COGNITIVE SCIENCE **19,** 563–574 (1995)

Correlations Between Input and Output Units in Neural Networks

NORMAN D. COOK
Kansai University, Osaka, Japan

Correlation analyses of recent back-propagation neural networks show that network results are due to imbalances in stimulus input. Conclusions concerning the effects of receptive field size, hemispheric specialization, and other issues of relevance to psychology cannot therefore be drawn until the dominating effects of low-level correlations are removed. Statistical techniques for evaluating the stimulus materials for neural networks are introduced.

1. INTRODUCTION

Neural network computations provide a second empirical technique for exploring questions concerning how organic brains function. Their potential importance for human psychology and artificial intelligence is widely appreciated, but meaningful results can be obtained only if as much care is taken in designing neural networks as is normally taken for psychological experimentation. Used uncritically to produce results with a surface similarity to human psychological data, neural nets can be worse than wasted effort, because they suggest that there is empirical evidence where none exists and imply a computational precision that may be illusory.

Kosslyn and colleagues have reported several experimental studies supporting a theoretical view (Kosslyn, 1987) on the specializations of the cerebral hemispheres in man. The empirical support for the theoretical position is mixed: An insignificant trend in the predicted direction was found six times and the reverse trend once (reviewed in Kosslyn, Chabris, Marsolek,

My thanks go to Hideki Kawahara, Yoh'ichi Tohkura, and the members of the Human Information Processing Research Laboratories of ATR (Kyoto), where most of this work was carried out.

Correspondence and requests for reprints should be sent to Norman D. Cook, Faculty of Informatics, Kansai University, Takatsuki, Osaka 569, Japan. E-mail: <cook@res.kutc.kansai-u.ac.jp>

& Koenig, 1992). A sign test then suggested that the overall results approached significance ($p = .06$). Such findings from a variety of experimental situations are sometimes referred to as "converging" evidence and viewed optimistically as suggesting diverse support. A more cautious interpretation would be that conclusions cannot be drawn from many weak lines of evidence—an error of statistical interpretation traditionally referred to as the "fagot fallacy" (Skrabanek & McCormick, 1990)—a *fagot* being a bundle of twigs tied together and having more apparent than actual substance.

The experimental situation is thus somewhat uncertain, but a second line of evidence has been offered on the basis of the results of neural network simulations (Kosslyn et al., 1992). Colleagues and I have previously shown that those results can be explained solely on the basis of correlation coefficients (Cook, Früh, & Landis, 1995), but Jacobs and Kosslyn (1994) have nonetheless recently published similar networks with similar correlational problems. In order that such mistakes can be avoided in the future, details of an appropriate analytical technique are provided in this article.

Qualitatively, the basic criticism is that accidental strong correlations between input and output units in a neural network can completely dominate network performance and prevent the network from accomplishing anything of interest. As a consequence, although variables of potential relevance to psychology may be the intended focus of research, the obtained differences in performance in fact reflect only accidental differences in the correlations that are present in the input stimuli for the various tasks. Statistical analysis of the stimulus material alone is then sufficient to explain the results, which cannot therefore be considered as a kind of "empirical" evidence in support of theories of brain functions.

2. THE ARCHITECTURE OF THE JACOBS AND KOSSLYN NETWORKS

The simulations reported recently by Jacobs and Kosslyn (1994) were based upon a three-layer back-propagation network with an additional retinal layer prior to the input layer, as is depicted in Figure 1. Because the degree of activation of Layer Two units was based upon the summation of activity in a variable number of Layer One ("retinal") units, as defined by a receptive field of given size and shape, the focus of the correlational analysis is on the relationship between Layer One and Layer Four activity.

There were 95 input units in the retinal layer of all four simulations, with two output units in the first three simulations and eight output units in the last. In Simulation 1, the output units signified the position of an input shape as lying above or below a central bar (in the terminology of Jacobs & Kosslyn, 1994, a "catagorical" task for deciding where the shape was located, "Where/Cat"). In Simulation 2, the outputs signified that the shape was a variation of one of two prototype shapes, a T shape or an upside-down ⊥

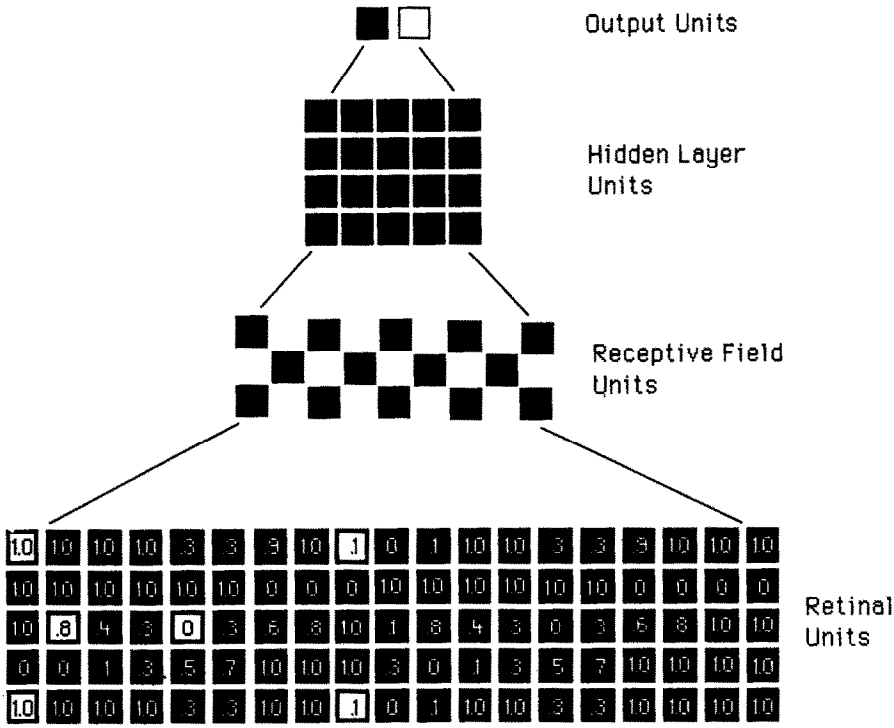


Figure 1. The four-layered back-propagation network used in the first three Jacobs and Kosslyn (1994) simulations. They were intended to show the relative ease of learning the left/right or near/far position of the T/L-like shapes, relative to the central "bar." (In the original work, the retinal units were displayed vertically and the T/L-like shapes were shown above or below the central bar represented as two firing units near the middle of the retinal layer. Here the T is shown lying on its side to the left or right of the central bar, but the patterns of pixel excitation are identical to those used by Jacobs and Kosslyn, 1994). The numbers printed on each input unit are the absolute values of the r_{phi} coefficients for the What/Cat simulation. Note the large number of $r_{phi}=1.0$ units, signifying the presence of individual pixels that, when activated, are alone unambiguous indications of the correct response, regardless of the configuration of the T or the location of the bar.

shape ("What/Cat"). In Simulation 3, the outputs signified the coordinate location of the shape as near to or far from the central bar ("Where/Coo"). Finally, in Simulation 4, eight separate output units signified eight distinct exemplars of the two prototype shapes ("What/Coo").

As is evident from Figure 1, the network was designed to have a two-dimensional "visual field" within which objects could be represented as patterns of on/off pixels. Similar stimuli can of course be shown to human participants, and responses and response latencies can be recorded for comparison with neural network performance on similar tasks. The theoretical

situation is therefore quite attractive in so far as network architecture and parameters can be manipulated until the human performance is simulated, and then implications can be drawn about information processing in living brains. The question that must be addressed, however, is whether the network performs the task using information of the kind that biological systems use. If we are interested in the perception of visual patterns, it is essential that the network deals with information of that kind as a consequence of the design of the neural network and the chosen stimuli. For example, if the task requires human participants to evaluate a combination of color and texture information, but a network successfully performs a seemingly analogous task using lower order information (e.g., absolute position in the visual field), then network results will have no relevance to human information processing (Minsky & Papert, 1969). For this reason, it is essential to determine the statistical order of the information inherent to the stimulus materials so that modelers can avoid the possibility that inappropriate stimuli have allowed the artificial network to perform the task using lower order information than that used by biological systems.

3. THE PHI CORRELATION COEFFICIENT

Because the activity of units at both the retinal and the output layers of many neural networks is binary (0/1), the proper statistic to use in examining their correlated firing is the so-called "phi coefficient" (Carroll, 1961; Cureton, 1959). Phi is designed to show the strength of association between two sets of dichotomous variables. It is defined as follows:

$$\text{phi} = \frac{bc - ad}{\text{sqrt}[a+b)(c+d)(a+c)(b+d)]},$$

where *a, b, c,* and *d* are the numbers of 0/0, 0/1, 1/0, 1/1 combinations of input/output unit activity, as defined in the following contingency table.

		output y: 0	output y: 1
input x	0	a	b
	1	c	d

To obtain a phi correlation coefficient, r_{phi}, that is comparable to the product-moment correlation coefficient, phi is divided by phi_{max}:

$$r_{phi} = phi/phi_{max}, \text{ where } phi_{max} = sqrt[(p_x/q_x)(q_y/p_y)],$$

with $p_x = (a+b)/n$, $q_x = (c+d)/n$, $p_y = (a+c)/n$, and $q_y = (b+d)/n$, and the choice of the assignment of 0 and 1 to the outputs is made such that $p_x \leq q_x$ and $q_y \leq p_y$.

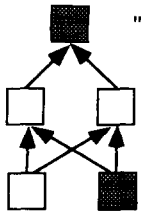
So doing, r_{phi} can have values between -1.0 and 1.0, but the sign of the coefficient is arbitrary since the assignment of the meaning of 0 and 1 at the output is entirely arbitrary. With or without a sign, the significance of r_{phi} lies in its capacity to reflect the strength of association between each input/output pair, a relationship that is not apparent using the product-moment correlation coefficient (suitable only when the variables represent continuous rather than dichotomous values); this is discussed by, e.g., Kurtz & Mayo (1979).

For the purpose of illustration, let us examine the input/output correlations found in the classical XOR network (Figure 2a) and in two slightly more complex nets (Figure 2b and c). Note that the size and complexity of the hidden layer architecture is irrelevant: of interest is only the nature of the stimulus materials. The Truth Table for the XOR problem is shown in Figure 2a and provides the numerical values for calculating the phi coefficients. It is found that the correlations between the input and output units are zero for both input units. In other words, the firing of either input unit is associated as frequently with an "off" output as with an "on" output. If the network is expanded and trained to become a "two-inputs on" detector (Figure 2b), the correlation coefficients for all input/output pairs remain zero. The neural network performs well and there is no dominating influence of the activity of one or several input units, as indicated by the fact that all r_{phi} values are zero.

Now consider a network trained to detect if either of the end units in the input layer has fired (Figure 2c). There are as many synaptic connections in this network as that in Figure 2b, but the input/output relations have been drastically simplified because the firing of either of the end units in the input layer indicates that the output should be "on." This corresponds to r_{phi} coefficients of 1.0 for both end units, whereas all other input units have r_{phi} values of zero because their activity is irrelevant to the correct response. These statistics alone are clear indication that the neural net can solve the task using only first-order correlations, and that higher-order statistics are not required. Instead of requiring the network to develop synaptic weights that reflect complex configurations of input unit firing, the net learns only to rely on the activity of either end unit and ignore the other units.

4. CORRELATIONS IN THE JACOBS AND KOSSLYN STUDY

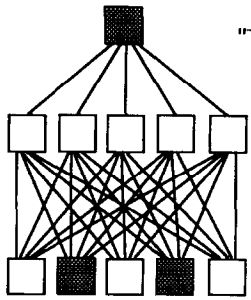
In the somewhat more complex networks used in the Jacobs and Kosslyn (1994) simulations, a nonrandom set of 192 input stimuli were selected from among the 2^{95} possible patterns in the input array. What that meant in prac-



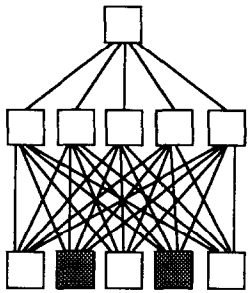
In	Out
0 0	0
0 1	1
1 0	1
1 1	0

phis: 0 0

(a)



(b)



(c)

In	Out (b)	Out (c)
0 0 0 0 0	0	0
0 0 0 0 1	0	1
0 0 0 1 0	0	0
0 0 0 1 1	1	1
0 0 1 0 0	0	0
0 0 1 0 1	1	1
0 0 1 1 0	1	0
0 0 1 1 1	0	1
0 1 0 0 0	0	0
0 1 0 0 1	1	1
0 1 0 1 0	1	0
0 1 0 1 1	0	1
0 1 1 0 0	1	0
0 1 1 0 1	0	1
0 1 1 1 0	0	0
0 1 1 1 1	0	1
1 0 0 0 0	0	1
1 0 0 0 1	1	1
1 0 0 1 0	1	1
1 0 0 1 1	0	1
1 0 1 0 0	1	1
1 0 1 0 1	0	1
1 0 1 1 0	0	1
1 0 1 1 1	0	1
1 1 0 0 0	1	1
1 1 0 0 1	0	1
1 1 0 1 0	0	1
1 1 0 1 1	0	1
1 1 1 0 0	0	1
1 1 1 0 1	0	1
1 1 1 1 0	0	1
1 1 1 1 1	0	1

"Two on" phis: 0 0 0 0 0

"Ends on" phis: 1 0 0 0 1

Figure 2. Simple neural nets and their Truth Tables. In the classical XOR problem (a), it is clear that both input units participate to an equal extent in producing firing of the output unit, and both have r_{phi} coefficients of 0. In the slightly more complex case of a network that fires only when two inputs are excited (b), input/output r_{phi} values are again 0. In network (c), however, the "end units" in the input layer are strongly correlated with output firing, and have r_{phi} values of 1.0 (even if only a random subset of stimulus patterns are used).

tice was that the firing of certain input units was strongly associated with certain outputs. This can be seen in the r_{phi} values for each input unit in the What/Cat network (Figure 1) and the pattern of r_{phi} values $= \pm 1.0$ in the first three simulations (Figure 3).

The mean r_{phi} coefficients for the four different simulations are shown in the top row of Table 1. Noteworthy is the fact that, when small receptive fields were used, the network performance was inversely proportional to the average correlation between input and output units. In other words, the networks found tasks easy if the mean correlation for all input/output pairs was high and, conversely, found them difficult if the correlation was low. The central arguement is that the proportionalities seen in Table 1 are not coincidental, but, on the contrary, are indication that the network performance was a direct function of input/output correlations.

It should be noted that the r_{phi} summary of a network, as shown in Table 1, is a function of the input and output vectors given to the network and reflects the structure inherent to the stimulus materials, regardless of various subtleties of network architecture, hidden layer connectivity, learning rules, and so forth. Small changes in network structure and dynamics, as well as in the criteria used to evaluate network performance, will have small effects upon the precision of the correspondence between the statistics of the stimulus materials and actual performance of the network, but the statistics on input/output relations remain an accurate reflection of the difficulty of the task that has been given to the network.

Table 1 presents the correlational data relevant to an a priori argument suggesting that the performance of nets on tasks with such large differences in the statistics of input/output relations cannot be meaningfully compared. Neural networks containing such correlations, which are then used in simulations, will invariably perform the tasks by utilizing this correlational information, and the stimulation "results" will be nothing more than restatement of the correlational structure of the input stimuli. This can always be demonstrated for individual nets by examining the synaptic weights that emerge after training.

5. RECEPTIVE FIELDS

The main theme of the Jacobs and Kosslyn (1994) study concerned the effects of changes in receptive field size on the performance of these tasks, so let us consider the meaning of receptive fields in light of the input/output correlations inherent to the networks. First of all, it is clear that the performance of networks without the receptive field layer (Layer 2) is dominated by the strength of input/output correlations per task. Given a particular set of input stimuli and desired outputs, the correlational structure, as summarized in Table 1, is the baseline from which net performance can be altered by

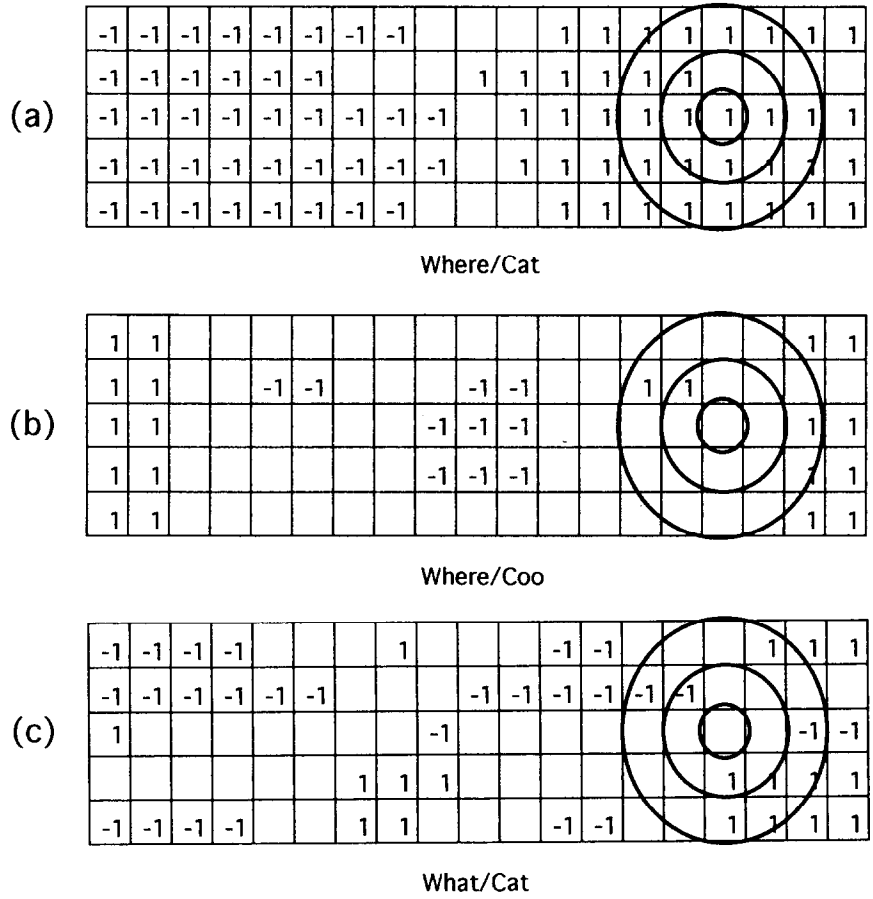


Figure 3. The correlational structure of the first three Jacobs and Kosslyn (1994) stimulations. The input layer of each simulation is shown, with r_{phi} coefficients of ± 1.0 as indicated. Input units that do not contain unambiguous information concerning the correct output ($r_{phi} \neq \pm 1.0$) are empty. The circles in each diagram show three levels of receptive field. In (a), most receptive fields include units with consistent r_{phi} values. In (b), a gradual increase in receptive field implies the inclusion of more and more units with $r_{phi} = \pm 1.0$. Network performance will gradually improve until contradictory information is included within each receptive field. In (c), an increase in receptive field size immediately leads to the inclusion of contradictory information (some units with $r_{phi} = +1.0$ and other units with $r_{phi} = -1.0$). Network performance suffers correspondingly.

manipulating various network parameters. By increasing receptive field size (whether implemented by a second layer with fewer units than the retinal layer or by simply connecting retinal units to a larger number of second layer units), the number of retinal units contributing to the activation of second layer units will be increased. If a large number of strong correlations between the retinal layer and the output layer is not found in the correlational

TABLE 1
Correlations Between Input and Output Units in the Jacobs and Kosslyn (1994) Simulations

	Where/Cat Simulation (up vs. down)	What/Cat Simulation (T vs. ⊥)	Where/Coo Simulation (near vs. far)	What/Coo Simulation (shapes 1–8)
M of the absolute values of phi correlation coefficients for the 95 input units	0.851	0.627	0.486	0.139
N of input units with phi = ±1.0 with the output units (out of 95 input units)	80	45	30	~2 (per shape)
Ease of learning [defined as the product of the mean phi and the number of in input units with $r_{phi} = \pm 1.0$]	68	28	15	~1
Epochs[a] until successful learning (rf=small) (Jacobs & Kosslyn, 1994)	3	22	100	260

[a] Each epoch was 192 learning trials.

analysis, then the effects of receptive field changes on network performance will be complex and the results potentially interesting. In the Jacobs and Kosslyn study, however, there were distinct patterns of $r_{phi} = 1.0$ coefficients in the retinal layers for the different tasks (Figure 3), and those patterns alone explain the effects of changing receptive field sizes.

As seen in Figure 3a, the r_{phi} coefficients for the Where/Cat task were divided neatly into two distinct regions. By increasing the size of the receptive fields, the second layer units simply received information from a greater number of retinal units, all of which consistently indicated that the T/⊥ shape was above or below the bar. An increase in the number of input units therefore brings about no change in network performance (as shown in Figure 4 of Jacobs and Kosslyn, 1994). Only when the receptive field is enlarged to such an extent that most receptive field units receive contradictory information will there be a decrement in performance.

In Figure 3b, it can be seen that the pattern of r_{phi} coefficients found in the Where/Coo task is such that an increase in receptive field size will involve a greater number of retinal units with ± 1.0 r_{phi} values—units which contain unambiguous information concerning the near/far location of the shapes from the central bar. As a consequence, starting from the relatively ambiguous situation of having few r_{phi} values of ± 1.0, there will be a gradual improvement in performance as receptive field size increases to include $r_{phi} = \pm 1.0$ units. Clearly, learning the correct response to patterns located very far

from or very close to the central bar will remain easy because of the presence of ± 1.0 r_{phi} values, but the response to in-between shapes will gradually improve as the unambiguous information of the $r_{phi} = \pm 1.0$ units can be exploited.

In Figure 3c, the more complex r_{phi} pattern of the retinal layer in the What/Cat task is shown. Unlike the two previous cases, it can be seen that some retinal units contain unambiguous information ($r_{phi} = \pm 1.0$) that is directly contrary to the information in neighboring retinal units. This means that, as the receptive field size is increased, a greater number of second layer units will receive such contradictory information. The net is thus forced to make judgments, not on the basis of absolute ($r_{phi} = \pm 1.0$) information, but rather on the basis of combinations of retinal unit activity. Compared with the learning process when r_{phi} correlations of ± 1.0 alone can be exploited, learning with units that contain only ambiguous information is more difficult, as indicated by the large increase in the number of learning cycles required for success (Figure 4 in Jacobs & Kosslyn, 1994).

Only the fourth simulation contained input stimuli that did not contain dominating input/output correlations and, significantly, network performance was approximately 100-, 10-, and 3-fold worse than the three networks that did contain strong correlations. Unlike the first three simulations, the changes in performance due to changes in receptive field size in the fourth simulation cannot, therefore, be explained solely on the basis of the effects just discussed. The significance of those effects relative to the other simulations cannot be evaluated, however, unless the correlational problems of the first three simulations are eliminated.

6. COMPARISONS AMONG NETWORKS TESTED WITH DIFFERENT STIMULI

The correlations listed in Table 1 are indication that the performance of the Jacob and Kosslyn networks was determined by differences in the magnitude of input/output correlations among the different tasks. This conclusion is the same one drawn previously (Cook et al., 1995) with regard to similar simulations reported by Kosslyn et al. (1992). In that study as well, neural nets were used in support of hypotheses concerning hemispheric specialization, receptive field size, and visual information processing, but similar imbalances in the stimulus materials were present. The principal difference between the two sets of simulations lies in the fact that, in the former case, four different sets of stimuli were paired with output responses in four different networks, whereas, in the latter case, the same stimuli were used in all cases, and only the required outputs differed.

Because identical stimuli had been used for all four tasks in the Jacobs and Kosslyn study, a comparison of the mean r_{phi} coefficients for the four tasks suffices to show the performance implicit to the input stimuli. In the

TABLE 2
Correlations Between Input and Output Units in the Kosslyn et al. (1992) Study

Simulation Task	EasyCat	EasyCoord	DiffCat	DiffCoord
N of input units with $r_{phi} = \pm 1.0$ with the output units (out of 28)	18	16	10	4
N of stimuli in which an input unit with $r_{phi} = \pm 1.0$ was activated (out of 40)	32	24	16	20
Ease of learning (defined as the product of the number of input units with $r_{phi} = \pm 1.0$ times the number of such stimuli)	576	384	160	80
M error after 30 learning epochs,[a] as reported by Kosslyn et al. (1992)	.016	.034	.069	.087

[a] Each epoch was 40 learning trials.

Kosslyn et al. (1992) study, however, different stimuli were used for the different tasks: 40 relatively easy stimuli or 40 relatively difficult stimuli were used in each of four separate nets. Individual input units were involved with different frequencies for the different tasks. Therefore, as shown in Table 2, a better measure of network performance than the mean correlation is the product of the number of input units with r_{phi} values of ± 1.0 and the number of stimuli in which $r_{phi} = \pm 1.0$ input units were activated. A clear inverse relation between the actual performance and this product is again seen. The fact that these differences in the correlational structure reflect the network performance indicates that the chosen labels of *easy* and *hard, categorical* and *coordinate* tasks are less appropriate expressions of what the networks were actually doing than simply to say: Network performance is determined by the strength of input/output correlations.

7. CONCLUSIONS

Without the aid of correlation coefficients, Scalettar and Zee (1988) have previously shown empirically that three-layer back-propagation networks are unsuitable for detecting the geometrical configuration of input units in three-layer back-propagation networks. By adding a retinal layer prior to the input layer, some geometrical information can in principle be learned by such networks, but there is still no guarantee that those features, which are to the human observer the salient features of the stimuli, will be the

information that the neural net uses to obtain the correct output. On the contrary, neural nets are generally "smart" enough to exploit first-order correlations between input and output, and to ignore higher-order configurational information unless lower-order correlations will not suffice to attain correct performance.

Whether a neural network has actually performed a task in a manner similar to living brains is never an easy question, but it is sometimes possible to demonstrate that a neural net has performed in a computationally trivial way. A necessary, but not sufficient, test of a neural net is to compute the strength of input/output correlations during the design of the neural network and prior to testing its performance. When individual pixels of the input layer are strongly correlated with output (especially $r_{phi} = \pm 1.0$), a different set of input stimuli should be created so that, for every input stimulus, the net is required to obtain the correct output on the basis of the firing of combinations of input units, for example, on the basis of the geometrical configuration of the stimuli. Even when correlations of 1.0 are not found, if large differences in the absolute mean of input/output r_{phi} values are found for different tasks (Tables 1 and 2), then differences in network performance will be due to those correlational differences, and not due to differences in various other network parameters. Incorrect conclusions will be drawn if attention is focused on such parameters while ignoring the dominant input/output correlations.

REFERENCES

Carroll, J.B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372.

Cook, N.D., Früh, H., & Landis, T. (1995). The cerebral hemispheres and neural network simulations: Design considerations: *Journal of Experimental Psychology: Human Perception and Performance, 21,* 410–421.

Cureton, E.E. (1959). Note on phi/phi(max). *Psychometrika, 24,* 89–91.

Jacobs, R.A., & Kosslyn, S.M. (1994). Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representations. *Cognitive Science, 18,* 361–386.

Kosslyn, S.M. (1987). Seeing and imagining in the cerebral hemispheres. A computational approach. *Psychological Review, 94,* 148–175.

Kosslyn, S.M., Chabris, C.F., Marsolek, C.J., & Koenig, O. (1992). Categorical versus coordinate spatial relations: Computational analyses and computer simulations. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 562–577.

Kurtz, A.K., & Mayo, S.T. (1979). *Statistical methods in education and psychology.* New York: Springer.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

Scalettar, R., & Zee, A. (1988). Perception of left and right by a feed forward net. Biological *Cybernetics, 58,* 193–201.

Skrabanek, P., & McCormick, J. (1990). *Follies and fallacies in medicine.* Buffalo, NY: Prometheus.