Clustering with Neural Networks using Hugging Face Datasets

Submitted to : Moin Mostakim

May 19, 2025

1 Introduction

This project explores unsupervised clustering using neural networks and open datasets from the Hugging Face Hub. Clustering refers to the task of grouping similar data points without predefined labels. The project utilizes image data from the MNIST dataset and applies neural network-based feature extraction followed by clustering algorithms.

2 Objective

The goal is to develop a neural network that transforms data into a compact latent space suitable for clustering and to evaluate the resulting clusters using metrics such as Silhouette Score and visualizations via t-SNE.

3 Tools and Libraries

- **Programming Language:** Python 3.8+
- **Frameworks:** PyTorch, scikit-learn, Hugging Face `datasets`
- **Visualization:** Matplotlib

4 Dataset

We use the MNIST dataset from Hugging Face (`mnist`), consisting of 60,000 handwritten grayscale images of size 28x28.

5 Neural Network Design

A simple feedforward autoencoder is designed with a latent space of dimension 64. The encoder compresses the 784-dimensional input into this latent space. Only the encoder is used for clustering.

Architecture Summary

- **Encoder:** Linear → ReLU → Linear (784 → 128 → 64)
- **Decoder (used only during training):** Linear → ReLU → Linear (64 → 128 → 784)
- **Loss:** Mean Squared Error (reconstruction loss)

6 Training

The autoencoder was trained for 10 epochs using the Adam optimizer with a learning rate of 0.001 and batch size of 256. After training, the encoder was used to extract embeddings for the entire data set.

7 Clustering Algorithms

7.1 K-Means Clustering

K-Means clustering was applied to the 64-dimensional embeddings. The number of clusters was set to 10 (for 10 digit classes).

- **Clusters Found:** 10
- **Silhouette Score:** 0.0903

7.2 DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Noise Applications) was also applied to explore density-based clustering without predefining the number of clusters.

- **Clusters Found (excluding noise):** 9
- **Silhouette Score:** -0.3330

The negative silhouette score suggests that DBSCAN did not form well-separated clusters in the current latent space without parameter tuning. This outcome is expected in high-dimensional embeddings when the density of data points varies across classes.

8 Visualization

To visualize the clusters in two dimensions, t-SNE was applied to a random subset of 1000 embeddings. The resulting 2D projection was color-coded by cluster assignments.

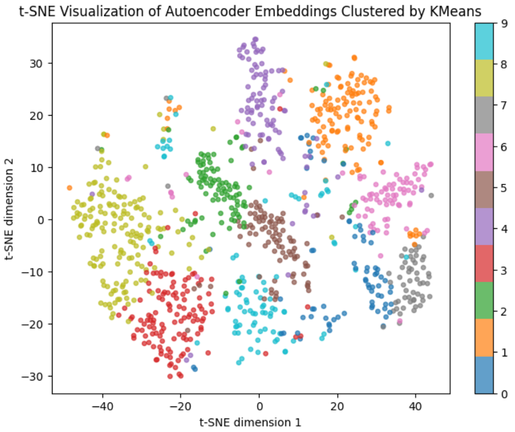

Figure 1: t-SNE Visualization of K-Means Clusters (Subset of 1000 Samples)

9 Evaluation

- **Quantitative:** Silhouette Score was highest for K-Means (0.0903), indicating moderate intracluster compactness.
- **Qualitative:** t-SNE plots visually confirmed that some digit classes form distinct clusters, while others overlap.

10 Conclusion

This project demonstrated how neural networks can transform raw input data into a latent embedding space suitable for clustering. K-Means outperformed DBSCAN in this setting, although further tuning of DBSCAN parameters could yield better results. The project confirms the value of combining representation learning with clustering algorithms for unsupervised tasks.

11 References

- Hugging Face Datasets: `https://huggingface.co/docs/datasets`
- PyTorch Documentation: `https://pytorch.org`
- Scikit-learn Clustering: `https://scikit-learn.org/stable/modules/clustering.html`