Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision

Paul Davidsson
Department of Computer Science, Lund University
Box 118, S-221 00 Lund, Sweden
e-mail: Paul.Davidsson@dna.lth.se

Abstract

The symbol grounding problem, described recently by Harnad, states that the symbols which a traditional AI system manipulates are meaningless to the system, the system thus being dependent on a human operator to interpret the results of its computations. The solution Harnad suggests is to ground the symbols in the system's ability to identify and manipulate the objects the symbols stand for. To achieve this, he proposes a hybrid system with both symbolic and connectionist components. The first section of this article presents a framework for a more general solution in which a composite concept description provides the critical connection between the symbols and their real-world referents. The central part of this description, referred to here as the epistemological representation, is used by the vision system for identifying (categorizing) objects. Such a representation is often referred to in computer vision as the object model and in machine learning as the concept description. Arguments are then presented for why a representation of this sort should be learned rather than preprogrammed. The second section of the article attempts to make explicit the demands on the system that arise when the learning of an epistemological representation is based on perceiving objects in a real-world environment, as well as the consequences this has for the learning and representation of the epistemological component.

1 The Symbol Grounding Problem

The symbol grounding problem is described by Harnad in [Har90]. It concerns the meanings of the symbols in (physical) symbol systems. Traditional AI systems manipulate symbols that are systematically interpretable as meaning something. The problem of concern is that the interpretations are made by the mind of an external interpreter rather than being intrinsic to the symbol manipulating system. The system itself has no idea of what the symbols stand for, their meaning being totally dependent on the external operator.

A possible solution to this problem would be to attempt to *describe* the meaning of the symbols in a more powerful language. However, this would only lead to another set of symbols, ones which would likewise need to be interpreted, and in the end to an infinite regression.

1.1 Harnad's Solution

Harnad suggests in the same article ([Har90]) a solution to this problem. According to him, the meaning of the system's symbols should be grounded in its ability to identify and manipulate the objects that they are interpretable as standing for. He proposes a hybrid system with both symbolic and connectionist components, stating: "In a pure symbolic model the crucial connection between the symbols and their referents is missing ..." (p.344)

He argues that three kinds of representations are necessary: *iconic representations*, which are the sensor projections of the perceived entities, *categorical representations*, which are "learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections", and *symbolic representations*, which consist of symbol strings describing category membership. Both the iconic and the categorical representations are assumed to be non-symbolic.

He concludes that a connectionist network is the most suitable for learning the invariant features underlying categorical representations and for thus connecting names to the icons of the entities they stand for. The function of the network then is to pick out the objects to which the symbols refer. Concerning Harnad's approach, one can remark that, although it seems clear that a pure symbolic system does not suffice (since sensors do not provide symbolic representations), regarding connectionist networks alone as being capable of serving this function appears too limited.

1.2 A More General Solution

The problem of symbol grounding becomes easier to resolve if one views it in terms of the general concept representation framework presented in [Dav93]. In this framework a concept is represented by a composite description consisting of several components. The main idea here is that one should have different kinds of representations for different purposes, a single description not being able to provide all the functions desired. For example, one should normally not use the same concept description for both perceptual categorization and high-level reasoning.

The most important parts of the composite description are the following: the *designator*, which is the name (symbol) used to refer to the category, the *epistemological representation*, which is used to recognize instances of the category, and the *inferential representation*, which is a collection of "encyclopedic" knowledge about the category and its members, one that can be used to infer non-perceptual information and to make predictions. Representations corresponding to the epistemological representation are often referred to as object models in computer vision and concept descriptions in machine learning.

The category "chair" can be used to illustrate this. The concept's English name "chair" could serve as the designator and some sort of 3-D object model of a typical chair as the epistemological representation. The encyclopedic knowledge in the inferential representation can include such facts as: that chairs can be used to sit on, that they are often made of wood, and the like. Note that the framework does not specify how the different components are represented.

It is essentially the vision system here, through its use of epistemological representations that are parts of the same structure as the corresponding symbols, which permits grounding, or the connection between symbols (designators) and their referents (objects in the world), to be carried out. The fundamental idea here bears strong resemblance with one that inspired Minsky [Min75] to introduce the concept of *frames*, namely that: "When one encounters a new situation ... one selects from memory a substantial structure ..." (p. 212). In present terms, an object that is encountered (perceived) is matched with the epistemological representation. This activates a larger knowledge structure, the composite concept representation, of which the epistemological representation is a part. This structure contains, among other things (encyclopedic knowledge, for instance), the designator.

The main point to be made is that the epistemological representation does not have to be a connectionist network. Rather, it can be virtually any representation the vision system can successfully use to identify (categorize) objects.

1.3 Why Learning is Important

Two arguments for the necessity of learning have been raised. The first is that, since the designers tend to program their own grounding based on their own experience, it is difficult for them to program the representations that ground the symbols explicitly. Thus, a system should perform the grounding itself by learning the representations from its own experience (through its own sensors and effectors).

The second argument is based on the fact that at the time of design it is often impossible to foresee all the different kinds of objects the agent will encounter. Although some object models may be preprogrammed, an agent acting in a real-world environment will probably encounter objects that do not fit any of these models.

2 What and How to Learn

Since the symbol manipulation capacities of a system are grounded in its object manipulation capacities, the system must (in Brooks' [Bro91a] terminology) be both *situated* and *embodied*. Thus, some kind of physical agent (i.e., a robot) is needed that can interact with its environment by means of its sensors and effectors. This will, of course, influence the choice of a learning algorithm.[1]

The learning task to be considered here is one of learning 3-D object models (concept descriptions) by perceiving real objects.[2] Since the task is therefore not restricted to 2-D image analysis, one can use multiple sensors (visual, range, tactile and others), stereo-vision, active vision, or whatever.

Several restrictions a real-world environment places on the concept acquisition process constrain the choice of a learning algorithm. One such restriction is that the algorithm must be *incremental*. Since the robot cannot control the environment, it probably will not encounter all instances of a category at one point in time. Instead, it encounters an instance now and then, incorporating it into the bulk of its knowledge of concepts. Thus, concepts are acquired in a gradual fashion through interaction with the environment over time. Fortunately, the focus of machine learning (ML) research has been moving from batch-oriented learning towards incremental

[1] Wrobel [Wro91] has also pointed out the importance of grounding in concept formation, and the need of studying it within the context of acting agents. In contrast to the approach here, however, he chose to study the problem in a process-control setting, since he found vision and robotics too difficult and, surprisingly enough, unrelated to the problem. (Similar arguments have been put forward by Etzioni [Etz93], who studied the problem within the context of *softbots*, that is, intelligent agents in real-world software environments such as operating systems or databases.) Wrobel admits, nevertheless, that since concept formation by physical agents takes place in a three-dimensional environment, it may require qualitatively different methods. His model allows only simple real-valued or nominal sensors.

[2] Some learning, however, may also occur on lower levels.

learning. Still, most connectionist networks (e.g., back-propagation networks) are not suitable for incremental learning.

Another constraint on the choice of a learning algorithm is that the concepts must be acquired in a *parallel* fashion. This is because the robot does not learn just one concept at a time as it is sometimes assumed to in ML.

2.1 Relevant ML Paradigms

Several concept learning paradigms exist in ML. They differ mainly in the amount of information the learning algorithm is provided with. In *deductive learning*, which needs a large amount of information, the learner acquires a concept description by deducing it from the information given. *Explanation-based learning* [MKKC86], which transforms a given abstract concept description (often based on non-perceptual features) into an operational description (often based on perceptual features) by using a category example (described by its operational (perceptual) features) as well as background knowledge for guidance, is the type of deductive learning investigated most.

In *learning from examples*, which is the paradigm in ML studied most [Win75, Mit82, Qui86], the learner induces a concept description from a set of pre-classified examples of the category. There must thus be some kind of teacher present who can classify the examples. The examples are almost always symbolic descriptions of the actual examples. In the connectionist literature this kind of learning is often referred to as supervised learning.

In *learning by observation* (unsupervised learning), where the examples are not pre-classified, the learner must also *form* categories. (In learning from examples the categories already exist and are known to the teacher.) Typically, the learner is given a number of descriptions of entities. Based on their features it groups the entities into categories. When this is done, the system creates descriptions of the categories much in the same way as the systems that learn from examples do. Such a system is commonly called a *conceptual clustering* system [FL85].

Since it is desirable that the agent is able to learn (and form) concepts both with and without human supervision, both learning from examples and learning from observation are relevant for the task at hand. On the other hand, it is questionable whether real-world situations exist in which explanation-based learning can be applied, i.e. in which all the information that is required can be provided.

2.2 Choice of Representation

Most existing ML-systems require as input descriptions of the examples in the form of feature vectors, and produce concept descriptions based on the values (of some) of these features. This is rather inappropriate for the recognition of 3-D objects, however, since there are often constraints on the relationships between the features, such as spatial relationships. Instead, *structural descriptions* of some kind are needed. Although some studies on the learning of structural concept descriptions have been carried out [DM81, TL91], interest in further research in the area seems to be declining. In computer vision, on the other hand, object models may have put too much emphasis on the shape of the objects (cf. generalized cylinders, surface boundaries, or volumetric representations), ignoring other, more global features such as color and surface texture that could be of help in the object recognition process.

Also questionable is the fact that the concept descriptions that most ML systems learn are classical definitions, i.e. descriptions based on necessary and sufficient conditions for category membership. In the cognitive science literature (cf. [SM81]) it has often been noted that there are various problems with such definitions. The most serious problem is probably that, in contrast to the *artificial* categories often used in AI experiments, it is often not possible to find necessary and sufficient features for *natural* categories. This is sometimes referred to as the ontological problem [Ams88].

Moreover, even if a classical definition for such a category exist, non-necessary features are often used to categorize (identify) the objects of the category. Sometimes one is even forced to do this since some of the necessary and sufficient features are not perceivable. For instance, in recognizing a piece of gold, one cannot generally perceive the atomic structure of the material directly. Instead, one uses such features as color, weight, and so on. Thus, it seems that humans, at least, do not use classical definitions for the epistemological representation. Instead, prototype-based concept descriptions appear more appropriate (cf. [SM81]). These are based either on the memorization of specific instances (the exemplar approach), or on the construction of a probabilistic representation (the probabilistic approach). Rather than applying a definition, categorization becomes a matter of assessing similarity to the prototype(s). Various studies in ML on prototype-based representation have been carried out (cf. [AKA91] and [dlM91]).

One of the key problems for an autonomous agent is to decide when to create a new concept. In particular, it needs to know when an instance of an unknown category is encountered. Somewhat surprisingly, this demand constrains the choice of representation radically. An assumption often made (sometimes implicitly) in creating ML-systems is that all the relevant categories are exemplified in the learning set. This assumption has led to the construction of algorithms that learn to *discriminate* between categories. By concentrating on differences between the categories rather than on

the categories themselves, they just learn the boundaries between categories. Moreover, they partition the entire description space into regions, so that every region belongs to a certain category. Thus, the algorithm cannot detect the occurrence of instances of unknown categories. Instead, such instances are categorized in a rather unpredictable manner. This problem is dealt with by Smyth and Mellstrom in [SM92] who considers decision trees and multi-layer neural networks as examples of discriminative models. As a solution to this problem they suggest the use of *generative* or *characteristic* [DM81] models which aim at discriminating the instances of a category from *all* other possible instances. These models concentrate on the similarities between the members of a category, the category boundaries being an implicit by-product of this. Various logic-based and prototype-based representations serve as examples of such models.[3] Smyth and Mellstrom also make quite a provoking statement: "In fact one could even conjecture that *only* generative models can be truly adaptive and that discriminative models are impossible to adapt in an incremental on-line manner. This is certainly true in the general case for the class of discriminative models which includes decision trees and fixed-structure neural networks." Thus, Harnad's connectionist solution to the symbol grounding problem may not even be adequate.[4]

3 Discussion

Some researchers (e.g., Brooks [Bro91b, Bro91a]) argue that reconstructionist (general-purpose) vision is too difficult and not even necessary for autonomous robots. They also maintain, in fact, that there is no need for symbolic representations at all. This would imply that the symbol grounding problem is irrelevant. However, they cannot escape the fact that the robots need to have concepts,[5] or at least, be able to recognize instances of a category. This implies that they must have an epistemological representation. Thus, the same problem appears again.

In the second section of this paper I have tried to make explicit the demands placed on the learning and representation of 3-D object models (concept descriptions) when the learning involved is based on perceiving real-world objects. First, the learning algorithm must be incremental and be able to learn many concepts at the time. Secondly, incorporating both learning from examples and learning from observation into it would be desirable. Finally, the learning system should be able to learn characteristic descriptions that can include structural information, preferably in a prototype-based representation. Some of these topics are treated in greater detail in [Dav92].

It is reassuring to note that each of these demands is met by some existing learning system. However, no system satisfies all of them. Thus, the (non-trivial) task remaining is to create such a system and integrate it with a computer-vision, or multi-sensor, system.

[3] Being discriminative or characteristic is actually not an intrinsic property of the representation, its also being dependent on the learning algorithm. For instance, both discriminate and characteristic logic-based descriptions exist (cf. [Mic77]).

[4] This at least applies to the learning of object models. On lower levels, however, it may turn out to be desirable to use connectionist networks.

[5] The concepts are not always explicitly represented, however. In Brooks' robots, for instance, the concepts are only implicitly present in the circuits and in the mechanical devices. It is argued in [Eps92], on the other hand, that there are several advantages to having explicit concept representations. For example, it facilitates the organization of knowledge and the focusing of attention.

Acknowledgements

I wish to thank Eric Astor, Christian Balkenius, Peter Gärdenfors, Robert Pallbo and Olof Tilly for helpful comments and suggestions.

References

[AKA91] D.W. Aha, D. Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[Ams88] J. Amsterdam. Some philosophical problems with formal learning theory. In *AAAI-88*, pages 580–584, St. Paul, MN, 1988.

[Bro91a] R.A. Brooks. Intelligence without reason. In *IJCAI-91*, pages 569–595, Sidney, Australia, 1991.

[Bro91b] R.A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1):139–159, 1991.

[Dav92] P. Davidsson. Concept acquisition by autonomous agents: Cognitive modeling versus the engineering approach. Lund University Cognitive Studies 12, ISSN 1101–8453, Lund University, Sweden, 1992.

[Dav93] P. Davidsson. A framework for organization and representation of concept knowledge in autonomous agents. In *Scandinavian Conference of Artificial Intelligence – 93*, pages 183–192. IOS Press, 1993.

[dlM91] M. de la Maza. A prototype based symbolic concept learning system. In *Eighth International Workshop on Machine Learning*, pages 41–45, Evanston, IL, 1991.

[DM81] T.G. Dietterich and R.S. Michalski. Inductive learning of structural descriptions. *Artificial Intelligence*, 16(3):257–294, 1981.

[Eps92] S.L. Epstein. The role of memory and concepts in learning. *Minds and Machines*, 2(3):239–265, 1992.

[Etz93] O. Etzioni. Intelligence without robots (a reply to Brooks). To appear in *AI Magazine*, 1993.

[FL85] D. Fisher and P. Langley. Approaches to conceptual clustering. In *IJCAI-85*, pages 691–697, Los Angeles, CA, 1985.

[Har90] S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.

[Mic77] R.S. Michalski. A system of programs for computer-aided induction: A summary. In *IJCAI-77*, pages 319–320, Cambridge, MA, 1977.

[Min75] M. Minsky. A framework for representing knowledge. In P.H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, 1975.

[Mit82] T.M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[MKKC86] T.M. Mitchell, R.M. Keller, and S.T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80, 1986.

[Qui86] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[SM81] E.E. Smith and D.L. Medin. *Categories and Concepts*. Harvard University Press, 1981.

[SM92] P. Smyth and J. Mellstrom. Detecting novel classes with applications to fault diagnosis. In *Ninth International Workshop on Machine Learning*, pages 416–425, Aberdeen, Scotland, 1992.

[TL91] K. Thompson and P. Langley. Concept formation in stuctured domains. In *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pages 127–161. Morgan Kaufmann, 1991.

[Win75] P. Winston. Learning structural descriptions from examples. In *The Psychology of Computer Vision*, pages 157–209. McGraw-Hill, 1975. Also in Readings In Knowledge Representation, ed. R. Brachman and H. Levesque, Morgan Kaufmann, 1985.

[Wro91] S. Wrobel. Towards a model of grounded concept formation. In *IJCAI-91*, pages 712–717, Sidney, Australia, 1991.