Very Rapid Induction of *General* Patterns

Robert F. Hadley
School of Computing Science
and Cognitive Science Program
Simon Fraser University
Burnaby, B.C., V5A 1S6
hadley@cs.sfu.ca

Abstract

Marcus (1998) and Phillips (2000) each have produced examples of human *generalizations* which, they argue, cannot be matched by the best known connectionist architectures and training algorithms. However, I argue that humans perform the crucial generalizations *without being trained* on the exemplars that Marcus and Phillips cite. So, in a sense, the issue whether networks can be trained to perform the crucial generalizations is a red herring. I argue further that humans achieve the dramatic generalizations in question as a side-effect of a variety of *pre-existing* skills, working in concert. Finally, it is shown that the "hard cases" displayed by Marcus and Phillips do in fact provide the basis for a serious challenge to *pure* (non-modular) connectionist architectures.

Introduction

In Marcus (1998, in press) and Phillips (2000), intriguing refinements on the (1988) Fodor-Pylyshyn "generalization challenge" are presented. Both Marcus and Phillips argue that linguistically competent humans exhibit important forms of *generalization* that backpropagation-trained networks (both recurrent and feedforward) cannot attain. Though neither author categorically asserts that their negative conclusions apply to every form of connectionist training, both authors argue that commonly recognized varieties of *eliminativist* architectures (i.e., those eschewing classical representations) are at stake.[1]

In this paper, I examine two instances which typify the "hardest challenges" produced by these authors. While I agree that they have each exposed some important training limitations of backpropagation networks, I shall argue that humans perform the crucial generalizations *without being trained* on the exemplars that Marcus and Phillips cite. So, in one sense, the issue whether networks can be *trained* to perform the crucial generalizations is a red herring. As I argue, humans possess the relevant generalization capacity because they have previously acquired separate skills which, working in concert, allow for nearly instantaneous pattern induction and reasoning. To be sure, the prior acquisition of these separate skills may involve "training up" various sub-networks in our brains, but this prior training may well *not* involve the "hard" kinds of generalization at issue.

Having said that, I would emphasize that my eventual, general conclusion supports both the positions of Marcus and Phillips. For, in the final section I consider whether connectionist architectures are capable (without implementing classically symbolic methods) of orchestrating the application of our "prior skills" in a fashion that permits very rapid pattern induction and reasoning. My conclusion favors the classicist position on this issue. Moreover, I propose a new challenge for eliminative connectionists which, in my view, formulates the deeper difficulty posed by the aforementioned "hard cases" of Marcus and Phillips.

The Hard Generalization Tasks

Generalizing Outside the Training Space

Marcus (1998) defines a network's training space as the N-dimensional vector space created by the non-zero training values of the N units comprising the network's input array. A datum presented during the network's (post-training) test phase lies "outside the training space" if and only if that datum does not fall within the vector space just mentioned. In effect, this entails that the datum is *novel* relative to the training corpus. For example, any datum would be novel in the relevant sense if it presented non-zero values to the input array in units that contained only zero values during training.

This "generalization hurdle" differs somewhat from the hierarchy of systematicity given in Hadley (1994a), but it appears equivalent to one of several levels of generalization formulated in Niklasson and van Gelder (1994). Interestingly, in the latter paper, the authors claim to satisfy this particular generalization challenge. Their claim is questioned in Hadley (1994b) and wholly disputed by Marcus (1998). Moreover, Marcus discusses a number of specific ways in which a network can fail to generalize outside its training space, and we now consider a particular "hard case" which Niklasson and van Gelder had not addressed.

Suppose a linguistically competent human is presented with the following series: "A rose is a rose", "A frog is a frog", "A pencil is a pencil". Humans will typically have no difficulty inducing the general pattern and complet-

[1] Here I am following Fodor and Pylyshyn's (1988) usage, according to which a composite (or complex) representation is classical in structure if one cannot activate (or token) that representation without, at the same time, tokening its syntactic constituents.

ing the following sentence: "A blicket is a ...". Humans will succeed here even though 'blicket' is a novel word which is outside their training space. In contrast, Marcus offers persuasive arguments, based upon the training-independence of output nodes, to show that backpropagation networks necessarily fail to match this success. These arguments are buttressed by several connectionist experiments conducted by Marcus.

On the basis of the above and related tasks, where a strong discrepancy exists between human performance and that of eliminative networks, Marcus concludes **C**: that human success in such cases is *not purely* due to any training of putative eliminative networks within our brains. This conclusion forms a keystone of Marcus' larger thesis – that the human ability to discover general patterns in cases such as these involves symbolic rule induction, and the application of such rules entails variable binding.

Now, while I agree with conclusion (**C**), I accept this conclusion for reasons other than any offered by Marcus. For one thing, I suspect that some Hebbian-competitive networks *can* generalize outside their training spaces on at least some tasks. This suspicion derives from recent experimentation with an architecture I have reported in (Hadley, et al, to appear). Another difficulty is that Marcus himself notes that when *distributed*, rather than local, representations are assigned to input tokens, backpropagation networks will, at first blush, provide the appearance of generalizing outside their training spaces. For example, in the "A blicket is a .....' test, a backpropagation network can successfully produce the distributed representation for 'blicket', *provided all the separate features* encoding blicket had, at some point, been employed in various nouns during the training phase. Admittedly, one could argue that this last proviso undermines any well founded claim to generalization outside the training space, but in doing so, one would undercut the entire force of the 'blicket' test case. For the word 'blicket' itself possesses only phonetic and graphemetic features that humans have often encountered prior to being presented with the 'a blicket is a ...' test phrase. That is, a plausible distributed representation for 'blicket' does not contain any features novel to English-speaking humans.

Marcus himself does not stress the objection I have assigned to some anonymous "one". Rather, he primarily objects (Marcus, in press, appendix 1) to the use of distributed representations on the grounds that they fail "... to unambiguously represent all and only the possible continuations to a given string ...". That is, when both nouns and verbs share several features in common (as indeed they would if we employ phonetic or graphemetic features), we run into the *superposition catastrophe* (crosstalk). (I would argue, however, that there is, *at most*, very little overlap between *semantic* features belonging to nouns and those belonging to verbs. For this reason, among others, the system described in Hadley et al, 2000, employs semantic features.) Be that as it may, it remains true that the phonetic and graphemetic features of 'blicket' are not novel. Moreover, *those* features are shared by both nouns and verbs, and, being a nonsense word, 'blicket' has no semantic features. So, if Marcus objects to the deployment of distributed representations in these network experiments, it seems incumbent upon him to demonstrate that humans are using only *local* representations when they successfully generalize from "A rose is a rose", etc. to "A blicket is a blicket". In the absence of such a demonstration, there seems no reason to grant that humans are in fact generalizing outside their training space in cases such as this.

For all the above reasons, I have serious reservations about Marcus' argument for conclusion **C**. Nevertheless, as mentioned, I believe there is a compelling reason to accept (**C**). And, if I am right about this latter reason, then the disputed capacity of eliminative networks to generalize outside their training spaces may be irrelevant *as the task is presently formulated*.

Here is the situation: humans clearly are able to perform very rapid pattern induction, not only in the various cases that Marcus cites, but in many other instances. In the above case, humans are able to induce a general pattern, and supply 'blicket' in response to the test phrase "A blicket is a ... ", within mere seconds after hearing "A rose is a rose", and the few remaining sample sentences. Given the very short time span involved, we may be quite certain that human success in this and similar cases does not stem from some extremely rapid training of "neural networks" (whether eliminative or not). As emphasized in Hadley (1993), in cases where humans make virtually instantaneous inferences, and when they acquire general rules in a matter of mere seconds, rapid synaptic weight change can be ruled out. Synapses simply do not grow fast enough to permit the acquisition of coherent functionality within the span of a few seconds. Functionally coherent synaptic changes occurs within spans of hours or days, not in a few seconds.

Now, it might be objected that in the case of the 'blicket' generalization, humans have in fact had entire days or even years to "train up" their networks, since, arguably, they have frequently heard phrases of the precise form, "an X is an X", in the past. However, this objection falters when we reflect that English-speaking adults have no difficulty inducing a *novel* pattern, and completing the final "sentence" in the following series: "Rose biffle biffle zarple zarple rose"; "Frog biffle biffle zarple zarple frog"; "Blicket biffle biffle — — —". In this case, the pattern being induced is clearly novel, since the pattern (template) itself not only includes the words 'biffle' and 'zarple', but involves a "syntax" that employs a double repetitive pattern not found in English. Yet, humans perform this induction in mere seconds. We must conclude, therefore, that the ability to perform rapid pattern inductions of this kind does not derive from some instantaneous training of a neural network, but must rely on at least some pre-existing skills. Certain of these prior skills involve the capacity to recognize phonemes or graphemes, which doubtless entails modification of synaptic "weights", which in turn (presumably) amounts

to the training of sub-networks within the brain.

Note, however, that this *prior* network training is not specifically directed to the generalization task just considered. The *novelty of the pattern* being induced ensures that very rapid, successful induction of this pattern must arise as a side-effect of prior skill acquisition. An appropriate challenge, therefore, for eliminative connectionism, is not whether a single network can be trained to generalize successfully from the few samples of data cited above, but *whether an essentially non-classical network can exercise its hitherto acquired skills* in a manner that yields, **as a side-effect**, the kind of rapid pattern induction considered above. Clearly, these are deep waters; I shall return to this issue in section 3.

Generalization in Rapid Inference

We turn now to consider an apparently "hard case", presented by Phillips (2000). This case is one of a series of generalization tasks considered by Phillips. Each task in the series possesses features which, at first blush, render it unlearnable by backpropagation methods in feed-forward and recurrent networks. However, Phillips engages in a dialectical process in each case, and *seems* to conclude that, provided overlapping distributed representations are assigned to functionally similar atomic constituents within the input data, then, with one exception, each task becomes learnable. The apparent exception is discussed below.

It is noteworthy, though, that even in the case of this seeming exception, Phillips describes a network capable of performing the task. He produces a carefully designed, fragile (and hand-crafted) network whose prescribed weight configuration suffices to display appropriate generalization behaviour. However, Phillips neither argues that the requisite weights could be acquired by learning, nor that the network possesses any cognitive plausibility. Given the precise and fragile nature of the requisite weight vectors, it seems unlikely that the particular network Phillips discusses could in fact be engendered through training.

Presently, I consider details of Phillips "recalcitrant case", but before doing so it will be helpful to consider a partially analogous example. Let us assume that Fiffle, Giffle, and Kiffle are names of propositions that have truth values. (I assume these three names, *qua* names, are novel for most readers.) Also suppose that the following three statements are true.

> If Fiffle is true, then Giffle is true.
>
> If Giffle is true, then Kiffle is true.
>
> If Kiffle is true, then Fiffle is true.

Finally suppose that Kiffle is true. What else can then be known to be true? Before reading further, I invite the reader to discover what can be inferred.

Doubtless, without effort, you have rapidly inferred the truth of the two remaining propositions, Fiffle and Giffle. Any number of literate humans, who have no training in formal logic, could similarly succeed at this task. Clearly, in the elapsed time between your having read the problem statement and your having derived the remaining propositions *no neural network was trained* within your brain to perform the relevant inferences. Rather, your success stems from a prior ability to engage in iterative processing and *modus ponens* inferences. Arguably, in the case of humans who lack formal logic training, the latter capacity derives from prior training in language use (with sentences of the form: if P then Q).

Of course, from a connectionist perspective, the capacity to apply inference patterns to novel data (Fiffle, Giffle, and Kiffle) is a significant achievement, and it is questionable whether any cognitively plausible ANN experiment has succeeded in this task.[2] However, just as in the case of 'blicket', 'Fiffle', 'Giffle', and 'Kiffle' possess only *non-novel* phonetic and graphemetic features. Given that Marcus was able to train a simple recurrent net to predict 'blicket' in the 'a Y is a Y' formula, there would seem no obstacle, in principle, to the *modus ponens* inference pattern being applied to nonsense words, provided the latter are represented by distributed representations of the right kind.

With this in mind, we now consider the problematic case that Phillips describes, viz., *transverse patterning*. Phillips defines transverse patterning as follows:

> Transverse patterning is an example of a stimulus-response task that depends on *between constituent* relations (my emphasis) . A task instance or problem set consists of three unique patterns (e.g., strings, shapes, etc.) *A, B* and **C**, such that: A predicts B; B predicts C; and C predicts A. Once the transverse patterning task structure has been learnt from the first few problem sets, subjects require only one of the three stimulus-response pairs to predict the remaining two, for any new transverse patterning problem set.

At first glance, there may appear to be an ambiguity in the last of the sentences just quoted. However, carefully read, the sentence tells us that human subjects can predict, when given a single *novel* stimulus-response pair (of the form "shape X predicts the appearance of shape Y") what the two remaining novel S-R pairs, having this general form, will be. In personal communication with Phillips I have verified that the sentence is *not* describing human predictions of the two remaining geometric shapes, given the first geometric shape.

Phillips goes on to relate that *trained* feed-forward and recurrent networks are not able to match the impressive kind of generalization just described. This is not surprising. What is surprising, initially, is that *humans can* predict what the two specific novel S-R pairs will be, given exposure only to one of the three novel S-R pairs. This surprise evaporates, though, when we learn (as I did in further personal communication with Phillips) that human subjects are told in advance what the three geomet-

[2] From a cognitive standpoint, I have serious qualms about Boden's and Niklasson's (2000) recent results on this issue.

ric shapes will be in the novel test situation. Given this, and given that the subjects will have learned the overall structure of the training experiment (following their first few sessions), they are able to *reason* analogically, and to derive by a process of elimination, what the remaining two S-R pairs must be. For example, in the new test situation, subject Kim learns that the novel shapes will be a star, an ellipse, and a hexagon. After being presented with the first S-R pair, Kim is able to *infer* immediately, that (say) A corresponds to the star, and that B corresponds to the ellipse. Knowing this, Kim can reason analogically that the ellipse (corresponding to B) must predict the third geometric shape, the hexagon. Reasoning further, again by analogy, Kim discovers that the hexagon (corresponding to C) must be the predictor of (A), the star.

Now, the crucial point to realize here is that human success in this task involves powerful reasoning skills (both analogical and reasoning by elimination) which the human possessed *prior to any* of the S-R conditioning induced in Phillips experiment. In all likelihood, these prior reasoning skills reside in separate modules which were unaffected by the S-R reasoning presently being considered (see Hadley, 1999, for arguments on the modularity issue). In contrast, the non-modular feedforward and recurrent networks which Phillips contrasts with the human success, possess no prior skills in reasoning of any kind, much less the powerful reasoning capacities that humans bring to the experiment.

The situation is complicated, and confused, by the fact that various of Phillips' remarks create the impression that he is contrasting a human ability to generalize an inference pattern, *which has been acquired in the S-R conditioning sessions*, with an inability, on the part of widely used connectionist architectures, to exhibit comparable generalization. At various points, Phillips explicitly states that the transverse patterning task amounts to the task of *generalizing* logical inference patterns. For example, he says,

> Under controlled conditions, subjects consistently make inferences implied by the underlying logical rules (Halford *et al*. 1998a). Indeed such tasks are ideal tests for systematicity in connectionist networks (Phillips and Halford 1997, Phillips 1999).

Given the type of S-R conditioning employed in Phillips experiment, one naturally supposes that the 'underlying logical rule' that Phillips currently has in mind is tantamount to the rule of *modus ponens* employed in the example I offered above. However, as we have now seen, the crucial human success that Phillips highlights in *not* dependent on a simple application of a given inference pattern (or even repeated applications of that pattern) as occurs in the Fiffle example I provided. Rather it depends upon the composition of prior reasoning skills (a composition involving both analogical reasoning and deduction by a process of elimination) *combined with* an ability to extend inference patterns to novel data.

Phillips believes that his transverse patterning case demonstrates that similarity in distributed representations (of atomic constituents) does not suffice to enable certain kinds of networks to generalize a particular kind of inference patterns to novel data. To the contrary, I have argued that Phillips has conflated the challenge of having a network generalize the application of a single inference pattern with several larger issues. While I certainly agree that the use of distributed representations cannot compensate for the absence of separate, previously acquired reasoning skills (together with the considerable prior training that would engender those skills), this tells us nothing about the efficacy of deploying distributed representations when attempting to apply a *single* known inference pattern to novel data. It is crucial to realize that, in the "transverse patterning" experiment discussed above, humans are doing far more than generalizing the application of a single inference pattern to novel data. They are engaged in an elaborate process involving meta-observations and the composition of separate, sophisticated inference skills.

Moreover, it is questionable whether the S-R training sessions have *trained* human subjects in any *new* inference pattern at all. It seems more likely that the sessions merely provided opportunities for subjects to acquire the base atomic facts (of the form X predicts Y, analogous to the simple "if-then" premises in my *modus ponens* example) which provide fodder for the capacity of humans to apply pre-existing inference skills to novel data.

In any case, I believe it is clear that Phillips' "hard case", like that of Marcus, involves the composition and application of pre-existing skills.

Discussion

In the foregoing, I have argued that, for the generalization tasks in question, the challenges posed to eliminative connectionism have not been felicitously formulated. For, in the tasks considered, we have seen that human success gives every appearance of either arising through the composition of multiple prior skills (viz., language comprehension, analogical reasoning, and deduction by process of elimination, in the case of transverse patterning) or arising as a side-effect of the capacity for language processing (as in the case of 'a blicket is a ...'). Human success in these cases is clearly *not* due to some virtually instantaneous "training" of our synaptic weights. I submit, therefore, that the fundamental challenge posed by these "hard cases" should be formulated essentially along the following lines:

> Demonstrate that a *single* holistic ANN, deploying eliminativist, non-classical representations could, as a manifest *side-effect* of its prior training, perform successfully on either of the "hard" tasks we have considered here.[3]

[3] I regard a network's success on a task, T, as a manifest side-effect of prior training just in case the following two conditions hold: (1) it is clear that prior training had in some way contributed to the success; (2) the network's prior training in-

It might now be objected that the challenge just formulated is unfair, because my wording clearly precludes any solution founded upon the *interactions* of multiple connectionist modules. However, solutions predicated upon the interactions of separate connectionist modules represent a radical departure from the pure connectionist paradigm. Such modular architectures share much in common with traditional, symbolic AI approaches to induction and problem solving, in that much of their processing power derives *not* from the *vector* and *settling* operations that characterize the "new" paradigm (involving weight and activation vectors), but from cooperative data exchanges between separate modules.

I return to these issues presently, but let us first consider a different sort of objection that may arise. It might be argued that the "challenge" I pose above is not especially worrisome for the connectionist. After all, it is well known that connectionist networks often display emergent side effects. Furthermore, we know that some networks trained via backpropagation have already displayed some degree of compositionality, as evidenced in the capacity of the St. John & McClelland network (1990) to assign correct semantic interpretations to novel sentences. In reply, it should be noted that the degree of *skill* compositionality, required to solve the transverse patterning task, is of a radically different kind than any compositionality displayed by networks that assign semantic representations to novel sentences. The degree of semantic compositionality displayed by Hadley & Hayward's (1997) Hebbian network is markedly greater than that evidenced by St. John and McClelland, but even the Hadley-Hayward network displays no compositionality of entirely separate skills.

Indeed, I know of no non-modular connectionist network, whether eliminativist or not, which exhibits skill compositionality remotely approaching the level required in the transverse patterning problem. Admittedly, in the case of the Marcus generalization task (*a blicket is a ...*), it may not be obvious at first blush that humans employ multiple, *separate* prior skills to solve the task, but we know that, at the very least, a capacity to understand a range of natural language is presupposed. Moreover, *this* capacity is so complex and multi-faceted, that any number of competent linguists would affirm that a variety of separate skills are involved.[4]

Returning now to earlier comments on a *modular* approach to skill compositionality, I would stress that, in my view, such an approach is promising. Indeed, I have argued in (Hadley, 1999) that whenever a variety of markedly distinct skills are involved in a task (such as the skills I have noted above), it is likely that separate modules are involved. Such modules may very well be spatially distributed, and subject to some degree of noisy interactions with other modules, but for computational reasons they should still be regarded as distinct modules. However, in that same paper, I argued that humans are demonstrably able to employ their skill modules in *novel combinations*. To place the issue in a very small nutshell, the mere fact that humans can follow specific kinds of novel rules, within mere seconds after being told such a rule, suffices to show that the brain can transfer information (or data) along sets of *combinatorially adequate pathways* between the separate skill modules. I argued further, by an examination of logically possible cases, that the existence of such combinatorial pathways entails that at least one of several types of classically recognized architectures is present in the human cognitive system. Space limits do not permit a detailed recapitulation of these arguments. However, I submit that a *modular* approach to achieving the impressive "side effects" noted in the "hard cases" which have concerned us, does not represent the type of solution that would appeal to researchers who view connectionism as a radically new paradigm. In any case, neither the hybrid-modular approach, nor the single-holistic-network approach has yet been shown to yield side-effects even approaching those involved in the transverse patterning example.

volved no task possessing an underlying structure identical to that of task T.

[4]Examples of such separate skills include: (1) the ability to recognize distinct words, (2) the ability to recognize highly ungrammatical sentences, (3) the ability to form the past tense of verbs.

References

Boden, M. & Niklasson, L. (2000). Semantic systematicity and context in connectionist networks. *Connection Science*, 12, 111-142.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28, 3-71.

Hadley, R.F. (1993). Connectionism, explicit rules, and symbolic manipulation. *Minds and machines*, 3, 183–200.

Hadley, R.F. (1994a). Systematicity in connectionist language learning. *Mind and Language*, 9, 247-272.

Hadley, R.F. (1994b). Systematicity revisited: reply to Christiansen and Chater and Niklasson and van Gelder. *Mind and Language*, 9, 431-444.

Hadley, R.F. & Hayward, M.B. (1997). Strong semantic systematicity from Hebbian connectionist learning, *Minds and Machines*, 7, 1-37.

Hadley, R.F. (1999). Connectionism and novel combinations of skills: implications for cognitive architecture. *Minds and Machines*, 9, 197-221.

Hadley, R.F., Rotaru-Varga, A., Arnold, D.V., & Cardei, V.C. (to appear). Syntactic systematicity arising from semantic predictions in a Hebbian-competitive network, *Connection Science*.

Marcus, G. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, Vol. 37.

Marcus, G. (in press). *The Algebraic Mind*. (Cambridge, MA: MIT Press).

Phillips, S. (2000). Constituent similarity and systematicity: the limits of first-order connectionism. *Connection Science*, 12, 1-19.

Niklasson, L.F. and van Gelder, T. (1994). On being systematically connectionist. *Mind and Language*, 9, 288-302.

St. John, M.F. and McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-257.