Neural Network 'Surgery': Transplantation of Hidden Units

Axel J. Pinz, Horst Bischof
Department for Pattern Recognition and Image Processing
Institute for Automation, Technical University of Vienna
Treitlstr. 3, A-1040 Wien, Austria
email: api@prip.tuwien.ac.at

Abstract

We present a novel method to combine the knowledge of several neural networks by replacement of hidden units. Applying neural networks to digital image analysis, the underlying spatial structure of the image can be propagated into the network and used to visualize its weights (WV-diagrams). This visualization tool helps to interpret the behaviour of hidden units. We notice a process of specialization of certain hidden units, while others remain apparently useless. These units are cut out of one network and replaced by units taken from other networks trained for the same task using different parameters. We achieve better prediction accuracies for the new, combined network than for any of the two original ones. This constitutes a special kind of information fusion in image understanding.

We give an application example from the field of remote sensing, where neural networks are used to interpret the species of trees in aerial photographs. The interpretation accuracy is raised from 85% to 90%.

1 Neural networks in image analysis

A large amount of information about the world we live in is supplied by our visual system. Humans perform the task of vision effortlessly, without being

aware of this complex process. The goal of computer vision (image understanding) is to build computers which are able to see, a goal that has not been reached yet [Kro91]. Remarkable attributes of the human visual system are parallelism, robustness, and adaptivity. Similar attributes apply for neural networks (e.g. [Pao89, RM86]), a fact that makes them interesting for computer vision. The success of neural networks for pattern recognition tasks has been demonstrated in a variety of examples (e.g. [Bis89, SL88, Rot90, QS88]). A severe limitation of neural networks is the increasing learning time when scaling the problem up in size [Hin89]. In image analysis it seems intractable to use all pixels of an image as a neural network input, which would, in the case of a fully connected architecture, result in networks with several millions of connections per unit. One solution is to use modular neural networks (e.g. [Bis91, BP92]), another one, to use only small portions of an image as an input. In many cases, when the observed objects are compact and small compared to the image size, this is a reasonable approach.

Figure 1: A neural network for image analysis

Fig.1 sketches a neural network architecure, where a $n \times n$ pixel window is used as an input to a three layer feed forward neural network. The pixels form the input vector (n^2 input units), h hidden units and o output units are used to recognize o different classes of objects whenever they are present in the input window [BP89, BP90, PB90].

To catch our line of argument, it is important to emphasize the propagation of spatial structure through a network of this kind. The pixels of the input window can be assigned a location in 2d space, and, in the case of the one to one mapping used here, this location can be propagated through the input units, so that the same location is valid for the input units and the weights of the hidden units. This was discussed in detail in [BP91] and leads to the weight visualization diagrams discussed below.

2 Weight visualization

Having a network trained for a certain task, it has accumulated its knowledge distributedly in form of its weights. In the case of the image analysis network (Fig. 1), we can arrange the $n \times n$ weights of each hidden unit in the same spatial manner as the input window and visualize them as an image. Fig. 2 shows such a WV-diagram (weight visualization diagram [AJP90, BP91]). Positive weights are shown in red, negative weights in blue, with high intensity of the colors for high absolute values of the weights. The five pixels at the top of the WV-diagram visualize the weights of the hidden unit to the five output units. Thus, WV-diagrams visualize both the receptive field (the weights of the incoming connections of a unit) and the projective field (the weights of the outgoing connections) of a unit [SL88]. In comparison with Hinton diagrams, WV-diagrams provide a better visual impression (especially if color is used) and require less space, so that the weights of a complete network like Fig.1 can be visualized on one screen (see Fig. 4).

3 Specialization of hidden units

Looking at the WV-diagrams in Fig.4, we find that many diagrams have structures, but some also have none (the amount of structure can be roughly quantified by the empirical variance of the weights). Compared with the less

structured ones, these WV-diagrams with structure usually show large and diverse output weights, indicating good discriminatory ability.

The receptive field gives us hints about the features a hidden unit is extracting. The more structured the receptive field, the more specific is the feature extracted by this unit: Hidden units with different structures will also compute different features. For these reasons one can talk about a kind of specialization of hidden units. One should note, however, that though specialization takes place, the network does not necessarily develop a local representation in the hidden units. More than one hidden unit can be active for a given input, so that a set of hidden units constitutes a class. These hidden units cannot be seen independently, they mutually depend on each other, and it is very difficult to devise a measure for the importance of a particular unit. Most measures proposed (e.g. [CDS88, MS88b]) are not sensitive to these mutual dependencies. In cluster analysis [HB90], the activations of the hidden units are also treated as independent variables, which makes this approach at least questionable for our purpose.

4 Cutting out unnecessary hidden units

To be able to remove superfluous hidden units, it is crucial to identify the 'good' ones (i.e. hidden units which contribute to the discrimination of objects). In addition to a well structured receptive field, certain hints can be gained from the inspection of the projective field. If the receptive field of a unit is unstructured, or the weights of the projective field are small or uniform, this hidden unit does not contribute to the discrimination. Such units can be cut out of the network. After a brief training phase, the pruned network shows similar prediction accuracies as the original one. This performance cannot be reached when we start with a network with fewer hidden units and train it for the same task using the identical training set. In this case we get lower prediction accuracies and the smaller network also has some unstructured units.

5 Combination of the knowledge of several networks by transplantation of hidden units

The training phase of a neural network is influenced by the training set and a few tunable parameters. Examining several networks of identical architecture, which are trained for the same task using slightly different training sets or parameter settings, these networks may show considerably different performances. The idea is to combine the knowledge of these networks to yield better prediction accuracies. While our method has similarities to McMillan and Smolensky [MS88a], our goal is a different one. They tried to show that there were rules (which they called 'soft rules') in their network; we want to improve the prediction rate of our model.

Assume that we have two networks, that network 1 needs an improvement in predicting class x and that network 2 is good in predicting class x. We want to incorporate this knowledge into network 1.

- The first step is to prune network 1 by removing its unstructured hidden units.
- Now, the free places in network 1 are substituted by 'good' hidden units of network 2. Guidelines to find such a hidden unit are its structure (high variance) and a good projective field (strong weight for class x). These good hidden units are cut out of network 2 and placed into network 1.
- Finally, the new network has to be adapted by post-transplantational training. The weights of the projective field of all hidden units are initialized randomly, and the weights are trained by a two layer learning rule (e.g. Widrow Hoff rule [WS85]). Since such a rule converges much faster than e.g. backpropagation, only a few steps are necessary to train the projective field.

An examination of the projective field of the new network shows how the new units fit into it. If the weights of the projective field are small or have almost equal values, the integrated hidden units will not improve the prediction accuracy. This can happen even if such a unit has a good projective field in its original net and it can happen to every unit in the new

network regardless of whether it is an original or a replacement unit. This demonstrates the mutual dependencies of hidden units and the difficulty in locating distributed knowledge. It is not sufficient to know the projective and receptive fields of a unit to properly interpret its behavior, information about the surrounding units is also required.

6 An application example: tree species interpretation

We demonstrate the transplantation method by giving an example from the field of remote sensing. The task of the neural network is to interpret the species of trees in color infrared aerial photographs. This application is discussed in detail in [PB90, BP89]. Fig.3 shows a small portion of an aerial photograph, where the trees are marked by circles and the species is specified (S=spruce, P=pine, B=beech). The trees are located by a different system (the Vision Expert System [Pin89]) and the pixels of a 15×15 window are used as the neural network input (similar to the architecture shown in Fig.1).

The network is trained to distinguish between five different species of trees, and it requires 13 hidden and 5 output units to solve this task. Table 1 compares the results of the best network without transplantation (86% for the training set and 85% for the test set) with the best network after several transplantations (93% and 90%). Fig.5 shows the WV-diagrams for the improved network. Compared with Fig.2, we find more structured hidden units with better projective fields.

Table 1: Comparison of results

	Training Set	Test Set
Normal Training	86%	85%
Transplantation	93%	90%

7 Discussion

A novel method, the transplantation of hidden units, was presented and applied to an example from remote sensing. It yields better prediction rates for the new, combined network than for any of its predecessors. This method constitutes a kind of information fusion - a current research issue in image understanding. The transplantation is possible, because a kind of specialization of hidden units is developed during the training phase. The proper selection of promising transplantation candidates is currently guided by weight visualization diagrams and is extremely difficult. A formalism for the automatic selection of such candidates is still required.

Since relevance measures for units are not sensitive to mutual depencies of units, other formalisms have to be developed. A possible approach is to use genetic algorithms for the transplantation. One can start with a population of already trained networks. When two networks are combined, several units are exchanged and the new network is a further individual of the population. The networks with the best prediction accuracies survive and are considered for reproduction in the next generation, where the process continues.

It should be pointed out that the proposed method is different from pruning [CDS88]. The transplantation of hidden units aims at the implantation of foreign hidden units to improve the prediction rates of a network, whereas the goal of pruning is to reduce the size of the network at comparable performance. Only the first step in the transplantation process, where those hidden units which do not contribute to the classification accuracy are identified, is similar to pruning.

Figure 2: A WV-diagram

Figure 3: Trees in an aerial photograph

Figure 4: WV-diagrams of a complete network

Figure 5: WV-diagrams of a complete network after transplantation

References

[AJP90] Horst Bischof Axel J. Pinz. The IVF Weight-Visualization Diagrams (WV-Diagrams). Technical report, Inst. of Surveying and Remote Sensing, Univ. of Bodenkultur, July 1990.

[Bis89] Horst Bischof. Interpretation von Fernerkundungsdaten mit Hilfe von Backpropagation Netzwerken am Beispiel der Baumerkennung aus Farb-Infrarot Luftbildern. Master's thesis, TU-Vienna, 1989.

[Bis91] Horst Bischof. Modular, Hierarchical, and Geometrical Neural Networks. Technical Report 9, Dept. for Pattern Recognition and Image Processing, TU-Vienna, December 1991.

[BP89] Horst Bischof and Axel J. Pinz. Verwendung von neuralen Netzen zur Bestimmung der Baumart. In Axel J. Pinz, editor, *Knowledge Based Pattern Recognition*, number 49 in OCG-Schriftenreihe, pages 149–161. Oldenbourg, 1989.

[BP90] Horst Bischof and Axel J. Pinz. Verwendung von neuralen Netzwerken zur Klassifikation natürlicher Objekte am Beispiel der Baumerkennung aus Farb- Infrarot-Luftbildern. In Georg Dorffner, editor, *Konnektionismus in Artificial Intelligence und Kognitionsforschung*, number 252 in Informatik Fachberichte, pages 112–120. Springer Verlag, 1990.

[BP91] Horst Bischof and Axel J. Pinz. Visualization Methods for Neural Networks. In Holger G. Ziegeler, editor, *Konnektionismus:*

Beiträge aus Theorie und Praxis, number 3 in Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence, pages 1–14. ÖGAI, 1991.

[BP92] Horst Bischof and Axel J. Pinz. Neural Networks in Image Pyramids. To appear in the proceedings of the Int. Joint Conference on Neural Networks 92, 1992.

[CDS88] Y.L. Le Cun, J.S. Denker, and S.A.Solla. Optimal Brain Damage. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, number 2. Morgan Kaufmann, 1988.

[HB90] S.J. Hanson and D.J. Burr. What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences*, 13:471–518, 1990.

[Hin89] G. Hinton. Connectionist Learning Procedures. *Artificial Intelligence*, 40:185–234, 1989.

[Kro91] Walter G. Kropatsch. Image Pyramids and Curves an Overview. Technical Report 2, Dept. for Pattern Recognition and Image Processing, TU-Vienna, March 1991.

[MS88a] C. McMillan and P. Smolensky. Analyzing a Connectionist model as a system of soft rules. In *Proc. Tenth Conference of Cognitive Science Society*, 1988.

[MS88b] M.C. Mozer and P. Smolensky. Skeletonization: A Technique for Trimmming the Fat from a Network via Relevance assesment. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, number 1. Morgan Kaufmann, 1988.

[Pao89] Yon Han Pao. *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley, first edition, 1989.

[PB90] Axel J. Pinz and Horst Bischof. Constructing a Neural Network for the Interpretation of the Species of Trees in Aerial Photographs. In *Proc. Tenth International Conference on Pattern Recognition*, pages 755–757. IEEE Computer Society Press, 1990.

[Pin89] Axel J. Pinz. Final results of the Vision Expert System VES: Finding Trees in Aerial Photographs. In Axel J. Pinz, editor, *Knowledge Based Pattern Recognition*, number 49 in OCG-Schriftenreihe. Oldenbourg, 1989.

[QS88] N. Qian and T.J. Sejnowski. Predicting the Secondary Structure of Globular Proteins using Neural Network Models. *Journal of Molecular Biology*, 202:865–884, 1988.

[RM86] David E. Rumelhart and James A. McClelland. *Parallel Distributed Processing*, volume 1. MIT Press, first edition, 1986.

[Rot90] M.W. Roth. Survey of Neural Network Technology for Automatic Target recognition. *IEEE Transaction on Neural Networks*, 1(1):28–43, 1990.

[SL88] T.J. Sejnowski S.R. Lehky. Network model of shape-from-shading: Neural functions araises from both the receptive and projective field. *Nature*, 333:452–454, 1988.

[WS85] B. Widrow and S.D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, New York, 1985.