Resource Constraints on Computation and Communication in the Brain

Sashank Varma
Stanford Center For Innovations in Learning
450 Serra Mall, Building 160
Stanford, CA 94305-2055
sashank@stanford.edu

Abstract

This paper contributes to the emerging literature at the border between AI and cognitive neuroscience, analyzing the resource constraints that shape brain function. The brain is conceptualized as a set of areas that collaborate to perform complex cognitive tasks. Both (1) computation within individual areas and (2) communication between collaborating areas are viewed as resource-consuming activities. The efficient deployment of limited resources is formalized as a Linear Programming problem which the brain is hypothesized to solve on a moment-by-moment basis. A model of language processing is analyzed within this framework and found to exhibit resource utilization profiles consistent with those observed in functional neuroimaging studies of humans.

1 Introduction

In his 2002 AAAI presidential address, Tom Mitchell called for a rapprochement between AI and the rapidly emerging discipline of cognitive neuroscience. AI researchers have begun to answer this call. Some are modeling the function of brain areas, e.g., Dean [2005] modeled primary visual cortex as a hierarchical Bayesian network. Others are applying classification algorithms to functional Magnetic Resonance Imaging (fMRI) data to predict cognitive states, e.g., which word a person is thinking of [Mitchell *et al*., 2004]. Still others are using clustering algorithms [Dimitriadou *et al*., 2004] and multi-agent approaches [Richard *et al*., 2004] to find patterns in large fMRI data sets.

The research reported in this paper contributes to this effort. It focuses on the resource constraints that shape brain function. Specifically, the brain is viewed as a set of areas that collaborate to perform complex cognitive tasks such as language processing, problem solving, and spatial reasoning. There are limitations on the computational power of individual brain areas and on the communication bandwidth between different brain areas. Both kinds of limitations can be conceptualized as resource constraints, and the brain can be seen as allocating resources in a way that respects these constraints while maximizing cognitive throughput.

This view is developed and evaluated below. First, resource constraints on brain function are formalized using the machinery of operations research. Second, this machinery is analyzed and three canonical patterns of resource allocation are identified. Third, this machinery is applied to a model of language processing. The model is shown to exhibit the patterns and to account for brain activations as measured by fMRI. Finally, three avenues for future research are sketched.

2 Formalizing Resource Constraints

Cognition is the emergent product of multiple collaborating brain areas. It is profoundly shaped by resource constraints on computation within individual brain areas and communication between brain areas. These resource constraints can be formalized as a linear programming (LP) problem.

Each brain area is termed a *center*. Each center is an encapsulated production system containing declarative elements and production rules. Declarative elements are lists of attribute-value pairs annotated with continuous activation levels. They are processed by productions, which have activation thresholds on their condition sides and whose primary action is to direct activation from one declarative element to another modulo a weight that is either positive (representing excitation) or negative (representing inhibition or suppression). The control structure is fully parallel: at each point in time, all productions are matched against all declarative elements in all centers and fired. We assume a decomposition of the brain into *M* centers.

A *function* is a primitive cognitive process, such as detecting an edge or retrieving the meaning of a word. Functions are abstractions that enable a higher level of description; each is implemented by a combination of declarative elements and productions. We assume a decomposition of cognition into *N* functions.

Each center is *specialized* for multiple functions, and conversely, each function can be performed by multiple centers. This proposal is intermediate between a pure modularity that maps functions one-to-one to centers and a pure distributivity that maps every function to every center. The specialization of center *i* for function *j* is denoted S_{ij}, where $S_{ij} \in [1,\infty)$. It specifies the amount of the center's resources (i.e., activation, the common currency of storage and processing) that are required to perform one unit of the function. A value of 1.0 represents perfect specialization, larger values represent lesser specializations, and a value of ∞ represents a complete inability of center *i* to perform function *j* (because resources are limited, as we will see below).

At each point in time during task performance, there exist a number of functions to be performed. The amount of function *j* performed by (i.e., assigned to) center *i* is denoted A_{ij}. The *assignment problem* is to determine the A_{ij}.

The assignment problem is a constraint satisfaction problem. One set of constraints specifies the resource demands of the functions to be performed. Specifically, the resource demands of function *j* are denoted R_j and the following constraint is enforced at all times:

$$\sum_{i=1}^{M} A_{ij} \leq R_j \tag{1}$$

It ensures that the resources supplied by all centers to the function are as close as possible to the resources demanded. Another set specifies *intra-center* constraints on computation within individual centers. Specifically, the resource supply of center *i* is denoted C_i and the following constraint is enforced at all times:

$$\sum_{j=1}^{N} \left(A_{ij} \times S_{ij} \right) \leq C_i \tag{2}$$

It ensures that the resources supplied by the center to all functions do not exceed its resource supply. The final set specifies *inter-center* constraints on the joint resource consumption of multiple centers, which are interpreted as bandwidth limitations on communication between these centers. The number of inter-center constraints and the nature of each one (i.e., the centers whose resource supplies it jointly constrains) are empirical matters. For ease of exposition, we assume a single inter-center constraint on the joint resource consumption of all centers:

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \left(A_{ij} \times S_{ij} \right) \leq C_{CORTEX} \tag{3}$$

It ensures that the resources supplied by all centers to all functions do not exceed the resource supply of the entire cortex, denoted C_{CORTEX}.

The constraints (1), (2), and (3) only partially determine the assignment of functions to centers because they are satisfied by many different assignments, e.g., A_{ij}=0 for all *i* and *j*. This ambiguity is resolved by defining a measure of the goodness of an assignment, expressed as a linear combination of the A_{ij}, to be maximized:

$$\sum_{i=1}^{M} \sum_{j=1}^{N} \left(W_{ij} \times A_{ij} \right) \tag{4}$$

Defining $W_{ij} := 1/S_{ij}$ ensures that all other things being equal, function *j* will be assigned to the center *i* most specialized for it (i.e., whose S_{ij} is minimal and W_{ij} is maximal).

(1), (2), (3), and (4) define an LP problem that the brain is hypothesized to solve at each point in time. The result is an assignment of functions to centers that (1) satisfies the resource demands of functions to be performed (to the degree possible), (2) respects resource constraints on computation within individual centers, (3) respects resource constraints on communication between centers, and (4) maximizes cognitive throughput. The simplex algorithm is used to solve the assignment problem.

This account of the resource constraints on brain function can be empirically evaluated via the following measurement assumption: The *capacity utilization* of center *i*, denoted CU_i, is the proportion of its resources currently being used.

$$CU_i := \frac{\sum_{j=1}^{N} \left(A_{ij} \times S_{ij} \right)}{C_i} \tag{5}$$

The capacity utilization of a center will be used to predict activation (as measured by fMRI) in the corresponding brain area. This is a reasonable assumption because the fMRI signal reflects the vascular response to the consumption of neurobiological resources [Logothetis, 2003].

3 Three Patterns of Resource Allocation

Solving the assignment problem produces three canonical patterns of resource allocation. This section describes and illustrates them in the context of a toy model. Specifically, consider the case of two centers with equal resource supplies (C_1=C_2=6) where the first center is more specialized for a function 1 than the second center (S_{11}=1 and S_{21}=2). Further assume that the resource demands of the function are relatively light (e.g., R_1=3). This defines the LP problem:

maximize:	$A_{11} + {}^{1}/_{2}A_{21}$	(objective function)
subject to:	$A_{11} + A_{21} \leq 3$	(function)
	$A_{11} \leq 6$	(first center)
	$2A_{21} \leq 6$	(second center)
	$A_{11}, A_{21} \leq 0$	(default LP constraints)

depicted in Figure 1a. The constraints, shown as dashed lines, demarcate the boundaries of the shaded *feasibility region* – the set of satisfactory assignments (A_{11}, A_{21}). The simplex algorithm effectively positions the objective function line, shown as a solid line, so that it touches a vertex of the feasibility region. This vertex represents the values of A_{11} and A_{21} that maximize the objective function. When the resource demands are relatively light, the objective function is maximized at a vertex that assigns the function entirely to the well-specialized first center (i.e., A_{11}=3 and A_{21}=0).

3.1 Pattern 1

As task difficulty increases from low to moderate, task performance consumes an increasing amount of resources. For example, as the resource demands of the function (R_1) increase from 4 to 5 to 6 units, the shape of the feasibility region changes, and consequently so does the vertex at which the objective function is maximized, as shown in Figure 1b.

Over this range of task difficulty, the well-specialized first center continues to possess a resource supply sufficient to satisfy the increasing resource demands, and therefore continues to be assigned all of the function (i.e., A_{21}=0).

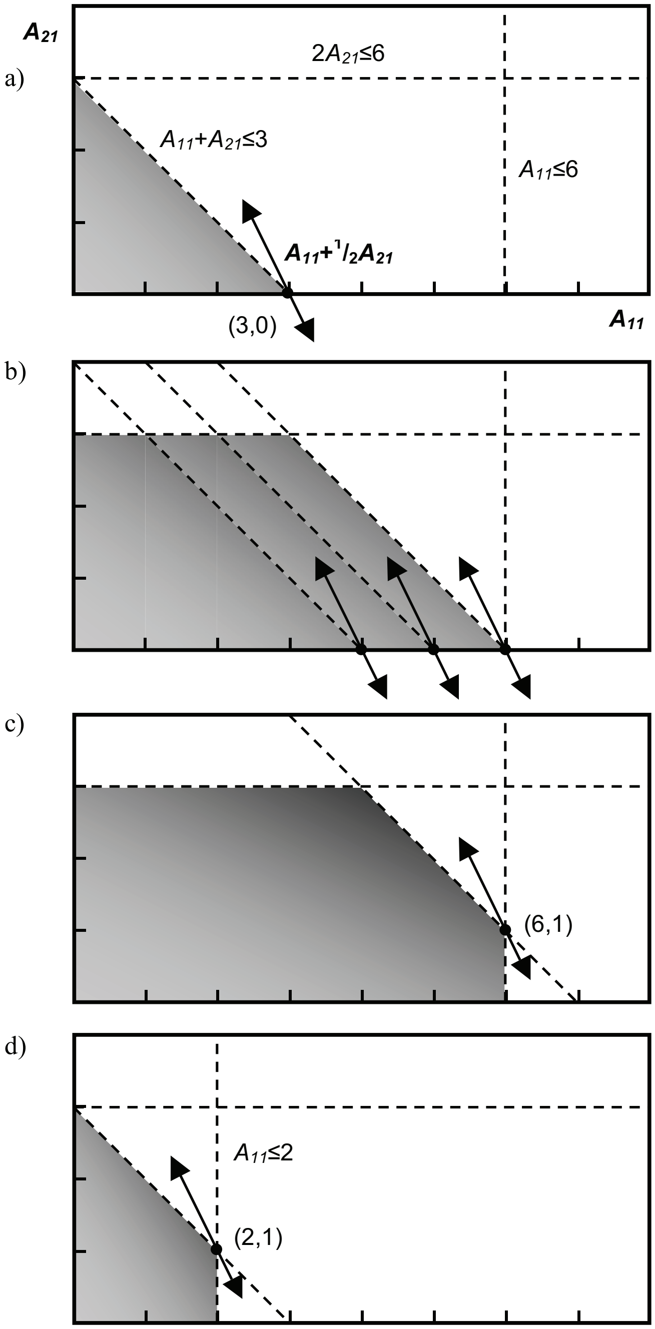

Figure 1. (a) Low resource demands. (b) Pattern 1: moderate resource demands. (c) Pattern 2: high resource demands. (d) Pattern 3: reduced resource supply following lesion.

3.2 Pattern 2

Next, consider a task of high difficulty. If the resource demands of the function (R_1) increase to 7 units, the feasibility region will change shape yet again, as shown in Figure 1c. The vertex that maximizes the objective function assigns 6 units to the well-specialized first center and 1 unit to the less-specialized second center. In other words, resource demands now exceed the resource supply of the first center, and excess processing spills over to the second center.

When resource demands decrease following some peak, performance of the function will revert entirely to the first center. This corresponds to a transition back to pattern 1 shown in Figure 1b.

3.3 Pattern 3

Finally, consider the case of a focal lesion to a brain area, which is simulated by drastically reducing the resource supply of the corresponding center. As a result, even easy tasks will result in the spillover of excess resource demands to less-specialized centers. For example, if the resource supply of the first center is reduced from 6 units to 2 units, then even light resource demands (R_1=3) will require recruitment of the second center to help perform the function. This is shown in Figure 1d, and it stands in contrast to pattern 1 shown in Figure 1a, where in the undamaged system the same resource demands can be entirely satisfied by the first center.

It should be noted that all patterns are bottom-up consequences of solving the assignment problem on a moment-by-moment basis. There is no executive making resource allocation decisions in a top-down fashion.

4 Application to Language Processing

This section situates the discussion in the context of a model of language processing. The model is shown to exhibit the three resource allocation patterns and to provide a good account of relevant fMRI data.

The model is composed of four centers. Left posterior superior temporal gyrus (L. STG; Wernicke's area) and left inferior frontal gyrus (L. IFG; Broca's area) are the core components of the human language network. They are well-specialized for lexical, syntactic, and semantic/thematic functions. The homologous areas in the right hemisphere, R. STG and R. IFG, also belong to the language network. They are less-specialized for these functions (although they are well-specialized for linguistic functions not considered here, such as prosodic processing). The four centers collaborate to process language. It is beyond the scope of this paper to describe the model any further; the interested reader is directed to Just and Varma (2006) for the details.

4.1 Patterns 1 and 2

The time to comprehend a sentence is an increasing function of its length, of course, but also of its structural complexity. For example, consider the following three sentences:

- Conjoined Actives: [$_{\text{clause-1}}$The senator attacked the reporter] [$_{\text{clause-2}}$and admitted the error].
- Subject-Relative: [$_{\text{main-clause}}$The senator [$_{\text{relative-clause}}$that attacked the reporter] admitted the error].
- Object-Relative: [$_{\text{main-clause}}$The senator [$_{\text{relative-clause}}$that the reporter attacked] admitted the error].

They contain the same number of words (nine) and clauses (two). The *conjoined actives* sentence concatenates the two clauses. Its resource demands are relatively light because when the second clause begins, the first clause has been completely processed, and therefore no associated declarative elements must be maintained. By contrast, the other sentences embed one clause (the "relative clause") within the other (the "main clause"). This imposes additional resource demands because the declarative elements associated with the main clause must be maintained while the relative clause is comprehended. The *object-relative* sentence is more resource demanding than the *subject-relative* sentence because the declarative elements associated with the relative clause must be maintained deeper into this clause (until the verb *attacked*) before thematic representations can be computed.

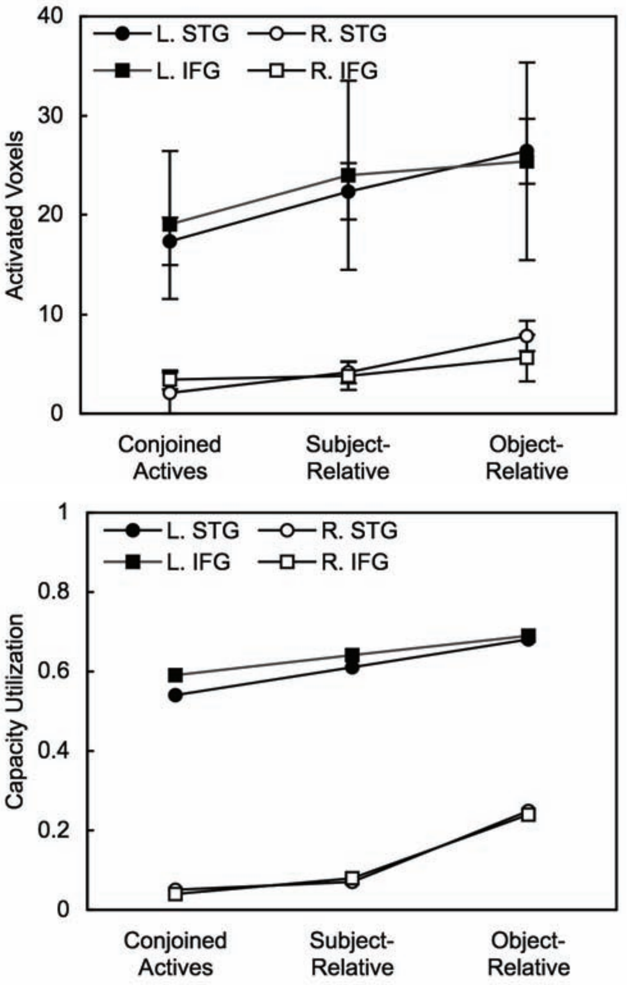

Figure 2. (top) Sentence comprehension data (Just et al. 1996). (bottom) Model capacity utilizations.

Just *et al*. [1996] had participants read sentences of the three types and answer comprehension questions. They collected activations in the four brain areas simulated by the model: left and right STG and IFG. These activations are shown in the top panel of Figure 2. There are two patterns to notice. First, for the well-specialized left-hemisphere areas, activation increases from moderate to high levels with increasing syntactic complexity. This is an example of pattern 1 illustrated in Figure 1b. Second, for the less-specialized right-hemisphere areas, activation is initially negligible, but increases to a moderate level with increasing syntactic complexity. This is an example of pattern 2 illustrated in Figure 1c. The capacity utilizations of the corresponding model centers while processing the three sentence types are shown in the bottom panel of Figure 2. (Four free parameters – the resource supplies of the four centers – were estimated to maximize the fit to the human data.) The model correctly displays pattern 1 in its left-hemisphere centers and pattern 2 in its right-hemisphere centers. This qualitative correspondence is confirmed by the 0.98 (*p*<.01) correlation between human and model performance.

4.2 Pattern 3

An often-observed phenomenon in patients with stroke-induced lesions is *contralateral takeover*: when a brain area is damaged, the functions it used to perform migrate permanently to the homologous area in the other hemisphere. This was observed in a patient tested by Thulborn *et al.* [1999] who had suffered a stroke six months earlier that damaged his L. IFG. The patient initially experienced a dense expressive aphasia but subsequently recovered much of his language function. Activation was measured in left and right STG and IFG, the four brain areas simulated by the model, while he read simple five- and six-word sentences. These data are shown in the top panel of Figure 3. The patient showed the normal pattern in STG, with much more activation in the well-specialized left area than the less-specialized right area. Strikingly, this pattern reversed in IFG: there was no activation in the damaged left area and substantial activation in the healthy right area. This is an example of pattern 3 illustrated in Figure 1d. Thulborn et al. interpreted this as evidence that R. IFG had taken over the functions of the damaged L. IFG. To simulate the L. IFG lesion, the corresponding model center was stripped of its resources. The model processed the same sentences the patient read. The capacity utilizations of its four centers are shown in the bottom panel of Figure 3. (Three free parameters – the resource supplies of the undamaged centers – were estimated to maximize the fit to the human data.) The model correctly displays pattern 3, with normal left-lateralization in STG and striking right-lateralization in IFG. Its capacity utilizations provide a good quantitative account of the observed activations (r=0.99, *p*<.01).

4.3 Inter-Center Resource Constraints

The fits to the data reported thus far have been due to intra-center resource constraints, i.e., limitations on computation within individual centers. We turn now to inter-center resource constraints, which are interpreted as bandwidth limitations on communication between different centers. Inter-center constraints do not normally operate during language processing because it recruits a relatively small number of collaborating centers. However, they do operate when language processing is paired with a second task such as spatial reasoning because this greatly increases the number of collaborating centers, and therefore the joint resource demands on the system.

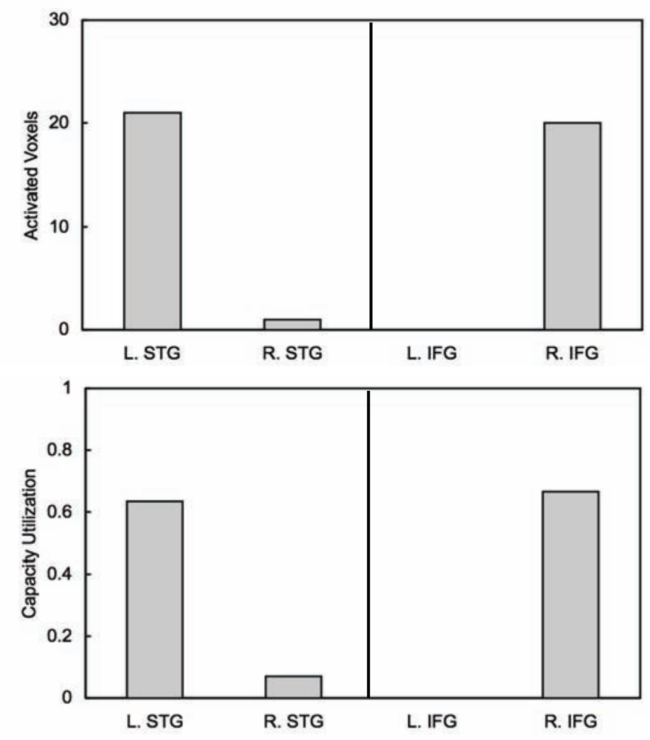

Figure 3. (top) Lesion patient data (Thulborn et al., 1999). (bottom) Lesioned model capacity utilizations.

For example, Just *et al.* [2001] had participants perform sentence comprehension and mental rotation tasks individually as single-tasks and concurrently as dual-tasks. (The mental rotation task required rotating 3D figures to determine whether they are the same or mirror-images [Shepard and Metzler, 1972].) The activations in one component of the language network, bilateral STG, and one component of the spatial network, bilateral superior parietal lobule, during the single-tasks and dual-task are shown in the top panel of Figure 4. The striking result is *underadditivity*: the activations during dual-tasking are less than the sum of the activations during the single-tasks (even though behavioral performance remains at high levels).

In separate work, we developed a model of mental rotation consisting of centers corresponding to bilateral superior parietal lobule (SPL) and bilateral dorsolateral prefrontal cortex. (The implementation details are found in Just and Varma [2006].) The sentence comprehension and mental rotation models were paired to form a *base* model of dual-tasking. This model contains only intra-center constraints on resource consumption within individual centers. An *augmented* model was also constructed by adding to the base model a single cortex-wide inter-center constraint on the joint resource consumption of all centers.

The goal was to evaluate the necessity of inter-center constraints for accounting for the Just *et al*. [2001] data by comparing the fits of the base and augmented models. The capacity utilizations of the bilateral STG and SPL centers of the base model during the single-task and dual-task conditions are shown in the middle panel of Figure 4; those of the augmented model are shown in the bottom panel. (The free parameters are the intra-center resource supplies of the individual centers and, in the case of the augmented model, the inter-center resource supply.) Qualitatively speaking, only the augmented model correctly exhibits the observed underadditivity during dual-tasking. Quantitatively speaking, the augmented model accounts for reliably more variance than the base model, 95% vs. 79% (p<.01).

5 Future Directions

The research described here represents a promising first step towards formalizing the resource constraints that shape brain function. It also suggests a number of avenues for future exploration.

5.1 The Assignment Problem

The assignment problem is currently formalized as an LP problem and currently solved using the simplex algorithm. However, the centralized nature of simplex is neurally implausible. One avenue for future research is therefore to develop algorithms that solve the assignment problem in a de-centralized manner. Such algorithms will likely lack the optimality of simplex but will be more neurally plausible. A number of promising candidates have been proposed by computer scientists interested in the parallel solution of LP problems [Alon and Megiddo, 1994; Lustig and Rothberg, 1996; Maros and Mitra, 2000] and by AI researchers interested in cooperative computation among resource-constrained agents [Wooldridge and Dunne, 2006; Wellman, 1993].

5.2 Inter-Center Communication

The communications infrastructure of the human brain is slowly being revealed by diffusion tensor imaging studies of the white matter tracts through which neural information flows. This infrastructure is currently modeled by inter-center resource constraints on the joint computation of multiple centers. The benefit of this approach is a unified account of intra- and inter-center resource constraints. The drawback is that it is conceptually too far from the point-to-point communication network of the brain. Therefore, an avenue for future research is to directly capture the connectivity between brain areas. Luckily, this can be done within the LP formalism. Specifically, the *transshipment* problem is an LP problem for the optimal transport of goods over routes of varying carrying capacity. By formalizing intra-center computation as a conventional LP problem and inter-center communication as a transshipment problem, a unified and neurally-plausible account of resource constraints on brain function should still be possible.

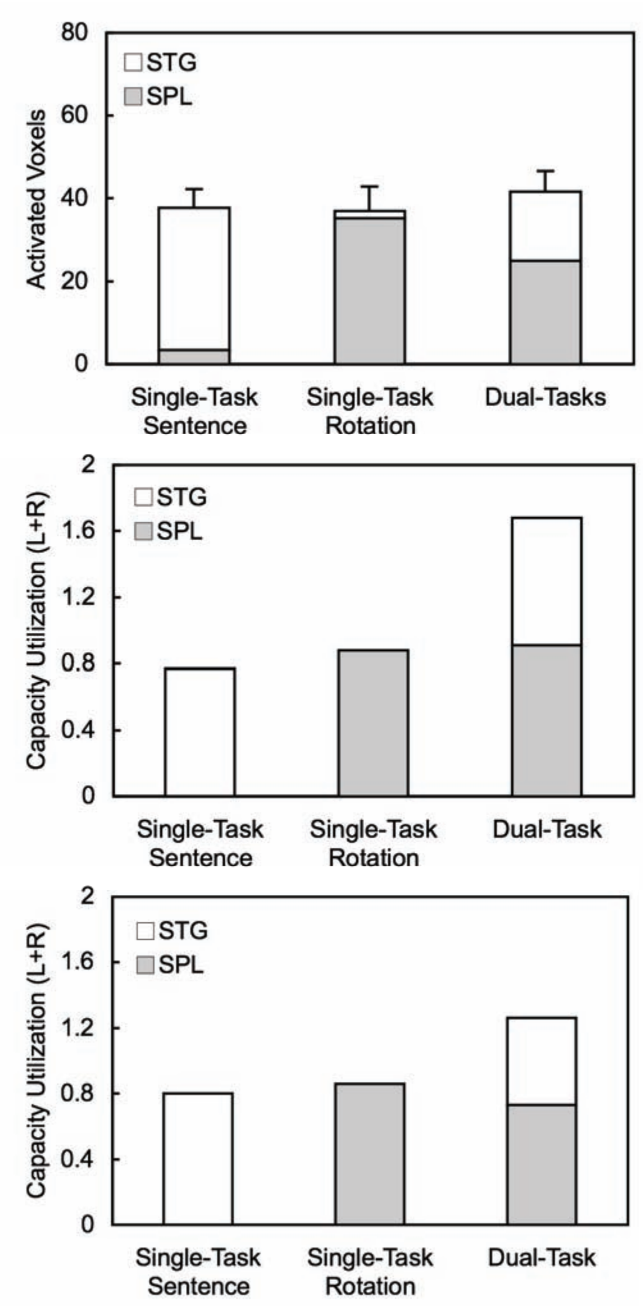

Figure 4. (top) Dual-task data (Just et al., 2001). (middle) Base model capacity utilizations. (bottom) Augment model capacity utilizations.

References

[Alon and Meggido, 1985] Noga Alon and Nimrod Megiddo, N. Parallel linear programming in fixed dimension almost surely in constant time. *Journal of the ACM*, 41:422-434, 1994.

[Dean, 2005] Thomas Dean. A computational model of the cerebral cortex. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, pages 938-943, Cambridge, MA, 2005. MIT Press.

[Dimitradou *et al.*, 2004] Evginia Dimitriadou, Markus Bart, Christian Windischberger, Kurt Hornik, and Ewald Moser A quantitative comparison of functional MRI cluster analysis. *Artificial Intelligence in Medicine*, 31:57-71, 2004.

[Just *et al.*, 1996] Marcel A. Just, Patricia A. Carpenter, Timothy A. Keller, William F. Eddy, and Keith R. Thulborn. Brain activation modulated by sentence comprehension. *Science,* 274:114-116, 1996.

[Just *et al.*, 2001] Marcel A. Just, Patricia A. Carpenter, Timothy A. Keller, Lisa Emery, Holly Zajac, and Keith R. Thulborn. Interdependence of non-overlapping cortical systems in dual cognitive tasks. *NeuroImage*, 14:417-426, 2001.

[Just and Varma, 2006] Marcel A. Just and Sashank Varma. *The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition*. CMU Center for Cognitive Brain Imaging Technical Report CCBI-2006-1.

[Logothetis, 2003] Nikos K. Logothetis. The underpinnings of the BOLD functional magnetic resonance imaging signal. *The Journal of Neuroscience*, 23:3963-3971, 2003.

[Lustig and Rothberg, 1996] I. J. Lustig and E. Rothberg. Gigaflops in linear programming. *Operations Research Letters*, 18:157-165, 1996.

[Maros and Mitra, 2000] I. Maros and G. Mitra. Investigating the sparse simplex algorithm on a distributed memory multiprocessor. *Parallel Computing*, 26:151-170, 2000.

[Mitchell *et al.*, 2004] Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, F. Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57:145-175, 2004.

[Richard *et al.*, 2004] Nathalie Richard, Michel Dojat, and Catherine Garbay. Automated segmentation of human brain MR images using a multi-agent approach. *Artificial Intelligence in Medicine*, 30:153-175, 2004.

[Shepard and Metzler, 1971] Roger Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701-703, 1971.

[Wellman, 1993] Michael P. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research*, 1:1-23, 1993.

[Wooldridge and Dunne, 2006] Michael Wooldridge and Paul E. Dunne. On the computational complexity of coalitional resource games. *Artificial Intelligence*, 170:835-871, 2006.