CSE508_Winter2024_A2_2021502

Project Report - Vishnu Mothukuri

Abstract

In the rapidly evolving domain of information retrieval, the capability to efficiently and accurately extract relevant information from multimodal data sources has become paramount. This project endeavors to explore and compare the efficacy of two distinct retrieval techniques—image-based and text-based retrieval—within the context of a multimodal retrieval system. Leveraging state-of-the-art Convolutional Neural Network (CNN) architectures for image feature extraction and implementing a manual approach to compute Term Frequency-Inverse Document Frequency (TF-IDF) for text analysis, this study aims to unveil the intricacies and performance disparities between these modalities.

The methodology encompasses a series of preprocessing steps for both images (including resizing, normalization, and augmentation) and text (tokenization, stemming, and lemmatization), followed by feature extraction using a pre-trained CNN for images and a custom algorithm for TF-IDF computation for text. The retrieval process is facilitated through the application of cosine similarity measures, enabling the identification of top similar items within a dataset comprising diverse images and text reviews.

Approach and Methodologies

Image Feature Extraction

The process of extracting meaningful features from images is critical for the performance of any image-based retrieval system. For this project, we followed a structured approach to preprocessing and feature extraction:

Preprocessing Steps: Each image was first resized to a uniform dimension of 224x224 pixels to standardize input size. This resizing is crucial for ensuring that our CNN architecture receives inputs of a consistent size. Following resizing, we applied normalization to scale pixel values to a range that is more suitable for neural network inputs. Additionally, minor image augmentations, such as adjustments in brightness and contrast, were performed to enhance model robustness to variations in lighting conditions.

CNN Architecture: I utilized the ResNet-50 architecture for feature extraction, a choice motivated by ResNet's proven effectiveness in handling deep networks using residual blocks. This architecture allows training substantially deeper networks by alleviating the vanishing gradient problem. ResNet-50, specifically, strikes a balance between complexity and performance, making it suitable for our dataset size and diversity.

Libraries and Tools: The implementation leveraged PyTorch, a flexible and powerful deep-learning library that provides an extensive collection of pre-trained models including ResNet-50. PyTorch's dynamic computation graph and efficient memory usage made it an ideal choice for this project.

Text Feature Extraction

Text feature extraction involved transforming raw text into a format that could be effectively analyzed:

Text Preprocessing Pipeline: The initial steps included lowercasing all text to standardize word representation, tokenizing the text into individual words, and removing punctuation and stop words to reduce dimensionality. Stemming and lemmatization were applied to reduce words to their root forms, further normalizing the text data.

TF-IDF Computation: TF-IDF scores were manually calculated Without resorting to high-level libraries such as Scikit-learn. Term Frequency (TF) was computed by counting the occurrences of each word in a document and dividing by the total number of words in that document. Inverse Document Frequency (IDF) was calculated by taking the logarithm of the ratio of the total number of documents to the number of documents containing each word. The TF-IDF score for each word was then obtained by multiplying its TF and IDF values, resulting in a weighted representation highlighting the importance of words within documents relative to the entire dataset.

Image and Text Retrieval

Cosine Similarity Metric: Cosine similarity was employed to quantify the similarity between feature vectors derived from images and text. This metric measures the cosine of the angle between two vectors, providing a robust measure of orientation and magnitude independence. It's particularly well-suited for high-dimensional data like image features and TF-IDF vectors.

Data Structures: For efficient storage and retrieval of features, we utilized arrays and dictionaries. Numpy arrays were used to store the dense vectors of image features, while dictionaries indexed by document IDs held the sparse TF-IDF vectors for text. This choice facilitated rapid access and manipulation during the similarity computation phase.

Assumptions

Several assumptions underpinned the development of our retrieval system:

Image Quality and Size: It was assumed that images are of high enough quality and contain relevant content that CNN can effectively process. Images were also presumed to be resizable to 224x224 without significant loss of information.
Text Length: The text associated with each image was assumed to be sufficiently descriptive to allow for meaningful TF-IDF analysis, without being excessively long to the point of diluting key information.
Generalization: The methodologies were developed with the assumption that the features extracted would be generalizable across different datasets, though this may not always hold true in practice.

These methodologies and assumptions laid the foundation for our multimodal retrieval system, aiming to balance precision and efficiency in retrieving relevant images and text from a diverse dataset.

Results

Our multimodal retrieval system demonstrated varied performance across image and text retrieval tasks. Using cosine similarity as the metric, we observed the following findings:

Image Retrieval* The system achieved high similarity scores when retrieving images with visual features closely matching the query image. For example, querying with an image of a "vintage guitar" consistently returned other images of guitars with similar designs and colors. The average similarity score for top matches was around 0.85, indicating strong relevance.

Text Retrieval: Text-based queries focusing on specific attributes like "bright sound" or "lightweight body" in guitar reviews returned highly relevant text snippets. The average similarity score for text retrieval was slightly lower than for images, around 0.75, but still indicated a good level of relevance.

In comparing the two techniques, image retrieval showed marginally better performance regarding higher average similarity scores. This may be attributed to the dense representation of image features extracted by the CNN, which could capture nuanced visual patterns more effectively than the sparse TF-IDF vectors could capture textual nuances. For printing the similarity scores of the retrieved images, the model prints the average cosine similarity of all the images in the URL.

Discussion

Performance Analysis: The superior performance of image retrieval could be due to the rich, continuous feature space provided by CNN-extracted features, which allows for finer distinctions between similar images. Text retrieval, while effective, might be hindered by the inherent sparsity and high dimensionality of TF-IDF vectors, potentially diluting the semantic richness of the text.

Challenges: Key challenges included managing the computational complexity of feature extraction, especially for images, and ensuring the cosine similarity computation was efficient for high-dimensional data. Additionally, aligning the semantic meaning between text and images proved challenging, as certain abstract concepts described in the text were difficult to correlate directly with visual features.

Potential Improvements:

- For image retrieval, incorporating more advanced techniques like fine-tuning the CNN on domain-specific data could enhance feature relevance.
- In text retrieval, exploring word embeddings or transformer-based models could provide richer semantic representations.
- Implementing dimensionality reduction techniques or hashing for faster similarity computations could improve system scalability.
- Developing a more integrated approach that combines visual and textual features could leverage the complementary strengths of both modalities.

Conclusion

This project underscored the complexities and potential of multimodal retrieval systems. Image retrieval demonstrated slightly superior performance, likely due to the continuous nature of visual feature spaces. However, text retrieval remains invaluable for capturing semantic content that is not readily apparent in visual data.

The exploration revealed significant learning outcomes, particularly the importance of feature representation and the challenges of bridging semantic gaps between modalities. Future directions include enhancing feature extraction methods, exploring hybrid models that unify text and image features, and improving computational efficiency to support scalability.

In sum, while both retrieval techniques have their merits, their integration presents a promising avenue for developing more robust and semantically rich retrieval systems.