CA6005 Assignment 2 – Image retrieval using transformers and vector support machine-based search engine

Daniel Verdejo

daniel.verdejo2@mail.dcu.ie

ABSTRACT

In this paper, we present a search engine architecture that leverages the power of vector space models (VSMs) to efficiently retrieve images. The system uses vector representations of the text descriptions generated using multimodal generative AI of images. These vector representations are used to calculate cosine similarities between user queries and documents. A custom-built VSM-based search engine is implemented using industry standard libraries to gather, process and retrieve data. The search engine's ability to rank multimedia data based on query similarity scores is demonstrated through a proof-of-concept implementation. This approach offers retrieval accuracy, scalability, and simplified indexing mechanism.

Keywords

Image retrieval, Transformer model, Large Language and Visual Assistant, Multimodal, Generative AI, Vector Support Machine, search engine, information retrieval, Web scraper.

1. INTRODUCTION

The explosion of images on the internet has created an unprecedented challenge for image retrieval systems. Traditional methods often struggle to keep up with the sheer volume and complexity of visual content, sometimes requiring labor-intensive manual annotations to bridge the gap [1]. This approach harnesses the power of a transformer model known as Large Language and Visual Assistant (LLaVA) to generate rich descriptions or annotations for images, combined with a Vector Support Machine (VSM) based search engine designed for image retrieval.

This paper explores whether transformer models, renowned for their success in natural language processing (NLP), can create rich annotations for visual content. It investigates how this approach impacts indexing and retrieval performance, shedding light on the potential of machine learning (ML) techniques in processing and indexing large image datasets.

As images are often subjectively interpreted by different users, we examine how our system ensures that retrieved results align closely with varied user queries. This variance in user queries highlights the potential of transformer models to provide rich nuanced annotations and interpret visual content, setting the stage for a comprehensive discussion on the fusion of NLP and ML in image retrieval.

The variety of images is somewhat random but focuses on wildlife, landscapes, cities, and vehicles. The terms used to scrape images are permutations of ~30 nouns and ~50 adjectives. to create search terms which were subsequently injected into google image searches. Using Selenium a browser session would start and an instance of the 'Undetected Chromedriver' is used to find all elements using the XPATH and then cherry pic the image src and alt properties. Once all the image URLs and titles are collected, they are then retrieved and written to .jpg files and stored in a common folder for later processing. During data gathering, about 1000 images are retrieved, with the first 1000 used in the dataset.

2. ANNOTATION

Using an Ollama Docker container enables pulling and running the LLaVA model locally, this could also be run in the cloud if doing so locally is not possible. Developed by Microsoft, the LLaVA model is an end-to-end trained large multimodal model that combines a vision encoder and language model for visual and language understanding, allowing for an image to text pipeline.

Following the data collection process a separate script for image processing is executed. The script first defines a data structure (a pandas Data Frame) to organize the data in a well-known meaningful data structure format using the following columns: image, title, description. The images stored in the folder mentioned above are iterated over and one by one are converted from .jpg file to a base64 encoded string representation of the image which is stored in the "image" column. The name of the file is used for the "title" column (excluding the filetype), and finally to get annotations for each image, the aforementioned LLaVA model is used. To the LLaVA model we simply pass a message of the following structure: {

role: 'user',

content: 'Give a description of this image',

'images': [<base64 encoded image>]

}

The model returns a rich description of the image is which is placed into the 'description' column of the Data Frame. These annotations will be used to generate embeddings or vector representations for a VSM based search engine.

3. INDEXING

Following the annotation process, the dataset is loaded from a CSV file and preprocessed. The preprocessing step involves refining text inputs by eliminating non-alphanumeric characters, converting all letters to lowercase, tokenizing the purified text into individual words, filtering out common stop words to reduce noise, and finally, reassembling the processed tokens into coherent strings. This refined text forms the foundation of the search engine's indexing strategy, ensuring that the generated embeddings accurately reflect the semantics of the image descriptions.

The SearchEngine class uses the "all-MiniLM-L6-v2" model to generate embeddings of the descriptions. This model was chosen for its efficacy in capturing semantic representations of text, which enables understanding the nuanced relationships between query texts and dataset entries. The class defines a function for creating embeddings which plays an important role in indexing the data, as entries with missing fields are dropped to ensure dataset integrity, preprocesses descriptions, and computes their embeddings. These embeddings are stored back in the DataFrame, providing a semantically-rich, indexed representation of each entry's description.

Using the embeddings as an indexing strategy for information retrieval (IR) offers several advantages over more traditional methods like keyword matching. One of the primary reasons is embeddings' ability to capture the semantic meaning of words or sentences beyond the surface level. This means that embeddings can understand context, synonyms, and varying sentence structures, which traditional methods may overlook. Onal et al. Highlights how embedding representations have advanced the field by enabling more semantically meaningful searches [2].

4. RETRIEVAL

The semantic search functionality is used through a search function, which computes a query's embedding using the same preprocessing pipeline as before and compares it against the indexed embeddings (the vector representations of the image descriptions) using cosine similarity. This measure quantitatively assesses the distance between the query and dataset entries, enabling the retrieval of the most relevant results. Scores are sorted in descending order of relevance, and the top n results are packaged into a JSON structure—comprising base64-encoded images, titles, and descriptions—ready to be presented to the user. This approach streamlines the retrieval of relevant information based on content semantics but also enhances the search experience by moving beyond the confines of keyword matching to understanding the user's search intent, potentially allowing for more nuanced search queries.

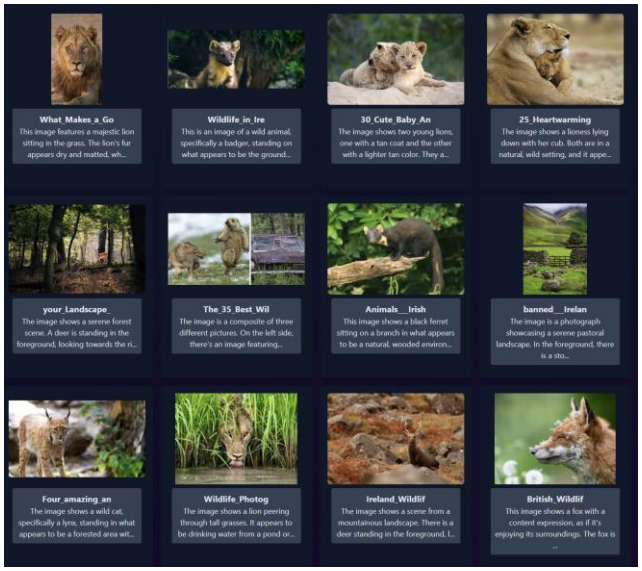
To retrieve and display results, the user is provided with a user interface built using popular modern tooling. It is built using VITE, React, Typescript and TailwindCSS. This allows for a rich responsive visual interface where the user can type a query into the search bar, click a button to invoke the search and see the semantically relevant results in real-time. The project is provided with Docker containers for a reproducible portable environment to deploy locally or in the cloud.

5. Evaluation

Evaluating the efficacy of retrieving relevant images based on user queries is a crucial step. This evaluation considers factors such as the accuracy and relevance of search results, the diversity of images returned, and the speed of the search process. To evaluate the search system these were some of the queries used to check how well the search met the criteria of the were as follows:

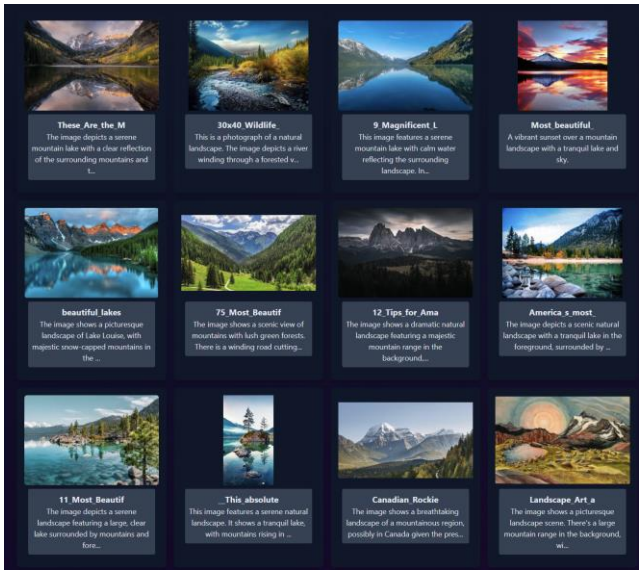
Query 1: “Animals in the wild relaxing”

Result**:**



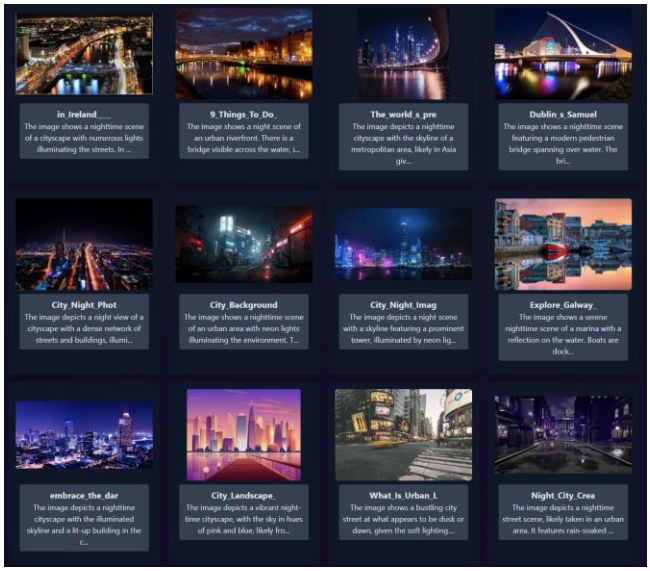
Query 2: “Mountains with a lake surrounded by forests”

Result:



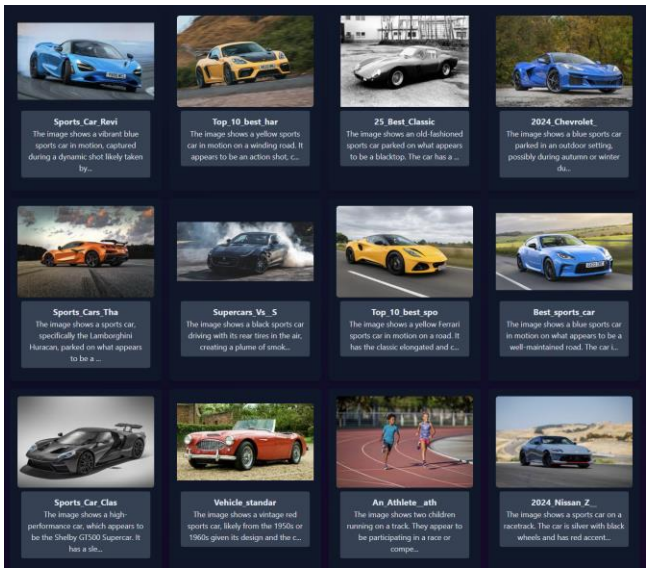
Query 3: “Large cities at night with water”

Result:



Query 4: “Sports cars driving fast”

Result:



As shown above in some of the sample queries and results, the images returned from each query contain highly relevant images. It can be noted that in each query there are hits present which are not as relevant to the user query but instead show a similar concept, for example in the query “Sports cars driving fast” an image of 2 people is shown running on a track, with the description containing “participating in a race or competition, as indicated by the numbered lane markers along the side of the track.”. These concepts of “race”, “competition” and “track” are semantically like the concept of “Sports,” “car” and “fast.”.

Additionally, as the user paginates to later pages of results the relevancy of the results begins to drop until such a point where the results are not relevant whatsoever.

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white

6. CONCLUSION & FUTURE WORKS

This paper has demonstrated the potential of combining transformer models, such as LLaVA, with VSM-based search engines to create a powerful image search engine. By leveraging the capabilities of these models, we have shown that it is possible to generate rich and nuanced annotations for images, which can be used to index and retrieve relevant results in response to user queries.

The proposed approach has several advantages over traditional methods, including its ability to capture semantic meaning beyond surface-level text matching. This enables the search engine to understand context, synonyms, and varying sentence structures, resulting in more accurate and relevant searches.

Through a combination of annotation generation, indexing, and retrieval, we have developed an image search system that can effectively retrieve relevant results based on user queries. The system has been evaluated using sample queries, which demonstrate its ability to return highly relevant images while also capturing similar concepts and nuances.

While there are still areas for improvement, the proposed approach offers a promising direction for advancing the field of information retrieval and computer vision. Future work could focus on refining the annotation generation process, improving the indexing strategy, or exploring new ways to integrate transformer models with VSM-based search engines.

In conclusion, this paper has presented an image search engine that leverages the capabilities of transformer models and VSM-based search engines to provide insight into a use case of transformer models such as LLaVA and the search experience as a result of doing so.

To further expand the capabilities of our system, we propose integrating search capabilities, such as:

- Sentiment analysis to capture sentiment and emotional tone in the image.
- Named entity recognition (NER) to identify and extract specific entities like people, locations, and organizations present in the image.
- Event extraction to identify and categorize events shown in images.

7. AKNOWLEDGEMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

8. REFERENCES

[1] Hanbury A. 2008. Journal of Visual Languages & Computing, Volume 19, Issue 5, Pages 617-627, https://doi.org/10.1016/j.jvlc.2008.01.002

[2] Onal KD, Zhang Y, Altingovde IS, et al. 2018. Neural information retrieval: at the end of the early years. Inf Retrieval J, Volume 21, Pages 111-182, https://doi.org/10.1007/s10791-017-9321-y

Lashkari F, Bagheri E, Ghorbani AA. 2019. Neural embedding-based indices for semantic search. Information Processing & Management, Volume 56,

Issue 3, Pages 733-755, https://doi.org/10.1016/j.ipm.2018.10.015

[3] Liu H, Li C, Li Y, Lee YJ. 2023. Improved Baselines with Visual Instruction Tuning. Computer Vision and Pattern Recognition, https://doi.org/10.48550/arXiv.2310.03744.