Harnessing Massive Parallelism for Tractable Reasoning — a cognitively motivated approach*

Lokendra Shastri
Department of Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
(215) 898-2661; shastri@cis.upenn.edu

February 1993

Extended Abstract

For discussion in the 1993 AAAI Spring-Symposium on "Innovative Applications of Massive Parallelism".

1 Introduction

We can understand language at the rate of *several hundred words per minute* (Carpenter & Just 1977) even though, doing so involves establishing referential and causal coherence, generating expectations and making predictions. This suggests that we can (and do) draw a wide range of inferences very rapidly, automatically and without conscious effort — as though they are a *reflexive* response of our cognitive apparatus. In view of this such reasoning may be described as *reflexive* (Shastri 1991). As an example of reflexive reasoning consider the inference 'John owns a car' upon hearing 'John bought a Rolls-Royce'. We can make this inference effortlessly even though it requires multiple steps of inference using background knowledge such as Rolls-Royce is a car and if x buys y then x owns y.

Not all reasoning is, and as complexity theory tells us, cannot be, reflexive. We contrast reflexive reasoning with *reflective* reasoning — reasoning that requires reflection, conscious deliberation, and at times, the use of external props such as paper and pencil (e.g., solving logic puzzles, doing cryptarithmetic, or planning a vacation).

1.1 Reflexive reasoning introduces a strong notion of tractability and necessitates the use of massive parallelism

Let us make a few observations about reflexive reasoning.

- *Reflexive reasoning occurs with respect to a large body of background knowledge.* Guha & Lenat (1990) have suggested that our background knowledge base may contain more than 10^7 items. This is not surprising given that this knowledge includes our knowledge of naive physics and naive psychology; facts about ourselves, our family, friends, colleagues, history and geography; our knowledge of artifacts, sports, art, music; some basic principles of science and mathematics; and our models of social, civic, and political interactions.

*This work was supported by NSF grants IRI 88-05465 and the ARO grant DAAL 03-89-C-003

Items in the background knowledge base are fairly stable and persist for a long-time once they are acquired. Hence this knowledge is best described as *long-term* knowledge and we will refer to this body of knowledge as the long-term knowledge base (LTKB).

- Episodes of reflexive reasoning are triggered by 'small' inputs. In the context of language understanding, an input (typically) corresponds to a sentence that would map into a small number of assertions. The critical observation is that *the size of the input, $|In|$, is insignificant compared to the size of the long-term knowledge base $|LTKB|$.*[1] [2]

- The vast difference in the magnitude of $|LTKB|$ (10^7– 10^8) and $|In|$ (a few) becomes crucial when discussing the tractability of common sense reasoning and we have to be careful in how we measure the time and space complexity of the reasoning process. Given the actual values of $|In|$ that occur during common sense reasoning, there is a distinct possibility that the overall cost of reasoning may be dominated by the "fixed" but much higher contribution of $|LTKB|$. Thus we cannot ignore the cost attributable to $|LTKB|$ and *we need to analyze the complexity of reasoning in terms of $|LTKB|$ as well as $|In|$.*

In view of the magnitude of $|LTKB|$, even a cursory analysis suggests that any inference procedure whose time complexity is quadratic or worse in $|LTKB|$ cannot provide a plausible computational account of reflexive reasoning. However, a process that is polynomial in $|In|$ does remain viable. We can place tighter bounds on the admissible time and space complexity of reflexive reasoning (see below).

1.1.1 Time complexity of reflexive reasoning

Observe that although the size of a person's $|LTKB|$ increases considerably from, say, age seven to thirty, the time taken by a person to understand natural language does not. This suggests that the time taken by an episode of reflexive reasoning does not depend on the $|LTKB|$. In view of this it is proposed that a realistic criteria of tractability for reflexive reasoning is one where the time taken by an episode of reflexive reasoning is *independent of* $|LTKB|$ and only depends on the depth of the derivation tree associated with the inference. This notion of tractability entails that the inferential process must be highly parallel and support the simultaneous evaluation of a large number of inferential paths.[3]

1.1.2 Space complexity of reflexive reasoning

The expected size of the LTKB also rules out any computational scheme whose space requirement is quadratic (or higher) in the size of the KB. For example, the brain has only about 10^{12} cells most of which are involved in processing of sensorimotor information. Hence even a linear space requirement is fairly generous and leaves room only for a modest 'constant of proportionality'. In view of this, it is proposed that the admissible *space requirement of a model of reflexive reasoning be no more than linear in $|LTKB|$.*

To summarize, it is proposed that as far as (reflexive) reasoning underlying language understanding is concerned, the appropriate notion of tractability is one where

- the reasoning time is independent of $|LTKB|$ and is only dependent on $|In|$ and the depth of the derivation tree associated with the inference, and

- the associated space requirement, i.e., the space required to encode the LTKB plus the space required to hold the working memory during reasoning should be no worse than *linear* in $|LTKB|$.

[1] A small input may, however, lead to a potentially large number of elaborate inferences. For example, the input 'John bought a Rolls-Royce' may generate a number of reflexive inferences such as 'John bought a car', 'John owns a car', 'John has a driver's license', 'John is perhaps a wealthy man', etc.

[2] Some of these inferences may be 'soft' inferences, but the issue of deductive versus evidential nature of inferences is irrelevant to our current concerns.

[3] The restriction that the reasoning time be independent of $|LTKB|$ may seem overly strong and one might argue that perhaps logarithmic time may be acceptable. Our belief that the stronger notion of effectiveness is relevant, however, is borne out by results which demonstrate that there does exists a class of reasoning that can be performed in time independent of $|LTKB|$.

In spite of the apparent significance of reflexive reasoning there have been very few attempts at developing a computational account of such inference. Some past exceptions being Fahlman's work on NETL (1979) and Shastri's work on a connectionist semantic memory (1988). However these models dealt primarily with inheritance and classification within an *IS-A* hierarchy. Hölldobler (1990) and Ullman and van Gelder (1988) have proposed parallel systems for performing more complex logical inferences, however, both these systems have unrealistic space requirements. The number of nodes in Hölldobler's system is quadratic in the the size of the knowledge base (KB) the number of processors required by Ullman and van Gelder is even higher.[4]

A significant amount of work has been done by researchers in knowledge representation and reasoning to identify classes of limited inference that can be performed efficiently (e.g., see Frisch & Allen 1982; Brachman & Levesque 1984; Patel-Schneider 1985; Dowling & Gallier 1984; Levesque 1988; Selman & Levesque 1989; McAllester 1990; Bylander, Allemang, Tanner, & Josephson 1991; Kautz & Selman 1991). This work has covered a wide band of the complexity spectrum but none that meets the strong tractability requirement discussed above. Most results stipulate polynomial time complexity, restrict inference in implausible ways (e.g., by excluding chaining of rules), and/or deal with limited expressiveness (e.g., deal only with propositions).

2 A tractable reasoning class

Below we describe a class of backward reasoning that is amenable to the use of massive parallelism and is tractable with reference to the criteria stated above. The characterization of the corresponding class of forward reasoning is analogous.

Some definitions:

Let us define *rules* to be first-order sentences of the form:

$$\forall x_1, ..., x_m \; [P_1(...) \wedge P_2(...) ... \wedge P_n(...) \Rightarrow \exists z_1, ...z_l \; Q(...)]$$

where the arguments of P_i's are elements of $\{x_1, x_2, ...x_m\}$, and an argument of Q is either an element of $\{x_1, x_2, ...x_m\}$, an element of $\{z_1, z_2, ...z_l\}$, or a constant. □

Any variable that occurs in multiple argument positions in the antecedent of a rule is a *pivotal* variable. □ (Note that the notion of a pivotal variable is local to a rule).

A rule is *balanced* if all pivotal variables occurring in the rule also appear in its consequent. □

For example, the rule $\forall x, y, z \; P(x, y) \wedge R(x, z) \Rightarrow S(y, z)$ is not balanced because although x is a pivotal variable (it occurs twice in the antecedent) it does not occur in the consequent. On the other hand, the rule $\forall x, y, z \; P(x, y) \wedge R(x, z) \Rightarrow S(x, z)$ is balanced because the pivotal variable x does occur in the consequent. The fact that y does not appear in the consequent is immaterial because y occurs only once in the antecedent.

Facts are partial or complete instantiations of predicates. Thus facts are atomic formulae of the form $P(t_1, t_2...t_k)$ where t_i's are either constants or distinct existentially quantified variables. □

Queries have the same form as facts. We distinguish between queries that are ground instances of predicates (i.e., all the arguments are bound to constants) and queries that are partially instantiated (i.e., queries with one or more existentially quantified variables). We will refer to the former as *yes-no* queries and the latter as *wh*-queries. □

Consider a query Q and a LTKB consisting of facts and balanced rules. A derivation of Q obtained by backward chaining is *threaded* if all pivotal variables occurring in the derivation get bound and their bindings can be traced back to the bindings specified in Q. □

[4] Ullman and van Gelder treat the number of nodes required to encode the background KB base as a fixed cost, and hence, do not refer to its size in computing the space complexity of their system. If the size of such a KB is taken into account, the number of processors required by their system turns out to be a high degree polynomial.

Given a LTKB consisting of facts and balanced rules, a *reflexive* query is one for which there exists a threaded proof. □

A class of tractable reasoning

The worst-case time for answering a reflexive *yes-no* query, Q is proportional to $|In|^{V+1}d$, where:

- $|In|$ is the number of *distinct* constants in Q.
- V is as follows: Let V_i be the arity of the predicate P_i. Then V equals $max(V_i)$, i ranging over all the predicates in the LTKB.
- d equals the depth of the shallowest derivation of Q given the LTKB.

Observe that the worst-case time is i) *independent* of $|LTKB|$ and ii) only polynomial in $|In|$. If we make the closed world assumption, we can show that a *no* answer can be produced in time proportional to $|In|^{V+1}D$, where D equals the diameter of the inferential dependency graph associated with the rule-base.[5] Answers to *wh*-queries are also computed in time proportional to $|In|^{V+1}D$, except that $|In|$ now equals the arity of the query predicate.

Lower bound nature of above result: It can be shown that derivations that involve unbalanced rules or those that do not satisfy the *threaded* property may not be computable in time independent of $|LTKB|$ if the available space is no more than $linear$ in $|LTKB|$. This result follows from the observations that i) the *common-integer* problem, i.e., the problem of determining whether two lists of integers share a common element, can be reduced to the problem of computing a derivation involving unbalanced rules and/or non-threaded derivations and ii) the *sorting* problem can be reduced to the common-integer problem (Rajasekaran 1992).
The *space* requirement is $linear$ in $|LTKB|$ and polynomial in $|In|$. This includes the cost of encoding the LTKB as well as the cost of maintaining the dynamic state of the 'working memory' during reasoning.

Note that while $|LTKB|$ may be about 10^7, $|In|$ is simply the number of (distinct) 'entities' referred to in the input. In the context of natural language understanding, $|In|$ would typically be quite small (less than 5). Furthermore we expect V, the maximum arity of predicates in the LTKB to be small.

Worst-case versus expected case

The above result offers a worst-case characterization which assumes that during the derivation, *all variables will get instantiated with all possible bindings involving constants in Q.* This will not be the case in a typical situation. In fact it may be conjectured that in a typical episode of reasoning, the actual time will seldom exceed $50d$ (see next section).

3 A massively parallel model of tractable reasoning

We have proposed a neurally plausible model (SHRUTI) that can encode a LTKB of the type described above, together with a term hierarchy and perform a class of forward as well as backward reasoning with extreme efficiency (Shastri & Ajjanagadde 1990; Ajjanagadde & Shastri 1991; Mani & Shastri 1991; Shastri 1992). SHRUTI can draw inferences in time that is only proportional to the *depth* of the shallowest derivation leading to the conclusion. A SHRUTI like model has also been used by Henderson (1992) to design a parser for English. The parser's speed is independent of the size of the lexicon and the grammar, and it offers a natural explanation for certain center embedding and long distance dependency phenomena.

If we set aside SHRUTI's ability to perform terminological reasoning, the class of reasoning that SHRUTI can perform efficiently is a subclass of the class of reasoning specified in the previous section. The additional restrictions placed on SHRUTI's reasoning ability are motivated by gross constraints on the speed at which humans can perform reflexive reasoning and gross neurophysiological parameters such as:

[5] The inferential dependency graph refers to the graph obtained by representing each predicate by a node and each rule by a hyperarc from the antecedent predicates to the consequent predicate.

i) π_{max}, the maximum period at which nodes can be expected to sustain synchronous activity, ii) ω, the tolerance or the minimum lead/lag that must be allowed between the spiking of two nodes that are firing in synchrony, and iii) the time it takes a cluster of synchronous nodes to drive a connected cluster of nodes to fire in synchrony.

The details of the model are beyond the scope of this paper and the reader is referred to (Shastri & Ajjanagadde 1990). Let us however, state the additional constraints on the class of reasoning SHRUTI can perform.

3.1 Additional constraints on the reasoning performed by SHRUTI

SHRUTI can encode a LTKB of facts and *balanced* rules and answer yes to any *reflexive yes-no* query in time proportional to the *depth* of the shallowest derivation leading to a derivation of the query provided:

1. The number of distinct constants specified in the query does not exceed k_1, where k_1 is bounded by π_{max}/ω) (biological data suggests that k_1 is small, perhaps between 5 and 10).

   The model suggests that as long as the number of entities introduced by the query is 5 or less, there will essentially be no cross-talk among the facts inferred during reasoning. If more than 5 entities occur, the window of synchrony would have to shrink appropriately in order to accommodate all the entities. As this window shrinks, the possibility of cross-talk between bindings would increase until eventually, the cross-talk would become excessive and disrupt the system's ability to perform systematic reasoning. The biological data suggests that a neurally plausible *upper bound* on the number of distinct entities that can occur in the reasoning process is about 10.

2. During the processing of the query, each predicate may only be instantiated at most k_2 times.

   Note that this restriction only applies to run-time or 'dynamic' instantiations of predicates and not to 'long-term' facts stored in the system. As argued in (Shastri 1992) a plausible values of k_2 is around 3. Also, k_2 need not be the same for all predicates. The application of a SHRUTI-like model to parsing by Henderson also suggests that k_2 equal to 2 may be sufficient for parsing English sentences.

Some typical retrieval and inference timings

If we set system parameters of SHRUTI to some neurally motivated values, SHRUTI demonstrates that a system made up of simple and slow neuron-like elements can perform a wide range of inferences (both forward, backward and those involving a type hierarchy) within a few hundred milliseconds.

If we choose the period of oscillation of nodes to be 20 msecs., assume that nodes can synchronize within two periods of oscillations and pick k_2 equal to 3, SHRUTI takes 320 msecs. to infer 'John is jealous of Tom' after being given the dynamic facts 'John loves Susan' and 'Susan loves Tom' (this involves the rule 'if x loves y and y loves z then x is jealous of z). The system takes 260 msecs. to infer 'John is a sibling of Jack' given 'Jack is a sibling of John' (this involves the rule 'if x is a sibling of y then y is a sibling of x). Similarly, the system takes 320 msecs. to infer 'Susan *owns* a car' after its internal state is initialized to represent 'Susan *bought* a Rolls-Royce' (using the rule 'if x buys y then x owns y' and the *IS-A* relation, 'Rolls-Royce is a car').

If SHRUTI's long-term memory contains the fact 'John bought a Rolls-Royce', SHRUTI takes 140 msecs., 420 msecs., and 740 msecs., respectively, to answer 'yes' to the queries 'Did John buy a Rolls-Royce', 'Does John own a car?' and 'Can John sell a car?' (the last query also makes use of the rule 'if x owns y then x can sell y). Note that the second and third queries also involve inferences using rules as well as *IS-A* relations.

The above times are independent of $|LTKB|$ and do not increase when additional rules, facts, and *IS-A* relationships are added. If anything, these times may decrease if a new rule is added that leads to a shorter inference path.

4 Conclusion

We have proposed a criteria for tractable reasoning that is appropriate in the context of common sense reasoning underlying language understanding. We have suggested that an appropriate measure of tractability for such reasoning is one where the time complexity is independent of, and the space complexity is no more than linear in, the size of the long-term knowledge base. We have also identified a class of reasoning that is tractable in this sense. This characterization of tractability can be further refined by cognitive and biological considerations. This work suggests that the expressiveness and the inferential ability of a representation and reasoning systems may be limited in unusual ways to arrive at extremely efficient yet fairly powerful knowledge representation and reasoning systems.

References

Ajjanagadde, V. G. & Shastri, L. (1991). Rules and variables in neural nets. *Neural Computation*, 3:121-134.

Brachman, R. & Levesque, H. (1984) The tractability of Subsumption in frame-based description languages. In *Proceedings of AAAI-84*, the fourth national conference on artificial intelligence. Morgan Kaufman.

Bylander, T., Allemang, D., Tanner, M. C. & Josephson, J. R. (1991) The computational complexity of abduction. *Artificial Intelligence*, 47(1-3), 25–60.

Carpenter, P. A. & Just, M. A. (1977) Reading Comprehension as Eyes See It. In: *Cognitive Processes in Comprehension.*

Dowling, W. F. & Gallier, J.H. (1984) Linear time algorithm for testing the satisfiability of propositional horn formula. *Journal of Logic Programming*, 3:267-284.

Fahlman, S. E. (1979) *NETL: A system for representing real-world knowledge*, MIT Press.

Frisch, A.M. & Allen, J.F. (1982) Knowledge retrieval as limited inference. In: *Notes in Computer Science: 6th Conference on Automated Deduction* ed. D. W. Loveland. Springer-Verlag.

Guha, R.V. & Lenat, D.B. (1990) Cyc: A Mid-Term report, *AI Magazine*, Volume 11, Number 3, 1990.

Henderson, J. (1992). A Connectionist Parser for Structure Unification Grammar, In the *Proceedings of ACL-92.*

Hölldobler, S. (1990) CHCL: A Connectionist Inference System for Horn Logic based on the Connection Method and Using Limited Resources. *TR-90-042*, International Computer Science Institute, Berkeley, CA.

Kautz, H. A. & Selman, B., (1991) Hard problems for Simple Default Logics. *Artificial Intelligence*, 47(1-3), 243-279.

Levesque, H. J. (1988) Logic and the complexity of reasoning. *Journal of Philosophical Logic*, 17, pp 335-389.

Mani, D. R. & Shastri, L. (1991) Combining a Connectionist Type Hierarchy with a Connectionist Rule-Based Reasoner, *Proceedings of the Thirteenth Conference of the Cognitive Science Society.* Lawrence Erlbaum.

McAllester, D.A. (1990) Automatic recognition of tractability in inference relations. Memo 1215, MIT Artificial Intelligence Laboratory, February 1990.

Patel-Schneider, P. (1985) A decidable first-order logic for knowledge representation. In *Proceedings of IJCAI-85*. Morgan Kaufman.

Rajasekaran, S. (1992) Personal communication.

Selman B., & Levesque H.J. (1989) The tractability of path-based inheritance. In the *Proceedings of IJCAI-89*. pp. 1140-1145. Morgan Kaufmann.

Shastri, L. (1992) Neurally motivated constraints on a working memory capacity of a production system for rapid parallel processing. To appear in the *Proceedings of the Fourteenth Conference of the Cognitive Science Society*. Lawrence Erlbaum.

Shastri, L. (1991) Why Semantic networks? In Principles of Semantic Networks. Edited by John Sowa. Morgan Kaufman Los Altos. 1991.

Shastri, L. (1988) *Semantic networks : An evidential formulation and its connectionist realization*, Pitman London/ Morgan Kaufman Los Altos. 1988.

Shastri, L. & Ajjanagadde, V. G. (1990). From Simple Associations to Systematic Reasoning: A connectionist representation of rules, variables and dynamic bidings using temporal synchrony. Technical Report MS-CIS-90-05, Department of Computer and Information Science, Univ. of Pennsylvania. (Revised January 1992). To appear in *Behavioral and Brain Sciences*.

Ullman, J. D. & Gelder, A. V. (1988) Parallel Complexity of Logical Query Programs. *Algorithmica*, 3, 5-42.