Image Compression: An Artificial Neural Network Approach

Anjana B[1], Mrs Shreeja R[2]

[1] Department of Computer Science and Engineering, Calicut University, Kuttippuram
[2] Department of Computer Science and Engineering, Calicut University, Kuttippuram

Abstract

Image compression has become the most recent emerging trend throughout the world. Image compression is essential where images need to be stored, transmitted or viewed quickly and efficiently. The artificial neural network is a recent tool in image compression as it processes the data in parallel and hence requires less time and is superior over any other technique. The reason that encourage researchers to use artificial neural networks as an image compression approach are adaptive learning, self-organization, noise suppression, fault tolerance and optimized approximations. A survey about different methods used for compression has been done. From the above study, recently used network is multilayer feed forward network due to its efficiency. The choice of suitable learning algorithm is application dependent. A new approach by modifying the training algorithm to improve the compression is proposed here. Protection of image contents is equally important as compression in order to maintain the privacy. If any malicious modification occurs either in storage or in transmission channel, such modifications should be identified. So authentication and protection are incorporated into the proposed system to enhance the security.

Keywords: Jacobian, Levenberg-Marquardt, Multilayer perception, Neural network, Radial basis function.

1. Introduction

Image compression has become the most recent emerging trend throughout the world. Some of the common advantages of image compression over the internet are reduction in time of web page uploading and downloading and lesser storage space in terms of bandwidth. Compressed images make it possible to view more images in a shorter period of time. Image compression is essential where images need to be stored, transmitted or viewed quickly and efficiently. Image compression is the representation of image in a digitized form with a few bits maintenance only allowing acceptable level of image quality. A high quality image may require 10 to 100 million bits for representation. The large data files associated with images thus drive the need for extremely high compression ratio to make storage practical. Compression exploits the following facts.

* Imagery data has more redundancy than we can generally find in other types of data.
* The human eye is very tolerant of approximation error in an image. This tolerance has to be exploited in order to produce increased compression at the expense of image quality.

Artificial neural networks are simplified models of the biological neuron system. A neural network is a highly interconnected network with a large number of processing elements called neurons in an architecture inspired by the brain. Artificial neural networks are massively parallel adaptive networks which are intended to abstract and model some of the functionality of the human nervous system in an attempt to partially capture some of its computational strengths. A neural network can be viewed as comprising eight components which are neurons, activation state vector, signal function, pattern of connectivity, activity aggregation rule, activation rule, learning rule and environment. They are considered as the possible solutions to problems and for the applications where high computation rates are required. The BPNN has the simplest architecture of ANN that has been developed for image compression but its drawback is very slow convergence.Image processing is a very interesting and are hot areas where day-to-day improvement is quite inexplicable and has become an integral part of own lives. It is the analysis, manipulation, storage, and display of graphical images. Image processing is a module primarily used to enhance the quality and appearance of black and white images. It enhances the quality of the scanned or faxed document, by performing operations that remove imperfections. Image processing operations can be roughly divided into three major categories, image enhancement, image restoration and image compression. Image compression techniques aim to remove the redundancy present in data in a way, which makes image reconstruction possible. Image compression continues to be an important subject in many areas such as communication, data storage, computation etc.The report begins with an introduction to image compression following the need for the compression. The next section describes some of the underlying technologies for performing the image compression follows its observation and analysis. Last section is the future scope and conclusion.

2. Related works

2.1 Back Propagation Neural Network [1]

The neural network is designed with three layers, one input layer, one output layer and one hidden layer. The input layer and output layer are fully connected to the hidden layer. Compression is achieved by designing the number of neurons at the hidden layer, less than that of neurons at both input and the output layers. Image compression is achieved by training the network in such a way that the coupling weights scale the input vector of N-dimension into a narrow channel of K-dimension with K less than N, at the hidden layer and produce the optimum output value which makes the quadratic error between input and output minimum. Basic neural network used for compression is shown in Figure 1. The basic back-propagation network is further extended to construct a hierarchical neural network by adding two more hidden layers into the existing network.

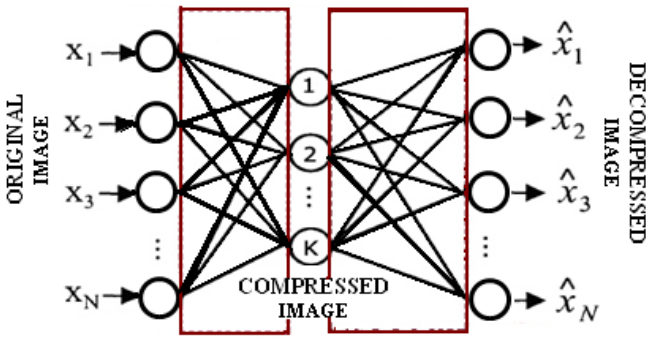

Fig 2.1:- Back Propagation Neural Network

2.2 Hierarchical and adaptive back-propagation neural network [2]

The basic back-propagation network is further extended to construct a hierarchical neural network by adding two more hidden layers into the existing network. All three hidden layers are fully connected. Nested training algorithm is proposed to reduce the overall neural network training time. The neuron weights are maintained the same throughout the image compression process. Hierarchical neural network for compression is shown in Figure 2. Adaptive schemes are based on the principle that different neural networks are used to compress image blocks with different extent of complexity. The basic idea is to classify the input image blocks into a few subsets with different features according to their complexity measurement. A fine tuned neural network then compresses each subset. Prior to training, all image blocks are classified into four classes according to their activity values which are identified as very low, low, high and very high activities. The network results in high complexity.

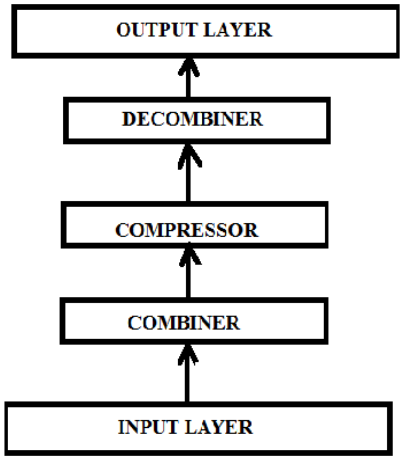

Fig 2.2:- Hierarchical Neural Network

2.3 Multi layer Feed Forward Artificial Neural Network [3], [4]

The network is designed in a way such that N will be greater than Y, where N is input layer/output layer neurons and Y is hidden layer neurons. Divide the training image into blocks. Scale each block and apply it to input layer and get the output of output layer. Adjust the weight to minimize the difference between the output and the desired output. Repeat until the error is small enough. The output of hidden layer is quantized and entropy coded to represent the compressed image. Two categories of optimization algorithms are considered i.e., derivative-based and derivative-free [5]. Derivative based methods include gradient descent, conjugate-gradient,Quasi Newton and Levenberg-Marquardt methods. Gradient descent indicates the direction to move. The conjugate-gradient method reduces oscillatory behavior and adjusts weight according to the previously successful path directions as it uses a direction vector which is a linear combination of past direction vectors and the current negative gradient vector. LM and QN algorithm-based back propagation neural networks are equally efficient. Under derivative free, two of the popular developed approaches are Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).

2.4 Multilayer Perception [6]

Basic multilayer perception (MLP) building unit is a model of artificial neuron. This unit computes the weighted sum of the inputs plus the threshold weight and passes this sum through the activation function usually sigmoid. In a multilayer perception, the outputs of the units in one layer form the inputs to the next layer. The weights of the network are usually computed by training the network using the back propagation algorithm. The basic computational unit, often referred to as a neuron, consists of a set of synaptic weights, one for every input, plus a bias weight, a summer, and a nonlinear function referred to as the activation function. Each unit computes the weighted sum of the inputs plus the bias weight and passes this sum through the activation function to calculate the output value as

$$y_j = f(\sum w_{ji} x_i + \emptyset_i) \quad (1)$$

2.5 Radial Basis Function Network [6]

Radial basis function networks are feed-forward networks. They are trained using a supervised training algorithm. They are typically configured with a single hidden layer of units whose output function is selected from a class of functions called basis functions. The input layer is made up of source nodes (sensory units) whose number is equal to the dimension N of the input vector. The second layer is the hidden layer which is composed of nonlinear units that are connected directly to all of the nodes in the input layer. Each hidden unit takes its input from all the nodes at the input layer. The hidden units contain a basis function, which has the parameters centre and width. Observation and AnalysisThe back propagation neural network is generally used as a basic network through which different variations of image compression schemes can be implemented with different error functions and using overlapped blocks, which include hierarchical and adaptive back propagation neural networks. Later came neural network based adaptive image coding which was basically developed from the mathematical iterations for obtaining the K-L transform conventionally. To improve the compression performance, multi layer feed forward network is used. It uses different optimization methods of which Quasi Newton is better but takes a long time. There are different optimization techniques which can be combined with basic networks in order to improve the compression efficiency. Survey is concluded by giving a brief idea about how the authentication and protection to be incorporated into the neural network to enhance the security.

3. Proposed System

Two different categories for improving the compression methods and their performance have been suggested. In the first case, conventional methods like SPIHT, vector quantization (VQ) etc., can be used with some enhancements. Secondly, apply neural network to develop the compression scheme, so that new methods can be developed and further research possibilities can be explored in future.In this work, image compression using multi layer neural networks has been proposed. In the proposed system, there is a testing set consists of sub images that are not included in the training set. Levenberg-Marquardt algorithm is used for training purpose. Image pixels are normalized before the compression process. If the learning factor ∞ is very large, the LM algorithm becomes the steepest decent. This parameter is automatically adjusted for all iterations in order to secure convergence. Here, a modified version of LM algorithm is proposed that provides a similar performance, while lacks the inconveniences of LM. It is more stable. The MSE between the target image and reconstructed image should be as small as possible so that the quality of reconstructed image should be near to the target image. The proposed method gives high compression ratio.

(a) One to one mapping:
For incorporating protection of the data, one to one property of the neural network can be used. If there are interactions of two parameters, resultant should be a unique value stated as:

$$\emptyset(x_i, y) \neq \emptyset(x_j, y); \forall j if j \neq i; \quad (2)$$

(b) One way property:

For authentication, the property allows to compute output from the input easily while very difficult to compute input from the output. The input P is composed of n elements $[p_1, p_2, \dots \dots, p_n]$ while the output is unique C as:

$$C = f\left(\sum_{j=1}^{n} w_j p_j + b\right) \quad (3)$$

It is easy to compute C from a given P, but difficult to compute P from C.

3.1 Neural Network Compression

The compression process is described below:-

1. Read image pixels and then normalize it to range [0-1].
2. Divide the image into non-overlapping blocks.
3. Rasterize the pixels and apply to the input layer.
4. Compute the outputs of hidden layer units by multiplying the input vector by the weight matrix (V).
5. Store the outputs in a compressed file after renormalization.
6. If there are more image vectors go to (4).
7. Stop.

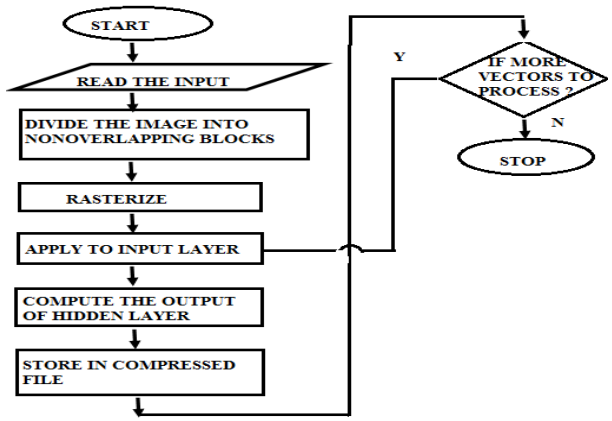

Fig 3.1 Compression

The decompression process is described below:-

1. Take one by one vector from the compressed image.
2. Normalize this vector.
3. The outputs of output layer units by multiplying outputs of hidden layer units by the weight matrix.
4. Derasterize the outputs of output layer units to build the sub image.
5. Return this sub image to its proper location.
6. Renormalize this block and store it in the reconstructed file.
7. If there are more vectors go to (1).

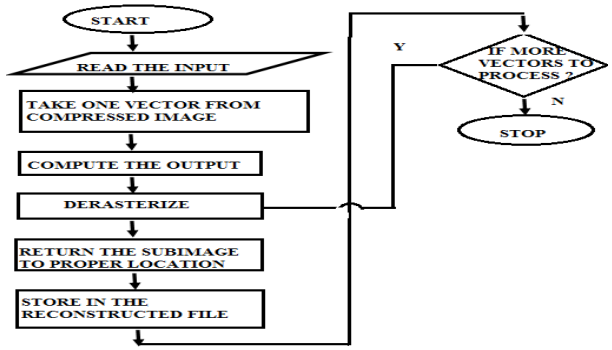

Fig 3.2 Decompression

4. Implementation

4.1 Preprocessing

The neural network requires inputs with real type and the sigmoid function of each neuron requires the input data to be in the range [0-1]. For this reason, the image data values must be normalized. The normalization is the process of linearly transformation of image values from the range [0-255] into another range that is appropriate for neural network requirements. Segmentation is the process of dividing it into non overlapping blocks with equal size to simplify the learning/compressing processes. Image rasterization is the process of converting each sub image from a two dimensional block in to a one dimensional vector, to speed up the learning.

4.2 Neural Network Design

Multilayer feedforward network is used for compressing the images. Neural network is designed in such a way that the numbers of input and output layer neurons are set to 64. Hidden layer neurons are set to 16. The two weight matrices are selected to small random numbers.

4.3 Training

The input image is split up into blocks or vectors of 4X4, 8X8 or 16X16 pixels. These vectors are used as inputs to the network. The network is provide by the expected output, and it is trained so that the coupling weights, {wij}, scale the input vector of N -dimension into a narrow channel of Y -dimension, which is less than N, at the hidden layer and produce the optimum output value which makes the quadratic error between output and the desired one minimum.

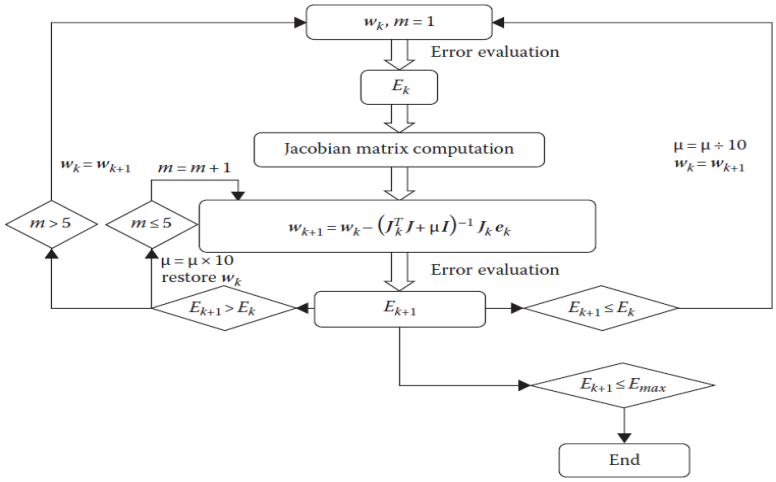

Fig 4.1 LM algorithm

The LM algorithm has got some disadvantages. If the learning factor is very large, the LM algorithm becomes the steepest decent. This parameter is automatically adjusted for all iterations in order to secure convergence. The LM algorithm computes the Jacobin J matrix at each iteration step and the inversion of square matrix. In the LM algorithm must be inverted for all iterations. Hence for large size neural networks, the LM algorithm is not practical. Here, a modified version of LM algorithm is proposed that provides a similar performance, while lacks the inconveniences of LM. A new performance index is introduced,

$$F(w) = \sum_{k=1}^{p} \left[\sum_{p=1}^{p} \left(d_{kp} - o_{kp} \right)^2 \right]^2 \quad (4)$$

where d_{kp} is the desired value of k^{th} output and o_{kp} is the actual value of k^{th} output and the p^{th} pattern is the number of the weights, P is the number of patterns, and K is the number of network outputs. This index represents a global error, will later lead to a significant reduction of the size of a matrix to be inverted at each iteration step [6]. The learning factor, α is modified as 0.01 $E^T E$, where E is a k×1 matrix. If the error is small, then actual output approaches to desired output.

The trained network is now ready to be used for image compression which, is achieved by dividing or input images into normalization and segmentation. To decompress the image; first the compressed image is renormalized then applies it to the output of the hidden layer and get the one vector of the hidden layer output is normalized then it rasterization to represent the reconstruct the image.

MSE and PSNR are the parameters which define the quality of an image reconstructed at the output layer of neural network.

a) Mean Square Error (MSE)

The MSE between the target image and reconstructed image should be as small as possible so that the quality of reconstructed image should be near to the target image. Ideally, the mean square error should be zero for ideal decompression. The compression ratio is defined by the ratio of the data fed to the input layer neurons to the data out from the hidden layer neurons. In a structure 1, 016 neurons were used in the hidden layer. So it will results in the fixed 4:1 compression ratio.

b) Peak Signal to Noise ratio (PSNR)

The term peak signal-to-noise ratio (PSNR) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation. The PSNR computes by the following equation:-

$$PSNR = 10\ log_{10}\ {255^2}/{MSE} \quad (5)$$

The compression ratio performance can be computed by,

$$CR=(1 - N_h/N_i) \times 100\% \quad (6)$$

where N_h is the input layer neurons and N_i is the hidden layer neurons.

5. Conclusion

The need for effective data compression is evident in almost all applications where storage and transmission of digital images are involved. Neural networks offer the potential for providing a novel solution to the problem of compression by its ability to generate an internal data representation.Multilayer feed forward network is used due to its efficiency. Learning algorithms has significant impact on the performance of neural networks, and the effects of this depend on the targeted application. The choice of suitable learning algorithms is therefore application dependent. The performance can be increased by modifying the training algorithm which outperforms the existing method.

Protection of image contents is equally important as compression in order to maintain the privacy. If any malicious modification occurs either in storage or in transmission channel, such modifications should be identified. So the authentication and protection can be incorporated into the proposed system in future by utilizing the other properties of the neural network.

References

[1] J. Jiang, Image compression with neural networks–a survey, Signal processing: image Communication, 1999, 737–760.

[2] M. Egmont-Petersen, D. de Ridder, and H. Handels, Image processing with neural networks–a review, Pattern recognition, vol. 35, no. 10, 2002, 2279–2301.

[3] F. Ibrahim, Image compression using multilayer feed forward artificial neural network and dct, Journal of Applied Sciences Research, vol. 6, no. 10, 2010, 1554–1560.

[4] V. Gaidhane, V. Singh, Y. Hote, and M. Kumar, New approaches for image compression using neural network, Journal of Intelligent Learning Systems and Applications, vol. 3, no. 4, 2011, 220–229.

[5] N. Relhan, M. Jain, V. Sahni, J. Kaur, and S. Sharma, Analysis of optimization techniques for feed forward neural networks based image compression, International Journal of Computer Science and Information Technologies, vol. 3, no. 2, 2012.