Statistical Analysis of Information Content for Training Pattern Recognition Networks

C L. Wilson, National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

Statistical models of neural networks predict that the difference in training and testing error will be linear in network complexity and quadratic in the feature noise of the training set. Models of this kind have been applied to the Boltzmann pruning of a large MLP (3786 weights) trained on 10,000 and tested on 10,000 Karhunen-Loève (K-L) features sets derived from images of handprinted characters and to a fingerprint classification problem which has 17,157 weights and is trained and tested on 2,000 K-L feature sets. Using the information content to optimize network size,the pruned networks have achieved high rates of recognition and at the same time been reduced in size by up to 90%. In this pruning process the product of the network capacity and the recognition error can be used effectively to select an optimum pruned network. If, in addition to conventional Boltzmann weight reduction, a weight reduction method which takes the variance content of the K-L by weighing the features using the K-L eigenvalues is used, networks with optimal size and information content can be constructed.

1 Introduction

The focus of most neural network applications has been on error minimization. A standard method of error minimization for real world problems is backpropagation [1] although more powerful methods of optimization have also been used [2],[3]. In addition to the problem of error reduction, effective generalization also requires that the information content of the network be reduced to some minimum value [4], [5],[6]. The resulting reduced network has the advantage of increased speed achieved by using fewer connections and is more effective in terms of the use of information capacity to achieve a specified pattern recognition accuracy.

The optimization strategy used in this research focuses on information content and the efficiency of information transferred to the network from the training set. This results in a smaller network with a very high information content that allows the use of a reasonably small training set. We have used the Boltzmann method as a secondary method of optimization to prune the networks used here [4], [5]. The method can be used in conjunction with a primary method of optimization such as Scaled Conjugate Gradient scheme [3]. The resulting optimized Multi-Layer Perceptron (MLP) network has been used for both fingerprint pattern level classification (PCA) and handwritten character recognition (OCR). Each recognition problem is briefly described in section 2. The method used in the statistical characterization

is discussed in section 3. The results of this size optimization for Boltzmann pruned networks is discussed in section 4. The results of the Eigen weighted pruning are discussed in section 5.

2 Pattern Recognition Experiments

The smaller of the two systems is a MLP based character recognition system. When isolated characters are used, recognition rates of 10,000 characters/second have been achieved and recognition accuracy of 98.9% with 10% rejection has also been achieved using a massively parallel computer. This speed contrasts sharply with the integrated system speed of 30 seconds/page, recognizing 130 characters/page, or 4.33 characters/second for total systems recognition time. The recognition time, using neural networks, is 0.34% of the total time in this system. At the same time, the module which loads image data into the parallel processor uses 30% of the time and segmentation uses 58% of the total time. Details of this system are given in [7].

The larger, more complex system, is a system for pattern classification of fingerprints. The system uses ridge-valley direction to convert the fingerprint image into alignment and classification features, alignment of fingerprint cores from ridge-valley directions as the image alignment method, K-L transforms of ridge-valley directions as a feature extraction method and a MLP as a classification method. The ridge-valley direction detection takes 0.4 s/image, the alignment 0.1 s/image, the K-L transform 20 ms/image, and the classification 1ms/image on a massively parallel computer. A classification accuracy of 93% is achieved with 10% rejects. The image processing prior to classification takes more than 99% of total processing time; the classification time is 0.03% of the total system time. Details of this system are given in [8].

The K-L method is used [9] for feature extraction. This method is a self-organizing method [10] that maximizes the variance in a feature set by using the principal eigenfunctions of the covariance matrix of the feature set. In the fingerprint system, local ridge directions are extracted from the image and used in subsequent processing. In the character recognition system, character images are used directly as input to the K-L transform. A similar technique has also been used with wavelets for face recognition [11] and for Kanji character recognition [12]. For characters, the 1024 bit image is converted to 48 features. For finger prints, 640 ridge valley direction components are converted to 128 features.

In this work K-L features were used to train MLP's using different methods of statistical size reduction. Only the training and recognition parts of the system were involved in the test. For the OCR problem 10,000 K-L feature from characters taken from NIST special data base 3 [13] were used. For the PCA problem 2,000 K-L features taken from fingerprints from NIST special data base 4 [14] were used.

3 Statistical Pruning of Networks

The SCG method is used as a starting network for the Boltzmann weight pruning algorithm. For the OCR problem the network has an input layer with 48 nodes, a hidden layer with 64 nodes, and an output layer with 10 nodes. For the fingerprint problem, the network has an input layer with 128 input nodes, a hidden layer with 128 nodes and an output layer with five nodes. In both cases the initial network is a fully connected network. The pruning using the

Boltzmann method was carried out by selecting a normalized temperature, T, and removing weights based on a probability of removal:

$$P_i = \exp(-w_i^2/T).$$

The values of P_i are compared to a set of uniformly distributed random numbers, R_i, on the interval $[0, 1]$. If the probability P_i is greater than R_i then the weight is set to zero. The process is carried out for each iteration of the SCG optimization process and is dynamic. If a weight is removed it may subsequently be restored by the SCG algorithm; the restored weight may survive if it has sufficient magnitude in subsequent iterations.

This method can be modified to include information about the strength of the input features so that:

$$P_i = \exp(-\lambda_j w_i^2/T),$$

where λ_j is the eigenvalue associated with the jth K-L feature for weights connected to these features in the input layer and $\lambda_j = 1$ for weights connecting the hidden and output layers. This method of pruning is referred to as eigenvalue-weighted pruning.

During this optimization process two important measures of information content are calculated [15]. The information capacity of the network, C, is given by:

$$C = N_{wts}((\log_2(|w_{max}| - \log_2(|w_{min}|) + 1)$$

where N_{wts} is the number of non-zero weights, w_{max} is the weight with the largest magnitude, and w_{min} is the weight with the smallest magnitude. The entropy is given by:

$$H = C - (\sum_{i=1}^{N_{wts}} \log_2 |w_i| + N_{wts}(1 - \log_2(w_{min})))$$

The effect on the information content of the network can be evaluated by examining the distribution of weights in the network as a function of temperature or by evaluation of the information capacity of the network.

4 Statistical Characterization of Network Pruning

The results of using Boltzmann and eigenvalue weighted pruning during the training of a network for the solution of the OCR problem is shown in table 1. The results of using Boltzmann and eigenvalue weighted pruning during the training of a network for the solution of the PCA problem is shown in table 2. The statistical evaluation of each network were carried out using the equations provided in the previous section. Examination of these results shows two distinct results. The OCR problem is easier to solve than the PCA problem and the efficiency of information transfer in both cases is improved by the eigenvalue weighting of the pruning.

In every statistical measure of network capacity and accuracy the OCR network pruned with the eigenvalue weighted pruning function is superior to the Boltzmann pruned network. Recognition accuracy is higher at all temperatures for both testing and training as shown in figures 1 and 2. At the two critical temperatures accuracy is 93.5% for the Boltzmann case and 93.6% for the eigenvalue case. The number of weights used is 1186 in the Boltzmann case and 1065 for the eigenvalue weighted case. The capacity-error product is lower in the

eigenvalue case and the bits per weight are higher. This indicated that the information transfer during training is more efficient for eigenvalue weighted training. The reduction in network capacity and entropy are much more gradual in the eigenvalue weighted case as can be seem by comparing figures 3 and 4. This capacity reduction results in a clear minimum in the capacity-error product as shown in figure 5 for Boltzmann pruning but results in a gradual reduction in the capacity-error product for the eigenweighted case as shown in figure 6. This makes selection of the critical temperature by minimization of the capacity-error product more difficult but much less critical for the eigenvalue weighted case.

Parameter	Boltzmann	Eigen
T_c	0.07	0.77
Weights	1186	1065
Max. Weights	3786	3786
Capacity(bits)	11146	10281
Max. Capacity(bits)	41646	41646
Accuracy(%)	93.5	93.6
Max. Accuracy(%)	94.8	94.8
Minimum Error×Capacity	658	560
Bits per weight	8.00	9.79

Table 1: Parameters of the pruned network for the OCR problem using Boltzmann pruning and Eigenvalue weighted Boltzmann pruning.

Parameter	Boltzmann	Eigen
T_c	0.404	0.737
Weights	667	1046
Max. Weights	17157	17157
Capacity(bits)	4120	6632
Max. Capacity(bits)	171570	171570
Accuracy(%)	71.8	78.1
Max. Accuracy(%)	84.3	84.3
Minimum Error×Capacity	1177	1447
Bits per weight	6.34	7.14

Table 2: Parameters of the pruned network for the PCA problem using Boltzmann pruning and Eigenvalue weighted Boltzmann pruning.

The result for the PCA problem are more complex. Some statistical measures of network capacity and accuracy for the PCA network pruned with the eigenvalue weighted pruning function are superior to the Boltzmann pruned network and some are not. Recognition accuracy is higher at all temperatures for both testing and training as can be seen by comparing figures 7 and 10. At the two critical temperatures accuracy is 71.8% for the Boltzmann case and 78.1% for the eigenvalue case. The number of weights used is 667 in the Boltzmann case

and 1046 for the eigenvalue weighted case. The Boltzmann pruned network is smaller but less accurate. The capacity-error product is high in the eigenvalue case both because there are more weights and because the number of bits per weight is higher. This indicates that the information transfer during training is more efficient for eigenvalue weighted training and that more information is retained. The reduction in network capacity and entropy are much more rapid in the eigenvalue weighted case as can be seen by comparing figures 8 and 11. This is a significant difference from the OCR problem. This capacity reduction results in a very gradual minimum in the capacity-error product as shown in 9 for Boltzmann pruning. No clear minimum is seen in the capacity-error product for the eigenweighted case as shown in figure 12. This makes selection of the critical temperature by minimization of the capacity-error product more difficult for the eigenvalue weighted case. The sharp drop in accuracy seen in figure 10 is used to obtain T_c.

5 Conclusions

Statistical evaluations of error and efficiency of information storage in pruned MLP networks for OCR and PCA have been made for Boltzmann pruning and eigenvalue weighted pruning. Both methods remove weights from the network in a self-organized way designed to optimize information content in the network. This optimization is carried out without knowledge of the effect of these size reductions. The eigenvalue weighted method is shown to be more effective in reducing network size without corresponding reductions in classification accuracy. We also show that the transfer of information from the training set using these methods is more efficient for the eigenvalue weighted method.

In addition to evaluating the statistical efficiency of the pruning methods, some comparisons of the difficulty of the OCR and PCA problems can be made. The OCR problem is reasonably well specified by the set of 10,000 K-L image features. This training set is adequate to allow the construction of a high accuracy, well minimized network. The PCA problem is still not adequately specified by the ridge direction features of 2000 finger prints. This suggests that solution of the PCA problem will require a larger training set, better features, and more efficient transfer of information from the training set to the neural network.

Acknowledgement

The author would like to acknowledge Rama Chellappa for suggesting the eigenvalue weighted method of network pruning.

References

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, et al., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 8, pages 318–362. MIT Press, Cambridge, MA, 1986.

[2] M. F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. Technical Report PB-339, Aarhus University, 1990.

[3] J. L. Blue and P. J. Grother. Training Feed Forward Networks Using Conjugate Gradients. In *Conference on Character Recognition and Digitizer Technologies*, volume 1661, pages 179–190, San Jose California, February 1992. SPIE.

[4] O. M. Omidvar and C. L. Wilson. Optimization of neural network topology and information content using boltzmann methods. In *Proceedings of the IJCNN, volume IV*, pages 594–599, June 1992.

[5] O. M. Omidvar and C. L. Wilson. Topological Seperation versus Weight Sharing in Neural Network Optimization. In Su-Shing Chen, editor, *Neural and Stochastic Methods in Image and Signal Processing*, volume 1766. SPIE, San Deigo, 1992.

[6] I. Guyon, V. N. Vapnick, B. E. Boser, L. Y. Botton, and S. A. Solla. Structural risk minimization for character recognition. In R. Lippmann, editor, *Advances in Neural Information Processing System*, volume 4, pages 471–479. Morgan Kauffman, 1992.

[7] M. D. Garris, C. L. Wilson, J. L. Blue, G. T. Candela, P. Grother, S. Janet, and R. A. Wilkinson. Massivelly parallel implementation of character recognition systems. In *Conference on Character Recognition and Digitizer Technologies*, volume 1661, pages 269–280. San Jose California, February 1992. SPIE.

[8] C. L. Wilson, G. T. Candela, P. J. Grother, C. I. Watson, and R. A. Wilkinson. Massively Parallel Neural Network Fingerprint Classification System. Technical Report NISTIR 4880, National Institute of Standards and Technology, July 1992.

[9] P. J. Grother. Karhunen Loève feature extraction for neural handwritten character recognition. In *Proceedings: Applications of Artificial Neural Networks III*. Orlando, SPIE, April 1992.

[10] R. Linsker. Self-organization in a perceptual network. *Computer*, 21:105–117, 1988.

[11] M. V. Wickerhauser. Fast approximate factor analysis. In *Proceedings October 1991*. SPIE, Washington University in St. Louis, Department of Mathematics, 1991.

[12] T. P. Vogl, K. L. Blackwell, S. D. Hyman, G. S. Barbour, and D. L. Alkon. Classification of Japanese Kanji using principal component analysis as a preprocessor to an artificial neural etwork. In *International Joint Conference on Neural Networks*, volume 1, pages 233–238. IEEE and International Neural Network Society, 7 1991.

[13] M. D. Garris and R. A. Wilkinson. Handwritten segmented characters database. Technical Report Special Database 3, **HWSC**, National Institute of Standards and Technology, February 1992.

[14] C. I. Watson and C. L. Wilson. Fingerprint database. *National Institute of Standards and Technology*, Special Database 4, **FPDB**, April 18, 1992.

[15] J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Networks*, 3(2):213–251, 1992.

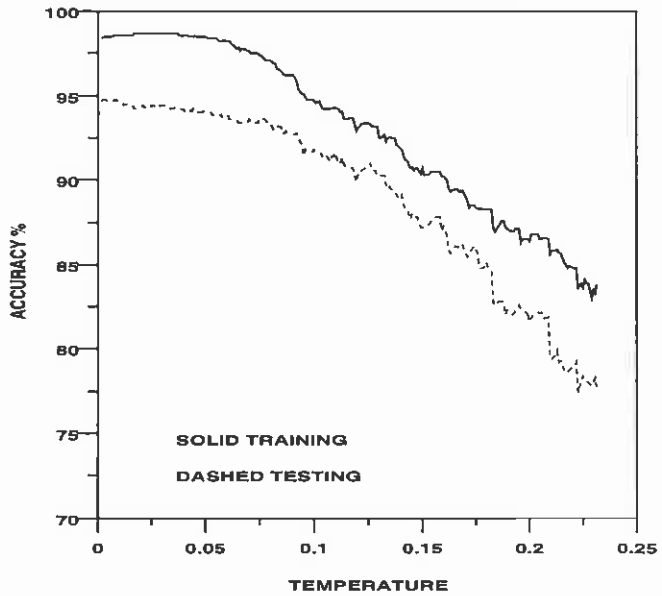

Figure 1: Network testing and training accuracy as a function of temperature for the OCR problem using Boltzmann pruning.

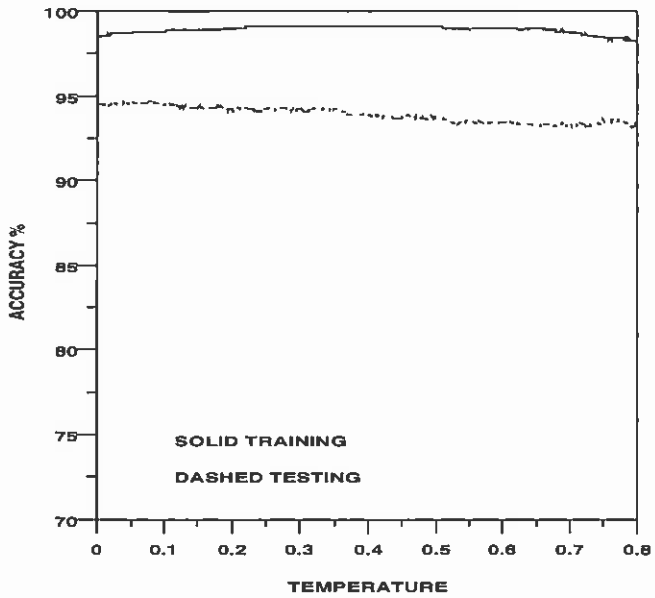

Figure 2: Network testing and training accuracy as a function of temperature for the OCR problem using Eigenvalue weighted pruning.

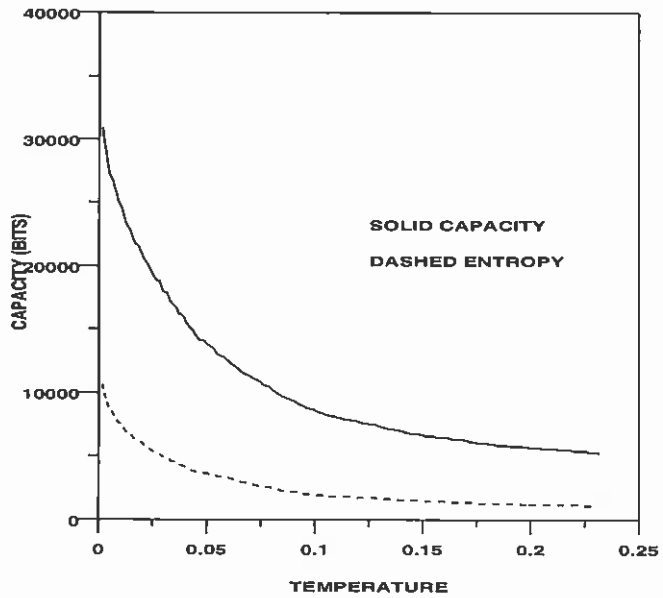

Figure 3: Network capacity and entropy as a function of temperature for the OCR problem using Boltzmann pruning.

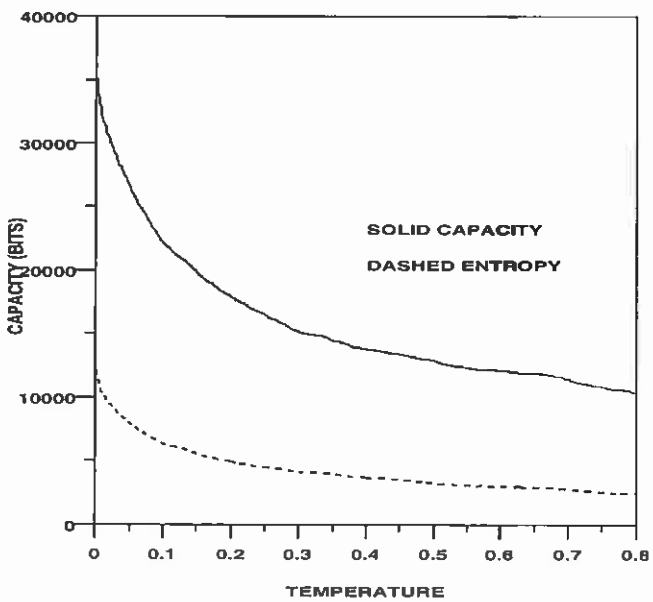

Figure 4: Network capacity and entropy as a function of temperature for the OCR problem using Eigenvalue weighted pruning.

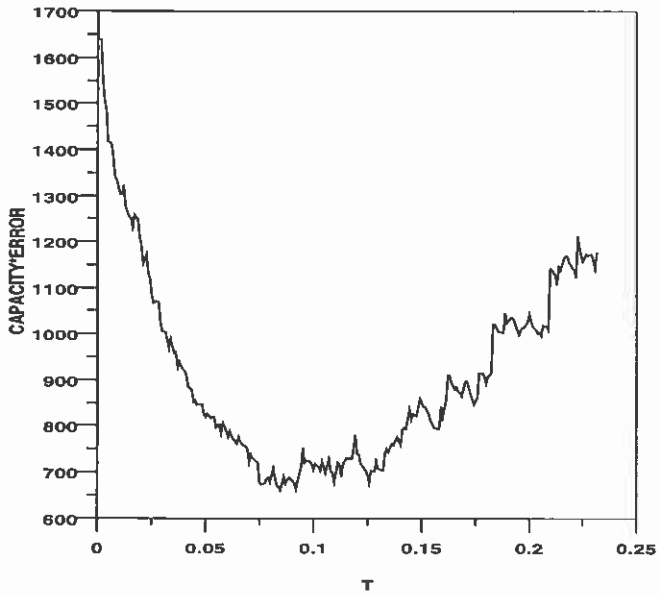

Figure 5: The product of capacity and error as a function of temperature for the OCR problem using Boltzmann pruning.

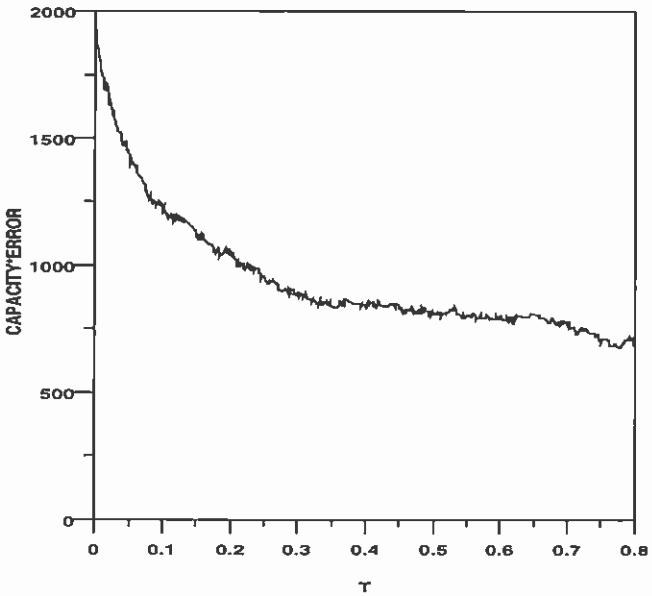

Figure 6: The product of capacity and error as a function of temperature for the OCR problem using Eigenvalue weighted pruning.

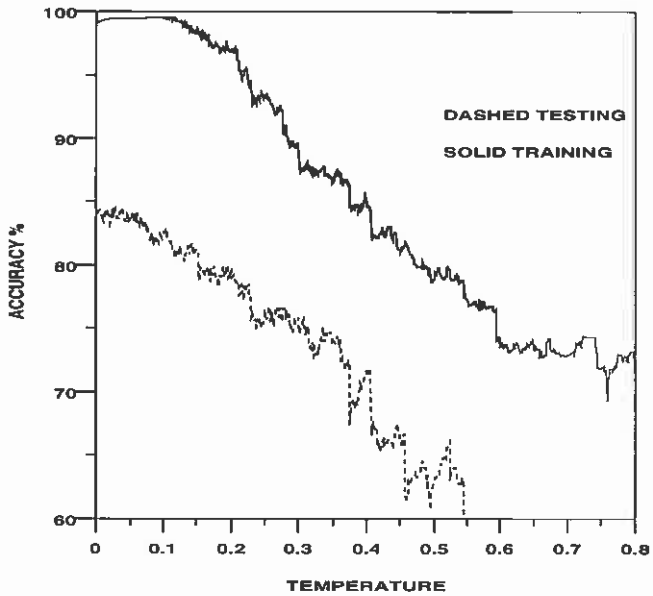

Figure 7: Network testing and training accuracy as a function of temperature for the PCA problem using Boltzmann pruning.

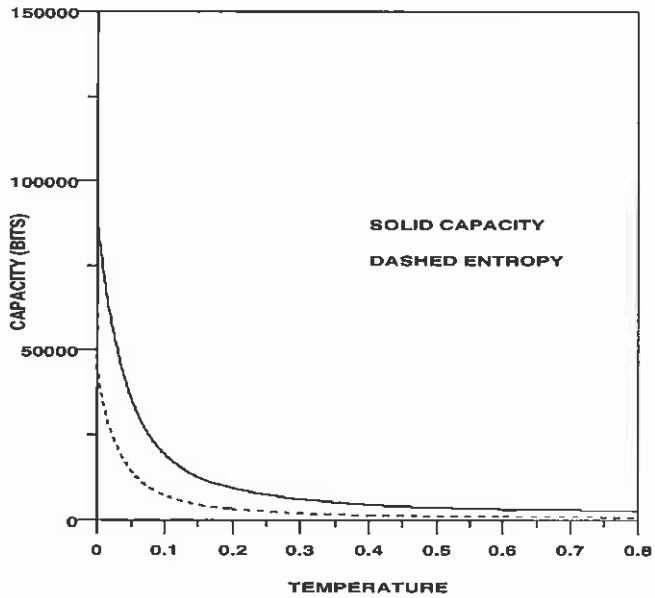

Figure 8: Network capacity and entropy as a function of temperature for the PCA problem using Boltzmann pruning.

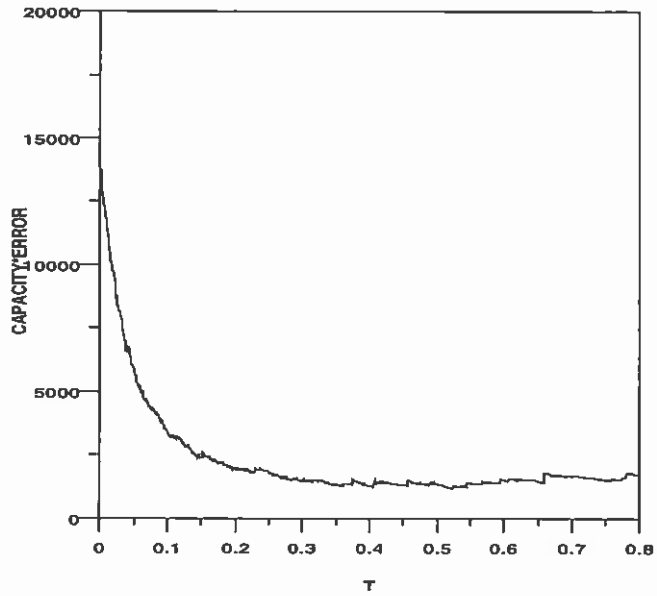

Figure 9: The product of capacity and error as a function of temperature for the PCA problem using Boltzmann pruning.

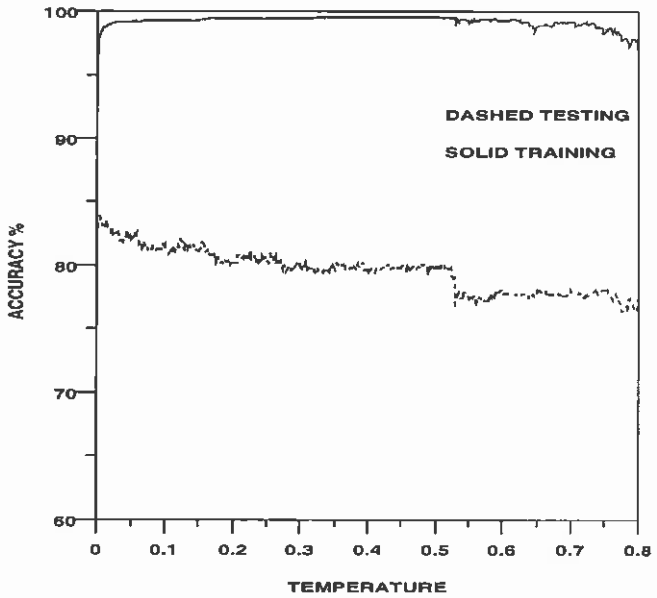

Figure 10: Network testing and training accuracy as a function of temperature for the PCA problem using Eigenvalue weighted pruning.

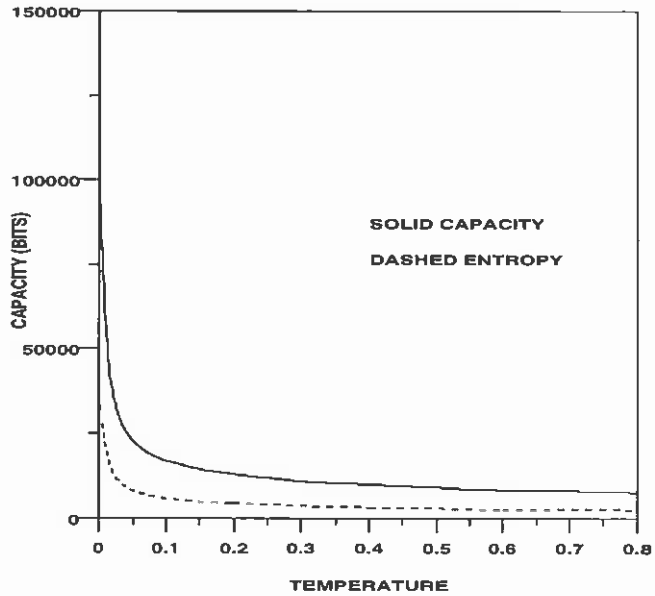

Figure 11: Network capacity and entropy as a function of temperature for the PCA problem using Eigenvalue weighted pruning.

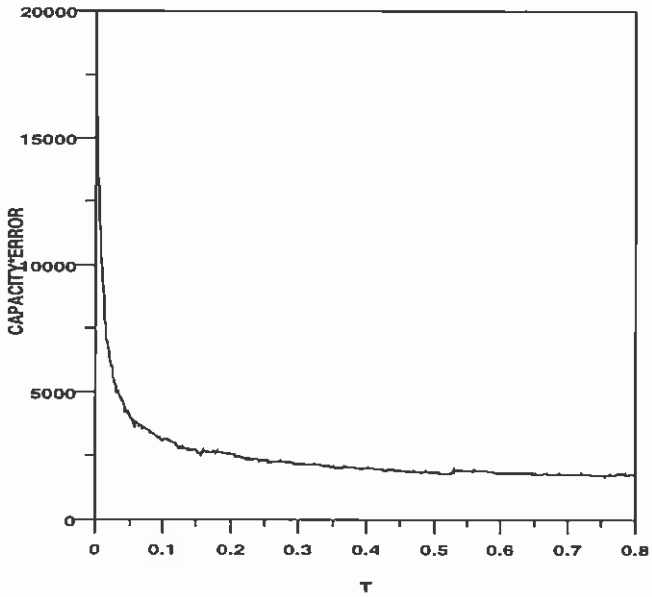

Figure 12: The product of capacity and error as a function of temperature for the PCA problem using Eigenvalue weighted pruning.