Feature extraction in shared weights neural networks

Dick de Ridder, Aarnoud Hoekstra and Robert P.W. Duin *

Pattern Recognition Group
Faculty of Applied Physics, Delft University of Technology
Lorentzweg 1, 2628 CJ Delft, The Netherlands
e-mail: `dick@ph.tn.tudelft.nl`

Abstract

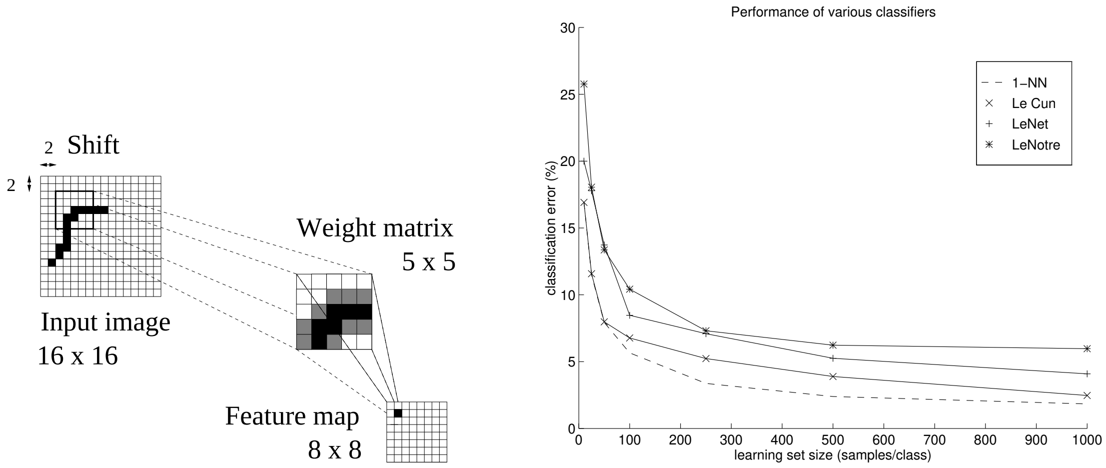
Since 1989 there have been numerous reports about shared weights neural networks: feed-forward networks performing very well in image analysis problems. As an explanation of these results, the authors of the corresponding papers propose that the networks somehow extract local features from the image, combining them at deeper levels in the network. Experiments on a large real-life dataset of handwritten digits show these networks to be able to achieve a low test set error. However, the filters or templates implemented by trained networks are uninterpretable and do not conform to conventional approaches in the image processing field. To study the feature extraction process, we trained a simple network on a small 2-class digit dataset. The results indicate that a comparison between these networks and traditional image processing techniques is not straightforward, since the two are based on different mechanisms and goals.

1 Introduction

Our research has been focused on the shared weights network (or TDNN, Time Delay Neural Network) paradigm proposed by among others Le Cun et al. in various papers (e.g., [LBD$^+$90]). These architectures differ from conventional feed-forward artificial neural networks (ANNs) in that they combine a large number of cells, connected using non-trivial schemes, with a relatively low number of free parameters. One or more layers consist of maps of units, in which each unit is connected to a unique square region in a map in the previous layer. Neighbouring units are connected to neighbouring regions. All units are connected to their regions using the same set of weights. In this way, a convolution is performed on the map in the previous layer. The way in which these ANNs are built leads to a classification process which closely resembles well-known image processing techniques, filtering and template matching. For an example, see figure 1 (a).

Neural networks, when used as classifiers, are able to implement highly non-linear classification functions. Normally however, when a certain problem is tackled using a neural approach, no insight is gained into the basic structure of the problem. Shared weights ANNs give the impression that certain sensible features are extracted from the raw image data

*This research is partly supported by the Foundation for Computer Science in the Netherlands (SION) and the Dutch Organization for Scientific Research (NWO).



(a) (b)

Figure 1: *(a) A set of shared weights acting as a convolution filter. (b) Classification errors in % on the* `TEST` *dataset, for the 1-nearest neighbour rule and the three ANN architectures used in the experiments: Le Cun, LeNet and LeNotre.*

and stored in sets of weights. However, because of the nature of neural network training, the problem remains that the features may be distributed over these sets of weights in an unclear, non-systematic way.

2 Handwritten digit recognition experiments

The work of Le Cun et al. [LBD$^+$90, LJB$^+$89] and Viennet [Vie93] was reproduced. Several shared weights network architectures were simulated and tested on a set of images of handwritten digits [Rid96].

2.1 The NIST database

The NIST (*National Institute for Standards and Technology*) distributes a CD-ROM containing images of handwritten forms and digits: Special Database 3 [WG92]. From this database, pre-segmented digit images were taken and pre-processed. The pre-processing consisted of:

- shearing the image to put the digit upright;
- eroding or dilating to normalize the line width;
- scaling down to a 16×16 image size, using a Gaussian kernel;
- scaling the pixel values to fall in the range $[-1, 1]$.

During these operations the aspect ratio was preserved. The resulting 2,500 images per digit were split into a training set (`LEARN`) of 1,000 images/class, a test set of 1,000 images/class (`TEST`) and a validation set (`VALIDATION`) - used in network training - of 500 images/class.

Name	Input size	Layers	Units	Connections	Weights	Biases
Le Cun [LJB+89]	19×19	5	1,361	63,660	8,760	1,000
LeNet [LBD+90]	28×28	6	4,634	94,952	2,584	3,850
LeNotre [Vie93]	16×16	4	394	2,210	626	138

Table 1: *The ANN architectures tested. Note the relatively small number of weights, compared to the number of connections, in in each ANN.*

2.2 The networks

Three shared weights network architectures were trained. They are listed in table 1. The 16×16 digit images were presented to the network with added borders of background pixels (value -1) to avoid filter overlap problems. The training goal was a vector using place coding: 10 elements indicating classes "0", "1", ..., "9". The training targets were 0.1 and 0.9. The transfer function used was the standard sigmoid; training was performed using the backpropagation rule, with a learning rate of 0.5 and no momentum. Training was stopped when the MSE on the `VALIDATION` set did not decrease for 50 cycles.

The networks were simulated and trained using the SPRLIB/ANNLIB software library [HKRS96].

2.3 Experiments

The ANNs described above were trained using a number of subsets of the training set sized 10, 25, 50, 100, 250, 500 and 1,000 images/class. Furthermore, the dataset was classified using the 1-nearest neighbour rule (1-nn). No feature selection was performed: the rule was used on all image pixels, stored in a vector of length 256. The results are shown in figure 1 (b).

The results were satisfactory, but did not completely match the original authors' claimed performance. However, our results can probably be improved upon by repeating the experiments and tweaking initialization and learning parameters. Nevertheless, the performance of the ANNs is good, considering the NIST database is a real-life dataset: the Le Cun network reaches a 2.5% `TEST` set error when trained on the entire `LEARN` set. Note that the LeNotre network performs quite well, although its number of parameters is far less than that of Le Cun and LeNet. This indicates that ANN architecture is at least as important as size.

Interesting is the excellent behaviour of the 1-nn method. It performs better than any of the ANNs used in the experiments, having a `TEST` set error of only 1.8% when using the entire `LEARN` set. Clearly, the 1-nn method does not perform any local shift-invariant feature extraction. A hypothesis is that the success of this method is partly due to the filtering step in the pre-processing, which increases the correlation between neighbouring pixels. The effect of filtering is studied in [Dui96].

3 A small network

Unfortunately, the learned weights (the filters) are hardly interpretable in terms of traditional image processing filters or templates. This is probably due to the large number of

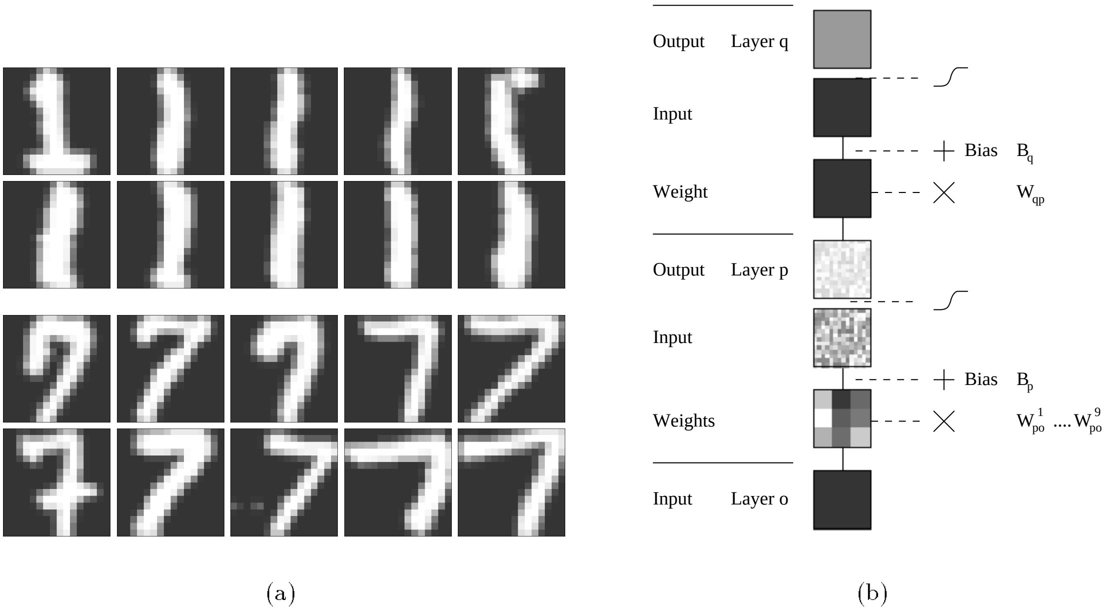


Figure 2: *The small digit dataset (a) and the network used (b). Note that layer o is the input layer, layer p contains the feature map and q is the output layer. For each layer, input and output is shown.*

filters in a network over which the information is distributed, and the complexity of the dataset. As the network adapts itself to certain peculiarities in the data, the filters become less and less general.

Therefore, to study the filter coefficient learning a small network was constructed which was trained on a dataset containing only 20 digits: $10 \times$ "1" and $10 \times$ "7". The network consisted of a 16×16 input layer o, one feature map p connected through a 3×3, 5×5 or 7×7 filter to the input and 1 output unit q. Training was performed with a goal of 0.1 for a "1" and 0.9 for a "7". The network and the dataset are depicted in figure 2.

Training the networks showed the following:

- when the use of bias in the feature map p was allowed, it would immediately be used to store place-dependent information;
- when the weights between the feature map p and the output unit q were not shared, they too were used for coding place-dependencies (see figure 3);
- for differently sized filters, the trained coefficients can be very different (see figure 4);
- even when trained without the use of bias in the feature map p and only one shared weight between the feature map p and output layer q, the filters are still hard to interpret.

The reason underlying the lack of interpretability lies in the fact that a shared weights network - at least in this form - is not forced to find an optimal and understandable filter. When there is only one shared weight between layer p and q, network output (and therefore classification) is based on the summation of the outputs of the feature map p. Assuming

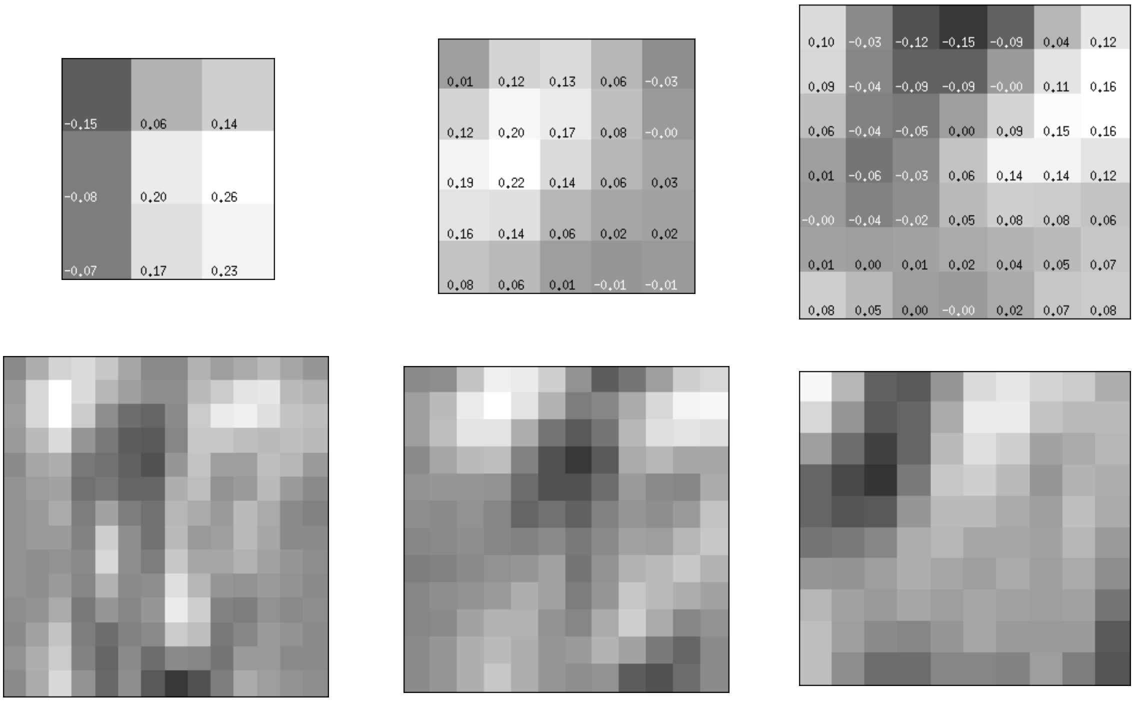
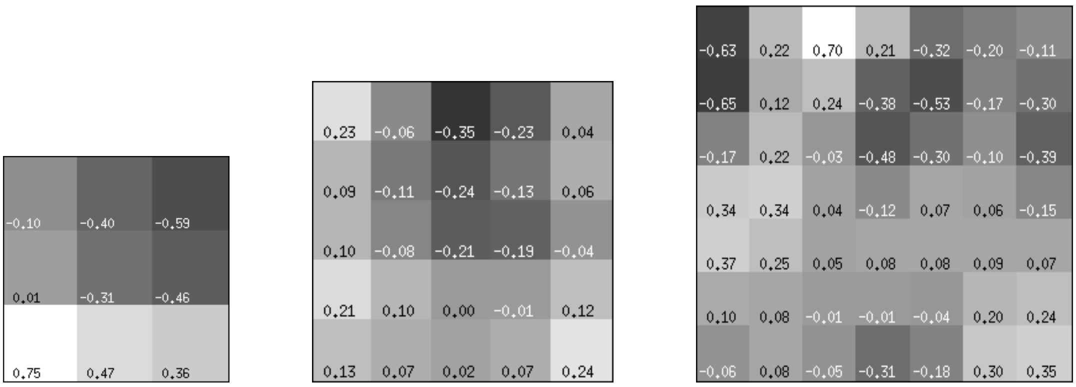


Figure 3: *Training results for the 1/7 dataset. The upper picture shows the trained filter, the lower the weights between the second layer p and output layer q. In the latter, mixed 1/7 templates are visible. Feature map sizes are* 3×3, 5×5 *and* 7×7.



(a) (b) (c)

Figure 4: *Training results for the 1/7 dataset. The pictures show the trained filters. The networks had only one shared weight between the second layer p and output layer q. Feature map sizes are* 3×3 *(a),* 5×5 *(b) and* 7×7 *(c).*

the sum of pixel values in the original image (layer o) is roughly equal for both digits, the sum of the convolutions (i.e., the input to p) is also equal. The network has to rely on the non-linearity of the sigmoid to scale the convoluted image differently for both classes, so that different values can be obtained in the output unit for both classes. Therefore, any filter which pushes the input values of feature map p further into the non-linear region of the sigmoid for one class than for the other, suffices.

4 Conclusion

Three shared weights ANNs have been tested on a real-life dataset, NIST Special Database 3. The ANNs performed quite well, although not as well as the 1-nearest neighbour method, which outperforms all ANNs studied.

To simplify the study of the filter coefficient learning, a small network was trained on a small dataset. It was shown that although these networks seem to perform operations related to the filtering and template matching techniques, the underlying mechanisms are quite different and need not lead to filters optimal in the image processing sense.

In light of these results, it seems that the original authors' idea that these ANNs are local, shift-invariant feature extractors perhaps deserves some further research. It is possible that local, shift-invariant feature extraction leads to a good performance. However, the reverse of that statement, i.e. concluding that a network must perform sensible feature extraction *since it performs well*, need not be true. The 1-nn method gives the best performance, yet it is clear that this method performs no feature extraction.

References

[Dui96] Duin. R.P.W. The influence of spatial pixel relations on the image recognition performance. In *Proc. Second Annual Conference of the Advanced School for Computing and Imaging*, Delft, 1996. ASCI.

[HKRS96] Hoekstra, A., Kraaijveld, M.A., Ridder, D. de, and Schmidt, W.F. *The Complete SPRLIB & ANNLIB.* Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, 1996.

[LBD+90] Le Cun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. Handwritten digit recognition with a back-propagation network. In Touretzky, D., editor, *Advances in Neural Information Processing Systems 2.* Morgan Kaufmann Publishers, San Mateo, CA., 1990.

[LJB+89] Le Cun, Y., Jackel, L.J., Boser, B., Denker, J.S., Graf, H.P., Guyon, I., Henderson, D., Howard, R.E., and Hubbard, W. Handwritten digit recognition: applications of neural network chips and automatic learning. *IEEE Communication*, pages 41–46, november 1989.

[Rid96] Ridder, D. de. Shared weights neural networks in image analysis. Master's thesis, Delft University of Technology, March 1996.

[Vie93] Viennet, E. *Architectures Connexionistes Multi-Modulaires, Application à l'Analyse de Scène.* PhD thesis, Université de Paris-Sud, Centre d'Orsay, 1993.

[WG92] Wilson, C.L. and Garris, M.D. Handprinted character database 3, february 1992. National Institute of Standards and Technology; Advanced Systems division.