Deep Learning Using Tiled Convolution Feature Extraction Technique For Content-Based Image Retrieval

Ganta. Vijaya Lakshmi[1], SK. Naseema[2]
[1]Dept of ECE
[2]Assistant Professor, Dept of ECE
[1, 2] Indira Institute of Science and Technology, Markapur, A.P, INDIA

Abstract- *In the last few years, deep learning has led to very good performance on a variety of problems, such as visual recognition, speech recognition and natural language processing. Deep learning using tiled convolution feature extraction technique for content-based image retrieval has been successfully implemented in this paper. Among different types of deep neural networks, convolutional neural networks have been most extensively studied. Leveraging on the rapid growth in the amount of the annotated data and the great improvements in the strengths of graphics processor units, the research on convolutional neural networks has been emerged swiftly and achieved state- of-the-art results on various tasks. In this paper, we provide a tiled convolution neural network, mixed pooling layer, leaky ReLu activation function layer, hinge loss function, dropout regularization layers have been used to extract the high content features using AlexNet database. Besides, we also introduce various applications of convolutional neural networks in computer vision, speech and natural language processing. As compared to existing techniques this method provides better results in terms of precision, recall and F-score.*

Keywords- Tiled convolution nerural network, prescion, recall, F-score.

I. INTRODUCTION

In the recent past the advancement in computer and multimedia technologies has led to the production of digital images and cheap large image repositories. The size of image collections has increased rapidly due to this, including digital libraries, medical images etc. To tackle this rapid growth, it is required to develop image retrieval systems which operates on a large scale. The primary aim is to build a robust system that creates, manages and query image databases in an accurate manner. CBIR is the procedure of automatically indexing images by the extraction of their low-level visual features, like shape, color, and texture, and these indexed features are solely responsible for the retrieval of images [8]. Thus, it can be said that through navigation, browsing, query-by-example etc. we can calculate the similarity between the low-level image contents which can be used for the retrieval of relevant images. Images are a representation of points in a high dimensional feature space and a metric is used to measure the similarity or dissimilarity between images on this space. Therefore, those images which are closer to the query image are similar to it and are retrieved. Feature representation and similarity measurement are very crucial for the retrieval performance of a CBIR system and for decades researchers have studied them extensively. A variety of techniques have been proposed but even then, it remains as one of the most challenging problems in the ongoing CBIR research, and the main reason for it is the semantic gap issue that exists between the low-level image pixels captured by machines and high-level semantic concepts perceived by humans. Such a problem poses fundamental challenge of Artificial Intelligence from a high-level perspective that is how to build and train intelligent machines like human to tackle real-world tasks. One promising technique is Ma- chine Learning that attempts to address this challenge in the long-term. In the recent years there have been important advancements in machine learning techniques. Deep Learning is an important breakthrough technique, which includes a family of machine learning algorithms that attempt to model high-level abstractions in data by employing deep architectures composed of multiple non-linear transformations. Deep learning impersonates the human brain that is organized in a deep architecture and processes information through multiple stages of transformation and representation, unlike conventional machine learning methods that are often using shallow architectures. By exploring deep architectures to learn features at multiple level of abstracts from data automatically, deep learning methods allow a system to learn complex functions that directly map raw sensory input data to the output, without relying on human-crafted features using domain knowledge. In the recent studies like Hinton et al [5] and Wan et al [10] encouraging results have been reported for applying deep learning techniques in applications like image retrieval, natural language processing, object recognition among others. The success of deep learning inspired me to explore deep

learning techniques with application to CBIR task for annotated images. There is limited amount of attention focusing on CBIR applications even though there has been much research attention of applying deep learning for image classification and recognition in computer vision.

Content based image retrieval is described in the section II, Proposed method is explained in the section III, experimental results described based different databases in section IV, and conclusion has been discussed in section V

II. CONTENT BASED IMAGE RETRIEVAL AND CNN

Content based image retrieval is composed the feature extraction on color, shape and texture. a combination of dominant color, average color, dominant channel and fuzzy color histogram to describe the color feature. The dominant color uses representative colors to characterize the color information in the required region of an image thus making it a compact and efficient descriptor. Local features of an image can be well represented by a dominant color descriptor which helps in fast and efficient retrieval of images from large datasets. The average color descriptor returns the average of all colors present in the image and compares to it. The dominant channel descriptor takes into consideration the dominant tone per channel and returns the percentage of the dominant channels. Fuzzy 3D color histograms are required to compute dominant color. Fuzzy version is more balanced for colors that fall between color bins. We have used only 8 color bins in this project. Gabor Filter as a texture feature descriptor. Gabor Filter is a linear filter used for edge detection. It is an image filter that can be used to describe texture of the image. The Gabor Filters are of any arbitrary size and orientation and are good to detect edge orientations in images. The only drawback of Gabor Filters is that it is scale-sensitive. We alsoaddedtheaverageandstandarddeviationofbrightnessforeachregiontocomplement the information provided by the GaborFilter. Also known as Query By Image Content (QBIC), presents the technologies allowing to organize digital pictures by their visual features. They are based on the application of computer vision techniques to the image retrieval problem in large databases. Content-Based Image Retrieval (CBIR) consists of retrieving the most visually similar images to a given query image from a database of images. Learn more in: Using Global Shape Descriptors for Content Medical-Based Image Retrieval

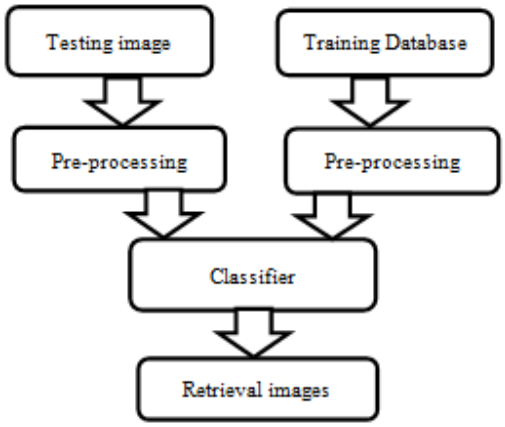


Fig 1: Content-based image retrieval block diagram

the input of a CNN, in this case an image, is passed through a series of filters in order to obtain a labelled output that can then be classified. The specificity of a CNN lies in its filtering layers, which include at least one convolution layer. These allow it to process more complex pictures than a regular neural network. Whereas the latter is well adapted for simple, well-centered images such as hand-written digits, the use of CNNs in image analysis ranges from Facebook's automatic tagging algorithms, to object classification and detection, in particular in the field of radiology.

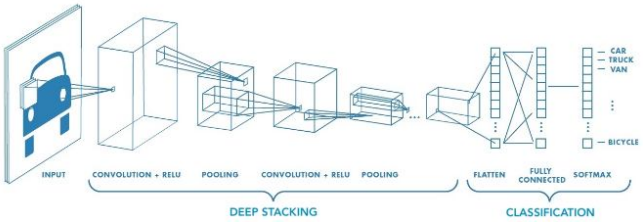


Fig 2: Convolution neural network structure

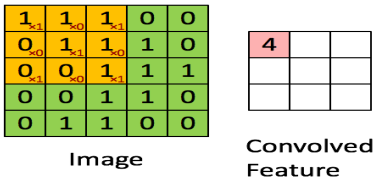


Figure 3: A diagram showing a convolution

Max-pooling a form of non-linear down-sampling is an important concept of CNNs. The input image is partitioned into a group of non-overlapping rectangles and a maximum value is given for each such sub-region. We use max-pooling in vision for the following reasons- The computation of upper layers is reduced by the removal of non-maximal values. Suppose a max-pooling layer is cascaded with a convolutional layer. The input image can be translated by a single pixel in 8 directions. 3 out of 8 possible configurations produce

exactlythe same output at the convolutional layer if max-pooling is done over a 2x2 region. This jumps to 5/8 for max-pooling over a 3x3 region [6]. A form of translation invariance is provided by this. The dimensionality of intermediate representations is reduced by max-pooling because it provides additional robustness to position.

III. PROPOSED METHOD

The convolution neural network has been implemented as following steps.

Step 1: The input image is taken from the COREL database. Each image size is 256*256*3. Where number of rows are 256, number of columns are 256 and directional range is 3.
Step 2: The input image has to decompose with original image using various convolution filters such as sharpen, edge, line and emboss etc. the possible image filters sre 3*3 filters, 5*5 filters and 7*7 filters.
Step 3: To extract the features using padding technique. There are two types of paddings one is zero padding and another one is one's padding.
Step 4: To reduce the dimensionality the image can be implemented by using pooling method. There are different types of pooling methods. One is max pooling, secondly, min pooling and average pooling etc.
Step 5: To extract the high content of the features from large set of data it can be applied to number of layers based on the number of filters. It is also called as kernel of the input image.
Step 6: after completion of number of layers, it is composed in the domain of fully connect layer.
Step 7: The final layer is to connected to the classifier method to classify the test image from the large set of databases.

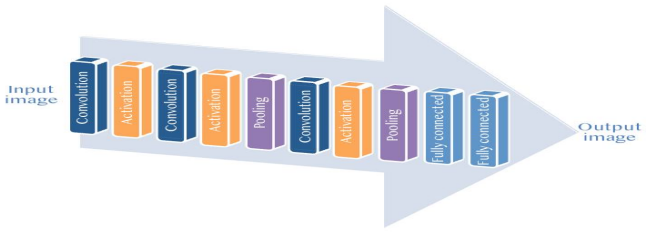
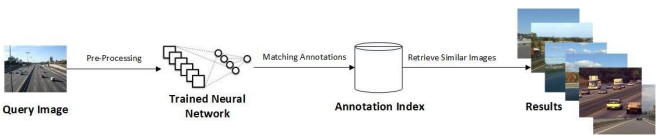


Figure 4: The basic flow of the LeNet-5 Layer for the content-based image retrieval



the image decomposition is represented as below based on the pooling layers.

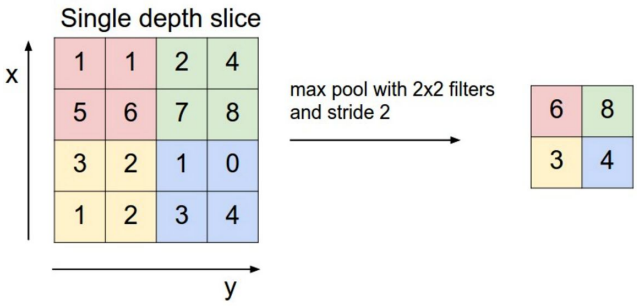


Figure 5: Convolution process

IV. EXPERIMENATAL RESULTS

The dataset I chose for this thesis is from the SUN database [12]. The major reason for choosing this dataset was that the images in it were pre-annotated and had annotations as XML files for each image. The SUN database is huge so I had to choose a small subset of it for this study. In this study I am trying to classify images based on 8 classes namely: water, car, mountain, ground, tree, building, snow, sky and unknown which contains all the rest of the classes. I chose only those sets of images which I felt were more relevant to these classes. I collected a database of 3000 images from 41 categories. Each image has its annotations in an XML file. I randomly divided the dataset into 80% training set and 20% testing. There are 1900 training images, 600 testing images and 500 validation images. The training set was further divided into 80% training set and 20% validation set. The major drawback of this dataset is that the images are annotated by humans and the annotations are not perfect thus it may have some effect on the results. I try to handle this problem by getting as many synonyms as I can for each class label. A few examples of the synonyms are lake, lake water, sea water, river water, wave, ripple, river, sea, river water among others which all belong to the class label water. I mapped these synonyms to their respective class labels which are being used. Not all images in every category were annotated. I filtered out the annotated images from the dataset and used only them for this study. Fig.4.5 shows an example of an image from the dataset and its annotation file where it can be seen how a river is annotated by the user.A little pre-processing was required on the dataset before it could be trained because of the way the code for CNN training was written. The images were converted to grayscale and resized to 28x28 pixels. I used the annotation files to get a flag for each class present or absent from the image and using the flags I compressed the dataset into a 1D array which contains the image dimensions and binary values for each class where 1 states that class is present and 0 states the class is absent. The compressed data is then trained by the neural. The major constraint I had to put is the downsizing of images due to memory issues. The images had to be resized

which resulted in loss of data. This problem can be avoided as shown in the work of Krizhevsky et al [5]. In their model the image can be of any square size and can have any number of channels (color channels). Removing this constraint will make the system more robust. One thing that needs to be kept in mind while using original dimensions of the image is that a good GPU will be required for training the convolutional neural network otherwise there will be memory issues which I faced.

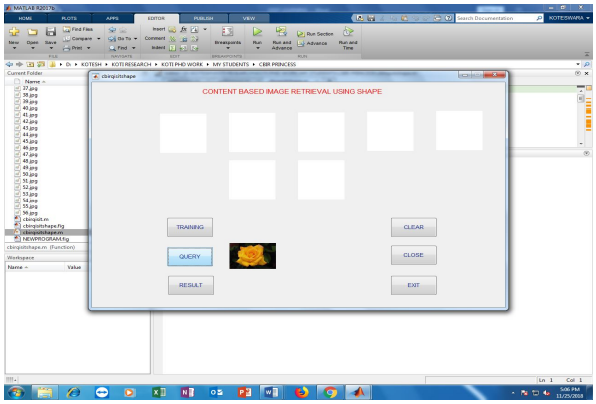
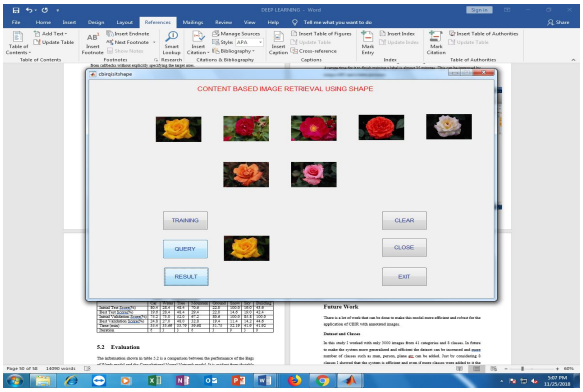
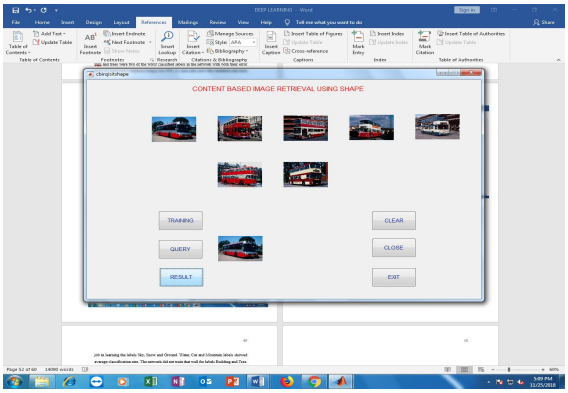


Figure 6: Testing image





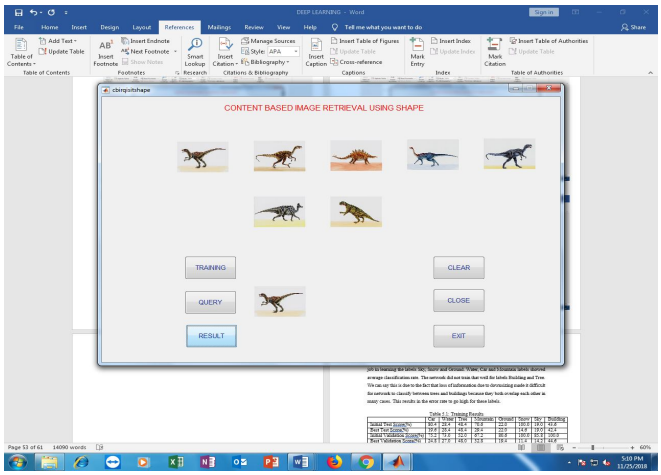


Figure 7: Retrieval images

The major constraint I had to put is the downsizing of images due to memory issues. The images had to be resized which resulted in loss of data. This problem can be avoided as shown in the work of Krizhevsky et al [5]. In their model the image can be of any square size and can have any number of channels (color channels). Removing this constraint will make the system more robust. One thing that needs to be kept in mind while using original dimensions of the image is that a good GPU will be required for training the convolutional neural network otherwise there will be memory issues which I faced.

Table 1: Content based image retrieval comparison

	Flower	Beach	Bus	Elephant
Initial Test Score	80.4	28.4	48.4	22.0
Best Test Score (%)	19.6	28.4	48.4	22.0
Initial Validation Score(%)	75.2	73.0	52.0	80.6
Best Validation Score(%)	24.8	27.0	48.0	19.4
Time (min)	35.4	33.69	33.79	31.75
Iteration	6	3	3	3

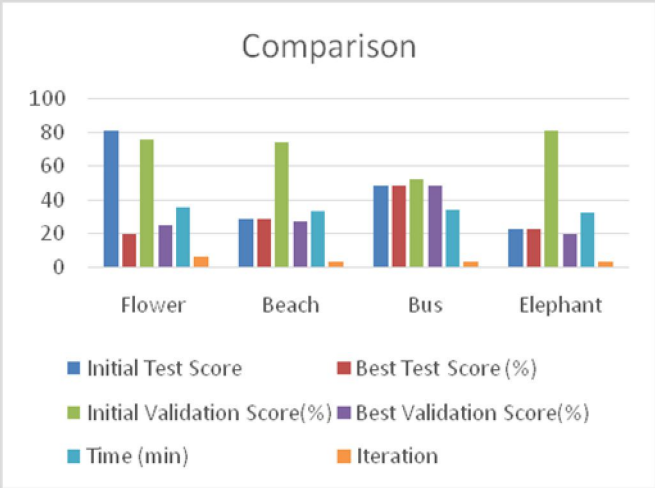


Figure 8: Comparison chart for Content Based Image Retrieval.

IV. CONCLUSION

There is a lot of work that can be done to make this model more efficient and robust for the application of CBIR with annotated images.In this study I worked with only 3000 images from 41 categories and 8 classes. In future to make the system more generalized and efficient the dataset can be increased and more number of classes such as man, person, plane etc can be added. Just by considering 8 classes I showed that the system is efficient and even if more classes were added to it the performance of the system would only increase.

V. ACKNOWLEDGMENT

Authors would like to thank management of Indira institute of technology, Markapur for providing facilities to finish this work.

REFERENCES

[1] Fre´de´ric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfel- low, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[2] James Bergstra, Olivier Breuleux, Fre´de´ric Bastien, Pascal Lamblin, Razvan Pas- canu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In Proceedings of the Python for Scientific Computing Conference (SciPy), June 2010. Oral Presentation.

[3] Yining Deng and B. S. Manjunath. Unsupervised segmentation of color-texture re- gions in images and video. IEEE Trans. Pattern Anal. Mach. Intell., 23(8):800–810, August 2001.

[4] Yahong Han, Fei Wu, Qi Tian, and Yueting Zhuang. Image annotation by input 2013;output structural grouping sparsity. Image Processing, IEEE Transactions on, 21(6):3066–3079, June 2012.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
University of Montreal LISA Lab. Deep learning tutorial, 2008.

[6] Ying Liu, Dengsheng Zhang, and Guojun Lu. Region-based image retrieval with high-level semantics using decision tree learning. Pattern Recogn., 41(8):2554–2570, August 2008.

[7] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content- based image retrieval with high-level semantics. Pattern Recogn., 40(1):262–282, January 2007.

[8] Chih Fong Tsai. Bag-of-words representation in image annotation: A review. ISRN Artificial Intelligence, 2012(1), 2012.

[9] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehen-sive study. In Proceedings of the ACM International Conference on Multimedia, pages 157–166. ACM, 2014.

[10] Wikipedia. Convolutional neural network — wikipedia, the free encyclopedia, 2015. [Online; accessed 9-May-2015].

[11] Jianxiong Xiao, Jianxiong Xiao, James Hays, J. Hays, Krista A. Ehinger, K. A. Ehinger, Aude Oliva, A. Oliva, A. Torralba, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. pages 3485–3492. IEEE, 2010.

[12] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. Evaluat- ing bag-of-visual-words representations in scene classification. In Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, MIR '07, pages 197–206, New York, NY, USA, 2007. ACM.

[13] Dengsheng Zhang, Md Monirul Islam, Guojun Lu, and Jin Hou. Semantic image retrieval using region based inverted file. In Proceedings of the 2009 Digital Image Computing: Techniques and Applications, DICTA '09, pages 242–249, Washington, DC, USA, 2009. IEEE Computer Society.

[14] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[15] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
[16] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
[17] K. Elissa, "Title of paper if known," unpublished.
[18] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[19] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[20] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.