE` latent space. **Left:** The proposed Geodesic KMedoids. **Right:** Standard Euclidean K-Means baseline. The geodesic method produces clusters that better align with the data's manifold structure.

4.2. Comparison of Full Generative Pipelines

Next, the complete generative pipelines for both approaches are compared across three key metrics: codebook utilization, reconstruction quality, and sample fidelity.

Codebook Utilization. A critical aspect of discrete representation learning is the effective use of the codebook. Figure 2 compares the codebook usage for both methods after training. The a posteriori method (Approach A), by design, results in 100% codebook utilization. In contrast, the end-to-end trained VQ-VAE (Approach B) suffers from severe codebook collapse, utilizing only a small fraction of its available codes. This demonstrates a key stability advantage of the a posteriori approach.

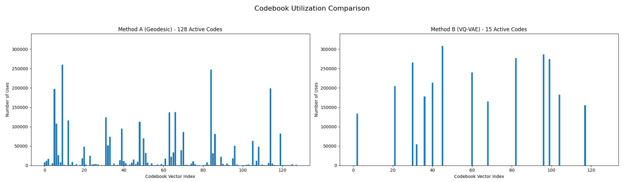

Figure 2. Codebook utilization comparison. **Left (Method A):** All 128 codes are used by design. **Right (Method B):** The baseline suffers from severe codebook collapse, with only 15/128 codes being active.

Reconstruction Quality. The ability of each autoencoder pipeline to reconstruct images from the test set is compared. As shown in Figure 3, the end-to-end VQ-VAE (Method B) produces slightly sharper reconstructions, which is confirmed by its lower Mean Squared Error (MSE). This is an expected outcome, as its architecture is fully optimized for this specific task.

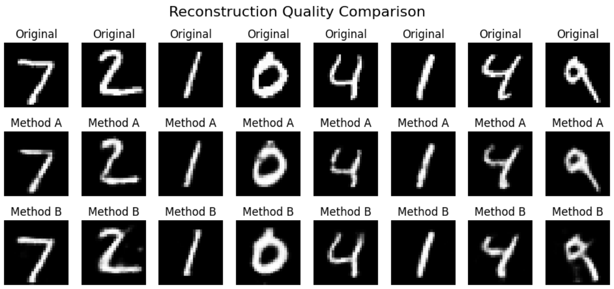

Figure 3. Qualitative comparison of reconstruction quality. From top to bottom: Original images, reconstructions from Method A (Geodesic), and reconstructions from Method B (VQ-VAE).

Sample Fidelity and Perplexity. Finally, the quality of newly generated samples from both full pipelines is evaluated. The results are shown in Figure 4. The proposed method (Approach A) is able to generate varied, recognizable digit-like shapes. The baseline method (Approach B) fails to generate coherent images, producing abstract patterns. This failure in sample fidelity is a direct consequence of the codebook collapse diagnosed previously. A summary of the quantitative metrics is provided in Table 1.

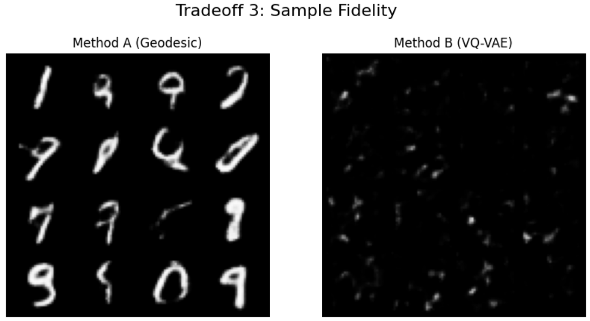

Figure 4. Comparison of generated samples. **Left (Method A):** Recognizable, diverse digits. **Right (Method B):** Failure to generate coherent images due to codebook collapse.

Table 1. Final quantitative comparison of the two pipelines.

Metric	Method A (Geodesic)	Method B (VQ-VAE)
Reconstruction MSE	0.016125	**0.011385**
Codebook Utilization	**100%**	11.7%
Transformer Perplexity	4.9448	**4.2686**

References

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Van Den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in neural information processing systems*, volume 30, 2017.

A. Appendix: Investigation of VQ-VAE Training Dynamics

This appendix details the experimental process and challenges encountered during the training of the end-to-end VQ-VAE baseline (Approach B), which led to the final, stable model presented in the main report.

A key challenge identified in early experiments was a severe case of **codebook collapse**, where a very small fraction of the available discrete latent codes were utilized during training. The initial diagnosis revealed that fewer than 20% of the 128 codebook vectors were active, providing an impoverished latent representation for the subsequent generative model.

To mitigate this issue, two primary strategies were investigated. The first was a hyperparameter search on the `commitment_cost`. A series of experiments were conducted, with this value being progressively increased from the default of 0.25 up to an aggressive value of 2.5. While higher values led to a marginal increase in the number of active codes, the collapse remained severe, indicating that this hyperparameter alone was insufficient to solve the problem.

The second strategy involved implementing a more advanced technique: **periodic resampling of unused codebook vectors**. This method was designed to directly combat codebook collapse by identifying "dead" codes and replacing them with encoder outputs from the current data batch. However, this approach introduced a critical **training instability**. The training would proceed normally for several epochs, but upon activation of the resampling mechanism, the reconstruction loss would catastrophically explode. This instability persisted even when delaying the start of the resampling to later epochs.

Given these findings, the conclusion was reached that the resampling technique, while powerful, was too disruptive for the training dynamics of the model. Therefore, the final, stable VQ-VAE model presented in this work reverted to the simpler architecture, relying solely on a moderately-tuned `commitment_cost` as a regularizer. This approach provided a stable, convergent training process, which was deemed a necessary prerequisite for a fair comparison.