

LOCAL FEATURE EXTRACTION—WHAT RECEPTIVE FIELD SIZE SHOULD BE USED?

Rita Kovordanyi

Chandan Roy

Mohammad Saifullah

Department of Computer and Information Science
Linköping University
Linköping, Sweden

ABSTRACT

Biologically inspired hierarchical networks for image processing are based on parallel feature extraction across the image using feature detectors that have a limited Receptive Field (RF). It is, however, unclear how large these receptive fields should be. To study this, we ran systematic tests of various receptive field sizes using the same hierarchical network. After 40 epochs of training, we tested the network both by using similar but novel images of the same tropical cyclone that was used for training, and by using dissimilar images, depicting different cyclones. The results indicate that correct RF size is important for generalization in hierarchical networks, and that RF size should be chosen so that *all* RFs at least partially cover meaningful parts of the input image.

Index Terms—Hierarchical artificial neural networks, local feature extraction, receptive field size, generalization, activation-based receptive field analysis

1. INTRODUCTION

One approach to artificial neural networks for image processing employs a network architecture that is inspired by the human visual system. These biologically-inspired networks have a hierarchical structure. At the base of this structure are the input and the first hidden layer, which detects local features that are extracted from part of the input image [1] [2] [3] [4]. These features typically constitute oriented line segments and/or color patches that are located in a limited part of the input image.

Receptive field (RF) denotes the part of the input image which a unit receives input from. In this study, we focus on the RF of units in the first hidden layer. At the second and subsequent hidden layers, the local features are combined into increasingly complex patterns located in progressively larger parts of the input image [1] [2] [3] [4]. Hence, RFs become larger for each hidden layer in the network.

This is achieved through a hierarchical connectivity, by connecting each hidden layer $k+1$ group-wise to the previous layer k , so that each unit group in layer $k+1$ receive signals from a small number of groups in layer k (Figure 1). This connectivity pattern makes up a hierarchical communication structure, where patches in the input image are processed by dedicated receiving groups in the first hidden layer, and these groups in turn project to dedicated groups in the subsequent layers, until a single group of units can cover the complete input image.

In hierarchical networks, each unit receives signals that directly or indirectly originate from a limited part of the input image. This part makes up the unit's receptive field, in the following called RF. Of particular importance are the RFs of the first hidden layer units, as these serve subsequent processing in the network. Also, RF size in the first hidden layer and the rate of convergence of RFs in subsequent hidden layers determine the number of hidden layers that are necessary to obtain a full connection hierarchy, where object recognition works on the whole input image. Hence, it is especially important that RF size is chosen carefully in the first hidden layer.

Theoretically, on the one hand, larger RF size at the first hidden layer should yield smaller networks, containing a fewer number of layers, decreasing the network's complexity, which in turn ought to decrease the risk for overfitting.

On the other hand, smaller RF size would encourage the development of micro-feature detectors in the network, which would presumably allow for a greater recombination possibilities, and support combinatorial generalization [5].

To clarify the role of RF size, we conducted systematic generalization tests where we varied the RF size of the first hidden layer. In particular we wanted to relate the optimal RF size to the size of the meaningful portion of the input image, in our case, the oval contended by the cyclone's spiral shape. We hypothesized that optimal RF size is dependent on the size of this meaningful portion of the input image.

2. THE NETWORKS USED IN THIS STUDY

Hierarchical networks for image processing have been demonstrated to be especially well suited for extracting local features from an image. Feature extraction is often achieved using various filters [1] [4] [6]. These filters are pre-defined, which simplifies learning in the network. As pre-defined filters were ill-suited for the amorphous cloud shapes in our tropical cyclone images, we let feature detectors in the first hidden layer develop through training.

In contrast to standard hierarchical networks where feature integration and translation invariance is achieved by two separate but intertwined hierarchies of layers, we implemented both types of computation using one hierarchy. Kovordanyi and Roy [7] provide a detailed description of this architecture and its usefulness for image processing.

We used the same convergence rate for all networks that were tested, which inevitably meant that the networks contained a different number of layers, depending on the initial RF size at the first hidden layer. Hence, the networks comprised of five to six layers: Input, V1, V2, V3, and Dir, alternatively Input, V1, V2, V3, V4, and Dir, depending on how many layers were needed to make the connections converge into a final layer where units received input from the complete image (cf. Figure 1). The networks also contained an additional layer Correct_dir. This layer did not partake in computation, but allowed us to visually compare the network's output with the desired output.

3. TESTING PROCEDURE

A series of recurrent hierarchical networks with a 66 x 66 input layer size was developed in the neural network tool Emergent [8]. The network's task was to predict cyclone movement based on cloud patterns in a satellite image. We used satellite images of one cyclone for training. The

original image was rotated and shifted in up to four steps in eight directions, producing about 200 images. We wanted to relate the optimal RF size to the size of the meaningful portion of the input image, in our case, the oval contended by the cyclone's spiral shape. For this reason, the size of the cyclone relative to the image frame was kept constant during training and testing. About 95% of the 200 translated images were used for training. The remaining 5% of the images were set aside for testing. In addition, we tested the network using all possible translations (except zooming) of four novel cyclones.

3.1 Simulations

We ran systematic tests of nearly identical networks, the main difference being the RF size at the first hidden layer.

Hidden layer units were organized into groups, each group sharing the same receptive field, and thus looking at the same part of the input image. RF size of the first hidden layer units was varied from 16 x 16 pixels, to 20 x 20 pixels, and finally 26 x 26 pixels. Using a few pixels' overlap between neighboring RFs, the chosen RF sizes yielded a group structure in the first hidden layer, so that V1 consisted of 5 x 5, 4 x 4, and 3 x 3 receiving groups, respectively (Figure 1).

The number of units in each receiving group in layer V1 was $6 \times 6 = 36$. This number, as well as other network parameters, was kept constant across the networks that were tested. Hence, the only variation between networks was the RF size at the first hidden layer, which in turn determined the number of layers that were required to build a full connection hierarchy with the top hidden layer receiving input from the entire image.

Emergent offers a sigmoid-like activation function, and saturating weights limited to the interval [0, 1]. Learning in the network was based on a combination of Conditional

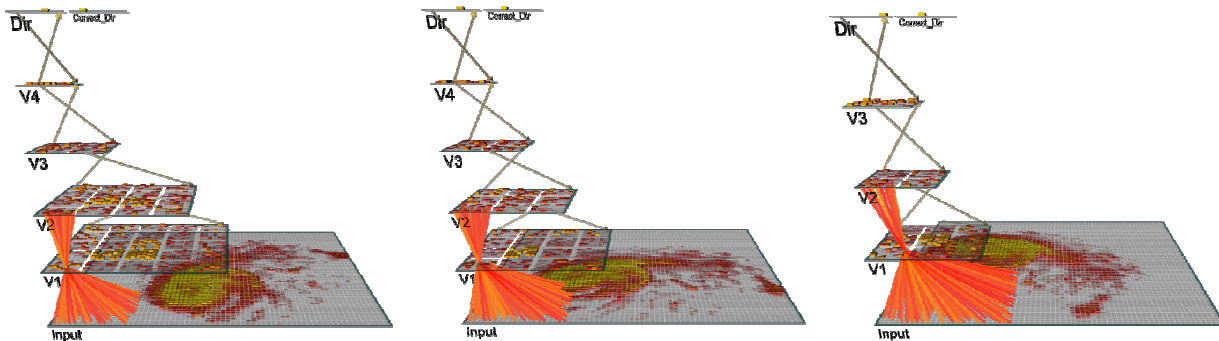


Figure 1. The three types of recurrent hierarchical network that were used in this study. The dense connecting lines between layers Input and V1 illustrate the RF for receiving group (1, 1) in layer V1, as well as the projection field of this group in layer V2. As can be seen, RF at the first hidden layer (V1) was varied across the three networks, from 16 x 16, through 20 x 20, to 26 x 26 pixels (input image dimension was 66 x 66 pixels in each case).

Principal Component Analysis (CPCA), which is a Hebbian learning algorithm and Contrastive Hebbian learning (CHL), which is an error-driven algorithm, a biologically-based alternative to backpropagation of error [9]:

$$\text{CPCA: } \Delta w_{ij} = \epsilon y_j (x_i - w_{ij}) = \Delta_{hebb} \quad \text{Eq. 1}$$

x_i = activation of sending unit i
 y_j = activation of receiving unit j
 w_{ij} = weight from unit i to unit j

$$\text{CHL: } \Delta w_{ij} = \epsilon (x_i^+ y_j^+ - x_i^- y_j^-) = \Delta_{err} \quad \text{Eq. 2}$$

x_i = activation of sending unit i
 y_j = activation of receiving unit j
 x^+, y^+ = act when also output clamped
 x^-, y^- = act when only input is clamped

$$\text{Learning mix: } \Delta w_{ij} = \epsilon [c_{hebb} \Delta_{hebb} + (1 - c_{hebb}) \Delta_{err}] \quad \text{Eq. 3}$$

ϵ = learning rate
 c_{hebb} = proportion of Hebbian learning


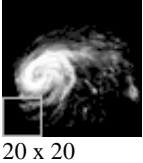
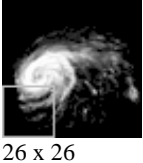
The amount of Hebbian learning used was based on previous systematic testing [10]. We used $c_{hebb} = 0.01$ for connections between Input and V1, and $c_{hebb} = 0.001$ at subsequent layers.

Recurrent networks tend to get over-activated during settling due to feedback signals that can induce an uncontrollable spread of activation across the entire network. Recurrent networks must therefore use some inhibitory mechanism to reduce this tendency for over-activation. In order to keep at least some units active, the amount of inhibition must be adapted to the actual net input coming into the layer. Emergent offers a k -Winners-Take-All (k WTA) mechanism, which allows at most k units to stay active in each layer or unit group. Assuming that all units within a layer or unit group have been sorted in ascending order according to their activation level, the objective of the k WTA-algorithm is to keep units 1- k active, while inhibiting units $k+1$ - N . The amount of inhibition g_i to be delivered to a layer or a unit group is defined to lie somewhere between the *inhibition threshold* of unit $k+1$, $g_i^\ominus(k+1)$, which is the amount of inhibition that is required to drive unit $k+1$ below its activation threshold, and the inhibition threshold, $g_i^\ominus(k)$, of unit k :

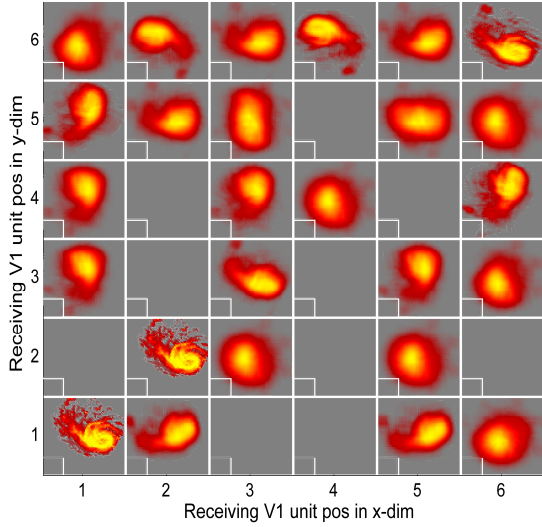
$$g_i = g_i^\ominus(k+1) + q(g_i^\ominus(k) - g_i^\ominus(k+1)) \quad \text{Eq. 4}$$

$g_i^\ominus(k)$ = inhibition threshold for unit k
 q = margin above required inhibition level

Table 1. Results from the generalization tests using both a 5% testing set from the same cyclone that was used for training and images of four new cyclones. Testing with the four new cyclones was based on the two best weight sets selected from the five listed below.

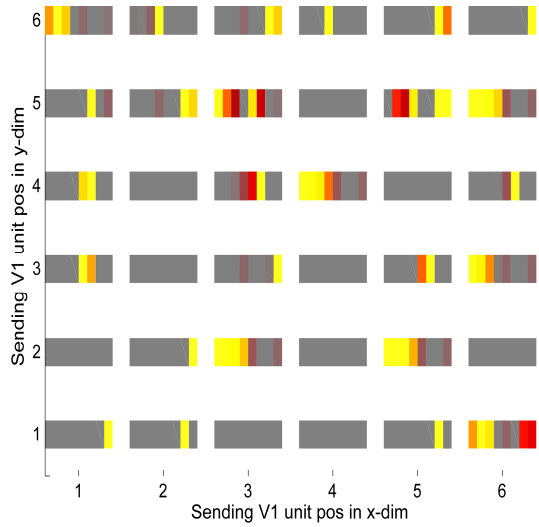
Cyclones		Generalization with test images		Generalization with new cyclones							
Cyclone used for training	RF size, and size and number of the receiving groups in V1 layer			Total 7		Total 200		Total 200		Total 200	
		Corr (%)	Wrong (%)	Corr (%)	Wrong (%)	Corr (%)	Wrong (%)	Corr (%)	Wrong (%)	Corr (%)	Wrong (%)
 16 x 16	RF size 16 x 16, group size 6 x 6, 25 groups	85.7 100 57.1 85.7 100	14.3 0 42.9 14.3 0	64.5 71.5	35.5 28.5	30 7.5	70 92.5	48.5 57.5	51.5 42.5	45 59	55 41
 20 x 20	RF size 20 x 20, group size 6 x 6, 16 groups	85.7 100 71.4 100 85.7	14.3 0 28.6 0 14.3	71.5 76.5	28.5 24.5	5.5 20.5	94.5 79.5	58.5 58.5	41.5 41.5	52 66.5	48 33.5
 26 x 26	RF size 26 x 26, group size 6 x 6, 9 groups	100 85.7 100 100 100	0 14.3 0 0 0	87 96.5	13 3.5	36 53	64 47	68 74	32 26	88 88	12 12

Activation based receptive fields from Input layer into target layer V1



a.

Activation based projection fields from target layer V1 to Dir layer



b.

Figure 2. Activation-based receptive and projective field analysis for layer V1 in the 16 x 16 receptive field network architecture (leftmost network in Figure 1). The plots show the thirty-six (6 x 6) units in the lower leftmost group in V1 organized in the same order as they appear within the layer. **a.** The plot on the left shows average activation mediated from the Input layer into each of the thirty-six units in the receiving group. So, in a sense, the plots show the average of all those training and testing images that a particular unit has learnt to react to. Note that all thirty-six units have the same RF, that is they receive input from the same part of the image (marked with white squares). **b.** The plot on the right shows projection fields (output) from the same unit group in V1 into the output layer Dir (8 x 1 units). The two plots reveal that about ten units in V1 are never activated during processing. In addition, nine or so units have not developed any useful feature representations (round blobs) and become activated for all inputs. These units also project to a large number among the eight directional units (those direction units that receive projection are marked with yellow-red squares).

In this study, inhibition g_i was calculated separately for each unit group within a layer. q was set to 0.25, which is the default value used for the standard k WTA algorithm in Emergent ([9]).

4. TEST RESULTS

We tested the network’s generalization capability in two ways: First, by using the 5% testing set containing novel translations (orientation and position variations) of the same cyclone that was used for training, and second by using a full set of images (all orientation and position variations) of four novel cyclones. We recorded the number of errors and calculated the percent errors that were made (Table 1).

As illustrated in Table 1, generalization performance was worse when a small RF of 16 x 16 pixels was used in layer V1. Generalization performance was intermediate for 20 x 20 pixels RF in layer V1. Generalization was best for an RF size of 26 x 26 pixels. This indicates that RF is an important factor for generalization, and that RF needs to be big enough so that all receiving groups in V1 cover some portion of the meaningful parts of the image. The meaningful parts would be those parts of the image that

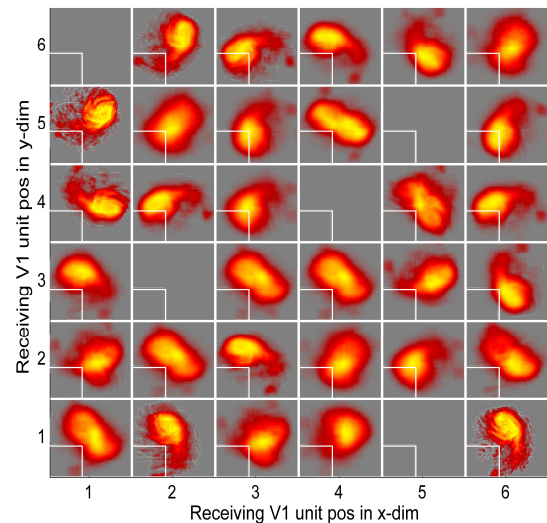
contain task-relevant information, such as information about object identity. Non-meaningful parts could, for example, be background information and/or noise.

In addition to the above tests, we analyzed the weight structure developed in the network during training, using an indirect method called activation-based receptive field analysis. This analysis is based on the co-activation of input-units and a particular receiving unit. The average co-activation taken over all input images reflects the tendency of this particular receiving unit to react to particular features in the input.

The activation-based receptive field analysis reveals that the feature detectors that were developed in layer V1 for the small RF sizes (16 x 16) were often not indicative of cyclone direction (which was an important task-relevant dimension of the input). Also, units were to a large extent dead that is, they did not partake in the processing of any input (Figure 2).

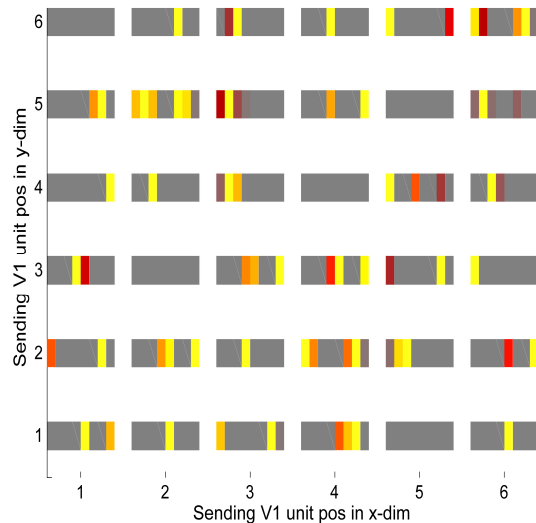
In contrast, V1 units with large RFs (26 x 26 units) developed useful representations, and tended to react to input depicting cyclones in a few specific directions (which can be seen in the elongated shapes in Figure 3a). In addition, the V1 feature detectors turned out to indirectly

Activation based receptive fields from Input layer into target layer V1



a.

Activation based projection fields from target layer V1 to Dir layer



b.

Figure 3. Activation-based receptive and projective field analysis for layer V1, when 26 x 26 pixels RF were used in layer V1. The same unit group is shown as in Figure 2 (depicting the lower leftmost receiving group in V1). Note that although average incoming activation is shown across the whole image, units actually receive input from a small RF within the image (marked with white squares). **a.** As illustrated in the plot at the left, most units have developed useful feature representations with some form of direction indication, **b.** and these feature detectors project to a small number of directional units in the output layer (those direction units that receive projection are marked with yellow-red squares).

project to a few direction units in the output layer, indicating that the V1 units were to a greater extent useful for the task of predicting the direction of cyclone movement (Figure 3).

5. CONCLUSIONS

Our results indicate that RF-size has to be adapted to the type of input images that are used. In particular, RF size in the first hidden layer must be chosen so that each individual RF will cover meaningful information; meaningful in the sense that the information contributes to the task the network has to accomplish. Meaningful information makes up the foreground of the image, while other information and/or noise are part of the image's background.

For the images that we have used, optimal RF size turned out to be relatively large, 26 x 26 pixels, which is about one third of the image size. When we used smaller RFs, peripheral receiving groups, for example covering the lower left corner of the input image, did for some image translations not see any part of the foreground. At the same time, these unit groups were encouraged by the *k*WTA algorithm to produce an activation pattern consisting of *k* active units.

It may be the case that part of the observed effect can be attributed to the *k*WTA-algorithm that we used, which encourages a certain level of activation (*k* active units) in each receiving group. This entails that those units whose RF

does not, at least partially, cover foreground information will develop a tendency for spurious activations. Instead of mediating useful information in the feed-forward direction, these units will be driven by feedback signals, which may hamper learning and subsequent generalization in the network. This effect will, of course, not occur if larger RFs are used that cover part of the image foreground.

6. REFERENCES

- [1] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, 1999, pp. 1019-1025.
- [2] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, 1980, pp. 193-202.
- [3] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 994-1000.
- [4] J. Mutch and D. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," *International Journal of Computer Vision*, vol. 80, 2008, pp. 45-57.

- [5] R. O'Reilly, "Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning," *Neural Computation*, vol. 13, 2001, pp. 1199-1241.
- [6] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, 2007, pp. 411-426.
- [7] R. Kovordanyi, C. Roy, "Cyclone Track Forecasting Based on Satellite Images Using Artificial Neural Networks," *International Journal of Photogrammetry and Remote Sensing*, In press, 2009.
- [8] B. Aisa, B. Mingus, and R. O'Reilly, "The emergent neural modeling system," *Neural Networks: The Official Journal of the International Neural Network Society*, vol. 21, Oct. 2008, pp. 1146-52.
- [9] R.C. O'Reilly and Y. Munakata, *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*, MIT Press, 2000.
- [10] R. Kovordanyi, M. Saifullah, C. Roy, "Improved generalization in a recurrent hierarchical network using a mixture of error-driven and Hebbian learning," *ICIP2009 IEEE International Conference on Image Processing*, submitted.