A Neural Architecture for a Class of Abduction Problems

Ashok K. Goel* J. Ramanujam†

Abstract

The general task of abduction is to infer a hypothesis that best explains a set of data. A typical subtask of this is to synthesize a composite hypothesis that best explains the entire data from elementary hypotheses which can explain portions of it. The synthesis subtask of abduction is computationally expensive, more so in the presence of certain types of interactions between the elementary hypotheses. In this paper, we first formulate the abduction task as a nonmonotonic constrained-optimization problem. We then consider a special version of the general abduction task that is linear and monotonic. Next, we describe a neural network based on the Hopfield model of computation for the special version of the abduction task. The connections in this network are symmetric, the energy function contains product forms, and the minimization of this function requires a network of order greater than two. We then discuss another neural architecture which is composed of functional modules that reflect the structure of the abduction task. The connections in this second-order network are asymmetric. Finally, we discuss how the second architecture can be extended to solve the general task of abduction.

1 Abductive Inference

Abduction is inference to the best explanation for a given set of data. The general task of abduction takes as input a set of data and gives as output a hypothesis that can best explain the input data. A typical subtask of the abduction task is *classification* of the given data into potentially relevant elementary explanatory hypotheses stored in memory [19]. In the classification task, the stored elementary hypotheses are matched with the data, and, depending on the degree of match, a *prima facie* belief value for each explanatorily relevant hypothesis is determined. For simple abductive problems, for instance diagnosis under the single fault assumption, the classification subtask may yield elementary hypotheses that can individually explain the entire data. For such problems, the elementary hypothesis with the highest belief value represents the best explanation.

In general, however, an elementary hypothesis that can explain the entire data may not be available in memory. Instead, a composite explanation has to be *synthesized* from elementary hypotheses that can explain various portions of the data [19]. What makes one composite hypothesis operationally better than others are factors such as explanatory coverage, plausibility, parsimony and internal consistency. Synthesizing a composite hypothesis that satisfies these criteria for a best explanation, however, is computationally very expensive, more so in the presence of certain types of interactions between the elementary hypotheses [1]. This raises the issue of how to rapidly synthesize composite explanations from elementary explanatory hypotheses.

In [14] we proposed the exploitation of concurrency in the synthesis of composite explanations and described a distributed-memory message-passing architecture for this purpose. Our work on concurrent

*College of Computing, Georgia Institute of Technology, Atlanta, Georgia 30332-0280 (goel@gatech.edu).

†Department of Electrical and Computer Engineering, Lousiana State University, Baton Rouge, Lousiana 70803-5901, (jxr@ee.lsu.edu).

synthesis of composite explanations led us to think in terms of artificial neural networks for the abduction task. From a concurrent-processing viewpoint, neural networks form an attractive proposal for an efficient, fine grained, massively parallel machine dedicated to some special class of problems [29]. In [13] we described a preliminary neural architecture based on Hopfield networks [16, 17] for synthesizing composite explanations. In this paper, we formalize the abduction task as a constrained optimization problem, elaborate on the neural architecture for the synthesis subtask of abduction, compare our computational model with similar proposals, and discuss its implications.

2 Abduction as a Combinatorial Optimization Problem

In this section we provide a characterization of the general task of abduction. We begin with a formal definition of the abduction task, then specify the classification [1] and synthesis subtasks of abduction, next specify the criteria that make a explanatory hypothesis a best explanation for a set of data, then characterize some of the interactions that can occur between elementary hypotheses, and finally describe a special version of the general abduction task.

2.1 Characterization of the Abduction Task

Let $D = \{d_i | i = 1, \ldots, N\}$ be a finite set of N data (effects, facts, etc.).
Let $H = \{h_j | j = 1, \ldots, M\}$ be a finite set of M elementary (causal) hypotheses.
Let q be a map from subsets of H to subsets of D; $q : \wp(H) \rightarrow \wp(D)$. $\wp(S)$ denotes the *power set* of a set S, *i.e.,* the set of all subsets of S. We define $q(\{h_j\})$ as $q(h_j)$ for $h_j \in H$; thus the map q is also defined from an elementary hypothesis to subsets of D. We interpret $q(h_j) = D_j$, where $h_j \in H$ and $D_j \subseteq D$, as the *explanatory coverage* of h_j, *i.e.* h_j can explain only and all members of D_j.
Let D_o be a set of n *observed* effects and *given* facts, where $D_o \subseteq D$.
The information-processing task of abduction may be characterized as a five-tuple $\langle D, H, q, D_o, H_c \rangle$, where D, H, q, and D_o are the inputs to the task, and H_c, the output of the task, is a subset of H, $H_c \subseteq H$, that best explains D_o.

2.2 The Classification and Synthesis Subtasks of Abduction

Let $B = \{b_k | k = 1, \ldots, l\}$ be a finite set of l belief values.
Let $H_e \subseteq H$ such that each $h_j \in H_e$ can explain some non-empty subset of D_o.
Let p be a map from H_e to B; $p : H_e \rightarrow B$. The map p is also defined from an elementary hypothesis to belief values. we define $p(\{h_j\})$ as $p(h_j)$ and interpret $p(h_j)$ as the *prima facie* belief value for h_j.
The classification subtask takes D, H, D_o, and r as input, where r is a map from $\wp(D_o) \rightarrow \wp(H)$, and gives H_e and p as output [11, 15]. The synthesis subtask of abduction may be characterized as a five-tuple $\langle D_o, H_e, q, p, H_c \rangle$, where D_o, H_e, q, and p constitute the input to the task, and H_c is the output of the task.

Since the classification problem is relatively well understood, and a number of well known techniques for solving it already exist, we will not discuss it any further in this paper (Bylander *et al.* [1] and Chandrasekaran and Goel [3] provide surveys of some classification techniques). Instead, we will focus on the task of synthesizing composite explanations.

2.3 Characterization of the Best Explanation

The best composite explanation can be operationally characterized based on the following three optimization criteria:

1. Maximal explanatory coverage of data: A composite hypothesis H_1^c is a better explanation of D_o than another composite hypothesis H_2^c if

   $$q(H_1^c) \cap D_o \supset q(H_2^c) \cap D_o$$

   Ideally, the assembled composite hypothesis, H_c, would provide complete explanatory coverage of D_o, *i.e.,* $q(H_c) \supseteq D_o$.

2. Maximal belief in hypothesis: A composite hypothesis H_1^c is a better explanation of D_o than another composite hypothesis H_2^c if

   $$p(H_1^c) \geq p(H_2^c)$$

   This specifies that among composite hypotheses that explain the data, the one with the highest belief value is the "best" explanation.

3. Minimal hypothesis: A composite hypothesis H_1^c is a better explanation of D_o than another composite hypothesis H_2^c if

   $$|H_1^c| < |H_2^c|.$$

   This *global* optimization condition specifies that H_c should be *parsimonious*.

Of these, the first and the third criteria are independent of belief values. Deciding on the "best" set of belief values for a problem at hand is extremely difficult. We choose to interpret the second criterion as follows: A composite hypothesis H_1^c is a better explanation of D_o than another composite hypothesis H_2^c if

$\forall d \in D_o$,
$\forall h_2 \in H_2^c$ such that $d \in q(h_2) \cap D_o$,
$\exists h_1 \in H_1^c$ such that $(d \in q(h_1) \cap D_o) \wedge (p(h_1) \geq p(h_2))$.

This specifies that the component hypotheses in H_c should be *locally* optimal in terms of their belief values. In the presence of interactions among elementary hypotheses, local optimality is critical in deriving heuristic solutions to the abduction problem (See the discussion in Section 2.4).

The general abduction task is over constrained due to the potential conflicts among the above criteria for a best explanation. This conflict may be resolved by imposing a *precedence relation* according to which maximal coverage of the data has the highest precedence and parsimony of the composite hypothesis has the lowest [20]. However, depending on the functions q and p, the synthesis task may now be underconstrained, in which case the synthesized explanation would only be a best explanation.

2.4 Interactions Among Elementary Explanations

Several distinct types of interaction are possible between two elementary explanatory hypotheses $h_1, h_2 \in H_e$ [1, 19]:

- *Associativity:* The inclusion of h_1 in H_c suggests the inclusion of h_2. Such an interaction may arise if there is knowledge of, say, a statistical association between h_1 and h_2.

- *Additivity:* h_1 and h_2 cooperate additively where their explanatory capabilities overlap. This may happen if h_1 and h_2 can separately explain some datum $d \in D_o$ only partially, but collectively can explain it fully.

- *Incompatibility:* h_1 and h_2 are mutually incompatible, *i.e.,* if one of them is included in H_c then the other should not be included.

- *Cancellation:* h_1 and h_2 cancel the explanatory capabilities of each other in relation to some $d \in D_o$. For example, h_1 might imply that an increase in some data value, while h_2 may imply a decrease in the value, thus canceling each others explanatory capability with respect to that datum.

The abduction problem is *nonlinear* in the presence of incompatibility interactions and *nonmonotonic* in the presence of cancellation interactions. The general (nonlinear, nonmonotonic) abduction task is NP-complete [1].

2.5 A Special Version of the General Abduction Task

Let us now consider a special version of the general problem of synthesizing composite hypotheses that is linear, and, therefore, also monotonic. The synthesis task is *linear* [1] if

$$\forall h_i, h_j \in H_e, \quad q(h_i) \cup q(h_j) = q(\{h_i, h_j\})$$

and it is *monotonic* if

$$\forall h_i, h_j \in H_e, \quad q(h_i) \cup q(h_j) \subseteq q(\{h_i, h_j\})$$

In this special version of the problem, we assume that the elementary hypotheses are non-interacting, *i.e* they offer mutually compatible explanatory alternatives where their explanatory coverages overlap. We also assume for simplicity that the belief values found by the classification subtask of abduction for all $h \in H_e$ are equal to 1 (or *True*).

Under these conditions, the synthesis subtask of abduction can be represented by a bipartite graph consisting of nodes in the set $D_o \cup H_e$. There are no edges between the nodes in D_o, nor are there any edges between the nodes in H_e. The edges between the nodes in D_o and the nodes in H_e can be represented by a matrix Q of m rows and n columns, where the rows correspond to data $d_i \in D_o$ and the columns correspond to hypotheses $h_j \in H_e$. The entries in the matrix may be denoted Q_{ij} and indicate whether a given data is explained by a specific hypothesis. The entries are defined as:

$$Q_{ij} = \begin{cases} 0 & \text{if datum } \mathrm{d_i} \text{ is not explained by hypothesis } \mathrm{h_j} \\ 1 & \text{if datum } \mathrm{d_i} \text{ is explained by hypothesis } \mathrm{h_j} \end{cases}$$

Given the matrix Q for the bipartite graph, the synthesis subtask of abduction can be modeled as a set-covering problem, which is to find the minimum number of columns that cover all the rows. This ensures that the composite hypothesis will explain all elements of D_o and will be parsimonious. (Since the belief values for all $h \in H_e$ are assumed to be 1, these are the only two remaining criteria for characterizing a best explanation.) Note that the general set cover problem is NP-complete [10].

3 The Hopfield Model of Computation

Hopfield [16, 17] has proposed a neural network in which highly interconnected neurons collectively compute good solutions to difficult optimization problems such as the Traveling Salesman Problem (which too is NP-complete) [18]. The neurons in the network are analog devices which may make them closer to biological neurons than strictly digital models. The power of this model of computation comes from the rapidity with which acceptable solutions are found, though the solutions are not guaranteed to be the globally optimal. The emphasis in the model is on exploitation of massive parallelism as opposed to the pursuit of the best solution, the meaning of which often carries a certain degree of arbitrariness for many real world problems.

The processing elements (or neurons) can be modeled as amplifiers having a sigmoid input-output relationship defined by $V_j = g(u_j)$, where the output voltage V_j of amplifier j is a function of the input voltage u_j. I_j represents the input bias current to the j^{th} neuron. Connectivity between pairs of neurons is defined by a connectivity matrix $\{W_{ij}\} : i, j = 1, \ldots, N$, where a negative value of connectivity indicates that the connection is inhibitory. Hopfield [17] has shown that in case of symmetric connections between neurons *i.e.*, $\forall i, j \quad W_{ij} = W_{ji}$, the network evolves to stable states in which the outputs of all neurons reach a constant value. The time evolution of the individual neurons is given by

$$\frac{du_j}{dt} = -u_j + \sum_{i=1}^{N} W_{ij} V_i + I_j$$

The specific values of the steady-state output voltages V_j obtained from the time evolution of the network are determined by the bias currents I_j and the initial values of the input voltages u_j.

If the diagonal elements of the W matrix are zero (with symmetric connectivities) and the amplifiers operate in a high gain mode, *i.e.*, their gain functions are good approximations to threshold functions, then the stable states of the network are the local minima of the *energy function* defined by

$$E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} V_i V_j - \sum_{j=1}^{N} V_j I_j$$

The state space of the network of N neurons is defined by the interior of a hypercube of dimension N, and the set of all local minima of energy E is defined by the values of W_{ij}. By a proper choice of the gain function, the local minima of E can be constrained to occur at the corners of the N-dimensional hypercube *i.e.*, with all the values of V_j = 0 or 1. This is especially useful when one is interested in digital solutions (0 or 1) to a given problem. An appropriate choice for the gain function is

$$V_j = \frac{(1 + \tanh(u_j))}{2}$$

If instead, we use purely digital neurons, problems such as cycling through a set of states may arise. In addition, with purely digital neurons, the deterministic search for the solution will have to proceed along the edges of the hypercube that defines the search space. However, by using analog neurons, one expects to *cut through* the search space.

4 A Neural Network for the Abduction Task

In this section, we consider the special (linear, monotonic) version of the general synthesis subtask of abduction discussed in Section 2.5, and describe two neural networks for solving it. The first network is based on the Hopfield model of computation for optimization problems, and the second network uses Tank and Hopfield's computational architecture for linear programming [33].

To solve problems using neural networks, one must cast them into the neural network model of computation. For the synthesis subtask of abduction, we associate a neural variable V_j with each hypothesis $h \in H_e$ to indicate if the hypothesis is included in the composite hypothesis C. We minimize the cardinality of the composite hypothesis

$$\sum_{j=1}^{m} V_j$$

subject to the constraint that all the data $d \in D_o$ are completely explained *i.e.*,

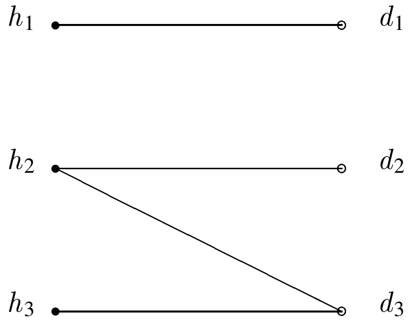

Figure 1: An Instance of Abduction

$$\forall i = 1, 2, \ldots, n, \quad \sum_{j=1}^{m} Q_{ij} V_j \geq 1$$

The term in the energy function that represents the problem constraint must evaluate to zero when the constraint is satisfied and must evaluate to a large positive value when the constraint is not satisfied, forcing the network to evolve accordingly. We chose to use a term expressed as a sum of expressions, one for each datum d_i, such that the expression evaluates to zero, when a hypothesis h_j that can explain the datum d_i is in the composite, i.e., $V_j = 1$. Given that Q is an incidence matrix (with elements having a value 0 or 1), the expression

$$\sum_{i=1}^{n} \prod_{j=1}^{m} \{(1 - Q_{ij}) + (1 - V_j)\}$$

satisfies the following conditions:

1. Each product expression can never evaluate to a negative number,
2. The sum of the product terms, thus, can never evaluate to a negative number,
3. Each product term evaluates to zero when a hypothesis that can explain the datum is in the composite; otherwise, it evaluates to a large value,
4. The sum of the product term, thus, evaluates to zero when a composite set of hypotheses that can explain all the data is found.

To illustrate, let us consider the very simple example shown in Figure 1. This example contains three hypothesis, h_1, h_2, h_3, and three data elements, d_1, d_2, d_3. In this example, d_1 can be explained only by h_1 and d_2 can be explained only by h_2, where as d_3 can be explained by both h_2 and h_3. If we choose h_1 and h_3 but not h_2 (i.e., $V_1 = 1$ and $V_3 = 1$ but $V_2 = 0$), the value of the expression for datum d_2 is 1 and hence the constraint term has a value 1 (the other two data result in a value zero, since hypothesis that can explain them is in the composite). If we include all the three h_j's, the constraint term evaluates to zero; for the case with h_1 and h_2 in the composite, the term again evaluates to zero. However, the second case results in a lower value of the actual function $\sum_{j=1}^{m} V_j$ (the cardinality of the composite hypothesis) being minimized; thus the network would choose the second case as the solution for the abduction instance in Figure. 1.

Thus, we derive the energy function as follows:

$$E = \alpha * \sum_{j=1}^{m} V_j \quad + \quad \beta * \sum_{i=1}^{n} \prod_{j=1}^{m} \{(1 - Q_{ij}) + (1 - V_j)\}$$

where α and β are positive constants. The first term in the energy function represents the cardinality of the composite hypothesis and the second term represents the penalty for lack of complete coverage. The second term will have a value 0, which is its lowest, in the case of complete coverage – for each datum $d \in D_o$, the product will be 0 if there is at least one hypothesis $h \in H_e$ that can explain it, and among those that can explain it, at least one of them is included in the composite hypothesis. Since, ensuring complete coverage of the data has a higher precedence than parsimony of the composite hypothesis, the constant β should be much larger than α.

We note the appearance of the *product* form in the energy function. This requires a k-th order neural network, where $k = \max_i \sum_{j=1}^{m} Q_{ij}$, the largest number of hypotheses $h \in H_e$ that can explain any one of $d \in D_o$. A k-th order neural network admits up to k-way connections among the neurons [31]. A k-way connection could be thought of as a bus connecting those neurons which participate in the connection. With symmetric k-way connections, the energy function of such a network can be shown to be non-increasing with time. An example of a 3-way symmetric connection is one where

$$\forall i, j, k \qquad W_{ijk} = W_{ikj} = W_{jki} = W_{jik} = W_{kij} = W_{kji}$$

i.e., the connectivity W is the same for all permutations of the subscripts. Note that such a *higher-order network* can solve the synthesis task using only m neurons, one for each $h \in H_e$.

4.1 Another Neural Architecture

We now describe a second neural architecture for the special version of the general synthesis subtask of abduction discussed in Section 2.5. Unlike the first neural network, this architecture is based on a second order neural network with asymmetric connections between the neurons.

The key idea in this architecture is to explicitly represent the constraints of the abductive task on the neural network. The network is composed of different subnets, each responsible for ensuring the satisfaction of the different criteria for the best explanation: one subnet represents the hypotheses $h \in H_e$ and ensures global parsimony of the composite hypothesis while another subnet explicitly represents the local constraints of explaining the data elements $d \in D_o$ and ensures maximal explanatory coverage of D_o. These subnets may be viewed as *functional modules* that reflect the structure of the problem and cooperatively perform the task of abduction. Note that assigning different functions to subnets comprising the network makes for a *structured neural network* [8].

In fact, this is the functional organization that we initially used for the distributed-memory message-passing model for abduction [14]. In this model, one layer of processes (neurons) representing the m elementary hypotheses are responsible for ensuring parsimony of the composite hypothesis as well as for accommodating interactions between the hypotheses for the general abduction problem. Another layer of processes (neurons) corresponding to the n data elements are responsible for ensuring maximal coverage of data (as well as ensuring maximal belief in the composite hypothesis for the general abductive task). Each process (neuron) representing an explanatory hypothesis makes local decisions based on the knowledge available to it and communicates (sends a signal) its results to the relevant data processes. Similarly, each process (neuron) corresponding to a datum to be explained makes local choices and communicates its results to the appropriate hypotheses processes. This flow of information back and forth between the two layers of processes continues until a composite hypothesis is fully synthesized. The *adaptive resonance theory*

[2] uses a similar functional organization of processes which allows feedback of information between two layers of processes.

Tank and Hopfield [33] have proposed a neural network for the linear programming problem that implicitly captures many of these ideas, and our second neural architecture can be mapped on to their scheme. The network contains $m + n$ neurons. The output voltages of the m neurons which represent hypotheses $h \in H_e$ in the module responsible for ensuring parsimony are denoted as $f_j, j = 1, \ldots, m$. These neurons operate in the high gain mode in order to provide $0/1$ solutions. The module responsible for satisfying the constraints of explaining each $d \in D_o$ is made up of n neurons denoted by $g_i, i = 1, \ldots, n$, one for each constraint, $\sum_{j=1}^{m} Q_{ij} V_j \geq 1$. The output of the neuron g_i is zero if the corresponding datum d_i is explained by some elementary hypothesis in the composite hypothesis.

The pattern of connectivity among the modules represents the interaction between the competing criteria which characterize the best explanation. Thus, in the network for abductive inference, the output of each f_j hypothesis neuron is transmitted to the input of those constraint neurons g_i which represent a constraint that the hypothesis can satisfy. Similarly, the output of each g_i neuron is transmitted to those f_j neurons that can help satisfy the constraint. The relative speeds of operation of the neurons in the different modules reflect the precedence relationship among the criteria. Thus the g_i neurons operate at a much higher speed than the f_j neurons reflecting the precedence relation between the criteria of complete coverage and parsimony.

5 Discussion

Abductive inference appears to be ubiquitous in cognition. Abduction occurs, for instance, in diagnostic problem solving, where the data is in the form of manifestations (or symptoms) and the explanatory hypotheses are about malfunctions (or diseases) [25, 26]. Data interpretation, where the data is in the form of observations (or sensor readings) and the explanatory hypotheses are about objects (or events); and situation assessment, where the data is in the form of events and the explanatory hypotheses are goals and plans ascribed to agents, are also instances of abduction. Some aspects of natural language processing, visual image processing, and explanation-based learning appear to be abductive in character as well ([6] gives an account of how these tasks involve abductive inference).

5.1 Related Research

Previous research on abduction has led to the development of several general computational theories of inference to the best explanation. Our characterization of the abduction task and its decomposition into the subtasks of classification and synthesis directly builds on [19] with one notable difference. In [19], Josephson et al view the property of parsimony of the composite explanation like this: a composite hypothesis H_1^c is a better explanation of the observed data D_o than another composite hypothesis H_2^c if $H_1^c \subset H_2^c$ (see [1]). In contrast, we view parsimony directly in terms of the cardinality of the composite hypothesis: H_1^c is a better explanation of D_o than H_2^c if $|H_1^c| < |H_2^c|$. This characterization of parsimony is similar to that of [28].

Theories of abductive inference range from "heuristic" mechanisms to "normative" models. Heuristic mechanisms include the early and pioneering work of Pople [25, 26], and the more recent work of Charniak [5] and Josephson et al [19]. In [14, 20], we describe a general heuristic useful in synthesizing composite explanations based on the notion of *explanatorily essential* elementary hypotheses. An elementary hypothesis is explanatorily essential if it provides an unique explanation of some $d \in D_o$. The presence of explanatorily essential elementary hypotheses can be used to decompose the synthesis task like this: first, include the explanatorily essential elementary hypotheses in the composite hypothesis H_c; then, remove all elementary hypotheses from H_e that are incompatible with the essential hypotheses in H_c, and also remove all $d \in D_o$ that are explained by H_c; finally, from the remainder of H_e, include those elementary hypotheses in the composite hypothesis that are needed to explain the remainder of D_o. [9] describes an empirical study of how this heuristic results in more efficient synthesis of better composite explanations.

Normative models of abduction include Reggia, Nau and Wang's [28] set cover model, Pearl's [23] probablistic reasoning model, de Kleer and Williams' qualitative reasoning model [7], and Reiter's model of reasoning from first principles [27]. Although our characterization of the abduction task is based on [19], our computational model for synthesizing composite explanations too is a normative model of abductive inference. Unlike other normative models, however, our model of abduction is based on a connectionist model of computation.

The goal of exploiting concurrency to rapidly synthesize composite explanations has motivated much work both with in the classical symbolic paradigm and in the connectionist paradigm. In the classical symbolic paradigm, Sekar and Reggia [30] describe a two-processor architecture in which one processor represents D_o and incrementally removes those $d \in D_o$ that already have been explained by some $h \in H_e$, and another processor, representing H_e, removes those $h \in H_e$ that are no longer needed for explaining the remaining data elements in D_o. Charniak [4] describes the use of marker-passing techniques on spreading activation networks for composing an explanation. Pearl's [23] probablistic reasoning model for abduction also is distributed. In his scheme, the elementary hypotheses are organized in a bayesian network. Sticklen [32] describes the use of multiple classification agents for retrieving elementary hypotheses from memory but he does not explore the issue of concurrency. Our distributed-memory message-passing architecture for abduction described in [14] is somewhat similar to Sekar and Reggia's multiprocessor scheme. The difference lies in that our model associates a process with each $h \in H_e$ and each $d \in D_o$. This enables considerably more distribution of processing and allows for exploitation of parallelism with in H_e and D_o.

In the connectionist paradigm, Peng and Reggia [24] describe a competition-based neural network for solving a class of problems in medical diagnosis. Like our second network, their network contains connections among the neurons representing $h \in H_e$ and $d \in D_o$ where the connections represent the explanatory relations between H_e and D_o. In addition, their network contains excitory and inhibitory connections among the neurons representing $h \in H_e$; these connections represent the associativity and incompatibilty interactions among the elementary hypotheses described earlier. Thagard [34], and, more recently, Wang and Ayeb [35] have described similar competition-based neural networks for solving a class of abduction problems.

In contrast, our two neural networks directly build on Hopfield's model of computation. We agree with Thagard's argument that excitory and inhibitory connections among the neurons representing $h \in H_e$ are needed to ensure that the synthesized explanation is internally consistent and coherent. Indeed, such connections are natural to Hopfield networks, and, thus, it is quite possible to extend our second neural architecture for the purpose of ensuring the internal consistency of the composite explanation. First, the *prima facie* belief values of the elementary explanatory hypotheses are incorporated into the architecture by assigning different weights to the connections from the f_j neurons to the g_i neurons. Thus the weight of the connections from the f_j neuron representing the elementary hypothesis $h \in H$ represents the belief value $p(h)$. Next, the associativity interaction between two elementary hypotheses $h_1, h_2 \in H_e$ is accommodated by providing an excitatory connection between the neurons representing h_1 and h_2. Similarly, the incompatibility interaction between $h1$ and $h2$ is accommodated by providing an inhibitory connection between the neurons corresponding to them. Further, the mechanism for handling additive interaction between two hypotheses $h_1, h_2 \in H_e$ is implicit in the architecture.

Nevertheless, we have chosen not to commit to competition-based neural networks such as that used by Peng and Reggia. This is due to several reasons. Firstly, it is not clear whether competition-based networks indeed can converge to the best explanation for abduction problems. Peng and Reggia apparently found that their network converged to the best explanation only for relatively small and simple abduction problems. Secondly, it is not clear whether associativity and incompatibility interactions among the elementary hypotheses can be adequately represented by static connection weights as used by Thagard. Thirdly, the

Hebb's rule of learning employed by Wang and Ayeb for learning the right set of connection weights for abduction problems is very sensitive to the initial conditions of the network. In [21], we describe a series of experiments on the use of Hebb's rule for learning the right set of connection weights in a multilayered backpropagation network for the exclusive disjunction (XOR) problem. We found that the number of training sessions needed for learning the right weights varied from a few thousand to several hundred thousand depending on the initial conditions of the network. Finally, it is not clear how to accommodate cancellation interactions among elementary hypotheses in a competition-based scheme. In fact, none of the competition-based schemes described in [24, 34, 35] can handle this type of interaction. This is important because abductive inference is non-monotonic in the presence of this interaction. The ability to accommodate cancellation interactions in abduction problems appears to require some kind of additional processing.

5.2 Critique of the Neural Architectures

Our first neural network for abduction is directly based on Hopfield's model of computation. The network contains only m neurons, one for each $h \in H_e$. The connections in the network are symmetric but of order greater than two. While the notion of a higher order neural network for abduction appears theoretically attractive, it is not clear how plausible this network is from the biological and engineering viewpoints. From the biological perspective, the evidence for the existence of higher order synaptic connections is unclear. >From an engineering point of view, regular two-dimensional (or those embeddable in a plane) *VLSI* networks exhibiting nearest neighbor interconnection are especially attractive for implementing neural architectures [22]. In these networks, the connectivity among neurons is realized in terms of resistors and capacitors between nodes. In order to realize a kth order network, the size of the network of resistors and capacitors that reliably simulates a kth order network is extremely high. A network of n units would require $O(n^k)$ connections to implement all terms of order k. In the case of the abduction task, it is not clear how to design a single network that would work for different instances of the same problem. If the size of the network, and hence the size of the *RC* network, is variable, the accuracy of the simulation will be lost.

The second neural architecture of abduction evolves from our distributed-memory message-passing scheme for concurrent synthesis of composite explanations that can be mapped on to Tank and Hopfield's model of computation. The second architecture contains $m + n$ neurons, one for each $h \in H_e$ and each $d \in D_o$. Unlike the first network, however, it requires neither symmetric connections between the neurons nor a higher order network. The use of asymmetric connections is closer to the current models of synaptic connections. The reduction in the complexity of the connections in the architecture is due to the explicit representation of the constraints of the abduction task and the functional modularization of the network.

The computational advantages of explicit representation of information-processing abstractions (such as the problem constraints) and functional organization of processing (such as the modularization of the network into specific subnets for accommodating the problem constraints) are independent of the underlying neural mechanisms of their realization. In [3, 4, 12], we argued that it is these abstractions and organizations that provide the computational theory of performing an information-processing task such as that of abduction. Also, in [21] we empirically showed that the computational power of neural networks often arises from their organization and from the abstractions explicitly embedded in them. Our work on the use of neural networks for the abduction appears to confirm this.

Acknowledgements

This work has benefited from many discussions with P. Sadayappan and John Josephson at the Ohio State University.

References

[1] T. Bylander, D. Allemang, M. Tanner and J. Josephson, "The Computational Complexity of Abduction," *Artificial Intelligence*, 49:25-60, 1991.

[2] G. Carpenter and S. Grossberg, "The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network," *IEEE Computer,* 21(3):77-88, March 1988.

[3] B. Chandrasekaran and A. Goel, "From Numbers to Symbols to Knowledge Structures: Artificial Intelligence Perspectives on the Classification Task," *IEEE Transactions on Systems, Man, and Cybernetics*, 18(3):415-424, May/June, 1988; IEEE Press.

[4] B. Chandrasekaran, A. Goel and D. Allemang, "Connectionism and Information Processing Abstractions: The Message Still Counts More Than the Medium," *AI Magazine,* Winter 1988, pages 24-34.

[5] E. Charniak. Motivation Analysis, Abductive Unification, and Non-monotonic Equality. *Artificial Intelligence*, 34(3):275-295, 1988.

[6] E. Charniak and D. McDermott, *Introduction to Artificial Intelligence,* Addison-Wesley, Reading, MA, 1985.

[7] J. de Kleer and B. Williams. Diagnosing Multiple Faults. *Artificial Intelligence*, 32(1):97-130, 1987.

[8] J. Feldman, M. Fanty, N. Goddard and K. Lynne, "Computing with Structured Connectionist Networks", *Communications of ACM* 31(2):170-187, February 1988.

[9] O. Fischer, A. Goel, J. Svirbely, and J. Smith, "The Role of Essential Explanations in Abduction," *Artificial Intelligence in Medicine*, 3:181-191, Amsterdam, Netherlands: Elsevier, 1991.

[10] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* W.H. Freeman and Company, San Francisco, CA, 1979.

[11] A. Goel and T. Bylander. "Computational Feasibility of Structured Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(12):1312-1316, December 1989.

[12] A. Goel and G. Montgomery. "What is Abductive Reasoning?", *Neural Network Review*, 3(4):181-187, June 1990, Hillsdale, NJ: Lawrence Erlbaum Publishers.

[13] A. Goel, J. Ramanujam and P. Sadayappan, "Towards a Neural Architecture for Abductive Reasoning," in *Proc. Second IEEE International Conference on Neural Networks,* San Diego, CA, July 1988, Vol. II, pages 681-688.

[14] A. Goel, P. Sadayappan and J. Josephson, "Concurrent Synthesis of Composite Explanatory Hypotheses," in *Proc. 17th International Conference on Parallel Processing,* August 1988, Vol. III, pages 156-160.

[15] A. Goel, N. Soundararajan and B. Chandrasekaran, "Complexity in Classificatory Reasoning," in *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, Washington, July 1987, pages 421-425.

[16] J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," in *Proceedings of the National Academy of Sciences, USA* 79:2554-2558, 1982.

[17] J. Hopfield, "Neurons with Graded Response have Collective Computational Properties like those of Two-State Neurons," in *Proceedings of the National Academy of Sciences, USA* 81:3088-3092, 1984.

[18] J. Hopfield and D. Tank, "Neural Computation of Decisions in Optimization Problems," *Biological Cybernetics,* 52:141-152, 1985.

[19] J. Josephson, B. Chandrasekaran, J. Smith and M. Tanner, "A Mechanism for Forming Composite Explanatory Hypotheses," *IEEE Transactions on Systems, Man, and Cybernetics,* SMC-17(3):445-454, May/June 1987.

[20] J. Josephson and A. Goel, "Practical Abduction," in *Abduction: Computation, Philosophy, and Technology*, J. Josephson and S. Josephson (editors), New York: Cambridge University Press, in press.

[21] J. Kolen and A. Goel. Learning in Parallel Distributed Processing: Computational Complexity and Information Content. *IEEE Transactions on Man, Systems, and Cybernetics*, 21(2):359-367, March/April 1991.

[22] C. Mead, *Analog VLSI and Neural Systems,* Addison-Wesley, Reading, MA, 1989.

[23] J. Pearl, "Distributed Revision of Composite Beliefs," *Artificial Intelligence* 33(2):173-215, October 1987.

[24] Y. Peng and J. Reggia, "A Connectionist Model for Diagnostic Problem Solving," *IEEE Trans. Systems Man and Cybernetics*, 19:285-289, 1989.

[25] H. Pople, "On the Mechanization of Abductive Logic," in *Proc. Third International Joint Conference on Artificial Intelligence,* Stanford, CA, 1973, pages 147-152.

[26] H. Pople, "The Formation of Composite Hypotheses in Diagnostic Problem Solving: An Exercise in Synthetic Reasoning," in *Proceedings of the Fifth International Joint Conference in Artificial Intelligence*, Cambridge MA, August 1977, pages 1030-1037.

[27] R. Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 1987.

[28] J. Reggia, D. Nau and P. Wang, "A Formal Model of Diagnostic Inference," *Information Sciences* 37:227-285, 1985.

[29] D. Rumelhart, J. McClelland and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Volume. 1, The MIT Press, Cambridge, MA, 1986.

[30] S. Sekar and J. Reggia, "Parallel Set Covering Algorithms," in *Proc. Fourth IEEE International Conference on Artificial Intelligence Applications,* San Diego, CA, March 1988, pages 274-279.

[31] T. Sejnowski, "Higher-Order Boltzmann Machines," in *Neural Networks for Computing, AIP Conference Proceedings No. 151*, 1986, pages 398-403.

[32] J. Sticklen. Distributed Abduction in MDX2. In *Artificial Intelligence in Scientific Computation: Towards Second Generation Systems*, R. Huber, C. Kulikowski, J. Krivine, and J. David (editors), Basel, Switzerland: AG Scientific.

[33] D. Tank and J. Hopfield, "Simple Neural Optimization Networks: An A/D Converter, Signal Decision Circuit, and a Linear Programming Circuit," *IEEE Transactions on Circuits and Systems,* CAS-33(5):533-541, May 1986.

[34] P. Thagard, "Explanatory Coherence," *Brain and Behavioral Sciences,* 12:435-502, 1989.

[35] S. Wang and B. Ayeb. "Diagnosis: Hypothetical Reasoning With A Competition-Based Neural Architecture," in *Proc. International Joint Conference on Neural Networks*, Volume I, pages 7-12, Baltimore, 1992.