Image Retrieval with Siamese Networks

November 10, 2023

Marti JIMENEZ
m.jimenez@student.utwente.nl

Barbara Noemi SZABO
b.n.szabo@student.utwente.nl

Abstract

This paper explores the effectiveness of a Siamese Neural Network (SNN) in developing a robust image retrieval model by leveraging image key points. Through comprehensive experimentation and analysis, we assess the SNN's capabilities in capturing intricate image similarities, emphasizing its potential in addressing challenges associated with image retrieval tasks. By leveraging image key points and employing the SNN architecture, our findings decisively underscore the model inadequacy in facilitating robust image retrieval. Rather than showcasing effectiveness, our analysis illuminates the inherent flaws and limitations of the approach when applied to this context. Consequently, it is evident that pursuing this model direction, even with potential adjustments or simplifications, may not yield viable improvements in image retrieval accuracy or efficiency.

1 Introduction

In our data-driven world, Information Retrieval (IR) systems play a pivotal role. These systems are designed to help users search, retrieve, and make sense of vast amounts of information from diverse sources, such as documents, images, and multimedia content. With the exponential growth of digital data, the importance of these systems has only become more pronounced.

In the context of this project, the goal was to implement and evaluate methods aimed at improving image retrieval performance. Specifically, a neural network was implemented, wherein we harnessed the power of machine learning to enhance the accuracy and effectiveness of image retrieval. The importance of retrieving images is underscored by its real-world applications. For instance, in the medical field, IR systems with image retrieval capabilities can aid doctors in quickly accessing and comparing medical images, facilitating timely diagnosis and treatment decisions (Qayyum et al., 2017). In the realm of e-commerce, such systems enable users to find products visually, enhancing the shopping experience.

Machine learning has emerged as a crucial component in this field. It has revolutionized how we approach the task of retrieving information by enabling systems to learn from data and adapt to user preferences.

These innovations are motivated by several key factors that enhance the functionality and efficiency of IR systems. For example, machine learning-driven relevance ranking plays a pivotal role in the success of these systems. These models learn the relevance of documents to specific queries, enabling more accurate and context-aware retrieval, thereby increasing the likelihood of users finding precisely what they seek. Indeed, this is precisely what we are accomplishing with our network.

In the realm of state-of-the-art approaches within information retrieval, machine learning techniques have redefined how we navigate the digital landscape. Learning to rank methods, driven by machine learning models, are adept at predicting the relevance of documents to specific queries, making them a cornerstone of search engine result ranking. Neural networks, particularly deep learning techniques like Convolutional Neural Networks (CNNs) revolutionize the representation of documents and queries, leading to more accurate retrieval outcomes (Guo and Li, 2015). Finally, embedding-based approaches such as Word2Vec and Doc2Vec, are used to craft dense vector representations of words and documents that facilitate semantic understanding and retrieval (Roy et al., 2018).

Our model represents a contribution in this context, implementing a Siamese Network-based solution for image retrieval. It employs a Bag of Words approach, which represents and searches for images based on their visual content. The ORB feature detection and binary description algorithm is utilized for feature extraction, enabling the learning of vector representations for images. These vectors are employed to measure image similarity, enhancing image retrieval accuracy.

The research questions guiding this project are centered on the effectiveness of a Siamese Neural Network that leverages image key points in developing an image retrieval model. We aim to explore how well this approach performs in improving the precision and accuracy of image retrieval.

All the code can be found in the GitHub repository by Jimenez and Szabo (Jimenez and Szabo, 2023) present in the references.

2 Data

The dataset used for this project focuses on the retrieval of city images, with a particular emphasis on London. We used the full dataset that consists of two fundamental components: a collection and a set of query images ($N = 2692$) and a collection of city (map) images ($M = 3291$). These images serve as the reference dataset for the retrieval task, acting as the repository from which relevant images are sought. Conversely, the query images are employed to identify and retrieve the most similar images from the map dataset. The listing of the images included in the map and query datasets, respectively is provided. The order of images in these files corresponds to their respective positions in the similarity matrix. Lastly, the relevance judgments are structured in the form of a matrix. The matrix has N rows, where N is the number of query images in the query folder, and M columns, where M is the number of images in the map database. Each row of the matrix corresponds to a query image, and each column corresponds to a map image. This serves as the ground truth dataset for relevance judgments.

There are two sets of relevance judgments provided, specified by the keys ’sim’ and ’fov.’ ’Sim’ judgments use a binary score, with 0 indicating non-relevance and 1 indicating relevance between the corresponding map and query images. ’Fov’ judgments specify similarity as a degree, represented as a value in the interval $[0, 1]$, which indicates the degree of similarity between the query and map images.

The dataset is used to train and evaluate image retrieval models using two different approaches: one based on binary relevance scores and the other using the degree of image similarity. These relevance judgments are essential for assessing the performance and effectiveness of the image retrieval model in finding images that are relevant to given queries, which is a fundamental aspect of this project.

3 Method

In this section, we provide a comprehensive overview of the methodology adopted for the development of our image retrieval model using a Siamese Neural Network (SNN) with key-point vectors.

3.1 Data Preprocessing

During the data preprocessing phase, the images were transformed into key-point vectors using the Oriented FAST and Rotated BRIEF (ORB) algorithm, resulting in a concatenated vector representation that encapsulated the essential features of each image. The binary nature of the vector components denoted the presence or absence of specific key-point features within the image.

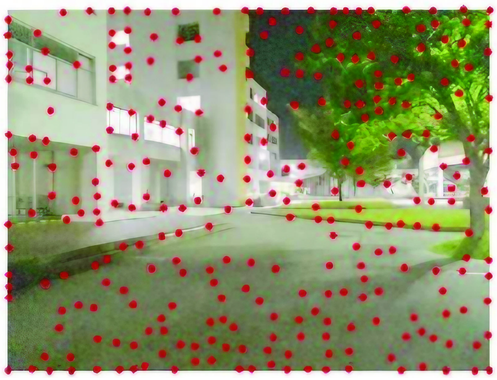

Figure 1: Image Key-point Extraction

A critical challenge encountered in our dataset was the substantial class imbalance, with a considerably larger proportion of non-relevant image pairs compared to relevant ones. To mitigate the impact of this class imbalance, we adopted a systematic approach wherein the dataset was balanced by maintaining a consistent ratio of relevant to non-relevant image pairs. Specifically, we ensured that each training set comprised an equal distribution of relevant and non-relevant image pairs, enabling the model to learn from a diverse range of data points and fostering a more equitable learning process.

By employing this strategy, we aimed to facilitate a more comprehensive exploration of the model’s learning dynamics, enabling it to discern intricate patterns and relationships within the data while addressing the challenges associated with class imbalance effectively.

3.2 Siamese Neural Networks

A Siamese Neural Network (SNN) is a specialized architecture that consists of two identical subnetworks that share the same set of weights and parameters. This design allows the network to process two different input data points, enabling direct comparison and similarity measurement between the inputs. SNNs are particularly relevant in tasks requiring similarity assessment, such as image comparison and retrieval, as they can effectively capture intricate relationships and patterns within the data.

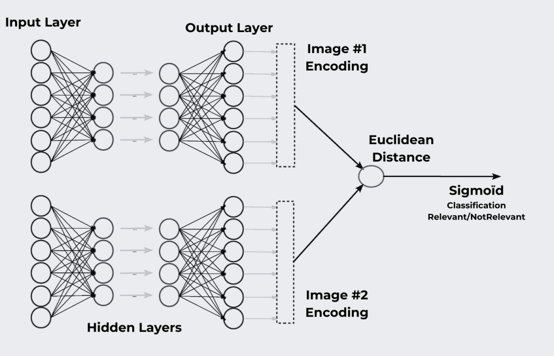

Figure 2: Our Siamese Neural Network Model

The SNN architecture was meticulously designed to accommodate the input key-point vectors and predict the relevance scores between image pairs.

During the experimentation phase, various configurations of the SNN architecture were tested, encompassing different activation functions, layer sizes, and regularization techniques. The shared layers within the SNN architecture were carefully selected and fine-tuned to capture the nuanced relationships and patterns embedded within the input vectors, ensuring the optimal performance of the model.

The training process involved rigorous experimentation with different methodologies, including variations in the number of key-points extracted and the nature of the target variables. By training the SNN using two different target configurations, one with continuous relevance scores and the other with binary relevance labels, we aimed to assess the model's adaptability to different target representations and evaluate its robustness in handling varying degrees of relevance.

3.2.1 SNN Architecture Details

The Siamese Neural Network (SNN) architecture comprised shared dense layers with specific configurations. The architecture included four shared dense layers, with layer sizes of 100, 50, and 10 neurons, respectively. Each shared dense layer was equipped with a 'sigmoid' activation function, ensuring non-linearity and facilitating the modeling of complex relationships within the data. Additionally, a dropout layer with a rate of 0.2 was incorporated, enhancing the model's robustness and preventing overfitting during the training process.

The SNN architecture employed the Euclidean distance metric to quantify the similarity between the encoded vectors.

The output layer of the SNN consisted of a single neuron with a 'sigmoid' activation function, facilitating the prediction of relevance scores between image pairs. The 'mean_squared_error' loss function and the 'Adam' optimizer were utilized to train the model, ensuring efficient convergence and accurate relevance score predictions.

4 Experiment and Results

4.1 Experiment Setup

The experiments were conducted on a standard laptop with an Intel i7-12th generation processor, utilizing Python 3.8 along with TensorFlow, Keras, and NumPy for neural network implementation.

We employed the Adam optimizer with a learning rate of 0.001 and a batch size of 32 for efficient model convergence. The training phase comprised 30 epochs, with early stopping and model checkpoints to prevent overfitting.

The dataset was split into training, validation, and test sets (80-10-10 ratio) to ensure a balanced representation of classes. Regular monitoring of training and validation loss curves was undertaken to assess the model's convergence and generalization capabilities.

4.2 Evaluation Process

The evaluation of the image retrieval model was performed using three key metrics, namely Precision at K, Accuracy within a Range, and Manual Check, ensuring a comprehensive assessment of the model's performance.

Precision at K was computed to measure the precision of the top K retrieved images, providing insights into the model's ability to retrieve relevant images within the top retrieved results.

The Accuracy within a Range metric was employed to assess the consistency of the predicted relevance scores with the ground truth within a specified range (e.g., 0.05). This evaluation metric offered valuable insights into the model's precision and reliability in predicting relevance scores within a predefined tolerance.

Furthermore, a Manual Check was conducted, wherein several queries were executed, and the relevance of the retrieved images was manually examined. This qualitative assessment allowed for a nuanced evaluation of the model's performance in retrieving relevant images based on real-world scenarios and user expectations.

By employing a combination of quantitative and qualitative evaluation metrics, we aimed to comprehensively assess the model's effectiveness in accurately retrieving relevant images and its ability to align with human judgment in relevance prediction.

4.3 Results

The evaluation of the image retrieval model yielded noteworthy findings, underscoring the intricate nature of the relevance prediction task and the efficacy of the model under distinct evaluation metrics.

When utilizing the dataset containing the relevance scores, the model's performance remained inconclusive, yielding no discernible results. However, with the dataset featuring binary relevance judgments, the model exhibited a commendable performance, demonstrating its robustness in handling the binary classification task.

k	Precision
5	0%
10	2%
20	2%
50	0.75%

Table 1: Model Accuracy within the Margin of Error

The Precision at K metric indicated a notably low precision. Despite this observation, an examination of the retrieved images revealed a significant number of potentially relevant images not captured by the metric. Notably, the images shared common elements, such as trees, people, buses, taxis, and road lines, underscoring the complexity of the image retrieval task and the limitations of the Precision at K metric in fully reflecting the model's performance.

Margin of Error	Accuracy
0.1	3%
0.05	0.5%
0.01	0%

Table 2: Model Accuracy within the Margin of Error

Regarding the Accuracy within Range metric, the best-performing model only achieved a 3% accuracy within a tolerance of 0.01 for the binary classification dataset. The accuracy dropped to a mere 0.5% for a tolerance of 0.005 and further plummeted to 0% for a tolerance of 0.01. These dismal results highlighted the model's incapacity to accurately classify images within a specific relevance score range, revealing substantial limitations in retrieving similar images and indicating severe deficiencies in the model's capabilities.

The Manual Check revealed that approximately 10 to 50% of the first 10 images retrieved were similar, reflecting the model's ability to capture some relevant images effectively. However, the subsequent images displayed notable similarities to the query, as observed in the shared elements among the images. This highlighted the challenges associated with capturing nuanced similarities and dissimilarities in image content and the limitations of the model in certain retrieval scenarios.

In a rigorous comparison utilizing the sign test to assess the performance of our model against an alternative approach employing Bag-of-Words (BOW) vectors, k-means clustering, and Euclidean distance for similarity measurement, intriguing insights have emerged. Over 20 iterations, our model exhibited superior precision at 10 in a modest 5% of the cases, suggesting a nuanced advantage. Our model did not perform worse than the alternative in any of the remaining instances. In the remaining scenarios both models achieved a precision at 10 score of 0. The absolute superiority of our model is limited, which suggests further optimizations.

The results indicated the model's competence in addressing certain aspects of the image retrieval task, while also emphasizing the existing challenges and nuances inherent in evaluating the performance of image retrieval models within real-world contexts.

5 Discussion and Conclusion

The evaluation of the image retrieval model revealed the complexities inherent in relevance prediction and image retrieval tasks. While the Precision at K metric provided valuable insights, its limitations were apparent, especially in scenarios where images shared common visual elements. This highlighted the need for more advanced feature extraction techniques and sophisticated similarity metrics to comprehensively evaluate the model's performance.

The Accuracy Metric and Manual Check demonstrated the challenges in capturing nuanced similarities and dissimilarities between images, particularly in scenarios where contextual similarities complicated the retrieval process.

These observations emphasize the crucial role of advanced feature descriptors and convolutional layers in enhancing the model's ability to capture intricate visual relationships and improve its performance in complex image retrieval tasks.

In **conclusion**, our investigation into employing Siamese Neural Networks (SNNs) alongside keypoint extraction for image retrieval has revealed inherent flaws and limitations within this approach. Although the SNN displayed marginal effectiveness in identifying relevant images, the overall performance, coupled with its complexity, signifies that this model is not a viable solution for the information retrieval task at hand.

Moving forward, a more streamlined approach involving a simple SNN with convolutional layers is anticipated to yield superior results. This alternative strategy aims to capitalize on the power of convolutional layers for feature extraction while minimizing unnecessary complexities, offering a promising avenue for improved image retrieval accuracy and efficiency.

References

Jinma Guo and Jianmin Li. 2015. Cnn based hashing for image retrieval.

M. Jimenez and B. Szabo. 2023. Image retrieval with siamese networks [github repository]. https://github.com/Marti2405/Image-Retrieval-with-Siamese-Networks.

Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. 2017. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20.

Dwaipayan Roy, Debasis Ganguly, Sumit Bhatia, Srikanta Bedathur, and Mandar Mitra. 2018. Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Examples of Image Retrievals

Dataset of 3291 images.

A.1.1 Example 1: Good Performance

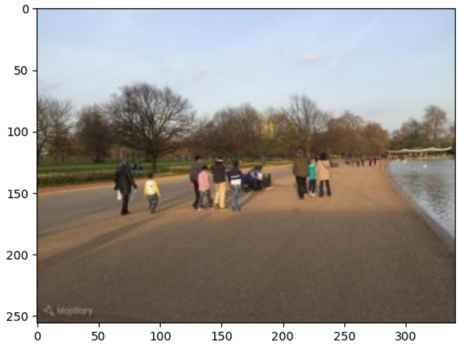

Figure 3: Query Image 1

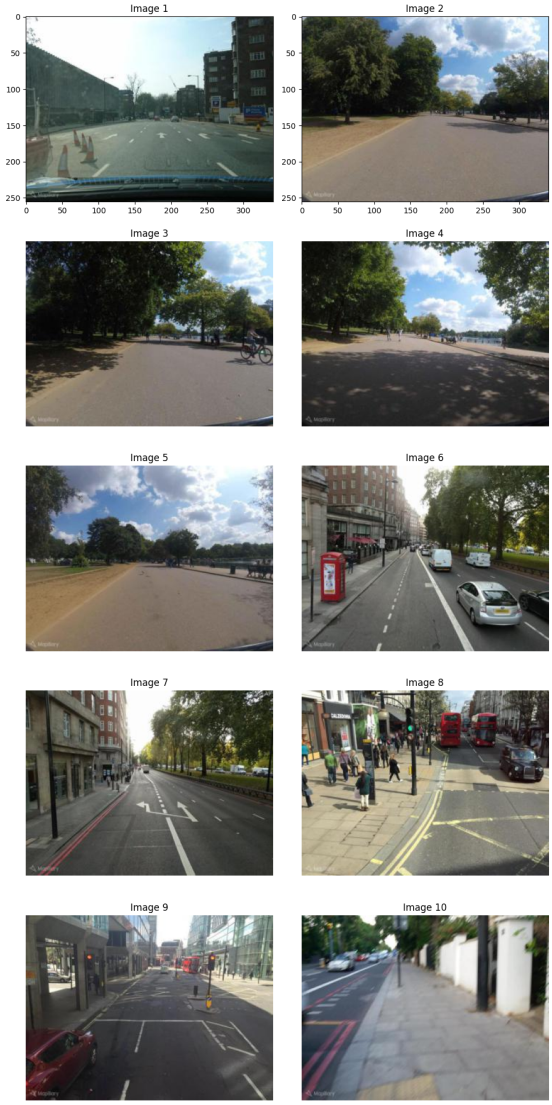

Figure 4: Retrieved Images 1

A.1.2 Example 2: Bad Performance

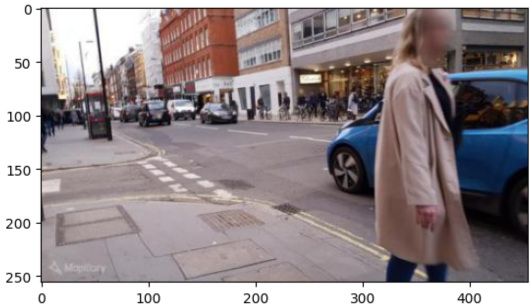

Figure 5: Query Image 2

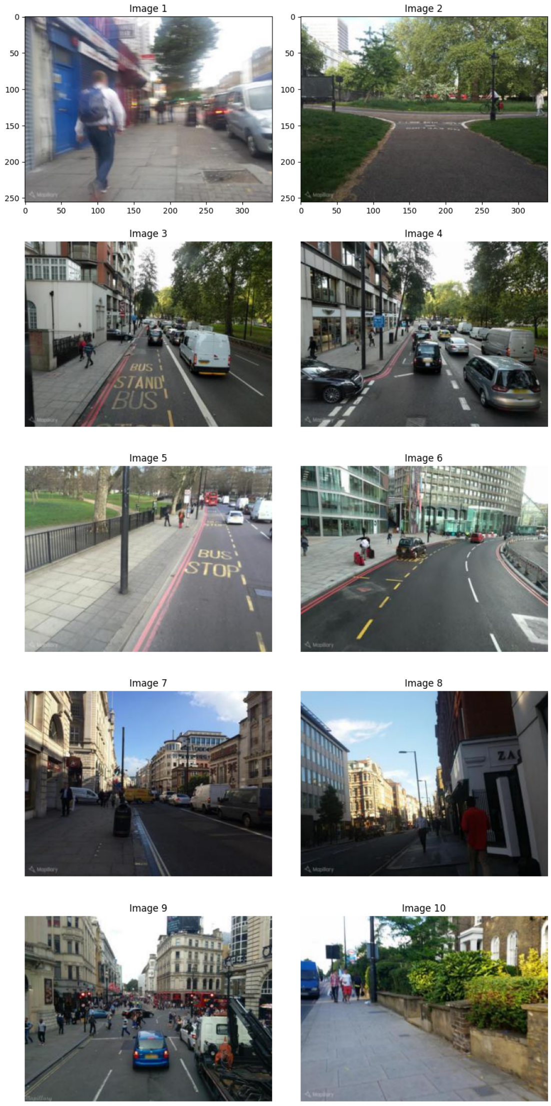

Figure 6: Retrieved Images 2