Chapter 5: Related Work

Work in selective attention can draw from two major fields for inspiration and guidance. The first is work in machine learning, primarily that concerned with causal model builders and active agents, and secondarily passive learners and the reinforcement learning literature in general.

The second is the study of attentional processes in the cognitive science literature. There has been considerable work on attention in cognitive science, and the questions asked and insights gained into attentional processes, while themselves often insufficiently well-specified to serve as computational theories, may serve as inspiration for approaches to machine implementation.

5.1 Related Work in Machine Learning

5.1.1 Introduction

As illustrated earlier, a typical agent in the world cannot perceive every part of the world at once, nor should it—even “perceiving” without “learning” is expensive if the agent must perceive everything. However, not perceiving the whole world at once can lead to a phenomenon that [Whitehead and Ballard 90] calls *perceptual aliasing*, in which different world states can appear identical to the agent, and which causes most reinforcement learning mechanisms to perform poorly or not at all. Both they and [Woodfill and Zabih 90] propose systems which combine selective visual attention (which is used to "ignore" certain parts of the world at certain times) with special algorithms to attempt to overcome the aliasing problem.

Many of the methods described in earlier chapters may be available to other machine learning systems, and several such extensions are discussed in later sections. While Drescher's schema system keeps exhaustive statistics and is thus easy to adapt in the manner shown, *any* agent that tries to correlate its actions with results must keep around *some* sort of statistics regarding those results from which to learn, even if they are only available for the instant of perception, and those stored statistics are candidates for pruning. Further, any such agent must somehow perceive the world, and its sensory inputs are likewise candidates for pruning.

For example, the techniques used in the most-focused of the goal-independent strategies shown in Chapter 3 (bottom line of Figure 8, on page 63) are likely to be available to most learning systems operating in a discrete microworld. They require being able to keep track of which sensory items have changed recently, and which facts depend upon (e.g., make predictions concerning) those items. This does not seem an insurmountable obstacle for many algorithms. It is even possible that particular algorithms which do not possess absolute knowledge about, for example, which sensory items are mentioned in any given learned fact (such as the hidden nodes of a neural net) might nonetheless be able to yield a probabilistic estimate of how likely it is that some particular part of the internal knowledge base might depend on a particular sensory input. If so, such algorithms might also allow cognitive pruning to take place.

Selective attention and goal-directed learning have recently been getting considerably more attention in the literature than previously. Several researchers have advanced frameworks or architectures for thinking about and taxonomizing such systems, usually based either upon a model of *filtering* or *discarding* information deemed unnecessary or harmful for the learner [Markovitch and Scott 93], or upon explicitly modelling goals, using goals both to inform the learning process and as input to metareasoning strategies which reason about the performance of the learner [Ram and Leake 94].

The next few sections will discuss some of these issues. Section 5.1.2, examines the selective attention side of the issue, and Section 5.1.3, on page 117, addresses goal-directed learning aspects.

5.1.2 Selective attention as filtering

[Markovitch and Scott 93] propose an information filtering framework for evaluating and specifying selection mechanisms in machine learning systems. As shown in Figure 17, they propose five different places in which filtering may be employed as a selectional mechanism:

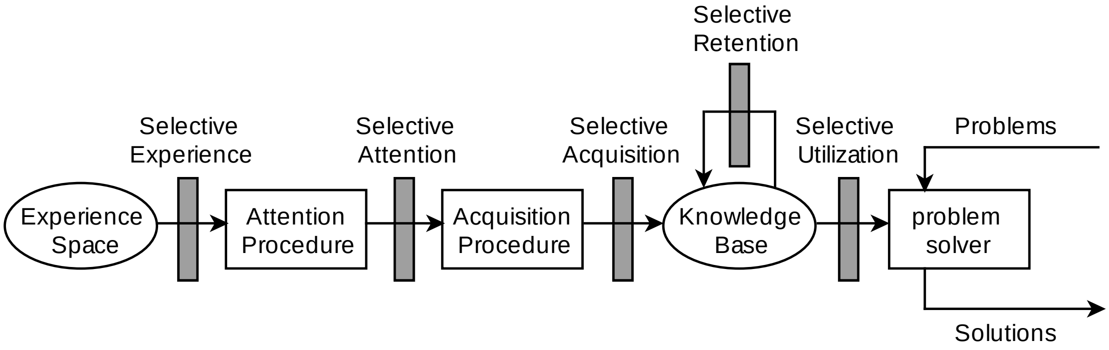

Figure 17: Markovitch and Scott’s information-flow filtering model

- *Selective experience* reduces the acquisition of knowledge when the number of possible training experiences is large.
- *Selective attention* reduces the acquisition of knowledge when individual training experiences are complex.
- *Selective acquisition* reduces knowledge that has been acquired from some training example(s) from reaching the permanent knowledge base, but can be somewhat limited by not knowing how the knowledge might be used.

- *Selective retention* allows “forgetting” knowledge that will not be worth its storage cost (or is actively harmful), for any later problem.
- *Selective utilization* allows “ignoring” parts of the knowledge base that will not be useful for solving the current problem.

Using this framework, they summarized several well-known machine learning systems. Figure 18, on page 115, is from [Markovitch and Scott 93], amended to include the current research. According to their framework, the work described in previous chapters employs *selective attention* (in the form of how sensory information is accepted, e.g., sensory pruning) and *selective utilization* (in the form of which schemas are updated or spun off from, e.g., cognitive pruning). Selective utilization, as is used here, is not often used in current machine learning systems; IB4 [Aha and Kibler 91] and EGGS [Mooney 89], both relatively recent systems, make use of it, but few others.[1]

[Kaelbling and Chapman 90] propose a technique (the G algorithm) for using statistical measures to recursively subdivide the world known by an agent into finer and finer pieces, as needed, making particular types of otherwise intractable unsupervised learning algorithms more tractable. One could view that as an example of perceptual selectivity: the agent gradually increases the set of state variables that are considered, as needed, when selecting actions and learning (updating statistics).

Along these lines, [Moore and Atkeson 93] propose a technique called *prioritized sweeping*. This approach concentrates learning effort in those regions of the world that are likely to be least well-understood, creating a tree of which questions should be answered and in what order, and shows promise in substantially decreasing the computational com-

[1] In IB4, which functions similarly to memory-based reasoning [Stanfill and Waltz 86], instances which perform poorly are simply discarded from the database (selective retention), and newly-acquired instances are prevented from contributing to decisions until sufficient evidence has accumulated to demonstrate that they are reliable (selective utilization). In EGGS, an explanation-based learner, the system only uses learned rules that completely solve a problem, and no use is made of learned knowledge to prove subgoals.

System	Filter	Description	Evaluation Metric
Checker Player [Samuel 59]	Retention	Discards least useful board position	Frequency of use
Genetic algorithms [Holland 86]	Retention	Randomly retains elements with probability proportional to their fitness	Fitness defined in a domain-specific manner
MetaDENDRAL [Buchanan and Feigenbaum 82]	Acquisition	Attempts to find smallest set of rules that accounts for data	Rules that correctly predict peaks not predicted by other rules score higher
Version Space [Mitchell 82]	Experience	Chooses experience that will reduce version space by greatest amount	Selects experience that comes closest to matching half remaining hypotheses
ID3 [Quinlan 86]	Experience	Selects only misclassified instances	Correctness of classification
INDUCE [Dietterich and Michalski 81]	Acquisition	Eliminates candidate generalizations using several evaluation criteria	Includes coverage, specificity, and user defined function
LEX [Mitchell, Utgoff, and Banerji 83]	Experience	The Problem Generator constructs new practice problems	Prefer problems that will refine partially learned heuristics
	Attention	The Critic marks positive and negative instances in the search area	Select search steps on the lowest cost solutions as positive
MetaLEX [Keller 87]	Retention	Removes subexpressions that are estimated to be harmful	A weighted combination of estimated cost and estimated benefit
DIDO [Scott and Markovitch 89]	Experience	Performs experiments on classes with high uncertainty	Prefer experiences involving objects of classes with higher uncertainties
PRODIGY [Minton 88]	Experience	Generates experiments when discovers incomplete domain knowledge	Incompleteness
	Attention	The OBSERVER selects training examples out of the trace tree	Training example selection heuristics eliminate “uninteresting” examples
	Acquisition	Estimates utility of newly acquired control rules and deletes those unlikely to be useful	Eliminate rules whose cost would outweigh saving, even if always applicable
	Retention	Empirical utility validation by keeping the running total of the costs and frequency of application	Estimated accumulated savings minus accumulated match cost; if negative, discard rule

Figure 18: Selection mechanisms in some existing learning systems (Sheet 1 of 2)

System	Filter	Description	Evaluation Metric
MACLEARN [Iba 89]	Attention	The macro proposer uses peak-to-peak heuristics	Propose only macros that are between two peaks of the heuristic function
	Acquisition	Static filtering; only macros estimated to be useful are acquired	Redundancy test (primitive) and limit on length and domain specific test
	Retention	Dynamic filtering; invoked manually	Frequency of use in solution
FUNES [Markovitch and Scott 88]	Retention	Various heuristics to decide what macros to delete	Random, Frequency of use x Length
CLASSIT-2 [Gennari 89]	Attention	Attributes with low salience are ignored	Salience
Hypothesis Filtering [Etzioni 88]	Retention	Runs a test on sample population; passes only hypotheses which are PAC	For a given ε and δ, compute an upper bound on the distance between the hypothesis and the target concept
IB4 [Aha and Kibler 91]	Acquistion	Acquires misclassified instances	Correctness of classification
	Retention	Removes instances that appear to be noisy	Confidence interval of proportions test
	Utilization	Only instances that have proved reliable are used for classification	Confidence interval of proportions test
EGGS [Mooney 89]	Utilization	Learned macros are used only if they solve the problem	Macros that do not solve the problem are worth nothing
This research [Foner 94]	Attention	Sensory bits not relevant to typical world behavior or current goals are not perceived	Spatial and temporal proximity (goal-independent); relevance to current goal (goal-dependent)
	Utilization	Schemas which do not make predictions concerning currently useful sensory bits, actions, or goals, are not updated or spun off	Spatial and temporal proximity (goal-independent); relevance to current goal (goal-dependent)

Figure 18: Selection mechanisms in some existing learning systems (Sheet 2 of 2)

plexity of many common learning situations. They demonstrate it in system that learns a world model in a world of stochastic Markov chains; it is somewhat similar to DYNA [Sutton 90] .

As another way to look at the problem, consider classifier systems, such as in [Holland 86]. Classifier systems have a built-in mechanism for generalizing over situations as well as actions and thereby perform some form of selective attention. In particular, a classifier may include multiple “don’t care” symbols which will match several specific sensor data vectors and actions. This makes it possible for classifier systems to sample parts of the state space at different levels of abstraction and as such to find the most abstract representation (or the set of items which are relevant) of a classifier that is useful for a particular problem the agent has. [Wilson 85] argues that the classifier system does indeed tend to evolve more general classifiers which “neglect” whatever inputs are irrelevant.

Others have also addressed the problem of finding the proper tradeoff between *efficiency* (the number of measurements a robot must take, for example) and *accuracy* (number of prediction errors) when attempting to build a world model. For example, [Tan 93] proposes a unified framework for learning from examples, based on four distinct spaces in concept learning: example space, feature space, concept space, and concept description space. This framework is applied to analyzing CS-ID3 and CS-IBL in detail, which are *learning-cost-sensitive* (hence “CS-”) algorithms which can trade off accuracy for efficiency in decision-tree-based (CS-ID3) or instance-based (CS-IBL) learning.

5.1.3 Selective attention as goal-driven learning

There has been considerable work in goal-driven learning in recent years. For example, Ram, Leake, Cox, and Hunter have between them produced on the order of thirty papers quite recently which all bear in some way or another on this topic. Some of them are discussed below; other relevant papers include [Ram 90a], [Ram and Cox 90], [Cox and Ram 94], [Ram 90b], and [Cox and Ram 92].

For example, [Leake and Ram 93] describe aspects of goal-directed learning from the perspectives of AI, psychology, and education in a survey paper that reported on a workshop involving participants from all three areas. They summarized four taxonomies of learning goals: by overarching tasks, by knowledge gap or failure-necessitated learning, by the learning results, and by the learning activity. They also talk briefly about how goals for learning arise, how they affect the learning process, how different types of learning goals relate to each other, and how they are represented.

[Ram and Leake 94] propose a general framework for describing goal-driven learning systems. This survey paper discusses how goals guide task performance, task learning, and knowledge storage, with a strong emphasis on using plans to manage the growth of complexity in all these areas. They propose a two-step framework for managing the learning, in which the first step attempts to reach some particular goal, maintaining a trace of the reasoning performed. Plan failures or deficiencies during this reasoning are then used in the second step, which uses credit/blame assignment to find the source of the failure. Thus, learning in the second step is guided by a knowledge of *what* must be learned and *why*, stemming from the information resulting from analysis of plan failures. Such explicit reasoning about goals is also discussed in [Ram and Hunter 92].

In addition, they point out the importance of *multistrategy learning*. Large, complicated learning systems that operate on real-world problems are increasingly being implemented as multistrategy learners. Such a technique allows using the appropriate learning strategy for the particular piece of the problem which is currently of interest, but are often hard to control or program without some automated way of determining when to use particular algorithms. Learning systems that can reason about their goals and use this information to select particular learning modules can make multistrategy learning more feasible. In addition, they increase the feasibility of systems that can actively seek out additional sources of information, rather than having to be spoon-fed information from a small number of hand-picked sources. For example, reasoning involving multistrategy learning is employed by [Ram, Narayanan, and Cox 93] in a system that learns to troubleshoot and is based on observations and a model of human operators engaged in a real-world troubleshooting task.

In recent work, [Hunter 94] provides two examples of these strategies in action. Both are drawn from the domain of biology, which is becoming increasingly important as a source of real-world applications of machine learning due to the large number of interesting datasets, the possibility of external verification and grounding of results via physical experiments, and the discipline imposed by having to cope with very large and scaled-up systems from the start. Molecular biology is also increasingly in need of advanced computational tools to accomplish knowledge discovery.

The first of these examples concerns situations in which there is too *little* data for many learning systems to operate effectively. Hunter's example in this case concerns determining the causes of lethality in osteogenesis imperfecta, a sometimes-fatal bone disorder involving point mutations in the amino-acid coding sequences for collagen. The dimensionality of the space is vast (approximately 243-dimensional), yet only approximately 70 relevant sequences have been determined from sufferers of the disease. With such a small dataset, conventional clustering techniques are useless. However, systems such as RELIEF [Kira and Rendall 92], a Focus/Induce/Extract system, can attempt to eliminate irrelevant features in the dataset using statistical tests. (The system described uses C4.5 [Quinlan 86] to extract condition/action rules from the resulting decision trees; see also [Dietterich 89] and [desJardins 92] for related work.)

The system described by Hunter has successfully discovered previously-unknown information about this disease. It has a large collection of machine-learning algorithms in it, and determines which of them to use for the part of the problem at hand by consulting lists of *preconditions* and *applicability* conditions: a tool's preconditions must be completely satisfied for it to be eligible, whereas a tool's applicability conditions are used to help determine which of the eligible tools to use (in general, the cheapest, fastest such tool is chosen).

The second example concerns situations for which there is too *much* data, such as in megaclustering of protein sequences. The current database set being produced by biologists consists of approximately 100,000 sequences, comprised of 20 million amino acids, and doubles about every 18 months. Hunter's system actively determines, based on the current subgoals, which of multiple data sources to contact over the network, how to communicate with the variety of different databases, what sort of analytical tools should be used to analyze the data, which platform(s) to use to do so (since some tools require very large machines, whereas others do not), and so forth.

This system, too, has discovered new scientific results,[2] which is a strong claim of the utility of the reasoning techniques employed.[3]

In both of these examples, and in the goal-directed learning literature in general, the problem of learning has been transformed from a search problem to a planning problem. A naive approach is quite likely to result in simply turning one intractable problem into another, so this transformation should not devolve into first-principle planning, but should instead yield what Hunter calls a *discovery strategy*, or skeletal plans for learning, an espe-

[2] Resulting in publications in the biological literature of its discoveries.

[3] Hunter also makes the point that such a multistrategy approach is similar in spirit to the multiple-competence-modules model proposed in the Society of Mind [Minsky 86]; this point is taken up again in the next section.

cially important feature in learning systems that must query databases—since otherwise the expected outcome of any plan is sufficiently unpredictable that almost all plans of non-trivial length will fail.

Such transformations show great promise in making machine learning in large, real-world problems both tractable and useful in true knowledge discovery applications.

5.2 Related Work in Cognitive Science

5.2.1 Introduction

The cognitive science literature about attentional processes is vast. This overview will examine some of the high points of that literature that seem most salient and that seem to be of most utility in producing ideas about implementations of focus of attention in a machine learning system.

One of the most interesting things about the literature of the last two or three decades is that many of the same questions have been asked for all that time. It is slowly becoming apparent that, in some cases, the questions themselves are likely to be misguided; in others, while it has become clear that certain mechanisms are *not* responsible for attention, it is still unclear which mechanisms *are*.

The confusion exists on many levels, from what constitutes a reasonable theory (e.g., there is disagreement about whether it need be computational, in the sense of [Marr 82]) even to simple aspects of terminology (many have railed against the vague use of supposedly well-defined terms, and, amusingly enough, [White 64] dedicates an entire and rather delightfully readable book to philosophical definitions of terms such as “attention,” “realizing,” “noticing,” and so forth).

Many people have summarized important parts of this literature and the questions surrounding it, with an eye towards computer modeling of the mechanism (one example would be [Chapman 90]) or to debate and clarify certain of the confusions of the field (see, for example, Allport's excellent retrospective [Allport 90]). The debates over whether the questions even make sense are not particularly new; for example, many of the concepts and ideas that [Kinchia 80] feels necessary to bury still have enough life in them that [Allport 90] is still driving stakes through their hearts. [Van der Heijden 92] spends an entire chapter of a book doing likewise, exploring and denigrating theories such as the belief in [Broadbent 71] (for example) of *central* and *limited* capacity.

Such guides to the other literature have been very useful in determining how the field has progressed, and in which directions to proceed in examining this enormous array of work; in some of the discussion that follows, I am particularly indebted to the keen and sometimes provocative thinking of Allport and Kinchia.

A large part of the problem with much of the literature on attention is due to its treatment of "attention" as a single, unitary, central process, rather than as a variety of cognitive mechanisms that mediate human information processing. [Kinchia 80] proposes several illustrative theoretical processes, summarized as:

- *All-or-none attention model*, and *weighted integration model*. These models posit a sort of zero-sum information processing paradigm, in which any attention devoted to one stimulus necessarily robs attention from an unattended stimulus. The former model assumes that this works like a switch—attending to one stimulus essentially completely ignores another stimulus—whereas the latter assumes a sort of linear transfer function between two stimuli, where attention can be "shared" between them, albeit with lower processing efficiency for each.

- *Serial coding models*. Many influential models of perception characterize the initial internal representation of a stimulus as being held in a sensory register, which is assumed fairly rich but unprocessed, and which must be processed relatively quickly lest it decay while in temporary storage in this register. Some view attention in this regard as a switch [Broadbent 58] or a filter [Treisman 69] which determine what information was retrieved from the register for further processing. [Rumelhart 70] characterizes this mechanism as a feature-extraction process, in which individual features are serially extracted and coded from this register.

Note that much of the debate about a single locus of attention seems rooted in more fundamental conceptions of how the mind works, in areas unrelated to selective attention per se. There is a long history of dealing with the mind as if it possessed a homunculus somewhere, leading to theories whose explanatory powers are negligible. Dennett and Kinsbourne, for example, feel this problem keenly [Dennett and Kinsbourne 92]. In reporting research results concerning the perception of subjective time, they spend considerable effort first demolishing the *Cartesian Theater* model of the mind, in which the mind is presumed to have some place where “it all comes together.” Their research instead supports what they call the *Multiple Drafts* model, in which discriminations in multiple modalities are not registered and synchronized before “presentation” to “consciousness,” but instead are distributed in both space and time in the brain. The arguments that they present in demolishing the Cartesian Theater model are of the same sort required to demolish the “single, serial” model of attentional processes, as delineated below.

Two of the major aspects of the problem concern arguments over *early* versus *late* attention (e.g., whether attentional selection occurs before or after stimuli are coded into categories), and which cognitive processes require attention, and are hence limited by attention, and which do not. The major thrust of these arguments is to determine possible constraints in the processing architecture of the brain, so as to determine the overall and detailed architecture of how the brain processes information. Unfortunately, many of these architectural models are imperfect at best, and are surprisingly unhelpful in generating useful architectures with sufficient explanatory power to permit either further analysis of the system, or its reproduction (e.g., as a program).

The vast majority of studies of attention concentrate on visual attention. Of those, many are detailed neuroanatomical studies of either humans or other primates (often via lesion studies or examination of pathological cases). Others are performance tests of healthy human volunteers. A small percentage deal with auditory attention, with an almost insignificant percentage examining other forms of attention. This means that examples of attention are heavily biased towards human (or at least primate) visual attention only.

Such studies of attention are examining a system (namely, human cognition) which is far more complicated than those yet investigated in machine learning. Consequently, while they serve as interesting inspirations for approaches to try, it is not claimed that the research in this thesis either explains anything about mammalian visual attention, or that such studies necessarily will lead to a direct implementation.

5.2.2 The plausibility of attention as a system of limitations

There is a very widespread view that the need for selective attention stems from fundamental limitations in cognitive processing power in particular portions of the brain, and that, if the brain were to have infinite computational power, such attentional limitations would be unnecessary. This is argued over a span of decades by [Broadbent 58] [Broadbent 71] [Broadbent 82], among many others. He and others view attention itself, therefore, as a limited-capacity system, one which must be shared by many processing stages and whose capabilities are therefore competed for by various portions of the brain.

This view also espouses that some tasks are "automatic" and hence do not require attention, presumbly by using portions of the brain whose capacity is not as severely limited [Kahneman 73]. The appeal of dividing cognition up in this rather intuitive fashion, of course, is that, if one could identify the bottlenecks, one might begin to get a handle on how cognition is structured.

Allport [Allport 90] describes Treisman's feature-integration theory (FIT) [Treisman and Gelade 80], [Treisman 88] as one of the best-known of the theories that equate attentional mechanisms with intrinsic bottlenecks in processing. It includes careful characterization and theoretical arguments that, according to the theory, necessitate serial focusing on each item to be perceived in turn in order to correctly perceive objects that must be distinguished by conjunctions of separable features. Yet, immediately after describing Treisman's (and others') theories of attention, Allport (quite rightly) takes issue with much of the terminology of the field; even what is meant by the word *selection* is ambiguous: Does it mean "any task-dependent modulation of sensory neuronal responses?" "Selective facilitation?" "Attentional tagging?" "Selective feature integration?" "Entry to a limited-capacity short-term memory store?"

The term *attention* has similar problems in cognitive science: [Johnston and Dark 86] ask whether attention is some hypothetical causal agency which can be directed or focused on an entity (with the result that this entity may be "selected"), or an *outcome*, characterizing the behavior of the whole organism. They point out that most current attentional theories postulate the above hypothetical causal agency, but that there is a great deal of drift between the two concepts; they also mention that, in most contemporary theories, this causal agency "has all the characteristics of a processing homunculus," which does not help us to understand the underlying mechanisms.

Many of the assumptions about cognitive architecture adopted by models of attention appear in the following list, adapted from [Allport 90]:

- Information processing follows a linearly ordered, unidirectional sequence of processing stages from sensory input to overt response. Parallel or reciprocal processing is disallowed in this model.
- Such a sequence is already known, or can be assumed a priori.
- The processing of nonsemantic attributes occurs before processing of semantic attributes.
- Spatial attribute and relation processing logically precedes categorical or semantic distinctions. There is just *one* locus of attentional selection, hence it can be *early* or *late*, but not both.
- Attentional selection therefore serves as a gate for *any* further processing to be performed; whatever does not make it past this gate will be remain unprocessed.
- There exists a single “central system” of limited capacity, responsible for all cognitive processes that “require attention,” which can only be bypassed for “automatic” processes (defined, of course, as those which do *not* “require attention”).

I will detail below only a few of the ways in which, as mentioned in the introduction to this section, this set of assumptions has begun to fall apart. But as an overall trend, there is growing pressure to develop a theory of attentional selectivity and *control*, rather than the current conception of attention as being a passive information *filter*.

Let us start at the top. If processing is inherently serial, why does the brain seem to have separate processing for “what” versus “where” information? Consider the primate visual system, composed of at least twenty different modules [Desimone and Ungerleider 89] [Ullman 91]. These modules can be broadly grouped into the *ventral* system, crucial for form-based object recognition, and the *dorsal* system, responsible for spatial vision and coordination [DeYoe and Van Essen 88] [Ungerleider and Mishkin 82]. This "what" versus "where" system is quite well established in the literature of visual attention, and poses a rather embarrassing problem for the "single, serial" assumption above. (Indeed, [Felleman and Van Essen 91], to pick only one of many similar papers, demonstrate that among 32 areas that are associated with visual processing in the primate visual cortex, approximately 40% of all possible connection pathways between the modules actually exist! This makes cortical visual processing organization look more like a bush than a hierarchical or serial system, and does not even include the straightforward reciprocal neural connections in each cortical area; see the discussion immediately below.)

Worse yet, almost everywhere in the cortex, the "forward," *afferent* connections that one would expect (leading from the retina toward higher centers of processing) are paralleled by equally rich, "backward," or *efferent* connections [Ullman 91]. If all processing proceeds in the afferent direction, what are all of those reciprocal connections *doing* there? Many have proposed ideas: for example, [Mumford 91] proposes that each cortical area is responsible for updating and maintaining knowledge of a specific aspect of the world, at any given level from low level raw data to high level abstract representations, and that the multiple, often conflicting hypotheses which result are integrated by thalamic neurons and then sent back into the cortex, making the thalamo-cortical loop a sort of "active blackboard" system and thereby explaining the density of reciprocal cortical connections.

Ullman [Ullman 91] has proposed a particularly interesting idea with his sequence-seeking counterstreams model, in which he posits that the neocortex searches for mappings between "source" and "target" representations, exploring both "top-down" and "bottom-up" a large number of alternative sequences in parallel.[4] Finally, even though most diagrams of the visual cortex show each module interacting with a few nearby ones in a semiwell-behaved processing mesh, almost every module *also* has a direct connection (e.g., an output pathway) to some motor or action system, forming a number of direct, parallel links between sensory and motor systems that could potentially bypass all levels of higher processing [Creutzfeldt 85]. What are those links doing there, if the "single, serial" model is correct?

It has also been argued by many that tasks conforming to what [Kahneman and Treisman 84] call the "filtering paradigm," in which the to-be-selected visual items are cued by *non*spatial visual attributes such as color or size, may instead depend on selective cueing by *location* [Butler and Currie 86, Johnston and Pashler 90], and that where this spatial separation is absent, performance drops [Johnston and Dark 86]. This tends to imply that many "early selection" paradigms of visual attention may instead correspond to spatial selection.

But the picture is murky even in spatial selection. For one thing, extensive experimental evidence reveals many *different* coordinate systems and corresponding transformations along the path from the retinotopic input through the cortex. For example, [Allport 90] provides a virtual laundry list of such transformations, mentioning some that take account of eye and head position, some that code location in terms of arm- or body-centered coordinates, and some based on environment- and perhaps object-centered coordinate systems; a small sampling of work in this area can be found in [Andersen 87] [Andersen 89] [Ellis et al 89] [Feldman 85] [Hinton and Parsons 88] [Marr 82] [Soechting, Tillery, and

[4] Such counterstream architectures, if they exist in the brain at all, may no longer be unique to it, however. Bob Sproull of Sun Microsystems has proposed [Sproull 94] a novel, "counterflow pipeline" architecture for advanced, pipelined RISC CPU's which shares many remarkable features with Ullman's counterstreams model of processing. Instructions and results propagate in opposite directions in a processing ladder, interacting with each other as they pass, and employ only local interaction (e.g., only within a ladder level, or between two adjacent rungs of the ladder). Such a design also admits an asynchronously-clocked implementation, making it more similar to possible cortical models such as Ullman's. However, the intersection between cognitive science and machine architecture is understandably not what it could be: neither Ullman nor Sproull had heard of the other's work.

Flanders 90] [Zipser and Andersen 88]. This does not even include the many lesion studies which investigate neglect in various coordinate systems after brain damage.

If recent results seem to put nails in the coffin of attention as a serial, feedforward, strictly limitation-based strategy, what models are proposed instead? [Allport 90] argues that the influence of attention in noncategorical, spatial-vision systems is of the form of *enhancement* of neuronal responsiveness in attended locations, rather than attentuation of unattended locations, and that many results involving delays in attending to stimuli reflect the time cost of disengagement from the cued location, rather than withdrawal of processing resources from the uncued location. (He also notes work, such as [Driver and Tipper 89], which points out the problems with equating “no processing” and “no interference.”)

Indeed, lesion studies such as in dorsal simultanagnosia, in which the patient perceives only one part of any given object even though his visual field is often full and complete, seem to indicate that such damage leads to an inability to *disengage* from one part of the visual field in order to shift attention to a different part of it: unilateral lesions [Posner et al 87] [Morrow and Ratcliff 88] can lead to problems shifting attention to the contralateral side, and full simultanagnosia can lead to problems shifting attention in *any* direction [Luria et al 63].

Viewing attentional processes as a process involving commitment of resources, rather than filtering, leads [Crick and Koch 90], for example, to suggest that attention facilitates local competition among neurons: in other words, when a local group of neurons is not attended to, it can have multiple (ambiguous) outputs, but attention then narrows down the possible outputs, forcing disambiguation. This view of attention is quite different from that of protecting the limited computational power of a single center from overload.

This argument for a multiplicity of attentional mechanisms, each “specialists” in a particular cognitive area, fits in nicely with the Society of Mind hypothesis [Minsky 86]. While Minsky posits that many of the aspects of attention have to do with limits (e.g., limitations in processing leading to the intuitively serial feeling of thought, or limitations in a particular agent’s ability leading to inability in tracking multiple locations simultaneously), the theory does not say that there is a *single* limit anywhere—only that different agents will likely contribute different limitations.

The existence of multiple loci of attentional control are reinforced dramatically by [Mangun, Hillyard, and Luck 90], who use a combination of MRI brain images, behavioral data, and event-related brain potential mapping. While this research comes down rather strongly on the side of “early” selection (since the effects cited occur very quickly, within 150 msec), [Sperling, Wurst, and Lu 90] introduce a new theoretical construct, attentional “tags,” through which visual item traces may be selected from short-term memory, rather than positing a single filter. Such an interpretation is completely in support of multiple loci of attentional control.

It is interesting to note that the majority of even current work in the machine learning community still treats attention as a single, serial pathway, and structures its systems accordingly. (See, for example, Figure 17, on page 113, from [Markovitch and Scott 93], and the discussion in Section 5.1, on page 111.)

One reason for this might be that current machine learning systems are still too primitive to take advantage of architectures rich in reciprocal connections, or that contain multiple loci of control or information processing. For example, the bulk of this thesis concerns itself with single-agency pruning, of the type of “limitations and bottleneck” school so denigrated above. In addition, the sensory and cognitive system modelled in this research is greatly simplified compared to even the most rudimentary levels of human cognition;

many insects might have more sensory processing abilities, and even the simplest mammal must be better at memory and generalization.[5]

Just because the evidence for a single, serial control of attention no longer appears to be compelling is no reason, of course, to discount much of the work that has been done in attention. While it may not be the case that an explanation of one particular aspect of attention explains all of attention (either in that modality or others), there are many probably-correct explanations of parts of the attention puzzle. Unfortunately, few have implemented their theories, possibly because many of them are insufficiently precisely described for such implementation. This makes it even more difficult to determine which theories might be correct.

For example, [Chapman 90] cites several aspects of visual attention, such as results in covert attention [Posner et al 80], and the winner-take-all addressing pyramid in [Koch and Ullman 85] in support of visual spotlight search behavior. [Treisman and Gormican 88] have done extensive research on visual pop-out behavior in visual search routines; for surveys of visual search in general, see [Julesz 84] and [Treisman and Gelade 80]. But Koch and Ullman did not implement their theory; in fact, Chapman's work is one of the few to implement several subtheories of visual attention.

It may be, as machine learning systems become more sophisticated, employing multiple processing strategies in a richly-connected information architecture, that they will be better positioned to take advantage of current thinking about attention in cognitive science.

[5] Although the lack of generalization in the schema system, as currently designed, does seem to put it on a par with certain insects. For example, bees apparently remember places retinotopically—if they learn a shape with one part of their eye, they can only recognize it again with that same part [Christensen 94]. Bees, which have magnetite in their abdomens as part of their navigation system, face magnetic south (or magnetic northwest in certain cases in which south is infeasible) when encountering and departing targets of interest. By doing so, they can image the target in the same orientation; rather than rotating a mental representation of the target, they simply rotate their real eyes instead until a match is acquired. Artificially imposed external magnetic fields lead to predictable perturbances of this behavior.

Organizations such as the subsumption architecture [Brooks 86], for example, or the Society of Mind [Minsky 80] seem as if they will be logical computational testbeds for implementing computational verification of multiple-loci attentional models.

Indeed, as shown by the some of the systems mentioned in Section 5.1.3, particularly the multistrategy systems of Hunter, Ram, Cox, and others, the increasing complexity of modern learning systems is forcing implementations of the control of their attentional focus down just the sort of pathways that the multi-locus models of modern cognitive science might lead one to expect.