Understanding as generalization not just representation*

Steven Phillips

Department of Computer Science,
The University of Queensland, QLD 4072 Australia. Email: stevep@cs.uq.oz.au

Halford and Wilson define understanding of a concept in terms of representing the relation to which that concept is equivalent. A model is said to have represented a relation when it performs the correct input-output mappings for all functions implicated by that relation. For example, in the balance scale domain, the concept of balance is a relation between two weight and two distance variables and a state variable (which has the possible values: balance, tip-left and tip-right). The concept of balance implicates a number of questions which can be used to evaluate a child's understanding of the concept.

For example, the *balance state* question is a function from the two weights and two distance variables to the state variable. The *missing weight* question is a function from two distance variables, a weight variable, and a state variable to a weight variable. The *missing distance* question is a function from two weight variables, a distance variable and a state variable to a weight variable. More generally, from any combination of four variables, the fifth variable can be predicted.

Halford and Wilson argue that just as a child's understanding of the concept balance is evaluated on a battery of implicated questions, so too can a model of that concept be evaluated by its performance on a number of implicated functions. That is, on the basis of whether or not the model performs the same input-output mappings. Their main point is that such a definition provides an adequate criterion against which candidate Connectionist models can be evaluated. With this definition the authors reject McClelland's feedforward network model on the basis that it cannot demonstrate adequate performance on all three questions (i.e., the model is incomplete).

The first point I want to make is that implicit in their definition is a notion of generalization (i.e., the capacity to take existing representations and apply them correctly to previously unexperienced situations). I suggest that by confounding these two issues (representation and generalization) some respondants to Halford's presentation have been lead to question the authors' justification for rejecting McClelland's feedforward network model.

The authors claim that McClelland's model cannot perform the additional missing weight and missing distance tasks. However, they do not state the basis for this claim. Clearly, McClelland's model cannot be rejected on the basis of the representational capacity of the feedforward network alone. Since multi-layer feedforward networks are universal function approximators (Hornik, Stinchcombe, & White, 1989) there is every reason to expect that given sufficient information and time the network could perform correctly on these additional tasks.

Presumably, however, the authors have some additional criterion in mind. I suspect that although they would agree that McClelland's model (suitably extended) *can* represent the required functions, to do so would require an extensive and unaccountable amount of (re)training. If it can be shown that such training is not available to (nor required by) children, then there is a basis for rejecting the feedforward network model. That is, on the basis of the model's generalization characteristics.

The question is, of course, how much training do children receive? Based on empirical evidence, Hadley (1993) presents an example of how generalization may be characterized in the linguistic domain.

*In *Collected papers from a Symposium on Connectionist Models and Psychology.* Tech. Report 289, The University of Queensland. A comment on Halford and Wilson's paper: How far do neural network models account for human reasoning?

He identifies generalization across syntactic position as a criterion against which Connectionist models can be evaluated. Connectionist models can then be differentiated on the basis of this generalization criterion. For example, feedforward networks without the assumption of weight tying cannot demonstrate generalization across syntactic position, yet recurrent networks in a limited case can demonstrate this form of generalization (Phillips, 1994).

Essentially, the issue is that although some models have the capacity to represent any function of interest they simply require too many training examples to be psychologically plausible. I suspect that it is on grounds of generalization that the authors want to reject McClelland's model. The point I want to make is that the degree of generalization considered as psychologically plausible must be made explicit; and in doing so, the authors would then have much stronger grounds for rejecting particular Connectionist models. That is, on the basis that these models make use of information either not available to or not required by children.

The issue of generalization brings me to my second point: if generalization is used as a criterion on which to accept or reject models then the tensor model that Halford and Wilson are proposing is incomplete. For, although it demonstrates how relational concepts may be represented, there is no story as to how these representations might arise from experience.

In Halford and Wilson's formulation, an appropriate arrangement of weights and connections are in place to represent relational concepts as tensor products. Presumably, however, these weights and connections were not always there otherwise this model could not account for the empirical fact that children below a certain developmental stage do not understand (in Halford and Wilson's sense) the concept of balance.

The two issues of representation and generalization place the authors in a dilemma. Explicitly, they want to distinguish between two models on the basis of representational capacity, yet both models (suitably extended) are capable of representing relations. Implicitly, I suspect they want to reject McClelland's model on grounds of generalization. However, if generalization is the distinguishing criterion then they are required to tell a story of how the *right* arrangement of weights and connections of a tensor model come into being; and that is a story not yet told.

Acknowledgements I would like to thank Julie Stewart and Danny Latimer for discussions on these issues, and Paul Bakker and Janet Wiles for helpful comments.

References

Hadley, R. F. (1993). Compositionality and systematicity in connectionist language learning. Technical Report CSS-IS TR 93-01, Simon Fraser University, BC: Burnaby.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Phillips, S. (1994). Systematicity and connectionism. In Tsoi, A. C., & Downs, T. (Eds.), *Proceedings of the Fifth Australian Conference on Neural Networks*, pp. 53–55. University of Queensland Electrical and Computer Engineering.