Geodesic Post-Hoc Quantization for Variational Autoencoders

Marco Galletti (2109043) [1]

Abstract

This project revisits the classic VQ-VAE by implementing a post-hoc quantization pipeline. Instead of training end-to-end, a continuous-latent VAE is first trained and then a discrete codebook is built by clustering its latent space. This work compares standard graph-based geodesic clustering with two novel variations based on a decoder-induced Riemannian metric. Under this metric, latent-space quantization error is reduced by three orders of magnitude, but this does not translate to improved image-space reconstruction quality, where simpler Euclidean-based graph geodesics perform best. Furthermore, the success of post-hoc quantization appears to be highly dataset-dependent: high-quality results are achieved on structured data like FashionMNIST, but performance degrades on more complex datasets such as CIFAR-10, revealing what appears to be a fundamental trade-off between post-hoc and end-to-end quantization schemes.

1. Introduction

Generative models, particularly Variational Autoencoders (VAEs), are powerful tools for learning low-dimensional representations of complex data. The Vector Quantized VAE (VQ-VAE) extends this by learning a discrete latent space, which is often more suitable for downstream tasks such as autoregressive modeling. Standard VQ-VAEs learn this discrete representation end-to-end with the encoder and decoder. This work explores an alternative post-hoc quantization pipeline, where a continuous-latent VAE is first trained, and a discrete codebook is then built by applying clustering algorithms to its latent space.

The main contributions are as follows:

- An end-to-end pipeline for post-hoc vector quantization of VAEs.
- The implementation and evaluation of a geodesic clustering method based on a decoder-induced Riemannian metric, compared against a standard Euclidean baseline.
- A detailed analysis showing that while geodesic clustering better respects the latent manifold's structure, simpler Euclidean clustering consistently yields superior image reconstructions.
- The finding that the viability of post-hoc quantization appears to be strongly dependent on dataset complexity, offering insights into the trade-offs between post-hoc and end-to-end learned quantization.

The code for this project is available at `https://github.com/m4rch1n0/vqvae`.

2. Related Work

The project builds upon three main areas of research: discrete latent variable models, the geometric analysis of latent spaces, and graph-based clustering methods.

Discrete Latent Representations. The Vector-Quantized Variational Autoencoder (VQ-VAE) (Van Den Oord et al., 2017) is the foundational model for learning discrete representations for generative tasks. VQ-VAE introduced the paradigm of learning a discrete codebook end-to-end with the encoder and decoder, and using a powerful autoregressive model over the resulting spatial grid of codes to generate high-fidelity images. The research in this project deliberately deviates from this by exploring a post-hoc quantization scheme, which separates the continuous representation learning from the codebook creation.

Geometry of VAE Latent Spaces. The core motivation for the geodesic method comes from works that model the latent space of a VAE as a Riemannian manifold. Arvanitidis et al. (Arvanitidis et al., 2017) established the use of the pull-back metric tensor, defined by the decoder's Jacobian ($M(z) = J(z)^T J(z)$), to better represent the geometry of the data manifold. They argued for using geodesic distances, rather than Euclidean ones, for tasks like interpolation and clustering. Shao et al. (Shao et al., 2018) further developed practical methods for geodesic computation, though they

[1]Sapienza University of Rome Email: Marco <galletti.2109043@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

noted that for some image VAEs, the learned manifolds can be nearly flat, making Euclidean distance a reasonable approximation. This work leverages these geometric insights directly, using Jacobian-vector product to approximate local geodesic distances for clustering.

Graph-Based Geodesic Approximation. The specific technique of approximating geodesic distances by computing shortest paths on a k-Nearest Neighbors (k-NN) graph was popularized by Isomap (Tenenbaum et al., 2000) in the context of manifold learning. This approach provides a computationally feasible way to estimate global manifold distances from local neighborhood information. The methods presented here are a direct application of this principle: a k-NN graph is constructed and shortest-path distances (via Dijkstra's algorithm) are used as the metric for the k-medoids clustering algorithm, with the key novelty being the re-weighting of the graph's edges with the learned Riemannian metric.

3. Method

The pipeline consists of three main stages: continuous representation learning, post-hoc quantization, and autoregressive modeling.

3.1. Model Architectures[1]

Spatial VAE. The VAE encoder consists of three convolutional layers (with 64, 128, and 256 channels respectively, stride 2, and Batch Normalization) that downsample the input image to a 4×4 spatial grid. Two final 1×1 convolutions produce the mean and log-variance for a latent space at each spatial location (16-dim for FashionMNIST, 32-dim for CIFAR-10). The decoder mirrors this architecture using transposed convolutions to reconstruct the image.

Transformer. A decoder-only Transformer is used for autoregressive modeling of the quantized latent codes. For both datasets, the model has 4 layers, 4 attention heads, an embedding dimension of 256, and is trained on flattened 16-token sequences.

3.2. Training Details

The `SpatialVAE` for both datasets was trained for 200 epochs using the AdamW optimizer with a learning rate of 10^{-3}, a weight decay of 10^{-5}, and a cosine learning rate schedule. The Transformer models were also trained for 200 epochs with AdamW, but with a learning rate of 3×10^{-4} and a weight decay of 0.01. The batch size for all experiments was 256.

[1]Model implementations are available at `vqvae/src/models/{spatial_vae.py,transformer.py}`.

3.3. Post-Hoc Quantization Methods

After training a VAE, the mean latent vectors (μ) are extracted from the training set and a discrete codebook is built by clustering this latent space. The analysis on FashionMNIST compares three distinct graph-based quantization schemes which progressively incorporate the learned manifold geometry.

Method 1: Graph Geodesic Clustering. This baseline clustering method first constructs a k-NN graph from the latent vectors, with edge weights defined by the standard Euclidean distance. K-medoids clustering is then performed, where the distance between any two points on the graph is calculated as the shortest path between them (a "graph geodesic"). This method was applied to the latent space of the `vanilla VAE`.

Method 2: Partial Riemannian Geodesic Clustering. This method aims to improve upon the first by incorporating the VAE's learned geometry. It begins with the same Euclidean-weighted k-NN graph, but then re-weights a small, stratified subset of the edges (5,000 by default) using the decoder-induced Riemannian metric described in the introduction. Clustering is then performed on this hybrid graph, which is mostly Euclidean but is locally corrected by the Riemannian metric. This was also applied to the latent space of the `vanilla VAE`.

Method 3: Full Riemannian Geodesic Clustering. This is the most geometrically rigorous approach. It re-weights *all* unique edges in the k-NN graph with the Riemannian metric. The subsequent k-medoids clustering is therefore performed on a graph that fully represents the learned manifold structure. This method was applied to the latent space of a `SpatialVAE`, which represents a change in the underlying VAE architecture as well as the quantization scheme.

3.4. Autoregressive Modeling

Once the codebook is built, the spatial latent grids for each training image are quantized, converting them into sequences of discrete integer codes. A decoder-only Transformer model is then trained to predict the next code in a sequence, **conditioned on the previous ones**. This allows for autoregressive generation of novel sequences of codes, which can then be mapped back to latent vectors via the codebook and rendered into images by the VAE decoder.

4. Experiments and Results

The pipeline was evaluated on two datasets of varying complexity: FashionMNIST and CIFAR-10. Performance was assessed using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

4.1. Comparative Analysis on FashionMNIST

The primary analysis compares the three quantization methods on the FashionMNIST dataset. As shown in Table 1, the results suggest a nuanced trade-off between geometric accuracy and reconstruction quality. The graph geodesic method with pure Euclidean weights achieves the highest performance. The introduction of a partial Riemannian correction provides a competitive result, while the full Riemannian method, though most geometrically faithful, yields a lower reconstruction quality.

Table 1. FashionMNIST Quantization Method Comparison.[3]

Method	PSNR (dB)	SSIM
Graph Geodesic (Euclidean)	**33.42**	**0.9589**
Partial Riemannian Geodesic	31.95	0.9432
Full Riemannian Geodesic	9.84	0.5168

4.2. CIFAR-10: Post-Hoc vs. End-to-End Training

For the more complex CIFAR-10 dataset, the analysis was focused on comparing our strongest post-hoc method (Full Riemannian Geodesic on a Spatial VAE) against a baseline, end-to-end trained VQ-VAE. The results, summarized in Table 2, suggest the superiority of the end-to-end approach. The baseline VQ-VAE outperforms the post-hoc method across all key metrics, including reconstruction quality (PSNR, SSIM) and perceptual similarity (LPIPS). This suggests the importance of jointly training the encoder, decoder, and codebook on complex data distributions.

Table 2. CIFAR-10 Method Comparison.[5]

Method	PSNR (dB)	SSIM	LPIPS
Post-Hoc (Full Riemannian)	8.72	0.005	0.564
Baseline VQ-VAE	**9.49**	**0.019**	**0.475**

5. Discussion and Conclusions

FashionMNIST results show that while Riemannian metrics achieve more geometrically faithful clustering (quantization error 3.40×10^5 vs. Euclidean's 3.55×10^8), this does not translate to superior image reconstruction. Latent-space distortion metrics do not perfectly correlate with image-space metrics like PSNR. Best performance comes from simple graph-geodesic clustering with Euclidean weights.

The partial Riemannian method suggests that limited application of the Riemannian metric may act as regularization, correcting local geometry without being dominated by noise in the fully-reweighted manifold. However, the full Riemannian method was tested on a `SpatialVAE` while others used vanilla VAE - an architectural confound requiring future work.

On CIFAR-10, the "training mismatch" in post-hoc methods creates a performance bottleneck overcome by joint, end-to-end optimization of standard VQ-VAE.

In conclusion, this project reveals a complex trade-off between geometric fidelity and reconstruction performance. While leveraging a decoder-induced Riemannian metric substantially reduces latent-space quantization error, this geometric accuracy does not translate to improved image quality. On FashionMNIST, simpler graph-geodesic methods with Euclidean weights outperformed more rigorous Riemannian approaches. On CIFAR-10, the "training mismatch" of post-hoc frameworks proved a significant bottleneck, with end-to-end trained VQ-VAE achieving superior results. Future work could address this by fine-tuning the decoder on discrete codebooks, combining benefits of geometric clustering with jointly-optimized performance.

References

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

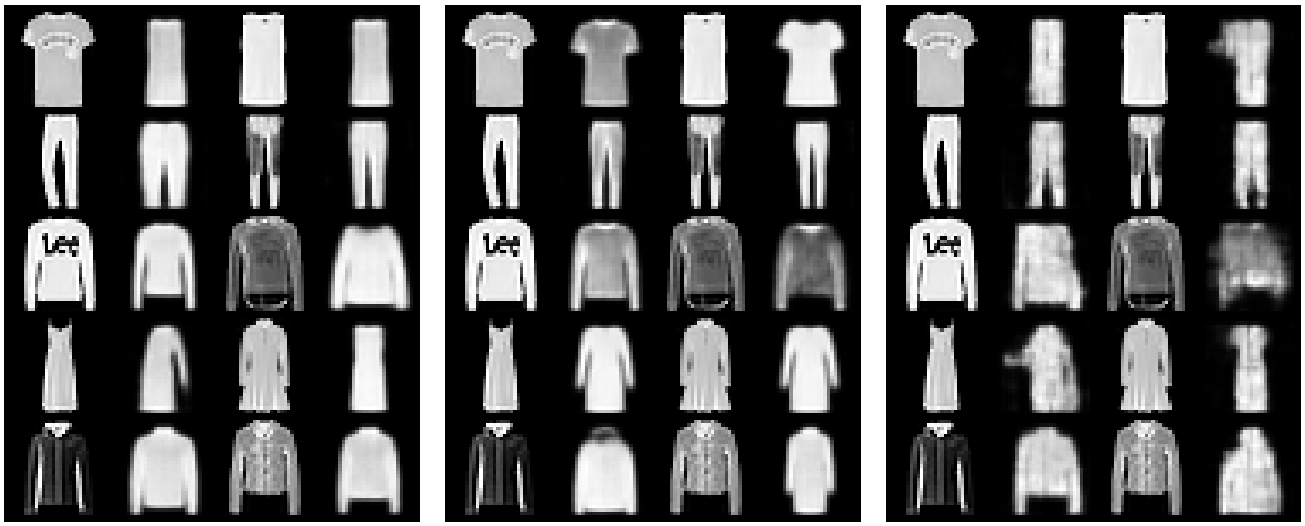
Shao, H., Kumar, A., and Thomas Fletcher, P. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 315–323, 2018.

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

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

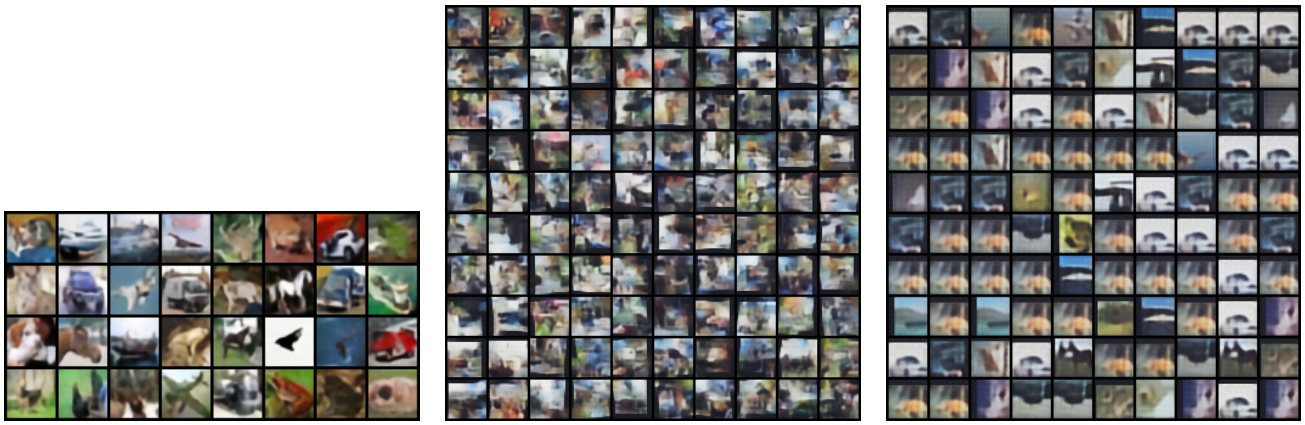
[3]Metrics are computed consistently across all methods. While absolute values may be implementation-dependent, the relative performance ranking is valid.

[5]Metrics are computed consistently across all methods. While absolute values may be implementation-dependent, the relative performance ranking is valid.



(a) Method 1: Graph Geodesic (Euclidean) (b) Method 2: Partial Riemannian Geodesic (c) Method 3: Full Riemannian Geodesic

Figure 1. Qualitative comparison of FashionMNIST reconstructions from the three different post-hoc quantization methods. In each subfigure, the top row shows original images and the bottom row shows their reconstructions. The visual quality directly corresponds to the quantitative results in Table 1, with the Euclidean-based method producing the sharpest images and the Full Riemannian method showing significant degradation.



(a) Baseline VQ-VAE (Reconstructions) (b) Post-Hoc Spatial Geodesic (Generated) (c) Post-Hoc Vanilla Euclidean (Generated)

Figure 2. Qualitative comparison on CIFAR-10. (a) Reconstructions from the end-to-end baseline VQ-VAE are coherent and recognizable. (b) Samples generated by the Transformer in the post-hoc spatial geodesic pipeline are less coherent. (c) Samples from the post-hoc vanilla pipeline show clear signs of mode collapse, where the model repeatedly generates the same limited set of outputs, failing to capture the data diversity.