Advances and Applications in Mathematical Sciences
Volume 20, Issue 11, September 2021, Pages 2613-2626

DEEP CONVOLUTIONAL NEURAL NETWORKS FEATURES FOR IMAGE RETRIEVAL

SURESH KUMAR KANAPARTHI and U. S. N. RAJU

Department of Computer Science
and Engineering National Institute
of Technology Warangal
Warangal, Telangana State, India
E-mail: sureshkonline@gmail.com
usnraju@nitw.ac.in

Abstract

Content-Based Image Retrieval (CBIR) has become one of the trending areas of research in computer vision. In traditional CBIR, the features are considered as hand-crafted features. The state-of-the-art technology for feature extraction is to use deep convolutional neural networks (CNN). In this paper, four deep convolutional neural networks (CNNs): AlexNet, VGG-16, GoogleNet, and ResNet-101 with transfer learning are used to extract and the features from the image. By using these features, d1-distance is used to compare the query images with the images in the image dataset. To evaluate the efficiency of these four models, five standard performance measures are calculated i.e., Average Precision Rate (APR), Average Recall Rate (ARR), *F*-Measure, Average Normalized Modified Retrieval Rank (ANMRR) and Total Minimum Retrieval Epoch (TMRE). Six benchmark image datasets: Corel-1K, Corel-5K, Corel-10K, VisTex, STex, and Color Brodatz are used to corroborate the performance of the four CNN models for CBIR.

I. Introduction

In the present days, an exponential increase in usage of digital cameras and mobile phones makes the size of the image dataset gigantic. Maintaining such kind of large image dataset is an extremely tedious and troublesome job. So, an efficient technique is required to retrieve desired images from such kind of huge image dataset. One of the effective solutions to such retrieval problem is Content-Based Image Retrieval (CBIR). The term "content"

2010 Mathematics Subject Classification: 68T07.
Keywords: CBIR, Deep CNN Features, Transfer Learning.
Received October 13, 2020; Accepted November 8, 2020

signifies that images are retrieved based on some features which can be calculated from the actual content of images. The retrieval process depends on the similarity between the query image and all the images of the image dataset. Feature vector comparison is one of the possible ways to find similarity between the corresponding images. In traditional CBIR, features of an image can be Color features, Local and Global texture features, Point features, and Shape features. As because of technology shift from traditional image processing to 'image processing with deep learning' the feature extraction process has changed drastically since the year 2012, after the proposal of a deep CNN model 'AlexNet' by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton.

1.1. Convolution Neural Networks for Image Retrieval and Classification

AlexNet: AlexNet secured the first position in ILSVRC 2012. It is mainly used for image classification problems. It takes one image from any of 1000 different types of classes and produces an output vector of length 1000. Each element of the output vector is nothing but a probability of an input image belonging to a particular class, i.e. k^{th} value of output vector is the value of expectation that the input image is from k^{th} class. Thus, the summation of all values of output vector results to 1. Image Net data was used to train AlexNet. This dataset contains 15 million annotated images that belong to 22,000 different classes. The total number of parameters and neurons of AlexNet is 60 million and 650,000 respectively. Five to six days were required to train AlexNet on the Image Net dataset using two GTX 580 3GB GPUs. AlexNet contains 8 layers: 5 convolutional layers and 3 fully connected layers [1]. To down sample the feature maps produced by the convolution layers, they are followed by the max-pooling layer. After every convolutional and fully connected layer, the Re Lu layer is applied. Before the first and the second fully connected layer, the dropout layer is used to solve the over fitting problem. Data augmentation [2] methods are also used to resolve the over fitting problem.

VGG: VGG Net achieved an error rate of 7.3% in ILSVRC 2014. VGG has two popular implementations using 16 and 19 layers. When compared to AlexNet, the filter size is reduced here so that the number of parameters can also be reduced. Re Lu layer is used after each convolution layer to make the

system non-linear. It was trained for three weeks on four NVIDIA Titan

Black GPUs. GoogleNet: GoogleNet is more suitable for high-level feature extraction because of its special inception module, which will give proper memory utilization and less computation complexity. GoogleNet has 22 layers and it follows a directed acyclic graph structure. It has an error rate of 6.7%. GoogleNet emphasized that CNNs can be created with layer structures rather than sequentially putting layers. It was trained on high configuration GPUs in one week.

Res Net: Residual Network being a 152-layered network architecture, gives a remarkable performance in classification, localization, and detection problem. Res Net won ILSVRC 2015 with 3.6% error rate [3]. An 8 GPU machine was used for 14 to 21 days to train Res Net. The degradation problem is resolved with the help of a deep residual learning framework [4]. There are different variations of Res Net is available having on the number of layers 18, 34, 50, 101, and 152. Among them, the last three versions of Res Net are much more accurate than the first two. In this work, we have used only the training part and feature extraction part of these CNNs.

Liang Zheng et al. [5] have given the survey on Image retrieval, which mentioned Scale Invariant Feature Transform (SIFT) and convolutional neural network (CNN) as two prime methods for image retrieval problem. One new CBIR model was proposed by Peizhong et al. [6] which combines both high-level and low-level features of an input image. To extract low-level features dot-diffused block truncation coding (DDBTC) is used. CNN is frequently used for extracting high-level features. Among various CNNs (AlexNet, GoogleNet, Res Net, VGG), Famao Ye et al. [7] proposed a new CBIR method that used CNN feature with weighted distance. The method is divided into two phases. In the 1st phase, known as the offline phase, feature extraction, and class leveling of an input image is done based on Fine-tuned CNN. In the 2nd phase, known as the online phase, the weightage factor of each class is calculated based on the probability of a query image belonging to that particular class. Distance between the retrieved images and the query image is determined based on the calculated weighted factor. Zar Nawab Khan Swati et al. [8] proposed a new CBIR method based on Deep CNN. VGG19 architecture is used as Deep CNN. In this paper closed-form, metric learning is used as a similarity measure between dataset images and the

query image. A novel CBIR framework is proposed by Adnan Qayyum et al. [9] which is divided into two stages. 1st stage is the classification stage, where the training of a Deep CNN is performed on a medical image dataset based on supervised learning. In the 2nd stage, the last three fully connected layers of trained Deep CNN is used to extract feature from the input image. As a single CNN is used by most of the CNN based CBIR methods for feature extraction, intermediate layers of CNN are not used properly to determined local patterns of an input image. To resolve this problem, a novel CBIR framework is proposed by Ahmad Alzubi et al. [10] where two CNN are used parallelly to extract features. In this proposed method Deep CNN is trained on a large image dataset followed by fined tuning. In addition to that one bilinear root pooling is used at low dimensional pooling layer for efficient dimension reduction of a feature vector.

1.2. Similarity Measures and Query Matching

Different types of distance measures can be used to calculate the similarity between the query image and other images in the dataset: Manhattan, Euclidian, d1 distance, Canberra, chi-square, and histogram intersection [11]. The process of calculating all these distances are given in Table 1. The result of applying the six distances on the two feature vectors, as shown in Figure 1.

Table 1. Different distance measures.

Manhattan Distance	$d(db_i, q) = \sum_{s=1}^{len} \left\| F_{db_1}(s) - F_q(s) \right\|$
Euclidian Distance	$d(db_i, q) = \left(\sum_{s=1}^{len} (F_{db_1}(s) - F_q(s))^2 \right)^{\frac{1}{2}}$
d1 Distance	$d(db_i, q) = \sum_{s=1}^{len} \left\| \frac{F_{db_1}(s) - F_q(s)}{1 + F_{db_i}(s) + F_q(s)} \right\|$
Canberra Distance	$d(db_i, q) = \sum_{s=1}^{len} \left\| \frac{F_{db_1}(s) - F_q(s)}{F_{db_i}(s) + F_q(s)} \right\|$

Chi-Square Distance	$d(db_i, q) = \frac{1}{2}\left(\sum_{s=1}^{len} \frac{(F_{db_1}(s) - F_q(s))^2}{1 + F_{db_i}(s) + F_q(s)}\right)$
Histogram Matching	$d_{HI}(db_i, q) = \sum_{s=1}^{len} \min\ (F_{db_1}(s), F_q(s))$

Feature Vector-1	2	6	7	5	4
Feature Vector-2	0	1	7	2	5

(a)

S. No.	Name of the Distance	Result
1	Manhattan Distance	11
2	Euclidian Distance	6.24
3	d1 Distance	1.77
4	Canberra Distance	2.25
5	Chi-Square Distance	3.49
6	Histogram Matching	14

(b)

Figure 1 (a). Example of two feature vectors. (b) Results of different distances for the feature vectors shown in (a).

II. Methodology

In this process, first, training a CNN model with transfer learning (using 70% image from the respective datasets) is done. Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. Then all the images of that dataset are given as input to the trained CNN to extract the feature vector. The length of the feateur vectors of the CNNs considerd are shown in Table 2. Different CNNs have a different number of layers for obtaining the feature vector. Then these feature vectors are used for image retrieval. Here the five performance measures: APR, ARR, FMeasure, ANMRR, and TMRE are calculated based

on the feature vectors extracted by the respective CNN. In this study, four different CNNs: AlexNet, VGG-16, GoogleNet, and ResNet-101 are used. The entire process is shown in Figure 2.

Table 2. Feature vector length of different CNNS.

CNN Model	Feature Vector Length
AlexNet	4096
VGG-16	4096
GoogleNet	1024
ResNet-101	2048

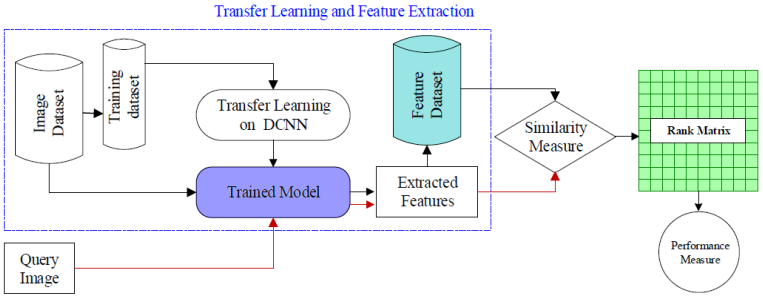

Figure 2. General CBIR Framework by using Deep CNN Features.

III. Dataset and Experimental Results

For performance evaluation of the different CNN features methods, six benchmark color datasets are used. Precision, Recall, ANMRR, TMRE, and F-measure are used for performance comparison. Each of the query images gives a feature vector by following the steps of the proposed method during evaluation. As per the distance measures, which are given in Table 1, the comparison between the query image feature vector and dataset images' feature vector is carried out. A rank matrix is obtained from the distances calculated as shown in Figure 3, of size $N \times N$, where the total number of images in the dataset is denoted by N. k^{th} similar image concerning i^{th} query image is referred by cell Rank (k, i) in the rank matrix.

3.1. Performance Measures for CBIR

Average Precision Rate (APR) and Average Recall Rate (ARR): Precision is defined as a ratio between the number of total relevant images retrieved and the number of total images retrieved for a given query. Recall is defined as the ratio between the number of total relevant images retrieved and the number of total images having the same class as a query image. Average precision for different step sizes $m_1 m_2, \ldots, m_k$ is known as APR. Similarly, average recall for different step sizes is known as ARR.

F-Measure: It is represented by a single value to reflect the relationship between precision and recall. It is obtained by assigning equal weight to both precision and recall in the harmonic mean calculation as given in equation (1).

$$F - \text{Measure} \quad (n) = \frac{(2 \times APR \times APR)}{(APR + ARR)}. \tag{1}$$

Average Normalized Modified Retrieval Rank (ANMRR): It is used to measure the retrieval accuracy. To calculate ANMRR for each image we consider only those images whose rank is less than 2× (number of images in the class). If an image's rank is less than 2×(number of images in the class) then score of that image is rank of the image, else it is a predefined fixed number. Now the average score is calculated and then normalized score.

Total Minimum Retrieval Epoch (TMRE): It is used to measure the minimum number of images to be traversed to retrieve all the relevant images. These five performance measures are obtained on the six image datasets. The details are explained individually of each image dataset.

		Query Iamges Considered from the Image Dataset									
		Class-1				…	…	…	Class-10		
		Image1	**Image2**	…	**Image100**	…	…	…	…	**Image999**	**Image1000**
Image Numbers in the Dataset	1	**Rank 1**									
	2	Rank 102	**Rank 1**								
	3	Rank 5		**Rank 1**							
	1000										**Rank 1**

Figure 3. Rank Matrix Representation for 1000 images dataset.

Dataset-1 (Corel-1K): This dataset [12], consists of a total of 1000 images

with 10 categories where each category consists of 100 images. The different categories are Africans, Beaches. Food. The size of each image in this image dataset is 384×256 or 256×384. Three images from each group, a total of 30 images of this dataset are shown in Figure 4. The five performance measure values for this image dataset are shown in Table 3.

Figure 4. Corel-1K Samples (three images per category).

Table 3. Performance measures for COREL-1K.

	Corel 1K				
	APR	ARR	*F*-Measure	ANMRR	TMRE
AlexNet	78.68	44.19	36.420	0.4600	9.84
VGG-16	85.06	55.51	44.970	0.3500	9.81
GoogleNet	85.42	41.79	35.550	0.4800	9.80
ResNet-101	80.05	37.95	32.450	0.5200	9.81

Dataset-2(Corel-5K): Corel-5K image dataset [13] consists of 50 categories, where each category is of 100 images, which total 5000 images. Figure 5 shows a total of 50 images, one image from each of 50 categories. As in the Corel-1K image dataset, here too all the five performance measures are evaluated and shown in Table 4 for the four CNNs respectively.

Figure 5. Corel-5K Samples (one image per category).

Table 4. Performance measures for CORE-5K.

	Corel 5K				
	APR	ARR	*F*-Measure	ANMRR	TMRE
AlexNet	44.50	22.76	18.61	0.71	49.26
VGG-16	73.72	39.84	33.57	0.52	48.98
GoogleNet	63.50	25.31	22.47	0.68	49.04
ResNet-101	64.48	25.89	22.91	0.68	49.04

Dataset-3 (Corel-10K): The third natural image dataset considered is the Corel-10K image dataset [13]. This dataset consists of 100 categories with 100 images in each category results in a total of 10000 images. This dataset was containing a total of 53 different sized images: 128×192, 192×128, etc. In Figure 6, a total of 100 images are given, one from each of 100 categories of this image dataset. Table 5 show all the five performance measure results of the Corel-10K dataset by using all the four CNNs models.

Figure 6. Corel-10K Samples.

Table 5. Performance measures for COREL-10K.

	Corel 10K				
	APR	ARR	*F*-Measure	ANMRR	TMRE
AlexNet	30.95	14.36	11.51	0.81	98.43
VGG-16	64.24	33.20	28.04	0.60	97.83
GoogleNet	62.05	23.62	21.40	0.71	98.10
ResNet-101	64.49	25.96	23.10	0.68	97.69

Dataset-4 (Vis Tex): This image dataset is the first texture image dataset considered Vis Tex texture dataset [14] consists of a total of 484 images. Out of these 484 texture images, 40 are considered for experimentation. The actual image dimension is 512×512. Each image of these 40 is made into 16 nonoverlapping sub-images where each sub-image is of dimension 128×128, which results in a total of 640 texture image datasets. From these 640, images 1, 17, 33, 49 …625 which are the 1st sub-image of each of 40 actual texture images, are shown in Figure 7. All the performances are obtained on these 640 texture image dataset. Table 6 show the five performance measure values.

Figure 7. Forty Vis Tex texture images considered.

Table 6. Performance measures for Vis Tex.

			Vis Tex		
	APR	ARR	*F*-Measure	ANMRR	TMRE
AlexNet	83.28	57.36	49.06	0.34	26.58
VGG-16	97.58	75.82	62.91	0.17	20.18
GoogleNet	94.38	64.07	56.11	0.28	24.80
ResNet-101	95.59	66.02	57.68	0.26	25.15

Dataset-5 (STex): The other color texture dataset considered is the Salzburg Texture Image Dataset (STex) [15] which contains a total of 476 texture images. All these 476 are considered to test the performance of different methods. Here also, each texture image is made into 16 non-overlapping sub-images which results in a total of 7616, where each sub-image is of dimension 128×128. Figure 8 shows forty of these 7616, where each of these forty is considered from 40 different actual texture images from 476 texture images. All the five performance measures are given in Table 7.

Figure 8. Forty of the STex texture images from 7616 images.

Table 7. Performance measures for STEX.

	Stex				
	APR	ARR	*F*-Measure	ANMRR	TMRE
AlexNet	79.77	47.15	42.91	0.46	309.76
VGG-16	91.92	60.84	53.74	0.32	285.89
GoogleNet	88.79	53.77	48.88	0.38	306.49
ResNet-101	92.80	60.66	53.88	0.31	253.09

Dataset-6 (Color Brodatz): This is the last image dataset considered, color Brodatz texture image dataset [16]. We made each of the images into 25 non-overlapping sub-images, which results in a total of 2800 images. The first sub-image from each of these 112 are shown in Figure 9 and the results are shown in Table 8.

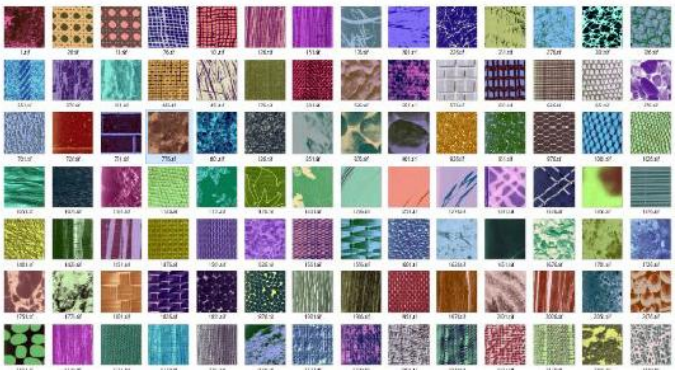

Figure 9. 112 Textures each from Color Brodatz texture images from 2800 textures images.

Table 8. Performance measures for color BRODATZ.

	Color Brodatz				
	APR	ARR	*F*-Measure	ANMRR	TMRE
AlexNet	92.76	66.52	55.38	0.27	76.14
VGG-16	97.29	76.64	61.28	0.18	63.56
GoogleNet	87.16	54.78	47.24	0.38	72.18
ResNet-101	88.11	54.85	47.56	0.38	74.91

IV. Conclusions

The features extracted by the four deep CNN models: AlexNet, VGG-16, GoogleNet and ResNet-101 are used for CBIR on a total of six image datasets, three of which are natural image dataset: Corel-1K, Corel-5K, and Corel-10K and three are color texture image datasets: VisTex, STex, and Color Brodatz. Five performance measures are applied: APR, ARR, *F*-Measure, TMRE, and ANMRR are obtained. Results show that the all the four CNN models are performing good in one or other datasets when compared with the traditional features. Future Extensions: In Future, the four CNN models used will be fine-tuned to get more accuracy for the given image dataset.

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, In Advances in neural information processing systems (2012), 1097-1105.

[2] D. Ciregan, U. Meier and J. Schmidhuber, Multi-column deep neural networks for image classification, In 2012 IEEE conference on computer vision and pattern recognition, IEEE June (2012) 3642-3649.

[3] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, In Proceedings of the IEEE conference on computer vision and pattern recognition (2016), 770-778.

[4] K. He and J. Sun, Convolutional neural networks at constrained time cost, In Proceedings of the IEEE conference on computer vision and pattern recognition (2015),

5353-5360.

[5] L. Zheng, Y. Yang, Q. S. I. F. T. Tian and S. M. CNN, A Decade Survey of Instance Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 40(5) (2018), 1224-1244.

[6] P. Liu, J. M. Guo, C. Y. Wu and D. Cai, Fusion of deep learning and compressed domain features for content-based image retrieval, IEEE Transactions on Image Processing 26(12) (2017), 5706-5717.

[7] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo and W. Min, Remote sensing image retrieval using convolutional neural network features and weighted distance, IEEE geoscience and remote sensing letters 15(10) (2018), 1535-1539.

[8] Z. N. K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed and J. Lu, Content-based brain tumor retrieval for MR images using transfer learning, IEEE Access 7 (2019), 17809-17822.

[9] A. Qayyum, S. M. Anwar, M. Awais and M. Majid, Medical image retrieval using deep convolutional neural network, Neurocomputing 266 (2017), 8-20.

[10] A. Alzu'bi, A. Amira and N. Ramzan, Content-based image retrieval with compact deep convolutional features, Neurocomputing 249 (2017), 95-105.

[11] M. Verma, B. Raman and S. Murala, Local extrema co-occurrence pattern for color and texture image retrieval, Neuro computing 165 (2015), 255-269.

[12] James Z. Wang, Modeling objects, Concepts, Aesthetics and Emotionsin Big Visual Data, http://wang.ist.psu.edu/docs/home.shtml (Accessed 15 October 2020).

[13] Guang-Hai Liu et al., Corel-10k dataset, http://www.ci.gxnu.edu.cn/cbir/Dataset.aspx (15 October 2020).

[14] Alex (Sandy) Pentland and Ted Adelson, VisTex Dataset, http://vismod.media.mit.edu/pub/VisTex/, (15 October 2020).

[15] Roland Kwitt, Salzburg Texture Image Dataset, http://www.wavelab.at/sources/STex/, (15 October 2020).

[16] Dong-Chen, Abdelmounaime Safia, Multiband Texture (MBT) dataset, https://multibandtexture. recherche.usherbrooke.ca/index.html (15 October 2020).