Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 2, pp. 427–437
DOI : 10.47974/JIOS-1925

Ensemble deep learning technique for optimized informative query assessment for visual images

Komuravelli Mounika *
Department of Computer Science and Engineering
Jawaharlal Nehru Technological University
Hyderabad 500085
Telangana
India

B. V. RamNaresh Yadav †
Department of Computer Science and Engineering
Jawaharlal Nehru Technological University
Hyderabad 500085
Telangana
India

Abstract

Traditional data analysis is not up to pace with the quick rate at which multimedia repositories are growing. For finding pertinent images with optimized accuracy, a sophisticated visual informative retrieval (VIR) model is necessary. Currently, the techniques of deep learning are playing a vital role for facing this challenging issue. In addition to ignoring the emphasis on particular channels or locations, deep learning models prioritize the most useful sections of the feature maps. This will result in optimal feature representations of visual images that are less effective since they have various levels of relevance across different channels or regions. This work aimed to address this issue by focusing on improving the feature maps through the use of channels. The experiments conducted with benchmarked datasets for validating the proposed work. The comparative analysis demonstrated that proposed ensemble deep learning technique outperformed the existing deep learning techniques with notable optimized informative query assessment results.

Subject Classification: *Primary 93A00, Secondary 94A00.*

Keywords: *Image retrieval, Convolution neural network, Image queries, Self attention, Ensemble deep model.*

* *E-mail:* **komuravellimounika33@gmail.com (Corresponding Author)**
† *E-mail:* **bvramnaresh@jntuh.ac.in**

© Taru Publications

1. Introduction

Technological advancements have considerably boosted the generation and accessibility of visual information on the global web. In this scenario, the retrieval (or assessment) of identical images [1] has become an important obstacle in a variety of uses, like research, education, e-business, and medical. Several distance measurement techniques are used to estimate image similarity, increasing retrieval precision as well as effectiveness [2]. There are two basic steps of image retrieval: extraction of features and similarity-based retrieval. In the initial stage, the extraction of features is done using a variety of techniques, including local patterns, SIFT, moments, and Convolutional Neural Networks (CNNs) [3], [4], [5], [6]. In the second phase, the obtained images are organized in descending sequence according to their likeness metric scores. In an effort to catch important details in the carefully composed image, the spatial attention mechanism has been used [7]. For the purpose of recalculating the weights to do image similarity, [8] research has used second-order spatial attention. However, the system has to perform additional calculations and delays due to the extremely large feature dimensions that they obtained. Pre-trained models are increasingly being used for fine retrieval in subsequent research, particularly for feature extraction [9]. Several pretrained deep learning models and transfer learning models are employed for image retrieval, taking into account the efficiency of feature extraction. It has been observed with great insight that these models treat all features equally, possibly including features that are noisy or irrelevant and lowering performance. Additionally, these models are less able to adjust to changes in the dataset, such as adjustments to scale, lighting, and object placement. This research has proposed a novel solution termed the attention mechanism to address these issues with a cohesive framework by combining the ResNet-50 model [10] with the Convolutional Block of Attention Module (CBAM) has made it feasible.

Summary of the key contributions of the work are as following:

1. Scaling pixel values to a range fit for the neural network and translating category labels into a format fit for training a classification model helps preprocessing the data for training.
2. Focusing on significant picture features, build CBAM component and integrate it into the ResNet-50 framework to improve the extraction of features.

3. Perform effective training with the proposed integrated deep model.
4. Utilizing the results from the integrated model's global average pooling layer, create the feature vector database after training.

The remaining part of the paper has been organized as follows: Section 2 explores the background study, while Section 3 presents the proposed approach. Section 4 describes experimental results and analysis, and Section 5 presents the conclusion of the work.

2. Background Study

The ResNet-50 deep learning model has taken prominence in the revolutionary development in visual computing. The residual network (ResNet), one of the various CNN models, is well-known for its architecture and was initially introduced by He et al. in [11]. Because it strikes a compromise between complexity and computing efficiency among the different ResNet variants, ResNet-50, which contains 50 layers, has become very popular. By using residual learning, ResNet-50 has become known for its ability to solve the vanishing gradient problem. The ability to incorporate shortcut connections that skip a few layers has made it feasible. The framework enables the training to delve into incredibly deep networks in order to extract the more correlated elements from visual content because of these skip connections. Using this ResNet-50 model, Zhang et al. [12] developed a novel method for identifying flaws in steel surfaces. In order to classify illnesses in potato plant leaves, Kashif et al. [13] introduced a novel method called EfficientRMT-Net, which combines the advantages of ResNet-50 for feature extraction and Vision Transformers for capturing long-range dependencies in the image data. The work of Lin et al. [14] focuses on improving the ResNet-50 model such that it may be specifically used to identify diabetic retinopathy (DR) in retinal pictures. Enhancing DR detection efficiency and accuracy is the goal in order to support early diagnosis and treatment. To classify the X-ray image, [15] uses an integration of CNNs, Inception v3 and ResNet-50. A modified version of the ResNet-50 architecture was published by Duan et al. [16] with the express purpose of forecasting the flow fields surrounding an underwater vehicle.

Deep learning models, often known as CNNs, have achieved notable success in categorization and prediction-based tasks. Numerous deep learning models are available for analyzing an image's visual content. Many studies have examined changing the depth, breadth, and

fundamental character of models in order to further enhance their performance. The CBAM model has achieved a noteworthy success rate across numerous applications. Numerous modifications have been made to it, such as the multi-level attention network [17] and the lightweight dual-path attention block [18]. Basically, CBAM enhances the feature representation and it is done by channel and spatial attention mechanisms as shown in Figure 1.

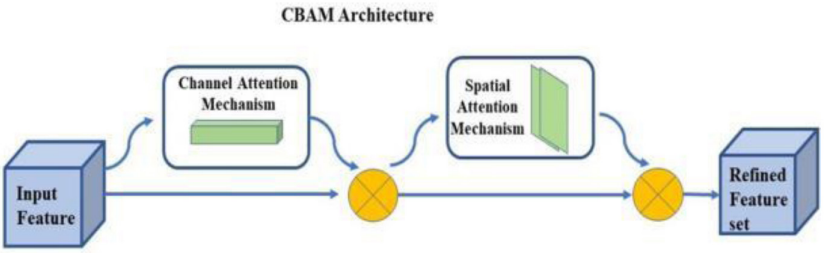


Figure 1
CBAM Architecture

Spatial Attention Mechanism

The Spatial attention module is explained here as shown in Figure 1. This attention is a harmonizing step to the CAM. The average and max pooling are applied again to channel-refined activation map *Ms (F)*, and it is given in Eqn. (1) to (5)

The inputs to SAM are

Channel-refined feature map $F_c \in \mathbb{R}^{HXWXC}$

The output of SAM is

Spatial attention feature map $M_s(F_c) \in \mathbb{R}^{HXWX1}$

1. Pooling Operations

$$F_c^{avg} = \frac{1}{C}\sum_{k=1}^{C} F_c(:, :, k) \quad (1)$$

$$F_c^{max} = max_k \; F_c(:, :, k) \quad (2)$$

2. Concatenation

$$F_{concat} = [F_c^{avg}; F_c^{max}] \in \mathbb{R}^{HXWX2} \tag{3}$$

3. Convolutional layer

$$M_s(F_c) = \sigma(Conv2D\ (F_{concat}; k = 7) \tag{4}$$

4. Element wise manipulation

$$F_{cs} = F_c \otimes M_s(F_c) \tag{5}$$

3. Proposed Optimized Image Query Approach

This section presents the proposed integrated (or ensemble) deep learning for optimization of image query results. To ensure peak performance throughout the training and testing stages, all database images are standardized and normalized during the pre-processing stage. Before moving on to the next phase of the method, this involves reducing the data to uniform dimensions and performing normalizations. Using CBAM on the Corel-10k and CIFAR datasets, a pre-trained ResNet-50 model is fine-tuned following the pre-processing phase. Table 1 displays ResNet-50 and CBAM residual block wise integration. Channel attention concentrates on the inter-channel relationships of feature maps at each designated convolutional layer, emphasizing important features, while spatial attention concentrates on the inter-spatial relationships of feature maps, emphasizing the important while suppressing the less important one. Because of this, it is possible to extract the fine-grained feature database from the images, which only slightly reduces the semantic gap in the model. In order to fine-tune the model for the particular job of image retrieval, it is trained on pre-processed images from the Training and Testing databases.

A feature vector database is created for the second phase of the picture retrieval job once the model has been trained and evaluated. The trained model extracts feature for every query image, and then it uses the cosine similarity measure to discover images in the database that are comparable. After that, the recovered images are indexed in descending order according to the similarity metric values. For finding the similarity features, the cosine distance metrics are used in the proposed approach. Every residual block of the ResNet is given both channel attention and spatial attention in order to focus more on the image's key areas.

Table 1

Detailed information for the Resnet-50 model integrated with specific attention types of CBAM

Stage	Layer Type	Configuration	Attention Type	Output Shape
Input	Input Layer	-	-	(224, 224, 3)
Conv1	Convolution	7x7, 64, stride 2	-	(112,112, 64)
	Batch Normalization	-	-	(112, 112, 64)
	ReLU	-	-	(112, 112, 64)
	Max Pooling	3x3, stride 2	-	(56, 56, 64)
Conv2_x	Residual Block	3x3, 64, stride 1	Channel Attention	(56, 56, 64)
	CBAM	Channel Attention	Channel Attention	(56, 56, 64)
Conv3_x	Residual Block	3x3, 128, stride 2	Spatial Attention	(28, 28, 128)
	CBAM	Spatial Attention	Spatial Attention	(28, 28, 128)
Conv4_x	Residual Block	3x3, 256 stride 2	Channel & Spatial	(14, 14, 256)
	CBAM	Channel Attention	Channel Attention	(14, 14, 256)
	CBAM	Spatial Attention	Spatial Attention	(14, 14, 256)
Conv5_x	Residual Block	3x3, 512, stride 2	Channel & Spatial	(7, 7, 512)
	CBAM	Channel Attention	Channel Attention	(7, 7, 512)
	CBAM	Spatial Attention	Spatial Attention	(7, 7, 512)
Output	Global Average Pooling	-	-	(512)
	Dense	1000	-	(1000)
	Softmax	-	-	(1000)

4. Experimental Results and Analysis

Two benchmark datasets, Corel 10k [19] and CIFAR-100 [20], were used in this research. The 100 unique categories that comprise the Corel 10k dataset each depict a different facet of daily existence and appearance, and there are 100 photographs in each category.

The proposed method's effectiveness is assessed using the mean of precision, recall, F1-measure, and accuracy given in Eqn. (6) and (7).

$$Mean\ Precision = \frac{1}{D_N}\sum_{i=1}^{D_N} P_i \tag{6}$$

$$Mean\ Recall = \frac{1}{D_N}\sum_{i=1}^{D_N} R_i \tag{7}$$

where Pi represents precision for each category, Ri represents recall for each category, and DN is the total number of images in the database. The pretrained ResNet-50 model weights were trained using the CIFAR-100 and Corel-10k datasets at a ratio of 0.2. Each residual block of the ResNet is given both channel attention and spatial attention in order to focus more on the key areas of the image. Cosine similarity is the distance metric utilized.

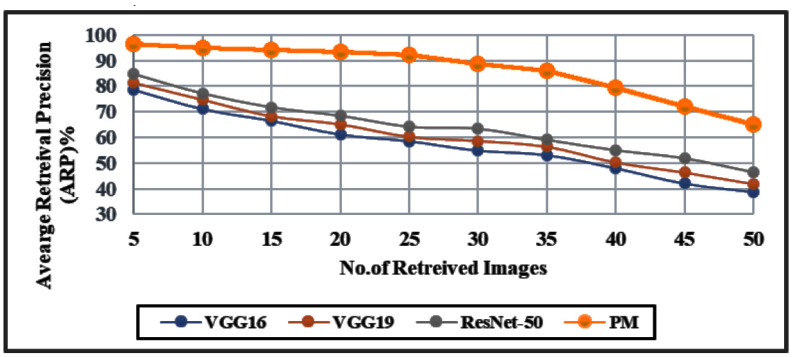


Figure 2

Mean Precision or Average Retrieval Precision (APR) on Corel-10k

The dataset is divided into categories such as objects, animals, and different man-made and natural scenes, which aids in assessing how reliable image categorization and retrieval techniques. The CIFAR-100 benchmark dataset is primarily intended for classification; however, due to its wide range of categories and substantial image count for each category, this database is better suited for image retrieval. The dataset has

600 distinct color photos divided into 100 classes. 60,000 photos in total, of which 50,000 are used for training and 10,000 for testing.

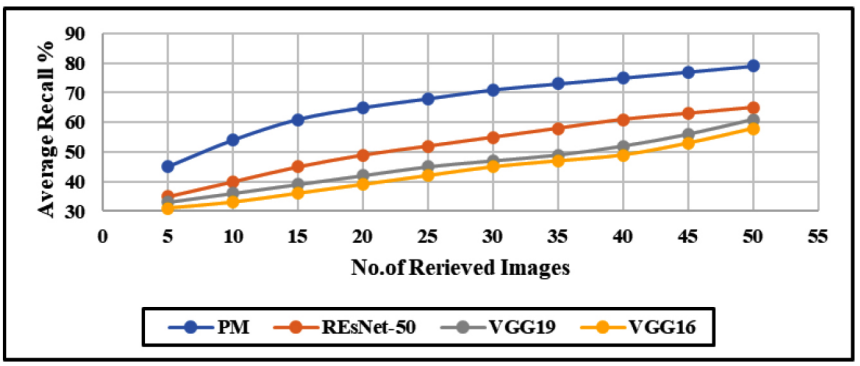


Figure 3

Mean Recall or Average Retrieval Recall (ARR) on Corel-10k

Table 2 presents the same comparative analysis using the accuracy for the dataset of CIFAR-100. In this analysis, it noted that proposed method of ensemble approach achieved the high rate of precision and recall values for Corel-10k dataset and achieved the highest rate of accuracy in image query assessment results for the dataset of CIFIR-100.

Table 2

Optimized Query Assessment Results of CIFAR-100 Dataset

No.of Images	VGG16	VGG19	ResNet-50	Proposed Approach
5	70.21	72.85	76.32	84.58
10	67.86	70.49	74.23	81.56
15	65.09	67.24	71.84	79.47
20	61.32	63.18	68.65	75.03
25	58.45	60.51	66.54	72.31
30	55.48	58.26	64.28	69.22
35	53.16	55.31	60.21	64.57
40	50.91	53.28	57.45	61.25
45	47.82	50.94	54.94	58.45
50	45.34	47.98	50.34	54.13

For each given number of retrieved images, the proposed approach significantly outperforms the state-of-the-art deep learning models in terms of precision and recall and its analysis shown in figure 2 and figure 3 for the top 10 retrieved results for a query image for the dataset of corel-10k.

5. Conclusion and Future Scope

This research proposed the efficacy of integrating the Convolutional Block Attention Module (CBAM) with the Residual Network (ResNet-50) for optimized visual query image assessment tasks. By combining the two models, can see how adding attention processes to deep learning models improves retrieval performance significantly in the experimental of this paper. Accurate visual query assessment relies on the model capturing relevant and discriminative features. Both recall and precision are improved upon by the CBAM-ResNet50 model compared to the ResNet-50 model when used separately. Better accuracy and recall are the results of CBAM's attention mechanism, which enhances the model's capacity for retrieving or assessing the predicted visual images in an optimized way.

References

[1] 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).

[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] 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).

[4] 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).

[5] M. S. Basha, K. R. Prasad, S. K. Mouleeswaran, R. C. Poonia, and S. Sebastian, "Multi-disease detection system with X-ray images using deep learning techniques," *Journal of Information and Optimization Sciences*, vol. 45, no. 5, pp. 1379–1388 (2024), doi: 10.47974/JIOS-1710.

[6] M. Sivalakshmi, K. R. Prasad, and C. S. Bindu, "Improved privacy protection technique for enhancing security of real-time video surveillance," *Journal of Information and Optimization Sciences*, vol. 45, no. 5, pp. 1389–1399 (2024), doi: 10.47974/JIOS-1711.

[7] 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).

[8] 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.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90 (2017).

[10] L. Zhang, Y. Zhang, and X. Liu, "A transfer residual neural network based on ResNet-50 for detection of steel surface defects," *Applied Sciences*, vol. 13, no. 9, p. 5260 (2023).

[11] D. Roy, T. Ishizaka, C. Krishna Mohan, and A. Fukuda, "Detection of collision-prone vehicle behavior at intersections using Siamese interaction LSTM," *IEEE Transactions on Intelligent Transportation Systems* (2020), doi: 10.1109/TITS.2020.3031984.

[12] D. Singh and C. Krishna Mohan, "Graph formulation of video activities for abnormal activity recognition," *Pattern Recognition*, vol. 65, pp. 265–273 (2017), doi: 10.1016/j.patcog.2017.01.001.

[13] 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, p. 9516 (2023).

[14] C.-L. Lin and K.-C. Wu, "Development of revised ResNet-50 for diabetic retinopathy detection," *BMC Bioinformatics*, vol. 24, no. 1, p. 157 (2023).

[15] M. Muhathir, M. Sivalakshmi, and S. K. Rajendra Prasad, "Convolutional neural network (CNN) of ResNet-50 with InceptionV3 architecture in classification on X-ray image," Computer Science On-line Conference, Cham: Springer International Publishing (2023).

[16] Z. Duan, X. Zhang, and Q. Li, "An adapted ResNet-50 architecture for predicting flow fields of an underwater vehicle," *IEEE Access* (2024).

[17] 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).

[18] J. Mo and L. Xu, "Weighted cluster-range loss and criticality-enhancement loss for speaker recognition," *Applied Sciences,* vol. 10, no. 24, p. 9004 (2020).

[19] Corel, "Corel 10K Dataset," Corel Corporation (2004). [Online]. Available: http://www.corel.com.

[20] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html. Received June 2016.

Received November, 2024