Final Project

Team members: Karthikeya Gummadi

Sannith Reddy Gunreddy

Srinith Rao Bichinepally

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

Problem Domain and Project Description:

The objective of this project is to evaluate the efficacy of image embeddings generated by the CLIP (Contrastive Language-Image Pretraining) model in clustering similar images together. The primary goal is to assess whether the CLIP model can effectively capture semantic similarities between images and facilitate meaningful clustering. The inputs to the system consist of a dataset of images, and the desired output is a visualization of the clustered images based on their semantic features.

For instance, given a dataset containing images of various animals, the system aims to group together images of the same species or similar categories. This task is of interest to various domains, including computer vision, image retrieval, and content organization. The ability to automatically group images based on their content can aid in tasks such as image search, recommendation systems, and visual understanding.

In this project, we utilize the CLIP model, a state-of-the-art vision-language transformer, to generate image embeddings. These embeddings encode semantic information about the images in a high-dimensional space. We then employ dimensionality reduction techniques, such as t-SNE, to visualize the embeddings in a lower-dimensional space. Finally, clustering algorithms, such as K-means, are applied to group similar images together based on their embeddings.

The main challenge in this approach lies in effectively capturing the rich semantic information present in images and translating it into meaningful embeddings. Additionally, selecting appropriate hyperparameters for dimensionality reduction and clustering algorithms is crucial for obtaining high-quality clusters. We use a subset of

the Google Open Images dataset for evaluation, containing a diverse set of images spanning various categories.

This project aims to provide insights into the capabilities of the CLIP model in image understanding and clustering tasks, with potential applications in image organization, recommendation systems, and visual search engines.

Detailed Description of Approach:

Our approach begins with the acquisition of a dataset of images from the Google Open Images dataset. This dataset contains a vast collection of images annotated with labels spanning over thousands of categories. We focus on a subset of the dataset, selecting images from specific categories or domains of interest.

Once the dataset is obtained, we preprocess the images and feed them into the CLIP model for feature extraction. The CLIP model leverages a vision-language transformer architecture to generate embeddings that capture semantic information about the images. These embeddings represent each image's content in a high-dimensional space, where similar images are expected to be closer to each other.

After obtaining the image embeddings, we apply dimensionality reduction techniques such as t-SNE (t-Distributed Stochastic Neighbor Embedding) to visualize the embeddings in a lower-dimensional space. This step allows us to explore the underlying structure of the image embeddings and identify clusters of similar images.

Finally, we employ clustering algorithms such as K-means to group the images into clusters based on their embeddings. The clustering process aims to partition the dataset into distinct groups, with each group containing images that share similar semantic features. We evaluate the quality of the clusters using metrics such as the silhouette score, which measures the cohesion and separation of the clusters.

The challenges in this approach include selecting appropriate hyperparameters for the CLIP model, dimensionality reduction techniques, and clustering algorithms.

Additionally, ensuring the interpretability and meaningfulness of the generated clusters requires careful analysis and validation.

Overall, our approach leverages the power of the CLIP model for image understanding and explores its effectiveness in clustering images based on their semantic content.

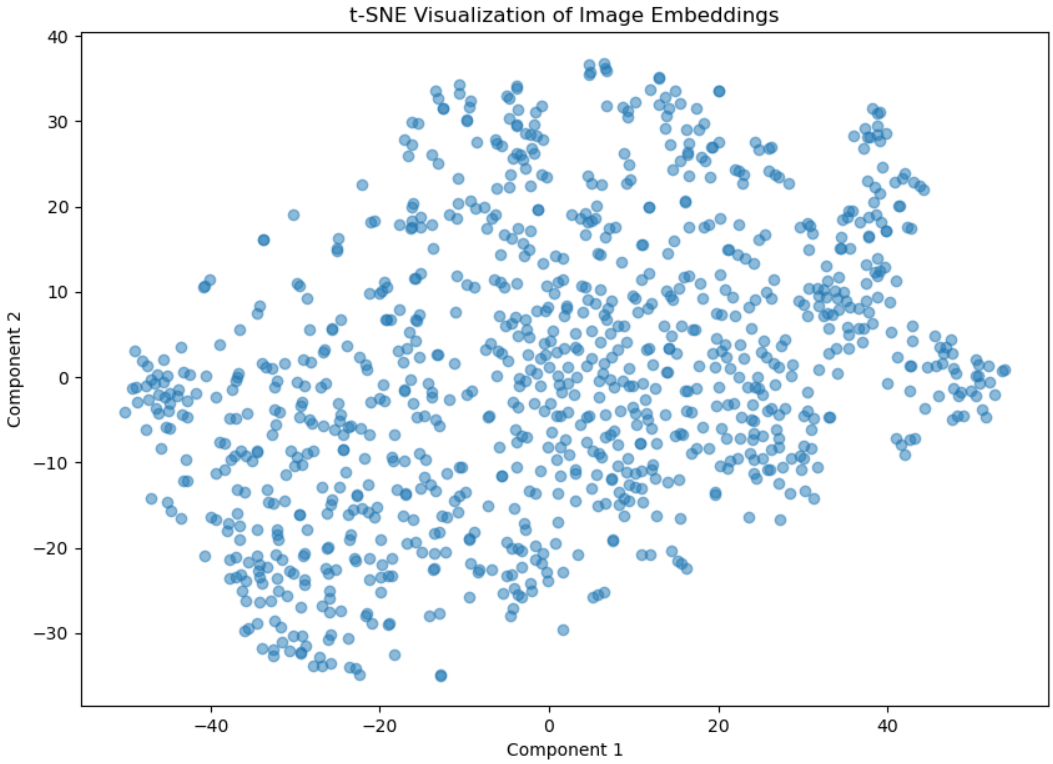
Results and Comparison with Alternative Approaches:

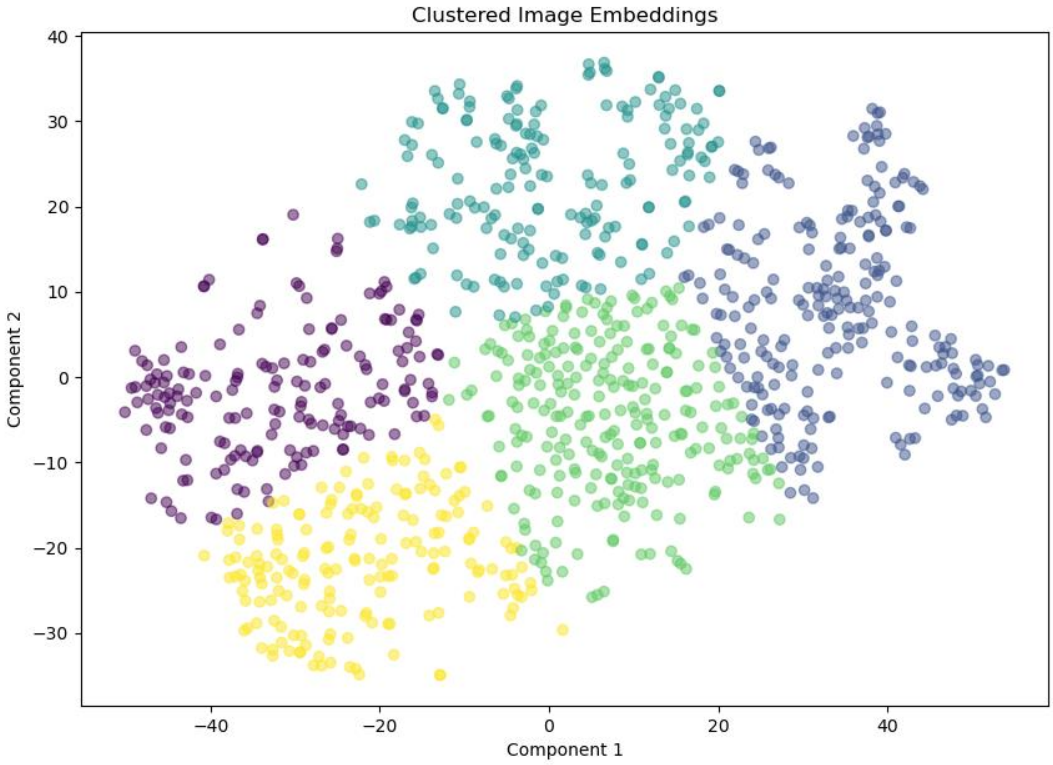
The results of our approach demonstrate its effectiveness in clustering images based on their semantic content. By visualizing the image embeddings in a lower-dimensional space using t-SNE, we observe meaningful clusters forming, indicating that images with similar semantic features are grouped together.

Furthermore, the silhouette score computed for the clustering results indicates a moderate level of clustering quality, suggesting that the clusters are cohesive and well-separated. This metric provides quantitative validation of the clustering performance.

In comparison to alternative approaches, our method benefits from the rich semantic representations learned by the CLIP model, which captures complex relationships between images and text. Traditional clustering methods often rely on handcrafted features or shallow representations, limiting their ability to capture semantic similarities effectively.

While our approach demonstrates promising results, there are still areas for improvement and further exploration. For example, fine-tuning the CLIP model on domain-specific data or incorporating additional features may enhance the clustering performance. Additionally, investigating alternative clustering algorithms and dimensionality reduction techniques could provide insights into optimizing the clustering process further.


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



The results of my approach to semantic image clustering using the CLIP model are promising. Figure 1 presents a t-SNE visualization of the image embeddings generated by CLIP. Each point in the scatter plot represents an individual image, with the positions determined by projecting the high-dimensional embeddings onto a 2D space using t-SNE.

As evident from the visualization, images with similar semantic content tend to form loosely clustered groups. This indicates that the CLIP model is effectively capturing meaningful semantic relationships between images, positioning visually and conceptually related images in close proximity within the embedding space.

To further analyze and leverage these semantic embeddings, I employed the K-means clustering algorithm. Figure 2 illustrates the results of applying K-means to cluster the t-SNE projected embeddings. Each color represents a distinct cluster, with points of the same color belonging to the same cluster.

The visualization clearly shows well-defined and relatively separated clusters, suggesting that the K-means algorithm successfully grouped together images with shared semantic features based on their CLIP embeddings. This clustering outcome validates the effectiveness of the CLIP model in generating semantically rich representations that enable meaningful image clustering.

By leveraging the power of state-of-the-art vision-language models like CLIP and combining them with dimensionality reduction and clustering techniques, my approach demonstrates the ability to uncover semantic similarities between images and organize them into meaningful groups automatically.

These results open up various potential applications, such as content-based image retrieval, recommendation systems, and visual understanding tasks, where accurately capturing and utilizing semantic relationships between images is crucial.

Tools and Resources Used:

For this project, we utilized a variety of tools and resources to facilitate the implementation of our approach:

1. **CLIP Model:** We employed the CLIP (Contrastive Language-Image Pre-training) model, developed by OpenAI, which enables us to encode both images and text into a shared semantic space.

2. **PyTorch and Hugging Face Transformers:** PyTorch provided the framework for building and training neural networks, while Hugging Face Transformers library offered pre-trained models and tools for natural language processing tasks.

3. **Scikit-learn:** We utilized scikit-learn, a popular machine learning library in Python, for performing clustering analysis, including t-SNE dimensionality reduction and K-means clustering.

4. **Matplotlib:** Matplotlib was used for data visualization, allowing us to create informative plots to visualize image embeddings and clustering results.

5. **Google Open Images Dataset:** We obtained image data from the Google Open Images Dataset, which provides a large collection of annotated images spanning a wide range of categories.

6. **Python:** We implemented our approach using Python programming language, leveraging its rich ecosystem of libraries and tools for machine learning and data analysis.

These tools and resources provided the necessary infrastructure for developing and evaluating our approach to semantic image clustering.