Limits of generalization: An error surface view

Steven Phillips
stevep@etl.go.jp
Information Science Division, Electrotechnical Laboratory, Tsukuba, Japan

Abstract: *Feedforward networks with shared weights transfer learning to isomorphic tasks, but the degree of transfer is not the same as humans [3]. As an addendum to [3], error surface plots clarify the problem: the global minimum for training cannot be constrained to coincide with the test set minimum, hence generalization cannot be guaranteed. Such networks are rejected as the mechanism for transfer in human cognition.*

Introduction

Apparent demonstrations of learning transfer over isomorphic tasks by feedforward networks [2] suggest that such networks provide a solution to the symbolic aspects of human cognition, without resort to the ungrounded notions of symbol systems. However, recent simulations showed that the degree of transfer exhibited by these networks is not the same as psychological experiments on human subjects [3]. As an addendum to [3], further analysis here clarifies *why* transfer is not possible with these networks. Essentially, the global minimum for the training set cannot be constrained to coincide with the global minimum for the test set in order to guarantee generalization.

Learning transfer: Klein Group

A series of psychological experiments [1] tested the capacity to transfer knowledge learned on one task to improve learning on isomorphic tasks. In one experiment, subjects are given a series of four task instances derived from the Klein Group:

Klein	x_1	x_2	x_3	x_4
H	x_2	x_1	x_4	x_3
V	x_4	x_3	x_2	x_1
D	x_3	x_4	x_1	x_2
N	x_1	x_2	x_3	x_4

where x_i are the states; "H" and "V" are the operations used in the experiments; and "D" and "N" are the other two operations in the Klein Group. When states are depicted as vertices of a square, operations appear as horizontal, vertical, diagonal and no state transitions, respectively. A task instance consists of four unique three-letter strings (states) crossed with two distinguishable shapes (operations) generating eight stimulus-response pairs. Given a string and shape, subjects must predict the response string. A new set of strings and shapes is randomly generated for each new task instance. The pertinent result for connectionist models is that by the fourth task instance, all subjects made perfect predictions (i.e., six out of eight) after seeing correct responses for the first two stimulus-response pairs. Perfect prediction is achieved by aligning task and structure elements identified from the first two stimulus-response pairs [3][1].

Feedforward networks (FFN)

A natural approach to learning transfer is to use shared weights to encode common structure and separate weights for task specific features[1] (Figure 1(a)). Learning of subsequent tasks is easier since some weights are already set to suitable values from previous tasks. With this type of network, Hinton [2] demonstrated transfer between isomorphic family trees in his *family-relations* task. Yet, the same type of network did not show the same degree of transfer as human subjects on the Klein Group experiments [3].

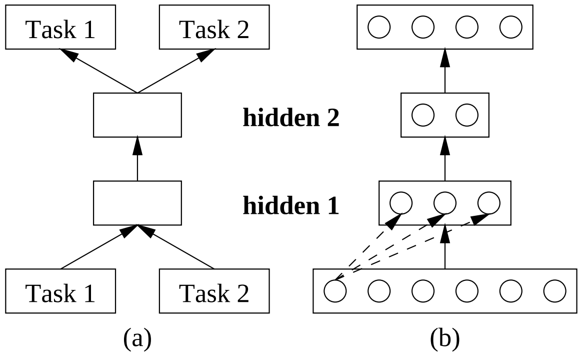


Figure 1: Feedforward network with different (a) and same (b) input/output units for each task.

Given the host of parameters accompanying backpropagation, negative learning results may simply be attributed to unsuitable parameter settings (e.g., large step sizes may lead to overstepping minima, etc). Rather than attempt some "reasonable" coverage of the parameter space, a potentially more informative approach is to plot the error surface of different regions of weight space to gain an understanding for the difficulty of the learning task.

The network was trained to correct response on all patterns in a task instance. Weights from one input unit to all first hidden layer units were reset (Figure

[1] Or, the same input-to-hidden and hidden-to-output weights reset for each new task instance (Figure 1(b)).

1(b), dashed lines). The network was retrained on seven patterns from the same task instance with all other weights fixed, constituting the minimum test for learning transfer between task instances [3].

Figure 2(a) shows the error surface for a network trained under this condition (seven patterns). The vertical z-axis indicates total error, as a function of two weights (x,y-axis). The flat region (grey area) is the global minimum for the training set. The error surface for the same two weights, but for all eight patterns (including the single test case) is shown in Figure 2(b). The global minimum (grey area) for this case is much smaller than for the training condition. Clearly, the training set does not constrain the error surface to be the same as for the test set. In other words, the network cannot guarantee generalization to even a single test case. Similar results where found for the other two weight pairs, and for other trials.

Insert Figure 2 here

Figure 2: Error surfaces for 7 (a) and 8 (b) patterns.

Enforcing generalization

As explained in [3], some degree of generalization can be enforced by restricting the number of possible hidden unit activation states, and using alternative activation functions. The minimum number of activation states needed in the second hidden layer is four (i.e., one for every possible response). Any fewer means that one state must be mapped to more than one output. All four states can be represented by a unique activation value for a single hidden unit. Since this organization no longer makes the states linearly separable by the output units, double threshold activation functions (e.g., Gaussian, pulse) must be used. Under these conditions, how many training examples are needed for generalization? The lower bound is four: one training pattern for each possible response to set the weights to each of the four output units. Yet, this lower bound is still significantly higher than human subjects, who required only two training patterns (averaged over 12 participants).

Discussion

The use of error surface plots clarifies the difficulty of the transfer task. The network was reduced to the smallest number of weights in order to maximize generalization, yet solve the task. However, the resulting error surface also permits solutions to the training set that do not contain the single test example. Under this condition, transfer of learning is not guaranteed.

The analysis emphasises two points raised in [3]: one theoretical, and the other meta-theoretical. The theoretical point pertains to the nature of cognitive architecture. Despite the capacity to demonstrate transfer and the intuitive appeal of the weight sharing technique, the computational properties of this network discount it as a mechanism for transfer in humans.

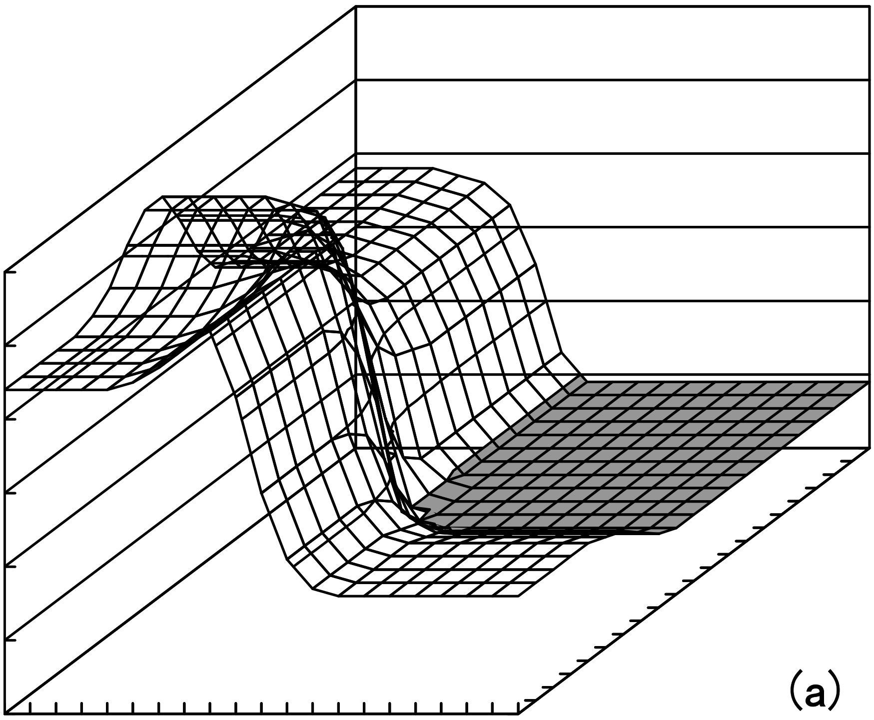
The meta-theoretical point is an objection to the claim that quantitative differences between (connectionist) models and human performance are just a matter of scaling or fine tuning of parameters. As argued in [3], attempting to support the same degree of transfer as human subjects leads to quite different sorts of networks: ones based on "one-shot" construction of representations (e.g., tensors, synchronous activation), rather than iterative learning. There are hard limits to generalization. And when these limits are surpassed it is cause enough to reject the model.

References

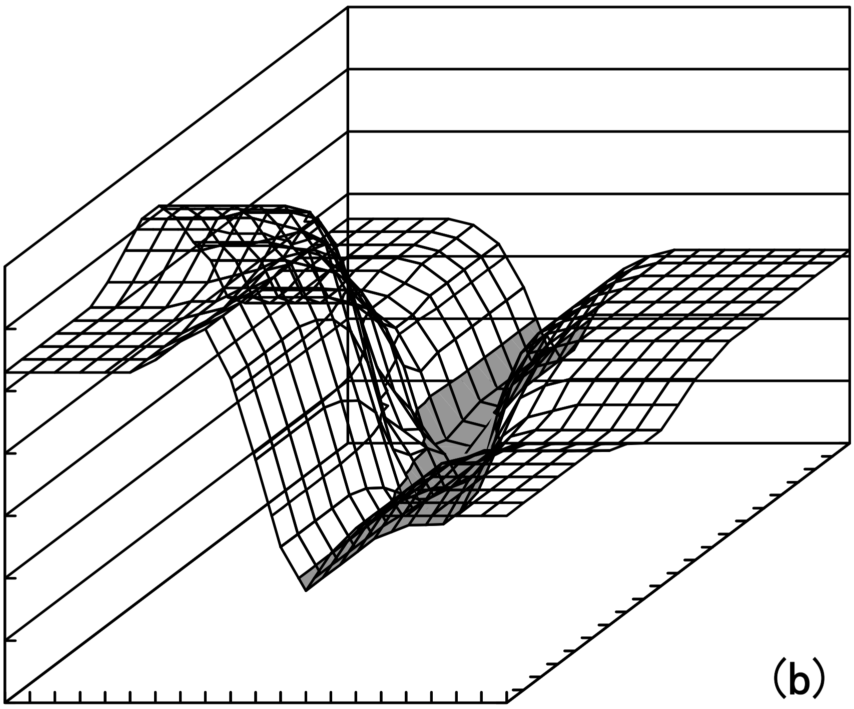
[1] G S Halford, J D Bain, and M T Maybery. Induction of relational schemas: Common processes in reasoning and learning set acquisition, submitted. http://www.psy.uq.edu.au/Department/Staff/gsh.

[2] G E Hinton. Mapping part-whole hierarchies in connectionist networks. *Artificial Intelligence*, 46(1-2):47–76, November 1990.

[3] S Phillips and G S Halford. Systematicity: Psychological evidence with connectionist implications. In M G Shafto and P Langley, editors, *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 614–619, 1997. http://www.etl.go.jp/etl/ninchi/stevep@etl.go.jp.



(a)



(b)