Proceedings of the European Cognitive Science Conference, Delphi, Greece, May 2007, pp. 312–317

A Generalization of Hebbian Learning in Perceptual and Conceptual Categorization

Harry E. Foundalis (harry@cogsci.indiana.edu)
Center for Research on Concepts and Cognition, Indiana University
Bloomington, IN 47408, USA

Maricarmen Martínez (m.martinez97@uniandes.edu.co)
Department of Mathematics, Universidad de los Andes
Bogotá, Colombia

Abstract

An algorithm for unsupervised competitive learning is presented that, at first sight, appears as a straightforward implementation of Hebbian learning. The algorithm is then generalized in a way that preserves its basic properties but facilitates its use in completely different domains, thus rendering Hebbian learning a special case of its range of applications. The algorithm is not a neural network application: it works not at the neural but at the conceptual level, although it borrows ideas from neural networks. Its performance and effectiveness are briefly examined.

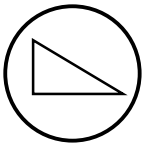
Introduction

Traditionally, learning has been considered to be one of the foundational pillars of cognition. In 1949, Donald Hebb expressed the basic idea of *association learning* as follows: when two neurons are physically close and are repeatedly activated together, some chemical changes must occur in their structures that signify the fact that the two neurons fired together (Hebb, 1949). Neuroscientists and cognitive science researchers dubbed this idea "Hebbian learning", and took it to mean that whenever two percepts appear together, we (or animals in general) learn an association between them. The "how" is usually left to algorithms in artificial neural networks (ANN's). James McClelland points out in a recent study that the potential of Hebbian learning has been largely underestimated (McClelland, 2006). McClelland focuses on the neuronal level, but refers also to work at the conceptual level (the focus of the present article); specifically, in *categorization* (e.g., Schyns, 1991) and in self-organizing topological maps (Kohonen, 1982).

In what follows, an algorithm that uses Hebbian learning at the conceptual level is introduced through an example that appears in human cognition. It is then generalized in a way that, while retaining its basic properties, makes it applicable to problems that involve the detection (and hence the learning) of sets of entities that are most closely related to each other.

The Basic Algorithm

Suppose we are given input consisting of a pair of elements, i.e., a drawing depicting some familiar objects, and a phrase with identifiable words, the meaning of which is unknown (Figure 1). The phrase is supposed to be "about" the objects, their properties, and/or relations. The problem is to find the correct associations between words and percepts; or, stated otherwise, to learn the *meaning* of the words (in some rudimentary sense of the word "meaning"[1]).



nae triogon aems es nae cycol

Figure 1: A drawing, and its associated phrase

The drawing in Figure 1 shows two familiar geometric objects, and the phrase is given underneath the objects in italics. The learner is not expected to know any words for these objects or their properties in some other language; indeed, the phrase might originate from what will turn out to be the learner's native language.[2] The learner will receive a number of such examples, pairing visual and linguistic data, and the algorithm described below will discover the correct associations between words and visual percepts. This problem has been examined also by Deb Roy (Roy, 2002), but Roy's learning succeeds by batch processing, whereas the learning described in the present article is incremental.

Suppose the learner is capable of perceiving some objects, features, and/or relations — collectively called *percepts* — by looking at the visual input, which activate corresponding *concepts* in long-term memory (LTM). Also, the learner can identify (i.e., separate from each other) the words in the given phrase.[3] In reality, an infant learning a language will not identify all the words in a phrase, but it does not harm the generality of our algorithm to assume so.

As a first — perhaps naïve — step in our effort to discover which word corresponds to which percept, let us adopt the following simplistic strategy: associate every concept with every word. Figure 2 depicts this idea. LTM concepts that were activated by percepts in the input are listed on the top row, and the words of the given phrase are added to LTM and shown on the bottom row. Notice that

[1] Contrary to our simplifying assumption, in reality words and percepts are not associated according to a 1–1 correspondence.

[2] However, infants do not become native speakers by hearing phrases that refer to geometric objects. What is described here is an abstraction of a real-world situation.

[3] Thus, suppose the "word-segmentation problem" is solved.

each word is shown only once (for instance, "*nae*", which is repeated in the input, is not repeated in Figure 2).

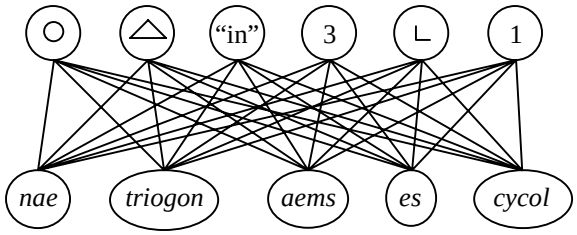
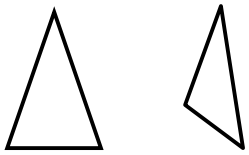


Figure 2: First uninformed step in the algorithm

The top row of nodes in Figure 2 shows only *some* of the LTM concepts that could be activated by the input; specifically, "circle", "triangle", "in", "three-ness", "right angle", and "one-ness". A look by a different agent (or by the same agent at a different time) might activate different concepts, such as "round", "surrounds", "slanted line", etc. Those shown in Figure 2 are the ones that happened to be activated most strongly in the given learning session.

Next, the learning session continues with a second pair of visual and linguistic input.



doy triagon, ot nae ainei scelisoes

Figure 3: A second pair of visual and linguistic input

The image in Figure 3 activates again some concepts in LTM, some of which were activated also by the input of Figure 1 (e.g., "triangle"). Linguistically, there are some old and new words. Also, one word looks very much like an old one ("*triagon*" vs. "*triogon*"). A suitable word-identification algorithm would not only see the similarity, but also analyze their morphological difference, which would later allow the agent to make an association between this morphological change and the visual percept that was responsible for the change. But at this stage no handling of morphology is required by the algorithm; suffice it to assume that the two words, "*triagon*" and "*triogon*", are treated as "the same".

The algorithm now adds the new concepts to the list at the top (that is, it forms the union of the sets of the old and new concepts) and the new words to the list at the bottom.

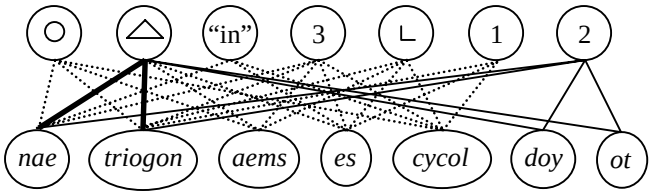


Figure 4: Some associations are reinforced

For lack of horizontal space, only a few of the new words and concepts are shown in Figure 4. The important development in Figure 4 is that those concept – word pairs that are repeated in inputs 1 and 2 (such as △ – "*triogon*" and △ – "*nae*") have their associations *reinforced* (shown in Figure 4 with bold lines), whereas the associations of those pairs of input 1 that did not co-occur in input 2 have *faded* slightly (shown in Figure 4 with dotted lines). At the same time, as before, the added words and concepts form all possible associations (lines of normal thickness).

The expectation is now clear: given more instances of paired visual and linguistic input (not too many — at most a few dozen) the "correct" associations (i.e., those intended by the input provider) should prevail, while those that are "noise" (unintended) should be eliminated, having faded beyond some detection threshold.

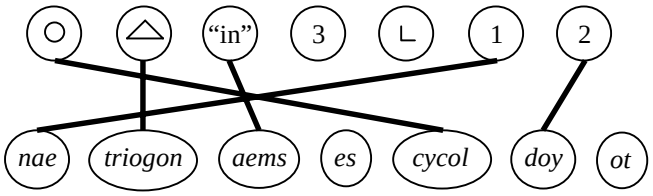


Figure 5: Desired (final) set of associations

Figure 5 shows an approximation of the desired result of this procedure. (There will be many more concepts and words introduced on the two rows of nodes, and most, but not all associations will be "correct": there will be spurious associations that did not fade to the point of elimination.)

Activations of Associations

For the above-outlined algorithm to converge to the desired associations, a precise mechanism of association fading and reinforcement must be established. If the associations fade too fast, for example, they will all be eliminated before additional evidence arrives making them strong enough to survive to the end of the process; if they fade too slow, all associations (including "noise") will eventually survive; finally, if a minute degree of fading is allowed even after the system receives no further input (waiting, doing nothing), then given enough time all associations will drop back to zero and the system will become amnesic. To counter these problems, the "strength" of each association, which we call *activation*, is expected to have the following properties.

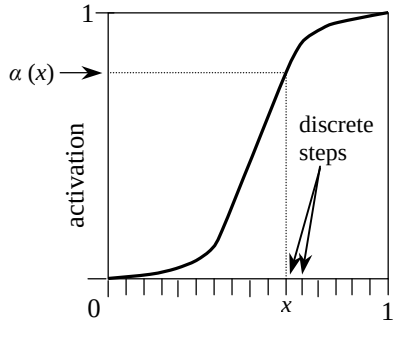


Figure 6: How activation changes over time

The activation of an association is the quantity $\alpha(x)$ on the y-axis of Figure 6. Suppose that, at any point in time, there is a quantity x on the x-axis ranging between 0 and 1, but that can take on values only along one of the discrete points shown on the x-axis. (The total number of these points is an

important parameter, to be discussed soon.) Thus, x makes only discrete steps along the x-axis. Then the quantity $\alpha(x)$ is given by the sigmoid-like function α, also ranging between 0 and 1. Below, we explain how and when x moves along the x-axis, and why α must be sigmoid-like.

Activations *fade* automatically, as time goes by. This means that if a sufficient amount of time elapses, x makes one discrete step backwards in its range. Consequently, $\alpha(x)$ is decreased, since α is monotonic. "Time" is usually simulated in computational implementations, but in a real-time learning system it is assumed to be the familiar temporal dimension (of physics).

Activations are *reinforced* by the learning system by letting x make a discrete step forward in its range. Again, the monotonicity of α implies that $\alpha(x)$ is increased. However, when x exceeds a threshold that is just before the maximum value 1, the number of discrete steps along the x-axis is increased somewhat (the resolution of segmenting the x-axis grows). This implies that the subsequent fading of the activation will become slower, because x will have more backward steps to traverse along the x-axis. The meaning of this change is that associations that are well established should become progressively harder to fade, after repeated confirmations of their correctness. The amount by which the number of steps is increased is a parameter of the system.

Finally, an explanation must be given for why α must be sigmoid-like. First, observe that α must be increasing strictly monotonically, otherwise the motion of x along the x-axis would not move $\alpha(x)$ in the proper direction. Now, of all the monotonically increasing curves that connect the lower-left and upper-right corners of the square in Figure 6, the sigmoid is the most appropriate shape for the following reasons: the curve must be initially increasing *slowly,* so that an initial number of reinforcements starting from $x = 0$ does not result in an abrupt increase in $\alpha(x)$. This is necessary because if a wrong association is made, we do not want a small number of initial reinforcements to result in a significant $\alpha(x)$ — we do not wish "noise" to be taken seriously. Conversely, if x has approached 1, and thus $\alpha(x)$ is also close to 1, we do not want $\alpha(x)$ to suddenly drop to lower values; α must be *conservative,* meaning that once a significant $\alpha(x)$ has been established it should not be too easy to "forget" it. This explains the slow increase in the final part of the curve in Figure 6. Having established that the initial and final parts must be increasing slowly, there are only few possibilities for the middle part of a monotonic curve, hence the sigmoid shape of curve α.

Implementation

The above-described principles were implemented as part of *Phaeaco,* a visual pattern-recognition system (Foundalis, 2006), in which the visual input consisted of a 200×200 rectangle of black-and-white dots. A "training set" consisted of 50 pairs of the form [image, phrase], where the images were similar to those shown earlier in Figure 1 and Figure 3 (they contained geometric figures), and phrases (in English) were likewise relevant to the content of their paired images. Phaeaco is capable of perceiving the geometric structure of such figures, building an internal representation of the structure in its "working memory" according to its architectural principles, and letting the parts of this representation activate the corresponding concepts in its long-term memory (LTM). Thus, if a square is drawn in the input, the following concepts might be activated in Phaeaco's LTM: *square, four, four lines, parallel lines, equal lengths, interior, four vertices, right angle, equal angles,* etc. (listed in no particular order). Some of these concepts will be more strongly activated than others, due to the principles in Phaeaco's architecture (e.g., *square* is more "important" than *vertex,* because a vertex is only a part of the representation of a square). Normally, Phaeaco starts with some "primitives", or *a priori* known concepts (such as *point, line, angle, interior,* etc.), and is capable of learning other concepts (such as *square*) based on the primitives.[4] However, for the purposes of the described experiment we suppressed Phaeaco's mechanism of learning new concepts, so as to avoid interference with the learning of associations between words and concepts. Thus, we worked with an LTM that already included composite concepts such as *square, triangle,* etc. — i.e., anything that might appear in the visual input of a training set.

An additional simplifying assumption was that the rudimentary morphology of English was ignored. Thus, words such as "triangle" and "triangles" were treated as identical; so were all forms of the verb "to be", etc.

The entire training set can be found in Martínez (2006).

Performance

The following graph (Figure 7) presents the progress of the learning of correct associations over time.

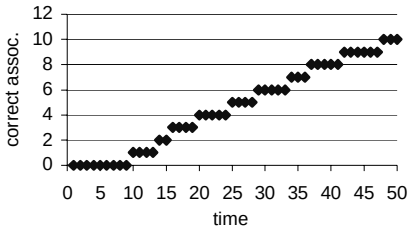


Figure 7: Progress of learning over time

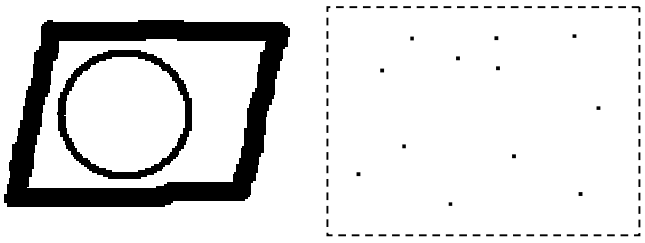
The graph in Figure 7 shows that the average number of correctly learned associations (y-axis) is a generally increasing function with respect to the number of input pairs presented (x-axis). Since the input pairs were presented at regular intervals in our implementation, the x-axis can also be seen as representing time. The y-values are averages of a large number (100) of random presentation orderings of the training set. The gradient of the slope of the curve (the "speed of learning") depends on a variety of factors, including the settings of the activation parameters and the content of the visual and linguistic input. Under most reasonable settings, however, the learning curve appears to slowly increase with respect to time, and this was the main objective of our implementation. In addition, notice that our algorithm requires only one presentation of the training set, as opposed to multiple epochs typically required in ANN's.

[4] This learning ability is completely unrelated to the algorithm of learning associations, discussed in the present text.

A Seemingly Different Application

The above-described process can be generalized in an obvious — and rather trivial — way, to any case in which data from two different sets can be paired, and associations between their members can be discerned and learned. For example, consider discovering the cause for an allergy. One set is "the set of all possible pathogens that could cause me this allergy", and the second set has a single member, the event (or fact) "I have this allergy". What is needed is an association of the form "pathogen X causes this allergy to me". We might observe the candidate causes during a long period of time, and *subconsciously* only, without actively trying to discover the cause. Over time, some candidates are reinforced due to repetition, whereas others fade. Given the right conditions (sufficient number of repetitions and not too long an interval of observation time forcing all associations to fade back to zero), we might reach an "Aha!" moment, in which one of the associations becomes strong enough to be noticed consciously.[5] Similarly, some examples of scientific discovery ("what could be the cause of phenomenon X?") can be implemented algorithmically by means of the same process. However, the generalization discussed in what follows goes beyond the pairing of elements of two sets.

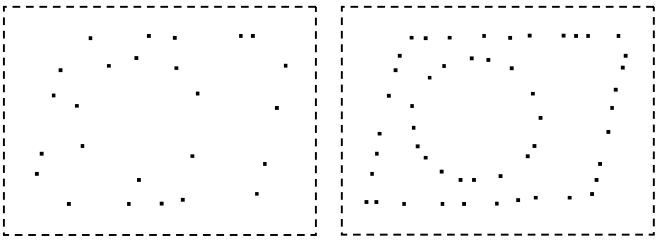
Suppose that the input is in the visual modality only, and consists of the shapes shown in Figure 8 (a).



(a) (b)

Figure 8 (a): Sample visual input; (b): a few pixels seen

Suppose also that a visual processing system examines the black pixels of this input, not in some systematic, top-to-bottom, left-to-right fashion, but randomly. Indeed, this is the way in which Phaeaco examines its input (Foundalis, 2006). After some fast initial processing, Phaeaco has seen a few pixels that belong to the central ("median") region of the parallelogram and/or the circle, as shown in Figure 8 (b).



(a) (b)

Figure 9 (a) – (b): More pixels seen

[5] It might be incorrect to make this association, thus concluding the wrong cause, but it is the process that concerns us here.

Shortly afterwards, a few more pixels become known, as shown in Figure 9 (a), where the outlines of the figures are barely discernible (to the human eye). Within similarly short time, enough pixels have been seen — as in Figure 9 (b) — for the outlines to become quite clear, at least to the human eye. What is needed is an algorithm that, when employed by the visual processing system, will make it as capable and fast in discerning shapes from a few pixels as the human visual system. To achieve this, Phaeaco employs the following algorithm.

As soon as the first pixels become known (Figure 8 (b)), Phaeaco starts forming hypotheses about the line segments on which the so-far known pixels might lie (Figure 10).

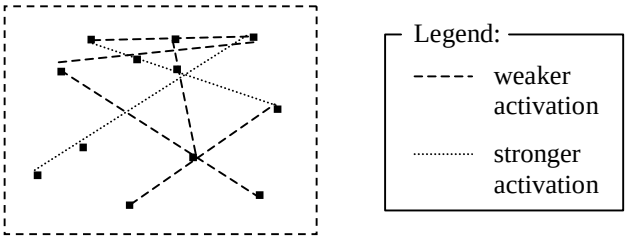


Figure 10: First attempts to "connect the dots" with lines

Most of these hypotheses will turn out to be wrong ("false positives"). But it does not matter. Any of these initial hypotheses that are mere "noise" will not be reinforced, so their activations will fade over time; whereas the correct hypotheses for line segments will endure. Thus, Phaeaco entertains *line-segment detectors*, which are line segments equipped with an activation value. The method of least squares is used to fit points to line segments, and the activation value of each detector depends both on how many points (pixels) participate in it, and how well the points fit. Note that only one of the detectors in Figure 10 is a "real" one, i.e., one that will become stronger and survive until the end (the nearly horizontal one at the top); but Phaeaco does not know this yet.

Subsequently more points arrive, as in Figure 9 (a). Some of the early detectors of Figure 10 will receive one or two more spurious points, but their activations will fade more than they will be reinforced. Also, a few more spurious detectors might form. But, simultaneously, the "real" ones will start emerging (Figure 11).

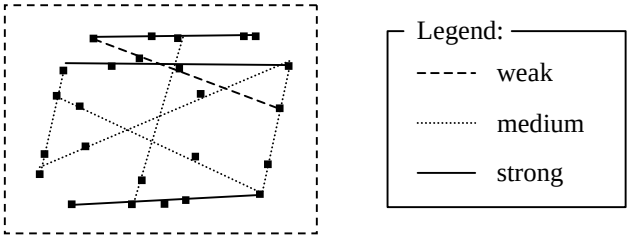


Figure 11: More points keep the detectors evolving

Note that several of the early detectors in Figure 10 have disappeared in Figure 11, because they were not fed with more data points and "died". Thus, the fittest detectors survive at the expense of other detectors that do not explain

the data sufficiently well. The last (but not final) stage in this series is shown in Figure 12, below.

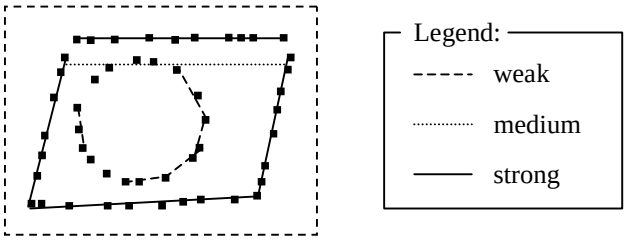


Figure 12: Most of the surviving detectors are "real"

In Figure 12, the desired detectors that form the sides of the parallelogram are among the survivors. Also, the tiny detectors that will be used later to identify the circle have started appearing. In the end, all "true" detectors will be reinforced with more points and deemed *the* line segments of the image, whereas all false detectors will fade and die.

The particular reasons why Phaeaco employs this rather unconventional (randomized) processing of its visual input are explained in Foundalis (2006). For the purposes of the present article it is important to point out that the above procedure is extendible to any case where the detection of the most salient among a group of entities (objects, features, etc.) is required. The reason why this process is similar to the Hebbian association-building will be discussed soon.

Generalizing to Categorization

The process described in the previous section is a specific application of the more general and fundamental process of *categorization*. An example will suffice to make this clear.

Suppose we are visitors at a new location on Earth where we observe the inhabitants' faces. Initially all faces appear unfamiliar, and, having seen only a few of them, we can do no better than place them all in one large category, "inhabitant of this new place", as abstracted in Figure 13.

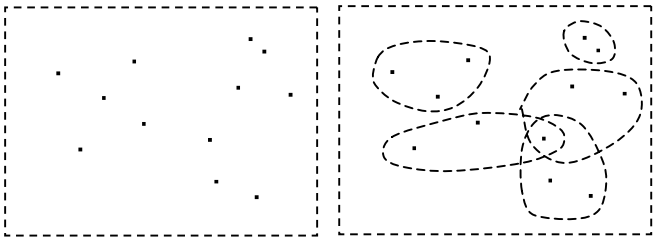


Figure 13: Categorization in abstract space of facial types

Note that the face-space on the left side of Figure 13 includes only two facial dimensions, *x* and *y*, since each dot represents a face. In reality the space is multidimensional, and we become capable of perceiving more dimensions as our experience is enriched with examples. An initial stage at categorization is shown on the right side of Figure 13. The dashed outlined regions are detectors of categories, almost identical in nature with the detectors of lines of the previous section. As before, new examples are assigned to the group in which they best fit, statistically. Over time, as more examples become available, the categories that are "noise" fade, and we end up with a clearer view of the correct categories in this space, as shown in Figure 14. (The space dimensionality has been kept constant, equal to 2, for purposes of illustration.)

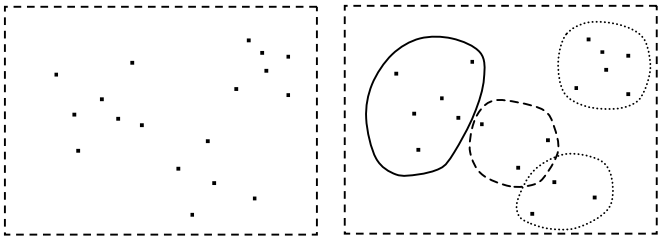


Figure 14: More points help to better discern the categories

The algorithm is described in detail in Foundalis (2006, pp. 228–235). In Phaeaco, categories formed in this way (see Figure 15) are defined by the barycenter ("centroid") of the set, its standard deviation, and a few more statistics. They also include a few of the initially encountered data points. Thus, they are an amalgam of the prototype (Rosch, 1975; Rosch and Mervis, 1975) and exemplar (Medin and Schaffer, 1978; Nosofsky, 1992) theories of category representation, following the principles of the FARG family of cognitive architectures (Hofstadter, 1995).

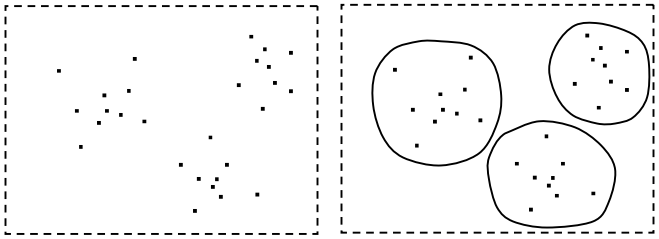


Figure 15: The categories finally become clear

The connection with the earlier discussion of line segment detection should now be clear: the same mechanism can be employed in both cases for the gradual emergence and discerning of the categories, or *concepts*, in the given space. The two cases, perceptual detection of lines and conceptual detection of categories, are very nearly *isomorphic* in nature. In general, the discerning of concepts (of any kind, whether concrete objects or abstract categories), might be thought to rely on a process similar to the one described here. For concrete objects, this process has been completely automated and has been hardwired by evolution, so that very few parameters of it need be learned; for other, more abstract and gradually discerned categories, it might look more like the mechanism outlined above.

Discussion

Two seemingly different case-studies were discussed in the previous sections: a Hebbian-like "association-building by co-occurrences" process, and one of gradual discerning of categories in a multidimensional space. What is the deeper commonality between the two?

Both are about *discerning* entities. In the case of Hebbian learning the entities are the associations, whereas in the

other examples mentioned the entities are categories that appear as clusters of objects. The discerning is gradual, and initially there is a lot of "noise" in the form of wrongly detected entities. But over time the "noise" fades away and the correct entities emerge highly activated.

This is an unsupervised learning algorithm, the success of which depends on the appropriate settings of the parameters of the *activations* of entities. Activations (illustrated in Figure 6) must have the following properties:

- The initial increase in activation must be *conservative*, to avoid an early commitment to "noise".
- The activation strength must be fading automatically as time goes by, to avoid having all activations of associations or categories eventually reach the value 1.
- Those associations or categories that appear to stand consistently above others in strength must curb the fading speed of their activations, to avoid forgetting even the most important ones among them over time.

It might be thought that our choice of setting the starting value of an activation to 0 results in a strictly deterministic process. However, this criticism is superficial. Although our model is indeed deterministic, in a real-world situation the order in which the input is encountered is practically never predetermined. Indeterminism arises from the real world. Thus, what is required is that any (random) order allows our algorithm to run, and indeed, in our measurements we varied the input presentation order randomly, observing no dependence of the algorithm on any particular input order.

Although ideas similar to the above have traditionally been viewed as belonging to the neural level that inspired ANN's, our work supports the idea that it is possible that the same principles have been utilized by human cognition at a higher conceptual level. This is not without precedent in material evolution and has been noted also in other scientific disciplines (e.g., physics, chemistry, biology). For example:

- The structure of a nucleus with surrounding material is found in atoms at the quantum level, and in planetary systems and galaxies at the macroscopic level. It is also found in biology in eukaryotic cells, and in animal societies organized around a leading group, with "distances" of individuals from the leader, or leaders, varying according to their social status.
- The notion of "force": in the quantum world, forces are interactions of fermions through the exchange of bosons (e.g., Ford, 2004). Chemically, forces are responsible for molecular structure. In biology, a force is exerted usually by a muscular structure. By analogy, a "force" can be of psychological or social nature.
- The notion of "wave": in the microworld there are waves of matter, or waves of probability; in the macroworld there are waves of sound, fluids, gravity, etc. More abstractly, there are "waves" of fashion, cultural ideas, economic crises, etc.

In a similar manner, we suggest that, through evolutionary mechanisms, cognition abstracted from what was initially employed as a simple association-building mechanism in creatures that appeared early on in evolutionary history to a conceptual categorization method, which finds its most versatile expression and application in human cognition.

References

Ford, Kenneth W. (2004). *The Quantum World*: Harvard University Press.

Foundalis, Harry E. (2006). "Phaeaco: A Cognitive Architecture Inspired by Bongard's Problems". Ph.D. dissertation, Computer Science and Cognitive Science, Indiana University, Bloomington, Indiana.

Hebb, Donald O. (1949). *The Organization of Behavior*. New York: Wiley.

Hofstadter, Douglas, R. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Kohonen, Teuvo (1982). "Self-organized formation of topologically correct feature maps". *Biological Cybernetics*, no. 43, pp. 59–69.

Martínez, Maricarmen (2006). "Implementation and performance results of the learning algorithm presented at the 2007 European Cognitive Science Conference, Delphi, Greece": http://pentagono.uniandes.edu.co/~mmartinez97/EuroCogSci07

McClelland, James L. (2006). "How Far Can You Go with Hebbian Learning, and When Does it Lead you Astray?". In Y. Munakata and M. H. Johnson (ed.), *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI*, pp. 33–69. Oxford: Oxford University Press.

Medin, D. L. and M. M. Schaffer (1978). "Context theory of classification learning". *Psychological Review*, no. 85, pp. 207-238.

Nosofsky, Robert M. (1992). "Exemplars, prototypes, and similarity rules". In A. Healy, S. Kosslyn and R. Shiffrin (ed.), *From Learning Theory to Connectionist Theory: Essays in Honor of W. K. Estes*, vol. 1 pp. 149-168. Hillsdale, NJ: Erlbaum.

Rosch, Eleanor (1975). "Cognitive representations of semantic categories". *Journal of Experimental Psychology: General*, no. 104, pp. 192-233.

Rosch, Eleanor and C. B. Mervis (1975). "Family resemblance: Studies in the internal structure of categories". *Cognitive Psychology*, no. 7, pp. 573-605.

Roy, Deb K. (2002). "Learning Words and Syntax for a Visual Description Task". *Computer Speech and Language*, vol. 16, no. 3.

Schyns, Philippe G. (1991). "A Modular Neural Network Model of Concept Acquisition". *Cognitive Science*, vol. 15, no. 4, pp. 461–508.