



Content-based Image Retrieval for Color Logo Images using Deep Learning Model

Latika Pinjarkar¹, Poorva Agrawal², Gagandeep Kaur³

1. latika.pinjarkar@sitnagpur.siu.edu.in

2. poorva.agrawal@sitnagpur.siu.edu.in

3. gagandeep.kaur@sitnagpur.siu.edu.in

^{1,2,3}: Department of Computer Science and Engineering,
Symbiosis Institute of Technology Nagpur,

Symbiosis International (Deemed University), Pune, India

Abstract: Content-based image retrieval (CBIR) has gained significant attention in recent years due to the exponential growth of digital images and the need for efficient image search and retrieval systems. Deep learning, with its ability to automatically learn hierarchical representations from raw data, has emerged as a powerful tool for CBIR tasks.

This paper proposes a CBIR system by merging DarkNet-19 and DarkNet-53 information for logo image retrieval. Experimentation is done using the Flickr Logos-47 dataset. The results revealed a significant enhancement in the retrieval results in terms of precision as compared with the standard image retrieval methods and the Deep Convolutional Neural Network(DCNN).

1. Introduction

An application of machine vision called image retrieval looks for related images within a group of images. It is extensively employed in daily life as well as a number of preservative computing and machine vision fields. Number of approaches have been accumulated as a result of decades of research and development. The three main kinds of image retrieval techniques are text based image retrieval [4], semantics based image retrieval [6] and CBIR [5]. In text-based image retrieval, text annotations are used. Without manual annotations, this type of retrieval process is poorly handled, and such methods neglect the visual content of the image and thereby lose sensory information. Traditional content-based image retrieval frequently makes use of low contrast characteristics including color features [7], GIST, texture features [8] and edge features [7].

Deep learning's improved feature learning capability and hierarchical representation of features have recently had an influence on and altered almost each field of computer vision, impelling and transmuting people's lives. This development and influence is being driven by the widespread use of deep learning concepts in research. In a range of tasks requiring computer vision, such as image recognition [11], object detection [12] and image classification [10] deep learning offers a number of benefits. Deep learning's outstanding performance in machine vision applications is fast changing and influencing human life. Several renowned companies, including Microsoft Google and Facebook as well as some renowned start-ups, have developed deep learning technologies and research centers. These businesses have

commercialized technologies such as video analytic technology, vision-based technology, driver technology, aided driving and facial recognition leading to improved functioning and a slow change in people's lifestyle. Deep learning has also had an influence on the fields such as Machine Learning, Natural Language Processing etc.

The convolutional neural networks (CNN) performance, in particular, has improved as a result of deep learning, which has a significant influence on the research in computer vision. Due to difficulties in image categorization on ImageNet datasets, CNN performance has improved recently. Alexnet, VGG, Googlenet, and Resnet are a few notable CNN models that have been proposed. As model depth increases, newly proposed models update the database of image categorization jobs. Generally, pre-trained CNN employing ImageNet datasets can be used right away for applications like feature extraction of images.

Given that CNN has shown strong feature encoding abilities, network learning through the ImageNet classification task has roughly invariance and can be employed to a varied kind of applications with related image database. Due to the overall advancement of deep learning models and the significance of extraction of features in Content Based Image Retrieval(CBIR) systems, we present visual features in CBIR learning model by means of convolutional neural networks (CNN). The suggested CNN utilizes this knowledge to image retrieval by learning key features extraction from the given dataset of images.

2. Related Work

The authors in [1] measured the degree of resemblance between two logos. This is accomplished by creating a database of logo images that is saved and derived from different sources of existing logo image data. The authors used Content Base Image Retrieval (CBIR) method to search images from database using Convolutional Neural Network (CNN) type Residual Network (ResNet-18).

Images are typically stored in a compressed manner to reduce their storage sizes since remote-sensing (RS) picture archives are continually expanding. Therefore, the majority of content-based RS image retrieval systems in use today require complete picture decoding (i.e., decompression), and are computationally required for large-scale archives. In order to solve this problem, authors in [2] developed a revolutionary method known as SCI-CBIR, which stands for simultaneous RS image compression and indexing. The suggested SCI-CBIR eliminates the demand for RS image decoding prior to image search and retrieval.

The authors in [3] introduced the SegNet architecture, a useful deep fully convolutional neural network designed for semantic pixel-wise segmentation. An encoder network, a related decoder network, and a pixel-wise classification layer make up the core trainable segmentation engine. The network converts

full input resolution feature maps from low resolution encoder feature maps for pixel-wise classification.

In the initial attempt at deep learning image retrieval, training was given to the deep autoencoder [13]. The deep autoencoder is trained using a Deep Belief Network (DBN)[14] and then evaluated on a small image dataset[14]. Since then, image categorization has seen considerable advancements thanks to deep learning, particularly the CNN[10]. The researchers claim that there is a great level of generality in the CNN model. The CNN model created for picture classification could be successfully used for various applications requiring recognition [[15], [16], [17]]. Due to CNN's universality, its influence on many facets of computer vision has continuously increased. The first CNN-based image retrieval challenge was completed in [18], however the outcomes were not adequate. The performance of CNN methodologies that of VLAD encoding on the basis of traditional SIFT and BoW when there is more spatial information available. In addition to improved retrieval efficiency, "multi-scale orderless collection" (MOP) [19] combines images blocks with numerous scales rather than CNN properties extraction from the whole image.

Alternatively, this method has trouble estimating the window size. Academics are now looking for ways to CNN properties extraction. Few pooling strategies, like R-MAC [21], methods for spatial and channel-aware weighted pooling [22] and max-pooling [20] were created by repurposing ideas from the original CBIR structure. Consideration of each local feature within the 3 Dimensional array with its dimension of depth called "super SIFT" is a more usual CBIR-like technique. Then, known VLAD or Fisher vector models are used to encrypt these chosen features, and the values of encoding are combined on the feature map [[19], [23]]. As an alternative to traditional CBIR, spatial search method based upon deep feature is represented. [24].

Another method to learn improved representations is to fine-tune CNN using new datasets and cutting [[25], [26], [27], [28]]. Although retrieval performance is excellent, it takes time to gather and categories new datasets. This is because in order to train the network, additional datasets relevant to a certain picture retrieval task must be gathered, and careful data selection or cleaning procedures must be used. Two recent works [25] and [26] for the retrieval of photographs of iconic buildings used the fine-tuning method. These two pieces of work both rely on additional datasets. [25] uses the labelled datasets that were retrieved from the Web via a search engine, but in their cleanest form.

The literature study discussed above gives a fundamental understanding of the creation of deep learning models, the progress of image retrieval techniques, and current developments in image retrieval mechanisms using Deep Learning. It helps with concentration on several important research concerns. The remainder of this

article is structured as follows. In Section 3, an introduction of two models: DarkNet-19 and DarkNet-53, which are used extraction of features and measuring similarity among the images is proposed. The proposed method is presented in Section 4, and the experimental findings and performance analysis are discussed in Section 5. Section 6 presents final observations and conclusions.

3.Prefaces

Recent studies have shown that CBIR methods based upon CNN gain superior low-level feature extraction techniques and produce superior outcomes. Numerous neural connections in 3-dimension height, breadth, and depth make up each CNN layer.

I) DarkNet-19

In order to classify objects, the Darknet-19 model [29] uses YOLOv2, this consists of nineteen conv2D layers, five max-pooling layers, and a soft-max layer. The Darknet-19 is a neural network framework built on CUDA and written in C. It swiftly recognizes objects; this is essential for prediction in real-time application. This network employs 1000 diverse class images to categorise. Consequently, the model has gained knowledge about specific image features by classifying a variety of image sets. The network can accept images up to 256x256 in size.

II) Darknet-53

Since the key feature extraction technique of the real-time entity identification network YOLOv3(You Only Look Once), Darknet-53 was presented in 2018. This model's objective is feature extraction of an image given as input. The main idea of this network can be formulated as a combination of the residual module and the basic feature extraction of YOLOv2. This model comprises five repeating blocks, each of which has two convolution layers (sized 1 and 33) and one residual layer on top of them. It has 53 convolutional layers, followed by a batch normalisation layer and a Leaky ReLU activation layer for each one.

Multiple image filters are convolved using the convolution layer to create a variety of feature maps. The maps of feature are down sampled with a convolutional layer using stride 2 and there is no pooling. It helps reduce loss of low-level features, which pooling is commonly associated with. As seen in the image, 1000 totally linked layers, an activation function, SoftMax and the Average pool layer could be included if classification like that in ImageNet was the intended outcome. The proposed work is focusing upon extraction of the pertinent images from the group of logo images.

III) Principal Component Analysis(PCA)

By condensing a vast number of possible outcomes into a small subset while keeping the rest of the data from the original set, PCA is a dimension-reduction approach that is

widely used to reduce the dimensions of enormous amounts of data. Data in the form of a p-dimensional variable, such as $v = (v_1, \dots, v_n)$, can be stated in a lower dimension, such as $d = (d_1, \dots, d_m)$, where $m < n$. The objective of PCA is to retain as much information as possible while reducing the number of alternatives in a data set.

$$\text{Input } J \in \mathbb{R}^{n \times m} \quad \dots(1)$$

Here J denotes Images Features. Rows of J has observation and columns contains variables.

$$\begin{aligned} \text{Coeff} &= \text{pca}(J) \\ J &= J X \text{coeff} \quad \dots(2) \end{aligned}$$

The dimension of coefficient matrix is $m \times m$. Each column in coefficient contains a major component values and the columns are arranged in descending order by element variance. Data is automatically centered by PCA, which employs the SVD method. It calls for the rectangular matrix M (where J is $n \times m$).

$$J = X_{n \times n} Y_{n \times m} Z_{m \times m}^T \quad \dots(3)$$

$$\begin{aligned} &\text{Here,} \\ &X^T X = I_{n \times n} \end{aligned}$$

$$\text{and } Y^T Y = I_{m \times m}$$

i.e X and Y are orthogonal.

The first step in SVD calculation is the calculation of eigenvalues and eigen vectors of $J J^T$ and $J^T J$. The columns in Z contains $J^T J$ eigenvectors., and columns in X contains $J J^T$ eigenvectors. Additionally the singular values in Y represents eigen values square roots from $J J^T$ or $J^T J$.

4. Proposed CBIR system

The proposed framework is a CBIR system for color logo images through fusion of features. The conventional method manually chooses many pieces for fusion after extracting the image's texture, color or other significant aspects. The drawback of this approach is that the same attribute affects picture retrieval differently in other datasets. In advance of image retrieval operations, it might be challenging to determine the features best appropriate for the particular datasets and the optimum features combination. Different combinations over the entire

datasets can be tested to improve the outcomes, but the process is extremely time-consuming, inefficient, and unstable.

In this study, we use the pre trained DarkNet-19 and DarkNet-53 models to extract picture features. The extracted features are combined to create a vector with new feature. We use the last convolutional layer in DarkNet-53, which is linked to every input neurons and preserves spatial information, to extract features. The average pooling layer, known as "Avg 1," is employed in DarkNet-19 as a feature extraction layer, producing a vector with new feature.

5. Experimental Framework

5.1 Background

A Logo image retrieval framework must be able to find images that match a given query by grouping them according to similarity. By selecting just a small subset at the top of the list, the logo examiners would be able to determine whether it contains images that are sufficiently similar to the query. The straightforward criterion for evaluating a logo image retrieval system is to determine its ability to successfully and effectively achieve this goal.

5.2 Database Selection

How to obtain trustworthy datasets for conducting experiments is one of the most important considerations when creating any image retrieval system because all results and outcomes depend on this dataset. A fair experimental dataset must be obtained in order to produce results that can be trusted. There are two ways to access this information. The first choice is to obtain information from a group of subject-matter experts (Logo examiners in the application sector), and the second choice is to gather information from a somewhat broader group of human subjects. The experimentation in the suggested framework were carried out utilising a FlickrLogos-47 logo dataset as described in the following table Table1.

Table 1: Details of the Flickr Logos datasets used in the experimentation

Dataset	Total No. Images	Brand	Class
FlickrLogos-47 Dataset	9200	47	47

5.3 Description:

The dataset FlickrLogos-47 contains images of logos from popular brands and is intended for testing logo detection and recognition software on real-world images. It is constructed from the FlickrLogos-32 dataset images that have undergone re-annotation

to correct mission annotations and add new classes. There are 47 distinct classes for logos containing the brands like Adidas (Symbol), Adidas (Text), Aldi, Apple, Becks (Symbol), Becks (Text), BMW, Carlsberg (Symbol), Carlsberg (Text), Chimay (Symbol), Chimay (Text), Coca-Cola, Corona (Symbol), Corona (Text), DHL, Erdinger (Symbol), Erdinger (Text), Esso (Symbol), Esso (Text), Fedex, Ferrari, Ford, Foster's (Symbol) and Foster's (Text) etc.

Some sample trademark images used in the framework from FlickrLogos-47 Dataset are represented in Figure 1.1.



Figure 1.1: Sample Logo images from FlickrLogos-47 Dataset

The experimental steps in the algorithmic form are illustrated as follows:

Algorithm 1.1 image Retrieval Algorithm:

Input: The QI-designated query Image. By using DBimg, an image database is identified. N represents the overall number of images.

Output: R How many relevant images were found for QI?

i) N: overall picture count in DBimg

ii) For $j = 1$ to N

an image $J \in P^{N \times R \times 3}$

Resize image to 224 by 224 pixels.

Utilizing the activation of the two models of N images from DBimg, extract high-level features. Create a database index, DB Index utilizing the generated features for all of the images.

iii) Feature Extraction of QI

iv) Feature Reduction using PCA.

v) For each image in the DB Index

Compare the query image's feature with every image in the DB Index using the similarity measure.












vi) Arrange the pertinent images according to QI features

vii) Retrieve R closed images for the given QI.

6. Retrieval Results

Top ten retrieved images for an input query are depicted in Table 2.

Table 2: Output Retrieved Images (top 10) for the given query image

INPUT QUERY IMAGE	OUTPUT RETRIEVED IMAGES				
					
					

The experimentation is evaluated using the precision and recall evaluation parameters and can be defined as follows:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

where:

TP = True Positive: case was positive and predicted positive

TN = True Negative: case was negative and predicted negative

FP = False Positive: case was negative but predicted positive







FN = False Negative: case was positive but predicted negative.





Table 3 depicts the results in terms of average precision for sample 10 query images using the DarkNet-19 model and the DarkNet-53 model and proposed framework and the results in terms of average recall are presented in Table 4.

Table 3. Retrieval results in terms of average precision for sample 10 query images

Query Image	DarkNet-19	DarkNet-53	Proposed Method
	80.75	81.4	84.9
	79.8	82.9	86.5
	91.3	93.7	94.7
	89	90.4	92.5
	92.7	89.9	91.7
	95.4	97.8	95.4
	98.7	89.9	93.8
	92.3	95.7	89.9
	93.9	96.8	96.7
	95.9	93.2	97.1
Average	81.4	91.17	92.32

Table 4. Retrieval results in terms of average recall for sample 10 query images

Query Image	DarkNet-19	DarkNet-53	Proposed Method
	15.35	15.02	19.53
	15.05	16.73	19.54
	17.17	18.97	18.3
	19	19	18.8
	18.45	18.9	19.9
	17.34	19.98	19.8

	18.45	19	18.8
	19.02	18.93	19.08
	19.98	19	19.9
	18.9	19.24	19.43
Average	17.871	18.477	19.31

The suggested framework by employing the fusion of features mechanism through the DarkNet-19 model and the DarkNet-53 model has outperformed in terms of average precision when compared with the existing deep learning based models as shown in Table 5.

Table 5. Relative analysis with other deep learning models

<i>Results/ Deep Learning Model</i>	<i>AlexNet[17]</i>	<i>VGG16[10]</i>	<i>ResNet[30]</i>	<i>GoogleNet[31]</i>	<i>Modified AlexNet [32]</i>	<i>Proposed Framework</i>
<i>Average Precision</i>	76.22	89.08	93.27	95.61	92.37	92.32
<i>Average Recall</i>	41.58	58.7	71.99	79.43	70.92	19.31

7. Conclusion

The aim of this study is to develop a system using multi-feature fusion mechanism for enhancing CNN illustration of feature, ensuring accurate images retrieval in the CBIR system for Logo images. A major issue in CBIR is feature extraction and representation because it is so diligently tied to perception of human. The Euclidean distance is employed to find relevant images within the dataset of Logo images. In the suggested method, we combine the darknet-19 and darknet-53 feature vectors to produce a new feature vector for logo images retrieval. The FlickrLogos-47 dataset was used in this experiment, outperforming other models in terms of retrieval results when evaluated on the basis of average precision and average recall.

References:

- [1] L. N. Rani and Y. Yuhandri, "Similarity Measurement on Logo Image Using CBIR (Content Base Image Retrieval) and CNN ResNet-18 Architecture," *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, Jakarta, Indonesia, 2023, pp. 228-233, doi: 10.1109/ICCoSITE57641.2023.10127711
- [2] G. Sumbul, J. Xiang and B. Demir, "Towards Simultaneous Image Compression and Indexing for Scalable Content-Based Retrieval in Remote Sensing," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-12, 2022, Art no. 5630912, doi: 10.1109/TGRS.2022.3204914.
- [3] V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
- [4] Rui Y, Huang T S and Chang S F 1999 *Journal of visual communication and image representation* **10** 39–62
- [5] Smeulders A W, Worring M, Santini S, Gupta A and Jain R 2000 *IEEE Transactions on pattern analysis and machine intelligence* **22** 1349–1380
- [6] Bradshaw B 2000 *Proceedings of the eighth ACM international conference on Multimedia* pp 167–176
- [7] Jain A K and Vailaya A 1996 *Pattern recognition* **29** 1233–1244
- [8] Manjunath B S and Ma W Y 1996 *IEEE Transactions on pattern analysis and machine intelligence* **18** 837–842
- [9] Mezaris V, Kompatsiaris I and Strintzis M G 2003 *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)* vol 2 (IEEE) pp II–511
- [10] Krizhevsky A, Sutskever I and Hinton G E 2012 *Advances in neural information processing systems* **25** 1097–1105
- [11] Lee H, Grosse R, Ranganath R and Ng A Y 2009 *Proceedings of the 26th annual international conference on machine learning* pp 609–616
- [12] Girshick R, Donahue J, Darrell T and Malik J 2014 *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 580–587
- [13] Krizhevsky A and Hinton G E 2011 *ESANN* vol 1 (Citeseer) p 2
- [14] Hinton G E, Osindero S and Teh Y W 2006 *Neural computation* **18** 1527–1554
- [15] Oquab M, Bottou L, Laptev I and Sivic J 2014 *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 1717–1724

- [16] Deng C, Ji R, Liu W, Tao D and Gao X 2013 *Proceedings of the IEEE International Conference on Computer Vision* pp 2600–2607
- [17] Zeiler M D and Fergus R 2014 *European conference on computer vision* (Springer) pp 818–833
- [18] Sharif Razavian A, Azizpour H, Sullivan J and Carlsson S 2014 *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* pp 806–813
- [19] Gong Y, Wang L, Guo R and Lazebnik S 2014 *European conference on computer vision* (Springer) pp 392–407
- [20] Babenko A and Lempitsky V 2015 *Proceedings of the IEEE international conference on computer vision* pp 1269–1277
- [21] Tolias G, Sivic R and Jégou H 2015 *arXiv preprint arXiv:1511.05879*
- [22] Kalantidis Y, Mellina C and Osindero S 2016 *European conference on computer vision* (Springer) pp 685–701
- [23] Yue-Hei Ng J, Yang F and Davis L S 2015 *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* pp 53–61
- [24] Razavian A S, Sullivan J, Carlsson S and Maki A 2016 *ITE Transactions on Media Technology and Applications* **4** 251–258
- [25] Gordo A, Almazan J, Revaud J and Larlus D 2017 *International Journal of Computer Vision* **124** 237–254
- [26] Radenović F, Tolias G and Chum O 2016 *European conference on computer vision* (Springer) pp 3–20
- [27] Gordo A, Almazan J, Revaud J and Larlus D 2016 *European conference on computer vision* (Springer) pp 241–257
- [28] Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J 2016 *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 5297–5307
- [29] Redmon J and Farhadi A 2017 *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 7263–7271
- [30] Maji S and Bose S 2021 *ACM/IMS Transactions on Data Science (TDS)* **2** 1–24
- [31] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 *Proceedings of the IEEE conference on computer vision and pattern recognition* pp 1–9
- [32] Shakarami A and Tarrah H 2020 *Optik* **214** 164833