Final Project

Team members: Karthikeya Gummadi
Sannith Reddy Gunreddy
Srinith Rao Bichinepally

Link to code: https://github.com/karthikeya9296/CLIP-Image-Embeddings-Analysis

Evaluating the Efficacy of CLIP Model for Semantic Clustering of Randomly Selected Images from the Google Open Images Dataset

In this project, we investigate the performance of the CLIP (Contrastive Language-Image Pretraining) model in clustering images based on their semantic similarities. Using a randomly selected set of images from the Google Open Images dataset(We used 1,000 images from the Google image database. Initially, we attempted to upload them to our GitHub repository, but the file size was too large. Consequently, we opted to upload only a selection of these images.), we explore the model's ability to generate meaningful image embeddings and effectively cluster these embeddings using t-SNE and K-means algorithms. Our analysis focuses on the adaptability of the CLIP model to diverse and non-categorical visual data, assessing its potential applications in fields that require robust image understanding and organization.

1. Introduction:

The integration of vision and language models has revolutionized tasks requiring semantic understanding of images. The CLIP model by OpenAI, designed to comprehend and link images with textual descriptions, presents a promising avenue for semantic image clustering. In this project, we aim to test the model's versatility across a non-curated assortment of images, reflecting a realistic scenario where image data may not always be pre-sorted or labeled.

2. Problem Statement and Objectives:

Our primary objective is to evaluate how well the CLIP model performs in clustering a random set of images from the Google Open Images dataset. Through this evaluation, we aim to understand the model's capability to handle diverse visual information and to ascertain its utility in practical, real-world applications where image categorization is crucial yet challenging.

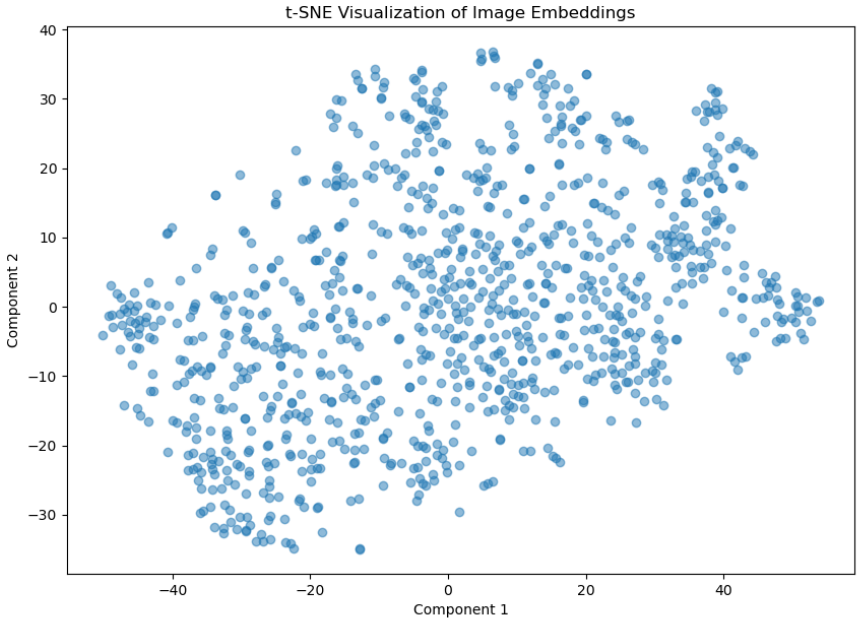
3. Methodology:

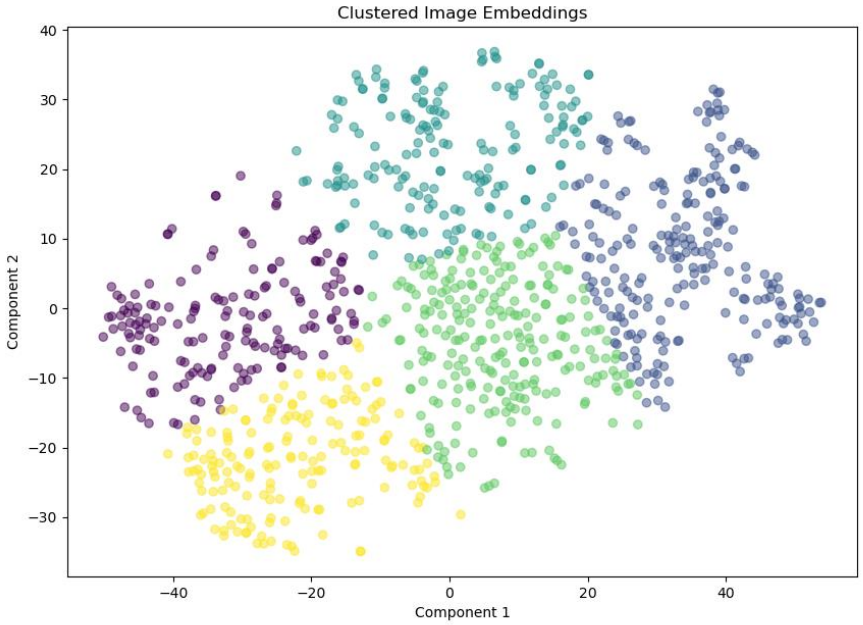
Data Collection: We used a randomly selected collection of images from the Google Open Images dataset. This selection was intended to simulate a realistic use-case scenario for the application of image clustering technologies.

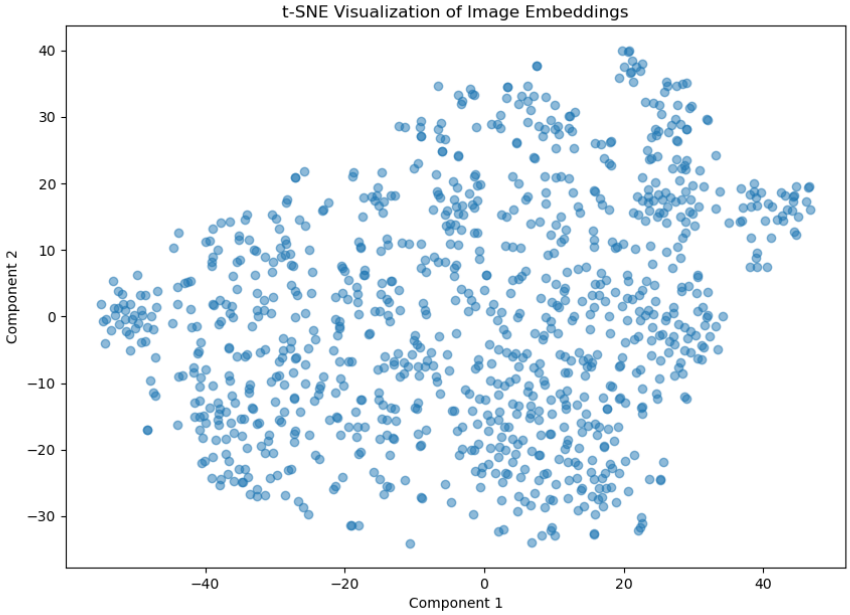
Feature Extraction: We utilized the CLIP model to generate image embeddings, capturing the high-dimensional semantic features of each image.

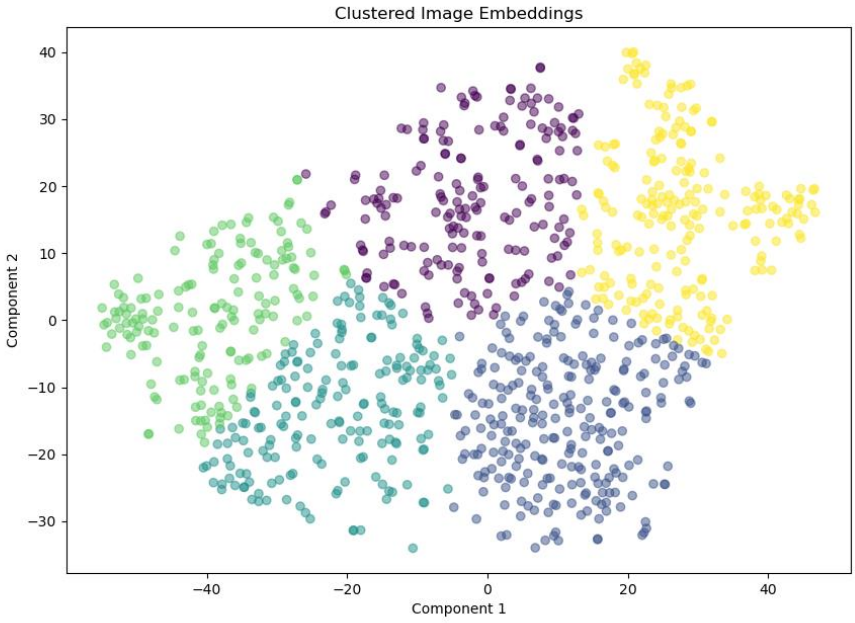
Dimensionality Reduction: We applied t-SNE to project these high-dimensional embeddings into a two-dimensional space for visualization and further analysis.
Clustering Analysis: We employed K-means clustering to group the images based on the proximity of their embeddings in the reduced space.
Evaluation: We calculated the silhouette score for each cluster to evaluate the quality and effectiveness of the clustering.

4. Results:
The t-SNE visualization indicated a varied distribution of images, with clusters forming around semantically similar content despite the random selection. However, the silhouette scores suggest that while some clusters were well-defined, others were less cohesive, indicating potential limitations of the model when dealing with highly diverse datasets.


t-SNE Visualization of Image Embeddings
Component 2
Component 1
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
40


Clustered Image Embeddings
Component 2
Component 1
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
40





5. Discussion:

Interpretation of Results: Our results showcase the CLIP model's strengths in identifying and grouping semantic similarities but also highlight challenges when applied to random and diverse image sets.

Limitations: Our study is limited by the randomness of the image selection, which might not provide a balanced representation of all possible image categories.

Future Work: We believe that future research could explore the impact of curated vs. random datasets on the model's performance and the integration of additional clustering algorithms to enhance cluster quality.

6. Conclusion:

This project confirms the capability of the CLIP model to cluster images semantically, even within a non-curated dataset. However, the varying quality of clustering underscores the need for further refinement in model application and dataset preparation for achieving optimal results in practical scenarios.