COGNITIVE SCIENCE **19,** 575–579 (1995)

Neural Network Models as Evidence for Different Types of Visual Representations

STEPHEN M. KOSSLYN
CHRISTOPHER F. CHABRIS
DAVID P. BAKER
Harvard University

Cook (1995) criticizes the work of Jacobs and Kosslyn (1994) on spatial relations, shape representations, and receptive fields in neural network models on the grounds that first-order correlations between input and output unit activities can explain the results. We reply briefly to Cook's arguments here (and in Kosslyn, Chabris, Marsolek, Jacobs, & Koenig, 1995) and discuss how new simulations can confirm the importance of receptive field size as a crucial variable in the encoding of categorical and coordinate spatial relations and the corresponding shape representations; such simulations would testify to the computational distinction between the different types of representations.

Does the brain encode different types of spatial relations between objects or parts of objects in a visual scene, and are different types of shape representations associated with corresponding types of spatial relations representations? Over the past several years, our group has developed and tested a theory that answers these questions by focusing on the distinction between *categorical* and *coordinate* spatial relations representations. Categorical spatial relations treat as equivalent a wide range of positions that share a defining characteristic with respect to a reference object; for example, one object can be said to be above, to the right of, or under another and yet be located at any point in a large region of space. Categorical spatial relations can be used to describe the shape of a multipart object in a manner that is preserved when the object moves or changes posture. In contrast, coordinate spatial relations specify metric spatial properties and are useful primarily for guiding movement. Moreover, we have hypothesized that representations of prototypical shapes normally are associated with categorical spatial

This research was supported by NIH grant no. NS 27950, ONR grant no. N00014-94-1-0180, an NSF Graduate Fellowship, and an ONR Graduate Fellowship. We thank Robert Jacobs for invaluable comments, criticism, and inspiration.

Correspondence and requests for reprints should be sent to Stephen M. Kosslyn, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138.

relations, and representations of specific exemplars are associated with coordiate spatial relations (see Jacobs & Kosslyn, 1994; Kosslyn, 1987, 1994).

Several independent sources of evidence based on experiments with human participants support this distinction between types of spatial relations representations (for reviews, see Kosslyn, Chabris, Marsolek, Jacobs, & Koenig, 1995; Kosslyn, 1994; for examples, see Hellige & Michimata, 1989; Laeng, 1994; Laeng & Peters, 1995). In addition, neural network modeling (Kosslyn, Chabris, Marsolek, & Koenig, 1992) suggests that it is computationally advantageous to separate processing of the two types of spatial relations and that differences in receptive field size can account for this phenomenon. Further computational work by Jacobs and Kosslyn (1994) replicated and extended the Kosslyn et al. findings, showing that differences in receptive field size can be associated not only with different spatial relations representations, but also with the corresponding types of shape representations.

Cook and his colleagues have criticized these simulations on several grounds. In particular, discussing the findings of Kosslyn et al. (1992), Cook, Früh, and Landis (1995) argued that neural network models do not process "spatial" information, and that even if they do, the Kosslyn et al. models were flawed because their training patterns contained so-called "definitive information," which made it possible for the networks to encode spatial relations without developing generalizable representations. Cook (1995) has now focused on the issue of input–output correlations with respect to the simulations reported by Jacobs and Kosslyn (1994).

In response to Cook et al. (1995), Kosslyn et al. (1995) addressed this last criticism in detail (see pp. 427–429), so we will not repeat that discussion here. One key point was made that such correlations might explain some aspects of the results, but not others. Indeed, Kosslyn et al. (1992) suggested that categorical spatial relations are encoded best when the input space can be carved into discrete bins such that the presence of a stimulus in a bin signals the existence of a specific spatial relation; in contrast, coordinate spatial relations are encoded using course coding, which will not profit from such information. However, our agreement that input–output correlations might explain some aspects of the results is not a concession that the models were methodologically flawed. In some cases such correlations can be more than just artifacts of poor experimental design: Successful perceptual computation relies on exploiting regularities in the input, and the presence of low-order regularities in any simulation or experiment (e.g., Kosslyn, Koenig, Barrett, Cave, Tang, & Gabrieli, 1989) does not necessarily invalidate its results. In some situations such correlations are accurate reflections of the perceptual problem of interest. In the final portion of this article we address the correlation issue directly by briefly summarizing new simulations that will definitively settle the issue; first we would like to point out some errors or misrepresentations made by Cook (1995).

Cook claims that the theoretical distinction between categorical and coordinate spatial relations made by Kosslyn (1987) is ill-supported by experimental results. Cook writes, "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" (p. 563), referring to a meta-analysis by Kosslyn et al. (1992). But Kosslyn et al. were discussing only the left-hemisphere advantage for categorical spatial relations, not the overall task-by-hemisphere interaction, which was significant in each of the experiments surveyed and is also the most relevant test of the theory (see Hellige, 1983). Later, Cook argues that "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.'" This is mystifying, because it contradicts the entire concept of meta-analysis, that independent effect sizes and probabilities may be statistically combined to refine and strengthen conclusions, which is exactly what Kosslyn et al. (1992) were doing. We will not attempt to defend the well-established logic of meta-analysis here, and refer the interested reader to Rosenthal (1991, 1994) for entree to this vast literature.

Later, Cook insists that network models should contain no first-order asssociations between input- and output-unit activities. He uses the XOR problem and its generalized forms as examples. However, XOR is notoriously difficult for networks (and human beings) to represent and solve properly. Mandating that all problems to be studied with neural network modeling be versions of generalized XOR imposes a peculiar constraint because it rejects a priori the possibility that the brain uses low-order associations to help solve complex problems heuristically rather than with "geometric principles" or other abstractions. Following this rule may paradoxically force network models to be less brain-like in some respects.[1]

The focus of Cook's critique rests on the possibility that our previous results are entirely an artifact of the fact that some input units carried more information than others about the particular shape or spatial relation in the input array.[2] This possibility can be addressed directly with new simulations. In particular, a critical feature of the theory underlying the models of Jacobs and Kosslyn (1994) and Kosslyn et al. (1992) is that categorical spatial relations are more effectively encoded via input units that have relatively

[1] It is also worth noting Cook's claim (p. 573) that in Kosslyn et al. (1992), "different stimuli were used for the different tasks." What he is really referring to, however, is Study 2 of Kosslyn et al. (1992), not the entire article, since Studies 1, 3, and 4 did in fact use identical stimuli for both the categorical and coordinate judgments.

[2] Note however that, as Cook admits, the phi/phi_{max} correlations never explained the pattern of results found by Jacobs and Kosslyn (1994) on the exemplar shape encoding task.

small receptive fields (which facilitate dividing space into "bins"),[3] whereas coordinate spatial relations are more effectively encoded via input units that have relatively large, overlapping receptive fields (which facilitate coarse coding). This can be tested directly by examining the receptive field effects found by Kosslyn et al. (1992) and by Jacobs and Kosslyn (1994) when input-output correlations in the networks have been eliminated.

Specifically, it seems that the issue can be settled definitively if models are constructed with the following characteristics. By creating a larger training set than Kosslyn et al. (1992) used, allowing the bar and stimulus (e.g., a dot, as in the models of Kosslyn et al., 1992) locations to range over the entire input array, and omitting certain patterns in which the stimulus appears near the edge of the array, the overall phi/phi_{max} correlations can be reduced greatly. More importantly, this procedure can create a large central region of the input array where those correlations are zero (in other words, a region within which the state of any given output unit cannot be predicted solely by the state of any single input unit).

If large receptive fields facilitate encoding coordinate spatial relations and specific shape exemplars more than categorical spatial relations and shape prototypes, as the results and analyses of Kosslyn et al. (1992), Jacobs and Kosslyn (1994), and Kosslyn et al. (1995) suggest, then an interaction between task and receptive field size should be found for both the entire input array and the uncorrelated central region (assuming appropriate receptive field sizes are tested). Moreover, this relation should be found even with stimuli that were not part of the training set. No account based on first-order input-output correlations would be able to explain such findings. We are currently conducting these simulations, and our preliminary results clearly support the Kosslyn et al. (1992) theory of receptive field sizes and spatial relations: In general, networks with input filtered through large Gaussian receptive fields perform the coordinate spatial relations task better than the categorical spatial relations task, an effect not found when the receptive fields are small.[4] If encoding the two types of spatial relations does not involve distinct computations, why—in the absence of an explanation in terms of first-order correlations—should the variable of receptive field size affect them differently? (See Hellige, 1983, for the same point made with respect to studies of hemispheric specialization.)

Cook has performed a service by pointing out a possible alternative account for our previous simulation results. It is now time to design and build new models that will resolve the issue.

---

[3] Indeed, the notion of "bins" for categorical spatial relations is compatible with a computational mechanism proposed by Logan and Sadler (in press; cf. Kosslyn et al., 1995, p. 424) in which a spatial template is centered on a reference point and "definitive information" is then used to determine the relation between that point and a target object.

[4] A full report of this work is in preparation.

REFERENCES

Cook, N.D. (1995). Correlations between input and output units in neural networks. *Cognitive Science, 19,* 563–574.

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,* 410–422.

Hellige, J.B. (1983). Hemisphere × task interaction and the study of laterality. In J.B. Hellige (Ed.), *Cerebral hemisphere asymmetry: Method, theory, and application* (pp. 441–443). New York: Praeger.

Hellige, J.B., & Michimata, C. (1989). Categorization versus distance: Hemispheric differences for processing spatial information. *Memory & Cognition, 17,* 770–776.

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. (1994). *Image and brain: The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Kosslyn, S.M., Chabris, C.F., Marsolek, C.J., Jacobs, R.A., & Koenig, O. (1995). On computational evidence for different types of spatial relations encoding: Reply to Cook et al. (1995). *Journal of Experimental Psychology: Human Perception and Performance, 21,* 423–431.

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.

Kosslyn, S.M., Koenig, O., Barrett, A., Cave, C.B., Tang, J., & Gabrieli, J.D.E. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 723–735.

Laeng, B. (1994). Lateralization of categorical and coordinate spatial functions: A study of unilateral stroke patients. *Journal of Cognitive Neuroscience, 6,* 189–203.

Laeng, B., & Peters, M. (1995). Cerebral lateralization for the processing of spatial coordinates and categories in left- and right-handers. *Neuropsychologia, 33,* 421–439.

Logan, G.D., & Sadler, D.D. (in press). A computational analysis of the apprehension of spatial relations. In M. Peterson & P. Bloom (Eds.), *Language and space.* Cambridge, MA: MIT Press.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science, 5,* 127–134.