STATUSSEMINAR
des BMFT

KÜNSTLICHE INTELLIGENZ

27.- 28. April 1993
Berlin

Herausgegeben
durch den
Projektträger Informationstechnik des BMFT
bei der DLR
Gottfried Wolf

Deutsche
Forschungsanstalt
für Luft-
und Raumfahrt e.V.

DLR

Die in der vorliegenden Informationsschrift zusammengefaßten Beiträge stellen teilweise gekürzte Fassungen von Referaten dar, die vom 27. bis 28. April 1993 im Rahmen des vom Bundesministerium für Forschung und Technologie (BMFT) veranstalteten Statusseminar für das Forschungsfördergebiet Künstliche Intelligenz in Berlin vorgetragen wurden.
Organisation und Ausrichtung dieses Statusseminars wurden vom Projektträger Informationstechnik des BMFT besorgt.
Für den Inhalt der Beiträge zeichnen die vom BMFT geförderten Zuwendungsempfänger verantwortlich.

(Als Manuskript gedruckt)

Herausgegeben vom
Projektträger Informationstechnik des BMFT
bei der Deutschen Forschungsanstalt für Luft- und Raumfahrt e.V.
im Auftrag des BMFT, Referat 413

Redaktion: Dr.-Ing. Gottfried Wolf

Zu beziehen durch:
Projektträger Informationstechnik des BMFT
bei der DLR,
Abteilung Informationsverarbeitung
D - 12484 Berlin
Tel.: +49-30-69 54 57 41
Fax.: +49-30-69 54 57 42

Inhaltsverzeichnis

I. Eingeladene Vorträge

II. Projekte des Fördergebietes Anwendung der Wissensverarbeitung

Toward a Synthesis of Symbolic AI and Connectionism

Stephen Omohundro
The International Computer Science Institute
1947 Center St, Suite 600
Berkeley, CA 94704
Email: om@icsi.berkeley.edu

July 20, 1993

1 Introduction

The successful development of artificial intelligence would have a tremendous technological and intellectual impact. The AI research effort over the past 30 years has had some important successes but has also shown many tasks to be more complex than first thought. Mainstream AI research through the 1970's was dominated by a symbolic approach. In the 1980's, connectionist approaches based on artificial neural network models gained in popularity. Both approaches have significant strengths and weaknesses. Because the areas of weakness are fairly complementary, the time is ripe to forge a new synthesis combining the positive aspects of symbolic AI and connectionism. This paper proposes a set of characteristics for such a combined approach and describes some of our work in that direction.

1.1 Strengths of Symbolic AI

Coherent semantics. Much of symbolic AI uses formal logic as a foundation for inference. This base gives the representation a coherent underlying semantics. One can treat inference as a variant of theorem proving. To the extent that the axioms and rules of inference correctly reflect the domain being modelled, the reasoning process must give coherent results.

Dynamic structure. The use of predicate logic allows symbol-based systems to represent dynamically structured situations. Knowledge bases typically contain *relational* information which can be combined in complex ways, not foreseen by the knowledge engineer.

Understandable representation. The symbolic representation of knowledge is close to the linguistic form used by people for communicating and manipulating knowledge. This allows symbolic systems to explain conclusions in a way that people find understandable. It also makes it possible for knowledge engineers to introspect and construct rules for a domain by hand.

1.2 Weaknesses of Symbolic AI

Brittleness. Because symbolic rules tend to work in an "all or none" fashion, the performance of a system can drop off dramatically at boundaries of its competence. This kind of "brittleness" has been seen repeatedly in symbolic expert systems.

Poor uncertainty representation. Because symbolic systems are fundamentally logic based, there isn't a natural semantically coherent way of representing uncertainty. Attempts to syntactically add uncertainty to these representations tend to give rise to incorrect assessments in only a few steps of an inference chain.

Poor quantitative representation. Symbolic systems are not natural for representing the quantitative data that arises in sensory domains. For example, symbolic systems tend to be quite stilted for visual representation or geometric reasoning.

Weak learning. While there have been advances in symbolic machine learning, it is still rather weak compared with quantitative approaches. Because symbolic systems don't represent uncertainty well, it is difficult for the system to determine what to modify for improvement during learning.

High development cost. Because symbolic knowledge bases must usually be entirely constructed by hand, building a system for a new domain is an expensive endeavor requiring many man-years.

1.3 Strengths of Connectionism

Naturally evidential. Connectionist representations have been fundamentally evidential from the start. A state of knowledge is represented by the analog activation state of neuron-like units. Typically, a greater strength of activation represents greater confidence in the presence of a particular feature.

Naturally quantitative. Because they are fundamentally quantitative, connectionist representations are well-suited to geometric and quantitative domains. Analog information from auditory or visual input is naturally represented and manipulated. Synthetic connectionist systems have been built which well-model the corresponding biological systems.

Quantitative learning. The adaptive weights in connectionist representations combined with the recent development of many powerful learning rules give rise to a powerful quantitative learning capability. It is now commonplace to use neural networks to learn simple quantitative and geometric relationships.

Naturally parallel. The computation of unit activation in connectionist systems is naturally parallel. This type of representation therefore appears to be natural for the next generation of computer hardware.

1.4 Weaknesses of Connectionism

Fixed structure. Unlike symbolic representations, most connectionist systems require the pattern of connectivity between units to be fixed in advance. This makes it difficult to bind symbols to variables and to represent relational information in connectionist systems.

Opaque. The connectionist representation of knowledge is often opaque to humans, especially in unstructured neural networks. Numerical learning procedures can spread the representation of a concept over a whole network. This makes it difficult for a person to follow the "reasoning" of the network and also makes transfer of knowledge to other systems problematic.

Poor semantics. Most connectionist systems do not have a coherent underlying semantic base. While the representations are evidential, it is often difficult to say exactly what a particular activation state means about the world.

Poor structural learning. While most connectionist systems support powerful quantitative learning, they usually don't support the learning of new relationships.

1.5 Toward a Synthesis

The time appears ripe to form a synthesis exhibiting the best characteristics of symbolic and connectionist systems while avoiding the weaknesses. Work in being done in both the symbolic and connectionist domains towards this goal.

Structured connectionism is an approach which aims to put more structure on connectionist representations. This structure makes the representations more understandable to people, provides a more coherent semantics, and allows the incorporation of more prior knowledge. *Probabilistic networks* are a similar kind of representation which starts from a probabilistic representation and explicitly encodes conditional independence relationships. Both of these representations have a static underlying graph structure, however.

There have also been several attempts to add evidential components to symbolic representations. Most of these attempt to preserve the syntactic computational structure that is characteristic of logic, however, and so cannot support the coherent semantics of probability theory.

We do not present a complete synthesis here, but do describe some of our research which aims in that direction. We believe that such a representation must:

- have a coherent underlying *probabilistic* semantics.

- be capable of representing and manipulating quantitative *geometric* and physical information.
- support powerful *learning* and generalization of both quantitative and structural relationships.
- support the representation and manipulation of dynamically structured *relational* representations.
- be computationally *efficient*.

2 Learning and Recognition

Two lessons that most researchers would agree have been learned in the pursuit of artificial intelligence are that: 1) Large amounts of knowledge are necessary to perform well in even slightly complex domains and 2) Symbols must be grounded in the real world of physical objects to avoid physically incoherent conclusions.

The most efficient way to develop large knowledge bases is through the extensive use of learning. Not only is learning more cost effective, it provides better guarantees of domain coverage than hand-built systems because the system itself can seek to improve weak areas. The most efficient way to ground symbols in the physical world is through perception and interaction.

The two tasks of *learning* and *perception* are therefore central to the development of effective AI. Both of these are fundamentally inductive model-building tasks. The system must generalize on the basis of noisy and impoverished data. In the case of recognition, the system must construct a perceptual model to account for sense data. In the case of learning, the system must construct a knowledge base to account for past experiences. In many ways, learning is a longer time version of recognition. Learning provides the model construction materials for recognition.

2.1 Simple Neural Net Learning

Many currently popular learning algorithms have certain properties which make them ill-suited to complex AI tasks. We will use back-propagation neural networks as an example but the criticisms are applicable to many other approaches as well. The characteristics of these models include:

- A fixed small space of possible models is chosen before any learning takes place. Eg., an experimenter might choose a particular neural network architecture to learn a mapping.
- A simple parameterization of the model space is chosen. Eg., back-propagation networks are usually parameterized by their synaptic weight values.

- The system starts at randomly chosen parameter value. Eg., the network weights are chosen randomly.
- The parameters are gradually modified to improve performance on the training data. Eg., the use of back-propagation for stochastic error gradient descent.
- The system might possibly include terms to help prevent overfitting. Eg., weight decay or cross-validation are often used in conjunction with back-propagation networks to stop training early or to modify its course.

There are a number of problems with this type of algorithm both in their absolute learning performance and as models for animal cognition.

Because the representation space starts out as large as it will ever be, the system always has a fully formed model of the world. This model is a poor one at first but gradually improves. The system always thinks it has a complete model, though. It doesn't have a measure of what is reliable knowledge based on experience and what is due to the random initial conditions. In this sense, the system doesn't know what it knows.

Such systems are usually incapable of one-shot learning. Single examples rarely have a big impact and many examples of a give type must be seen to pull the model parameters into a region which explains them. Tuning one part of the model space will often interfere with performance in another.

Many of the problems arise from the initial choice of model space. If it has too small a representational power then it may not be rich enough to represent the true model. If the space is too large, then the system is subject to *overfitting*. This means that specific examples are inappropriately generalized. Large numbers of examples are required to reliably fit the parameters of large model spaces. These systems also typically cannot allocate their resources. The model space may be large enough, but it can't put the representational power where it is needed.

These systems are often very slow to learn because they gradually vary the parameters. They are liable to get stuck in local minima because all parts of a complex space are simultaneously manipulated and different portions can interfere with one another. They are computationally expensive because the full model is evaluated for each example. Most of these systems do not have a coherent underlying semantics. Such a semantics is essential, for example, to compositionally build up complex models from simple ones.

3 Human and Animal Cognition

What is known of human and animal cognition appears to have a very different character from the models described in the last section. Animals are typically capable of one-shot learning. Individual experiences can have a big impact on

the future behavior of an animal. One experience with a noxious substance will keep rats from ever sampling food with a similar smell again. Humans have episodic memory for specific events, especially when first learning about a domain. Each early experience in a domain is critical for discovering regularities and it is worthwhile for the animal to expend the resources to remember it.

Early in learning, generalization appears to proceed by similarity. In [13], Shepard has studied generalization from a single experience in a wide variety of domains. There appears to be a universal law of generalization in which the effect of an experience at one sensory parameter value is generalized to other parameter values according to a decaying exponential, ie. it is based on the distance in the sensory space.

As experience accumulates, animals build more complex models as they are warranted. These models have an adaptive structure and may be complex in one part of the domain while they are simple in other parts. For the most part, animal and human learning seems to avoid overfitting and getting stuck in local minima. There are a variety of elaborate focus of attention mechanisms which allow an animal to only access relevant information.

People generally know what they know. They have a sense of how much confidence they should place in their experience in a given domain. Human cognition is adaptable to a wide variety of domains. Finally, in many natural domains human cognition is well-modelled by a coherent underlying Bayesian semantics [1]. (There is also an extensive literature documenting the failure of human probabilistic reasoning outside of the natural domains in which it evolved [5].) We will describe Bayesian induction in the next section and then present artificial induction algorithms which have many of these characteristics of biological systems.

4 The Bayesian Approach

Over the past several centuries, probability theory has evolved into a powerful mathematical model for representing and manipulating uncertainty. It is used both in "objective" situations in which uncertainty arises from physical noise processes and in "subjective" situations in which uncertainty is due to ignorance (there has been much debate over whether these are in fact distinct). Probability theory may be applied to uncertain reasoning in both of these circumstances and the resulting inference procedure has come to be known as *Bayesian decision theory* [2].

Consider a rational agent (such as a robot or biological organism) which must make decisions on the basis of its current perceptions, its past experience, and any built-in knowledge. To compare different possible actions, the agent must have a set of built-in preferences for certain states of the world over certain others. This is usually encoded in a *utility function* $U(a, s)$ which encodes the "goodness" of taking the action a when the true state of the world is s. The

agent would like to take that action which maximizes this utility. The problem is that the agent doesn't have direct access to the state of the world s. It only has access to perceptual data d which is noisily related to the state of the world by a *probability distribution* $p(d|s)$. When this probability is viewed as a function of s with d held fixed, it is called the *likelihood* $l(s|d)$.

First, consider the objective probability case. Assume that the state of the world is chosen from a distribution $\pi(s)$ which is called the *prior* distribution over the state of the world. For a given world state s, the agent is presented with sensory data d drawn from the distribution $p(d|s)$. How should the agent act so as to maximize his expected utility? Probability theory provides a precise and objective answer to this question using Bayes' theorem. The agent should first compute the *posterior probability* of each possible state of the world conditioned on the data:

$$p(s|d) = \frac{\pi(s)p(d|s)}{\int_S \pi(s)p(d|s)ds}$$

This encodes the agent's estimate of the probability of each world state. The agent should choose that action which maximizes the expected utility when averaged with respect to this posterior distribution:

$$a = argmax_a \int_s U(a,s)p(s|d)ds$$

Since the normalization factor is a constant, the agent may equivalently do:

$$a = argmax_a \int_s U(a,s)\pi(s)l(s|d)ds$$

ie. the agent should take that action which maximizes the utility times the prior times the likelihood integrated over the possible world states.

The subjective version doesn't make assumptions about the world but instead focusses on the agent. Under a set of very reasonable assumptions (eg. that the agent's preference relationship is transitive), there is a powerful theorem [12] which says that a rational agent must behave as if it were performing the Bayesian analysis described above with respect to *some* choice of prior distribution and utility function. If it does not, then it is subject to exploitation by other agents. For example, it will accept certain "*Dutch bets*" in which it loses regardless of the state of the world. In a competitive evolutionary setting, agents which don't behave in a Bayesian way will be outperformed by those that do.

In these two senses, the Bayesian procedure is a goal toward which rational agents should strive. It should be possible to construct artificial agents which outperform humans in many situations by approximating the Bayesian choice of action more accurately.

What are the impediments to directly implementing the Bayesian approach? The two primary problems are the choice of prior and the evaluation of the integral. We discuss these aspects in the next two sections.

5 Natural Priors for Induction

The prior encodes any specific knowledge about the domain. Asymptotically, the prior is unimportant as long as it doesn't vanish on the true model. This is because, under very general conditions [2], the posterior distribution peaks around the true model. The shape of the posterior is asymptotically Gaussian and the width decreases as the amount of data increases. The effect of the data eventually overwhelms the effect of the prior. Ultimately, the Bayesian approach will choose the correct model regardless of the choice of prior.

In practice, the prior can have a tremendous effect on the rate of learning. In recognition, the amount of data is fixed by the properties of the sense device. In learning, the amount of data is limited by the amount of experience the system has had.

We have discovered in a variety of settings that there are a small number of aspects of the physical world that can have a major impact on learning performance when incorporated into a prior. Bertrand Russell identified similar features in his work on the origin of human knowledge [11].

The most fundamental priors describe aspects of time and space. The most essential is the "time" prior. This says that the future is likely to be like the past. Without this, there is no reason that data collected in the past should be applicable to future situations. This leads to inductive procedures that combine data obtained on different trials.

The "continuity" prior prefers solutions in which model parameters vary continuously as perceptual data varies. It reflects the "geometric" nature of space. It leads to generalization by similarity and to "interpolation" procedures for filling in missing data.

The two priors with the most dramatic impact on learning performance are the "sparseness" and "locality" priors [10]. These prefer models in which the model space and the data space naturally break into distinct components.

The "sparseness" prior says that sparsely interacting model components are preferred to those which are highly interconnected. This prefers probability distributions with a large amount of "conditional independence". This kind of distribution is naturally represented in terms of an underlying independence graph. They are being actively studied in the form of "Bayesian networks" and "Markov networks". The prior that a Bayes' net be sparse is so strong, that nets representing 37 random variables interconnected by 46 arcs can be accurately learned with only 10,000 samples [4]. This prior suggests only introducing interactions between model components when there is enough data to validate them.

The "locality" prior is similar but relates to the relationship between model components and data components. The preference is for models such that each model component affects only a small number of data components. For example, each pixel in a visual scene is affected by only a few objects in the scene, each word in a sentence is affected by only a few grammar rules, each experience is affected by only a few components of a knowledge base, etc. This prior suggests using the data components themselves as the initial model components.

There are many other candidates for fundamental priors (eg. priors representing the notion of "natural kinds"). We have found, however, that these are sufficient to perform powerful learning and recognition in a variety of important and complex domains.

6 Avoiding Overfitting

The central problem of induction is appropriate generalization. The problem of "overfitting" arises when a model is inappropriately "over-tuned" for specific data and so fails to generalize well. The full Bayesian procedure described above produces the correct answer according to probability theory and is not subject to overfitting. Unfortunately, it is usually computationally impossible to perform the integral over all possible models. A variety of approximations to the integral are used and they give rise to overfitting.

The simplest approximation is the so-called "MAP" approach which chooses the single model whose posterior probability is highest. Typically, more complex models will fit the data better and so will have a higher posterior. When the integral is performed, however, the contribution of the neighborhood of the posterior peak introduces a so-called "Occam factor" which gives preference to simpler models.

Much of the research in learning theory has been devoted to techniques for avoiding overfitting and detecting when it occurs. There are a variety of measures of the "capacity" of a space of models. The central results of learning theory describe how the number of samples needed for reliable model induction increases as the capacity of the model space increases.

For a given amount of data, small model spaces won't overfit but are unlikely to contain the true model. Larger model spaces may include the true model, but are subject to overfitting. Vapnik [15] proposes an approach to learning that begins with a *nested* family of model spaces. As data is received, the smallest model space is tested. If a good-fitting model is found, it may be reliably validated with a small amount of data. If no such model is found, the next larger space is considered. In this way the amount of data required is determined by the complexity of only the model space which actually contains the true model. This allows one to induce models which are potentially arbitrarily complex with only a finite amount of data.

In [10] we extended this approach into what we call "model merging". This

combines this kind of incremental learning procedure with the features of the priors described above. It has proven to be a powerful approach to learning and recognition in a wide variety of domains including: classification, density estimation, mapping learning, surface learning, Hidden Markov model induction, as well as general stochastic grammar induction.

The idea is to always represent a model as a collection of sparsely interacting component models. When only a small amount of data is present, the data elements themselves are the model components. Generalization at this point is by similarity (eg. nearest neighbor classification or kernel-based density estimators).

As more data appears, more complex model components are formed by merging together existing model components. When two components are merged, the amount of data available for fitting the component increases. This allows the complexity class of the model components to increase as the amount of data increases. A model component is never chosen from a model class which is too complex to fit the data without overfitting. In this way the whole model can produce reliable generalization and yet can adapt to the data in arbitrarily complex ways.

This procedure is capable of inducing representations which satisfy each of the criteria listed above for the synthesis of symbolic and connectionist approaches. It is fundamentally *probabilistic* and serves as an approximation to the full Bayesian model. It can naturally represent both *symbolic* and *geometric* information. It can learn and generalize both *structural* relationships and *quantitative* ones. The structure of the representation is *dynamic* and determined by the data, adding higher representational complexity where it is needed. It can be made computationally efficient through the use of appropriate data structures.

We will briefly describe two application areas to which we have applied this approach: the induction of surfaces and grammars. Surface learning demonstrates the geometric and quantitative aspects while grammar learning demonstrates the dynamic structural capabilities.

7 Surface Learning

Learning in geometric domains has been a fundamental task to which many connectionist systems have been applied. The most common geometric induction problems are density estimation, classification, and mapping learning. We have applied the ideas described here, along with algorithmic techniques from computational geometry to these problems (eg. [6], [7], [8], [9]).

More recently we have used these techniques to learn and represent nonlinear constraint surfaces [3]. This task is more complex than the other geometric tasks because the system must determine the dimension of the constraint. The initial local surface models are linear patches determined by a local principle components analysis. The local patches are smoothly "glued" together using

a partition of unity. The "bumptree" data structure may be used to achieve extremely fast access supporting nearest point, partial information completion, surface interpolation, and other queries.

Neighboring patches can be merged to form larger patches. After several merges, there is enough data to reliably fit more complex patch models (eg. quadratic). The system adapts the representation to fit the data.

This approach does an excellent job with a small amount of data on a wide variety of synthetic surfaces such as spheres and cylinders. We have also applied the technique to inducing and modeling the "space of lips" for a visual lipreading task. Initial results produce a five dimensional surface which significantly improves the tracking performance and produces relevant features that improve recognition performance over a system based on auditory information alone.

8 Stochastic Grammar Learning

A non-geometric task to which we have applied the general approach described here is stochastic grammar learning [14]. This is especially interesting because parse trees and grammars have dynamic structures which depend on the data. We have had excellent results inducing the topology of Hidden Markov Models (HMM's) from data and are currently studying stochastic context-free grammars.

The starting point for our HMM induction approach is to use the sample strings themselves as the data. The initial HMM has a separate internal state for each symbol of each sample string. The paths through the model are in one-to-one correspondence with the the sample strings. This is also the maximum-likelihood model. The algorithm proceeds by merging pairs of states together. When two states are merged their transition and emission probability distributions are replaced by a weighted mixture of the originals. A Dirichlet [2] prior on these distributions provides a Bayesian criterion for stopping the merging.

The merged models do an excellent job of inducing the topology of the underlying model. The performance is much better than the standard Baum-Welsh approach based on the "EM" algorithm. This is especially true when there is only a small amount of data and the effects of overfitting are especially important.

9 Conclusions

We have described some of the strengths and weaknesses of symbolic AI and connectionist approaches. We have proposed criteria that an approach which synthesizes the best of both should satisfy. We have described an approach to learning, recognition, and evidential knowledge representation that we have used

in a variety of domains. We described the use of this approach in a geometric domain and a symbolic domain. Our current work involves applying it to high-level vision tasks which incorporate both of these aspects.

References

[1] Anderson, J. R., *The Adaptive Character of Thought*, Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey, (1990).

[2] Berger, J. O., *Statistical Decision Theory and Bayesian Analysis, Second Edition*, Springer-Verlag, New York, 1985.

[3] Bregler, C. and Omohundro, S., "Surface Learning with Applications to Lip Reading", submitted to *Advances in Neural Information Processing Systems 6* (1993).

[4] Cooper, G. and Herskovits E., "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, Volume 9, Number 4, October (1992).

[5] Kahneman, D., Slovic P., and Tversky Al, editors *Judgement under uncertainty: Heuristics and biases*, Cambridge University Press (1982).

[6] Omohundro, S. M., "Efficient Algorithms with Neural Network Behavior," *Complex Systems*, 1 (1987) 273-347.

[7] Omohundro, S. M., "Geometric Learning Algorithms", *Physica D*, 42:307-321 (1990).

[8] Omohundro, S. M., "Bumptrees for Efficient Function, Constraint, and Classification Learning", in Lippmann, Moody, and Touretzky, (eds.) *Advances in Neural Information Processing Systems 3*, San Mateo, CA: Morgan Kaufmann Publishers, 1991.

[9] Omohundro, S. M., "Building Faster Connectionist Systems with Bumptrees" in W. Brauer and D. Hernandez, (eds.) *Verteilte Kunstliche Intelligenz und kooperatives Arbeiten, the Proceedings of the Fourth International GI-Congress*, Berlin: Springer-Verlag, 459-466, (June 1991).

[10] Omohundro, S. M., "Best-First Model Merging for Dynamic Learning and Recognition", in Moody, J. E., Hanson, S. J., and Lippmann, R. P., (eds.) *Advances in Neural Information Processing Systems 4* pp. 958-965, San Mateo, CA: Morgan Kaufmann Publishers, (1992).

[11] Russell, Bertrand, *Human Knowledge, Its Scope and Limits*, Simon and Schuster, Brooklyn, New York, (1948).

REFERENCES

[12] Savage, L. J. et. al. *The Foundations of Statistical Inference.* Methuen, London (1962).

[13] Shepard, R. N., "Toward a Universal Law of Generalization for Psychological Science", *Science* (1987).

[14] Stolcke, A. and Omohundro, S. M., "Hidden Markov Model Induction by Bayesian Model Merging", *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann Publishers, San Mateo, CA. (1993).

[15] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, New York: Springer-Verlag (1982).