DLAI Project Template (replace with your title)

July 3, 2025

Ana del Río and Darío González

Abstract

In this project we will revisit the classical VQ-VAE pipeline by proposing an alternative approach which instead of learning a discrete latent space jointly with the encoder, first, we will train a standard Variational Autoencoder (VAE) on the MNIST dataset and perform vector quantization 'a posteriori'. In order to capture in a better way the intrinsic geometry of the latent space, instead of using Euclidean distances, we will use geodesic distances computed by a modified k-means graph so it can work with geodesic distances.

We will compare this model with a classic VQ-VAE trained in an end-to-end form, evaluating both models in terms of reconstruction quality and efficiency of the code.

The preliminary results suggest that geodesic quantization can produce latent representations with a better structure, and this might have a positive impact on posterior tasks.

1. Introduction

Non-supervised learning is a fundamental tool for discovering structured representations in complex data, without the need for explicit labels. In this category, we can find Variational Autoencoders (VAE) and their discrete variants, such as Vector Quantized Variational Autoencoders (VQ-VAE), and they are really effective in tasks such as compression or sample generation.

If we focus on the VQ-VAE model, we see that it introduces a quantization layer that transforms the continuous latent space into a discrete one. This approach has significant limitations: 1. As quantization is learned jointly with the encoder it can lead to suboptimal solutions. 2. It relies on euclidean distances in the latent space, completely ignoring its intrinsic geometry, that can be nonlinear or complexly curved.

In this project, we propose a modular reinterpretation of the VQ-VAE pipeline. Instead of training the entire system end-to-end, we first train a continuous VAE on the MNIST dataset. Next, we construct a k-nearest neighbor (k-NN) graph over the obtained latent vectors and then we apply a geodesic distance-based clustering algorithm to that graph.

By doing all of this, we want to capture better the structure of the latent space before applying quantization, so we will obtain more coherent and useful discrete representations.

Finally, we compare the performance of the alternative that we implemented with the classical VQ-VAE.

2. Related Work

Several works have explored alternatives to euclidean distances.

2.1 Papers One foundational contribution is the Isomap algorithm by Tenenbaum et al. (2000) (Tenenbaum et al., 2000). Their method "uses easily measured local metric information to learn the underlying global geometry of a data set" demonstrating that approximating geodesics using a k-NN graph captures the true shape of the manifold where data resides. This principle inspires our first step: constructing a neighborhood graph over the continuous VAE codes to estimate reliable geodesic distances.

Then, Asgharbeygi and Maleki (2008) (Asgharbeygi & Maleki, 2008) introduced Geodesic K-means. They extend the classical k-means algorithm and acknowledge that "we introduce a class of geodesic distances and extend the k-means clustering algorithm to employ this distance metric. Adopting their variant (which replaces the euclidean distance with the shortest path length in the graph) will allow us to form geodesic centroids-medoids that respect the latent topology before quantization. This work directly inspires the clustering strategy that we will apply to quantize the latent space of our model.

Finally, the work by Yang et al. (2018)(Yang et al., 2018),

Email: Ana del Río and Dario González <delriodelbarrio.2183763@studenti.uniroma1.it and gonzalezsanchez.2183763@studenti.uniroma1.it>.

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

investigate geodesic clustering in deep generative models, such as VAEs. They show that "taking the geometry of generative model into account is sufficient to make simple clustering algorithms work well over latent representations". Their result prove that, even in latent spaces distorted by a VAE, measuring distances on the embedded manifold is sufficient for simple algorithms (such as our geodesic k-means) to perform competitively.

Together, these three works support the choice of our a posteriori pipeline:

1. Graph k-NN on the latent continuum.

2. Geodesic distances (Dijkstra/ Floyd-Warshall).

3. Geodesic k-means to extract a discrete code book consistent with the internal geometry.

Then, the resulting discretization inherits the structure of the manifold and provides more informative codes for the auto-regressive model that we will complete in later stages.

2.2 Comparison with our approach Our project differs from the classical VQ-VAE in two main ways:

1. Instead of learning the code book jointly with the encoder, we decouple training and quantization: we train a continuous VAE and apply quantization afterwards.

2. Instead of using euclidean distance in latent space, we compute geodesic distances over a k-NN graph, which better captures the latent manifold's structure.

We then apply Geodesic k-means for clustering, aiming to produce more coherent and topology-aware code books. This modular and geometry-aware approach is expected to yield improved latent representations and more effective discrete encodings compared to the standard VQ-VAE pipeline.

3. Method

3.1 Baseline The baseline for this project is the classical Vector Quantized Variational Autoencoder (VQ-VAE), introduced by van den Oord et al. (2017)(van den Oord et al., 2017). VQ-VAE modifies the standard VAE by replacing the continuous latent space with a discrete code book of embeddings.

The code book is updated via Exponential Moving Averages (EMA) or using a separate loss term, depending on the variant. The model is trained end-to-end with a reconstruction loss and a code book commitment loss that encourages the encoder to stay close to the selected code.

While VQ-VAE is successful in discrete generative modeling, as it has been said before, it treats the latent space as euclidean and optimizes the code book jointly with the encoder, and this can lead to suboptimal representations.

In our experiments, we replicate this architecture and training pipeline as a baseline for comparison against our proposed method.

3.2 Proposed method and contribution In contrast to the standard VQ-VAE, our method adopts a modular and geometry-aware approach to latent space quantization. It is structured in two main phases:

1. Continuous VAE training: we first train a Variational Autoencoder (VAE) on the MNIST dataset. The VAE encodes each image into a low-dimensional latent vector sampled from a Gaussian distribution. The training is done using the standard ELBO loss, combining reconstruction loss (BCE) and KL divergence.

2. 'A posteriori' geodesic quantization: after training the VAE, we extract all the latent vectors from the training set and construct a k-nearest neighbors graph over this points. Instead of using euclidean distance, we use geodesic distances between points via shortest paths on the graph. Thanks to this, we can capture the intrinsic geometry of the latent space more accurately. Then, we apply Geodesic K-means, a clustering algorithm that uses these geodesic distances to define cluster centroids. Each latent vector is assigned to its nearest centroid based on the geodesic metric, producing a discrete code book.

This decoupled approach allows for greater flexibility in quantization, avoids inference during VAE training, and leverages manifold geometry to form more coherent clusters. Our contribution lies in applying and evaluating this post-hoc, graph-based quantization strategy, and comparing it empirically with the classical VQ-VAE baseline in terms of reconstruction quality and representational structure.

4. Experiments and results

4.1 Experimental setup To evaluate the effectiveness of geodesic quantization, we conducted experiments on the MNIST dataset consisting of 60000 training images and 10000 test images, all 28x28 in size and in grayscale.

We trained three models:

1. VAE (classic Variational Autoencoder).

2. VQ-VAE (Vector Quantized VAE).

3. Geo-VQ (our version with geodesic quantization).

All of them share the same encoder and decoder architecture, which consist of fully connected layers. The latent space dimension was set to 20 in all cases. A code book of 64 embeddings was used for VQ-VAE and Geo-VQ. The

optimizer was Adam with a learning rate of 10^{-3} and a batch size of 128, and training for 10 epochs for both VAE and VQ-VAE models.

In the case of the Geo-VQ model, the pre-trained VAE encoder was used to generate the latent vectors for the entire training set. A k-NN graph (with $k = 10$) was constructed using euclidean distances, and geodesic distances were calculated on this graph using Dijkstra's algorithm. Finally, an adapted version of K-means (Geodesic K-means) was applied to these distances to generate the quantization code book.

4.2 Metrics To compare the models we evaluated three metrics in the test set:

1. Binary Cross Entropy (BCE) between the reconstructed image and the real one.

2. KL Divergence (in the VAE) and Commitment Loss (in the VQ-VAE).

3. 'Total Loss' as the sum of both.

4.3 Quantitative comparison

Table 1. Comparison between total loss, reconstruction (BCE) and commitment term across the evaluated models.

Model	Total Loss	BCE	KL / Commit
VAE	1,055,250.44	801,876.88	253,373.56
VQ-VAE	2,424,601.00	1,959,660.88	464,940.11
Geo-VQ	1,126,944.75	1,126,944.75	—

In the table we can observe that Geo-VQ has a 'Total Loss' which is much lower than the one we get with the traditional VQ-VAE, so we avoid the compromise cost that penalizes the original model. With all of this we conclude that the use of geodesic distances allows us to preserve better the structure of the latent space and generate a more representative code book.

4.4 Visualizations (qualitative comparison) Now, we compare the reconstructions that have been generated by the three models.

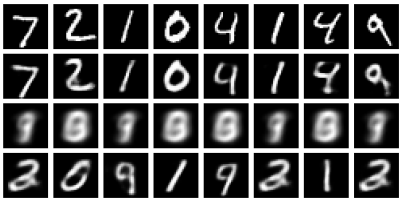

Figure 1. Qualitative comparison of reconstructions from each model. Top: original, 2nd: VAE, 3rd: VQ-VAE, bottom: Geo-VQ.

From Figure 1, we can observe the following from each model:

1. VAE (second row): the reconstructions are really similar to the input digits, but they are blurry and smooth, which is caused by the gaussian latent space and the pixel-wise reconstruction loss.

2. VQ-VAE (third row): we can see that the reconstructions are more blurry and distorted compared to VAE. In some cases, we cannot even distinguish the numbers (for example, in the case of the number 1). This degradation is due to discrete bottleneck and a large commitment loss weight, that makes the model rely less on reconstruction quality.

3. Geo-VQ (fourth row): the model achieves more accurate reconstructions that VQ-VAE, and in some cases it is better than the VAE. Shapes are sharper and better aligned with the original structure of the digits. The absence of commitment loss (as the latent space has a different structure) may contribute to a better gradient flow and reconstruction fidelity.

5. Conclusions

In this project, we have compared the performance of the models VAE, VQ-VAE and Geo-VQ on the MNIST dataset.

Quantitative results show that VQ-VAE suffers from a high commitment loss, Geo-VQ has a better balance between reconstruction accuracy and compact latent representation.

On the other hand, qualitative results show that Geo-VQ generates sharper and more faithful reconstructions. These findings highlight the potential of incorporating geometric structure into vector quantization for improved generative modeling.

References

Asgharbeygi, N. and Maleki, A. Geodesic clustering in deep generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1732–1743, 2008. URL https://arxiv.org/abs/0809.2251.

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

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural*

Information Processing Systems (NeurIPS), 2017. URL https://arxiv.org/abs/1711.00937.

Yang, G., Hypki, A., and Welling, M. Geodesic clustering in deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://papers.nips.cc/paper_files/paper/2018/file/0c576f7c923121e69edce3aa3734fffa-Paper.pdf.