Exploring Clustering in CLIP Latent Space: K-Means and PCA on the Stimuli Images of the Natural Scenes Dataset:

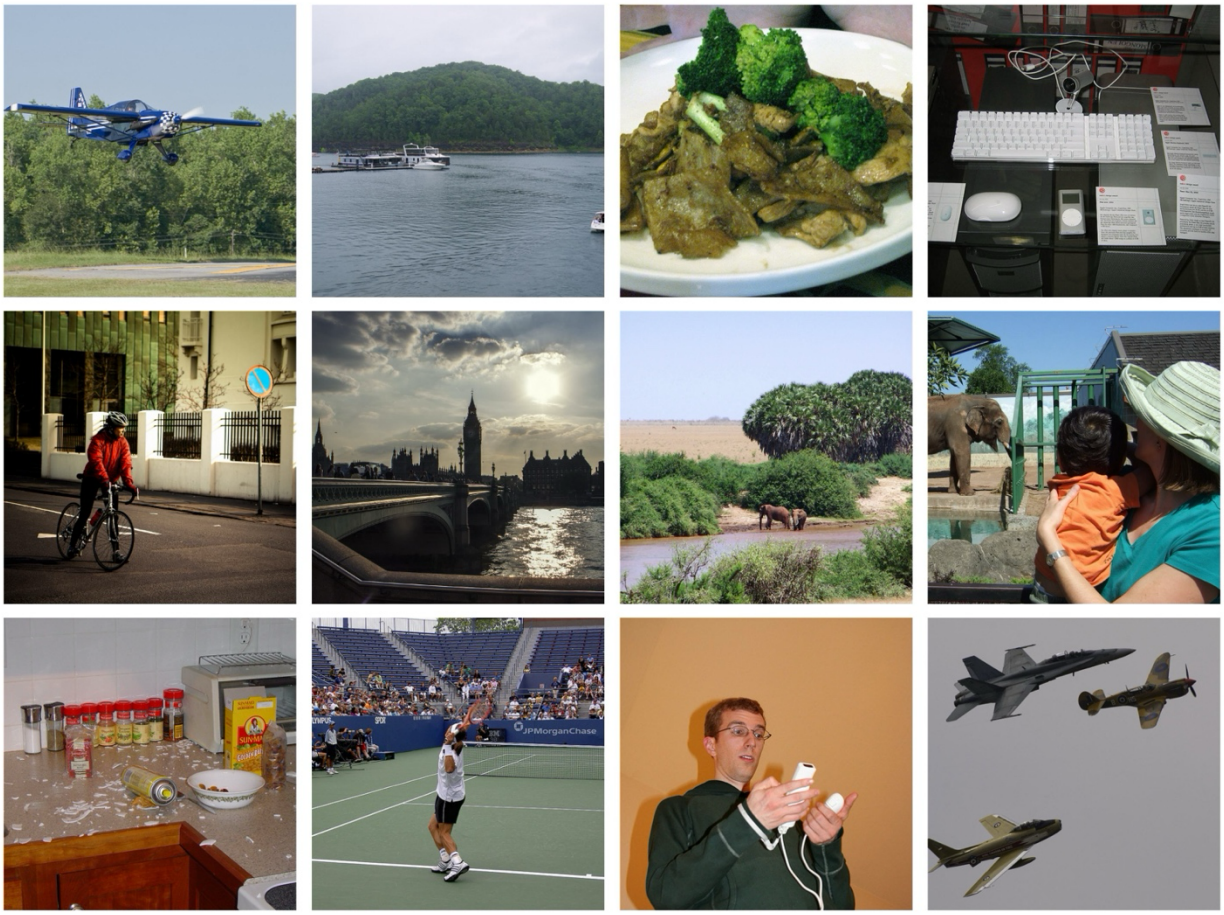


Figure 1. Twelve randomly selected images from the first 10.000 images of the Natural Scenes Dataset (Allen et al., 2022).

Abstract

This project applied k-means clustering to CLIP embeddings of the natural images from the Natural Scenes Dataset (NSD) to explore unsupervised machine learning techniques. Clustering the CLIP embeddings resulted in poorly defined clusters, even after scaling. Applying PCA improved dimensionality reduction, but silhouette scores remained low, reflecting overlapping

cluster boundaries. Visualization with t-SNE and UMAP provided improved insight into the structure of the embeddings. These findings highlight the challenges of clustering semantically rich, high-dimensional data and suggest future exploration with alternative clustering methods and qualitative analyses of decoded cluster group embeddings.

Introduction

The main objective of this project was to apply k-means clustering and Principal Component Analysis (PCA) to the CLIP latent space of visual images. This exploration was aimed at deepening my understanding of unsupervised machine learning methods, and evaluating the advantages and limitations, as well as to gain insights into the deployment of CLIP for feature extraction and the approach of mapping high-dimensional visual features into meaningful representations.

Inspired by the work of Takagi and Nishimoto (2023), the Natural Scenes Dataset (NSD) was selected for analysis due to its diversity of natural image stimuli, including objects, locations, faces, and more. A challenge arising from the dataset's rich diversity is that its underlying structure may not be easily or linearly separable. However, the NSD's relevance extends beyond just the stimuli images, as it includes 7T fMRI data from eight participants viewing this stimulus. (Allen et al., 2022).

For the project, the foundation was based on class material and resources provided in COGS 118B, complemented by "Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python" by Raschka et al. (2022). Additionally, I used large language models (LLMs) for assistance with repetitive code, debugging, and for help with

implementing advanced visualization techniques like t-SNE and UMAP, and to evaluate clustering performance using silhouette scores and inertia (via the elbow method).

This project was both engaging and a rich learning experience but did also prove to be challenging in terms of applying a centroid-based clustering algorithm, like k-means to a large and highly varied dataset like NSD, as well as including a large number of initial preprocessing steps.

Methods

The Natural Scenes Dataset (NSD), consisting of 73.000 natural image stimuli, was first loaded and inspected. To ensure computational efficiency, the dataset was divided into three subsets: small (100 images), medium (10.000 images), and large (20.000 images). This scalable approach allowed for a streamlined workflow simplifying debugging and iterative experimentation by initially applying methods on smaller subsets before scaling up.

The Contrastive Language-Image Pre-Training (CLIP) model was employed for feature extraction, generating embeddings for all subsets (Radford et al., 2021). Although CLIP applies input standardization and L2-normalization on its outputs, additional scaling using StandardScaler was tested to assess its impact on k-means clustering and PCA performance. Each of the subset datasets were therefore processed in both scaled and unscaled forms for comparative analysis.

For initial inspection, PCA and k-means was applied on the medium subset without CLIP, with example images from clusters then visualized. For exploratory visualization, T-SNE and UMAP were applied to the CLIP embeddings (scaled and unscaled) to evaluate clustering structures and overall data distribution. Thereafter, k-means clustering was applied to the embeddings without PCA using various cluster sizes (k = 3, 5, 10, 15, 25, 35, 50).

PCA was then applied to the CLIP embeddings to reduce dimensionality, first retaining 50 components and then preserving 95% of the explained variance. k-means clustering was then repeated on the PCA-reduced embeddings, and cluster visualizations were generated.

Additionally, T-SNE and UMAP were applied to the PCA-reduced embeddings for further analysis. This was further grouped by k-means clustering and then visualized with various cluster sizes (k = 3, 5, 8, 15, 25, 35, 50, 80), with silhouette scores calculated for each cluster size. At last, to evaluate clustering performance, silhouette scores and inertia (via the elbow method) were computed for the CLIP embeddings with PCA.

This pipeline provided a comprehensive exploration of the clustering and visualization strategies for the NSD image stimuli.

Results

For initial inspection, PCA and k-means were applied directly to the medium subset without using CLIP embeddings. Example images from selected clusters were visualized, and some clusters demonstrated partial qualitative coherence in their grouping. The clusters compositions reflect the algorithms ability to group visually similar elements. However, the results also reveal

the challenges of achieving meaningful clusters without using additional feature extraction methods like CLIP.

Figure 2. Cluster 26 examples with brown, beige, and circular objects, including a bear and two teddy bear-pictures.

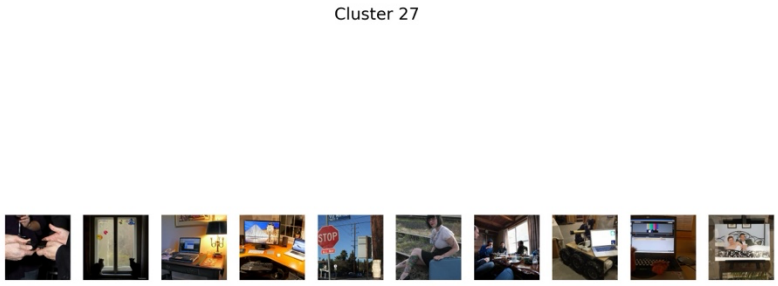


Figure. 3. Cluster 27 examples featuring dark images with central light elements.

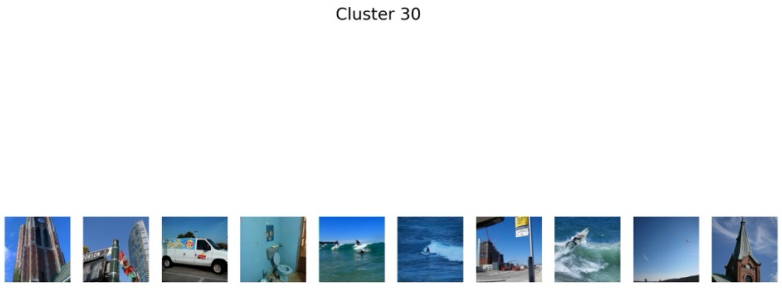
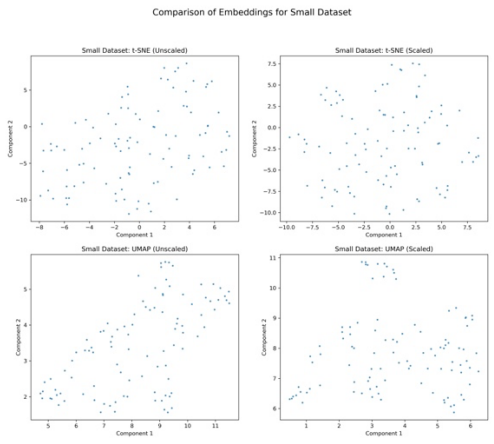
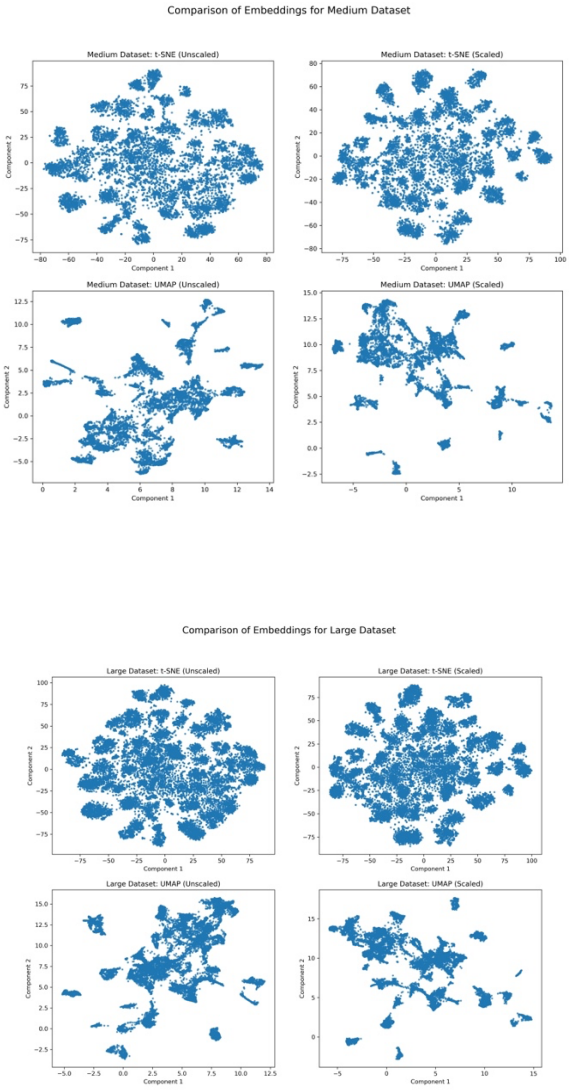


Figure 4. Cluster 30 examples highlighting blue-themed images.

t-SNE and UMAP visualization

The t-SNE and UMAP visualizations were able to find some underlaying structure in the dataset, however with increasing dataset size the overlap between clusters seem to increase, making separability more challenging. Scaling did slightly improve clustering performance of the larger datasets, but it did not notably solve issues with overlap between clusters.





(Figure 5-7. t-SNE and UMAP visualizations for the small, medium, and large datasets. Top row: Small dataset visualizations (unscaled and scaled). Middle row: Medium dataset visualizations (unscaled and scaled). Bottom row: Large dataset visualizations (unscaled and scaled).)

K-means on CLIP embeddings without PCA

The first step of our main objective was applying k-means directly on the CLIP embeddings without PCA. As seen in figure 8-10, this resulted in poorly separated clusters with significant overlap for both unscaled and scaled embeddings, indicating that the high-dimensional structure of the embeddings from the NSD-data does not naturally support effective clustering, at least not using centroid-based methods.

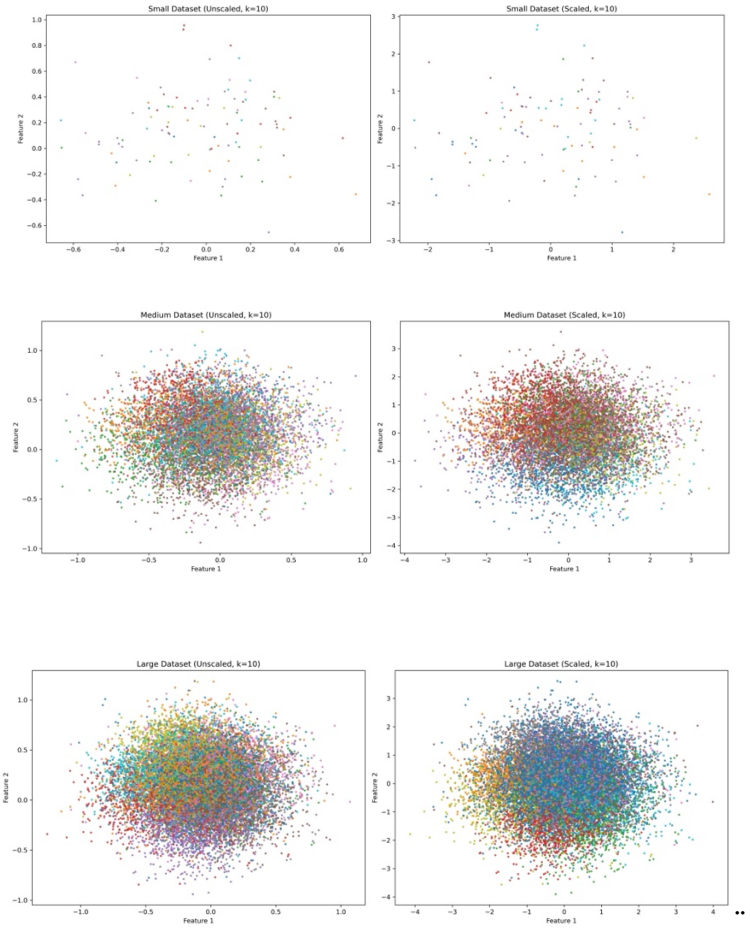


Figure 8-10. K-means clustering results for the small, medium, and large dataset (k=10) with unscaled and scaled embeddings.

The highest silhouette scores for k-means clustering on the CLIP embeddings without PCA were observed for the small dataset with unscaled embeddings (k=35, score: 0.095) and scaled embeddings (k=35, score: 0.080). For the medium dataset, the highest scores were unscaled (k=25, score: 0.072) and scaled (k=35, score: 0.072). Similarly, the large dataset showed the highest scores for unscaled embeddings (k=10, score: 0.072) and scaled embeddings (k=50, score: 0.070). Overall, these low silhouette scores across all datasets and cluster sizes indicate poor cluster separation and quality when applying k-means directly on the CLIP embeddings.

K-means on PCA-reduced CLIP embeddings (50 components)

Unscaled Embeddings

PCA was applied to the unscaled embeddings, which demonstrated slightly better cluster separability compared to clustering without PCA. With 50 components, PCA explained 85.73%, 62.83%, and 62.83% of the variance for small, medium, and large datasets. The explained variance from PCA retained much of the dataset's variability, but the silhouette scores remained relatively low (e.g., Small Dataset: 0.14 for 25 clusters). This suggests that while PCA reduced dimensionality effectively, the structure of the data was still not very susceptible to centroid-based clustering methods.

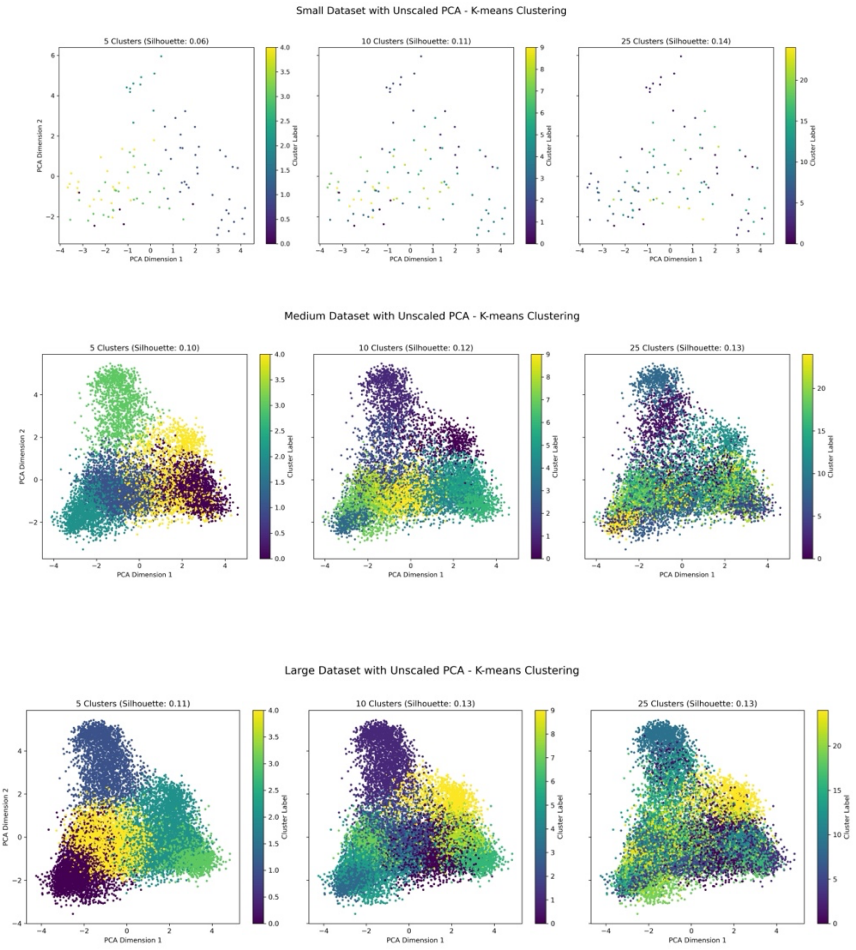


Figure 11-13. Unscaled PCA on CLIP embeddings for small, medium, and large dataset (k=5, 10, 25).

Scaled Embeddings

For the scaled embeddings, PCA and k-means clustering showed some improvements in silhouette scores, but the clustering did not exhibit considerably better separability. For scaled embeddings, explained variance dropped to 81.02%, 50.37%, and 50.20%. However, scaling before PCA gave a slight increase in silhouette scores for some datasets (e.g., Medium Dataset: 0.14, and Large Dataset: 0.16 for 25 clusters). Still, even with scaling, the clusters often overlap, indicating noise or complexity in the data distribution.

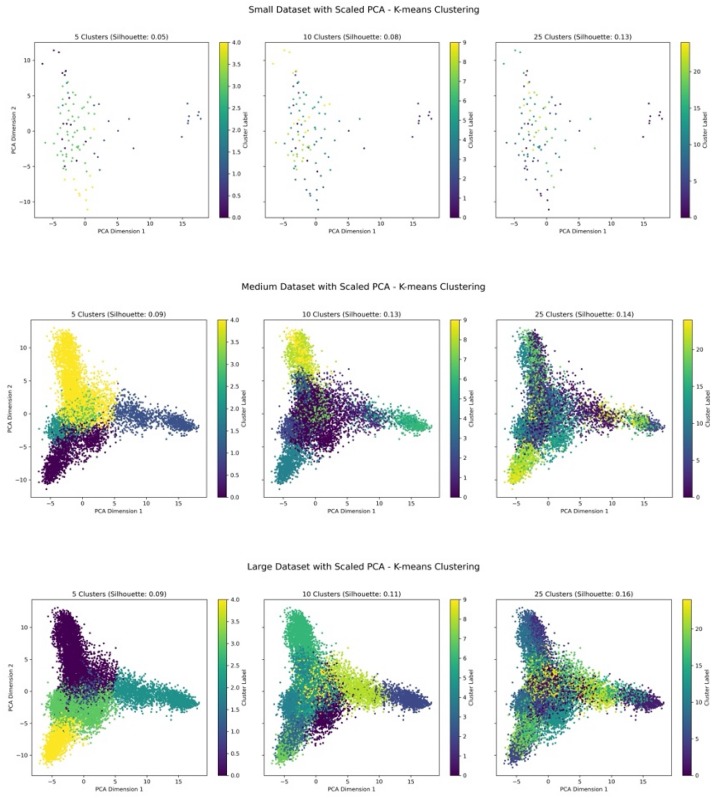


Figure 14-16. Scaled PCA on CLIP embeddings for small, medium, and large dataset (k=5, 10, 25).

PCA Results (for 95% Variance)

To retain 95% of the variance in the unscaled embeddings, PCA required 73 components for the small dataset, 296 components for the medium dataset, and 299 components for the large dataset. For the scaled embeddings, slightly more components were needed to achieve the same level of variance retention: 78 components for the small dataset, 324 components for the medium dataset, and 326 components for the large dataset.

t-SNE on the reduced components (PCA) of the CLIP embeddings

t-SNE was applied to PCA-reduced CLIP embeddings, followed by k-means clustering. The highest silhouette scores were observed with 35 clusters for the small dataset (0.44), medium dataset (0.48), and large dataset (0.46). These results indicate improved cluster separability when compared to k-means (with and without PCA) on the embeddings, highlighting the benefits of dimensionality reduction. PCA and k-means rely on linear transformations and centroid-based methods, while t-SNE captures non-linear relationships, which may be better for revealing underlying structures in the data. However, the clustering performance does remain limited and dependent on dataset size.

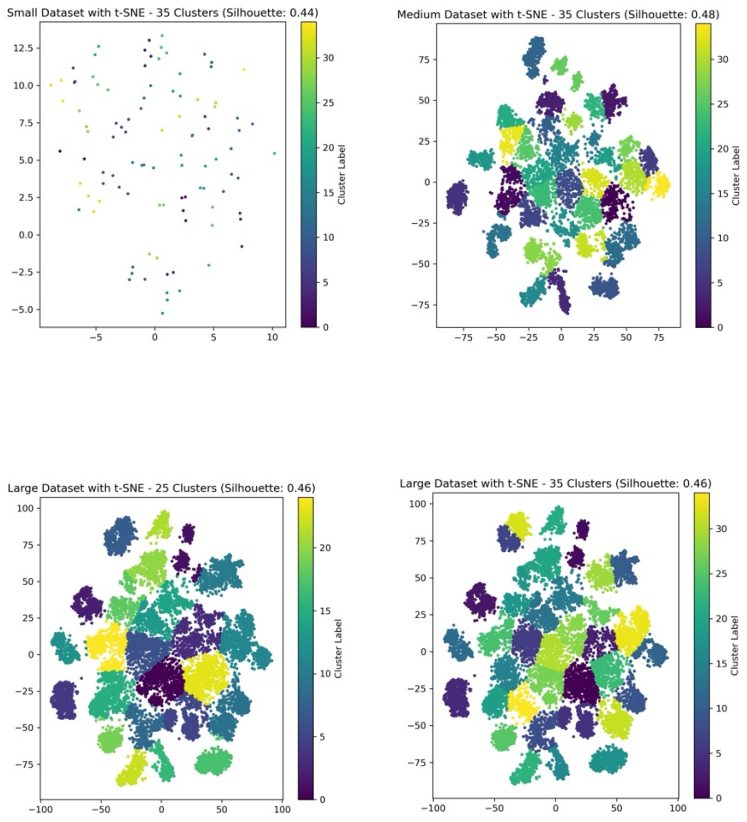


Figure 17-20. t-SNE visualizations of PCA reduced CLIP embeddings for small, medium, and large datasets with k-means clustering applied.

Silhouette Scores and Inertia (Elbow Method) for the CLIP embeddings with PCA

Silhouette scores and inertia (using the elbow method) were analyzed for k-means clustering on the PCA-reduced CLIP embeddings. The silhouette scores peaked at 35 clusters for the small and medium datasets (0.14 and 0.13) and at 25 clusters for the large dataset (0.13). Inertia values consistently decreased as the number of clusters increased, reflecting improved within-cluster compactness. However, the decreasing silhouette scores with higher cluster numbers indicate that the cluster separability is limited despite the PCA's dimensionality reduction. Also, the elbow points suggest that the optimal number of clusters is around 35 for the small and medium datasets and 25 for the large dataset.

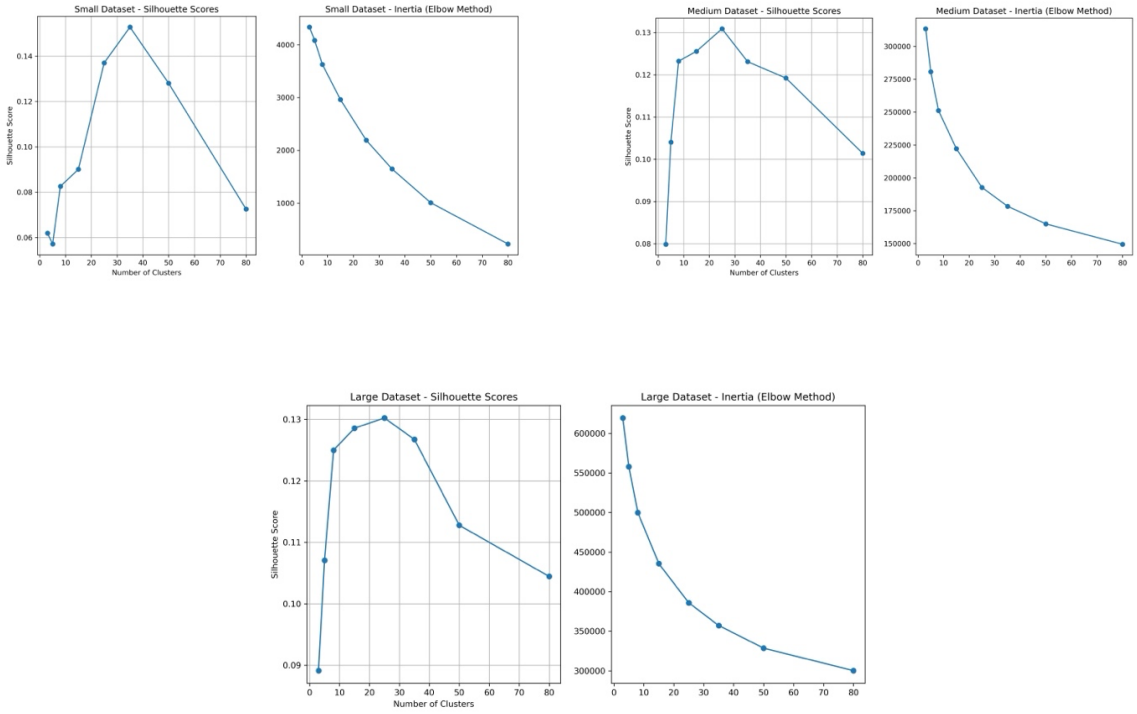


Figure 21-23. Silhouette Scores and Inertia (Elbow Method) for the CLIP embeddings with PCA.

Discussion

Clustering the Natural Scenes Dataset (NSD) embeddings was challenging due to its diverse and complex nature. While a simpler dataset might have provided clearer results, this analysis effectively highlighted the strengths and limitations of k-means and PCA for this type of data. The dataset consists of a wide range of natural images, from objects to faces to scenes, with large variations in visual and semantic features. While CLIP embeddings capture rich semantic information, the use of centroid-based clustering methods like k-means revealed clear limitations. The results showed that without dimensionality reduction, k-means struggled to create well-defined clusters, leading to poor silhouette scores. There was significant overlap between clusters, indicating that the embeddings' structure does not effectively support direct clustering.

Applying PCA improved the dimensionality reduction, as shown by explained variance scores above 80% for the small dataset. However, even with PCA, the results were weak, with silhouette scores still around 0.09 for the best-performing clustering.

Visualization methods like t-SNE and UMAP offered clearer insights into the structure of the embeddings compared to the other approaches. These techniques revealed improved cluster separability when compared to k-means clustering with and without PCA. While t-SNE and UMAP are primarily designed for dimensionality reduction and visualization rather than clustering optimization, they did contribute to highlight some of the underlying patterns in the NSD data that were not as apparent with linear and centroid-based methods.

Scaling the CLIP embeddings using StandardScaler had little to no impact on clustering performance. This suggests that CLIP’s inherent normalization pipeline already provides

embeddings that are sufficiently scaled. However, scaling did increase the number of components necessary for 95% variance retention, highlighting some of the difference in variance distribution between the scaled and unscaled data.

For future exploration, alternative clustering methods such as DBSCAN or hierarchical clustering might be more suitable to capture the structure of the CLIP embeddings. Additionally, qualitative evaluation of cluster centroids could provide insights into whether certain clusters correspond to distinct image features, such as textures or objects. This approach might also reveal clustering patterns that are meaningful to humans but not necessarily captured by statistical methods or similarity metrics, possibly resulting in lower quantitative scores despite their interpretive value. Unfortunately, this type of analysis could not be conducted due to time constraints. Lastly, combining non-linear dimensionality reduction methods like t-SNE or UMAP with density-based clustering could offer a more dynamic and suitable approach to the NSD embeddings.

Conclusion

This project explored the application of k-means clustering on CLIP embeddings generated from the natural images of the Natural Scenes Dataset (NSD), with and without dimensionality reduction using PCA. The results highlighted the challenges of clustering high-dimensional, diverse and complex data such as the NSD. The application of k-means to the CLIP embeddings resulted in poorly defined clusters, shown by low silhouette scores, underscoring its limitations for this type of data.

While PCA was able to reduce the dimensionality of the data, retaining over 95% of the variance in both unscaled and scaled embeddings, the clustering performance remained limited. Visualizations using t-SNE and UMAP revealed a dense and overlapping structure in the embeddings. Additionally, scaling demonstrated limited impact on clustering performance, which shows the robustness of CLIP's internal normalization pipeline.

These findings emphasize the limitations of linear and centroid-based methods for complex and semantically rich datasets like NSD. Future work should focus on exploring alternative methods, like density-based clustering or hierarchical approaches aimed at better capturing the relationships within the embeddings. Additionally, qualitative analysis of the cluster contents and also the use of non-linear dimensionality reduction methods together with clustering may provide new insights into the structure of the CLIP latent space.

This special project was a valuable learning experience through an exploration of the challenges associated with applying unsupervised machine learning techniques to high-dimensional data. While the complexity of the NSD data made it difficult to achieve meaningful clustering results, the project altogether, offered unique insights and learning opportunities that a simpler dataset would not have provided.

References:

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. N. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience, 25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

CLIP: Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Amodei, D. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2103.00020

Raschka et al.: Raschka, S., Liu, Y., & Mirjalili, V. (2022). *Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing.

Takagi & Nishimoto: Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. *bioRxiv*. https://doi.org/10.1101/2022.11.18.517004