Project Proposal:Instance Based Image Retrieval using Deep Learning

Akanksha Shrimal
Indraprastha Institute of Information and Technology
MT20055
akanksha20055@iiitd.ac.in

Shivam Sharma
Indraprastha Institute of Information and Technology
MT20121
shivam20121@iiitd.ac.in

Shivank Agrahari
Indraprastha Institute of Information and Technology
MT20096
shivank20096@iiitd.ac.in

Pradeep Kumar
Indraprastha Institute of Information and Technology
MT20036
pradeep20036@iiitd.ac.in

Sudha Kumari
Indraprastha Institute of Information and Technology
MT20098
sudha20098@iiitd.ac.in

1. PROBLEM STATEMENT

We propose an approach to solve the problem of instance retrieval. In instance retrieval task, we retrieve images from the database based upon the query or input image. There has been a lot of progress in the image classification and image retrieval domain in recent years. Recently, Instance retrieval has also gained a lot of attention. Its importance is felt in medical sciences, astronomy, security, autonomous vehicles among others. The task is to retrieve images from the database based upon the scene or object in the target image. It deals with the content inside the image such as color, shape and image structure. It can be of particular use when the required feature is partially obscured or the feature is inseparable in the reference image in the database.

2. MOTIVATION

Considering expansion of web data at cosmic level over time, image retrieval is a very essential problem. There are plenty of image retrieval techniques which are evolved. We found that state of the art techniques used for image retrieval(used by tech giants like Google, Bing etc.) perform well while finding relevant images. But we experienced that, the techniques followed there for retrieval consider image as a whole and automatically decide weights to different aspects of the image. There is no user intervention in deciding the part to focus on in the input image while retrieval. We tried to perform some queries and found that it does not consider part of the image which is of importance to the user, instead just returns the images which are visually similar to input image at large. Some of the example image query performed on Google Image search are shown in 1a where we have given the following as input images. 1) An animated house 2) A fortress 3) Samsung store. While Google is able to detect the key feature\concept, it overlooked the remaining features. 1b is the proposed retrieval system which allows the user to decide upon the key feature(s) in the query image. The images on left targets a particular subsection of the image and following three images are the expected retrievals.

Figure 1a: Current

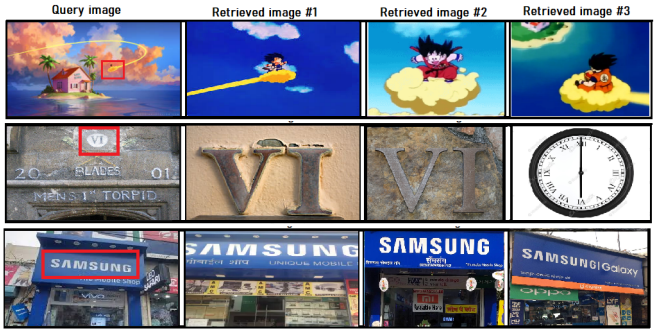

Figure 1b: Proposed

3. LITERATURE REVIEW

Image Retrieval is an important research area in computer vision in which similar images are retrieved from data base w.r.t. a given query image. Basically, the similarity between the query image and the database images is used to rank the database images in decreasing order of similarity.

In the past decade, a number of image retrieval types have been explored which include cross-modal retrieval, sketch based retrieval, multi-label retrieval, instance retrieval, object retrieval,semantic retrieval, fine-grained retrieval and asymmetric retrieval [1].

Image instance retrieval is the problem of retrieving images from a large database that contain or depict similar objects or scene to a target image. Recently, Instance Retrieval has gained popularity as users can retrieve images from database focusing on particular region of image.
Early Techniques on Instance Retrieval task used hand-crafted features for image matching but from past decade there is a shift to deep learning based features[1][2] .

Zheng, Liang and Yang, Yi and Tian, Qi [2] emphasize on a comparison between Scale Invariant Feature Transform(SIFT) based hand-crafted features and Convolutional Neural Networks(CNN) based features for Instance Retrieval. [3]. They stated that SIFT-based methods does not perform well for a specific object retrieval and for common object retrieval it is not as competitive as the CNN models are. On the other hand CNN-based methods with fixed-length representations have advantages in nearly all the benchmarking datasets. When sufficient training data is provided, the ability of CNN embedding learning can be fully utilized. Also the pre-trained CNN modals are competitive. Thus with CNN based methods, the results achieved for instance retrieval are far better then the traditional SIFT based methods and there is shifting from SIFT based methods to CNN based methods.

Followed by hand-crafted approaches, bag-of-features representations were largely used for instance retrieval. In [3] input image is first fed to several CNN-Pooling layers, this results a convolutional feature map of the image, Now, on application of K Means clustering on this feature map several clusters are generated, the quantity of feature-pixels in clusters are represented as a histogram(no. of pixel vs clusters), this representation of image is called Bag of Words(BoW) or Bag of local convolutional features. For finding similar images, BoW of input image is compared with that of other images in the database and images are ranked on the basis of highest value of localization score.

Chandrasekhar, Vijay and Lin, Jie and Morère, Olivier and Goh, Hanlin and Veillard, Antoine [4] stated that current State-of-the-art image instance retrieval pipelines consist of two major blocks: first, a subset of images similar to the query are retrieved from the database, and then geometric consistency checks are applied to select the relevant images from the subset with high precision. The first step is based on the comparison of global image descriptors thus global descriptors are key to improving retrieval performance.
[5] uses the above SOTA pipeline for instance retrieval. They used object detection CNN features from Faster R-CNN to extract both local and global features for given query image. Image wise pooling strategy is used to obtain image descriptors for both query and database images and a initial ranking is obtained based on cosine similarity between the two. After filtering stage, top N elements are locally analyzed using region-wise features from Faster R-CNN and re-ranked. This paper provides a simple baseline that uses off-the-shelf Faster R-CNN features to describe both images and their sub-parts.

One of the baseline model in instance retrieval is R-MAC[6] but it considers fixed spatial pooling which leads to higher computations as fixed grids considered to extract features may not even contain objects one is interested in. To overcome this problem [6] extended the R-MAC [7] approach for Instance Retrieval. Extended R-MAC produced a global image representation by aggregating the activation features of a CNN using a variable region pooling mechanism. They used three-stream Siamese network to optimize the weights of the R-MAC representation for the image retrieval task by using a triplet ranking loss.

Most of the approaches for instance retrieval were not able to produce promising results with low resolution images. *Razavian, Ali Sharif and Sullivan, Josephine and Carlsson, Stefan and Maki, Atsuto* proposed a solution to this in [8]. The paper [8] highlighted the availability of image representations based on convolutional networks and an efficient pipeline to extract local features by taking geometric invariance into explicit account. Variable footprints were used based on the size of the dataset and memory requirements. They made use of the last convolutional layer for the instance retrieval instead of taking the output of the fully connected layer which would have required to crop or edit the image again. Dimensionality of the features is reduced using spatial max-pooling and PCA which increases the efficiency of the model. Multi-resolution search and jittering is used to improve the instance retrieval capability. Using Multi-resolution search helps when the query image has a lower scale in the image and alternatively jittering helps when the query image has a larger scale in the image than the reference images in the database. Similarity between the feature vectors is calculated by calculating the sum of the distance of each query sub-patch.
The datasets used for Instance retrieval include INRIA Holidays, Oxford Buildings, UKBench and Graphics.
Among the work done for instance based retrieval it is observed that the results vary a lot when the query images provided are of low quality. We aim to improve the accuracy provided query images may not be similar to train images in terms of quality.
Bag-of-Visual-Words based methods fail to capture precise bounding box features for accurate results. The same limitation is observed in CNN-based features or SIFT based features. We look to improve this using current State of the art based techniques like Faster-CNN or VisualBert. Also Bag of words leads to a high dimensional feature vector which further delay the computation for large dataset.

4. AVAILABLE DATASETS

In our analysis, we found following datasets that we can use for training and testing our Deep Learning model:

a) Paris Dataset: This dataset contains 6412 images of 12 different landmarks of Paris, it was extracted from Flickr.

b) Oxford Dataset: This dataset contains 5042 images of 11 different Oxford buildings, collected from Flickr.

c) INSTRE Dataset: This is set of Datasets containing different pictures of architectures, buildings,toys, designs, paintings etc. There are three divisions of this datset S1, S2 and M.INSTRE-S1 and INSTRE-S2 contains 11011 images 12059 images respectively. Different objects/designs/scenes are annotated using bounding boxes. In these divisions,

a single image may contain multiple annotations(useful for complex Queries). On the other hand INSTRE-M dataset contains 5473 images strictly having only one annotation.

d) Sculpture Dataset: This dataset contains pictures of sculpters taken by Henry Moore and Auguste Rodin. It contains 6k images. The dataset is taken from Flickr. It contains equal number of images for train and test data. Query objects are chosen from 10 different sculptures for both sets. Accordingly query regions are defined for each of the objects, It provides 70 queries for performance evaluation, for query object 10 different sculptures are selected form each set and query regions and 7 images are defined these 10 objects.

5. BASELINE MODEL

We have used machine learning model for our proposed approach and model is trained on all the four mentioned datasets.

5.1 Machine Learning Based Model:

Initial Instance based retrievals by [2] used SIFT technique to extract features of images, after extracting features from each image of the dataset, we generate a visual codebook by applying K Mean Clustering on all features. Using this codebook we represent each image as Bag of these Visual Words(BoW). Retrieval is done in two steps, first we retrieve images which are having highest Cosine similarity of BoW with that of query image are highest. This phase gives us visually similar images, now we select top K images and re-rank them by finding cosine similarity within BoW of query image with sliding windows of that of top K images. This arranges the image containing the query instance in top rankings.

6. PROPOSED MODEL

We propose two model for our instance retrieval problem, both models are Deep Learning based model, and discussed in the upcoming subsections. Both the models are trained on the all four mentioned datasets.

6.1 Convolutional Features Based Image Retrieval:

Since ML based features are manually selected, they are not much efficient to extract relevant features of images of all domain.[3] provides a very efficient way of image retrieval which can be applied on large variety of images. In this retrieval technique features of an image is extracted using a pre-trained CNN layer. Following pipeline is followed for CNN based Image retrieval:

a)Image Preprocessing: Each image is fed into a pre-trained CNN model(VGG 16 in this case), and features are extracted from any of the Conv2D layers(usually later ones).

b)Clustering: Feature maps extracted from all images are clustered to generate codebook, each centroid of K Means algorithm is called a visual word.

c)Assignment Map and Bag of Words Representation: Each visual word is assigned a unique label and these labels are used to represent each feature map, this gives a 2D matrix called Assignment map. Now assignment map is used to count the number of visual word in each image, this count is used to make a vector histogram called Bag of Word vector.BoW is stored for each image of dataset for further retrieval.

d)Initial Image Retrieval: Image retrieval is performed by performing cosine similarity between BoW of query image and that of images of database.Top similarity score images are retrieved.

e)Local Re-ranking:In instance retrieval a bounding box region in query image is also given for finding images which contains this, so, for this we take top K images retrieved from first phase and apply a sliding window on database images,for each window we find cosine similarity of Bow of this image with that of query.On the basis of this score we return re-ranked images.

Bag of Word vectors size used is 200.

In this way we perform CNN based instance retrieval,results are given in Figure 3 and 4.

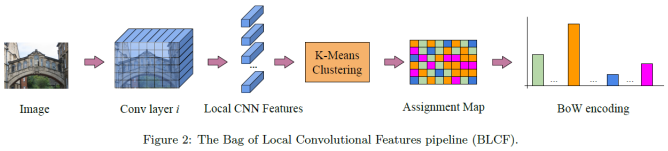

Figure 2: The Bag of Local Convolutional Features pipeline (BLCF).

Figure 2: The Bag of Local Convolutional Features pipeline (BLCF).

6.2 Faster RCNN

The proposed model in convolutional based feature retrieval requires high dimensional features for global re-ranking which is time consuming for real time ranking. To solve the problem of instance-based image retrieval we have used Region-proposal Faster RCNN network to capture only relevant information of the image. It retrieves top n images and re-ranks them based on the instance presence.

Methodology

Following pipeline is followed :

1)Filtering Stage: This stage uses Faster RCNN based model to extract the features form both query and database images.In this stage we built image descriptors using image-wise pooling for both the query image and the dataset. For retrieving the best similar images the descriptor of query image is compared with all descriptors of the database images. We used cosine similarity and weighted cosine similarity for ranking.

2)Spatial Re-ranking: In spatial re-ranking a sliding window with some defined window size and aspect ratio is used and it is slided all over the image and then compares it with bounding box features of the query image just like the bag of words model.

7. RESULTS

We have used MAP and NDCG for evaluation. The following are the results which we got using all three the models when tested on the datasets, the table represents the model's performance on different similarity measure. For figures, we have the first image as query image and the rest images are the retreived images against the query image.

DataSet	Mean Average Precision	NDCG
Oxford	0.0556	0.1824
Paris	0.506	0.3883
INSTRE	0.0627	0.2904
Sculpture	0.0654	0.3021

: Table 1: ML Based Model Similarity Cosine

DataSet	Mean Average Precision	NDCG
Oxford	0.0184	0.1667
Paris	0.0719	0.4296
INSTRE	0.0665	0.2826
Sculpture	0.1620	0.5384

: Table 2: ML Based Model Similarity Weighted Cosine

Figure 3: ML Method outputs: Instre Dataset

Figure 4: ML Method outputs: Oxford Dataset

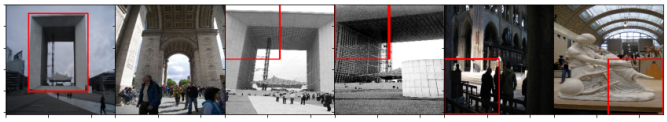

Figure 5: ML Method outputs: Paris Dataset

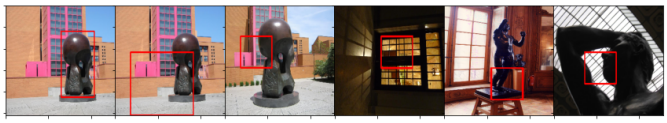

Figure 6: ML Method outputs: Sculpture Dataset

DataSet	Mean Average Precision	NDCG
Oxford	0.6963	0.7539
Paris	0.6296	0.8083
INSTRE	0.1862	0.3964
Sculpture	0.0654	0.3021

: Table 3: Convolutional Features Based Image Retrieval Based Model Similarity Cosine

DataSet	Mean Average Precision	NDCG
Oxford	0.6961	0.7535
Paris	0.6294	0.8061
INSTRE	0.0665	0.2826
Sculpture	0.1620	0.5384

: Table 4: Convolutional Features Based Image Retrieval: Similarity Weighted Cosine

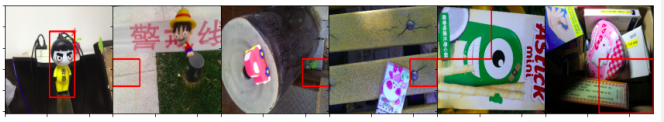

Figure 7: Convolutional Features Based Image Retrieval Method outputs: Instre Dataset

Figure 8: Convolutional Features Based Image Retrieval Method outputs: Oxford Dataset

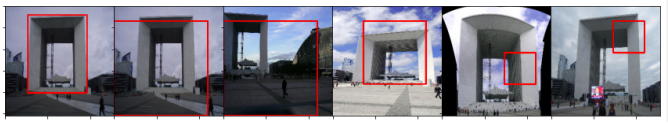

Figure 9: Convolutional Features Based Image Retrieval Method outputs: Paris Dataset

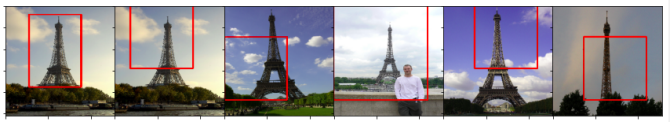

Figure 10: Convolutional Features Based Image Retrieval Method outputs: Sculpture Dataset

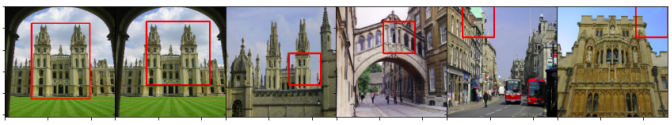

Figure 11: Faster RCNN outputs: Oxford Dataset

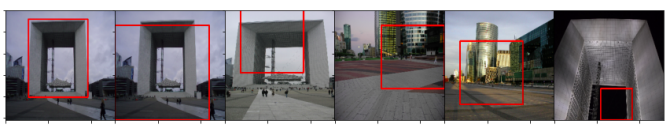

Figure 12: Faster RCNN outputs: Paris Dataset

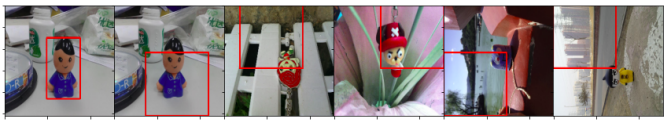

Figure 13: Faster RCNN outputs: INSTRE Dataset

Figure 14: Faster RCNN outputs: Sculpture Dataset

DataSet	Mean Average Precision	NDCG
Oxford	0.6515	0.7359
Paris	0.5608	0.7652
INSTRE	0.0909	0.2733
Sculpture	0.1324	0.5829

: Table 5: Faster RCNN Based Model Similarity Cosine

DataSet	Mean Average Precision	NDCG
Oxford	0.6214	0.7050
Paris	0.5008	0.7251
INSTRE	0.0709	0.2533
Sculpture	0.1314	0.5629

: Table 6: Faster RCNN Model Similarity Weighted Cosine

8. CONCLUSION

We have analysed three models for insatnce based retrieval, and we came to the conclusion that Faster-RCNN gives faster(due to local reranking) and better results as compared to the baseline ML and Dl models. But it has a limitation that when the query image is not of the domain of the trained dataset then the Faster RCNN can give results that are not much similar to the query image.

9. LIMITATIONS

Bag of word based model works on the similarity of visual words, For this representation,it requires storing features into a very high dimensional feature space, so more computation power is required to train and filter images, although it works flawlessly on GPU enabled systems.

10. FUTURE WORK

For the future prospects, we can take into considerations these techniques: 1.We can use LSTM based features with attention mechanisms, to filter and remember useful feature representation from the images. 2.We can meta train our models on dfferent datasets,so that feature extractor model can generalize itself to images of different domains and provide more robust feature maps.

11. REFERENCES

[1] S. R. Dubey, "A Decade Survey of Content Based Image Retrieval using Deep Learning," *arXiv:2012.00641 [cs]*, Nov. 2020.

[2] L. Zheng, Y. Yang, and Q. Tian, "SIFT Meets CNN: A Decade Survey of Instance Retrieval," *arXiv:1608.01807 [cs]*, May 2017.

[3] E. Mohedano, A. Salvador, K. McGuinness, F. Marques, N. E. O'Connor, and X. Giro-i-Nieto, "Bags of Local Convolutional Features for Scalable Instance Search," *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 327–331, Jun. 2016.

[4] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard, "A Practical Guide to CNNs and Fisher Vectors for Image Instance Retrieval," *arXiv:1508.02496 [cs]*, Aug. 2015.

[5] A. Salvador, X. Giro-i-Nieto, F. Marques, and S. Satoh, "Faster R-CNN Features for Instance Search," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 394–401.

[6] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "Deep Image Retrieval: Learning global representations for image search," *arXiv:1604.01325 [cs]*, Jul. 2016.

[7] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," 2016.

[8] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual Instance Retrieval with Deep Convolutional Networks," *arXiv:1412.6574 [cs]*, May 2016.