

An Attractor Neural Network Model of Semantic Fact Retrieval.

E. Ruppin

Department of Computer Science
School of Mathematical Sciences
Sackler Faculty of Exact Sciences
Tel Aviv University
69978, Tel Aviv, Israel

M. Usher

Department of Physics
School of Physics and Astronomy
Sackler Faculty of Exact Sciences
Tel Aviv University
69978, Tel Aviv, Israel.

October 23, 1996

Abstract

This paper presents an attractor neural network model of semantic fact retrieval, based on Collins and Quillian's original semantic network models. In the context of modeling a semantic network, a distinction is made between associations linking together objects belonging to hierarchically-related semantic classes, and associations linking together objects and their attributes. Using a distributed representation leads to some generalization properties that have computational advantage. Simulations performed demonstrate that it is feasible to get reasonable response performance regarding various semantic queries, and that the temporal pattern of retrieval times obtained in simulations is consistent with psychological experimental data. Therefore, it is shown that attractor neural networks can be successfully used to model higher level cognitive phenomena than standard content addressable pattern recognition.

1 Introduction

This paper presents an attractor neural network (ANN) model of semantic fact retrieval. The static structure of the model represents a Semantic Network (SN) and its dynamics a Spreading Activation process. Both these static and dynamic constituents are of paramount importance in artificial intelligence and cognitive sciences for the understanding and modeling of human memory retrieval processes. We shall therefore demonstrate that attractor neural networks can be successfully used to model higher level cognitive phenomena than standard content addressable pattern recognition.

Attractor neural networks are part of the more general Connectionist framework [1], that has been developed in recent years in parallel to the classic artificial intelligence mainstream of symbolic processing. While in the latter information is stored at specific memory addresses and processed by explicit rules, in the Connectionist framework no such distinction between the information and the processing algorithm exists [2].

The implementation of symbolic, graph-based structures such as semantic nets using connectionist architectures, is straight forward if a local representation is utilized; i.e., every node (representing a concept) in the SN is represented uniquely by a single processing unit (a neuron) in the NN. In accordance, every edge in the SN graph (representing an association) is to be represented by a connection (a synapse) in the NN, and the amount of activation of the various nodes is represented by using continuous valued neurons. Indeed, models constructed along these lines have already been presented [3, 4, 5, 6]. However, as noted by Ritter and Kohonen in regard with such one-to-one semantic modeling [7], "In view of the contemporary neurophysiological data, such a degree of specificity and spatial resolution is highly improbable in biology". Hinton [8] has claimed that an alternative approach, based on distributed representations is more promising. According to this approach, the semantic concepts are represented by groups (patterns) of neurons, such that each concept is distributed over many neurons and each neuron participates in representing many different concepts. Following Hinton's original conceptual framework, semantically related objects receive similar representations. A thorough discussion of the merits of distributed representation vs. a local one was presented by Hinton, McClelland and Rumelhart [1]. The specific advantages obtained by using distributed representations for our goals will be further described in the discussion.

In the context of modeling a semantic network, a distinction is made between associations linking together objects belonging to hierarchically-related semantic classes, and associations linking together objects and their attributes. This is made since we assume that hierarchically related objects receive closely related representations with respect to a certain distance measure, while no such encoding similarity is reasonable between objects and their corresponding attributes. An association of the first kind is 'zebra is-a mammal', while an association of the latter kind is 'zebra has

stripes'. A unique characteristic of a distributed representation is that if we assume that the encoding is a topology-conforming mapping, i.e., closely related concepts receive closely related representations, then associational linkage between such related concepts emerges implicitly from their "geometric" proximity in the encoding space. As opposed, the second kind of associations should be explicitly implemented into the synaptic connections. This distinction is realized by constructing the model from two interacting subnetworks, one storing the objects and the other the attributes. While the semantic proximity between objects is reflected in the degree of similarity of their corresponding encodings, the associations between objects and attributes are realized by projections between patterns belonging to different subnetworks.

Out of the various existing kinds of distributed connectionist models, we have selected the ANN paradigm. According to this paradigm, cognitive events are realized as sustained activity states of a neural net, which are attractors of the neural dynamics. These attractors are distributed activity patterns of the whole neural net, and learning is accomplished through modifying the synaptic connections' strength via a Hebbian-like mechanism. ANNs perform a content addressable memory retrieval and can be used for pattern recognition since they have the ability for performing error correction. For example, an ANN that models retrieval of information from short term memory, demonstrated in the high speed scanning experiments of Sternberg's type [9]. The advantage of using ANNs for such modeling results from the fact that the error correction is amplified at every processing phase which is based on convergence towards an attractor. In addition, biological data supporting the hypothesis that memory activation is achieved through sustained neural activity, has been found [10].

Our goal is to account for semantic fact retrieval using the ANN paradigm. The model is conceived so that response performance and retrieval times consistent with Collins and Quillian's experimental data, can be obtained. As a consequence, several constraints upon the model construction are imposed. The recognition and retrieval of facts (a relation composed of several concepts) would be straight forward if they would be stored as attractors in the network. Nevertheless, when modeling fact retrieval, in light of economy and language-productivity considerations [11], one cannot store every fact as an attractor on its own, and therefore another strategy should be pursued; facts will be represented as spatial-temporal combinations of their constituents, which themselves will continue to be represented as attractors. Subsequently, the retrieval of facts is described by a dynamical process performed on this attractor space.

Collins and Quillian [12] have originally modeled two fundamental types of queries:

1. An Object-Attribute (OA) query; i.e., asking wherever an item (object) A has a property (attribute) X, (e.g., does a canary sing?).
2. An Object-Object (OO) query; i.e., asking wherever an item A belongs to the category (superset) represented by item B, (e.g., is the canary a bird?).

In addition to the modeling of these queries, which are characterized by performing an upwards motion in the semantic tree along the path leading to the root, we have been able to demonstrate the modeling of other plausible types of queries, (been handled previously in Fahlman's NETL system, [13]):

3. The Property intersection query; i.e., what item has the properties X,Y, and Z, (e.g., who sings, flies, and has feathers?).
4. The Reverse Category query; i.e., what type (sub-category) of item A has property X, (e.g., what kind of bird sings?).

The network's dynamics should realize a process in which semantic queries are responded. The queries being modeled are composed of either an item (object) and an attribute, two objects, or two attributes. It is assumed that upon a presentation of a query, a preprocessing state exists, by which the two components of the presented query are distinguished and inputed into the corresponding subnetwork (items or attributes), by means of external fields. It has already been shown that ANNs can maintain their retrieval capabilities even if the input is applied as an external field [14]. It is assumed that a necessary condition for obtaining a positive response is that a stable state with high similarity to the second input component is achieved. If such a degree of similarity is not achieved during a certain fixed time, or the network has converged into a stable state with low amount of similarity with the second input, then a negative response occurs.

This paper is further divided into four sections; In section 2. we briefly review the Semantic Network and Spreading Activations concepts. In section 3. we outline the architecture of the model presented and describe its dynamical behavior. In section 4. the results of computer simulations of the model are described and in section 5. some computational characteristics of the model are analyzed and the model's performance in comparison with the existing psychological experimental data is discussed.

2 Preliminaries.

In this section we briefly discuss the static and dynamic features upon which the model is based. The model's static structure is a semantic network, originated by Quillian [15, 16, 17] as a mode of knowledge representation. Its primary motivation was the Economy principle stating that items are stored without redundancy, as is described further below. Information is represented by a set of nodes, representing concepts, connected to each other by a set of edges (pointers), which represent relationships among the concepts represented. Different models of semantic nets have later been used to represent a variety of kinds of knowledge [18, 19, 20].

The semantic network described by Collins and Quillian [12] is a directed tree linked structure (see fig. 1) where the nodes stand for concepts designating classes

of objects and the edges between them represent a special kind of relationship (designated an AKO ('a kind of') relationship in AI literature), denoting a category (or superset) relation; For example, such a directed edge pointing from node A to B designates that the concept represented by A (e.g., a dog) belongs to the category represented by concept B (e.g., the mammals). In addition, every node has another type of directed edges pointing to some properties represented by a set of nodes external to the tree structure. The existence of an edge between an object node to a property node denotes that the object has that property, (e.g., the dog barks).

denote object (item) nodes and

A chief characteristic of such a semantic net is that if there is a property X which is common to some of the concepts (items) defined in the tree, it can be stored at the highest possible concept, for which all concepts in the sub-tree spanned by it, still have this property X in common. Thus, a large amount of memory space can be saved, once a dynamic process is implemented characterized by motion along the path leading from a node through its ancestors. In the model presented, such a dynamic process has been implemented using spreading activation.

The theory of Spreading Activation [12, 21] has a major role in modeling the various phenomena gathered in psychological experiments, concerning the dynamics of semantic processing. It is based on the hypothesis that the cognitive activity evoked when a memory is queried about one of its stored facts could be characterized by a 'spreading out' process. In this respect, by 'facts' we mean statements about stored items or the relations between them, (e.g., 'the dog barks'). In this process, it is hypothesized that the activation originates at the nodes representing the concepts queried about ('dog, barks', denoted the 'initiating nodes') and further spreads in parallel to all the nodes that are connected to one of the initiating nodes. If several initiating nodes are connected to another node, then the latter could be considered as an 'intersection' node, receiving a high level of activation (above a certain threshold) from all initiating nodes.

Since our model is based on the ANN paradigm, the spreading of activation occurs in the concepts' and not in the neurons' space. The activity of a concept is defined as the overlap between the pattern representing it and the current network's state. Therefore, although in an ANN only one attractor can be fully activated at once, nevertheless, a restricted notion of spreading of activation can be manifested; when moving from one attractor to another, an intermediate stage consisting of mixtures of several (low activated) patterns, arises. This intermediate 'spreading' (divergence) of activity finally converges into an attractor. It will be further shown, that in some cases this convergence may be viewed as expressing the intersection phenomenon. In addition, since our model is based on two separate ANN subnetworks, objects and attributes may be fully activated at the same moment.

3 Description of the Model.

An ANN is an assembly of formal neurons connected by synapses. The neuron's state is a binary variable S , taking the values ± 1 denoting firing or resting states, correspondingly. The network's state is a vector specifying the binary values of all neurons at a given moment. Each neuron receives inputs from all other neurons to which it is connected, and fires only if the sum of the inputs is above its threshold. This process may include a stochastic component (noise) which is analogous to temperature T in statistical mechanics. When a neuron fires, its output, weighted by the synaptic strength, is communicated to other neurons and as a consequence, the network's state evolves. Mathematically this is described as follows; in the deterministic case

$$S_i(t+1) = \text{sgn}(h_i(t)) \quad (1)$$

and for the stochastic case, with temperature T by

$$S_i(t+1) = \begin{cases} 1, & \text{with prob. } \frac{1}{2}(1 + \tanh(\frac{h_i}{T})) \\ -1, & \text{with prob. } \frac{1}{2}(1 - \tanh(\frac{h_i}{T})) \end{cases} \quad (2)$$

where S_i is the state of neuron i and h_i is the local field (the postsynaptic potential) of neuron i , which is given by:

$$h_i = \sum_{j \neq i}^N J_{ij} S_j \quad (3)$$

where J_{ij} is the ij element of the connection matrix.

The network's trajectories, are therefore determined by the matrix describing the synaptic weights, which reflect prior learning. For specific synaptic weight matrices, the emergent motion is characterized by convergence to stable pattern configurations (representing the memorized concepts) which are attractors of the dynamics [22]. Various versions of ANNs, whose attractors form an hierarchic structure, have been developed [23, 24]: the hierarchical structure of the stored memory patterns reflects their proximity in the pattern space according to some distance measure.

The model we propose for fact retrieval from a semantic memory is composed of two ANN subnetworks of N neurons, (in our simulations we have chosen $N = 1000$).; the first subnetwork, denoted the items subnetwork, stores a semantic tree of items, illustrated as circles in fig. 1. The second subnetwork, denoted the attributes network, stores the items' properties, depicted as squares in fig. 1.

3.1 The items subnetwork

The stored items are organized in a three level tree. The top level consists of two patterns representing the most general items in our semantic set, denoted by ξ^μ , where $\mu = 1, 2$. Each of these two patterns is generated by selecting the value of

every neuron in the network to be 1 with probability q_1 and -1 with probability $1 - q_1$ independently. The average activity $\langle \xi^\mu \rangle$ for these patterns is therefore $a_1 = 2q_1 - 1$. From each of these patterns, two descendants, denoted by $\xi^{\mu\nu}$, ($\mu, \nu = 1, 2$) are generated by probabilistically flipping some $(1 - q_2)$ bits (neurons) of the ancestor pattern $\xi^m u$. Formally stated, this is performed in the following way;

$$\xi^{\mu\nu}_i = \xi^\mu_i \cdot \chi^\nu_i \quad (4)$$

where χ^ν is subject to the probability distribution

$$P(\chi^\nu_i) = q_2 \delta(\chi^\nu_i - 1) + (1 - q_2) \delta(\chi^\nu_i + 1) \quad (5)$$

where $\delta(x)$ is the Dirac delta-function which is zero for $x \neq 0$ and its integral equals one.

q_2 has to be greater than 0.5 so that a positive correlation will exist between the ancestor and its descendant patterns. In the same fashion, a third level of patterns, denoted $\xi^{\mu\nu\rho}$ ($\mu, \nu, \rho = 1, 2$), is generated by probabilistically flipping some of the ancestor's $\xi^{\mu\nu}$ bits, with a similarity parameter q_3 also greater than 0.5;

$$\xi^{\mu\nu\rho}_i = \xi^{\mu\nu}_i \cdot \chi^\rho_i \quad (6)$$

where χ^ρ is subject to the probability distribution

$$P(\chi^\rho_i) = q_3 \delta(\chi^\rho_i - 1) + (1 - q_3) \delta(\chi^\rho_i + 1) \quad (7)$$

Subsequently, each pattern belonging to the first level is in correlation $a_2 = 2q_2 - 1$ with its direct descendants, while the same holds for next generation with correlation $a_3 = 2q_3 - 1$. It is straightforward that the correlation between a pattern in the first level and its descendants in the third layer is $a_2 a_3$ and the correlation between two patterns belonging to the same ancestors is a_2^2 in the second layer and a_3^2 in the third one. Since we have chosen $q_2 = q_3 = 0.75$, we obtain an equal correlation of 0.5 between each ancestor and descendant. The correlation between two patterns at the same level, with a common direct ancestor, is 0.25. Observe that in this structure items closely related obtain similar encoding (where the degree of similarity is measured by the Hamming distance) which reflects the geometric proximity in the 2^N space.

The connectivity matrix which embeds this hierarchic structure of attractors, is constructed according to Feigelman and Ioffe [23];

$$J_{ij} = \sum_{\mu=1,2} \xi^\mu_i (\xi^\mu_j - a_1) + \frac{1}{1 - a_2^2} \sum_{\mu,\nu=1,2} (\xi^{\mu\nu}_i - a_2 \xi^\mu_i) (\xi^{\mu\nu}_j - a_2 \xi^\mu_j) \quad (8)$$

$$+ \frac{1}{1 - a_3^2} \sum_{\mu,\nu,\rho=1,2} (\xi^{\mu\nu\rho}_i - a_3 \xi^{\mu\nu}_i) (\xi^{\mu\nu\rho}_j - a_3 \xi^{\mu\nu}_j)$$

Using this connection matrix it can be easily shown that each of the stored patterns is a stable state since

$$h_i = \sum_{j=1}^N J_{ij} S_j = S_j \quad (9)$$

when S is one of the patterns stored $(\xi^\mu, \xi^{\mu\nu}, \xi^{\mu\nu\rho})$, whose total number is $2+4+8 = 14$.

3.2 The attributes subnetwork

This subnetwork stores the properties (attributes) belonging to the items, as attractors. The patterns representing the attributes denoted by η^μ are uncorrelated, i.e., they are generated by selecting randomly the values of each neuron to be ± 1 with equal probability. This network is therefore a Hopfield type network [22] and its connection matrix is: $\sum_{\mu=1}^p \eta^\mu_i \eta^\mu_j$. Since for every item stored in the items network we arbitrarily assign three distinct attributes, the total number of stored attributes p is $14 \cdot 3 = 42$.

3.3 The subnetworks interaction

The two subnetworks are weakly connected through two kinds of synapses. The first kind of synapses, J^{AI} , are projecting the activation from items subnetwork to the attributes one, while synapses of the second kind, J^{IA} , project the activation in the opposite direction; every item pattern projects through synapses of the first kind to three attribute patterns and each of those attribute patterns project their activity backwards to it. The first interconnectivity matrix is as follows:

$$J^{AI}_{ij} = \lambda_1 \sum_{\alpha} (\eta^{\alpha 1}_i + \eta^{\alpha 2}_i + \eta^{\alpha 3}_i) \xi^{\alpha}_j \quad (10)$$

In this formula, the neuron i belongs to the attributes network, and the neuron j belongs to the items network. ξ^α denotes an attractor in the items subnetwork so that the sum runs over all items patterns and connects each of them to its three attribute patterns in the attributes subnetwork. λ_1 denotes the strength of the projection from items to attributes patterns.

The second interconnectivity matrix is:

$$J^{IA}_{ij} = \lambda_2 \sum_{\alpha} \xi^{\alpha}_i (\eta^{\alpha 1}_j + \eta^{\alpha 2}_j + \eta^{\alpha 3}_j) \quad (11)$$

where i belongs to the item network, j to the attributes network and λ_2 is the strength of the projection from attributes to items.

3.4 Characteristics of the unified network

The behavior of the network as a whole depends on the values of the couplings λ_1 and λ_2 . In the extreme case when $\lambda_{1,2} \ll 1$, every combination of attractor pairs (of the two subnetworks) is an equally stable attractor. Therefore, the unified network can practically be regarded as two separate modules. On the other extreme, if $\lambda_{1,2} = 1$ and every item is coupled to a single attribute then the unified network reduces to a simple ANN in which every attractor is composed of the correspondingly coupled subnetworks attractors as substrings. For in between values of $\lambda_{1,2}$, an intermediate behavior is obtained; every combination of subnetwork attractors is an attractor of the unified network, but not with equal stability. The most stable attractors of the unified network (with the largest basin of attraction) are the ones composed of coupled subnetwork attractors. This is illustrated by the following examples; if two mutually coupled patterns are queried about, e.g., $S^A = \eta^{\alpha 2}$ and $S^I = \xi^\alpha$, then the local field in the items subnetwork is

$$h_i = \xi^\alpha(1 + \lambda_2) \quad (12)$$

while if the queried patterns are not mutually coupled, e.g., $S^A = \eta^{\beta 1}$ and $S^I = \xi^\alpha$, (where $\eta^{\beta 1}$ is an attribute of object ξ^β), we get

$$h_i = \xi^\alpha + \lambda_2 \xi^\beta \quad (13)$$

This means that in the first case the attractor is strongly stable while in the latter, it is less stable since for neurons at which $\xi^\alpha \neq \xi^\beta$ the local field is $\xi^\alpha(1 - \lambda_2)$. Therefore, for such intermediate values of $\lambda_{1,2}$, the unified network is characterized by weak modularity.

3.5 Input – output strategy

The need to account for the various experimental results of Collins and Quillian, has imposed several methodological constraints guiding the input–output strategy used in the model. In order to account for the fact that the response times to the various queries are graded, a uniform motion upwards the tree structure had to be implemented. This imposes a first constraint on the input to the objects subnetwork: the external field representing this input must decay fast enough so that the upward motion is not disturbed. Second, a criterion for discriminating between true and false queried facts is needed. Using only the network’s convergence into a stable state is not sufficient, since even for false facts such convergence could occur, therefore leading to a false response (a false fact positively answered). Thus, the identity of the stable state arrived should be examined. In the case of an Object-attribute query, this examination cannot be performed upon the Objects subnetwork, since in cases of property inheritance, (e.g., does the canary fly?) the identity of the ancestor pattern

to which the queried attribute is linked to, is not known in advance. Therefore, this examination should be performed upon the attributes subnetwork. This, in turn, leads to the conclusion that the input to the attributes subnetwork cannot decay too fast. This input has to last for sufficient time such that the upwards motion upon the objects tree structure reaches the object to which the queried attribute is linked. Otherwise, when this object is reached, no memory of the queried attribute is left.

In conclusion, the input-output strategy used is not symmetric: An input is realized by a local external field applied to the appropriate subnetwork. In particular, the first input is applied as an external field which is initially strong and then decays, (this is equivalent to initializing the network in the corresponding input state), while the second input is applied as a constant external field. Naturally, the subnetwork upon which the response is examined depends upon the type of query; In the case of an Object-attribute query, it is assumed that a necessary condition for obtaining a response is that the attribute subnetwork remains in a stable state for a sufficient duration; a response is considered positive if during this period a high degree of similarity between the stable state and the second input component is achieved. If such a degree of similarity is not achieved during a certain fixed time, or the network has converged into a stable state with low amount of similarity with the second input, then a negative response occurs. Similarly, In the case of an Object-object query, the second component of the query is applied as a non-decaying external field to the Objects subnetwork where convergence is examined. The initial state of both subnetworks is a “neutral” attractor, with no semantic assignment.

3.6 Expected network behavior

Four types of queries are defined and the following scenario are expected accordingly:

1. **Object-attribute query:** when a query is presented, the items network is initialized by a strong external field with the attractor representing the queried object, (ultimately, this field will decay as explained above). The equivalent external field applied to the attributes network is weaker but of constant value. By itself it is not sufficient to drive the attribute subnetwork out of its neutral attractor, nevertheless, due to the projection from the items to the attributes subnetwork, two scenario can occur:

- (a) In the case of a positive query (a query about a truly stored relationship (denoted ξ^γ, η^q) such as ‘does a canary fly ?’), the local field in attribute subnetwork will initially be

$$h_i = \eta^0_i + \lambda_1(\eta^{\gamma^1}_i + \eta^{\gamma^2}_i + \eta^{\gamma^3}_i) + \lambda_3\eta^q_i + R_i \quad (14)$$

where η^0 represents the neutral attractor, $\eta^{\gamma^{1,2,3}}$ represent attributes assigned to the queried item ξ^γ , η^q stands for the pattern representing the

queried attribute, and λ_3 denotes the strength of the external field caused by the queried attribute. R_i is a “cold” noise field originated by the projection from other attractors in the items network which overlap ξ^γ . Most of these projections will be negligible, except of the term originated by the direct ancestor, since a pattern’s strength of projection is proportional to its overlap with the actual state of the network. Two cases are further distinguished:

- i. If the queried attribute is directly associated with the queried item (e.g., ‘does the canary sing?’) then η^q is one of the three attribute patterns $\eta^{\gamma^{1,2,3}}$, for example η^{γ^1} . Therefore, for certain values of the parameters λ , since the component of η^{γ^1} in the field h_i is $\lambda^1 + \lambda^3$, the subnetwork’s state will be attracted into the pattern representing the queried attribute η^{γ^1} . Consequently, the reverse projection towards the items subnetwork will be directed toward the queried item enhancing its stability. Therefore, the unified network will remain in a stable state (composed of the two queried inputs) which is interpreted as a positive response.
- ii. If the queried attribute is directly associated with an ancestor of the queried item, (e.g., ‘does the canary have feathers?’), which is answered positively in light of the semantic inheritance), the external field caused by the queried attribute is directed towards an attribute of the queried item’s ancestor. Therefore the external field in the attribute subnetwork will be as in equation 14, where η^q is now a pattern representing an attribute belonging to the queried item’s ancestor. At first, since all the above terms are distinct (pseudo orthogonal) attractors, the attribute subnetwork is expected to arrive at a mixed stable state. We have chosen $\lambda_3 > \lambda_1$ and therefore the major component of the mixed state is η^q , but the other components $\eta^{\gamma^{1,2,3}}$ and the attribute attractor associated with the queried item’s direct ancestor (present in R_i) exist. Consequently, the reverse projection is diffuse, but its major component is directed towards the queried item’s ancestor because of the large relative weight of the η^q component. Hence, the items subnetwork state will be destabilized and a transition toward the ancestor corresponding to η^q will occur. In turn, the projection towards the attributes subnetwork is directed towards the ancestor’s attributes containing η^q , and therefore the unified network will be stabilized in a state composed by the queried attribute and its corresponding item. Thus a positive response is attained, however, the response time is larger due to the transition process. When the attribute queried belongs to a second level ancestor (a “grandfather”)

of the queried item, the transition passes (at least partially) through the direct ancestor and the response time will be even longer. This happens due to the direct ancestor's component in the noise term R_i and to the geometric proximity between the queried item and its direct ancestor (i.e., from the geometric aspect, the direct ancestor lies in between the item queried and the item corresponding to the queried attribute). Therefore we expect to obtain a cascade of response times, increasing with the semantic distance between the queried item and the item corresponding to the queried attribute.

- (b) In the case of a negative query, (a query about a non-stored relationship (denoted ξ^γ, η^r) such as 'does a canary bark?'), the external field caused by the queried attribute is not directed towards one of the attractors projected by the queried item. The external field is as in equation 14 and since all the terms are distinct (pseudo orthogonal) attractors the attribute subnetwork is expected to arrive at a mixed stable state. The major component of the mixed state is the η^r , but the other components $\eta^{\gamma_{1,2,3}}$ and the attribute attractor associated with the queried item's direct ancestor (present in R_i) also exist. Consequently, the reverse projection is diffuse and the state of the items subnetwork either remains in the initial state corresponding to the queried item or a transition towards the state corresponding to the queried item's direct ancestor occurs. The latter occurs when the queried attribute belongs to a direct sibling of the queried item, since the local field in the items subnetwork contains the major components pointing to ξ^{111}, ξ^{112} and ξ^{11} , where for example, ξ^{111} denotes the queried item, ξ^{112} the item pointed by the queried attribute, and ξ^{11} their direct ancestor. In light of the geometric structure, (the ancestor is in between its descendants) and using the majority rule we obtain a local field similar (with the same sign) with the ancestor pattern for most ($N(1 - q_3^2)$) of its neurons. This can be viewed as an instance of the spreading activation notion of intersection. However, also the above transition will not attract the attribute subnetwork into the state corresponding to the queried attribute, since using equation 14, all the terms continue to be orthogonal. Since the component of the queried attribute in the mixed state is not sufficiently large (we have chosen a threshold of 0.9), a negative response will be generated after a fixed amount of time. However, since the coefficients of these major components are unequal (they depend on the attribute subnetwork state) the situation is more complex; for the case when the coefficient of ξ^{112} is large (this occurs when λ_3 is large), the items subnetwork may be attracted into the ξ^{112} pattern, leading the attributes subnetwork into the attractor η^r representing the queried attribute, and an error is generated.

2. **Object-object query:** the behavior of the network for the case of an object-object query evolves in similar lines to the previous description, therefore we present only a brief overview of its main aspects; both inputs are introduced into the items subnetwork, as explained above, while the first input is initializing the items subnetwork, the second input is introduced as an external field of strength λ_4 . In the case of a positive direct query (e.g., ‘is a canary a bird ?’) the external field will enhance the stability of the network’s initial state. For the case of positive query in general, a transition towards the ancestor occurs due to the external field. When the second input is a two levels higher than the first, the transition passes through the first item’s direct ancestor. Therefore, a cascade of retrieval times is expected in accordance to the semantic distance between the inputs. In these cases a positive response is obtained since the stable states reached have a high degree of similarity (overlap) with the second input.

In the case of negative queries (e.g., ‘is a canary a mammal ?’), if the second input is a sibling of the first input, the network performs a transition towards the state corresponding to the common ancestor, (an intersection). This state is interpreted as a negative response since its overlap (similarity) with the second input is below the threshold previously defined. If the second input is semantically far apart (and not an ancestor of) the first input, a mixed state is achieved and again a negative response is produced. As in the Object-attribute case, an error may be produced since for large values of λ_4 the items subnetwork may be attracted to the pattern representing the second input.

3. **Reverse category query:** (what kind of bird sings ?) This query requires an explicit answer (which is not a yes/no response). The response will be achieved upon the convergence of the items subnetwork to a stable state. As in the previous case, the first input initializes the items subnetwork while the second one is introduced as an external field to the attribute network. However, in order to obtain a correct response, note that a downwards traverse should proceed on the tree structure.

4. **Reverse property query:** (who sings, flies and has feathers ?) As in the previous case, this query also requires an explicit answer. The query gains significance when there are attributes such that each belongs to several items, (otherwise the item can already be retrieved by presenting a single attribute). The query is modeled by presenting the queried attributes as external fields applied to the attributes subnetwork and the items subnetwork is initialized by a random pattern. As in fig. 4., when the queried attributes are $\eta^{\delta 1}$ and $\eta^{\delta 2}$, the attributes subnetwork reaches a mixed state whose projection to the items subnetwork leads to convergence into ξ^{111} (since ξ^{111} receives projection from

both attributes while ξ^{112} from only one of them). In this way an intersection process is used in order to retrieve the requested information.

3.7 An auxiliary bias mechanism.

A possible mechanism by which performance can be improved and the number of erroneous responses reduced (see section 3), is a dynamic bias field, introduced into the items subnetwork. It was shown [24, 25] that bias can enhance the performance of an ANN. This is done by introducing the additional field:

$$\Delta h_i = -\lambda_b \left(\frac{1}{N} \sum_j S_j - a \right) \quad (15)$$

where a denotes the average activity of a group of learned patterns. This additional field tends to weakly constraint the activity of the network's state around a . If the patterns in the tree structure are generated so that every level l is characterized by a distinct average activity a_l , then by modifying a_l in equation 15 we can enhance the stability of the pattern in the respective level. We have chosen the value of the parameter q_1 to be 0.25 and therefore the activity of the different levels is $a_1 = -0.5$, $a_2 = -0.25$ and $a_3 = -0.125$. In equation 15, the first term results from a constant decrement of the synaptic matrix values while the second is a constant field which can be interpreted as a uniform threshold. Thus, by controlling this external field the network's state motion can be guided upward or downward the tree structure. Such an external control mechanism can be psychologically motivated as involving an 'attention' process required in accordance with the specific query; for some kinds of queries (e.g. queries 1,2) the processing is characterized by upward motion while for others (e.g. query 3) a downward motion is required.

4 Computer Simulation of the Model

4.1 Technical details

Each of the two simulated sub-networks is composed of a 1000 neurons. The sub-networks were constructed according to the prescriptions presented in the model's description. Their attractor states were randomly generated before every new simulation run to ensure the independence of the results obtained of a specific set of attractors. The dynamics were characterized by an asynchronous, stochastic process; within one update cycle, the units are updated sequentially in random order. After all units have been updated the next update cycle begins.

We have run the simulations with and without the bias mechanism. An improved performance was found in the latter case, in the sense that a more uniform motion on the tree structure was achieved. In addition, the No. of retrieval errors was

significantly decreased, as will be described further on. The bias was implemented by modifying the parameter a of equation 15, so that every 10 iterations it equals the mean activity of the patterns in the following level, e.g., for an upwards motion a is:

$$\begin{cases} t < 10 & a = -0.125 \\ 10 \leq t < 20 & a = -0.25 \\ t \geq 20 & a = -0.5 \end{cases} \quad (16)$$

The values of the various parameters used are as follows;

- λ_1 , the strength of the projection from items to attributes was chosen equal to 0.3.
- λ_2 , the strength of the projection from attributes to items was chosen to be 0.55. λ_2 was chosen greater than λ_1 since the latter constitutes an important component of the force driving the activity in the items sub-network along the path leading from a queried item through its ancestors to the root of the semantic tree.
- The parameters denoting the strength of the external fields representing queried attributes and items λ_3 and λ_4 in correspondence are $\lambda_3 = 0.35$ and $\lambda_4 = 0.2$. λ_4 is relatively smaller since it acts directly on the items subnetwork. The bias parameter was set to $\lambda_b = 5$.
- The Temperature is $T = 0.15$. Since the dynamics of the network include stochastic noise, the networks' state is characterized by fluctuations even when a stable state is achieved. Therefore, a concept was considered to be activated if the the network's state S overlap with the pattern representing this concept is larger than a threshold of 0.9 for at least 5 consecutive iterations.

4.2 Illustration of a typical query run.

The dynamics of the unified network in response to the presentation of a query is exemplified graphically by the following figures, where item patterns are represented by circles connected by solid lines (corresponding to semantic relations) while the attribute patterns are represented by squares each of them connected by a dashed line to the item to which it belongs. The size of the various items and attributes presented is proportional to the amount of overlap between the current sub-network's state and the patterns representing them.

Fig. 2 describes the case when an item is queried about an attribute belonging to its 'grandfather' in the semantic tree; The initial state is depicted in fig 2a. where in the items subnetwork the pattern representing the queried item is dominant and the activity of the attributes subnetwork is dispersed among many patterns. Fig. 2b. (after 15 iterations) illustrates a transition of the items subnetwork to a state with

high degree of overlap with the queried item’s direct ancestor. In correspondence the activity of attributes belonging to it is increased. Fig. 2c. describes the situation (after 22 iterations) where a transition to the pattern representing the queried item’s ‘grandfather’ has taken place and shortly afterwards (after 24 iterations) the attribute networks is seen to converge into a stable state of high overlap with the queried attribute. Consequently, a correct positive response is generated.

Fig. 3 describes the case when an item is queried about an attribute belonging to its direct ancestor; The initial state (fig 3a., after 3 iterations) is as in the previous query. Fig 3b. (after 10 iterations) illustrates a transition of the items subnetwork to a state with high degree of overlap with the queried item’s direct ancestor. In correspondence the activity of attributes belonging to it is increased. This trend is strengthened as can be seen in fig. 3c., (after 13 iterations) where the network has converged and remains at a state with high overlaps with the patterns representing the queried components.

attribute belonging to

The dynamics of the network in response to a negative query are presented in fig. 4; observing the initial state (fig 4a., 3 iterations) one can see higher activity in the patterns corresponding with the query’s components. Since both inputs drive the items subnetwork to the queried item’s ancestor (fig 4b., after 14 iterations) it converges to its corresponding pattern, however the attributes network remains in a mixed state and the query is answered with a negative response (fig 4c., 22).

attribute which neither

The dynamics of the network’s response towards Object-Object queries follows similar lines; for example, when an input is introduced pointing to a second level ancestor, a two stage transition (passing through the direct ancestor) will occur.

When a Reverse Category query is presented, a downwards motion should be performed and in order to implement such motion the direction of the bias modification was inverted. Best performance was achieved when in addition the parameters λ_4 and λ_b were now set to the values of $\lambda_4 = 0.4$ and $\lambda_b = 7$.

Fig 5. illustrates the case of a Reverse Property query. It can be seen that the activity of the item corresponding to the intersection of all three inputted attributes presented grows.

‘intersect’ and their corresponding

4.3 Performance

In order to examine the performance of the network the simulations were run 30 times for each query. For every run, the patterns representing the items and attributes, are randomly generated. The results of the overall performance without introducing bias are presented in table 1, where for example, a zero-level object-attribute (OA) query denotes that the queried attribute directly belongs to the queried item, a one-level

object-object (OO) query denotes that the first component of the query is a direct descendant of the second one, etc. Best performance rate was achieved with parameter values of $\lambda_1 = 0.35$, $\lambda_2 = 0.75$, $\lambda_3 = 0.5$, $\lambda_4 = 0.4$ and $T = 0.15$.

The query	% of CR	MT
Positive zero-level OA query	100	2.7
Positive one-level OA query	86.5	9.4
Positive two-level OA query	63.5	15.2
Negative OA query	80	
Positive zero-level OO query	100	1
Positive one-level OO query	83	4.5
Positive two-level OO query	47	6.5
Negative OO query	30	

Table 1: Performance and response times for the various queries. CR denotes correct response and MT the mean-time of correct response counted by the No. of iterations.

It can be seen that for positive object-attribute queries a gradual linear cascade of retrieval times exists; i.e., the larger the semantic distance between the queries components, the longer it takes for a stable state to be reached. For object-attribute negative queries, no correlation was found between the semantic distance between the query’s components and the time a stable state was reached in the attribute subnetwork. Several possibilities for determining the response time of negative queries exist and will be discussed in the next section.

For object-object queries the network’s performance is considerably deteriorated as can be seen in table 1; depending on the value of λ_4 , either we obtain good performance for positive queries and poor performance for the negative ones, or vice versa. It seems that the difference originates from the non-symmetric modeling of the two queries.

The addition of an auxiliary bias mechanism improves the performance of the network with regard to its response accuracy while retaining the cascade of retrieval times described above. The results of the statistics performed on such runs are presented in table 2.

We have found that the network implicitly exhibits the Priming phenomenon; a fundamental characteristic inherent to the spreading activation theory [26]; Priming designates the phenomena by which cognitively active concepts cause the memory retrieval of related concepts to be faster. According to SA theory this is explained by the nodes representing the latter concepts receiving a certain amount of baseline activation from the former node representing the cognitively active concept and thus rendering it more prone for further sub-threshold activation. In the model presented,

The query	% of CR	MT
Positive zero-level OA query	100	3.1
Positive one-level OA query	100	9.1
Positive two-level OA query	100	15.3
Negative OA query	96.5	
Positive zero-level OO query	100	1
Positive one-level OO query	100	15.3
Positive two-level OO query	96.5	22.7
Negative OO query	100	

Table 2: Performance and response times for the various queries — Auxiliary bias mechanism is incorporated.

CR denotes correct response and MT the mean-time of correct response counted by the No. of iterations.

when an input is presented to the network its retrieval time depends on the residual state of the network. When the network was initialized with the pattern ξ^a and subsequently another pattern ξ^b is presented (as an external field, with $\lambda_4 = 0.6$), the convergence time of the network towards the latter pattern is proportional to the semantic distance between them. For example, for the patterns ξ^{111} and ξ^{112} we observe a convergence time of 9 iterations and for the case that the second pattern is ξ^{222} a convergence time of 13 iterations is observed.

5 Discussion

In this section we discuss several computational aspects of the ANN model presented and examine the fit between simulation results and the psychological data regarding fact retrieval.

Several advantages originate from using a distributed representation and ANN dynamics, in comparison with classic symbolic or local connectionist models:

1. Models based on distributed representations function well as content addressable memories since they are capable of performing error correction and are fault tolerant.
2. As have been described, the network ‘ascends’ the concept hierarchy in a distributed way, by having its state become more like the pattern for a higher level concept. In this way, it does not necessarily go through discrete steps corresponding to particular higher concepts, (as a symbolic program would do), but sort of ‘mushes’ its way up or down the hierarchy. This process seems to us less rigid and more ‘brain-like’. A further speculation is to view stored patterns with

low amount of overlap with the network’s current activity state as representing ‘sub-conscious’ concepts awaiting for their turn to become ‘conscious’, i.e., with high overlap with the network’s activity state.

3. Distributed systems automatically give rise to generalizations; i.e., when presented with a non-familiar pattern the network can still relate to it according to its relative similarity with the various familiar patterns. In our model, when a new semantic concept is presented, the pattern representing it is stored in the items subnetwork and its correct placing in the semantic tree, regarding its inheritance path and semantic relatedness to other concepts, is determined implicitly by the representation encoding without any need of explicit instructions. In general, the semantic structure follows directly from the assumption that the encoding has such ‘semantic geometry’ conservation. There is no need to specifically denote and learn pointers between items of the semantic tree.
4. The network has the ability to learn new concepts that were not anticipated at the time the network was constructed; while in classic symbolic models, or connectionist models entailing local representations, one has to add new nodes to the network every time a new concept is learned, in the model presented simply a new pattern is memorized. Our model has the characteristic that when a new concept is added, it becomes more and more significant as its definition gets more precise; i.e., if a new concept is added but it is not known yet to contain any properties then no attribute patterns project to this concept’s pattern and thus its basin of attraction would be very small. Only when a pattern in the item’s subnetwork receives considerable projection from the attributes subnetwork, it gains a sufficiently deep and wide basin of attraction, so that its corresponding item becomes significant.
5. The various attributes are implicitly dynamically reallocated to the items to which they belong. This is however a virtual process with no pointers been actually modified; observe for example a part of the semantic tree containing an item ξ^{11} and his two sons ξ^{111} and ξ^{112} . When the attribute η^l is first introduced, for example when the fact ‘item ξ^{111} has the attribute η^l ’ is first stored, then its projection is directed only to this item ξ^{111} . However, if the fact stating that ‘ η^l belongs to ξ^{112} ’ is subsequently stored, then the projection from η^l to the items subnetwork is now $\lambda_2 \eta^l (\xi^{111} + \xi^{112})$. Hence, in light of the geometric construction, if the attributes subnetwork will be at the state η^l it will contribute a local field aligned with the ancestor ξ^{11} pattern in $\approx N(1 - q_3^2)$ of its neurons, and therefore can now practically be considered as belonging to this ancestor. Such a scenario is illustrated in fig. 6. where after introducing a property belonging to two sibling items the items subnetwork converges into the pattern representing their direct ancestor, though no such fact was explicitly

stored by the network. By the same mechanism, if subsequently some of these stored facts will be deleted then the situation will be reversed accordingly. This generalization property is expected to be enhanced in a network where a pattern can have more than just two direct descendants.

6. By using a specific form of encoding representing different amounts of activity at different layers of the semantic tree, we were able to use the bias mechanism in order to model a general tendency of motion upon this tree. We do not have to perform a costly search process upon all the trajectories defined in the tree but move in a general direction till the network converges to the correct attractor. Thus, by modeling such an external ‘attention’ mechanism we were able to improve much the performance of the network. It may be that modifying the external bias according to some oscillatory function can create such continuous upwards and downwards motion until correct convergence is achieved.

The ANN paradigm is used to model the concept of ‘intersection’ underlining spreading activation theory; As shown, intersection is achieved when the integrated action of projections from some patterns leads to convergence to a stable state, not achievable by the activation of only one of the patterns solely. Thus, intersection is modeled without any need for explicitly implementing a labelling procedure as in the original spreading activation models.

The model presented is constrained by the assumption that hierarchical associations are implicitly implemented by the geometric similarity solely. Under this constraint, several additional strategies for modeling fact retrieval phenomena have been tried, unsuccessfully; we have found that introducing a fatigue component into the dynamics of the items subnetwork was not sufficient to create the requested motion upwards or downwards the semantic tree. The same applies to increasing the temperature T . In both cases, most of the time the network would converge to various mixed states, not assumed to be of cognitive significance. Another conclusion of our work is that without bias the network’s behavior is sensitive to small variations in the values of the model parameters presented, and we have performed quite a number of simulations in order to find the parameters leading to best performance. However, when bias is introduced performance becomes much more reliable and more robust via parameter values modification. However, the above mentioned strategies were successful when, removing the geometric constraint, the connections between the various patterns in the items subnetwork were explicitly implemented by pointers between the patterns.

Let us now compare the retrieval times (RT) of queries obtained in our model, with Collins and Quillian’s psychological data [12]:

1. The main similar feature is the cascade of RTs obtained for positive queries. As with the psychological data, when the RT’s are plotted against the level of the

query, straight lines are obtained. While using the bias mechanism, the slope of the lines can be controlled mainly by changing the bias. If a constant rate of bias increase is assumed, the lines representing the RT of OA and OO queries are predicted to be parallel.

2. The zero level OO query has a shorter RT than predicted by straight line extrapolation. This is explained by Collins and Quillian by assuming that this query is treated by pattern recognition and not by semantic retrieval. According to our model, this phenomena is expected since when the two inputs are identical the network is initialized to the desired stable end-state and no convergence time is needed.
3. According to the psychological experimental data, OA and OO queries are represented by two parallel lines, with the OA line translated upwards relative to the OO line by $\Delta t = 225ms$. This shift time, which is three times larger than the time of a one-level ascendancy in the tree structure, is explained by Collins and Quillian as the time needed for performing a search process for the queried attribute. While ANN implementations of this scheme are possible in principle, they were not realized in our model since no search process is explicitly defined. The relatively longer (OA) RT's mentioned above are explained using our model by the additional time needed to allow for the convergence of the attributes subnetwork. However, this leads to a shorter expected shift than the one experimentally observed. Yet, since we assume weak modularity, it is plausible that processes involving both subnetworks entail a constant delay due to the flow of information between the subnetworks and thus the actual value of the observed shift Δt can be larger.
4. As described, no correlation was found between the RT of a negative query and the semantic distance between its components. However, it should be noticed that besides the strategy used in the model presented, there are several other plausible strategies for determining the time when a query is considered 'unretrievable' and a negative response is generated;
 - (a) The negative response is generated upon convergence of the item subnetwork into an attractor which is the root of the structure tree and whose semantic meaning is 'negative response'. However, according to such a strategy negative RTs for patterns closer to the root are expected to be shorter, which was not observed experimentally.
 - (b) Negative facts could be directly stored in the network. This strategy is compatible with our model since projections from items towards the respective attributes (and vice versa) may be negative, and it could be interesting to investigate it in the future.

Considerable simplifications have been made in the construction of the model presented. It would be interesting to perform further research focused on modeling semantic networks not limited to simple tree structures and to investigate the dynamics of the model when associations between concepts that are not hierarchically related nor is one the property of the other (e.g., between a zebra and a crossroad), would be entered. We have shown that constructing a model for semantic fact retrieval using a distributed ANN approach is indeed feasible and enables a quite fair simulation of the corresponding psychological experimental results. Using such an approach, we have demonstrated that some properties of the cognitive dynamics predicted by the model are obtained implicitly without any need for their explicit formulation.

References

- [1] Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol I, D.E. Rumelhart and J.L. McClelland (Eds.) Cambridge, MA: MIT Press. 1986.
- [2] Smolensky P., Connectionist Modeling: Neural Computations/Mental Connections. Chapter 2, 49 - 67, Neural Connections, Mental Computation, Nadel L., Cooper L. A., Culicover P. and Harnish M., eds., A Bradford Book, MIT press Cambridge Mass, 1989.
- [3] Fahlman S., Representing Implicit Knowledge. In G.E. Hinton and J.A. Anderson (Eds.), Parallel Models of Associative Memory, Erlbaum.
- [4] Hollbach Weber S., A connectionist model of conceptual representation. IJCNN I, 477-483, 1989.
- [5] Shiue L.C. and Grondin R.O., Neural processing of semantic networks. IJCNN II, 589, 1989.
- [6] Jagota A. and Jakubowicz O., Knowledge Representation in a Multy Layered Hopfield Network IJCNN I, 435-442, 1989
- [7] Ritter H. and Kohonen T., Self-Organizing Semantic Maps. Biol. Cybern. 61, 241-254, 1989.
- [8] Hinton G.E., Implementing semantic networks in parallel hardware, In G.E. Hinton and J.A. Anderson (Eds.), Parallel Models of Associative Memory, Erlbaum.
- [9] Amit D.J., Sagi D. and Usher M., Architecture of Attractor Neural Networks Performing Cognitive Fast Scanning. To appear in Networks.

- [10] Miyashita Y. and Chang H.S., Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331, 68, 1988.
- [11] Fodor J. and Pylyshyn Z.W., Connectionism and cognitive architecture: a critical analysis. *Cognition* 28: 3-71, 1988.
- [12] Collins A.M. and Quillian M.R., Retrieval Time From Semantic Memory, *Journal of verbal learning and verbal behavior*, 8, 240-248, 1969.
- [13] Fahlman S., NETL: A system for representing real-world knowledge. MIT Press, Cambridge, MA, 1979.
- [14] Amit D.J., Parisi G. and Nicolis S., Neural potentials as stimuli for attractor neural networks. *Network*, 1, 75 - 88, 1990.
- [15] Quillian M.R., Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Sci.* 12, 410-430, 1967.
- [16] Quillian M.R., "Semantic Memory", in *Semantic Information Processing*, M.Minsky (Ed.), MIT press, Cambridge, Mass, 1968.
- [17] Quillian M. R., The Teachable Language Comprehender: A simulation program and theory of language. *Communications of the ACM*, 12, 459-476, 1969.
- [18] Schubert L., "Extending the Expressive Power of Semantic Networks," *Artificial Intelligence*, Vol. 7, No. 2, 1976.
- [19] Findler N.V., (Ed.), *Associative Networks: Representation and Use of Knowledge by Computer*, Academic Press, New York, 1979.
- [20] Schank R. C. and Colby K., (Eds.), *Computer Models of Thought and Language*, W.H. Freeman, San Fransisco, CA, 1973.
- [21] Collins A.M. and Loftus E.F., A spreading activation theory of semantic processing. *Psychological Review*, 82, 407-428, 1975.
- [22] Hopfield J.J., Neural networks and physical systems with emergent collective abilities. *Proc. Nat. Acad. Sci. USA* 79, 2554-2558 (1982).
- [23] Feigelman M.V. and Ioffe L.B., The augmented models of associative memory asymmetric interaction and hierarchy of patterns. *Int. Jour. of Mod. Phys. B*, Vol 1, 51-68 (1987).
- [24] Gutfreund H., *Neural Networks with Hierarchically Correlated Patterns*. ITP-SB-86-151 preprint 1987.

- [25] Amit D., Gutfreund H. and Sompolinsky H., Information storage in neural networks with low levels of activity. *Phys. Rev. A*, 35, 2293-2303 (1987).
- [26] Anderson J.R. Spreading Activation, in Anderson J.R. and Kosslyn S. M. (eds). *Essays in learning and Memory*. W.H. Freedman and Comp., New York (1984). pattern classification and universal recoding, II: expectation, olfaction, and illusions. *Biol. Cybern.*, 23, 187-202