MATH 5473 Final Project: Dimension Reduction and Visualization on MNIST Data Series

LEE Young Kyu[1], KIM Jaehyeok[2], MA Jiabo[2] and JIN Cheng[2] {ykleeac, jkimbf, jmabq, cjinag}@ust.hk
[1]: Department of Mathematics, HKUST [2]: Department of Computer Science and Engineering, HKUST

1. Introduction

In modern machine learning, visualizing high-dimensional data is key to understanding complex model behaviors, especially in image recognition tasks. Our study uses dimensionality reduction methods like MDS, ISOMAP, LLE, t-SNE, and UMAP to analyze the MNIST and Fashion-MNIST datasets and the features of models trained on them. We aim to evaluate the effectiveness of these techniques in revealing data structures and improving model interpretability.

2. Datasets

MNIST Dataset: The MNIST dataset consists of 70,000 images of handwritten digits, each a 28x28 pixel grid.

Fashion-MNIST Dataset: The Fashion-MNIST dataset includes 70,000 images of fashion items like T-shirts and trousers also in 28x28 pixels across 10 classes.

3. Method

Multidimensional Scaling (MDS):
MDS seeks to minimize the stress function $\sigma(X) = \sqrt{\sum_{i<j}(d_{ij} - \|x_i - x_j\|)^2}$ where d_{ij} are the distances in the original space and x_i, x_j are the coordinates in the reduced space.

Isometric Mapping (ISOMAP):
ISOMAP constructs a neighborhood graph $\min_X \sum_{i<j}(\delta_{ij} - \|x_i - x_j\|)^2$ and then uses graph distances to approximate the geodesic distances in the original manifold, where δ_{ij} is the geodesic distance between points i and j in the original data.

Contributions

LEE, Young Kyu: **Initial Pipeline Setup**
MA, Jiabo: **Result Refinement**
JIN, Cheng: **Poster Design**
KIM, Jaehyeok: **Poster Refinement & Presentation**

3. Method (continued)

Locally Linear Embedding (LLE):
LLE computes the linear coefficients that best reconstruct each data point from its neighbors, and then chooses the low-dimensional embeddings that ibest preserve these local properties by solving

$$\min_W \sum_i \|x_i - \sum_j w_{ij} x_j\|^2 \quad \text{subject to} \quad \sum_j w_{ij} = 1$$

where w_{ij} represents the contribution weight from neighbor j to point i.

t-Distributed Stochastc Neighbor Embedding (t-SNE):
t-SNE preserves local structures and reveals clusters at different scales by minimizing distributional differences between pairwise similarities of input data points and their low-dimensional embeddings:

$$\min_Y KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

where P is the probability distribution of pairwise similarities in the high-dimensional space, Q is the probability distribution of pairwise similarities in the low-dimensional embedding, and p_{ij} and q_{ij} are the elements of these distributions respectively.

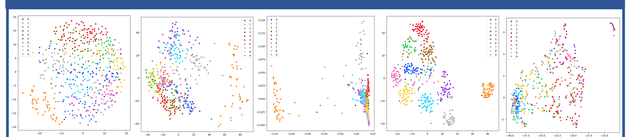
Uniform Manifold Approximation and Projection (UMAP):
UMAP constructs a fuzzy simplicial set representing the high-dimensional data and then optimizes a low-dimensional layout of this set, aiming to preserve the topological structure by:

$$\min_Y \sum_{i \neq j} (p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}})$$

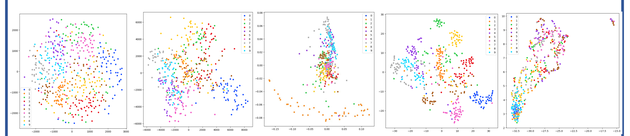
3. Method (continued)

where p_{ij} is the probability of connection between points i and j in the original data and q_{ij} is the probability of connection in the low-dimensional embedding.

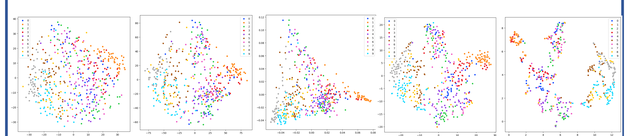
4. Experimental Results



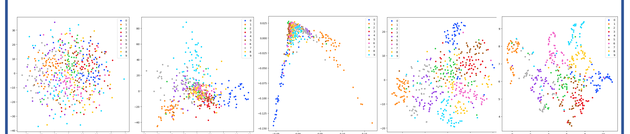
Left to right: Visualization results of MDS, ISOMAP, LLE, t-SNE, and UMAP on the **MNIST Images**



Left to right: Visualization results of MDS, ISOMAP, LLE, t-SNE, and UMAP on the **MNIST Features of ResNet18**



Left to right: Visualization results of MDS, ISOMAP, LLE, t-SNE, and UMAP on the **Fashion-MNIST Images**



Left to right: Visualization results of MDS, ISOMAP, LLE, t-SNE, and UMAP on the **Fashion-MNIST Features of ResNet18**

5. Analysis

MNIST Analysis:
MNIST, with simpler handwritten digits, shows well-separated classes when visualized using t-SNE, likely due to its proficiency in preserving local structures. Despite a high accuracy of 96% with a ResNet18 classifier, the feature visualization lacks the clarity of raw image inputs. This discrepancy may stem from the model's complexity and the transformation of raw pixels into a deep feature space, possibly capturing unnecessary noise.

Fashion-MNIST Analysis:
Fashion-MNIST, consisting of more complex clothing images, presents challenges in visualization, with UMAP providing the best results. The classifier achieves an accuracy of 86%, reflecting the dataset's complexity. UMAP's effectiveness in handling Fashion-MNIST could be due to its balance in capturing both local and global data structures, essential for datasets with subtle variations within classes and similarities across different categories.

6. Conclusion

In conclusion, our study highlights the importance of selecting appropriate dimensionality reduction techniques for different datasets to enhance data visualization and model interpretability. For the MNIST dataset, t-SNE provided clear visual separations of classes, whereas for the more complex Fashion-MNIST, UMAP excelled by effectively capturing both local and global structures. Despite high classification accuracies with the ResNet18 classifier, discrepancies in visualization quality emphasize the need for careful consideration in model deployment. This research underscores the necessity of aligning model interpretability with classification performance in complex machine learning tasks.

References

Rabadán, Raúl, and Andrew J. Blumberg. Topological data analysis for genomics and evolution: topology in biology. Cambridge University Press, 2019.