Distribution Clustering Using Fisher's Linear Discriminant for Improved Cluster Separability

I. Implementation Details

A. Process Flow

- The process begins by loading images from a specified directory, using a **ResNet Architecture** to extract features from each image.
- The extracted features are stored in a feature matrix, which serves as the input for various clustering algorithms.
- Multiple clustering methods are applied to the feature matrix, including:
  - KMeans Clustering
  - Gaussian Mixture Model (GMM) Clustering
  - Spectral Clustering
  - Distribution-based Clustering
  - **FLD-Enhanced Distribution-based Clustering** using affinity matrices.
- For each clustering method, the labels are generated, and the clustering results are saved in respective directories.
- The clustering outcomes are then evaluated using metrics such as:
  - Silhouette Score
  - Calinski-Harabasz Score
  - Davies-Bouldin Score
- Finally, the evaluation results are compiled into a **Pandas DataFrame** for easier comparison and visualization.

B. FLD-Inspired Feature Refinement

1) **Feature Extraction**: Using a pre-trained AlexNet model, the algorithm extracts deep features from images, creating a feature vector for each image.
2) **Initial Clustering and FLD Refinement**: After Initial clustering is performed, Fisher's Linear Discriminant (FLD) is then applied to the clusters, aiming to maximize inter-cluster distances and minimize intra-cluster variance:
   - **Between-Cluster Scatter Matrix** S_B:
   $$S_B = \sum_{i=1}^{K} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$
   where K is the number of clusters, N_i is the number of points in cluster i, μ_i is the mean of cluster i, and μ is the overall mean.
   - **Within-Cluster Scatter Matrix** S_W:
   $$S_W = \sum_{i=1}^{K} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T$$
   where C_i denotes points in cluster i.
   - **Optimal Projection Direction** w: Calculated as the principal eigenvector of $S_W^{-1} S_B$, refining feature vectors to improve separability.
3) **Affinity Matrix Construction with FLD-Refined Features**: Using the refined feature space, an affinity matrix is constructed, reflecting the enhanced separability from FLD.
4) **Distribution Clustering with Enhanced Affinity Matrix**: Clustering is performed using the refined affinity matrix, grouping points with high affinity to create compact clusters, while outliers are handled separately. The algorithm initializes clustering with the smallest non-zero element in the affinity matrix, forming a candidate cluster H . It expands H by including points with average second-order distances below a threshold tau , and validates clusters that meet a minimum size, assigning unique labels.

II. Datasets

A. Testing Dataset

The unit test dataset can be found at: Unit Test Dataset. It contains 100 images of the coliseum, with clustered images displayed in `Main.ipynb`.

B. Additional Evaluation

Additional datasets used include:

- **MNIST** - Sample from the digit recognition dataset.
- **CalTech 101** - Randomly selected folders containing various images of objects/animals.

III. RESULTS

Scoring results on the test and MNIST datasets for all clustering algorithms. Additional results are in the GitHub repository.

method	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
kmeans	0.025718	4.999815	2.248301
gmm	0.021279	4.790093	1.896970
spectral	0.025493	4.747467	2.313330
distribution	0.028362	3.192529	2.287611
improved_distribution	0.034347	3.462428	2.224356

method	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
kmeans	0.088517	27.638248	2.217181
gmm	0.100902	27.930472	2.153142
spectral	0.070360	26.062310	2.231486
distribution	0.189591	17.558190	1.549544
improved_distribution	0.186540	18.402878	1.805982

Fig. 1. Table 1: Sample Dataset Results; Table 2: MNIST Dataset Results

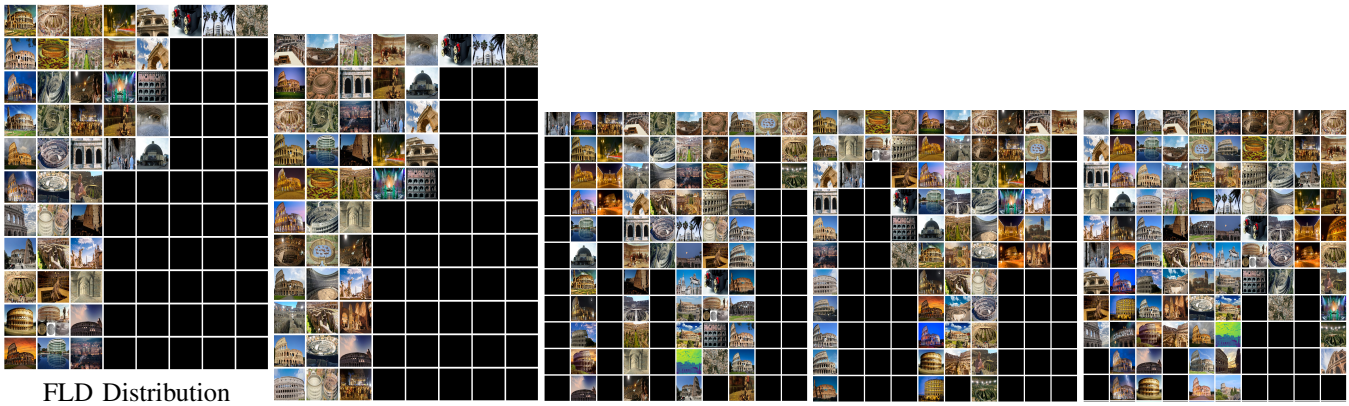
A. Analysis

- The Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score are used to assess clustering quality. Higher Silhouette and Calinski-Harabasz Scores indicate better-defined clusters, while lower Davies-Bouldin Scores indicate compact and well-separated clusters. FLD-enhanced distribution clustering shows improved separability.

B. Challenges

- Original code issues and missing files required rewriting the code.
- Dataset ambiguity led to the selection of datasets not exactly same as the paper.
- The absence of true labels precluded purity score calculation, leading to alternative evaluation metrics.

C. Sample Cluster Visualizations



- FLD Distribution Clustering
- Distribution Clustering
- KMeans
- GMM
- Spectral