Acquiring Visual↔Linguistic Associations and Reasoning with Them

Ganesh Mani and **Leonard Uhr**
Computer Sciences Department
University of Wisconsin—Madison
1210 W. Dayton St., Madison, WI 53706.
E-mail: {ganesh, uhr}@cs.wisc.edu

Introduction

Human beings combine information of several kinds from a number of modalities in the course of everyday activities. Visual and linguistic information are two major types of such information; these seem to be used almost effortlessly in the process of everyday activities which may involve seeing, talking, acting or a combination thereof.

Visual ↔ linguistic associations are central to the abilities of both children and adults to perform ordinary tasks and not-so-ordinary tasks in the world. For example, when a person (A) tells another (B) "Turn on the light in the kitchen," it is assumed that A and B are in agreement on what words such as "light," "kitchen" and "on" refer to. Further, B may have to find the light switch in the kitchen and turn it on. Sometimes B may respond with "The light is already on in the kitchen" or with "Which one of these switches is for the light?". Other examples may involve thought experiments and mental imagery. For instance, A may ask B "Is the living room in your new apartment bigger than this room?" and before responding, B may have to visualize his new living room in his mind's eye and resort to some subtle reasoning (e.g. as to the sizes of tables and chairs and distances between them) to compare the size of the imagined room with the size of the perceived room.

We are of the view that one way to look upon mental representations is to assume that visual and linguistic information may be intertwined in the same memory structure, or if they get processed into distinct memory structures there are rich links between them that facilitate going from one representation to another, as necessary. We have developed a computer simulation of such a model where inputs from separate visual and linguistic modalities are processed and then combined. Hierarchical representations of visual inputs and linguistic inputs are built and linked together; this makes possible reasoning with both representations and also *grounds* each representation in the other. For example, once the words "horse" and "striped" have been described to a person using visual inputs (pictures), simply being told that a "zebra is a striped horse" is often adequate without actually being shown one (example due to Harnad, 1990). To teach "striped" in the first place, we might say something like "striped means with long narrow ribbons more or less parallel to one another," but think of how much simpler and more informative it is to show a few pictures of striped objects. Similarly, we are able to visualize a "sphinx" or a "unicorn" quite well from verbal descriptions. Whenever we describe objects or narrate events, we are making use of the fact that words or linguistic symbols are grounded in their visual counterparts.

Central to all this is the issue of different kinds of representations and the efficiency associated with them. For our purposes, a representation can be thought of as a collection of percepts, compounds of percepts, and sets of mechanisms or operators that act on them to produce new percepts or symbols in a systematic fashion. For example, percepts such as "stem" and "round" sensed through the visual modality may be acted upon by an operator that looks for their relative positions to imply the percept that stands for the shape of an apple. Such operators may often be specific to the particular percepts that they act on. In many cases, pictorial percepts are far richer in meaning than verbal descriptions, and their use can greatly expedite recognition and learning.

Visual and Linguistic Representations

Considerable research effort in the overlapping fields of artificial intelligence and cognitive science has been devoted to analyzing and understanding the nature of representations (e.g. see Fodor, 1976; Kosslyn, 1990; Larkin & Simon, 1987; Newell, 1980; Pylyshyn, 1981). Major types of representations that have been discussed in the literature can be differentiated based on their origin: either visual or linguistic. While Kosslyn uses the terms propositional and depictive representations to talk about linguistic and visual representations, Larkin and Simon have used the terms sentential and diagrammatic representations to refer to them.

Kosslyn has argued that people indulge in mental imagery (i.e. imagine things as pictures or images rather than as propositional structures) in tasks such as judging which room is bigger (discussed above). Support for imagery via depictive representations comes from experiments where people tend to take longer to answer questions about details of objects (such as naming of parts or features, or verifying a particular spatial relationship between objects), if these details are farther from the point of focus. In these experiments, the subjects are often asked to mentally focus on one point in the image (the object or a picture of it is shown initially, and the subjects imagine it during the remainder of the experiment) or on one part of the object before they are asked subsequent questions. Also, these experiments are repeated by keeping constant the amount of material scanned over to ensure that the amount of material scanned over is not responsible for the increased response times.

Kosslyn claims that the results of such experiments compel one to postulate an internal representation where properties such as distances between objects and object parts, or relations among them, are kept intact implicitly. For example, an apple on a table may be encoded using something like "on(apple, table)" in a propositional representation. In a depictive representation, there is no explicit symbol that stands for the relation "on"; the relation emerges from the juxtaposition and relative location of the symbols for the apple and the table.

Larkin and Simon's distinction between sentential representations (sequential, like the propositions in a text) and diagrammatic representations (indexed by location in a plane) is similar to the propositional versus depictive distinction made by Kosslyn above.[1] Larkin and Simon argue for the (computational) efficiency of the diagrammatic representation using the examples of a pulley problem from physics and a geometry problem.

[1] One important point to note about Larkin and Simon's work is that it explicitly refers to external representations people use. It may be reasonable to posit that the internal representations people use are similar to the external representations they employ. Note that here we are primarily talking about internal representations that people consciously manipulate and hence, people can be quizzed about these representations.

It is not clear, however, whether the diagrammatic or depictive representation is generally more efficient than a sentential representation. General principles or specific facts may frequently be stored propositionally. Examples of this may include "It is always cold in Madison in December," "This computer has 24 MB of memory," and "The Mark is the currency in Germany". Often, a sentence may be called upon to denote or stand for a potentially complex, rich set of visual or depictive information. Such a mechanism may result in the generation of a quicker response by obviating the visual search that may be necessary to arrive at the general principle or the specific entity.

Thus the most powerful and efficient representation may well be one that combines pictorial and symbolic information in such a way that both can be used as appropriate. The rest of the paper deals with the issue of using multiple representations, in particular using the two different kinds of representation suggested by thought experiments and also analyzed by other researchers (e.g. along the lines discussed above). However, we take the view that learning the mapping from one representation to another may turn out to be one of the crucial determinants of accuracy and overall efficiency.

The Role of Learning

Since humans can reason easily and fluently with both linguistic and visual representations, it seems appropriate to analyze how the acquisition of different representations, plus mechanisms to translate from one representation to another, is useful for solving a rich set of problems. We often find that in everyday use, words and pictures can substitute for one another. Further, even though the visual medium is often richer (in terms of information content) than the linguistic medium, it is possible to describe and analyze events and perform reasoning tasks in terms of words alone. For example, most of the information in this paper is linguistic.

We often resort to the use of mnemonics for remembering things. Often these mnemonics are linguistic strings rather than a connected visual sequence of interesting objects and events. For example, a sentence whose words start with the initial letters of various colors is a helpful mnemonic for translating the color bands on a resistor to its numerical resistance value. Similarly, some people make up different mnemonics to remember the number of days in each month.

Almost certainly combinations are best of all. Often a few words make a picture far more meaningful. Words can very effectively name, focus on the highlights, point to unusual aspects. We do not merely gesture to "look" — rather we say things like "notice the sunset, and how she is looking at it, and how he is looking at her."

These examples highlight the importance of not only using different types of representations (which impact the ease of use of knowledge and the accuracy of problem solving), but also the need to develop effective mappings from one kind of representation to another.

Another important point to note is that learning to associate words with pictures and objects, object qualities, etc. starts at a very young age. This facilitates further learning through the linguistic medium via conversations, textbooks and verbal exercises. In the context of machine learning paradigms (Carbonell, Michalski & Mitchell, 1983), this can be looked upon as the *learning by example* paradigm acting as a foundation for further *learning by instruction or advice-taking.*

In this paper, we would like to argue that people may have knowledge about the world stored in terms of both propositional and depictive representations (*a la* Kosslyn); or sentential and diagrammatic representations (*a la* Larkin & Simon). Further, they may decide to go from one representation to another as they deem fit during the reasoning process (say, in the living room example discussed earlier, if the person knew the size of his living room to be 30 X 20, the person may decide to estimate the size of the novel living room he is seeing in terms of its width and length, and may perform the required arithmetic to decide which room is bigger; in addition, the person may use the distances between chairs and tables in the room as an aid to estimate the dimensions of the room). But to be able to go from one representation to another requires appropriate learned or acquired mechanisms.

Learning in the CHILDLIKE System

The CHILDLIKE[2] system (Mani & Uhr, 1991a,b) is a computational model for a learner that learns from simple experiences consisting of inputs from multiple perceptual and linguistic modalities. In the current version of the system, visual and linguistic information that has already been pre-processed to some extent is used — e.g. the color or texture of a particular region and the circular shape present in the image may be encoded as primitive symbols in the visual channel. Similarly, segmented words are assumed to be available on the language channel. The recognition–cone architecture proposed for visual tasks such as recognizing houses in TV-images (Li & Uhr, 1987; Uhr, 1987) is being extended to handle a larger variety of tasks, including language. In the process, we are focusing on the higher levels of this hierarchical architecture.

The basic algorithm for forming visual $\leftrightarrow$ linguistic associations is the following (for a more detailed description, see Mani & Uhr, 1991b). Features on each of the channels — visual and linguistic — are extracted and aggregated using a near-neighbor heuristic. The actual aggregation function may simply sum the strength of each extracted feature or may apply a more complex *IF–THEN* rule which searches for particular features in specific relative locations and orientations. At each level moving up this hierarchical aggregation process, lesser and lesser entities are implied; but the entities implied grow more and more complex. The only restriction we impose is that the functions (at each "node" that looks at a set of its "children nodes") should be locally computable, for example using a microcircuit of neuron-like nodes (also see Uhr, 1990). It may also be possible to order the aggregation functions within each microcircuit or parent node which may enable, to a limited extent, gradient-based searches (such as the one described in Mani, 1990).

Links are generated between the features (whether primitive or aggregated) from one channel to the other. Thus a green region in the visual representation may be linked to the word "green," to the word "apple" and also to the phrase "green apple" initially. Subsequent experiences help strengthen some of these links selectively (e.g. the link between visual and linguistic "green"s is strengthened by gradually raising its weight each time a green object is observed). The current algorithm used in CHILDLIKE is a first cut at learning such associations. Controls such as a resource constraint to contain the number of compound features and links generated with each training instance are used. Also implemented are generalization mechanisms (such as *turning constants into variables* and *closing intervals* from Michalski, 1983) in an effort to help the memory structures capture regularities that are present in the learner's environment. At the same time, care is taken to see that the sizes of these structures are kept reasonable. Currently, arbitrary resource bounds are used; it would be best to have the algorithm add, for example, $k * logN$ nodes and links for every N training instances or features perceived (where k is some small constant).

CHILDLIKE assumes that a number of candidate interpretations are possible from the start (for example, a new word

[2] which stands for Conceptual Hierarchies In Language Development and Learning In a Kiddie Environment.

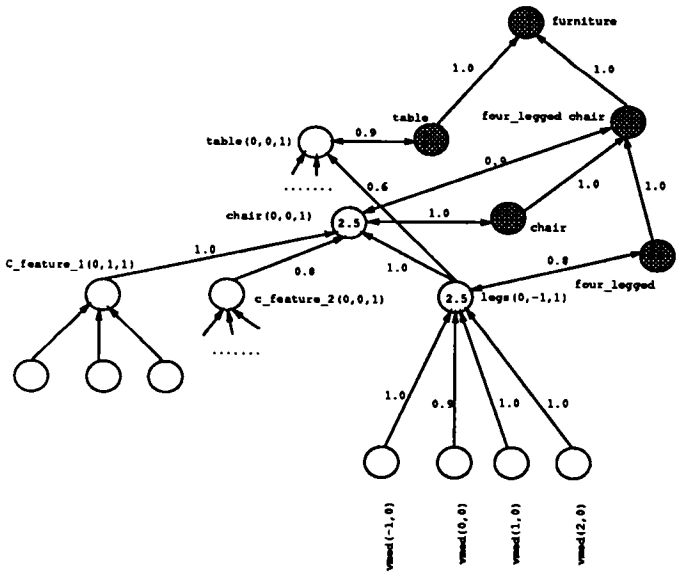

Figure 1: The Network Structure of Hypotheses

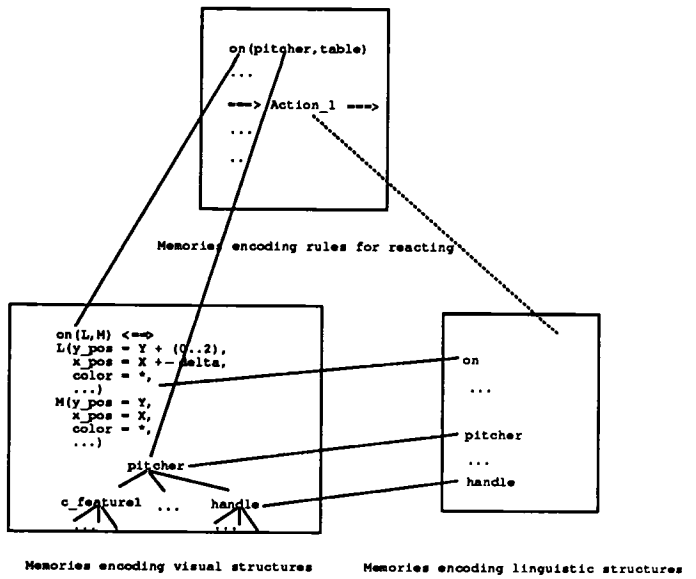

Figure 2: An Overview of the Memory Structures Acquired by CHILDLIKE

may be the name for an object, an object quality or for a relation) and hence generates a number of initial links (subject to a resource bound). Further training strengthens some of these associations selectively. In this sense, CHILDLIKE does not learn the easiest interpretation rapidly (like young children do in certain cases), but generates a number of candidate hypotheses with low weights. Out of these, a large number get weeded out as they are not strengthened by subsequent experiences.

What purpose does the acquisition of such visual $\leftrightarrow$ linguistic associations serve? Such associations can be very helpful, if not crucial, for reasoning using either the visual or linguistic representational mode.[3] For example, these associations can help the system understand what a zebra is after learning from experiences (both visual and verbal) about "striped" and "horse." They facilitate learning using utterances such as "apple is sweet" and "apple is fruit" after grounding words such as "apple" in the visual modality. A verbal question like "Are chair legs parallel and below the seat?" can be answered by examining the object-model for chair (learned as a function of visual experiences) that the system uses to recognize objects. In general, linking visual and linguistic structures allows the system to make use of pictorial information when discussing or reasoning verbally, and/or to use verbal information when reasoning about visual inputs.

The CHILDLIKE system builds up semantic memories that encode visual features, linguistic features, their compounds and rules formed from such entities. These memories help the system to reason about objects, object qualities, and the effect of certain actions on objects. For an example of memories built in response to inputs such as pictures of chairs and tables, along with words about them, see Figure 1 (only a cross section of the memories is shown). The hatched nodes correspond to structures elicited from the linguistic channel and the plain nodes correspond to visual features and their compounds. Some, but not all, of the highly weighted links are shown. The three numbers alongside a visual feature node represent, respectively, the x and y coordinates of the feature (in the *object*-centered coordinates of its parent node) and a relative size value.

An overview of the memory structures acquired after a series of experiences about everyday objects and actions[4] is shown in Figure 2. The dashed line represents links that would be formed as a result of learning words about actions. The mutual grounding of the representations derived from the sensory channels is one the salient aspects of the semantic memories acquired by our approach.

Learning Numbers and Elementary Arithmetic

Although arithmetic is often considered an abstract skill, very young children seem to be adept at performing visual arithmetic operations such as putting together several small piles of objects to form a larger pile (addition) or dividing a pile of objects approximately equally among three people (division by three). Bootstrapping on such abilities can ground learning of abstract numbers and simple arithmetic operations in perceptual experiences. Once the simple arithmetic operations are grounded in concrete experiences (e.g. via the visual modality), we conjecture that learning of more complex arithmetic operations becomes easier.

In a computational model for learning about numbers and operations such as addition, subtraction, multiplication and division, an approach similar to the one described in the previous section can be used. First, numbers can be taught by showing pictures of, e.g. two pears, two apples, or two chairs along with the corresponding language descriptions for them (e.g. using "2," "2 pears," "two apples," or "two chairs"). This results in links between the visually learned concepts of the numbers and the abstract symbols. Following this, on the visual channel, pictures representing three apples and two apples are presented (either alongside each other or in quick succession); along with the linguistic channel's input — an utterance like "3 + 2 = 5" or "3 and 2 make 5". The idea here is that experiences where several groups of apples (or sticks or whatever objects are shown) are merged together along with the symbolic description can lead to the association of the symbol "+" with the visual operation of combining two piles. Note that the learner is forced to come up with or identify a visual operation that results in a picture of 5 objects from pictures

[3] Typically, the more parsimonious representation (here language) can be advantageously used to *bootstrap* on already learned, complex associations.

[4] Currently, an abstract action channel is used to encode and input actions. Eventually, we would like CHILDLIKE to interpret significant changes in its perceptual states as actions and also to learn words describing simple actions.

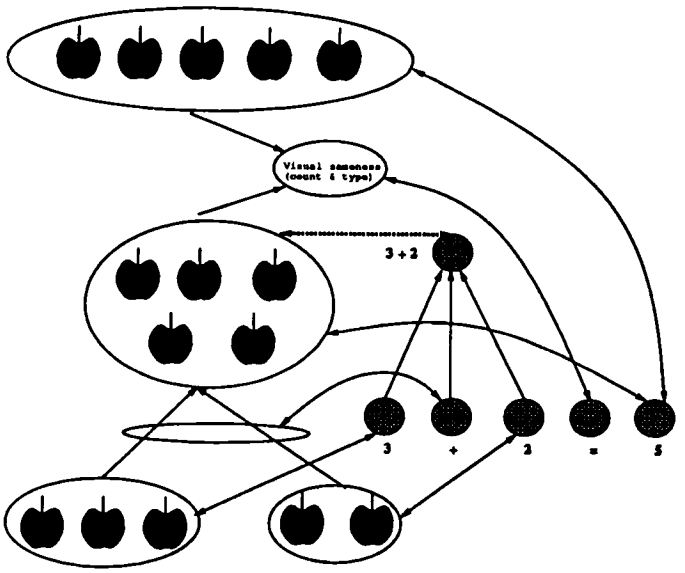

Figure 3: **Possible Memory Structures Acquired in Learning about Numbers and Elementary Arithmetic**

of 3 and 2 objects (since prior training has established the association between the numbers and their corresponding pictures). An example of what the memory structures look like following such a training regimen is shown in Figure 3. The hatched nodes represent features derived from the linguistic channel. Only the highly weighted links are shown. Figure 3 also assumes that the visual nodes are represented using one of the previously learned entities (apples here); following generalization, an abstract node for two things may be created and the number "2" from the linguistic channel may be linked to this node rather than being linked to the node representing two apples, to another node representing two chairs and so on. Similarly, the links from "3" and "+" will also undergo generalization.

Some elementary visual operations, such as visual juxtaposition, may already exist in the human visual system. More complex visual operations may have to be learned in terms of these elementary visual operations; exactly what is learned and stored may depend on the learner's experiences, the utility of the learned visual operations and the learning mechanisms that actually extract structures of information, generate new links between them, and search for good weights on these structures and links. The visual operations we are referring to are similar in spirit to the visual routines (such as shift, indexing and marking) postulated by Ullman (1984).

Subtraction can be taught in ways similar to addition— for example, by showing pictures of 5 things and 2 things, followed by 3 things along with the corresponding linguistic description (such as "$5 - 2 = 3$"). Multiplication can be taught as a series of additions linked together by the needed storage of intermediate results (including carries) and shiftings from each step to the next; and division as a series of subtractions (this is very similar to the training sequences needed to expedite the acquisition of successively more complex multi-step rules for action sequences in CHILDLIKE). It is important to point out that only arithmetic operations involving very small numbers need be visually grounded. For example, single-digit addition or subtraction acts as a crucial building block for learning the more abstract sequence of procedures needed for multi-digit addition or subtraction. The grounding of the simple arithmetic operations also helps build abstract knowledge of the form "When things are divided among people, each person gets an exactly equal part of the original pile ... hence division reduces or distributes things ...". Such abstract knowledge — some of which is related to and seems to be acquired under the guise of arithmetic — is used in a number of situations in everyday-reasoning (like allocating time across different activities, calculating approximate distances, or judging weights).

Conclusions

It appears that humans consciously build and manipulate representations while solving problems, including simple ones involving reasoning about everyday objects and events. A number of researchers have analyzed the possibly different representations used by humans in different cognitive tasks (e.g. see Fodor, 1976; Kosslyn, 1990; Larkin & Simon, 1987; Narayanan & Chandrasekaran, 1991; Pylyshyn, 1981) — these representations can be chiefly categorized as visual or linguistic (possibly based on the origin of the basic representational units).

In this paper, we have suggested that it is not crucial to postulate or choose one dominant (or desirable) type of representation. In particular, humans appear to learn and effectively use links that help mutually ground linguistic and visual information. A computational model (the CHILDLIKE system) that employs such a mechanism was briefly described. The CHILDLIKE system learns words about objects and object qualities, and also the effect of simple actions on object relationships. It appears that the same model can also be used for learning numbers and simple arithmetic operations.

References

1. Carbonell, J.G., Michalski, R.S. and Mitchell, T.M. 1983. An Overview of Machine Learning. *In:* R.S. Michalski, J.G. Carbonell & T.M. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach.* Palo Alto, CA: Tioga.
2. Fodor, J. 1976. *The language of thought.* Harvester Press, Sussex. (Harvard University Press paperback)
3. Harnad, S. 1990. The Symbol Grounding Problem. *Physica D, 42.*
4. Kosslyn, S.M. 1990. Mental Imagery. *In:* D.N. Osherson, S.M. Kosslyn & J.M. Hollerbach (eds.), *Visual Cognition and Action: An Invitation to Cognitive Science, Vol. 2.* Cambridge, MA: MIT Press/Bradford Books.
5. Larkin, J.H. and Simon, H.A. 1987. Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science, 11.*
6. Li, Z.N. and Uhr, L. 1987. Pyramid vision using key features to integrate image-driven bottom-up and model-driven top-down processes. *IEEE Trans. on Systems, Man, and Cybernetics 17.*
7. Mani, G. 1990. Learning by Gradient Descent in Function Space. *IEEE International Conference on Systems, Man and Cybernetics,* Los Angeles.
8. Mani, G. and Uhr, L. 1991a. Integrating Perception, Language Handling, Learning and Planning in the CHILDLIKE System. Working notes of the *AAAI Spring Symposium on Integrated Intelligent Architectures.* Also in the special issue of *SIGART* on *Integrated Cognitive Architectures,* August 1991.
9. Mani, G. and Uhr, L. 1991b. Perception–mediated Learning and Reasoning in the CHILDLIKE System. *Proceedings of the 13th Annual Cognitive Science Society Conference,* 1991.
10. Michalski, R.S. 1983. A Theory and Methodology of Inductive Learning. *In:* R.S. Michalski, J.G. Carbonell & T.M. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach.* Palo Alto, CA: Tioga.
11. Narayanan, N.H. and Chandrasekaran, B. 1991. Reasoning Visually about Spatial Interactions. *Proc. of the 12th IJCAI,* Mountain View, CA: Morgan Kaufmann.
12. Newell, A. 1980. Physical Symbol Systems. *Cognitive Science, 4.*
13. Pylyshyn, Z.W. 1981. The imagery debate: Analogue media versus tacit knowledge. *Psychological Review, 88.*
14. Uhr, L. (Ed.) 1987. *Parallel Computer Vision.* Boston: Academic Press.
15. Uhr, L. 1990. Increasing the Power of Connectionist Networks (CN) by Improving Structures, Processes, Learning. *Connection Science, 2 (3).*
16. Ullman, S. 1984. Visual routines. *Cognition, 18.*