Boosting Performance of Image Retrieval System through Advanced Convolutional Neural Network Techniques

Dang Thi Tuong Vy[1,2*], Nguyen Hoang Long[1,2], Le Hoang Long[1,2] and Nguyen Tu Luan[1,2]

[1]University of Information Technology, Ho Chi Minh City, Vietnam.
[2]Vietnam National Univeristy, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): 20522176@gm.uit.edu.vn;
Contributing authors: 20521568@gm.uit.edu.vn;
20521563@gm.uit.edu.vn; 20521583@gm.uit.edu.vn;

Abstract

In recent years, image retrieval systems based on Convolutional Neural Networks (CNNs) [4] have become increasingly popular due to their ability to provide highly discriminative and compact image representations that enable efficient searching. However, training CNNs for image retrieval requires a large amount of annotated data, which can be time-consuming and expensive to obtain. In this report, I use a fully automated method for fine-tuning CNNs on a large collection of unordered images using reconstructed 3D models obtained from state-of-the-art retrieval and structure-from-motion methods to guide the selection of training data We think that selecting both hard-positive and hard-negative examples from the 3D models, based on the geometry and camera positions, can significantly improve particular-object retrieval performance. [4]Additionally, we use Generalized-Mean (GeM) pooling layer that generalizes max and average pooling, which further enhances retrieval performance. The proposed CNN descriptor whitening method, which is discriminatively learned from the same training data, outperforms other methor. We Applying the proposed method to the VGG or ResNet network achieves state-of-the-art performance on standard benchmarks such as the Oxford Buildings, Paris, datasets [4]. The result suggest that the proposed automated method for fine-tuning CNNs for image retrieval can provide highly discriminative and efficient image representations, even in cases where a large amount of annotated data is not available.

Keywords: Image retrieval, CNNs, Oxford dataset, GEM

1 Introduction

Image retrieval is the process of searching for digital images in large databases based on their visual content. This technique involves extracting features from images such as color, texture, and shape, and then using these features to compare images and retrieve those that are similar to a query image. Image retrieval has a wide range of applications, from searching for images on the internet to more specific uses in areas such as medicine, surveillance, and forensics. For example, in the field of medicine, image retrieval can be used to compare a patient's medical images to a database of similar images to help doctors diagnose and treat diseases. There are various techniques used for image retrieval, including similarity-based and learning-based methods. Similarity-based methods compare the features of query and database images to compute a similarity score, while learning-based methods use machine learning algorithms to train a model to recognize patterns and retrieve relevant images. In recent years, deep learning techniques have been applied to image retrieval, resulting in significant improvements in retrieval accuracy.

Image retrieval using descriptors based on activations of Convolutional Neural Networks (CNNs) [3] has become very popular due to their powerful discriminative capability, compactness of representation, and search efficiency. However, training CNNs requires a large amount of annotated data, which is often difficult and expensive to obtain. In this paper, we propose a novel method to fine-tune CNNs for image retrieval using a large collection of unordered images and without the need for human annotation. Our approach utilizes 3D reconstructions of the scene obtained through Structure-from-Motion (SfM) methods and automatically retrieves hard-positive and hard-negative examples by exploiting the geometry and camera positions from these reconstructions. Additionally, we propose a novel Generalized-Mean (GeM) pooling layer that improves retrieval performance. The proposed method achieves state-of-the-art performance on standard benchmarks for image retrieval.

For the convenience of comparing the effectiveness of the method used in the article. We proceed to use both VGG16 and ResNet101 models, in addition, instead of only using the Oxford5k dataset, the team will conduct further evaluation on another dataset, Paris6k. I will present the report in detail in the next section.

2 Backgrounds

2.1 Content-Based Image Retrieval

Content-based Image Retrieval (CBIR) [12] is a computer vision technique that allows us to search for related images in large databases. This search is based on low-level features, such as color, texture, and shape, or any other features that can be extracted directly from the image. The selection of features greatly influences the performance of CBIR systems.

The fundamental concept of CBIR is to represent images as multidimensional feature vectors. The similarity between images in the dataset and the query image is measured using a distance metric between their feature vectors, such as the Euclidean distance. The k images with the shortest distance value will then be retrieved and returned to the user, where k is a user-defined parameter. A detailed description of the CBIR process is shown in Figure ().

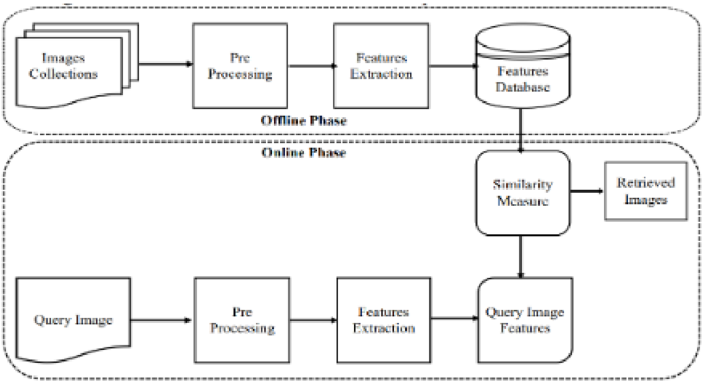


Fig. 1 Block diagram for an Content-Based Image Retrieval System

In detail, the image retrieval system consists of 2 stages: Offline and Online [12] **Offline Phase**:

- The offline phase of the CBIR system involves a series of preprocessing steps to prepare the images in the dataset for the training stage. These steps include tasks such as noise reduction, image enhancement, and feature extraction. The extracted features are then used to build an index that maps the feature vectors to their corresponding images.
- The preprocessing stage can be time-consuming, particularly when dealing with large datasets. The time required for this stage depends on various factors, including the number of images used in the training process and the complexity of the features being extracted. Nonetheless, this stage is critical to the overall performance of the CBIR system as it directly impacts the accuracy and speed of the image retrieval process.

Online Phase

- In the Online phase of CBIR, the user submits a query image and the system performs several operations on it. First, the query image undergoes a preprocessing stage, which typically includes resizing, normalization, and noise reduction. Then, the system extracts features from the query image using the same or similar feature extraction methods as in the Offline phase.
- Once the feature vector of the query image is obtained, the system uses a distance metric to measure the distance between the feature vector of the query image and the feature vectors of the images contained in the dataset. A popular distance metric used in CBIR is the Euclidean distance, but other metrics such as the cosine distance or the Manhattan distance can also be used.
- Based on the distance between the feature vectors, the system retrieves the k images that are most relevant to the query image. The value of k is a parameter that can be adjusted by the user. The k images with the shortest distance value will be returned as the query result. In practice, the relevance of the retrieved images can be evaluated by the user, and the query can be refined by adjusting the feature extraction or the distance metric used.

However, CBIR also faced some major challenges in its implementation. The biggest challenge is to reduce the semantic gap. This is the gap in terms of visual information captured by the imaging device and visual information perceived by a human vision system (HVS).

The semantic gap is a significant challenge in CBIR, as it can affect the accuracy and relevance of the retrieval results. Closing this gap requires bridging the differences between low-level image features and high-level semantics, which can be a difficult task. One approach to reducing the semantic gap is to incorporate user feedback, where users can rate or label retrieved images to improve the accuracy of future retrievals. Another approach is to use deep learning techniques to extract high-level features automatically, thereby reducing the dependence on handcrafted features.Moreover, CBIR systems often have to deal with large-scale datasets, which can be computationally intensive and time-consuming. To address this challenge, various optimization techniques have been developed to accelerate the search process, such as approximate nearest neighbor search and indexing methods. Another challenge is the robustness of the system, where noise or variations in the input image can affect the retrieval results. To improve the robustness, techniques such as feature fusion and ensemble learning have been proposed.

Despite many challenges, CBIR is still an efficient and improved date retrieval method today. It is a powerful technique that enables efficient image retrieval from large databases, with many applications. potential uses in areas such as healthcare, security, and e-commerce.

2.2 Convolutional Neural Networks

I will present and introduce the basics of Convolutional Neural Networks (CNNs). CNNs are a type of deep learning algorithm that is commonly used in image recognition and computer vision tasks, image retrieval. The structure of a CNN consists of convolutional layers, pooling layers, and fully connected layers. In convolutional layers, a filter or kernel is applied to the input image to extract features. Pooling layers downsample the image and reduce the number of parameters. Fully connected layers perform classification based on the extracted features. CNNs have achieved impressive results in various applications, including object recognition, face recognition, medical image analysis and image retrieval.In this report, we will only be interested in the feature extraction task of CNNs.

Over the past decades, many feature extraction methods have been proposed such as image extraction at the global level (shape, color). Some feature extraction methods have also been proposed such as SIFT, Bag of Words, etc [13]. However, they still cannot effectively solve the semantic gap problem in many problems. Recently, advanced machine learning tools have provided a new way to reduce semantic gaps, and CNNs are the hope for bride the gap by learning features directly from images. image without using any manual features because it is an end-to-end system.

The question here is how can we train CNNs to extract good features on images in the query dataset? Qayyum and colleagues proposed a method to use a single network of CNNs to perform image feature extraction [14]. They apply in the problem of medical images, here we use the same method to solve our feature extraction problem. They treat this as a normal classification problem. Suppose in the dataset there are 10 types of images (10 classes) CT such as houses, trees, gates... then we will train this deep learning network as a classification network with the last class having 10 neurons. After the training is done, remove the fully connected layer and it will become a specialized network for feature extraction only. The process of the method is depicted in figure (2)

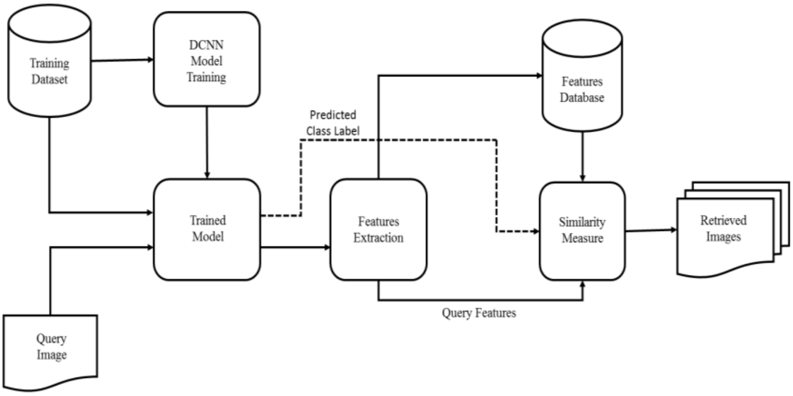


Fig. 2 Process of image retrieval system by CNNs. [14]

3 Methods

3.1 Generalized-Mean pooling

We will add a pooling layer [3] with X as an input and the output for the pooling process is a vector f. [3]The pooling step using the Generalized-Mean (GeM) pooling by using Equation 1 and 2 where k 1, ..., K.

Equation 1: $f^{(g)} = [f_1^{(g)}, f_k^{(g)}, ..., f_K^{(g)}]$

Equation 2: $f_k^{(g)} = (\frac{1}{\mathsf{X}_k \sum_{x \in X_k}^{x^{p_k}})^{\frac{1}{p_k}}}$

The pooling parameter p_k can be manually set or learned since this operation is differentiable and can be part of the back-propagation[2]. There is a different pooling parameter per feature map in (2), [2] but it is also possible to use a shared one. In this case $p_k = p$, k [1, K] and we simply denote it by p and not p_k. We examine such options in the experimental section and compare to hand-tuned and fixed parameter values.

Max pooling, [2] in the case of MAC, retains one activation per 2D feature map. In this way, each descriptor component corresponds to an image patch equal to the receptive field. Then, pairwise image similarity is evaluated via descriptor inner product.

The last network [2] layer comprises an $L2$-normalization layer. Vector f is $L2$-normalized so that similarity between two images is finally evaluated with inner product. [3][2]In the rest of the paper, GeM [6] vector corresponds to the $L2$- normalized vector f and constitutes the image descriptor.

3.2 Image descriptor

The MAC (maximum activation of convolutions) method employs max pooling to retain one activation for each 2D feature map. [1]This results in descriptor components that correspond to an image patch equal to the receptive field. MAC similarity is determined by evaluating pairwise image similarity via descriptor inner product. [2]Patch correspondences are formed implicitly and the strength of each correspondence is given by the product of the associated descriptor components. The image patches that contribute the most to the similarity are shown in Figure (). These implicit correspondences can be improved after fine-tuning.

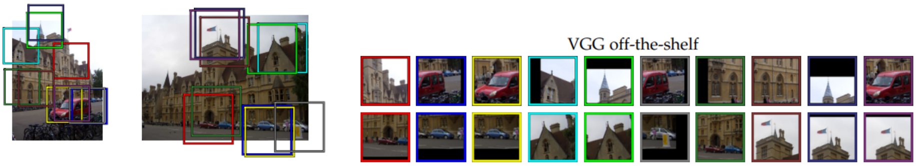


Fig. 3 Image descriptor.

3.3 Siamese learning and loss function

Siamese learning [15] is a type of deep learning that involves training a neural network to identify the similarity or dissimilarity between two objects[2], such as images, text documents, or audio recordings. Siamese networks only require a few photos to make a good prediction. This ability to learn from very little data has made Siamese networks increasingly popular in recent years. Siamese networks are not actually a completely new model; they are more like a technique to train the model. It is a neural network architecture containing two or more subnets, called "twins," which share the same hyperparameter configuration and share weights with each other. When updating the weights of a model, they will update both neural networks simultaneously [2]. The input Siamese networks are two images, and the output is the feature vector of those two images. We will use it in conjunction with Contrastive Loss in this article.

Contrastive loss is a loss function used in siamese neural networks for learning similarity between pairs of inputs. It is designed to push the embeddings of similar inputs closer together and push the embeddings of dissimilar inputs farther apart [3].The loss is computed as the sum of the distances between the embeddings of the similar pairs and the maximum of the distance between the embeddings of the dissimilar pairs and a margin value. The goal is to minimize this loss function to learn a useful similarity metric.

Summary, We apply siamese architecture and train two-branch network. [1]Each branch is a copy of the other, meaning they share the same parameters. [2]The training input consists of pairs of images (i, j) and labels Y(i, j) ∈ 0, 1 declaring whether a pair does not match (label 0) or match (label 1). We use the contrastive loss with the formula.

$$\begin{cases} \frac{1}{2}||\bar{f}(i)-\bar{f}(j)||^2 \text{ if } Y(i,j)=1 \\ \frac{1}{2}\left(max\left\{0,\tau-||\bar{f}(i)-\bar{f}(j)||\right\}\right)^2 \text{ if } x=Y(i,j)=0 \end{cases}$$

Fig. 4
[2]

[1][2]Where $\bar{f}(i)$ is the ‘$2-normalized$ GeM vector of image i, and is a margin parameter defining when non-matching pairs have large enough distance in order to be ignored by the loss.

3.4 Whitening and dimensionality reduction

The post-processing of fine-tuned GeM [6] vectors is considered. The projection is decomposed into two parts: whitening and rotation. The whitening part is the inverse of the square-root of the intraclass (matching pairs) covariance matrix $C_S^{-\frac{1}{2}}$ [2]
where $C_S = \sum_{Y(i,j)=1}(\bar{f}(i)-\bar{f}(j))(\bar{f}(i)-\bar{f}(j))^\top$ [2]
The rotation part is Principal Component Analysis (PCA) [7] of the

unmatched-pairs covariance matrix in the whitened space as in Equation The projection $P = C_S^{-\frac{1}{2}} eig(C_S^{-\frac{1}{2}} C_D C_S^{-\frac{1}{2}})$ [2] is then applied as $P^{\top}(\bar{f}(i) -)$, where is the mean GeM vector to perform centering. To reduce the descriptor dimensionality to D dimensions [2], only eigenvectors corresponding to D largest eigenvalues are used. Projected vectors are subsequently $L2$- normalied ing.Whitening consists of vector shifting and projection which is modeled in a straightforward manner by a fully connected.

3.5 3D reconstruction

The report proposes a method for automatic selection of training data for image search by coupling Bag-of-Words (Bow) image retrieval and Structure-from-Motion (SfM). It avoids the need for human-annotated data or any other assumptions for the training dataset. The approach uses geometrical camera positions from automatically reconstructed 3D models obtained via BoW-based image retrieval and exploiting the state-of-the-art retrieval-SfM pipeline. The SfM filters out mismatched images and provides image-to-model matches and camera positions for all matched images in the cluster, making the process fully automatic. The paper highlights the efficiency of the process, which is further improved by applying local features-based fast image clustering.

4 Experiments

4.1 Datasets

Created by a research group at the University of Oxford, the Oxford Building dataset (5K) [9] includes over 5,000 high-quality images of buildings in the central area of Oxford, United Kingdom. The dataset is divided into two main parts: the first part contains images of buildings taken from a distance, while the second part contains detailed images of the buildings.

The first part of the dataset includes over 3,000 images of buildings taken from a distance, where all the buildings can be fully seen from a far distance. These images are taken from a range of different angles and distances, with buildings of varying heights and architecture. Building names, addresses, GPS coordinates, and image attributes are provided for each image.

The second part of the dataset includes over 2,000 detailed images of the buildings [9]. These images are taken from closer distances and focus on the details of the architecture of the buildings. Building names, addresses, GPS coordinates, and image attributes are also included in these images.

The Oxford Building dataset (5K) can be used to develop and evaluate algorithms for image classification, object recognition, image search, and object detection in the fields of machine learning and artificial intelligence. Additionally, researchers and engineers can also use this dataset to implement applications related to architecture, tourism, and image-based localization systems. The Oxford Building dataset (5K) is provided for free for research and non-commercial purposes.

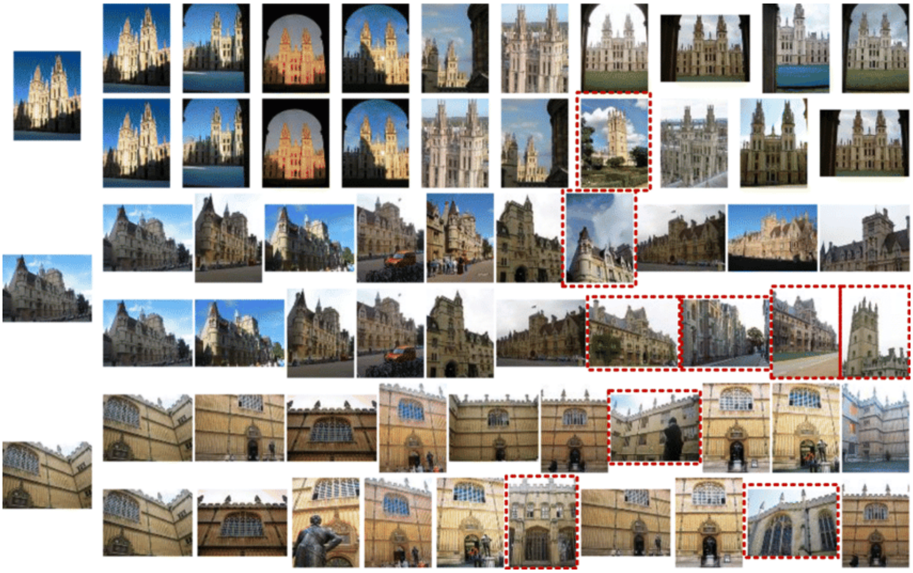


Fig. 5 Oxford5k dataset.

In addition, our team also evaluated and tested on the Pari6k dataset [10]. So let's talk a little more about this dataset. The Paris6k dataset is a rich collection of images of famous architectural landmarks in Paris, France. The dataset includes over 6,000 high-quality images, captured from various different angles and formats.

The images in the Paris6k dataset [10] were taken from a variety of different angles, including tilted angles, distant and close-up shots, with a resolution of up to 800x600 pixels. The architectural landmarks documented in this dataset include many famous historical monuments, tourist destinations, as well as modern buildings.Each image in the dataset is accompanied by detailed information about the name of the architectural landmark, address, GPS location, and other technical specifications of the image. The Paris6k dataset provides useful materials for researchers and engineers to develop algorithms for object recognition, image classification, image search, and image-based positioning systems.

4.2 Performance metric

Average precision (AP) [3] is a commonly used performance metric in information retrieval and machine learning to evaluate the accuracy of classification or ranking models. It measures the quality of a model's ranking or classification output by computing the area under the precision-recall curve. AP can range from 0 to 1, with a higher value indicating better performance. An AP of 1 means that the model achieved perfect precision and recall for all values of the classification threshold. AP is particularly useful when the positive class is rare, and precision is more critical than recall. It is commonly used in evaluating text and image retrieval systems, as well as in evaluating object detection and segmentation models.However, MAP requires many relevance judgements in text collection.

On the experimental process,[3] we use *Mean Average Precision* (MAP) as the primary metric for our work and MAP equal to average sum of all AP [1]

4.3 Experimental details

Training dataset: Our training dataset is drawn from a vast collection of million images that depict popular landmarks, cities, and countries from around the world. [5]To ensure the quality and relevance of the dataset, we have used an extensive retrieval Structure from Motion reconstruction process that involves clustering and removing overlapping 3D models. This has resulted in a refined dataset consisting of 713 3D models, comprising more than 163,000 unique images selected from the original dataset [11].And the dataset also, on purpose,consists of all images from Oxford5k datasets.

To prepare our training dataset. We have also employed image preprocessing techniques, including noise reduction, image enhancement, and feature extraction to improve the quality and consistency of the dataset. **Training pair** Describes the process of training an approach for image retrieval using a dataset of 713 3D models containing over 163,000 unique images of landmarks, cities, and countries around the world. The size of the 3D models in the dataset varies from 25 to 11,000 images, ensuring a diverse and comprehensive training set. 551 models have been randomly selected for training and 162 for validation, with around 133,000 and 30,000 images, respectively.. The training pairs are pre-processed using techniques such as image augmentation, normalization, and data balancing to ensure accuracy and reliability. The training pairs are designed and curated to provide a diverse and reliable resource for training and optimizing the approach. The approach is capable of accurately recognizing and retrieving landmarks and other visual features in real-world scenarios due to the quality and comprehensiveness of the training pairs.

Configuration For our image retrieval approach, we utilized pre-trained ResNet101-GeM [2] and VGG16-GeM [2] models, which are deep learning architectures that have been fine-tuned for our specific task. The fine-tuning process was conducted using the PyTorch framework, an open-source machine learning library that enables efficient development and training of deep learning models. We ran our experiments on Google Colab Pro, which provides access to high-end GPUs and TPUs for faster training and testing.

Test datasets and evaluations To evaluate the effectiveness of our image retrieval approach, we tested our model on several benchmark datasets, including Oxford5k b9, Paris6k [10]. These datasets consist of a wide variety of images of popular landmarks, objects, and scenes, and are widely used in the field of image retrieval for performance evaluation. We utilized the standard evaluation protocol for these datasets, which involves measuring the mean average precision (MAP) of the retrieval results. MAP is a widely used metric for evaluating the effectiveness of retrieval systems, as it takes into account both the relevance and the ranking of the retrieved images.

4.4 Results

In the section, we aim to provide a comprehensive analysis of our experimental results, covering various aspects of our research. Firstly, we will present our result in detail, discussing the specific methodologies and techniques that we employed. In the Oxford5k dataset, the proposed method achieved a maximum MAP of 88.22 using the ResNet101 model in combination with the ”whitening” technique. The MAPs for other methods used in the experiment remained stable, with marginal differences from the method with the highest results, all above 80.00. The runtime for the two models was also comparable, with ResNet taking slightly less time at 12.49 seconds. It is worth noting that the results were obtained using the standard evaluation protocol, where the query images were cropped, and the cropped zone was treated as input to the architecture.

	Oxford5k	Oxford5k - whiten	Oxford5k - elapse time
ResNet101-GeM	81.09	88.22	12m49s
VGG16 - GeM	82.51	87.34	14m59s

Table 1 MAP of Oxford5k

	Paris6k	Paris6k - whiten	Paris6k - elapse time
ResNet101-GeM	87.81	92.62	15m50s
VGG16 - GeM	82.32	87.82	18m29s

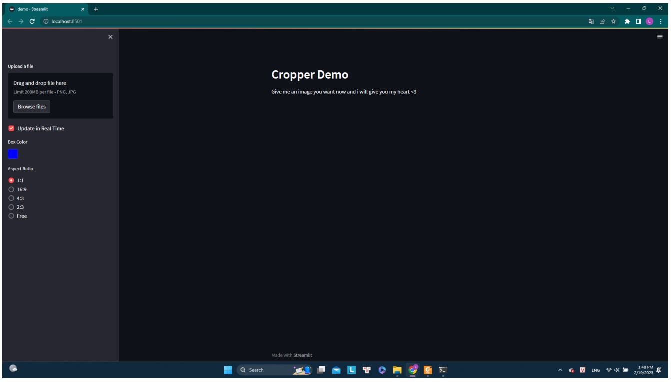
Table 2 MAP of Paris6k

The proposed method was evaluated on the Paris6k dataset, and the ResNet101 model with ”whitening” yielded the highest MAP of 92.62, while other methods used in the experiment showed stable MAPs with marginal differences from the method with the highest results, all above 80.00. The runtime for the two models was comparable, with ResNet taking slightly less time at 15.50 seconds.

To summarize, the proposed method achieved high mean average precision (MAP) results on both the Oxford5k and Paris6k datasets, with the ResNet101 model in combination with the ”whitening” technique producing the highest MAP results.Overall, the results indicate that the proposed method is effective for image queries and may be considered for use in related applications.

4.5 Demo

This is an application of using the image search system in a website that the team did.



Cropper Demo

Give me an image you want now and i will give you my heart <3

Upload a file

Drag and drop file here

Limit 200MB per file • PNG, JPG

☑ Update in Real Time

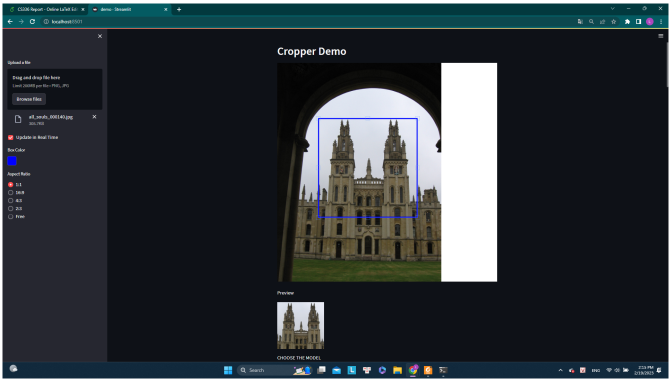
Box Color

Aspect Ratio

● 1:1
○ 16:9
○ 4:3
○ 2:3
○ Free

Fig. 6 Main user interface

We will feed into the system an image representing the information need, but we will have to crop the image to remove unnecessary features that save computational space and have the Aspect ratio available for user easily to choose.



Cropper Demo

Upload a file

Drag and drop file here

Limit 200MB per file • PNG, JPG

all_souls_000140.jpg 305.7KB

☑ Update in Real Time

Box Color

Aspect Ratio

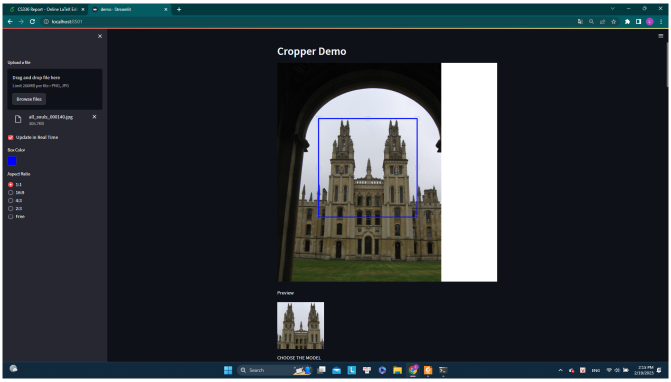
● 1:1
○ 16:9
○ 4:3
○ 2:3
○ Free

Preview

CHOOSE THE MODEL

Fig. 7 Cropper image to remove unnecessary part

After entering the image to be searched for in the system, the user needs to choose a model (there are 2 models, VGG16 and ResNet101) for the query and select the number of steps to return images. the best fit for the information need.
Here are some examples when we run a demo

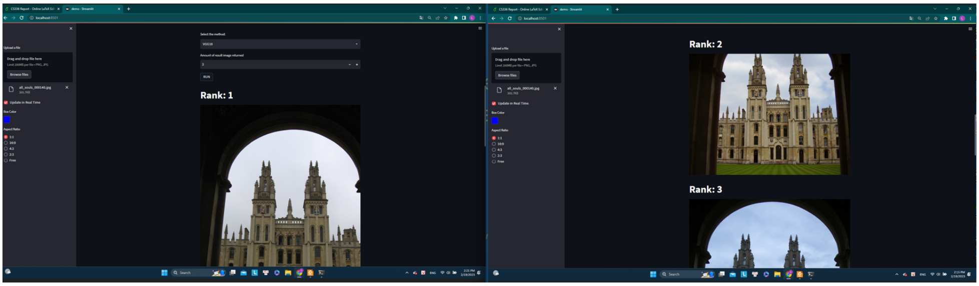


☑ Update in Real Time

Box Color 🟦

Aspect Ratio
● 1:1
○ 16:9
○ 4:3
○ 2:3
○ Free

Fig. 8 Choose Model and number of image in the result set



Select the method: VGG16

Amount of result image returned 3

☑ Update in Real Time

Aspect Ratio: ● 1:1 ○ 16:9 ○ 4:3 ○ 2:3 ○ Free

Fig. 9 Example of image retrieve

5 Conclusions

In this report, we employed fine-tuning of convolutional neural networks (CNNs) for image retrieval, using training data selected from an automated 3D reconstruction system applied to an unordered photo collection that includes buildings, landmarks, and other rigid 3D objects. The results brought on both Oxford5K and Paris6k episodes are quite good.The proposed method achieved competitive results among local feature-based systems while being faster and requiring less memory, without relying on any standard benchmark. Overall, this report provides a promising approach to improving image retrieval using CNNs and automated 3D reconstruction methods.

In the future, our team will use the method in the article to conduct testing and evaluating more data sets in larger quantities to make the most objective comment. But in general, with the current data set, Oxford5K or Paris6k, the results are quite positive.

Acknowledgements

The authors express their gratitude to Dr. Thanh Duc Ngo for his guidance and support during their project. The teacher taught and supervised them on multimedia information retrieval, providing invaluable insights that enriched their work. He shaped their understanding of the field and motivated them to pursue innovative approaches. He also spent countless hours reviewing and providing feedback on their work, and was always available and responsive to their questions and concerns. Without Dr. Ngo's support and dedication, their project would not have been possible, and the authors are grateful for his contributions to their academic and professional development.

References

[1] Filip Radenovi´c, Giorgos Tolias , Ondˇrej: Chum CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples

[2] Filip Radenovic, Giorgos Tolias Ond, ´ ˇrej Chum Fine-tuning CNN Image Retrieval with No Human Annotation

[3] Tan Ngoc Pham, An Vo, Dzung Tri Bui: https://github.com/DTA-UIT/ImageRetrievalSystem/blob/main/report.pdf

[4] Mingxing Tan, Quoc V. Le :EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten :Densely Connected Convolutional Networks

[6] O. Morere, J. Lin, A. Veillard, L.-Y. Duan, V. Chandrasekhar, and T. Poggio, "Nested invariance pooling and rbm hashing for image instance retrieval," in Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017, pp. 260–268

[7] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, vol. 2, no. 11, pp. 559–572, 1901.

[8] G. T. Y. A. Filip Radenovic, Ahmet Iscen and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," 2018.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in 2007 IEEE conference on computer vision and pattern recognition. IEEE, 2007,pp. 1-8

[10] ——, “Lost in quantization: Improving particular object retrieval in large scale image databases,” in 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008, pp. 1–8

[11] J. L. Schonberger, F. Radenovic, O. Chum, and J.-M. Frahm, “From single image query to detailed 3d reconstruction,” in Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5126–5134

[12] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. IEEE Transac- tions on Circuits and Systems for Video Technology, 2021.

[13] Maxim Mizotin, Jenny Benois-Pineau, Michèle Allard, and Gwenaelle Catheline. Feature-based brain mri retrieval for alzheimer disease diagnosis. In 2012 19th IEEE International Conference on Image Processing, pages 1241–1244, 2012

[14] Adnan Qayyum, Syed Muhammad Anwar, Muham- mad Awais, and Muhammad Majid. Medical image re- trieval using deep convolutional neural network. Neu- rocomputing, 266:8–20, 2017

[15] Gregory R. Koch. Siamese neural networks for one- shot image recognition. 2015