Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 2, pp. 415–425
DOI : 10.47974/JIOS-1924

Optimized approach for efficient image retrieval using Resnet50 and CBAM

Pilli Mounika *
Department of Computer Science and Engineering
Jawaharlal Nehru Technological University
Kakinada 533003
Andhra Pradesh
India

K. Venkata Subba Reddy †
Department of Computer Science and Engineering (AI&ML)
Vidya Jyothi Institute of Technology
Hyderabad 500075
Telangana
India

N. Ramakrishnaiah §
Department of Computer Science and Engineering
University College of Engineering
Kakinada 533003
Andhra Pradesh
India

Abstract

The rapid expansion of multimedia repositories surpasses the capabilities of traditional data analysis. An advanced image retrieval (IR) model is essential for accurately retrieving relevant images. Numerous machine learning and deep learning models have been proposed to address this challenge effectively. Even though the deep learning models are so powerful in extracting the image features, they are getting failed in prioritizing the most informative parts of the feature maps and also ignored the focus on specific channels or spatial regions. This will lead to the less effective feature representations of images, which contain varying levels of importance across different regions or channels. As a result, the

* *E-mail:* **mounika260328@gmail.com (Corresponding Author)**
† *E-mail:* **kvsreddy2012@gmail.com**
§ *E-mail:* **nrkrishna27@gmail.com**

retrieval performance may suffer because the model may not effectively differentiate between similar images. To address this point, this paper concentrated on refining the feature maps by applying channel and spatial attention, which helps in highlighting the most informative features and suppressing less important ones. This has become possible by integrating the Convolutional block of attention module (CBAM) and ResNet-50 models. CBAM is integrated with ResNet-50 by adding attention modules after each residual block. These modules apply channel and spatial attention mechanisms to the feature maps, refining them before passing to the next layer. The proposed model is verified on benchmark databases Corel-10k, CIFAR. The significant performance of the proposed method shown over existing deep learning models.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *CBAM, ResNet-50, Image retrieval, Attention mechanism, Corel-10k, CIFAR.*

1. Introduction

The advancement of technology has significantly increased the creation and availability of image data on the internet. In this context, the retrieval of similar images has become a major challenge across various applications, including research, education, e-business, and healthcare etc. To solve this problem urgently an efficient image retrieval (IR) technique or an algorithm is required [1]-[2]. Image retrieval technology is related to computer vision which is used to retrieve similar images from multimedia. Images generally consist of high-dimensional features that accurately represent the objects within them. Various distance measures are employed to determine the similarity between images, enhancing the accuracy and efficiency of the retrieval process [3]. Image retrieval involves two main phases: feature extraction and similarity-based image retrieval. In the first phase, various approaches such as local patterns, SIFT, moments, and Convolutional Neural Networks (CNNs) are employed for feature extraction [4]-[7]. In the second phase, the retrieved images are sorted in descending order based on their similarity metric values. The chosen metric should ideally be highly discriminative, consistent, and efficient to ensure optimal performance of the image retrieval model [8]. The CNN model's performance is negotiated when it gets transferred from the task of classification to retrieval [9]. The CNNs used for IR majorly focus on both training and testing phases to develop a strong image features set which represent the image accurately. Consequently, after addressing the challenges associated with model testing, the subsequent task involves identifying objects and shapes within the images [10]-[11]. A well-defined training dataset with diverse and representative patterns allows the model to generalize better and perform more accurately during retrieval tasks [12]. Once the model trained, it will be tested using test dataset to assess its performance in real-world scenarios. The precision of this stage reveals the significance of the model upon the image features which are extracted by it [13]. After successfully addressing the training and testing challenges, the focus shifts to more complex tasks such as object, face recognition, human pose invariant face recognition and also used in other domains like speech recognition, natural language processing [14]-[17]. Noh et.al., [18] has considered the spatial attention mechanism along with CNN model, has given good precision performance for image retrieval task. The spatial attention mechanism has tried to capture key points in the hand-crafted image [19]. Many pretrained DL models and transfer learning models are used for image retrieval on the consideration of effective feature extraction capability. In the keen observation it has been identified that, these models treat all features uniformly, potentially including irrelevant or noisy features that degrade performance. To overcome these problems in the conventional models, this paper has come with a new approach called attention mechanism. To overcome these challenges, this paper designs an integrated environment with efficient feature extraction model and attention modules. It has become possible by taking the CBAM along with ResNet-50 model. The paper contributions are as follows.

1. Preprocessing the data for training by scaling pixel values to a range suitable for the neural network and converting categorical labels into a format that can be used for training a classification model.
2. Define CBAM block and integrate it to the ResNet-50 architecture to enhance the feature extraction capabilities by focusing on important features of an image.
3. Compile and train the integrated model on two benchmark datasets.
4. After training, construct the feature vector database using the output from the global average pooling layer of the integrated model.

Other sections of the paper are described as follows: section 2 presents significance of the models under the related work, proposed approach details mentioned in section 3, section 4 with experimental analysis on two benchmark datasets and finally conclusion discussed in section 5.

2. Related Work

2.1 *ResNet-50*

In the revolution of computer vision the ResNet-50 deep learning model has occupied a significant position for many applications. Among many CNN models the residual network (ResNet) has a prominent architecture and it is introduced by

2.2 *Convolutional Block Attention Module (CBAM).*

Woo et al. [20] anticipated a Convolutional Block Attention Module (CBAM), which tries to improve the convolutional block (CB) performance by integrating the attention mechanism. CBAM model has crossed a significant success rate in many applications. It has undergone with many changes like light-weight dual-path attention block in multi-level attention network [21].

2.3 *Channel Attention Module (CAM)*

The CAM defined here by utilizing the Average pooling (AP) and max pooling (MP) to generate two features GAP and GMP as shown in Figure 1.

The inputs of CAM

$F \in \mathbb{R}^{HXWXC}$ is the input feature map, where H, W and C are height, width and no. of channels.

The outputs of CAM

Channel attention feature map $M_c(F) \in \mathbb{R}^{1X1XC}$

1. Global Average Pooling (GAP) and Global Max Pooling (GMP):

$$F_{avg} = GAP(F) = \frac{1}{HXW} \sum_{i=1}^{H} \sum_{j=1}^{W} F(i, j, k)$$

$$F_{max} = GMP(F) = max_{i,j} F(i, j, k)$$

2. Shared MLP layers

$$M_{lp} = ReLU(W_0 F_{avg})$$

$$M_{lp} = W_1 M_{lp}$$

$$M_{lp} = ReLU(W_0 F_{max})$$

$$M_{lp} = W_1 M_{lp}$$

3. Combination and Sigmoid Activation

$$M_C(F) = \sigma(M_{avg} + M_{max})$$

4. Element wise manipulation

$$F_C = F \otimes M_C(F)$$

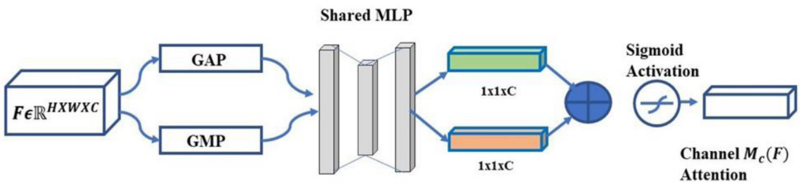

Figure 1

Framework of CAM

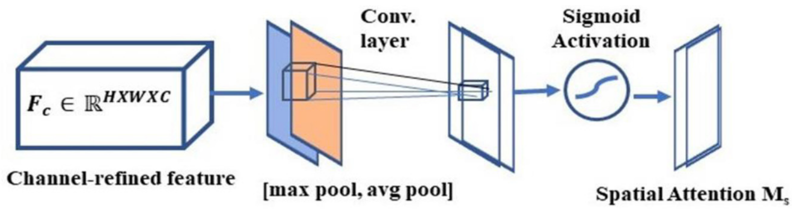

Figure 2

Framework of SAM

The framework of spatial attention module (SAM shown in Figure 2. The CBAM uses the both CAM and SAM for refining the features sets.

3. Proposed Approach

The proposed approach is described in this section for the implementing the optimized image retrieval and this framework is presented in Figure 3.

3.1 *Features Extraction*

After the pre-processing, the next step is to fine-tune a pre-trained ResNet-50 model using CBAM on Corel-10k and CIFAR datasets. At each specified convolutional layer, the channel attention focuses on their inter-channel relationships of feature maps, enhancing important features and the spatial-attention concentrates on inter-spatial relationships of feature maps, highlighting the important while suppressing the less important one. Hence the fine-grained feature database can be extracted from the images, which makes the model superficial in reducing semantic gap.

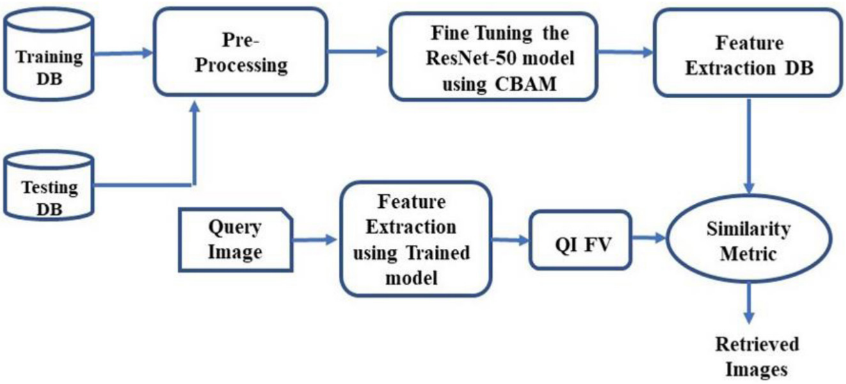

Figure 3

Framework of integrated proposed model

After the model is trained and tested, a feature vector database is constructed for the second stage of the image retrieval task. For each query image, features are extracted using the trained model, and a cosine similarity metric is applied to find similar images in the database. The retrieved images are then indexed based on the similarity metric values in descending order.

3.2 *Similarity Assessment*

Cosine similarity metric (CSM) is one of the popular metrics used to measure the distance between two non-zero vectors in an inner product space [22].

4. Result Analysis

4.1 *Experimental Setup*

To run the experiment, we have taken advantage of Google Colab Pro, utilizing an NVIDIA T4 GPU, along with an additional 100GB of disk space.

4.2 *Datasets*

This paper has taken two benchmark datasets Corel 10k [23] and CIFAR-100 [24] datasets. The Corel 10k dataset consists of 100 distinct categories, which represent different aspects of everyday life and visual appearances and for each category 100 different images. The other bench mark dataset, CIFAR-100 and it is designed primarily for classification and significant number of images per category made this database is more suitable for image retrieval.

4.3 *Performance Indicators*

Estimation of the performance for the methods is evaluated using mean average precision, recall, F1-measure and accuracy as given Eqn. (1) & Eqn. (2).

$$ARP = \frac{1}{D_N} \sum_{i=1}^{D_N} P_i \tag{1}$$

$$ARR = \frac{1}{D_N} \sum_{i=1}^{D_N} R_i \tag{2}$$

Where, DN is total count for the frames or images from the give database, and precision defined with the variable Pi for each category, Ri is recall for each category.

4.4 *Performance Analysis*

Training was carried out for the pretrained ResNet-50 model weights on Corel-10k and CIFAR-100 dataset on 0.2 ratio. Comparison of PM over existing methods on Corel-10k shown in Table 1.

The channel attention and spatial attention is applied on each residual block of the ResNet to concentrate more on the important regions of the image. The distance metric is used cosine similarity. Figure 4 shows a significant improvement of the proposed method in terms of precision and recall for every specific number of retrieved images over existing deep

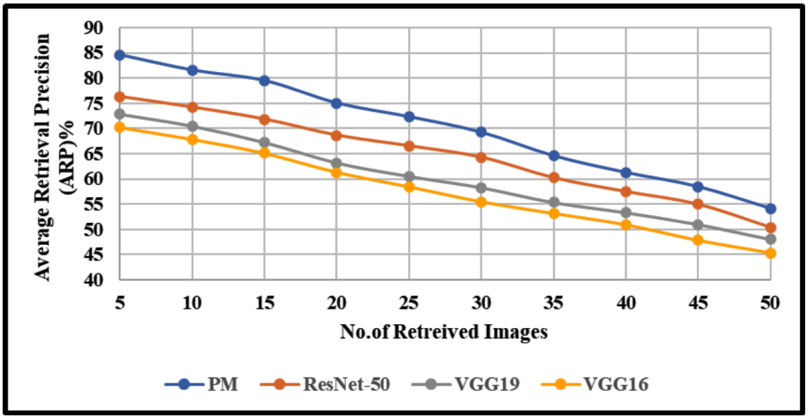

Figure 4

Average Retrieval Precision (ARP) on CIFAR-100 Dataset

Table 1

Comparison of PM over existing methods on Corel-10k

No.of Images	VGG16	VGG19	ResNet-50	PM
5	78.62	81.21	84.86	96.41
10	71.14	74.65	77.25	95.1
15	66.54	68.23	71.8	94.25
20	61.24	65.11	68.53	93.22
25	58.63	60.23	64.23	92.14
30	54.98	58.64	63.43	88.54
35	53.22	56.37	59.06	86.09
40	48.11	50.29	55	79.23
45	42.16	46.35	51.86	72.16
50	38.87	41.89	46.38	65.08

learning models. The same of the comparative analysis has given in Table 1.

5. Conclusion

This research investigates the effectiveness of integrating the Residual Network (ResNet-50) with the CBAM for image retrieval tasks, specifically focusing on the Corel-10k and CIFAR-100 datasets. The combined model demonstrates significant improvements in retrieval performance, highlighting the benefits of incorporating attention mechanisms in deep learning models. The attention mechanism allows the model to capture more relevant and discriminative features, which are crucial for accurate image retrieval. By emphasizing salient parts of the image, CBAM helps in distinguishing between similar and dissimilar images more effectively. The CBAM-ResNet50 model outperforms the standalone ResNet-50 model in terms of precision and recall. The attention mechanism in CBAM improves the model's ability to retrieve relevant images, leading to higher precision and recall scores. The precision-recall curves further validate the superiority of the CBAM-ResNet50 model, consistently providing more relevant results in top-N retrieval scenarios. The model performance has been tested on Corel-10k and CIFAR-100 datasets, and proved with significant performance over standalone methods. This research can be extended by adding modification text information for more accurate retrieval.

References

[1] D. Jiang and J. Kim, "Texture image retrieval using DTCWT-SVD and local binary pattern features," *Journal of Information Processing Systems*, vol. 13, no. 6, pp. 1628–1639 (2017).

[2] G. Sucharitha and R. K. Senapati, "Biomedical image retrieval by using local directional edge binary patterns and Zernike moments," *Multimedia Tools and Applications*, vol. 79, no. 3, pp. 1847–1864 (2020).

[3] Z. Zhu, C. Zhao, and Y. Hou, "Research on similarity measurement for texture image retrieval," e45302 (2012).

[4] G. Sucharitha and R. K. Senapati, "Local extreme edge binary patterns for face recognition and image retrieval," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, pp. 644–654 (2018).

[5] D. Sudarvizhi, "Feature based image retrieval system using Zernike moments and Daubechies Wavelet Transform," in 2016 International Conference on Recent Trends in Information Technology (ICRTIT), IEEE (2016).

[6] G. Sucharitha, R. K. Senapati, and A. B. Ranjan, "Secure and efficient content-based image retrieval using dominant local patterns and watermark encryption in cloud computing," *Cluster Computing*, pp. 1–17 (2024).

[7] M. S. Ghaleb, A. H. Rahman, and L. C. Park, "Image retrieval based on deep learning," *Journal of System and Management Sciences*, vol. 12, no. 2, pp. 477–496 (2022).

[8] S. R. Dubey, "A decade survey of content-based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704 (2021).

[9] Z. Rian, V. Christanti, and J. Hendryli, "Content-based image retrieval using convolutional neural networks," in 2019 IEEE International Conference on Signals and Systems (ICSigSys), *IEEE* (2019).

[10] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1309–1322 (2017).

[11] M. S. Basha, S. K. Mouleeswaran, and K. Rajendra Prasad, "Hybrid visual computing models to discover the clusters assessment of high dimensional big data," *Soft Comput*, vol. 27, pp. 4249–4262 (2023), doi: 10.1007/s00500-022-07092-x.

[12] S. A. Vassou, G. I. Papadopoulos, and C. D. Styliadis, "CoMo: a compact composite moment-based descriptor for image retrieval," Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (2017).

[13] C. Zhang and J. Liu, "Content Based Deep Learning Image Retrieval: A Survey," Proceedings of the 2023 9th International Conference on Communication and Information Processing (2023).

[14] S. Reddy K. and K. Rajendra Prasad, "An Extended Fuzzy C-Means Segmentation for an Efficient BTD with the Region of Interest of SCP," *IJITPM*, vol. 12, no. 4, pp. 11–24 (2021), doi: 10.4018/IJITPM.2021100.

[15] Y. Luo, S. Liu, and X. Zhang, "LSTM pose machines," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018).

[16] C. Vishnu, P. Jeripothula, R. Datla, S. Babu Ch., and C. Krishna Mohan, "mSODANet: A Network for Multi-Scale Object Detection in Aerial Images using Hierarchical Dilated Convolutions," *Pattern Recognition*, pp. 108548 (2022), doi: 10.1016/j.patcog.2022.108548.

[17] K. Subba Reddy, K. Rajendra Prasad, G. R. Kamatam, and R. K. Gupta, "An extended visual methods to perform data cluster assessment

in distributed data systems," *J Supercomput*, vol. 78, pp. 8810–8829 (2022), doi: 10.1007/s11227-021-04243-z.

[18] S. Wilson and C. Krishna Mohan, "An information bottleneck approach to optimize the dictionary of visual data," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 96–106 (2018), doi: 10.1109/TMM.2017.2716835.

[19] N. Perveen, D. Roy, and C. Krishna Mohan, "Facial Expression Recognition in Videos using Dynamic Kernels," *IEEE Transactions on Image Processing*, vol. 29, pp. 8316–8325 (2020), doi: 10.1109/TIP.2020.3011846.

[20] K. Shaheed, M. A. Yaseen, and M. A. Aslam, "EfficientRMT-Net—An Efficient ResNet-50 and Vision Transformers Approach for Classifying Potato Plant Leaf Diseases," Sensors, vol. 23, no. 23, pp. 9516 (2023).

[21] S. Woo, J. Park, and I. Y. Choi, "CBAM: Convolutional block attention module," Proceedings of the European Conference on Computer Vision (ECCV) (2018).

[22] L. Deng, Y. Li, Z. Zhan, and X. Xie, "Multi-level attention network: Mixed time–frequency channel attention and multi-scale self-attentive standard deviation pooling for speaker recognition," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107439 (2024).

[23] Corel, "Corel 10K Dataset," *Corel Corporation* (2004).

[24] A. Krizhevsky and G. Hinton, "Learning Multiple Layers of Features from Tiny," [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html.

Received November, 2024