DIMENSIONALITY'S BLESSING

CLUSTERING IMAGES BY UNDERLYING DISTRIBUTION

AUTHORS

Wen-Yan Lin, Siying Liu, Jian-Huang Lai, Yasuyuki Matsushita

CVPR 2018

EE798R COURSE PROJECT

INSTRUCTOR: DR. TUSHAR SANDHAN
MENTOR: HIDANGMAYUM BEBINA DEVI
BY: VISHAL HIMMATSINGHKA (211175)

INTRODUCTION

- **Contrast-Loss Challenge**: Traditional clustering in high-dimensional data is difficult due to "contrast-loss," where distances between data points converge, complicating separation.
- **New Perspective**: This paper reinterprets contrast-loss as an advantage, showing that it can help concentrate similar data points on thin, distinct hyper-shells, enabling better separability.
- **Transformative Insight**: By leveraging contrast-loss, the approach turns this limitation into a powerful tool for organizing and clustering high-dimensional data.

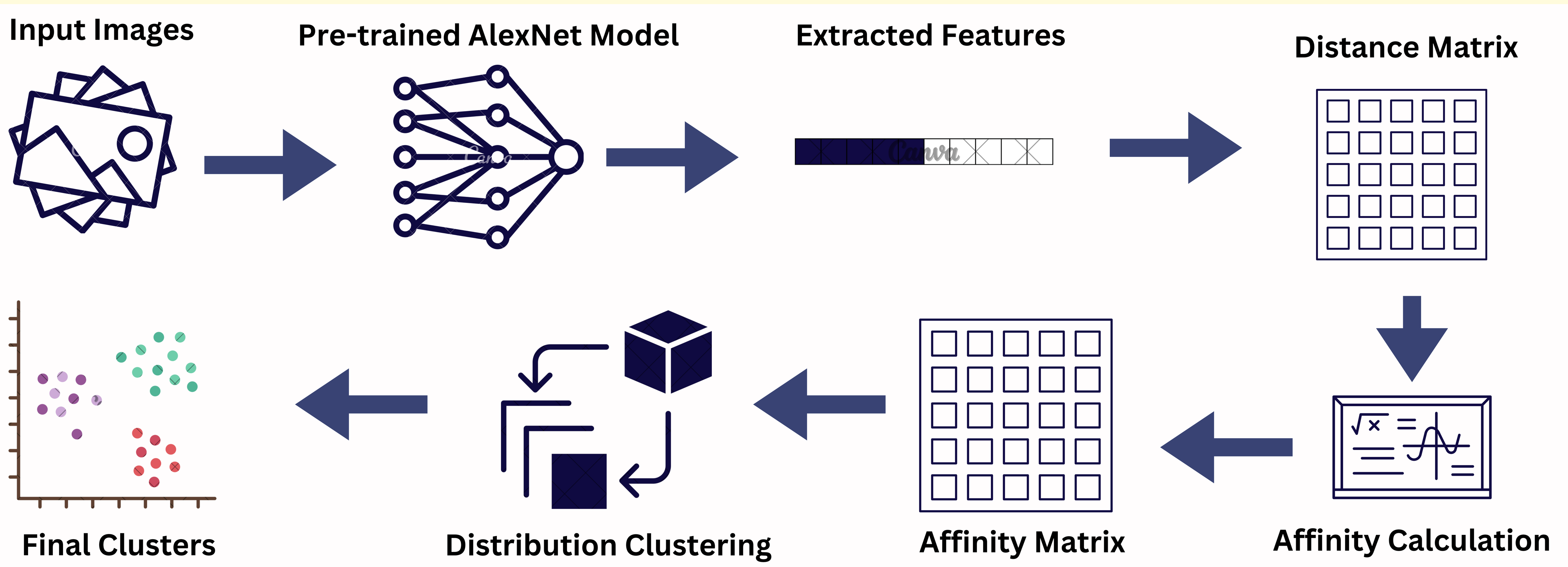
OBJECTIVE

- **Objective**: This research introduces a novel clustering algorithm, "Distribution-Clustering," designed to harness contrast-loss in high-dimensional spaces.
- **Features**: The algorithm automatically determines the number of clusters, groups data based on underlying distributions, and is robust to outliers.
- **Impact**: This approach enhances clustering accuracy, particularly for chaotic, high-dimensional datasets, by leveraging distribution-based similarities.

RELATED RESEARCH

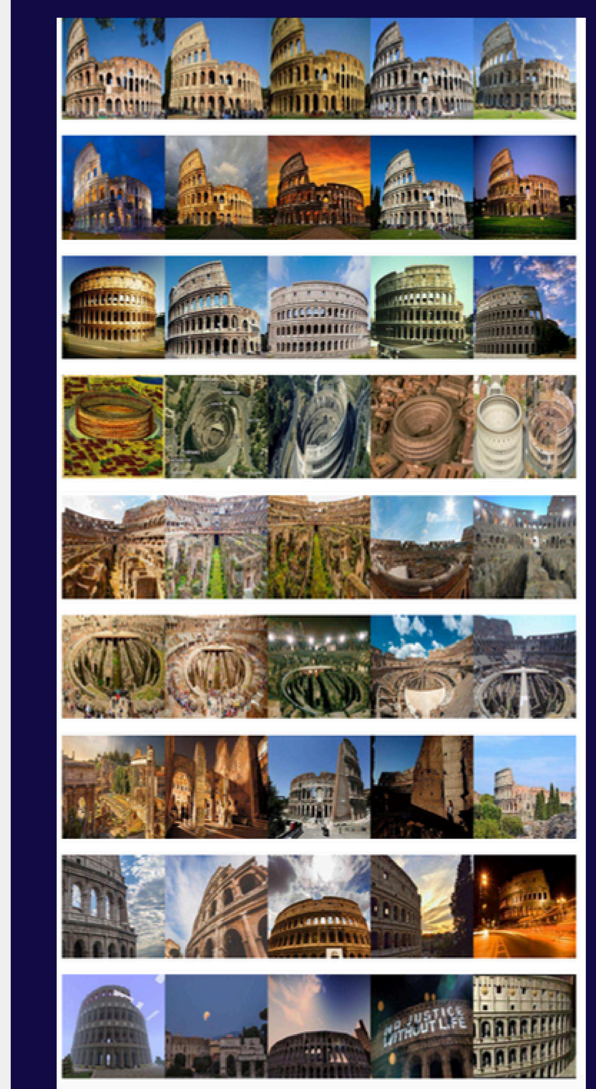
- **Existing Approaches:** High-dimensional clustering challenges have led to methods like sub-space and projective clustering, aimed at reducing contrast-loss effects.
- **Emerging Insights**: Recent studies indicate that contrast-loss might aid in cluster and outlier detection, presenting new advantages.
- **Contribution**: Building on Beyer et al.'s foundational work, this paper shows how contrast-loss can enhance natural data separability, providing a foundation for the proposed clustering algorithm.

METHODOLOGY



1. **Feature Extraction**: A pre-trained **AlexNet model** in the TestModel class extracts deep features for each image, converting them into one-dimensional vectors through adaptive pooling and flattening to capture essential characteristics.
2. **Distance Calculation**: The cluster function computes pairwise squared Euclidean distances between image feature vectors, laying the foundation for clustering by measuring the "**closeness**" of images in feature space.
3. **Affinity Matrix Computation**: Using these distances, the compute_affinity_matrix function calculates a second-order affinity matrix. This matrix reflects image similarities by computing and averaging the squared differences in distances of image pairs i and j to all other images, capturing second-order similarity in elements **A(i, j)**.
4. **Distribution-Clustering Algorithm**: 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 τ, and validates clusters that meet a minimum size, assigning unique labels.
5. **Outlier Handling**: Images that don't fit well into existing clusters are treated as outliers and placed separately, preserving the purity of primary clusters.
6. **Final Clustering Result**: This process repeats until all images are either clustered or marked as single-point outliers if they don't meet similarity criteria, resulting in clusters that group images based on their underlying distribution characteristics within the threshold tau.

RESULTS


Distribution-clustering

The clusttering results of Distribution clustering was compared with other known methods like **K-Means, GMM, Spectral clustering** on various performance metrices such tas **Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index**. Some sample results can be seen below on two given datasets, coloseum images and Mnist digit dataset are shown below.

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

PROPOSED ADVANCEMENT

- **FLD for Distance Calculation:** Fisher Linear Discriminant (FLD) replaces Euclidean distance in the distance matrix, projecting data into a space that maximizes inter-cluster separability and minimizes intra-cluster variance.
- **Initial Clusters:** The algorithm starts with initial clusters from a conventional method, then applies FLD to refine the clustering space.
- **Enhanced Affinity Matrix:** The FLD transformation yields a more accurate affinity matrix, improving the clustering process by better representing the true relationships between data points.

DATASET

Tests had been performed on a given unit dataset of 100 images alongwith the mentioned known datsets such as MNIST and Caltech 101 with multiple sets of images.

CONCLUSION

- **Novel Perspective on Contrast-Loss:** The paper reinterprets the traditionally negative contrast-loss phenomenon in high-dimensional clustering as a beneficial property, utilizing it in the Distribution-Clustering algorithm to group data based on distribution similarities.
- **Automatic Outlier Handling and Clean Clusters**: The algorithm naturally produces clean clusters with automatic outlier handling, enhancing its robustness in clustering tasks.
- **FLD Enhancement for Clustering:** The introduction of Fisher Linear Discriminant (FLD) refines the clustering process by improving cluster separability and robustness, particularly in complex, high-dimensional datasets. This results in enhanced accuracy and wider applicability to challenging datasets.