Controlled Growth of Cascade Correlation Nets*†

Lars Kai Hansen and Morten With Pedersen
CONNECT Electronics Institute B349
Technical University of Denmark, DK-2800 Lyngby, DENMARK

1 Introduction

The optimal selection of neural network architecture is a problem of significant theoretical and practical importance. The functional capacity of the network architecture determines the generalization properties and the computational complexity of the algorithm. For supervised training of feed-forward networks, current generalization theories tell us to minimize the functional capacity, *i.e.*, the number of weights subject to the constraints imposed by the rule implicit in the training examples see *e.g.*, (Hertz et al., 91). Several schemes have been proposed for *pruning* of the network architecture. The *optimal brain damage* method for network pruning proposed in (LeCun et al., 90) has been successful for pruning of classifiers as well as for pruning of regression networks for time series processing (Svarer et al., 93). The *optimal brain surgeon* (OBS) advanced in (Hassibi and Stork, 92) provides a refined estimate of the weight *saliency* by incorporating the effects of quadratic retraining; however, at the expense of significant additional computation. Likewise, growth algorithms are *legio*. Among these, Fahlman's *cascade correlation* stands out for solving hard classification problems (Fahlman and Lebiere, 89). In the standard implementation cascade correlation adds hidden units that are fully connected to all input units and to all previously added hidden units, thereby rapidly increasing the number of weights in the network with the obvious danger of overfitting.

In this communication we propose a controlled growth version of cascade correlation in which each hidden unit is pruned before it is inserted in the network. Since the cascade correlation scheme adds hiddens in the form of simple perceptrons the OBS scheme is feasible. For general networks the surgeon is problematic because it involves the full Hessian of the network costfunction. As the cascaded hidden units of the cascade correlation algorithm are trained according to a correlation measure rather than the usual least squares error function, we are forced to modify the surgeon in two directions: We give the modifications needed for the optimal brain surgeon operating with regularized costfunctions *e.g. weight decay*. Furthermore, we present the OBS saliencies for the cascade correlation measure. The weight decay regularization term plays a dual role in this work. First, it bias the learning process in the direction of "simple explanations"; secondly, it prevents the weights of the cascaded hidden units from diverging to infinite values that would otherwise maximize the correlation measure.

*This work is supported by the Danish Natural Science and Technical Sciences Research Councils. We thank Jan Larsen for valuable comments.
†Submitted for ICANN'94 Sorrento, Italy.

To illustrate the operation of the controlled growth scheme we apply it to a hard discrimination problem the so-called "Wieland spirals". In particular we show that we can grow cascade correlation nets with about two-thirds of the weights that we obtain from cascade correlation without pruning.

2 Optimal brain surgeon with weight decay

Let the training cost function (*e.g.* the mean squared error on the training set) be denoted $E(\mathbf{w})$. The augmented cost function for regularized learning is

$$C(\mathbf{w}) = E(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{D}\mathbf{w}, \tag{1}$$

where $\mathbf{D}$ is the regularization matrix. For a simple weight decay $\mathbf{D}$ is proportional to the unit matrix: $\mathbf{D}_{jj'} = \alpha\delta_{jj'}$. The OBS saliency of a weight is defined as the change in *training error* as the weight is eliminated and the remaining weights retrained to the new minimum of (1). The saliency is estimated through the following steps (Hassibi and Stork, 92)

- Expand the cost function to second order around an extremum.
- Find the change of the remaining weights resulting from elimination of the j'th weight, $\mathbf{w}_j$, and retraining of the remaining weights in the quadratic approximation.
- Compute the associated change in training error.

Let the extremum of the unconstrained cost be denoted $\mathbf{w}_0$. The expansion of the cost function around this point reads

$$C(\mathbf{w}) = C(\mathbf{w}_0) + \frac{1}{2}\delta\mathbf{w}^{\mathrm{T}}(\mathbf{A} + \mathbf{D})\delta\mathbf{w}, \tag{2}$$

where $\mathbf{w} = \mathbf{w}_0 + \delta\mathbf{w}$. The first order term vanish by the assumed optimality of $\mathbf{w}_0$. $\mathbf{A}$ is the second derivative matrix of the training error; the *Hessian* of the cost function is $\mathbf{H} = \mathbf{A} + \mathbf{D}$. Secondly, we derive the constrained extremum. Elimination of the j'the weight can be expressed as $\delta\mathbf{w}^{\mathrm{T}}\mathbf{e}_j = -\mathbf{w}_0^{\mathrm{T}}\mathbf{e}_j$, where $\mathbf{e}_j$ is the j'th unit vector. The constrained extremum is found using a Lagrange multiplier:

$$\tilde{C}_j(\mathbf{w}) = C(\mathbf{w}) + \gamma(\delta\mathbf{w} + \mathbf{w}_0)^{\mathrm{T}}\mathbf{e}_j \tag{3}$$

the extremum is given by $\delta\mathbf{w} = -\gamma\mathbf{H}^{-1}\mathbf{e}_j$, with

$$\gamma = \frac{\mathbf{w}_0^{\mathrm{T}}\mathbf{e}_j}{\mathbf{e}_j^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{e}_j} \tag{4}$$

We are now ready to compute the saliency of the j'th weight

$$\delta E_j(\mathbf{w}) = \gamma\mathbf{w}_0^{\mathrm{T}}\mathbf{D}\mathbf{H}^{-1}\mathbf{e}_j + \frac{1}{2}\gamma^2\mathbf{e}_j^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{A}\mathbf{H}^{-1}\mathbf{e}_j\,. \tag{5}$$

Note that this expression reduces to the one derived by (Hassibi and Stork, 92), if $\mathbf{D} = \mathbf{0}$. For further generalizations of OBS see also (Larsen, 93).

3 Cascade correlation

The basic idea of the cascade correlation algorithm is iteratively to reduce the error made by the output unit by inserting hidden units that correlate (or anti-correlate) well with the error signal. By freezing the network while optimizing the new hidden unit candidate the algorithm avoids the moving targets problem of standard backprop (Fahlman and Lebiere, 89). The current error (difference between network output and teacher signal) of example k is called E^k and the output of the candidate unit called V^k. Fahlman used the correlation measure: $S = \left|\sum_k (V^k - \overline{V})(E^k - \overline{E})\right|$ to quantify the quality of a hidden unit candidate computing: $V^k = \tanh\left(\sum_{j=1}^{n_I+n_H} w_j x_j^k\right)$ and n_I, n_H being the number of input units and previously inserted hidden units respectively. Further $\overline{V}, \overline{E}$ are the training set averages of V^k and E^k. The Hessian of the correlation measure S is approximated by

$$\frac{\partial^2 S}{\partial \mathbf{w}_j \partial \mathbf{w}_{j'}} = \mathrm{sign_S} \sum_k (E^k - \overline{E}) \frac{\partial^2 V^k}{\partial \mathbf{w}_j \partial \mathbf{w}_{j'}}, \tag{6}$$

with $\mathrm{sign_S} = \mathrm{sign}\left(\sum_k (V^k - \overline{V})(E_k - \overline{E})\right)$. This matrix plays the role of the second derivative of the training error ($\mathbf{A}$) in the OBS scheme, except for the trivial change of objective: here we *maximize* the magnitude of the correlation. The corresponding "correlation saliencies" read ($\mathbf{H} = \mathbf{A} - \alpha\mathbf{1}$):

$$\delta S_j(\mathbf{w}) = \frac{1}{2}\frac{(\mathbf{w}_0^{\mathrm{T}}\mathbf{e}_j)^2}{\mathbf{e}_j^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{e}_j} - \gamma\left(\frac{\mathbf{w}_0^{\mathrm{T}}\mathbf{e}_j\,(\mathbf{w}_0^T\mathbf{H}^{-1}\mathbf{e}_j)}{\mathbf{e}_j^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{e}_j} - \frac{1}{2}\frac{\mathbf{e}_j^{\mathrm{T}}\mathbf{H}^{-2}\mathbf{e}_j}{(\mathbf{e}_j^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{e}_j)^2}\right). \tag{7}$$

4 Experiments and concluding remarks

The generic benchmark for cascade correlation networks is the spirals classification problem involving classification of 192 (x, y) points forming two intertwined spirals (Fahlman and Lebiere, 89). On this problem the algorithm outperforms standard algorithms like feed-forward nets with backprop. Our primary interest here is to show that the controlled growth algorithm can reduce the number of weights needed to implement the spirals. It turns out that the performance of cascade correlation is quite sensitive to the search algorithms used for optimization of the hidden units. We use a dual scheme: Initially we use a simple standard backprop optimizer for the hidden weights, when improvement stops we refine the solution by a few second order Newton steps. Our results are presented in Figure 1 accumulating 20 runs of standard cascade correlation and 20 runs of the controlled growth algorithm. We trained a pool of 10 candidate units at each level in the cascade; the candidate with optimal correlation was pruned by OBS. The weight decay parameter was set to: $\alpha = 10^{-4}$. Pruning was terminated when the magnitude of the correlation had dropped by 5%. For the fully connected cascades all twenty runs succeeded with a number of hidden units ranging from 11 to 17. Likewise, all pruning runs succeeded with a number of hiddens in the range 10-16. The average number of weights used by the fully connected nets is 146, while the average number of weights in the pruned nets is 103.

In conclusion we have derived OBS saliency expressions for weight-decay regularized training and for training with Fahlman's correlation measure. The proposed controlled growth version of cascade correlation produces nets with approximately two-thirds of the number of weights used by fully connected nets.

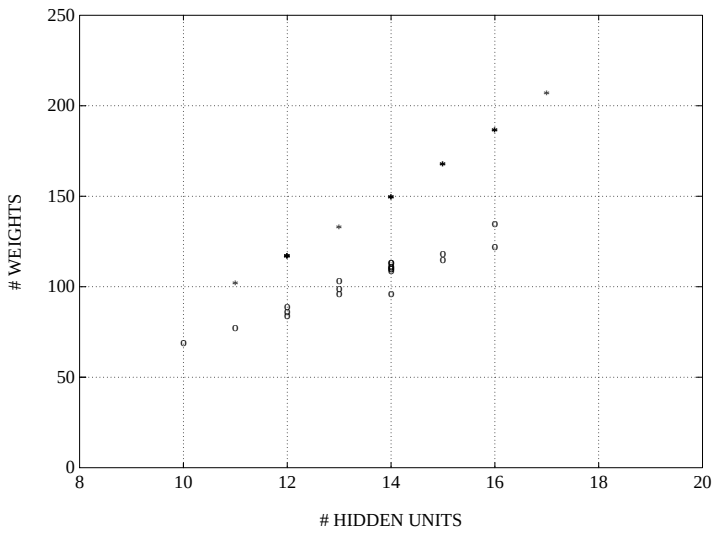

Figure 1: Number of weights versus number of cascaded hidden units for fully connected (∗) and OBS-pruned cascade correlation networks(○). Twenty runs of both algorithms resulted in networks with varying number of hidden units.

References

Fahlman, S.E. and Lebiere, C.: *The Cascade-Correlation Learning Architecture.* In NIPS 2, Ed. D.S. Touretzsky, 524-532. San Mateo, Morgan Kaufmann (1990).

Hassibi, B. and Stork, D.G.: *Second Order Derivatives for Network Pruning: Optimal Brain Surgeon.* In NIPS 5, Eds. S.J. Hanson, et al., 164. San Mateo, Morgan Kaufmann (1993).

Hertz, J., Krogh, A., and Palmer, R.G.: *Introduction to the Theory of Neural Computation.* Addison Wesley, New York (1991).

Larsen, J.: *Design of Neural Networks Filters.* PhD thesis, Electronics Institute, Technical University of Denmark (1993).

Le Cun, Y., Denker, J.S., and Solla, S.A.: *Optimal Brain Damage.* In NIPS 2, Ed. D.S. Touretzsky, 598-605. San Mateo, Morgan Kaufmann (1990)

Svarer, C., Hansen, L.K., and Larsen, J.: *On Design and Evaluation of Tapped-Delay Neural Network Architectures.* The 1993 IEEE Int. Conference on Neural Networks, San Francisco. Eds. H.R. Berenji et al., 45-51, (1993).