Content based Image Retrieval using Deep Learning Technique with Distance Measures

[1]M. Aswini Varalakshmi, [2] Geetanjali Nayak
[1] *PG Scholar, Dept of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering, Kommadi, Vishakhapatnam*
[2] *Assistant Professor, Dept of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering, Kommadi, Vishakhapatnam*

Abstract **- The combination of convolution neural networks (CNN) and deep learning generated a stunning result in a variety of image processing applications. For separating comparable images, CNN-based techniques to isolate image features from the last layer and the use of a single CNN structure could be used. The Content-Based Image Retrieval system is used to learn highlight extraction and efficient similarity examination (CBIR). Highlight extraction, like similarity tests, plays an important role in CBIR. The research is carried out using datasets such as the UC Merced Land Use Dataset. Using a pre-trained model that has been adjusted for the retrieval task and has been trained on a large number of photographs. For the retrieval process, pre-trained CNN models are used to generate image highlight descriptors. By using move learning and retrieval of highlight vectors using various similarity measures, this technique manages component extraction from the two completely connected layers present in the VGG-16 network. The proposed engineering has a fantastic presentation in terms of extracting features as well as learning features without any prior knowledge of the images. With vgg19 , we're going to be able to extend our work with CNN. Investigation of configuration and bunching for signs of improved execution. The outcomes were measured and execution correlation was completed using various execution metrics. In both completely connected layers, cosine similarity and Euclidean distance work better.**

Index Terms **- CBIR, CNN, Bag of visual words., vgg16, vgg19.**

INTRODUCTION

Finding the right set of images to go with an info image is a major concern for CBIR systems. CBIR is the next step in the evolution of keyword-based systems, in which images are recovered using data from their contents. The retrieval execution of a CBIR system is largely dependent on two variables: 1) feature representation and 2) feature representation. 2) estimate of similarity CNNs are a type of learnable model that can be used in applications such as image retrieval, classification, and recognition. Annotating images, recognising images, and so on. They used for image retrieval with the inspiration of the exceptional performance of deep learning algorithms to the innovation in this paper.

The aim of this paper is to develop a reliable image retrieval system for sifting through large amounts of data images retrieved from a database The proposed strategy starts with the use of a pre-trained model to obtain unaided learning from tweaked data in order to learn parameters for CBIR tasks. The feature extraction from two different completely linked layers present in the pre-trained CNN [1] is the main focus. CBIR is mostly used for browsing based on content rather than image annotations. It includes a method for representing and sorting images based on an information query image [1].

The method of feature extraction is at the heart of CBIR. Certain image features that can be resolved from the images in CBIR include shading, surface, and form. CBIR has two distinct modes of action. For instance, the first technique is ordering, and the second is looking. The extraction of features from an image using strategy ordering can be used to store these features in a feature database as feature vectors. In the second technique, for example, in looking, feature vectors are extracted from information images and these isolated features are compared to feature vectors stored in the database. This result is then used to find the most similar images in the database to the query image. In terms of image retrieval, there are two types: (1) exact image retrieval and (2) related image retrieval. For exact image retrieval, 100 percent

coordination with the query is required, whereas for relevant image retrieval, the retrieval is based on the contents or features of the image.

Deep Learning is a set of techniques in which artificial intelligence (AI) algorithms or methods are used to display substantial level impressions of data using deeper models. There are several ongoing studies in the field of learning information and complex capabilities without the use of human-crafted knowledge. In such instances, Deep Learning has proven to be a huge success in learning from raw data using a variety of procedures that can be applied to Speech Recognition and Natural Language Processing [2].

LITERATURE SURVEY

Image retrieval techniques are extremely useful in content management systems, according to Gopal and Bhooshan (2015). For retrieving identical images from databases, CBIR strategies integrate explicit image features such as hues, surfaces, key points, and so on. Grayscale information is used in a large number of key point detectors and key point descriptors. By adding additional shading information to the key point descriptors, the retrieval accuracy of these strategies can be enhanced.

An enhanced SURF descriptor for CBIR applications is proposed in this paper, which extracts image features by processing Hu minutes along with eigen values in the immediate vicinity of the specified key points. The use of an improved SURF descriptor improves image retrieval performance in the lab. Furthermore, the improved SURF descriptor can distinguish between images of the same object with identical grayscale properties but different hues.

J. Wan, D. Wang, and S.C.H. Hoi, 2014. Learning good feature representations and similarity measures is critical to a CBIR framework's retrieval efficiency. Despite decades of study, it remains one of the most difficult open issues that significantly impedes the progress of real-world CBIR frameworks. The main problem has been due to a significant semantic difference between low-level image pixels captured by machines and high-level semantic ideas seen by humans. Machine learning has been actively explored as a possible path to bridge the semantic gap in the long run, among other approaches. We investigate the state-of-the-art deep learning procedures for learning feature representations and similarity measures, sparked by recent successes of deep learning strategies for PC vision and other applications, to answer an open issue: whether deep learning is an expectation for connecting the semantic gap in CBIR and how much upgrades in CBIR tasks can be accomplished by investigating the state-of-the-art deep learning procedures for learning feature representations and similarity measures. Especially, By analyzing a state-of-the-art deep learning technique CNN for CBIR tasks under various set-chimes, we examine a theory of deep learning with application to CBIR tasks with a wide array of empirical examinations.

P. Nalini and B.L. Malleswari, 2016 CBIR is a technique that recovers similar images based on image content similarity for a given query image. The visual features of an image, which are mathematical representations of a digital image, are referred to as image material. The image retrieval task is primarily based on image feature extraction and feature vector similarity calculation. The output of the CBIR process is determined not only by the ideal features extracted from the picture, but also by the best possible decision of the CBIR process. Tests of similarity and dissimilarity (distance metrics). Since the image features vary so greatly in terms of colouring, surface, and form, using the same distance metric for all of them does not work well. We first provided an overview of the mathematical and statistical distance metrics used in CBIR, as well as a comparison of these measures on shading and surface features in this paper. Surface features are extracted by wavelet deteriorations and shading features are extracted by figuring shading histograms in HSV room. For feature similarity, geometrical distances such as Manhattan, Chebyshev, and Euclidean were analysed, as well as statistical distance metrics such as Cosine Similarity, Chi-square, KullbackLeibler, Jeffrey, and cumulative statistical distance metrics such as Kolmogorov-Smirnov, Cramer von Mises, and Earthmovers distances. With shading and surface features separately, we set specific goals for the performance of all of these distance metrics in terms of Mean Average Precision (MAP) and Recall rates.

PROBLEM DEFINITION

The image retrieval method looks for a tag that matches the graphic keyword or metadata associated with the image. The content-based image matching

technique is the name for this process. Content Dependent Image Retrieval is the process of retrieving images based on their content. The CBIR strategy produces much more reliable results than the picture ordering and bunching techniques. The aim of content-based image retrieval is to find more valid images from a large number of datasets that fit the query image or provided image. The images' attributes, as well as their values and lists, are extracted using function vector and stored in the database. Then, by using the list structure and comparing attributes to the query image, all irrelevant items should be filtered out. The related image feature is compared to the query image feature using a similarity test, and the recovered items are sorted by similarity The problem is that it uses the bunching and ordering approach to compare the structural similarity of the images. In any case, as the number of images increases, it becomes impossible to find a more valid image that is identical to the query image.

IMPLEMENTATION STUDY

The proposed strategy includes extracting image feature vectors from the pre-trained CNN model's two completely linked layers. Since the datasets contain fewer files, the pre-trained CNN loads of the ImageNet model can be used for retrieval. It is possible to use the loads legitimately and architecture learning and apply the learning to the CBIR tasks by using pre-trained models that have been trained on a large number of image datasets. This is transfer learning, which involves passing knowledge to a pre-trained model based on the problem statement that has been assigned.

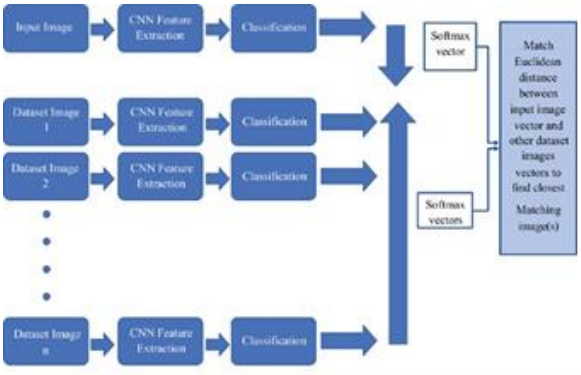


Fig 1: System Implementation architecture

VGG16 And VGG19 is the CNN model that has been pre-trained in this case. The data was arranged in distant detecting image datasets for the study. The testing protocol was given the entire dataset. The features are derived from a pre-trained model that includes convolution, max-pooling, and completely connected layers. The model was calibrated using both datasets, which was the main technique used. It is necessary to adjust the model with the two datasets while using transfer learning. To begin with, the loads of the pre-trained CNN are initialized randomly and then training of dataset was done accordingly. Then the loads obtained were legitimately used for the training of two datasets.

The information image provided as a query can come from any source, and it is not required to come from the datasets. The query image is taken from the datasets in this case. The features are derived from the AlexNet layers, and the yield classification layer is the final layer. Since transfer learning allows the pre-trained model to be adjusted. It is the process of removing the final yield classification layer from the architecture and using the remaining architecture as a fixed feature extractor. These extracted features were used to compare the query input image to image database features using a variety of similarity metrics. A limit value is set for sifting through the images that are identical to the information image and those that are not similar based on the value of similarity measures taken. As a result, the value of similarity above the edge will be sifted through, and the proposed approach will deal with image similarity assessment using different distance metrics.

IMPLEMENTATION MODEL ARCHITECTURE

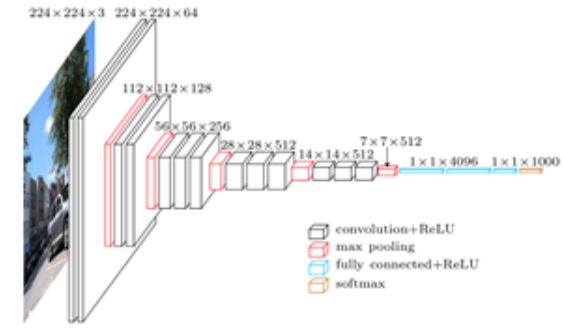


Fig 2: Image classification-based framework using DNN framework

Dataset

Land Use Dataset from UC Merced: This 21-class dataset [23] of land images is used for analysis. Each class has 100 256x256 pixel images. This dataset contains 500 images in total. The images were created

using vast images collected by the USGS National Map Urban Area Imagery amassment from districts around the country.

Fig 3: Sample Images in UC Merced Land Use Dataset

ALGORITHMS USED

Convolutional Neural Network(vgg16)

The network's input is a two-dimensional image (224, 224, 3). The first two layers have the same padding and 64 channels of 3*3 filter size. Then, after a stride (2, 2) max pool sheet, two layers of convolution layers of 256 filter size and filter size (3, 3). This is accompanied by a stride (2, 2) max pooling layer, which is the same as the previous layer. Following that, there are two convolution layers with filter sizes of 3 and 3 and a 256 filter. Following that, there are two sets of three convolution layers, as well as a max pool layer. Each has 512 filters of the same size (3, 3) and padding. This image is then fed into a two-layer convolution stack. The filters we use in these convolution and max pooling layers are 3*3 instead of 11*11 in AlexNet and 7*7 in ZF-Net. It also uses 1*1 pixels in some of the layers to manipulate the number of input channels. After each convolution layer, a 1-pixel padding (same padding) is applied to prevent the image's spatial attribute from being lost.

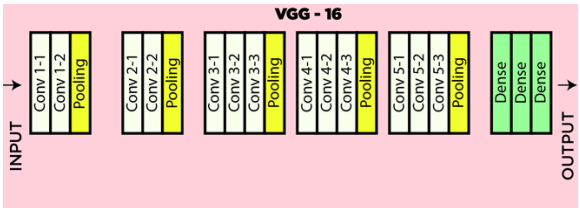


Fig 4: Convolutional Neural Network.(vgg16)

We got a (7, 7, 512) function map after stacking the convolution and max-pooling layers. This performance is flattened to create a (1, 25088) function vector. Following that, there are three completely connected layers: the first takes input from the last feature vector and outputs a (1, 4096) vector, the second layer also outputs a (1, 4096) vector, but the third layer outputs 1000 channels for 1000 ILSVRC challenge classes, and the output of the third fully connected layer is then transferred to the softmax layer to normalise the classification vector. Top-5 categories for evaluation after the classification vector output. The activation mechanism for all secret layers is ReLU. ReLU is more computationally efficient because it speeds up learning and reduces the chance of a vanishing gradient problem.

Vgg19:-

VGG-19 is a qualified Convolutional Neural Network from the Visual Geometry Group at the University of Oxford's Department of Engineering Science. The number 19 refers to the number of layers with weights that can be trained. There are 16 convolutional layers and 3 fully connected layers in this picture.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Fig 5: vgg19 Implementation Architecture

The input to the network is a (224, 224, 3) RBG image. Take note of how the layers are labelled. commutative As a result, conv3-64 denotes a total of 64 (3, 3) square filters. It's worth noting that in VGG-19, all of the conv layers use (3, 3) filters, with the number of filters increasing in powers of two (64, 128, 256, 512). Stride length is 1 (pixel) in all Conv layers, with a 1 (pixel) padding on each side. There are 5 sets of conv layers, two of which have 64 filters, two of which have 128 filters, four of which have 256 filters, and the other Between each set of conv layers are max pooling

layers. 2x2 filters with a stride of 2 are used in the maximum pooling layers (pixels) The output of the last pooling layer is flattened and fed to a 4096-neuron completely connected layer. The output is fed into a 4096-neuron fully connected layer, which in turn feeds into a 1000-neuron fully connected layer. Both of these layers have ReLU turned on. Finally, a softmax layer that employs cross entropy loss is employed. The convolution layers and the Fully Connected layers are the only layers with trainable weights. The input image is reduced in size using the maxpool layer, and the final decision is made using softmax. two sets each have four conv layers with 512 filters.

FEATURE EXTRACTION

In any item recognition algorithm, feature extraction is a critical step. It alludes to a method for extracting useful information from an information picture, which is referred to as features. The extracted features must be representative of the image, carrying essential and unique characteristics.

Step1: The key section point where feature extraction takes place. As input, this function accepts an 8-piece RGB or an 8-piece grayscale image. An array of extracted intrigue points is returned as a result. This function is made up of function calls that include computations that can be parallelized on the GPU.

Step2: The function converts an eight-piece RGB image into an eight-piece grayscale image. If the information provided is already in the 8-piece grayscale format, this step will be skipped. The 8-piece grayscale image is then converted to a 32-piece floating-point representation, allowing for faster computations on the GPU.

Step3: The integral image of the 32-piece floating-point grayscale image obtained in the previous step is calculated with this function. The integral image can be used to deconstruct the operation of calculating the total number of pixels contained within any rectangular region of the image. These aids in increasing the pace at which convolutions are performed in the next step.

Step4: The picture is convolutioned with box filters of different sizes, and the figured reactions are saved.

EXPERIMENTAL RESULTS

Fig 6 :-Vgg 16 training and validation loss graph

Training accuracy	Validation accuracy	Training loss	Validation loss
0.946	0.957	0.43	0.31

Table 1: training and validation accuracy for vgg16

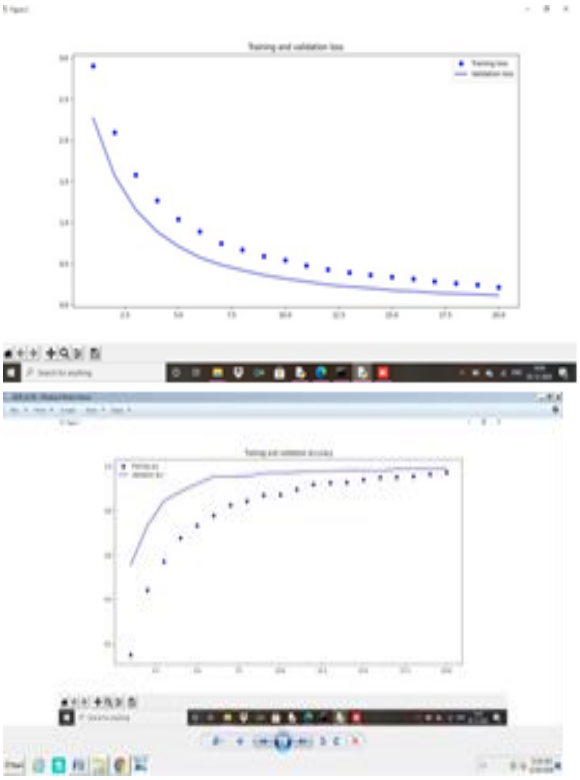
The total number of questions is 19, and there are 950 test images in the UC Merced Land Use Dataset. Images were retrieved by comparing to each query. The applied distance metrics measure the distance or similarity between the images when recovering related

Fig 7:- training and validation graph for vgg19

Training accuracy	Validation accuracy	Training loss	Validation loss
0.99	0.96	0.45	0.33

Table 2: training and validation accuracy for vgg19

images from a database of images. It should have a high precision and low recall for a better retrieval system. The UC Merced Land Use dataset provides 0.84 precision for Euclidean distance and 0.9 for cosine similarity in the Fc1 layer. As compared to other distance metrics, they perform better in the Fc1 layer. As compared to the Fc2 layer, the quantity of images returned for these two distance metrics is higher, and the output of the other distance metrics is better than the Fc1 layer. The SceneSat dataset yielded 0.92 precision for Euclidean distance and 0.95



precision for cosine similarity for extended implementation

	Precision (Existed)	Precision (Extended)
Euclidean Distance	0.84	0.92
Cosine Similarity	0.90	0.95

Table 3: precession for vgg16 and vgg19

CONCLUSION

The efficiency and accuracy of the image retrieval system will be improved by initializing the pre-trained model trained on ImageNet for new images. It is possible to recover new images with superior output using these pre-trained models as well as transfer learning. Similar images are retrieved by registering the similarity of features from both fully connected layers. The retrieval from the Fc2 layer is more efficient than the Fc1 layer retrieval. For all distance steps, the retrieval performed from the Fc2 layer has a higher precision rate. The Euclidean distance and Cosine similarity are used in both datasets to find the most related images. When recovering images from the Fc2 layer in the UC Merced Land Use dataset, the rate of precision for Euclidean Distance and Cosine Similarity, respectively, is 0.92 and 0.96. In the SceneSat dataset from the Fc2 sheet, both Euclidean Distance and Cosine Similarity are 0.96.

REFERENCES

[1] Singh, A.V. Content-based image retrieval using deep learning. Rochester Institute of Technology, 2015.

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

[3] Yasmin, M., Mohsin, S. and Sharif, M. Intelligent image retrieval techniques: a survey. Journal of applied research and technology 12 (1) (2014) 87-103.

[4] Bagyammal, T., and Parameswaran, L. Context Based Image Retrieval using Image Features. International Journal of Advanced Information Science and Technology 29 (2014).

[5] Bakar, S.A., Hitam, M.S. and Yussof, W.N.J.H.W. Content-Based Image Retrieval using SIFT for binary and greyscale images. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2013, 83-88.

[6] Gopal, N. and Bhooshan, R.S. Content Based Image Retrieval using enhanced SURF. Fifth National Conference on, Computer Vision, Pattern Recognition, Image Processing and Graphics, 2015, 1-4.

[7] Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y. and Li, J. Deep learning for content-based image retrieval: A comprehensive study. Proceedings of the 22nd ACM international conference on Multimedia, 2014, 157-166.

[8] Babenko, A., Slesarev, A., Chigorin, A. and Lempitsky, V. Neural codes for image retrieval. European conference on computer vision, 2014, 584-599.

[9] Nalini, P. and Malleswari, B.L. An Empirical Study and Comparative Analysis of Content Based Image Retrieval (CBIR) Techniques with Various Similarity Measures. 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences, 2016, 373-379.

[10] Liu, H., Li, B., Lv, X. and Huang, Y. Image Retrieval Using Fused Deep Convolutional

Features. Procedia Computer Science 107 (2017) 749-754.

[11] Chen, J., Wang, Y., Luo, L., Yu, J.G. and Ma, J. Image retrieval based on image-to-class similarity. Pattern Recognition Letters 83 (2016) 379-387.

[12] Wu, S., Oerlemans, A., Bakker, E.M. and Lew, M.S. Deep binary codes for large scale image retrieval. Neurocomputing 257 (2017) 5-15.

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

[14] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[15] Yang, H.F., Lin, K. and Chen, C.S. Cross-batch reference learning for deep classification and retrieval. Proceedings of the ACM on Multimedia Conference, 2016, 1237-1246.

[16] Zhu, H., Long, M., Wang, J. and Cao, Y. Deep Hashing Network for Efficient Similarity Retrieval. AAAI, 2016, 2415-2421.

[17] Cao, Y., Long, M., Wang, J., Zhu, H. and Wen, Q. Deep Quantization Network for Efficient Image Retrieval. AAAI, 2016, 3457-3463.

[18] Cao, Y., Long, M., Wang, J., Yang, Q. and Yu, P.S. Deep visual-semantic hashing for cross-modal retrieval. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 1445-1454.