Deep Autoencoder-Based Feature Learning for Enhanced Unsupervised Clustering of MNIST

Rantu Das
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
rantu.das@g.bracu.ac.bd

***Abstract*—This project introduces a deep learning framework using a Convolutional Autoencoder (CAE) for unsupervised clustering on the MNIST dataset. The CAE is designed to learn low-dimensional, meaningful representations of digit images by compressing and reconstructing input data. After training, the encoder's latent features are further reduced using Principal Component Analysis (PCA), and clustered via the KMeans algorithm. Dimensionality reduction and clustering are evaluated using visualization techniques like t-SNE and quantitative metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Silhouette Score. The results demonstrate the effectiveness of combining deep feature learning with classical clustering methods to uncover digit groupings in an unsupervised setting.**

***Index Terms*—Convolutional Autoencoder, MNIST, Clustering, PCA, KMeans, t-SNE, ARI, NMI.**

I. Introduction

Clustering is a cornerstone of unsupervised machine learning, allowing data to be grouped by underlying similarities without requiring labeled supervision. Traditional clustering methods like KMeans perform well on low-dimensional and structured data but often fail when applied to high-dimensional image datasets due to their inability to learn spatial hierarchies. The MNIST dataset, composed of handwritten digit images, represents such a challenge.

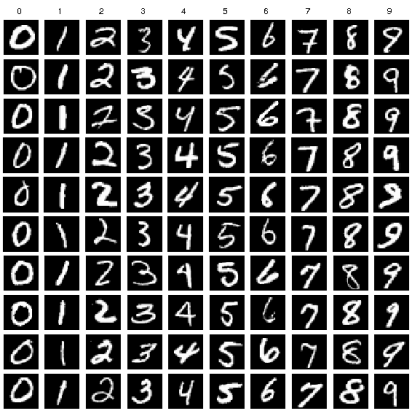

Fig. 1. MNIST dataset

Deep learning offers an effective solution through architectures like Convolutional Autoencoders (CAEs), which can automatically learn hierarchical and spatially-aware features [2]. In this work, we train a CAE on MNIST to extract compact features suitable for clustering. These features are then processed with PCA to reduce noise and dimensionality, followed by clustering using KMeans. To interpret the clustering quality, we visualize the results using t-SNE and evaluate them with established clustering metrics.

II. Methodology

A. Dataset and Preprocessing

The MNIST dataset consists of 70,000 grayscale handwritten digit images (60,000 training and 10,000 test), each sized 28×28 pixels [1]. For this unsupervised task, both sets are merged. Images are normalized to [0, 1] and loaded into PyTorch tensors using DataLoader with a batch size of 256.The MNIST dataset consists of 70,000 grayscale images of handwritten digits (0 through 9), each of size 28x28 pixels. These include 60,000 training samples and 10,000 test samples. Since our task is unsupervised, we combine both sets to form a unified dataset. The images are normalized to a [0, 1] range to enhance training convergence and are loaded into PyTorch tensors using DataLoader with a batch size of 256.

B. Convolutional Autoencoder Architecture

The CAE model consists of an encoder that compresses the input into a latent space and a decoder that reconstructs the image [3]. The model is symmetrical and uses Conv2D, BatchNorm2D, ReLU, and MaxPool2D in the encoder, and ConvTranspose2D layers in the decoder.

C. Training Strategy

The model is trained for 50 epochs using the Adam optimizer and Binary Cross Entropy Loss (BCELoss), which is ideal for binary pixel data. Training follows the typical pipeline of zeroing gradients, forward propagation, loss computation, backpropagation, and parameter updates. The model learns to minimize reconstruction loss, aiming to produce outputs as close as possible to the original input images.

D. Clustering on Latent Space

After training, we discard the decoder and extract features from the encoder. These latent features, flattened per image, are reduced to 50 dimensions using PCA. The reduced features are clustered into 10 groups using the KMeans algorithm. This

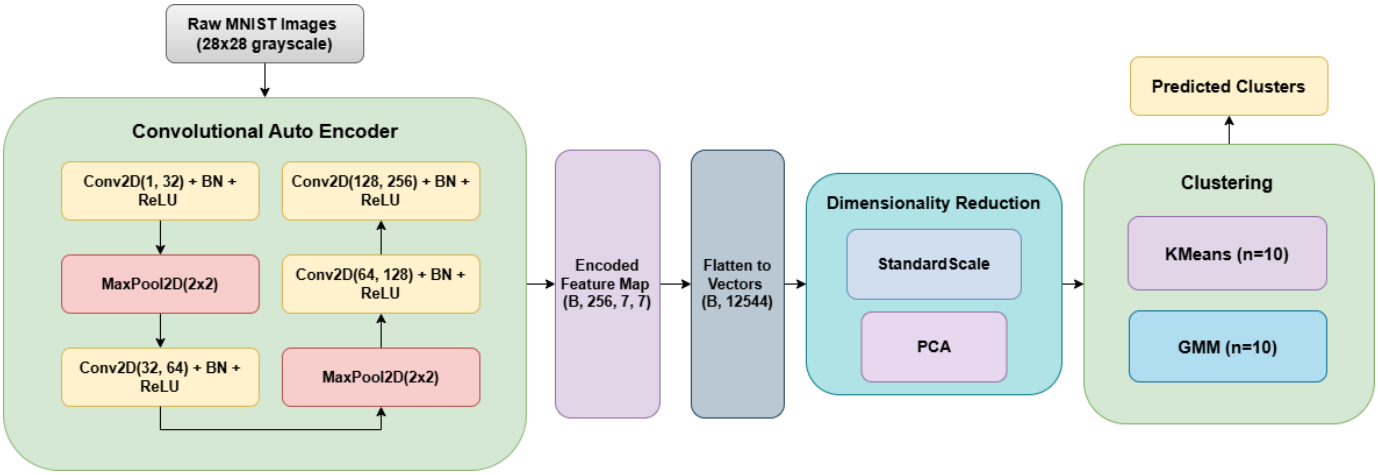

Fig. 2. Convolutional Autoencoder Architecture

step allows the model to group digit images based on their learned representations rather than raw pixels.

E. Evaluation Metrics

To quantitatively evaluate clustering:

- ARI (Adjusted Rand Index): Measures similarity to true labels, adjusted for chance.
- NMI (Normalized Mutual Information): Measures mutual dependence between labels and clusters.
- Silhouette Score: Measures cohesion and separation of clusters.

F. Visualization

t-SNE is used to project 2000 PCA-reduced vectors into 2D space, visualized using seaborn to compare true labels vs KMeans cluster assignments [4]. Cluster centroids are also overlaid to visualize their positions relative to data points.

III. Results and Discussion

A. Quantitative Results

The ARI and NMI indicate that the CAE successfully learns representations that preserve digit identity [5]. The modest Silhouette Score highlights some cluster overlap, which is expected in real-world image datasets.

TABLE I
Clustering Evaluation Metrics

Metric	Score
Adjusted Rand Index (ARI)	0.4296
Normalized Mutual Information (NMI)	0.5323
Silhouette Score	0.1108

B. Visual Insights

t-SNE plots show that the CAE effectively separates many digit classes. Clusters like 1, 0, and 8 form distinct groupings, whereas digits with similar structures (like 4 and 9) tend to overlap. The confusion matrix reveals strong alignment between KMeans clusters and true labels, though not perfect.

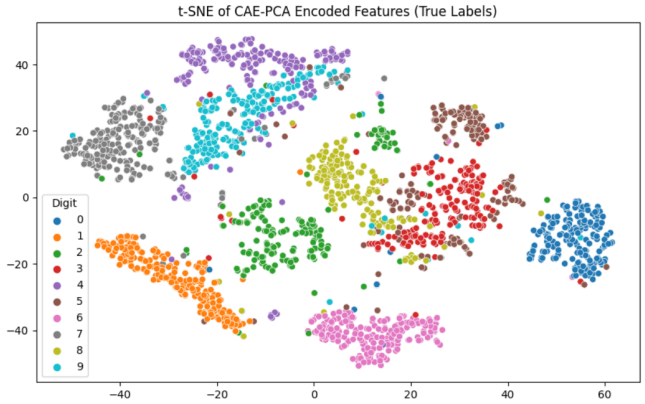

Fig. 3. t-SNE of CAE-PCA Encoded Features (True Labels)

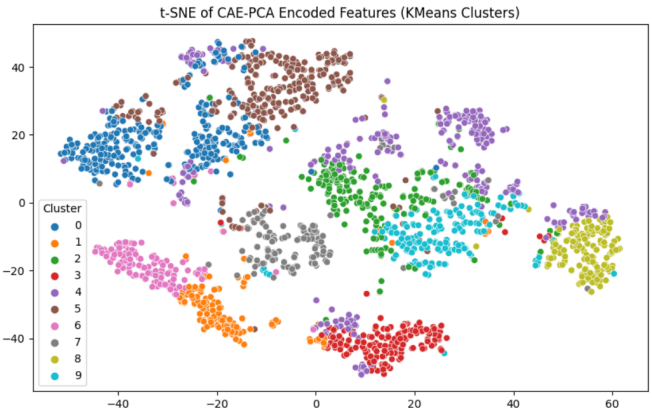

Fig. 4. t-SNE of CAE-PCA Encoded Features (KMeans Clusters)

C. Discussion

Despite strong performance, certain digits exhibit fragmentation (e.g., digit 7 spread across multiple clusters) or overlap (e.g., digits 3 and 5). This suggests the latent space could be enhanced further using contrastive or supervised constraints.

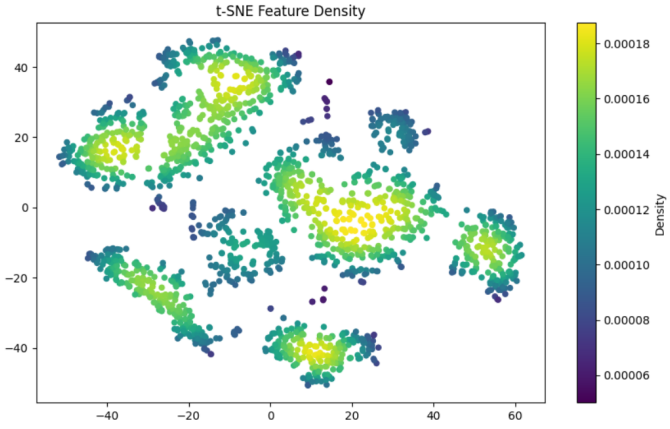

Fig. 5. t-SNE Feature Density

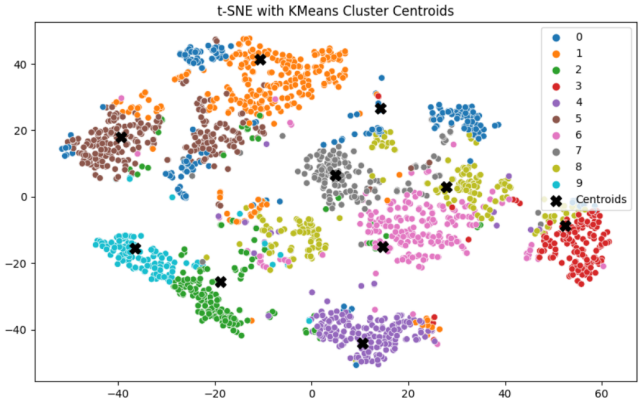

Fig. 6. t-SNE with KMeans Cluster Centroids

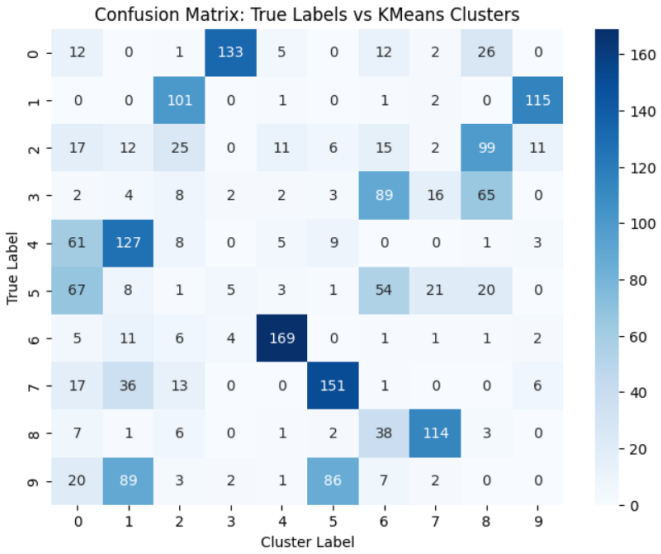

Fig. 7. Confusion Matrix

Also, KMeans assumes spherical clusters, which might not align with the true shape of latent data.

IV. Conclusion

This project demonstrates that a well-designed convolutional autoencoder can learn semantically meaningful representations of digit images, which significantly enhance clustering performance. The combination of CAE, PCA, and KMeans leads to clear cluster formation in both visual and quantitative assessments. While the model shows promise, future work can explore contrastive learning, variational autoencoders, or graph-based clustering methods to improve cluster purity.

References

[1] Y. LeCun, C. Cortes, and C. J. Burges, *"MNIST handwritten digit database"*, AT&T Labs, 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist

[2] G. Hinton and R. Salakhutdinov, *"Reducing the dimensionality of data with neural networks"*, Science, vol. 313, no. 5786, pp. 504–507, 2006.

[3] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, *"Extracting and composing robust features with denoising autoencoders"*, Proceedings of the 25th International Conference on Machine Learning (ICML), 2008.

[4] L. van der Maaten and G. Hinton, *"Visualizing data using t-SNE"*, Journal of Machine Learning Research, vol. 9, no. Nov, pp. 2579–2605, 2008.

[5] J. MacQueen, *"Some methods for classification and analysis of multivariate observations"*, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.