ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse217942020.pdf
https://doi.org/10.30534/ijatcse/2020/217942020

Deep Learning-based Self Organizing Map for Content-based Image Retrieval Applications

Eman Awny[1], Alaa Sagheer [1,2,*]
[1]Center for Artificial Intelligence and Robotics, Department of Computer Science, Aswan University, Egypt
[2] College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia
* asagheer@aswu.edu.eg

ABSTRACT

Image classification and retrieval are exciting topics in machine learning and computer vision applications. However, many algorithms can achieve these tasks; it is widely demonstrated that the best algorithms that attain high accuracy are deep learning algorithms. Inspired by the convolutional neural network (CNN) architecture, this paper presents a deep learning approach that replaces the convolution layer in the standard CNN with a layer of self organizing map neural network. At the same time, it keeps the activation function layer and pooling layer as they are in the standard CNN architecture. The proposed approach is employed to achieve image content-based image retrieval (CBIR), particularly in the case of large datasets used. Besides, it can alleviate the semantic gap problem, which refers to the difference between low-level and higher-level image representations. To assess the proposed approach, we employed the tiny ImageNet visual recognition dataset. The experimental results show that as we increase the depth of the network as the retrieval accuracy is improved.

Key words: Deep Learning, Convolutional Neural Network, Self Organizing Maps, Content-based Image Retrieval.

1. INTRODUCTION

Doubtless, this is the era of big data where rapid growth in the size of digital image datasets has been observed and will continue in the future [1]. We are shifting from how we collect data to how to process and retrieve information from them in real-time. Consequently, retrieval and querying of massive datasets of images efficiently are required to understand the visual content of an image. Content-based image retrieval (CBIR) provides the solution for efficient image retrieval [2]. CBIR definesgetting images similar to a query image, from such a massive collection, based on their visual contents.

Image retrieval is an old subject of research, where earlier achievements include manual annotation of images using search of a text or keywords. In contrast, given the feature representations of several images to be searched and the query image, CBIR achieves an automatic image retrieval in terms of similarity with the query image. The key factor here that associated with CBIR is to extract meaningful content from raw data to alleviate the so-called semantic-gap [3]. The semantic-gap refers to the difference between the low-level representations of images and their higher-level representations [4].

Traditional approaches employed to achieve CBIR are including scale-invariant feature transform [5], fisher vector descriptors [6], local binary pattern [7-8], and combine bag-of-words models [9]. Most of these traditional approaches were concerned with retrieving primitive features that were enough to describe the image content, such as shape, texture, and color [4]. However, the valuable results these approaches have attained, their performance degrades when processing big datasets. Unlike shallow neural network-based models or handcrafted features-based models, deep learning algorithms[10] use more-sophisticated architecture, incorporating several processing layers to yield robust local image representations with multiple abstraction levels. The remarkable success of deep learning algorithms is due to many reasons, including recent advances in the GPU-based computations and the availability of large annotated datasets [11].

Deep Convolutional Neural Networks (CNN) are considered the most efficient deep learning paradigm for visual content analysis [12]. CNNs comprise many convolutional, sub-sampling, and fully connected layers, with nonlinear neural activations. Such deep architecture of CNN hasshown a high capability to capture local features of image objects, e.g. corners and edges, much better than handcrafted feature-based models. Therefore, CNNs are very suitable to achieve image retrieval and widely used for matching local patterns of objects [13].

Despite the promising reported performance attained by CNNs, there are several limitations, and open questions require deep investigation. Foremost, there is no universal agreement, or a deep understanding, of how the intermediate hidden layers of CNN work. Most of the CNN-based models utilize the last layer to extract the image features with an orderless quantization approach, limiting the utilization of intermediate CNN hidden layers. Second, the discrimination accuracy of image features directly retrieved from the convolutional layers is against the features that traditional approaches retrieve. Third, a comprehensive investigation is needed on how the interactions among several CBIR aspects

should be, including similarity matching and retrieval performance in memory usage and search time [13].

All these limitations motivated us to propose a different deep CNN architecture to benefit from the intermediate CNN layers. The self organizing map (SOM) [14], a topological preserver of data space, can provide an efficient alternative to the convolution layer in modeling contextual constraints in the image contents and features [15]. In this paper, we present a deep learning architecture that replaces the convolution filter in CNN with a SOM layer. In this way, the local information is aggregated to represent more global information in the upper layers. Since the CBIR system retrieves relevant images based on the similarity of their features with the features of a query image, the SOM as an unsupervised learning-based approach can compare between images by measuring the similarities, which ensures implementation and efficiently improve CBIR. Henceforth, the proposed method will be denoted as DSOM-CBIR.

2. BACKGROUND

2.1 Self OrganizingMap (SOM)

The SOM algorithm defines a nonlinear competitive learning mapping from an input space into a topologically ordered low-dimensional map, or grid, of neurons [14]. It is a kind of artificial neural network approach that has proven extremely effective in converting and mapping data to represent it in low dimensions from a typically high dimensional [15]. This transform is called a code map, as demonstrated in Figure1.

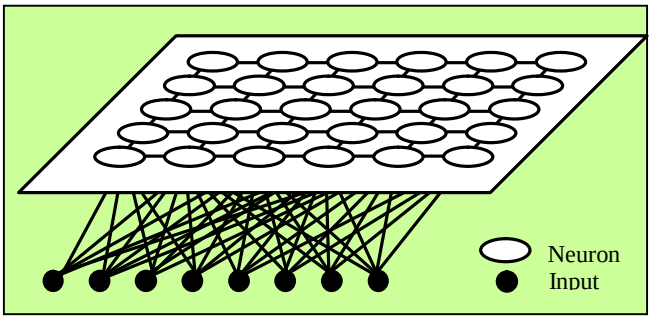

Figure 1: The SOM model

The remarkable advantage of SOM in computer vision applications is the consistency within the space of representation between the input data as represented in the input data space is maintained as accurate. SOM can learn and identify input regularities and similarities and predict input responses. The neurons of SOM learn to identify groups with similar input representations during the competitive phase of SOM and start recognizing the collection of identical input vectors to allow neurons in the SOM layer to adapt to the same input features vectors physically close to each other.

Competitive learning is a procedure in which the neurons of a neural network are tuned to specific input patterns by competing with each other during the learning phase. Assume that the SOM model consists of a single layer of N neurons, each of which is fully connected to a set of M inputs. Each neuron j stores (randomly initialized) a weight vector $\mathbf{w}_j = [w_{j1}, w_{j2},\ldots,w_{jM}]$; where $j = 1, 2,\ldots, N$ refers to the number of neurons. At every instant time t an input vector, $\mathbf{x} = [x_1, x_2,\ldots,x_M]$is presented to the network. Then, the best matching neuron, or the *winner*, "c" is selected if it has the smallest Euclidean distance from the input, of course, any other distance measure can be used. The distance can be given as:

$$\left\|\mathbf{x}(t) - \mathbf{w}_c(t)\right\| \leq \left\|\mathbf{x}(t) - \mathbf{w}_j(t)\right\| \quad (1)$$

Then, the *winner* weight vector is updated so that the distance between its weight vector $\mathbf{w}_j$ and the input vector $\mathbf{x}$is decreased by a certain fractional amount$\alpha(t)$. Then the update rule can be given as:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t)\left(\mathbf{x}(t) - \mathbf{w}_j(t)\right) \quad (2)$$

Such that $\alpha(t) = \gamma(t) \cdot h_{cj}(t)$ (3)

where $\gamma(t)$is the learning rate, and$h_{cj}(t)$isa neighborhood function represents a smoothing kernel function influences the update of the weight vectors of neurons in such a way that a neuron j that far away from the winner neuron c in the map experience less weight change than neurons close to the winner [15].The above two equations are denoted as the "winner-take-all" rule, which SOM computation is based upon them. After the training phase is exhausted, the SOM feature map neurons tend to become fine-tuned to different regions of the input. The results obtained using such a rule are identical if the input and weight vectors are normalized.

2.2Deep Learning

Deep learning is a kind of machine learning method, wherein hierarchical architecture, several layers of unit processing phases are used to identify patterns for learning features or representations. Deep learning intersects with many different research fields such as pattern recognition, artificial neural networks, computer vision, signal and image processing, optimization, etc. The Neural feed-forward model for multiples hidden layers is a clear example of the architectures in deep layers. Since the earlier work of Hinton et al. [16], this trend of research still attracts full attention in image classification research where deep learning models achieve high accuracy and performance in various tasks.

CNN is a primary deep learning model used widely to process visual contents in various computer vision and pattern recognition applications[17], [18]. It is a feed-forward neural network architecture inspired by the human beings' natural visual perception mechanism [19]. It comprises multi-layers of two repeated layers: the convolution layer and the pooling layer, besides the input layer in front and a fully connected layer at the end, as shown in Figure 2.

Such deep architecture of CNN has shown a high

capability on capturing local features of image objects, e.g., corners and edges, much better than handcrafted feature-based models. Therefore, CNNs are very suitable to achieve image classification and retrieval and widely used for matching local patterns of objects [13].

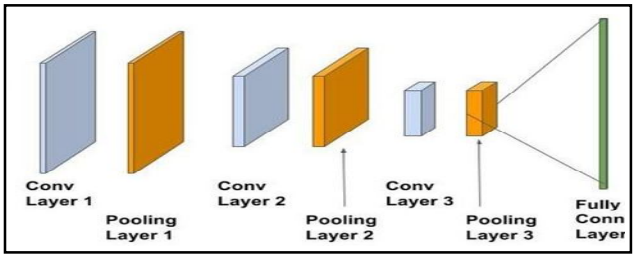

Figure 2: The standard CNN

Figure 3 shows the deep CNN model that appeared in the ILSVRC-2012 task of the image classification demonstrated, from which the proposed model is inspired.

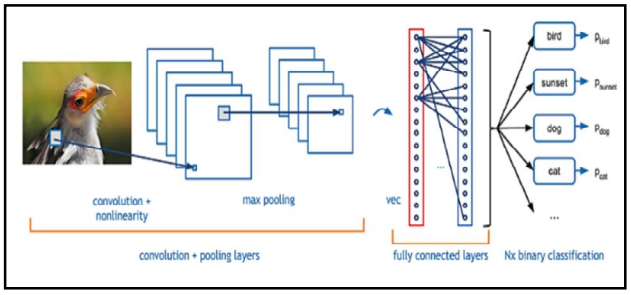

Figure 3: Deep CNN for image classification. The figure imported from the original article [12]

However, the reported results attained by CNNs in many applications, many open questions and limitations need more discussion. First, there is no clear evidence of how the convolution layers of CNN work. State-of-the-art models of CNN utilize the output layer to extract the image features with an orderless quantization approach, which limits the benefits of intermediate CNN hidden layers. Second, a comprehensive investigation is needed on how the interactions among several CBIR aspects should be, including similarity matching and retrieval performance in terms of search time and memory usage. Finally, the discrimination accuracy of image features that directly retrieved from the convolutional layers against the features extracted by traditional approaches [13].

2.3 Content-based Image Retrieval (CBIR)

CBIR is one of the major research problems in computer vision field [2-4]. CBIR seeks to scan images by evaluating their visual content, and thus the representation of images is the crux of CBIR. This task can be performed via many low-level descriptors for the description of images in recent decades. Examples of descriptors include color characteristics and edge features [20], texture features [21], GIST [22], and CENTRIST, and current local feature characteristics, such as Bow models [23] using local feature descriptors [24] (e.g., SIFT, and SURF [25]).

For multimedia similarity searches, traditional CBIR approaches typically prefer a rigid distance method on certain low-level extracted features, such as cosine function or Euclidean distance for similarity. Because of the significant problem of the semantic difference between low-level Machine-derived visual representation and high level perceived by humans, setting a static distance feature may not always be suitable for complicated image classification and recovery tasks. A CBIR model offers an effective way to retrieve images from image collections. CBIR aims to rescue images from an image of a dataset that is close to one or more sample images of their image content. The user submits such question images. The CBIR method then ranks the images of the database and displays the results obtained concerning their similarity to the images of the query [26].

The characteristics of CBIR systems:

(1) The gap between high and low -level features.
(2) The extraction of visual information perceived by humans.
(3) The image representation model itself.
(4) The structure of the learning method to promote interaction.

3. THE PROPOSED MODEL

We now introduce theproposed deep self organizing map learning model for content-based image retrieval (DSOM-CBIR).

3.1 The SOM Layer

As shown in Figure 4, the proposed model architecturecomprises two units: the triagingunit and the testing (retrieval) phase. The training part includes an inspired CNN block; each consists of a SOM layer, exponential linear unit (ELU) activation function layer, and amax pooling layer. This block can be repeated according to the application at hand.At the end of the training part, an output layer bears the features of the input image stream.

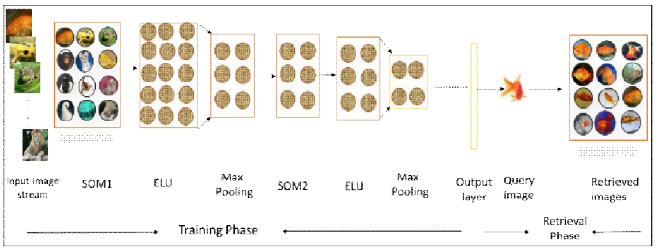

Figure 4: Architecture of the proposed DSOM-CBIR model

For the experiments of this paper, we used five blocks. In the beginning, we initialized the weight vector of the SOM neurons as well as other parameter learning rates and the neighborhood radius of the SOM feature map. Then, we implement the SOM formulas (1)-(2) until we find the best matching unit or the winner neuron. After the training data are exhausted, the map's neurons are automatically organized into a meaningful two-dimensional ordered map, usually

denoted as the codebook.

Each SOM layer is fitted with a 2 × 2 pixel max pooling kernel and stride as two.The ELU activation function's role is to speed up and adjust each SOM layer's performance.Each SOM layer processes and extracts various features from the input, the output of each block is then concatenated at the end before it is passed to the followingblock.Then, each layer isrepresenting different features extracted from the previous layer. This deeparchitecture dramatically reduces the number of parameters required for training. Besides, the network structure we have described so far can detect just a single kind of globalized feature.

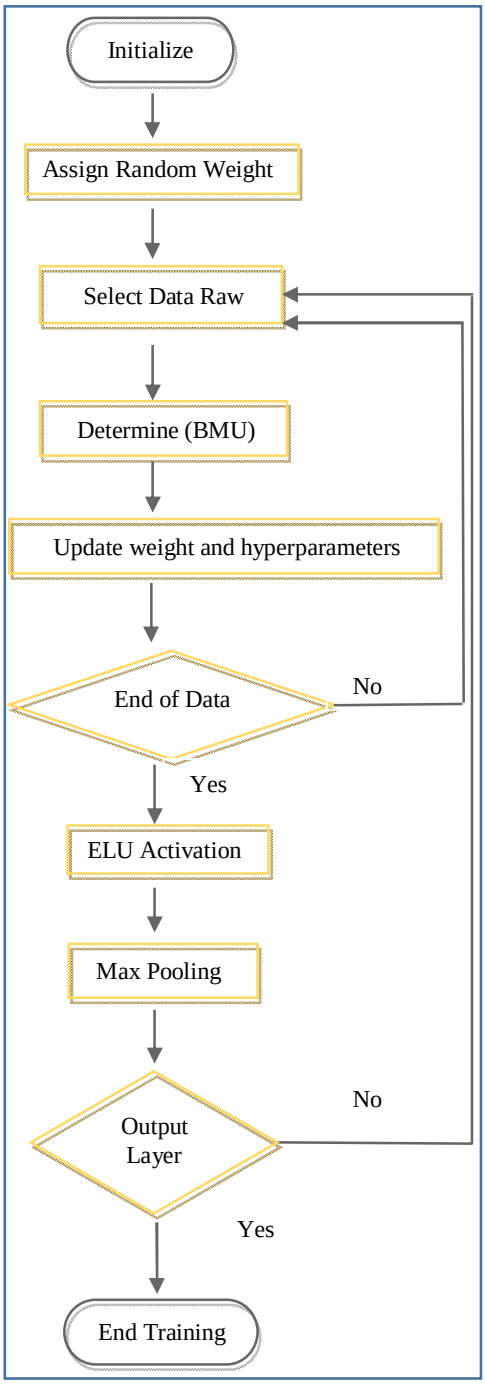

Figure 5: The DSOM-CBIR flowchart

3.2 The ELUs Layer

The ELUs tend to setthe mean activation closer to zero, which accelerates the learning process. It has been shown that ELUscan obtain higher classification accuracy than rectified linear units (ReLUs) [25]. The ELUscan be given as:

$$f(x) = \begin{cases} x & if\ x > 0, \\ \alpha(e^{x}-1) & otherwise, \end{cases} \quad (4)$$

Where the parameterαcontrols the value to which ELUs saturate for negative net inputs.Another benefit ELUs is that italleviates the effect of the vanishing gradient problem as ReLUs and leaky ReLUsare doing. The vanishing gradient problem is alleviated since the positive component of these functions is the identity, and its derivative is one and not contractive. In contrast, sigmoid and *tanh* activation functions are usually contractive. A most systematic method in finding a solution to the problem of bias shifts while at the same time mitigating the problem of vanishing gradients [27]. Figure 6 embeds such an attitude of these functions.

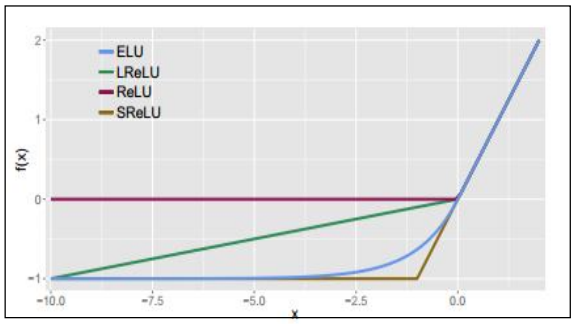

Figure 6: The ReLUs, the leaky ReLUs (α = 0:1), the shifted ReLUs (SReLUs), and the ELUs (α = 1:0)

3.4 The Max Pooling layer

The pooling layers aim to achieve spatial invariance by reducing the characteristics of the map resolution. As shown in Figure7, the input is a small*N* x *N*patch of units.This poolingkernel (window) can take various shapes, and kernels can overlap. Numerous pooling operations exist max pooling, average pooling, and subsampling. The subsampling function can be given as:

$$a_j = \tanh(\beta \sum_{N \times N} a_i^{n \times n} + b) \quad (5)$$

From all the inputs takes the average, multiplies by learning rateβ, keep adding a bias *b*, and transmits the result by the *tanh* nonlinearity function. The max poolingfunction can be given as:

$$a_j = \max_{N \times N}(a_i^{n \times n} u(n,n)) \quad (6)$$

It uses a kernel function $u(n,n)$ to the input patch and calculates the maximum number in the neighborhood. In all cases of applying pooling, the result is a feature codebook map of lower dimensionality. In our proposed model in this paper, we employ the max pooling activation function where the max pooling layer shrinks the map of DSOM by 2x2 max pooling kernel with stride 2.

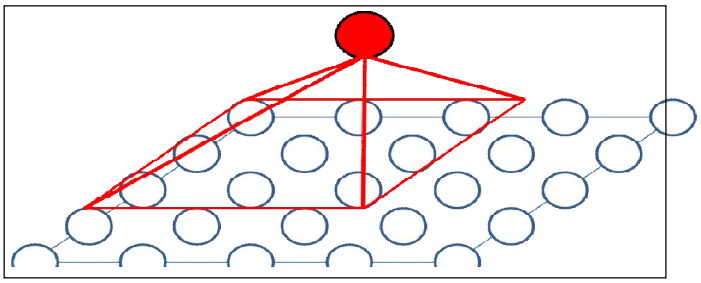

Figure 7: Max pooling 3x3 kernel shrinks map and for each such sub-region outputs maximum value from neurons values

4. Experiments and Results

This section shows experimental settings and the results of the proposed DSOM-CBIR model.

4.1 The Dataset

To assess the proposed approach, the experiments of this paper uses the dataset "Tiny ImageNet Visual Recognition Challenge" [28]. This dataset runs similar tothe original ILSVRC2016 dataset [29], an extensive collection of images with 100,000 labeled images representing 1000 classes. These images are obtained from Flickr and other search engines, tagged with the absence of 1000 categories. The Tiny ImageNet dataset is the thumbnail version of the ILSVRC2016 dataset, but the images are reduced to 64x64 from 224x224. It has 200 class, rather than 1000 ImageNet challenge; each class has 500 training images, 50 validation image, and 50 test image. Therefore, it has 10,000 validation images,10,000 test images, and 100,000 training images.

4.2 Experimental Setting

In our experiments, we used 20,000 images from the above-described dataset. We divided this number into 75% for training and 25% for testing, with the total number of classes 33, like dog, cat, bear, bee, monkey, chair, etc. as shown in Figure 8.The input to the first SOM layer is an image with a fixedsize as a 64×64 RGB image. No augmentation of the image dataset was performed during training. No batch normalization is used inside the DSOM-CBIR layers. The only preprocessing we did is the conversion from the RGB color images (i.e. 24 bits) into a grayscale image (8 bits) so that the color model will be more straightforward with each pixel grey level from 0 to 255, as shown in Figure 9.

The building block of the proposed DSOM-CBIR model includes five blocks such that each block contains a SOM layer, ELU layer, and Max pooling layer, as described in section 3.We improve the block's performance using the ELU activation, and max pooling with kernel size 2x2 is applied to every layer. Table 1 shows the configuration of each block in the DSOM-CBIR model.

The source code of the proposed approach is implemented in Python via the Tensorflow library with a scikit-learn. The experiments are run on a PC with spider IDE with 2.30 GHz speed Intel(R) Corei5 processor and RAM 4 GB.

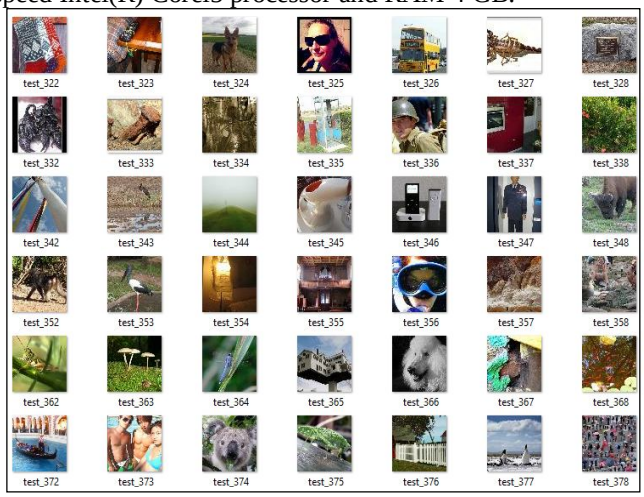

Figure 8: Sample images from different classes of the dataset

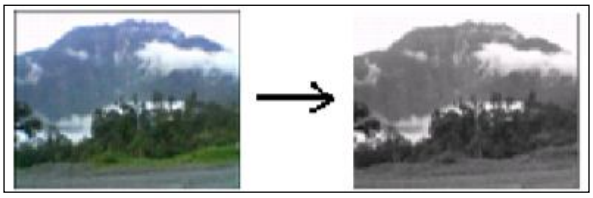

Figure 9: RGB color conversion into grayscale

We setthe first SOM layer with a size of 30×30 neurons, and the size reduced gradually until the size of 10x10 at the fifth SOM. We initialize the learningrate as 0.5 and initialize the Sigma parameter (i.e. the radius of the map) as 3. Generally, after each decay of the learning rate, the networks converge and reduce the learning rate and the sigma at each iteration.

Table 1: DSOM-CBIR Configuration and Accuracy of each layer

DSOM-CBIR Model Configuration					
Block	***Size***	**ELU**	**Pool*ing***	***Input* Size**	**Accuracy**
SOM1	30x30	64x64	2x2	64x64	64.6 %
SOM2	15x15	32x32	2x2	32x32	71.3 %
SOM3	10x10	16x16	2x2	16x16	74.96%
SOM4	10x10	8x8	2x2	8x8	81.76 %
SOM5	10x10	4x4	2x2	4x4	84.98 %

4.3 *Experimental Results*

Besides the configuration of each block,Table 1 shows the accuracy of each block separately, where this accuracy is the testing retrieval accuracy. The accuracy of retrieval is growing gradually from 64% in the first block, up to 85% in the last one. This is a clear indication of the impact of deep architecture on the learning process. A visual representation of the results shown in Table 1 is depicted in Figure 10.

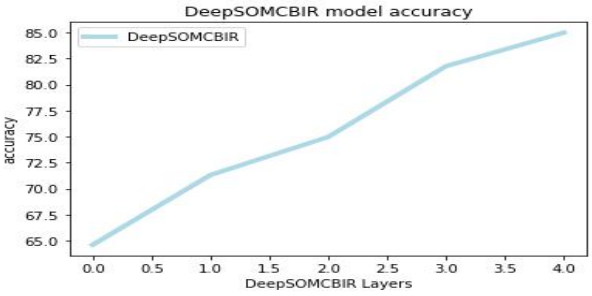

Figure 10: The accuracies for 5 DeepSOMCBIR model graph.

Table 2: Training and Testingtime (in *sec.*)for each block in DSOM-CBIR

Block	Training Time	Testing Time
SOM1	902.434	449.904
SOM2	119.196	74.467
SOM3	26.166	8.386
SOM4	24.956	6.183
SOM5	12.950	5.884

5. CONCLUSION

This paper addressed image retrieval, which isan interesting topic in the computer vision field. The literature contains many algorithms that can achieve these tasks; however,here we demonstrated that the best algorithm thatcan attain high accuracy isthe deep learning algorithm. In this paper, we presented a deep learning approach, inspired by the standard convolutional neural networkarchitecture. Here we replaced the convolution layer in the standard CNN with a layer of SOM neural network, keeping the other CNN elements as they are. The proposed approach is employed to achieve image content-based image retrieval (CBIR) application for a large dataset, namely, the tiny ImageNet visual recognition. The experimental results showed that as we increase the depth of the proposed deep architecture as the performance is growing, with a minor portion of execution time. As future work, we will apply the proposed approach in different image retrieval applications from the medical domain.

REFERENCES

1. L. Wei, Y. Dao-Ping, J. Yan and Y. Mao-Qiang. **The Era of Big Data of Image Retrieval and Recognition Technology.** The International Conference on Intelligent Transportation, Big Data and Smart City, Halong Bay, pp. 858-861, 2015. https://doi.org/10.1109/ICITBS.2015.217
2. V. Tyagi. **Content-Based Image Retrieval Techniques: A Review**. In: Content-Based Image Retrieval. Springer, Singapore, 2017.
3. A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, **Content-based image retrieval at the end of the early years.** in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, no. 12, pp. 1349-1380, 2000.
4. Maria Tzelepi, A. Tefas. **Deep convolutional learning for Content Based Image Retrieval**. Neurocomputing, Vol. 27531, pp. 2467-2478, 2018.
5. J. Sivic, A. Zisserman. **Video google: a text retrieval approach to object matching in videos**. Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1470–1477, 2003 . https://doi.org/10.1109/ICCV.2003.1238663
6. F. Perronnin, Y. Liu, J. Sánchez, H. Poirier. **Large-scale image retrieval with compressed fisher vectors**. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 3384–3391, 2010.
7. A. Sagheer, S. Saad. **Dominant Local Binary Pattern Considering Pattern Type for Facial Images Representation**. Proceedings of the 18th International Conference on Image Analysis and Processing (ICIAP), II, LNCS 9280, pp 252–263, 2015.
8. S. Saad, A. Sagheer. **Difference-based Local Gradient Patterns for Image Representation**. Proceedings of the 18th Inter Conf. on Image Analysis and Processing (ICIAP), II, LNCS 9280, pp. 472–482, 2015.
9. F. Jiang, H.-M. Hu, J. Zheng, B. Li. **A hierarchal bow for image retrieval by enhancing feature salience.** Neurocomputing, Vol. 175, 146–154, 2016.
10. J. Schmidhuber. **Deep learning in neural networks: An overview**. Neural Networks, Vol 61, pp. 85–117, 2015.
11. Y. Guo , Y. Liu , A. Oerlemans , S. Lao , S. Wu , M.S. Lew. **Deep learning for visual understanding: a review**. Neurocomputing, Vol. 187, pp. 27–48, 2016.
12. A. Krizhevsky, I. Sutskever, G.E. Hinton. **ImageNet Classification with Deep Convolutional Neural Networks.** Proceedings of the 25th International Conference on Neural Information Processing Systems NIPS, Vol. 1, pp. 1097–1105, 2012.
13. A. Alzu'bi, A. Amira, N. Ramzan. **Content-based image retrieval with compact deep convolutional features.**Neurocomputing, Vol. 2492, pp. 95-105, 2017. https://doi.org/10.1016/j.neucom.2017.03.072
14. T. Kohonen. **Self-organized formation of topologically correct feature maps**. Biological Cybernetics, Vol. 43, pp. 59-69, 1982.
15. A. Sagheer, N. Tsuruta and R. Taniguchi. **P1DSOM- A Fast Search Algorithm for High-Dimensional Feature Space Problems**. International Journal on Pattern Recognition and Artificial Intelligence, IJPRAI, Vol. 28, no. 2, 1459005, 2014.

16. G.E. Hinton, S. Osindero, Y-W Teh. **A fast learning algorithm for deep belief nets**. Neural Computation, Vol. 18, No. 7, pp. 1527–1554, 2006.
17. N. Manikandan,**Approximation Computing Techniques to Accelerate CNN Based Image Processing Applications – A Survey in Hardware/Software Perspective**. International Journal of Advanced Trends in Computer Science and Engineering,Vol. 9,pp. 3828-3846, 2020. https://doi.org/10.30534/ijatcse/2020/202932020
18. Alharthi, Adil**, Convolutional Neural Network based on Transfer Learning for Medical Forms Classification 34,** International Journal of Advanced Trends in Computer Science and Engineering,Vol. 8, pp.3405- 3411, 2019. https://doi.org/10.30534/ijatcse/2019/115862019
19. A. Li, M. Yuan, C. Zheng, X. Li,**Speech enhancement using progressive learning-based convolutional recurrent neural network**.Applied Acoustics, Vol. 166, 107347, 2020.
20. Sridhar, Gowri. **Color and Texture Based Image Retrieval ARPN Journal of System and Software**, Vol. 2, no. 1, pp 1−6,January2012.
21. M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. **Content-based multimedia information retrieval: State of the art and challenges**. TOMCCAP, Vol. 2, no. 1, pp.1–19, 2006. https://doi.org/10.1145/1126004.1126005
22. A. Oliva and A. Torralba. **Scene-centered description from spatial envelope properties**. In Biologically Motivated Computer Vision, pp. 263–272, 2002.
23. M. Norouzi, D. J. Fleet, and R. Salakhutdinov. **Hamming distance metric learning.** In NIPS, Vol. 6, no.5, pp. 1070–1078, 2012.
24. D. G. Lowe. **Object recognition from local scale-invariant features**. In ICCV, pp.1150–1157, 1999.
25. H. Bay, T. Tuytelaars, and L. J. V. Gool. Surf: **Speeded up robust features**. In ECCV, no. 1, pp. 404–417, 2006.
26. Deepak S. Shete1, M.S. Chavan "**Content Based Image Retrieval: Review**", International Journal of Emerging Technology and Advanced Engineering ISSN, Vol. 2, pp. 2250-2459, 2012.
27. Djork-Arne Clevert, Thomas Unterthiner , Sepp Hochreiter**.," Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUS)** ", Proceeding of the 4thInternational Conference on Learning Representations ICLR, 2016.
28. Tiny ImageNet Visual Recognition Challenge Available online at:https://tiny-imagenet.herokuapp.com/
29. ILSVRC2016 dataset. Available online at: http://image-net.org/challenges/LSVRC/2016/