OPTIMAL FEED-FORWARD NEURAL NETWORK ARCHITECTURES

George Bebis and Michael Georgiopoulos
Department of Electrical & Computer Engineering
University of Central Florida, Orlando, FL 32816 USA

1. Why network size is so important ?

One of the most important problems that neural network designers face today is choosing an appropriate network size for a given application. Network size involves in the case of layered neural network architectures, the number of layers in a network, the number of nodes per layer, and the number of connections. Roughly speaking, a neural network implements a nonlinear mapping of the from $u=G(x)$. The mapping function G is established during a training phase where the network learns to correctly associate input patterns x to output patterns u. Given a set of training examples (x, u), there is probably an infinite number of different size networks that can learn to map input patterns x into output patterns u. The question is, which network size is more appropriate for a given problem ? Unfortunately, the answer to this question is not always obvious. Many researchers agree that the quality of a solution found by a neural network depends strongly on the network size used. In general, network size affects network complexity, and learning time, but most importantly, it affects the generalization capabilities of the network; that is, its ability to produce accurate results on patterns outside its training set.

There is a very illustrative analogy between neural network learning and curve fitting which highlights the importance of network size. There are two problems in curve fitting: finding out the order of the polynomial and finding out the coefficients of the polynomial. Thus, given a set of data points, first we decide on the order of the polynomial we will use and then we compute the coefficients of the polynomial in order to minimize the sum of the squared differences between required and predicted values. Once the coefficients have been computed, we can evaluate any value of the polynomial given a data point, even for data points that were not in the initial data set. If the order of the polynomial chosen is very low, the approximations obtained are not good, even for points contained in the initial data set. On the other hand, if the order of the polynomial chosen is very high, very bad values may be computed for points not included in the initial data set. Figure 1 illustrates these concepts. Similarly, a network having a structure simpler than necessary cannot give good approximations

even for patterns included in its training set and a more complicated than necessary structure, "overfits" the training data, that is, it performs nicely on patterns included in the training set but performs very poorly on unknown patterns.

(somewhere here)

Fig. 1. Good and Bad fits

2. What does the theory say ?

Generally, the number of nodes in the input and output layers can be determined by the dimensionality of the problem. However, determining the number of hidden nodes is not straightforward and requires first the determination of the number of hidden layers. There is a number of theoretical results concerning the number of hidden layers in a network. Specifically, Nielsen has shown that a network with two hidden layers can approximate any arbitrary nonlinear function and generate any complex decision region for classification problems. Later it was shown by Cybenko that a single layer is enough to form an arbitrarily close approximation to any nonlinear decision boundary. (Furthermore, it was shown that one hidden layer is enough to approximate any continuous function with arbitrary accuracy where the accuracy is determined by the number of nodes in the hidden layer; also, one hidden layer is enough to represent any Boolean function). Recently, Hornik and Stinchombe have come up with a more general theoretical result. In specific, they have shown that a single hidden layer feed-forward network with arbitrary sigmoid hidden layer activation functions can approximate arbitrarily well an arbitrary mapping from one finite dimensional space to another. This tells us that feed-forward networks can approximate virtually any function of interest to any desired degree of accuracy, provided sufficiently many hidden units are available. Although the above theoretical results are of great importance because they demonstrate the powerful capabilities of feed-forward neural networks, they don’t give us an indication of how to choose the number of hidden units needed per hidden layer. In addition, even if for some problems one hidden layer may be enough theoretically, in practice more than one hidden layers should be utilized to solve the problem faster and more efficiently. For example, we mentioned before that one hidden layer is enough to approximate any continuous function. However, in certain problems, a large number of hidden nodes may be required in order to achieve the desired accuracy. Thus, a network with two hidden layers and much fewer nodes overall, should be able to solve the same problem more efficiently. Hence, choosing an appropriate network size for a given problem is still something of an art.

3. Why small and simple networks are better ?

In determining the network size, one is only guided by his/her intuition and by some specific knowledge about the problem. For example, when the input to the network is an image, it is more reasonable to define local receptive fields (that is, to use local connections), instead of full connectivity because nearby pixels in the image are probably more correlated than pixels located far away from each other. Unfortunately, when no a-priori knowledge about the problem is available, one has to determine the network size by trial and error. Usually, one has to train different size networks and if they don't yield an acceptable solution, then they are discarded. This procedure is repeated until an appropriate network is found. Experience has shown that using the smallest network which can learn the task, is better for both practical and theoretical reasons. Smaller networks require less memory to store the connection weights and can be implemented in hardware easier and more economically. Training a smaller network usually requires less computations because each iteration is less computationally expensive. Smaller networks have also very short propagation delays from their inputs to their outputs. This is very important during the testing phase of the network, where fast responses are usually required. Bigger networks generally need larger number of training examples for achieving good generalization performance since it has been shown that the required number of training examples grows almost linearly with the number of hidden units. However, in many practical cases we have only a limited number of training data and this may lead to a very poor generalization. Finally, although bigger networks can perform more complicated mappings, when trained with limited training data they exhibit poor generalization.

In order to illustrate the generalization capabilities of larger than the necessary networks, Reed performed the following experiment: he trained two different size networks in order to classify linearly separable data in a two dimensional plane, with some overlap close to the boundary. Both of the networks he used were three layer networks (two hidden layers and one output layer). The big network had 2 inputs, 50 nodes in the first hidden layer, 10 nodes in the second layer and a single node in the output layer. The small network had the same number of input-output nodes and 2 nodes in each one of the hidden layers. Then, he plotted the decision boundary found by each one of the networks. The results showed that although the classes were linearly separable, the decision boundary formed by the big network was highly nonlinear and it was very unlikely for this network to exhibit a good generalization performance. On the other hand, the small network found a much better solution by forming an almost linear decision boundary.

4. Modifying the network architecture

It should be more clear at this point that solving a given problem using the minimum possible network, poses a lot of advantages. However, choosing a smaller network over a larger one means that we actually restrict the number of free parameters in the network. Consequently, the error surface of a smaller network is more complicated and includes more local minima compared with the error surface of a larger network. It is well known that the existence of local minima in the error surface can seriously prevent a network from reaching a good solution. Thus, although smaller networks can prove to be very beneficial in terms of generalization, their training may require a lot of effort. Recently, Judd has proven that even if we were able to determine the optimal network size a-priori, the loading problem, that is, the problem of finding a set of weights for which that network performs the desired mapping, is NP-complete. Blum and Rivest proved later that the loading problem remains NP-complete even for nets containing only three nodes. Based on these theoretical results, Blum claimed that it is unlikely that any algorithm which simply varies weights on a network of fixed size can learn in polynomial time. During the last years, a number of techniques have emerged which attempt to improve the generalization capabilities of a network by modifying not only the connection weights but also the architecture as training proceeds. These techniques can be divided into two categories. The first category includes methods that start with a big network and gradually eliminate the unnecessary nodes or connections. These methods are called *pruning* methods. The second class includes methods that start with a small network and gradually add nodes or connections when it is necessary. These methods are called *constructive* methods. Figure 2, illustrates the classification of the methods.

(somewhere here)

Fig. 2. Algorithms for optimal network architectures

4.1 Pruning methods

These methods attempt to find a quick solution to the problem by starting with a large network and then reducing it to a smaller network in order to improve generalization. Considering the curve fitting problem, the above approach implies that we start with a high order polynomial and gradually eliminate the higher order terms which do not contribute significantly to the fit. There are two main subcategories of pruning methods: (i) pruning based on modifying the error function and (ii) pruning based on sensitivity measures.

4.1.1 Pruning by modifying the error function

The basic idea of methods falling in this subcategory is to modify the error function of the network in a way such that the unnecessary connections of the network will have zero weight (or near zero) after training. Then, these connections can be removed without degrading the performance of the network. These approaches, which are also called weight decay approaches, actually encourage the learning algorithm to find solutions which use as few weights as possible. The simplest modified error function can be formed by adding to the original error function a term proportional to the sum of squares of weights:

$$E = E_o + \gamma \sum_i \sum_j w_{ij}^2$$

where E_o is the original error function (sum of the squared differences between actual and desired output values), γ is a small positive constant which is used to control the contribution of the second term, and w_{ij} is the weight of the connection between node j of a layer and node i of the immediately higher indexed layer. The above error function penalizes the use of more w_{ij}'s than necessary. In order to show that, lets see how the weight updating rule is changed. Assuming that we apply the gradient descent procedure to minimize the error, the modified weight update rule is given by:

$$\Delta w_{ij}(t) = -\alpha\left(\frac{\partial E}{\partial w_{ij}}\right)(t) = -\alpha\left(\frac{\partial E_o}{\partial w_{ij}}\right)(t) - 2\gamma\alpha w_{ij}(t)$$

where t denotes the t-th iteration and α denotes the learning rate. The above expression can be written as:

$$w_{ij}(t+1) = -\alpha\left(\frac{\partial E_o}{\partial w_{ij}}\right)(t) + (1-2\gamma\alpha) w_{ij}(t)$$

It can be shown that the magnitude of the weights decreases exponentially towards zero by computing the weight values after t weight adaptations:

$$w_{ij}(t) = \alpha \sum_{i=1}^{t} (1-2\gamma\alpha)^{t-i} \left(-\frac{\partial E_o}{\partial w_{ij}}\right) + (1-2\gamma\alpha)^t w_{ij}(0)$$

(assuming $|1-2\gamma\alpha| < 1$). The above approach has the disadvantage that all the weights of the network decrease at the same rate. However, it is more desirable to allow large weights to persist while small weights tend toward zero. This can be achieved by modifying the error function in a way that small weights are affected more significantly than large weights. This can be done for example by choosing the following modified function:

$$E = E_o + \gamma \sum_i \sum_j \frac{w_{ij}^2}{1 + w_{ij}^2}$$

The weight updating rule then becomes:

$$w_{ij}(t+1) = -\alpha\left(\frac{\partial E_o}{\partial w_{ij}}\right)(t) + \left(1 - \frac{2\gamma\alpha}{(1 + w_{ij}^2(t))^2}\right) w_{ij}(t)$$

It can be shown that in this case small weights decrease more rapidly than large ones.

4.1.2 Sensitivity based methods

The general idea of these methods is to train a network in performing a given task and then to compute how important is the existence of a connection or node. Then, the least important connections or nodes are removed and the remaining network is retrained. In general, the sensitivity measurement does not interfere with training and requires an extra amount of computational effort. The key issue in the implementation of these techniques is finding a way to measure how sensitive is the solution to the removal of a connection or a node. Early approaches attempt to remove a connection by evaluating the change in the network's output error. If the error increases too much, then the weight must be restored back again. More sophisticated approaches evaluate the change in error for all the connections and training data and then remove the one connection which produces the least error increment. Obviously, both of the above approaches are extremely time consuming, especially when large networks are considered. A more heuristic approach is the "skeletonization" procedure proposed by Mozer and Smolensky. In their approach, the relevance of a connection is computed again by measuring the error increase when the connection is removed. However, the relevance of a connection is not computed according to the approaches described above but using information about the shape of the error surface near the minimum to which the network has currently settled down. This is performed using the partial derivative of the error with respect to the connection to be removed. Connections with relevance below a

certain threshold are then removed. "Optimal brain damage" is another approach proposed by Le Cun and his co-workers. In this approach, the saliency of connections is measured using the second derivative of the error with respect to the connection. In particular, the saliency s_{ij} of a connection w_{ij} is given by

$$s_{ij} = \left(\frac{\partial^2 E}{\partial w_{ij}^2}\right)\frac{w_{ij}^2}{2}$$

where the second derivative measures the sensitivity of the error to small perturbations in w_{ij}. In this way, connections with small weight values having a significant influence on the solution are not removed.

4.2 Constructive methods

These methods build the structure of a network by starting with a minimal network and by adding new nodes during training. Small networks get easily trapped to local minima, so new hidden nodes are added in order to change the shape of the weight space and to escape the local minima. Considering the curve fitting problem again, the constructive approach implies that we start with a very low order polynomial and we add higher order terms every time that the current polynomial cannot give a good fit. There are a lot of interesting algorithms falling in this category and various heuristics are employed during the network growth process. In order to give an idea how these algorithms operate, we have decided to focus on two of them: the *Upstart* algorithm which appears to be very successful for binary mappings and the *Cascade Correlation* algorithm which appears to be very successful for real valued mappings. Both of the algorithms build a tree-like network by dividing the input space successively.

The *Upstart* algorithm builds a tree-like network in a top-down fashion. The nodes of the network are linear threshold units (the output of a node is either 0 or 1). The number of input-output nodes is determined by the nature of the problem. In the following description we assume that the network consists of a single output node, that is, the network can assign an input pattern into two possible classes (one is represented by 0 and the other by 1). The steps of this algorithm can be summarized as follows:

Step 1. Initially, a single node is assumed which is connected to each input of the network. This node is trained to learn as many associations as possible.

Step 2. If, that node creates wrong classifications, two "child" nodes are created to correct the erroneous "0" and "1" classifications of their 'parents'.

Step 3. The weights from the inputs to the parent node are frozen and the child nodes are connected to the inputs of the network. Each child node is trained to correct the erroneous "0" classifications and the erroneous "1" classifications.

Step 4. The child node which corrects erroneous "0" is connected to its parent node with a large positive weight, while the child node which corrects the wrongly 1's cases is connected to its parent node with a large negative weight.

Step 5. Two nodes are added for each child node in order to correct their wrong classifications. The old child nodes are treated as parent nodes now and the added nodes as new child nodes. Training continues until all the data are classified correctly.

The *Upstart* algorithm is guaranteed to converge because each subnode is guaranteed to classify at least one of its targets correctly. This is true because for binary patterns, it is always possible to cut of a corner of the binary hypercube with a plane. The number of nodes grows linearly with the number of patterns and the resulting hierarchical architecture can be converted into an equivalent two layer network. It should be mentioned that the algorithm can easily be extended to a classifier with more than one classes.

The *Cascade Correlation* algorithm builds also a tree-like network but in a bottom-up fashion. The number of input-output nodes is determined a-priori based on the characteristics of the problem. The hidden units are added to the network one at a time and are connected in a cascaded way. The activation functions for the nodes may be sigmoidal functions or any mixture of non-linear activation functions. The main steps of the algorithm are the following:

Step 1. Connect each input node to each output node and train the network over the entire training set in order to learn as many associations as possible.

Step 2. When no significant error reduction has occurred after a certain number of training iterations, run the network one last time over the entire training set to measure the error.

Step 3. If the error is less than a threshold then stop, otherwise add a new hidden node (candidate), and connect it with every input node and every preexisting hidden node. Don't connect it to the output nodes yet.

Step 4. Freeze all the weights of the network and adjust only the new hidden unit's input weights by trying to maximize the magnitude of the correlation between the new unit's output and the network's output error.

Step 5. If the new hidden unit stops improving (i.e., the error doesn't decrease), freeze its input weights and connect it to the output nodes.

Step 6. Train the network adjusting only the connections from the new hidden unit to the output nodes. Then, go to step 3.

Step 4 of the algorithm actually maximizes the magnitude of the correlation between the candidate node's output and the network's error. Specifically, the function to be maximized has the form

$$S = \sum_{o,p} |(V_p - \bar{V})(E_{p,o} - \bar{E}_o)|$$

where V_p is the candidate node output when the p-th training pattern is presented to the network and $E_{p,o}$ is the output error of the o output node when the p-th training pattern is presented to the network. Furthermore, $\bar{V}$ and $\bar{E}_o$ are the averages of V_p and $E_{o,p}$ over all the training patterns. Each new node added to the network learns actually a mapping which has the best possible correlation with the errors of the previous network. The way that the hidden to output weights are modified is the following: if a hidden unit correlates positively with the error at a given output node, it will develop a negative connection weight to that node, attempting to cancel some of the error. Otherwise, it will develop a positive connection weight. The main advantages of the *Cascade Correlation* algorithm are: (i) it learns fast, (ii) it builds reasonably small networks, and (iii) it requires no back-propagation of error signals.

4.3 Just a few more words ...

So far we have only discussed certain pruning and constructive approaches without comparing them with each other. It is really difficult to say which of the two approaches performs better. Pruning has the disadvantage that since we don't really have any knowledge about the size of the proper

network, larger than the required size networks are often chosen as starting points. Since a lot of the training time is spent before pruning really starts, this may be computationally wasteful. In addition, since there are a lot of medium-size networks that can learn the same problem, the pruning procedure may not be able to find a small-size network because it may get stuck with one of these medium-size networks. Constructive approaches on the other hand have the disadvantage that they usually result with networks having long propagation delays from network inputs to network outputs. In addition, new nodes are usually assigned random weights which are likely to disrupt the approximate solution already found. A probably superior approach would be a combination of constructive approaches with pruning. For example, it has been suggested by the authors of the *Cascade Correlation* algorithm that a possible way to keep the depth of the network small and to minimize the number of connections to the hidden and output nodes, is to use a weight decay approach by adding a penalty term to the error function. A general procedure for coupling constructive and pruning approaches would be the following: allow a small network to grow enough during training until a reasonable solution is found and then prune the network in order to achieve a smaller and faster network which provides the desired solution more efficiently and accurately.

Our discussion would have been incomplete if we did not mention two recently emerged approaches which deal with the problem of network size: weight sharing and Genetic algorithms.

Weight sharing tries to reduce the number of weights in a network by first assigning a local receptive field to each hidden node. Then, the weights of hidden nodes, having receptive fields at different locations of their inputs, are given the same values. Thus, hidden nodes that have receptive fields with common weights try actually to detect the same kind of features but at different locations of their input. Weight sharing has been applied by Le Cun and his co-workers on a handwritten digit recognition task, illustrating very good performance.

Genetic algorithms are a class of optimization procedures inspired by the biological mechanisms of reproduction. A genetic algorithm operates iteratively on a population of structures, each one of which represents a candidate solution to the problem that the algorithm tries to solve. On each iteration, a new population is produced by first applying on the old population three fundamental operations: reproduction, crossover, and mutation. Then, each member of the population is evaluated through a fitness function and members assigned a bad evaluation are discarded, while members assigned a good evaluation survive in future populations. The key issue in the determination of a neural network architecture using a genetic algorithm is how an architecture should be translated into a proper representation to be utilized by the genetic algorithm, and how much information about an architecture

should be encoded into this representation. For example, Miller, Todd, and Hyde represent the network architecture as a connection matrix, mapped directly into a bit-string. Then, a number of different size networks are encoded in this way in order to form the initial population. The genetic operators act on this population and new populations are formed. A decoding procedure is applied on each member of the population in order to transform a bit-string into a legitimate network architecture. The fitness of each network is evaluated by training it for a certain number of epochs and recording the network's error. There is some preliminary success of Genetic algorithms associated with the problem of optimizing the network size but there is still a lot to be accomplished.

Read more about this

- J. Hertz, A. Krogh, and R. Palmer, *Introduction to the theory of Neural Computation*, Addison Wesley, 1991.
- D. Hush and B. Horne, "Progress in supervised Neural Networks", *IEEE Signal Processing Magazine*, Jan. 1993, pp. 8-39.
- R. Reed, "Pruning algorithms - a survey", *IEEE Transactions on Neural Networks*, vol. 4, no. 5, Sept. 1993, pp. 740-747.
- M. Wynne-Jones, "Constructive algorithms and pruning: improving the multilayer perceptron", in *Proc. of the 13th IMACS World Congress on Computation and Applied Mathematics*, R. Vichnevetsky and J. Miller, Eds., 1991, pp. 747-750.
- R. Hetcht-Nielsen, "Theory of the backpropagation neural networks", *Proc. Inter. Joint Conf. Neural Networks*, vol. I, June 1989, pp. 593-611.
- G. Cybenko, "Approximation by superpositions of a sigmoidal function", *Mathematics of Control, Signals, and Systems*, vol. 2, 1989, pp. 303-314.
- K. Hornik and M. Stinchombe, "Multilayer feed-forward networks are universal approximators", in f2Artificial Neural Networks: Approximation and Learning Theory, H. White et. al., Eds., Blackwell press, Oxford, 1992.
- A. Blum and R. Rivest, "Training a 3-node neural network is NP-complete", *Proc. of the 1988 Workshop on Computational Learning*, 1988, pp. 9-18.
- M. Frean, "The Upstart algorithm: a method for constructing and training feed-forward networks", *Neural Computation*, vol. 2, 1990, pp. 198-209.

- S. Fahlman and C. Lebiere, "The Cascade-Correlation learning architecture", in *Advances in Neural Information Processing Systems II*, ed. D. Touretzky, 1990, pp. 524-532.
- Y. Le Cun, B. Boser, J. Denker, D. Henderson. R. Howard. W. Hubbard, and L. Jackel, "Back-propagation applied to handwritten zip code recognition", *Neural Computation*, vol. 1, pp. 541-551, 1989.
- G. Miller, P. Todd and S. Hegde, "Designing Neural Networks using Genetic Algorithms", *Third International Conference on Genetic Algorithms*, pp. 379-384, 1989.