Experiment Report on Dimensionality Reduction using PCA and UMAP

Introduction

This report describes an experiment using PCA and UMAP on a dataset of handwritten digit images, analysing the difference in parameter choices affect the outcome of these dimensionality reduction techniques.

PCA Experiment

PCA was applied to the dataset of digit images to reduce the dimensionality while retaining the variance in the data. The cumulative explained variance was plotted against the number of components to identify the number of components that explain 80% of the variance.

The choice of 50 components for 80% variance retention was a balance between information retention and model simplicity. This number of components was enough to significantly reduce the dataset's dimensionality without losing most of the information.

UMAP Experiment

Three sets of hyperparameters were tested:

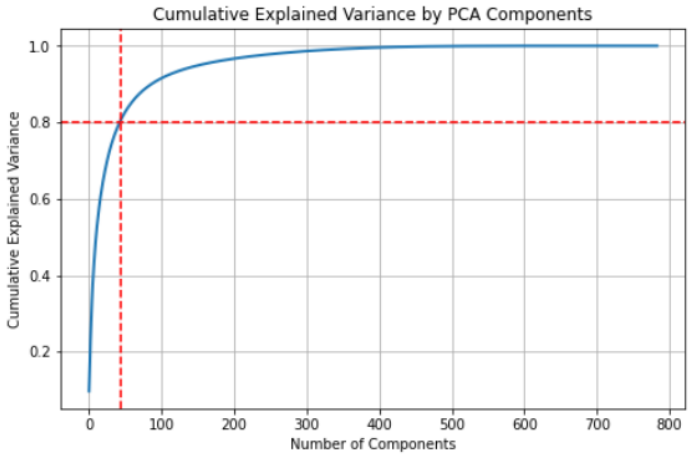
- Experiment 1: **n_neighbors**=5, **min_dist**=0.3, **spread**=1.0
- Experiment 2: **n_neighbors**=15, **min_dist**=0.1, **spread**=0.5
- Experiment 3: **n_neighbors**=30, **min_dist**=0.4, **spread**=1.5

Each set of parameters was chosen to observe the effects of local versus global data structure, cluster tightness, and point spread.

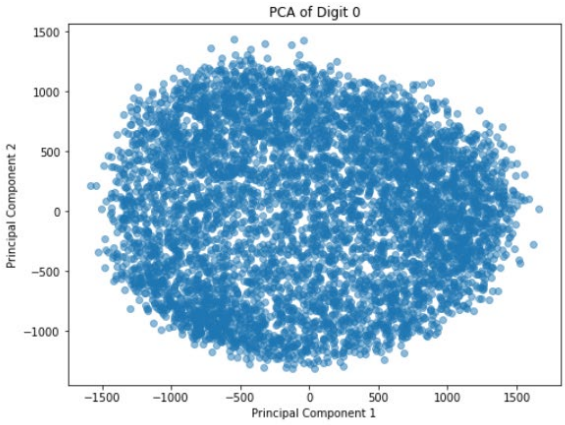
- The n_neighbors parameter significantly affects the visualization, with smaller values highlighting local data characteristics and larger values capturing more of the global data structure.
- The min_dist value affects how closely points in the same neighborhood are packed together. Lower values result in denser clusters.
- The spread parameter dictates the extent to which UMAP expands the neighborhood around each point. Higher values allow clusters to occupy more of the embedding space, which can aid in visual separation and potentially in classification tasks.

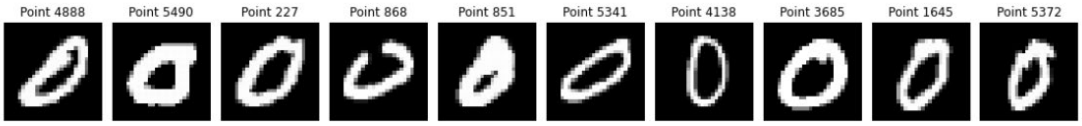
PCA Results

The PCA analysis revealed that approximately 50 components were required to capture 80% of the variance. The scatter plot of the first two principal components for digit '0' showed a wide spread, indicating diverse writing styles for the same digit.




```
((60000, 28, 28), (60000,))
```

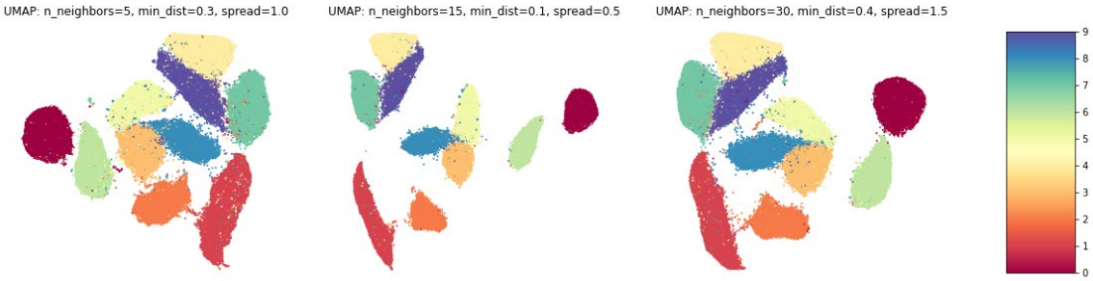




The scatter plot presents a two-dimensional PCA reduction of images of the digit '0'. In this representation, each point corresponds to a single image, positioned according to its values for the first two principal components. These components are linear combinations of the original pixel values that explain the most variance within the dataset.

Observing the scatter plot, images are spread out across the plane, suggesting a variety in the handwriting styles captured by the first two principal components. The density and spread of points could also indicate the presence of subgroups or clusters within the images of the digit '0', potentially corresponding to different writing styles.

UMAP Results



The UMAP visualizations displayed a clear distinction among the clusters corresponding to the ten-digit classes, as indicated by the color-coded points. The influence of hyperparameters on the clustering patterns is evident:

1. **UMAP 1** demonstrated distinct, well-separated clusters, which is indicative of the algorithm capturing the local structure due to a smaller **n_neighbors** value.
2. **UMAP 2** showed clusters that are closer together, reflecting the impact of a smaller **spread** and **min_dist**, leading to a denser embedding.
3. **UMAP 3** struck a balance between the preservation of local and global data structures, with the higher **n_neighbors** value revealing broader data trends and the higher **spread** leading to well-separated clusters.

Comparing PCA and UMAP

While PCA provided insights into the variance structure of the dataset, UMAP revealed intricate patterns and relationships within the data. The UMAP plots suggest that the algorithm can discern between different digit classes, capturing both local and global structures in a way that PCA cannot be due to its linear nature.