Embodied induction:
Learning external representations*

Mark Wexler
Laboratoire de la Physiologie de la Perception et de l'Action
Collège de France, 11, pl. Marcelin Berthelot, 75005 Paris, France
wexler@cdf-lppa.in2p3.fr

Abstract

The problem of inductive learning is hard, and—despite much work—no solution is in sight, from neural networks or other AI techniques. I suggest that inductive reasoning may be grounded in sensorimotor capacity. If an artificial system to generalize in ways that we find intelligent it should be appropriately embodied. This is illustrated with a network-controlled animat that learns n-parity by representing intermediate states with its own motion. Unlike other general learning devices, such as disembodied networks, it learns from very few examples and generalizes correctly to previously unseen cases.

Induction

In artificial inductive learning, the machine is trained on only part of the set of possible input-output pairs; once it reaches some criterion of success on this training set, the machine is tested for "correct" generalization on the remaining test set. The number of generalizations consistent with the training set is usually very large. The trouble is: no purely formal, syntactic criterion can systematically pick out what we consider as the "correct" generalization out of this large field of possibilities, as shown by Goodman in his "grue" argument (Goodman 1983). From a strictly unbiased ("objective") point of view any generalization that is consistent with the training set is as good as any other. Nevertheless, if all generalizations are equal, some generalizations are more equal than others: they strike *us* as more clever or perspicacious, and it is *these* generalizations that we want our learning machine to prefer.

The usual solution is to endow the learning system with systematic inductive bias. No formal system can be entirely free of bias, as it is obliged to formulate its hypotheses in some language; any such language excludes some hypotheses, and of the ones it does include makes some some easier and more natural to express than others. Another type of bias is due to the learning algorithm. This typically results in a norm, "badness", on the hypothesis language—the size of the classification tree or LISP expression, for example. Induction is then just a form of optimization or search, an attempt to minimize the badness of the hypothesis while maximizing consistency with the training set.

The practical problem with this picture of biased learning is the so-called bias/variance dilemma. Not enough bias underconstrains induction, while too much bias makes the system ad hoc or overly specialized, and deprives it of plasticity. It is very difficult to find robust forms of bias that avoid these two extremes inside particular problem domains, and crossing domain boundaries is even more difficult. The result is that learning systems must be restricted to specialized domains, and the inductive bias, to be put in by the programmer—or by nature—carries a heavy load of specific information. Such a solution to the problem of induction begs the cognitive question: if bias is to be put in by hand, whose hand is it? (If your answer is natural selection, keep in mind that phylogenetic learning is at least as difficult as the ontogenetic kind.) Moreover, both nervous system development and learning behavior are more flexible than such a picture would suggest.

The problem of induction has recently been discussed from the point of view of intermediate representations (Kirsh 1992, Clark and Thornton 1997). In the inductive learning of patterns, categories or rules, the distinction is made between low-order regularities that are present directly in the training data (such as conditional probabilities between individual input and output bits that are close to 0 or 1), and more subtle higher-order regularities that can only be expressed in a richer language and in terms of the lower-order regularities (such as relational properties, etc.). The lower-order regularities can certainly be picked up by formal, fairly unbiased techniques. As for the higher-

* Published in *Embodied Cognition and Action (Papers from the 1996 AAAI Fall Symposium)*, M. Mataric, ed. AAAI Press, 1996.

order regularities, if the learner is to succeed it must first re-represent the raw input in order to make the more subtle pattern manifest. (The learner can of course choose to treat the higher-order regularity as if it were lower-order, by, e.g., simply memorizing the input-output pairs, or by learning bit-by-bit associations. But then it would fail to generalize correctly: it will certainly have learned something, but not what we were trying to teach it.) The problem of perspicacious re-representation is thus seen as the critical component of higher-order learning.

Multi-layer neural networks have become popular mainly due to the assumption that the intermediate layers will somehow carry out this re-representation, and the magic is supposed to be that this is induced by means of a completely mindless procedure, hill-climbing. Clark and Thornton (1997) have shown that this is not so for a number of difficult learning problems; the one I will discuss is n-parity, as I also use this problem in my toy model (see below). The n-parity mapping receives n bits as input and outputs one bit, the sum of the input modulo 2. This mapping is hard to learn from examples because they provide no raw statistical information: all the conditional probabilities between input and output bits are 0.5. It is well known that multi-layer feed-forward networks can "learn" parity by means of hill-climbing—but this is only when they are trained on *all* the 2^n input-output pairs. No simple associationistic learning will do for this rule: changing any bit of the input flips the answer. Reproducing the training set is necessary but not sufficient for having a concept or a rule; simple memorization will lead to the same result, and we would not then say that the memorizer has learned a rule, because (among other reasons) no re-representation of the input has taken place. A sharper test for having learned a rule is of course correct generalization to previously unseen cases.

Clark and Thornton have found that *no* training algorithm on *any* network architecture leads to networks that generalize correctly to previously unseen cases of parity; in fact, even in the best cases it suffices to withhold a very small fraction of the problems from the training set for generalization to fail completely.[1] (Other, non-network general algorithms fail just as badly.) This is very bad news for neural networks. Practically, it shows how bad general network methods are at generalization: many complex problems can be expected to contain embedded parity or more complicated higher-order structures, and in cases of practical interest no system can be trained on *all* cases. More fundamentally, it shows that, at least in this case, that when a network learns to reproduce input-output pairs, what it has actually learned is entirely unlike the rule that *we* have used to arrive at the training set. Indeed, what reason do we have to suppose that a neural network, with its peculiar dynamics and bias, would generalize beyond the training set in the way *we* would? The only possible reason would be the biological verisimilitude of artificial neural networks; but this, apparently, is not enough.

Embodiment and action

I would like to suggest that the at least some of difficulties of induction in artificial systems are due to those systems' lack of embodiment, their incapacity for external action. By *action* I do not mean muscle contractions, or the planning of complex movements, though these are indispensable to it. What I do mean is: *the ability to act on, that is to modify one's environment and to perceive the results, direct or indirect, of the action.* Thus, issuing verbal commands to others counts as action on my definition (provided they get carried out, and you can perceive their effects), although in this case the immediacy of physical feedback—an important aspect of embodiment—is lost.

A more formal way to talk about action and cognition is through the action-perception cycle. A cognitive system consists of two components, the internal and external, at any given time each of which is in a particular state, S_i and S_e. At every moment the internal, mental state S_i is mapped into some action on the external component, executed by means of motor transducers. This action modifies the external, world state $S_e \rightarrow S'_e$. This leads to new perceptions (via sensory transducers), which modify the internal state $S_i \rightarrow S'_i$, which leads to new actions, etc. The view of cognition as "information processing" collapses this cycle to a linear process—sensory input, leading to cognitive processing and modification of the internal state, leading to motor output—which overlooks the fact that natural cognitive systems are themselves largely responsible for their "inputs". Even purely mental operations rely on this external control. The role of action and embodiment has been recently stressed in the situated robotics and animat communities (although with somewhat different motivations), in linguistics by Lakoff (1987) and Johnson (1987), in philosophy of mind by Putnam (1981); the view that the external half of the perception-action cycle plays an integral role in cognition has recently been dubbed

[1] To the author's knowledge, the closest that any neural network—or any other general learning system—has come to generalizing parity is Pollack's (1992) cascaded network, which, given input strings of length 1 and 2 as well as some longer strings for training, generalizes correctly to some longer input strings.

"active externalism" by Clark and Chalmers (1996).

Thus action, besides serving pragmatic ends, may be fundamental to what are generally considered "higher" cognitive functions. The main thesis of this work is that, in particular, sensorimotor routines or schemas may *ground* (Harnad 1990) the mental (and other) operations involved in induction. In the language of Clark and Thornton, the earliest re-representations of the raw input (and therefore the ones most difficult to learn) may very well be provided by action schemas. This somewhat paradoxical notion can be fleshed out by the following toy model.

Induction in an animat

The fine detail of one's embodiment is probably very important. It may turn out, for instance, that the degree of opposability of one's thumb constrains one's cognitive capacities. I believe, however, that embodiment also confers very general cognitive advantages on the embodied system, advantages that can, and should, be studied independently of the mechanical-physiological details. It is my goal, therefore, to create a *minimal* model to illustrate the general role that embodiment and action can play in cognition.

We begin therefore with a highly simplified—but not altogether unrealistic—embodied creature, an animat. This particular animat lives on a 2-dimensional plane; its external "world" state is simply its position on the plane and its heading. Time is taken as discrete. At each tick the animat's muscles receive a motor command, telling them the distance to travel forward and the angle by which to turn. The sensory input returned consists of the distance to a fixed landmark, and the angle between the animat's heading and the landmark (given as a sine and cosine, to avoid discontinuities).

The goal is for the system to learn n-parity. To this end we introduce two neural networks to connect the animat to its task (the system is illustrated in Fig. 1b). The animat always starts out in the same position and orientation (at (1,0) where the landmark is at the origin, and facing north). The parity problem is fed bit-by-bit to the input-motor network (a one-layer perceptron), whose job is not to give the answer to the problem but to issue motor commands to the animat. The sensory signals from the animat are fed to a second, sensory-output network (also a one-layer perceptron), which at the end of the presentation of the problem is supposed to output the answer—the parity of the input string. (To be in conformance with the action model given above, there should also be a recurrent connection between the sensory signal and the input of the input-motor network. Such a connection is not needed to learn parity, but can be added without substantially changing the results to be presented below. For more complex problems it becomes necessary.) Having no other representational methods at its disposal (such internal recurrent connections), the system is obliged to represent the problem, and to keep track of the intermediate results, by means of its action.

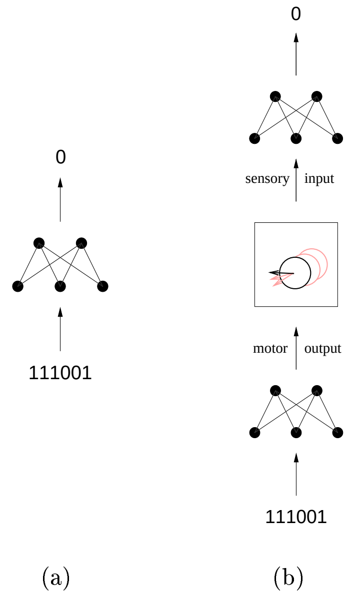

Figure 1: (a) A disembodied learning system. (b) An embodied system used in this work.

The networks are trained by means of a genetic algorithm. Each experiment proceeds as follows. A fraction f of the 2^n problems are assigned to the test set, and are never used in training (the training set always has an equal number of even and odd cases). In each generation of the GA each member of the (initially random) population is evaluated on the $2^n(1-f)$ training problems. (The weights and thresholds of the networks are coded as 10-bit strings in the genome.) The score on each problem is the absolute value between the output of the sensory-output network and the true answer (since logical 0 is represented by -1.0 and 1 by $+1.0$, the best score is 0, the worst is 2, and 1 is chance level); the score on the entire training set is the mean of the scores on each problem. The population (size 50, 10 bits/weight) is evolved by both 2-point crossover and mutation, with rank-based fitness. The experiment was stopped as soon as a member of the population reached criterion on the training set, a score of 0.001 or below. (Occasionally experiments ran over 200 generations without reaching criterion; these were discarded.) The best member of the population

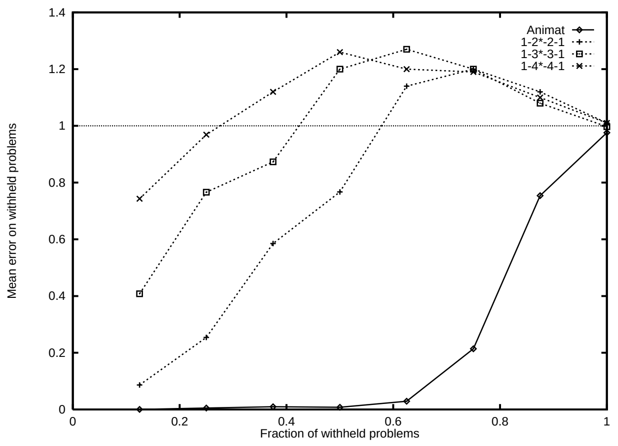

Figure 2: Generalization error (chance=1) for embodied systems and disembodied controls

is then tested on the $2^n f$ problems in the test set, which, I stress, the population had never seen during its training. The average score on the best population member on the generalization test is the score for the experiment. I ran 100 experiments for each value of f and averaged the scores, where new training and test sets were chosen for each experiment. (Further details available on request. All techniques used were very generic. The results were not sensitive to small variations in the parameters.)

For the control, I wanted to make things as hard as possible for myself by comparing the generalization performance of the embodied systems with that of the *best generalizing* disembodied networks. As shown by Clark and Thornton (1997), feed-forward networks for the non-temporal version of the problem are miserable at generalization. For the temporal version feed-forward networks won't do, as they do not preserve state, and therefore at least some recurrent connections are required. After experimenting with a number of architectures, I found that simple recurrent ("Elman") networks generalize best. Within this class, 1-a^*-b-1 architectures are best (* denotes a recurrent context layer), and as long as b is not too large the performance depends essentially on a; $b = a$ seems a good choice. The three best architectures are 1-2*-2-1, 1-3*-3-1, and 1-4*-4-1. These disembodied networks were trained by exactly the same method as the embodied systems. It should be noted that the disembodied systems got stuck in local minima much more often than the embodied.

The results for 4-parity are presented in Fig. 2, where the mean generalization error is plotted against f, the fraction of the 2^4 problems that were withheld from the training set. Error of 1 corresponds to chance level. (There is a good but uninteresting explanation for why the disembodied networks actually perform worse than chance for large values of f.) The embodied systems generalize almost perfectly up to $f = 0.75$. As for the disembodied networks, with the marginal exception of the 1-2*-2-1 architecture (which is almost pre-engineered to become a parity-calculating flip-flop), they generalize very poorly (as do Clark and Thornton's models): omitting just two problems gives very high error, and at four problems they are close to chance level. Even the 1-2*-2-1 architecture has errors that are 50-100 times greater than those of the embodied systems for f below 0.75. The problem length can be increased without substantially changing the results: keeping f fixed I got similar generalization performance for 5-, 6-, and 7-parity.

A final word concerning technique. The crucial aspect of this model is the sensorimotor embodiment, not the details of the two (input-motor, sensory-output) controllers. Although I used as controllers neural networks evolved by genetic algorithms—mainly because I am familiar with these techniques—other general purpose learning systems would probably do just as well, unless they are especially maladapted to the task. The point is that given a proper embodiment, the task of the controllers becomes very simple.

External representation

The interesting question is how the embodied systems managed to generalize so well. As I have already discussed, these systems had no *internal* means to represent the problem, therefore they had to perform all "computations" externally, i.e., by means of their movement. All the successfully generalizing systems adopted variations on the following strategy in the input-motor network: do not move forward, do nothing on 0, turn by 180° on 1. To calculate parity, the sensory-output network just has to give 0 if the animat is facing north (the landmark is to its left), and 1 if it is facing south (landmark to the right).

The representation of the parity problem spontaneously evolved by the embodied systems is closely akin to "epistemic action" recently investigated by Kirsh in the context of human problem solving (Kirsh 1995, Clark and Chalmers 1996). The idea is that for reasons having to do with human cognitive limitations, people choose to offload certain mental operations onto the external world; i.e., instead of performing an operation "in the head" one performs a much simpler physical action, lets the world evolve, and then reads off the answer. For instance, instead of mentally rotating the falling pieces in the computer game Tetris in order to see where they would best fit, people offload the rotation by physically rotating the pieces on the screen. On-line planning is another example: instead of planning where one will turn on the next street corner, one simply goes there, fully expecting that once there, it will be obvious where to go. In these examples and in many others, the cognitive operation is spread beyond the confines of the cognitive system into its external world. This kind of offloading, I suggest, may play an important role in grounding and channeling inductive learning. Even when overt physical action is absent (the usual case in human reasoning), such action and its benefits may be present in the covert, internalized form of mental imagery (visual and kinesthetic) and its transformations, often reported in reasoning.

The conclusion we can draw is as follows. The peculiar internal dynamics of artificial neural networks do not lead them to generalize "correctly" (i.e., in a way that makes sense to us) to previously unseen cases of difficult problems such as parity—indeed, why should they? If we trade this internal dynamics for an external, sensorimotor dynamics of a simple embodied creature, the generalization becomes correct. This suggests that induction is less a matter of formal selection criteria on hypotheses,[2] but rather a procedure grounded in one's sensorimotor relations with the external world. Indeed, we choose the hypotheses that we do, even in abstract contexts, partly because they have a certain *meaning* for us—for a human being parity is simply more meaningful than one of the "erroneous" hypotheses generated by the disembodied networks—and meaning is not present in formal, ungrounded, disembodied systems (Putnam 1981, Harnad 1990). The problem of induction is to discover what is it about certain hypotheses that makes them more meaningful than others in certain contexts. Perhaps a relation to sensorimotor schemas is part of the answer.

The obvious direction for future work is to vary both embodiment and learning problem, to discover the types of embodiment that can ground particular inductive hypotheses. In doing so one rapidly realizes that the simple architecture used here is insufficient. Its chief drawback is an inability to combine simple action schemas into more complex routines that would support more complex hypotheses. What is required is a robust, general framework for progressive modularization, combining simpler, successful schemas into larger wholes. Another interesting avenue is to study how the external routines of the type evolved here may be internalized. After all, we do not *always* use our hands when we think—sometimes we use mental images, an internalized form of perception and action. By varying the internal "mental" apparatus during the course of inductive learning, perhaps a mechanism can be found for the internalization of perception and action.

[2] Inductive learning in neural networks is no less formal than in classical AI; the formal computations simply occur on a lower and less explicit ("subsymbolic") level.

References

Clark, A. and Chalmers, D. 1996. The external mind. Manuscript, Dept. of Philosophy, Washington Univ.

Clark, A. and Thornton, C. 1997. Trading spaces: Computation, representation and the limits of uninformed learning. *Beh. Brain Sci.* 20(1).

Goodman, N. 1983. *Fact, fiction, and forecast.* Cambridge, Mass.: Harvard Univ. Press.

Harnad, S. 1990. The symbol grounding problem. *Physica* D42:335–46.

Lakoff, G. 1987. *Women, fire, and dangerous things.* Chicago: Chicago Univ. Press.

Johnson, M. 1987. *The body in the mind.* Chicago: Chicago Univ. Press.

Kirsh, D. 1992. From connectionist theory to practice. In M. Davis, ed., *Connectionism: Theory and practice.* New York: Oxford Univ. Press.

Kirsh, D. 1995. The intelligent use of space. *Art. Intell.* 73:31–68.

Pollack, J. 1981. The induction of dynamical recognizers. *Machine Learning* 7:227–252.

Putnam, H. 1981. *Reason, truth and history.* New York: Cambridge Univ. Press.

Sejnowski, T. and Quartz, S. 1997. The constructivist manifesto. *Beh. Brain Sci.*, to appear.