Similarity-based acquisition of spatial categories

Pantelis Papadopoulos

Computer Science Department
Indiana University, Bloomington
pantelis@cs.indiana.edu

Abstract

A very natural approach to categorization is similarity-based clustering. We propose a visual representation which can be used with such a mechanism for the acquisition of spatial relation categories. We also show how supervision can be helpful in cases where basic similarity is low by proposing a learning mechanism which operates both in the presence and in the absence of words.

Motivation

Words signifying spatial relationships between objects are quite common in human languages. Such words, like the English ABOVE, cannot be adequately explained in terms of other words (Harnad, 1990). The core of their meaning is rather directly associated with visual perception. This means that to a significant extent, the acquisition of spatial relation concepts has to be independent of language (Choi and Bowerman, 1991, Regier, 1992).

This, in turn, implies that humans have some innate capability to automatically form spatial relation concepts, by noticing regularities in their visual input (unsupervised learning). When words are present (supervised learning), they help refine those concepts to their intended meaning within the language being used.

Modeling

This acquisition process can be viewed as categorization of instances of various spatial configurations. And a very natural approach to categorization (especially unsupervised) is similarity : if two instances are "sufficiently" similar, they belong to the same category (Goldstone, 1994a, 1994b, Rosch and Mervis, 1975, Smith, 1992). In order to apply such a criterion in the case of spatial configurations, we have to specify representations for our instances and also a way in which their similarity can be judged.

Input Representation

On the representation issue, we could start with a picture ; this is what humans get as raw input. The first step in recognizing a spatial relation in the picture, is definitely to find the objects that take part in it. Then, we need a way to represent the locations of these objects within the picture, since this information is what matters for defining a spatial relationship.

It's interesting to observe that, for the kinds of spatial relation categories used in human languages, specific attributes of the participating objects are irrelevant (a "red elephant ABOVE a green mouse", is as good an instance of ABOVE as an "angry sky ABOVE a calm sea"). What is really essential to know is just which segments of the picture correspond to different objects.

Also, if we're just concerned with relations between two objects, which we'll call Trajector and Landmark (Langacker, 1987), the essential location information needed for specifying the spatial relationship, is position of the Trajector relative to the Landmark. The exact location of the Landmark in the picture is irrelevant ("a book (Trajector) is ON the table (Landmark)" no matter where "the table" is located in the visual field).

We chose to incorporate the abstractions over object specificity and Landmark position in our input representation. We assume the existence of a first processing stage that takes a 2D visual bitmap and outputs a 2D map where a number is attached to every map bit, which signifies the object it belongs to (Mozer, Zemel and Behrmann, 1991). There's a unique number for each object, and for the sake of simplicity, we'll assume that Trajector bits are marked with 1, Landmark bits with 2 and background bits with 0. These numbers tell us whether two bits belong to the same or different objects, but contain no information about the exact identities of the objects involved (a "red elephant" Trajector and an "angry sky" Trajector will just be two regions of 1-bits in the 2D map). The shape of the objects is retained in this representation.

A second processing stage, which takes the previous 2D map as input, produces a representation with explicit

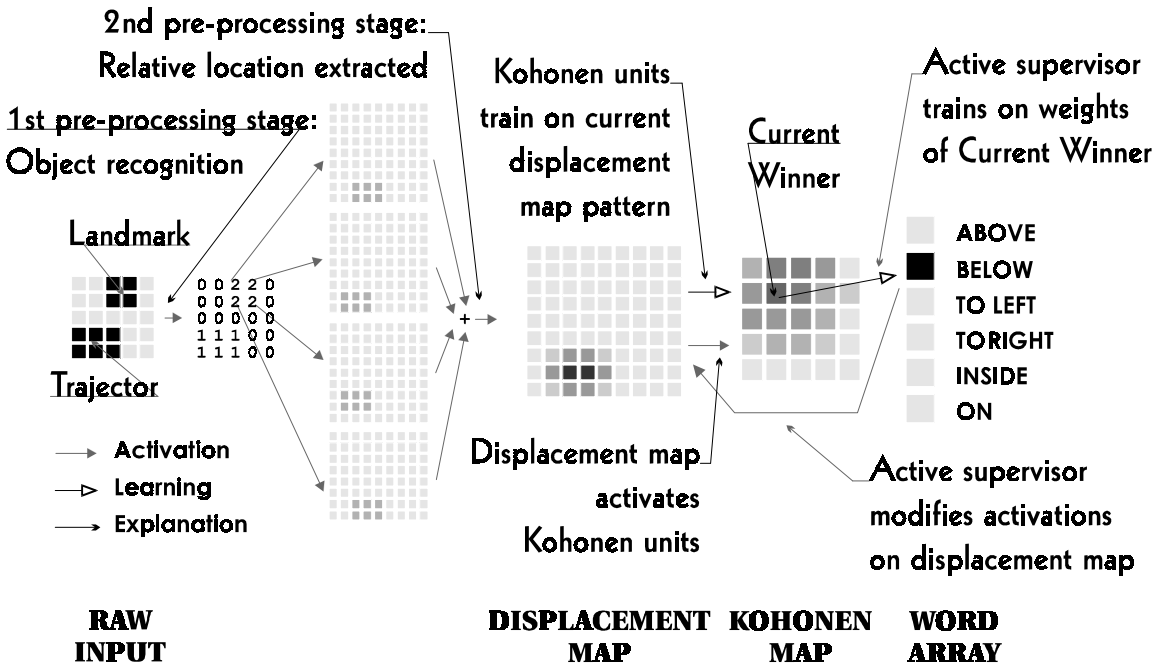

Figure 1: An overview of the architecture. Filled arrowheads indicate flow of activation and white ones direction of learning. The stage being pointed is always the one that is affected.

information about the position of 1-bits (Trajector) relative to 2-bits (Landmark). It's easy to imagine such a representation if there's just one Trajector bit and one Landmark bit : if their absolute coordinates are (Xt, Yt) and (Xl, Yl), we can use (Xt-Xl, Yt-Yl). This just tells us the way we have to go to find the Trajector, if we start from the Landmark. It's also relatively easy to represent location of a group of Trajector bits, relatively to a single Landmark bit. For each Trajector bit (Xti, Yti), we can use (Xti-Xl, Yti-Yl). This implies that to hold the relative location information, we could use a 2D displacement map, where if element (X, Y) is 1, it means that there's a Trajector bit which can be reached from the single Landmark bit.

But when there's more than one Landmark bit, it's not clear where we should be centering the coordinate system. One possibility might be some characteristic point in the Landmark, like its center of "mass". A possible problem with such approaches is that they abstract information about the size and shape of the Landmark, which is important for certain spatial relation categories (like SIKI vs SINI in Mixtec). A way to include such information in a simple manner, is to create displacement maps for all bits of the Landmark and then superimpose them into a single displacement map.

Similarity-based clustering

The superposition displacement map is used as input by both unsupervised and supervised learning.

Unsupervised learning. The unsupervised process, merely notices similarities between different inputs by evaluating their degree of overlap. Highly overlapping inputs should get clustered together, non-overlapping ones should be far apart. Such a process can be computationally realized in a very elegant manner through the use of a Kohonen map and Kohonen map learning (Kohonen, 1982).

Kohonen map units specialize to the detection of frequently occurring input patterns, by developing weight connections that maximally respond to them (a kind of Hebbian learning). In our version of learning, for each input pattern,there's a winner unit that is specialized the most, and a region of units around it that are specialized less and less as the distance from the winner increases. This ensures that winners of highly similar patterns are either going to eventually merge, or be physically close on the Kohonen map.

Supervised learning. For the supervised part, we directly represented spatial relation words with individual units. These units are either active or non-active and there can only be at most one active unit at any given time instant. The presence of an active unit, signifies the presence of the corresponding spatial relation word in the auditory input signal. These units are slowly associated with visual input.

Again, as in unsupervised learning, each word unit is specializing to respond maximally to the visual patterns that are present when it is active. Unlike unsupervised learning, such patterns may or may not have high degree of

overlap. In order to ensure maximal response, the word units specialize to the overlap, no matter how small that is.
This means that these units are sensitive to activation deviations over time. Visual input units that have wildly varying activations in different patterns that are present when the word unit is active, are going to be ignored. As consistency increases, so does the strength of association between the word unit and the corresponding visual input units. In the current implementation, high consistency in being active results in high valued weights and high consistency in being inactive receives weights close to 0 (which is the same as what happens with low consistency).

Interaction. The development of unsupervised (pure similarity driven) winners on the Kohonen map and supervised (word presence driven) winners are closely coupled. Every time a word is present, it affects the visual input by reinforcing input units to which it maximally responds, and suppressing ones that it has learnt to ignore or that it consistently found inactive in the past.
This flow of control from word units to visual input might correspond to a word-driven attention process. It results in increasing the similarity of different instances of the same language-relevant spatial relation and consequently making it more possible for a single -or closeby- Kohonen units to cluster them together.
Additionally, every time an active word unit is entrained, it does not directly examine the visual input, but, instead, examines the current Kohonen map winner, and performs weight adjustments based on the values of the winner's weights. This amounts to using not only current evidence for supervised learning, but, also, past experience.
A highly occurring pattern or group of similar (overlapping) patterns will develop a strongly responding Kohonen winner, with weights that closely match them. Then supervised learning will be fast, since it will be based on highly consistent input, the weights of the single Kohonen winner. It will take longer to learn categories with highly dissimilar instances, since they are going to initially develop different Kohonen winners, which will also be probably responding to patterns belonging to other categories. Then, at least initially, the relevant word units are going to be receiving misleading information from the Kohonen map winners.

Implementation

Simulation

In order to test the efficiency of our modeling specifications, we used the visual representations created by 75 7x7 picture bitmaps which produce 75 patterns on a 13x13 displacement map (displacement on a scale of 0-7 can be from -6 to +6) to train a Kohonen map of 5x5 units and an array of 12 word units, corresponding to certain spatial categories. Each picture bitmap was a clear example of one or more of those 12 categories. During training, there was a 25% probability that exactly one of the appropriate word units would be activated. This is a rough simulation of what happens in real life. Training was repeated until either the weights on the Kohonen map stabilized (change was consistently under a threshold) or an oscillation was detected.
In oscillations, the Kohonen weights cycle continuously through a number of value assignments. In each state in the cycle, at least as far as we could tell, the same clusters are formed but are represented in different places on the Kohonen map. Oscillations occur depending on the initial values of weights and the training set. For our particular training set, they occurred around once every ten times we trained.

Results

When training stops, clusters have been formed on the Kohonen map and the word units have developed certain specializations. In order to determine how closely these correspond to groupings made by humans, for the patterns of each category, we recorded their winners on the Kohonen map. Then we computed the average location of a winner for each category and the location dispersion. For successful category learning, we should get 0 or small winner location dispersions.
It's also interesting to notice whether similar categories have average locations that are closeby and whether their dispersion values define a high degree of overlap. To help such observations, for each pair of patterns we defined circles of influence centered at the location of the prototype winners and with radii equal to the respective location dispersions. Then, we computed the ratio of common area of the two circles over the total area they occupy.
Finally, direct inspection of the weights of the word units, would reveal how successfully the categories are learnt. It may be the case that our initial visual representation is too impoverished to allow learning of certain categories. In other cases, supervision might not be needed at all, because the visual representation forces instances of certain categories to have high degree of overlap by embodying the right kinds of abstractions. To determine the extent to which this is true, we performed an additional series of simulations in which supervision was turned off.

	with supervision			without supervision		
category	avg X	avg Y	dispersion	avg X	avg Y	dispersion
above	4	3.94	0.24	0.06	0	0.24
below	2.23	2	0.42	2	1.54	0.84
left	2	3.67	0.47	0.33	2	0.47
right	4	1.25	0.43	2.75	0	0.43
touch	3.17	2.80	1.28	1.07	0.72	1.35
inside	3.67	2.5	0.90	1.17	0.33	1.01
encircling	4	2.75	0.43	0.75	0	0.83
sideways	2.94	2.53	1.63	1.47	1.06	1.63
protect	4	4	0	0	0	0

Table 1: Dispersions get smaller with supervision. Certain categories (siki, sini, on) have been omitted since they are identical with the category PROTECT (where the Trajector is above the Landmark and completely covering it vertically).

	above	below	left	right	touch	inside	encirc	side	prot
above	100%	0%	0%	0%	1%	0%	0%	0%	100%
below	0%	100%	0%	0%	6%	0%	0%	7%	0%
left	0%	0%	100%	0%	3%	0%	0%	6%	0%
right	0%	0%	0%	100%	0%	0%	0%	3%	0%
touch	1%	6%	3%	0%	100%	43%	11%	62%	0%
inside	0%	0%	0%	0%	43%	100%	23%	30%	0%
encircling	0%	0%	0%	0%	11%	23%	100%	7%	0%
sideways	0%	7%	6%	3%	62%	30%	7%	100%	0%
prot. vert.	100%	0%	0%	0%	0%	0%	0%	0%	100%

Table 2: Ratios of circle area overlap over total circle area for all pairs of categories. The centers are (Avg X, Avg Y) from Table 1 and the radii are the corresponding dispersions.

The results indicate that certain categories (like TO_THE_LEFT, BELOW) can be acquired without supervision. Other categories (like ABOVE, TOUCH, SIDEWAYS) rely more on supervision. The rather intriguing difference in performance between ABOVE and BELOW should be attributed to the particular training set used. With a slightly expanded set (the 75 patterns plus 6 more) we got exactly the opposite result : ABOVE was better than BELOW.

Discussion

It is noteworthy that during purely unsupervised learning the network showed remarkable stability and always converged on the same clusters, even though these might be realized in different regions of the Kohonen map at different runs. Also, the supervised learning seems to be doing an excellent job of abstracting the best possible regularities for each category, and helping the Kohonen map converge to clusters that are closer to the intended meanings of words (contrast BELOW without and with supervision, also same with ENCIRCLING).

On the negative side, as the strength of supervision increases, it becomes more and more difficult for the network to converge to a stable state. Also, it seems that the coupling from unsupervised to supervised learning is really weak ; no matter what the strength of supervision is, the word unit weights always converge on the same optimal values. This would lead to the conclusion that the unsupervised learning component is really useless. Our only objection is that supervised learning is possibly faster when unsupervised learning is cooperating. But since we didn't measure speed of convergence, we can't support such a claim.

Clearly, the coupling of the two modes of learning needs further development. There is also the need to analyze how different values of the several operating parameters affect performance. Up to now, we have experienced a few cases where certain values lead to severe degradation in performance. We also need to train with more patterns to see how much our results (good and bad) depend on the specific patterns we've been using.

But, in overall, the tendency of the network is to become better with time -at least in the long run-, and that is an encouraging result, which strengthens our belief that the acquisition of spatial categories can be based on simple similarity, if the "right" input representation is used.

Our whole approach was largely motivated by the work of Schyns (Schyns, 1991) on general concept acquisition. The unsupervised part of our architecture is essentially identical to the one he uses. Schyns, however, worked mainly on the formation of object prototypes, from noisy input that is always centered. Such a process cannot be directly applied to raw visual input containing visual relations between objects. By assuming preprocessing that leads to our displacement map, unsupervised spatial concept formation becomes possible. Also, Schyns seems to underestimate the importance of supervised learning. He claims that a word can only be learnt after unsupervised learning has finished forming the appropriate concept. The presence of the word while unsupervised learning is in progress is irrelevant.

Dorffner (Dorffner, 1991) is also in favor of a strong unsupervised learning component. He allows for interaction between auditory and visual signals by assuming the formation of internal symbols that associate them. That interaction, however, seems to be going only one-way, from perception to higher cognition. Correct internal symbols can only be formed once the appropriate auditory and visual concepts have been independently formed.

We would agree that a strong unsupervised component is needed for such a perceptually oriented task as spatial concept acquisition, as far as it justifies our assumptions of preprocessing. There's just too much going on in what an infant perceives, for it to attain spatial categories just by hearing the correct words in the correct contexts, without any inborn attentive predispositions. How far these extend, is we think an open question. Regier (Regier, 1992) has achieved remarkable results in the field of spatial concept acquisition, both static and dynamic, across languages, by assuming a considerable amount of preprocessing, which is all hardwired in his architecture and precedes learning.

We think we have shown that with relatively simple and reasonable preprocessing assumptions, both unsupervised and supervised learning can become more efficient. We have also attempted to show how the acquisition process can benefit from the continuous interaction of these two modes of learning. This last issue is subject to future development.

Acknowledgments

Michael Gasser and Eliana Colunga have greatly contributed to the conception of the ideas behind this work. Doug Blank provided a sensible programming template, upon which all our simulations were based.

References

Choi, S., and Bowerman M. 1991. Learning to Express Motion Events in English and Korean: The Influence of Language-specific Lexicalization Patterns. *Cognition*, 41:83-121.

Dorffner, G. 1991. "Radical" Connectionism for Natural Language Processing. *AAAI Spring Symposium*.

Goldstone, R. L. 1994a. The role of Similarity in Categorization: Providing a Groundwork. *Cognition*, 52:125-157.

Goldstone, R. L. 1994b. Similarity, Interactive Activation and Mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):3-28.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D*, 42:335-346.

Kohonen, T. 1982. Self-organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43:59-69.

Langacker, R. 1987. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford University Press, Stanford.

Mozer, C. M., Zemel, R. S., and Behrmann M. 1991. Learning to Segment Images Using Dynamic Feature Binding, Technical Report, CU-CS-540-91, University of Colorado at Boulder.

Regier, T. 1992. The Acquisition of Lexical Semantics for Spatial Terms : A Connectionist Model of Perceptual Categorization. Ph.D. diss., Dept. of Computer Science, University of California at Berkeley.

Rosch, E., and Mervis, C. B. 1975. Family Resemblance: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7:573-605.

Schyns, P. 1991. A Modular Neural Network Model of Concept Acquisition. *Cognitive Science*, 15:461-508.

Smith, L. B. 1992. A model of Perceptual Classification in Children and Adults. *Psychological Review*, 96:125-144.