Lesioned Networks As Models Of Neuropsychological Deficits

John A. Bullinaria
School of Computer Science
The University of Birmingham
Birmingham, B15 2TT, UK
j.bullinaria@physics.org

RUNNING HEAD: Lesioned Networks and Neuropsychological Deficits

To appear in: *The Handbook of Brain Theory and Neural Networks, Second edition*, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002. http://mitpress.mit.edu

INTRODUCTION

Cognitive neuropsychology uses the patterns of performance observed in brain damaged patients to constrain our models of normal cognitive function. Historically, this methodology was rooted in simple "box and arrow" models, with particular cognitive deficits being taken to indicate selective breakdown of corresponding "boxes" or "arrows". Studying patients with complementary patterns of deficit allows us, in principle, to piece together a complete model of mental structure (Shallice, 1988). Of particular importance in this process has been the concept of *double dissociation*, which has been taken to imply modularity within many systems. If one patient can perform task 1 better than task 2, and another can perform task 2 better than task 1, then a natural explanation is in terms of separate modules for the two tasks.

In recent years, connectionist techniques have been employed to model the operation and interaction of these "modules" in increasing detail (Farah, 1994). Networks of simplified processing units loosely based on real neurons are set up with general architectures based on known physiology, trained to perform appropriately simplified versions of the human tasks, and iteratively refined by checking their performance against humans. Such network models can clearly be wired together in the manner of the old box and arrow models, with all the old explanations of patient data carrying through. The obvious advantage now is that we can look at the details of the degradation of the various components, and removing neurons or connections in our models constitute natural analogues of real brain damage. Moreover, as well as providing elaboration of the previous models, we can also question the validity of the old assumptions of neuropsychological inference, and explore the possibility that processing is actually more distributed and interactive than the older models implied.

This article reviews the general issues involved in lesioning neural network models to simulate neuropsychological deficits. I shall point out potential sources of misleading results, clarify apparent contradictions in the literature, and discuss some representative models.

LESIONING SIMPLE FEED-FORWARD NETWORKS

Many neural network models of human performance are based on simple feed-forward networks that map between conveniently simplified input and output representations via a single hidden layer, or have such systems as an identifiable sub-components. An important feature of these models is that they *learn* to perform the relevant tasks by iteratively adjusting their connection weights (e.g. by some form of gradient descent algorithm) to minimise the output errors for an appropriate training set of input-output pairs. Generally, we simply assume that the quick and convenient learning algorithms we choose will generate similar results to those produced by more biologically plausible procedures. Comparisons between Back Propagation and Contrastive Hebbian Learning by Plaut and Shallice (1993) provide some justification for this assumption. We can then compare the development of the networks' performance during training and their final performance (e.g. their output errors, generalization ability, reaction times, priming effects, speed-accuracy trade-offs, robustness to damage, etc.) with the performance of human subjects to narrow down the correct architecture, representations, and so on, to generate increasingly accurate models.

An obvious feature of network learning is that performance on one pattern will be affected by training on other patterns. It follows straightforwardly from adding up the network weight change contributions due to individual training patterns that:

1. Regular items will be learned more quickly than irregular items, because consistent weight changes combine and inconsistent weight changes cancel.

2. High frequency items will be learned more quickly than low frequency items, because the appropriate weight changes get applied more often.

3. Ceiling effects will arise as sigmoids saturate and weight changes tend to zero.

These fundamental properties of neural network learning not only result in human-like *age of acquisition* effects but indirectly account for realistic patterns of reaction times, speed-accuracy trade-off effects, and so on (Bullinaria, 1997). Having trained our networks, and confirmed that they are performing in a sufficiently human-like manner, we can then set about inflicting simulated brain damage on them. Small (1991) considered the various ways in which connectionist networks might be lesioned, and discussed their neurobiological and clinical neurological relevance. He identified two broad classes of lesion: *diffuse* such as globally scaling or adding noise to all the weights, and *focal* such as removing adjacent subsets of connections and/or hidden units. Which of these we choose will naturally depend on the type of patient we are modelling. Focal lesions would be appropriate for stroke patients, whereas diffuse lesions would be required for diseases such as Alzheimer's. Generally, for our simplified models, it is appropriate to examine all these possibilities. Finally, we should be aware that relearning after damage may affect the observed pattern of deficits, and so we must check this also (Plaut, 1996).

The relevant issues have been explored in an abstract setting by Bullinaria (1999) who trained a simple feed-forward network (with 10 inputs, 100 hidden units and 10 outputs, with binary inputs and output targets) on two sets of 100 regular items (different permuted identity mappings) and two sets of 10 irregular items (random mappings). One regular set and one irregular set appeared during training 20 times more frequently than the others. Figure 1 shows that both regularity and frequency do indeed affect the speed of learning in the expected manner.

Bullinaria and Chater (1995) explored the effects of damage on fully distributed, homogeneous, connectionist systems, and investigated the possibility that double dissociation between regular and irregular items could arise without modularity. They found that lesioning trained networks by removing random hidden units, removing random connections, globally scaling the weights, or adding random noise to the weights, all led to very similar patterns of deficits. They concluded that, assuming one successfully avoids small scale ar-

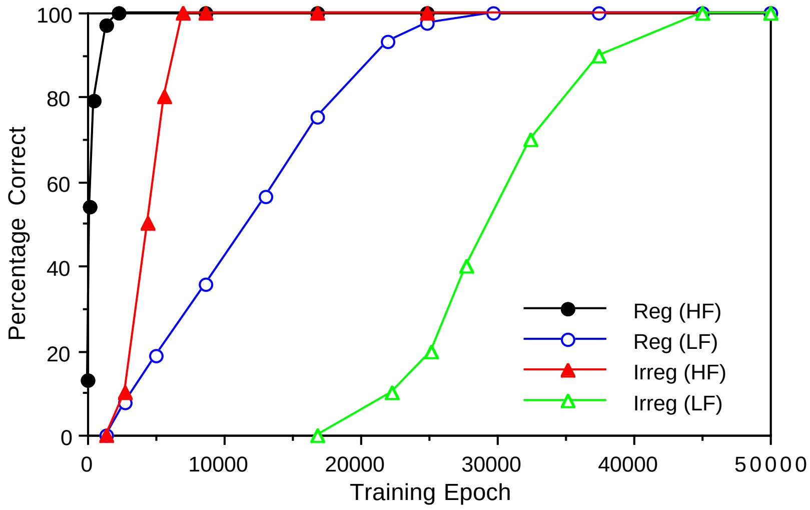

Figure 1: Regularity and frequency effects during the course of learning. HF, high frequency; LF, low frequency.

tifacts, and controls for all other factors, only single dissociations were possible. Moreover, these single dissociations were seen to be a natural consequence of the ease with which the mappings were originally learned. Plotting the patterns of activation feeding into the output units revealed why this should be the case. Each form of damage results in these activations either drifting in a random direction or falling to zero. For every output unit there will be some correct response threshold, and the items that are learned first during training will end up furthest past the thresholds when the training is stopped. They will consequently tend to be the last to cross over again and result in output errors during increased degrees of damage. We thus we get clear dissociations with the regulars more robust than frequency matched irregulars, and high frequency items more robust than regularity matched low frequency items. Figure 2 shows this explicitly for the network of Figure 1.

These basic effects extend easily to more realistic models, for example, surface dyslexia in the reading model of Bullinaria (1997). Here we not only successfully simulate the

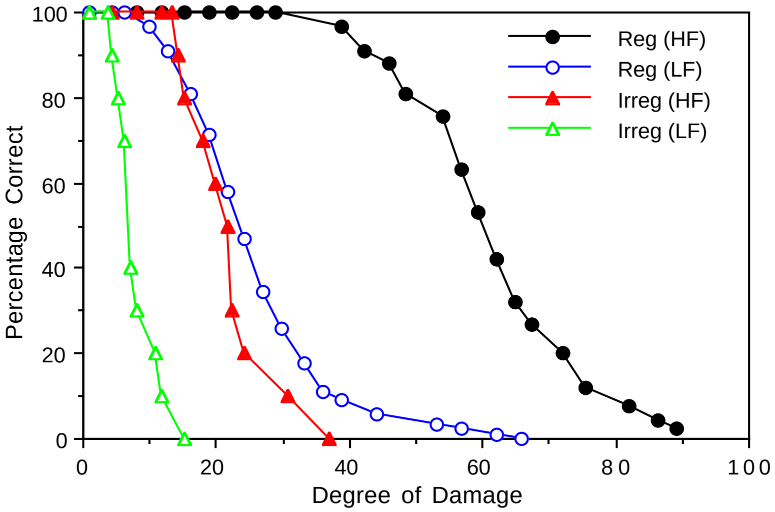

Figure 2: Regularity and frequency effects with increasing degrees of network damage. HF, high frequency; LF, low frequency.

relative error proportions for the various word categories (i.e. regular/irregular, high/low frequency), but also the types of errors that are produced. The closest threshold to an irregularly pronounced letter will be that of the regular pronunciation, and hence the errors will be predominantly regularizations of the lowest frequency irregular items, exactly as is observed in acquired human surface dyslexia.

Figures 1 and 2 also reveal what is behind a potential source of confusion. Bullinaria and Chater (1995) argued that network lesions would always result in single dissociations with the regular items more robust. Marchman (1993), however, studied models of past tense production and seemingly found dissociations with the irregular items more robust than the regulars. It is easy to see from the figures that sufficiently high frequency irregulars can be more robust than regulars. The English language has evolved to leave the irregulars with much higher frequencies than the regulars – otherwise they would have been lost from the language. Marchman built this into her models, with the expected consequences. This illustrates how important it is to control for all confounding factors when describing dissociations and drawing conclusions from them. Lavric et al. (2001) provide a review of the issues involved in understanding the dissociations of verb morphology.

It is also evident from Figure 2 that, if the frequencies and regularities are carefully matched, the performance on the high frequency irregulars can cross that of the lower frequency regulars. Initially there is a dissociation with better performance on the irregulars, and later the opposite dissociation. Such a "double dissociation" is a form of *resource artifact* that is well known not to imply underlying modularity (Shallice, 1988, p234). Patterns of deficits of this type are actually rather easily obtainable in neural network models. Devlin et al. (1998) present an interesting example involving a connectionist account of category specific semantic deficits. The finer grain of detail that connectionist modelling affords here allows explicit accounts of human deficits that would be difficult to accommodate in older "box and arrow" models.

The general point one can make about single, fully distributed, sub-systems is that some items are naturally learned more quickly and more accurately than others, and the effects of subsequent network damage follow automatically from these patterns of learning. There are actually many factors, in addition to regularity and frequency, that can cause differing learning and damage rates. We can explore them all in a similar manner, and use them in models of neuropsychological data in the same way. Consistency and Neighbourhood Density are the most closely related to regularity and are commonly found in models of language tasks such as reading and spelling (e.g. Plaut et al., 1996; Bullinaria, 1997). Representation Sparseness or Pattern Strength are often used to distinguish between concrete and abstract semantics, as in models of deep dyslexia (e.g. Plaut and Shallice, 1993). Correlation, Redundancy and Dimensionality are commonly used in models to distinguish the semantics of natural things versus artifacts, as in models of category specific semantic deficits (e.g. Devlin et al., 1998). At some level of description, all these factors act in a similar manner to frequency and regularity, and their effects can easily be confounded. Which we use will depend on exactly what we are attempting to model, but clearly, if we want to make claims about neuropsychological deficits involving one of them, we need to be careful to control for all the others.

Following brain damage, patients often show a rapid improvement in performance. This is important to connectionist modellers for two reasons. First, if relearning occurs automatically and quickly in patients, then we need to be sure that the same effects are observed in our models, and that we are comparing patient and model data at equivalent stages of the relearning process. Second, our models may be of assistance in formulating appropriate remedial strategies for brain damaged patients (Plaut, 1996). Since learning and damage have the same underlying regularity and frequency effects, relearning from the original training data is unlikely to reverse this pattern, indeed it is likely to enhance it (Bullinaria and Chater, 1995). However, if some rehabilitation regime is employed that involves a very different set of training examples to that of the original learning process, it is possible for different results to arise (Plaut, 1996). Here the models may be used to predict or refine appropriate relearning strategies, and the patients' responses can be used to validate our models.

SMALL SCALE ARTIFACTS

One should never forget that modelling massively parallel brain processes by simulating neural networks on serial computers is only rendered feasible by abstracting the essential details and scaling down the size of the networks. It is clearly important not to take the abstraction and scaling process so far that we miss important fundamental properties of the systems we are modelling, or introduce features that are nothing but small scale artifacts. The damage curves of Figure 2 are relatively smooth because our network has many more hidden units and connections than are actually required to perform the given mappings, and individual connections or hidden units make only small contributions to the network's outputs. For smaller networks, however, the effect of individual damage contributions can be large enough to produce wildly fluctuating performance on individual items, and this can result in dissociations in arbitrary directions. Often these small scale artifacts are sufficient to produce convincing looking double dissociations (Shallice, 1988, p254). Bullinaria and Chater (1995) showed that as we scale up to larger networks, the processing becomes more distributed and apparent double dissociations dissolve into single dissociations.

Our modelling endeavours would be much easier if some independent procedure could determine when networks were sufficiently distributed to obtain reliable results. In effect, we need to make sure that individual processing units are not acting as "modules" in their own right, and the obvious way to do this is by checking that all the individual contributions feeding into to each output unit are small compared to the total. In this case, many such lost contributions must conspire to result in an output change large enough to be deemed an error. This is the brain-like resilience to damage often known as *graceful degradation*. Fortunately, this distribution of information processing tends to occur automatically simply by supplying the network with a sufficiently large number of hidden units. However, in general, it seems that we really do need very many hidden units to avoid small scale artifacts – many times the minimal number required to learn the given task (Bullinaria, 1999). So, what can be done if limited computational resources make this impossible? Obviously, after removing a random subset of the hidden units or connections, the number of contributions will be reduced by some factor, but, in large fully distributed networks, the mean contribution will not change much and so the total contribution after damage is simply reduced by the same factor. Clearly we can achieve the same result by simply globally scaling all the weights by the same factor. In smaller networks, this equivalence breaks down because the means tend to suffer relatively large random fluctuations during damage. However, since global weight scaling does not suffer from such random fluctuations, it can be used to simulate a smoothed form of lesioning and give a reasonable approximation in small networks to what will happen in more realistic networks. Alternatively, if one wants to claim that each hidden unit corresponds to a number of real neurons, then the weight scaling can be regarded as removing a fraction of those neurons.

LESIONING ATTRACTOR NETWORKS

Many successful models of human performance and their associated neuropsychological deficits have been based on attractor networks (see COMPUTING WITH ATTRACTORS) rather than simple feed-forward networks. These are recurrent networks which develop *attractors* to appropriate patterns of activity, i.e. they have points in the *state space* of output activations to which the network settles. Lesions to this type of network can alter the settling behaviour by distorting or shifting the *basins of attraction*. Here the errors correspond to the network settling into the wrong attractor, rather than an output unit activation failing to reach a particular threshold. But still, the resilience to damage follows directly from how the particular items were originally learned.

One of the earliest applications of attractor networks to neuropsychology was the Mozer and Behrmann (1990) model of *neglect dyslexia*. But, perhaps the most successful models of this type are the Plaut and Shallice (1993) models of *deep dyslexia*, which were extensions of earlier work by Hinton and Shallice (1991) showing how both visual and semantic errors could arise from a single lesion. These attractor networks mapped from orthography to semantics via a layer of hidden units, and then from semantics to phonology via another set of hidden units, with layers of *clean-up units* at the semantics and phonology levels. One particular model was trained on 40 words, using back-propagation through time, until it settled into the correct semantics and phonology when presented with each orthography. Lesions at two different locations in the trained network were then found to produce a double dissociation between concrete and abstract word reading, where concreteness was coded as the proportion of activated semantic micro-features. Specifically, removal of orthographic to hidden layer connections resulted in preferential loss of abstract word reading, whereas removal of connections to the semantic clean-up units primarily impaired performance on the concrete words. Although the two damage locations do not constitute modules in the conventional sense, it is not difficult to understand how they contribute to the processing of the two word types to different degrees, and give opposite dissociations when damaged. It is simply a consequence of the sparser representations of the abstract words making less use of the semantic clean-up mechanism, and depending more on the direct connections, than the richer representations of the concrete words (Plaut and Shallice, 1993). This does not conflict with the claim of Bullinaria and Chater (1995) that only single dissociations are possible. The robustness of each location in the attractor network is fully consistent with the general discussion above, and the only disagreement concerns the appropriateness of using the word "module" to describe the two damage locations. As Plaut himself points out (Plaut, 1995), one of the problems when discussing "modularity" is that different authors use different definitions of the term. This is fine, but to avoid confusion one should be careful to quote the definitions along with the conclusions.

DISCUSSION

This article has covered the basic issues and complications involved in lesioning neural network models to provide accounts of neuropsychological deficits, and has provided pointers to a range of representative case studies. It seems clear that, despite all the abstractions and simplifications involved, connectionist modelling has a lot to offer in fleshing out the details of, or even replacing, earlier "box and arrow" models to provide a more complete picture of cognitive processing. The resulting enhanced models and the new field of connectionist neuropsychology are not only producing good accounts of existing empirical data, but are also already beginning to suggest more appropriate experimental investigations for further fine tuning of these models, and an ethical approach for exploring potential remedial actions for neuropsychological patients.

REFERENCES

Bullinaria, J.A., 1997. Modelling reading, spelling and past tense learning with artificial neural networks, Brain and Language, 59: 236-266.

Bullinaria, J.A., 1999. Connectionist dissociations, confounding factors and modularity, in Connectionist Models in Cognitive Neuroscience (D. Heinke, G.W. Humphreys and A. Olsen, eds), London: Springer, pp. 52-63.

Bullinaria, J.A. and Chater, N., 1995. Connectionist modelling: Implications for cognitive neuropsychology, Language and Cognitive Processes, 10: 227-264. *

Devlin, J.T., Gonnerman, L.M., Andersen, E.S. and Seidenberg, M.S., 1998. Category-specific semantic deficits in focal and widespread brain damage: A computational account, Journal of Cognitive Neuroscience, 10: 77-94.

Farah, M.J., 1994. Neuropsychological inference with an interactive brain: A critique of the locality assumption, Behavioral and Brain Sciences, 17: 43-104.

Hinton, G.E. and Shallice, T., 1991. Lesioning an attractor network: Investigations of acquired dyslexia, Psychological Review, 98: 74-95.

Lavric, A., Pizzagalli, D., Forstmeir, S. and Rippon, G., 2001. Mapping dissociations in verb morphology, Trends in Cognitive Science, 5: 301-308.

Marchman, V.A., 1993. Constraints on plasticity in a connectionist model of the English past tense, Journal of Cognitive Neuroscience, 5: 215-234.

Mozer, M.C. and Behrmann, M., 1990. On the interaction of selective attention and lexical knowledge: A connectionist account of neglect dyslexia, Journal of Cognitive Neuroscience, 2: 96-123.

Plaut, D.C., 1995. Double dissociation without modularity: Evidence from connectionist neuropsychology, Journal of Clinical and Experimental Neuropsychology, 17: 291-321. *

Plaut, D.C., 1996. Relearning after damage in connectionist networks: Towards a theory of rehabilitation, Brain And Language, 52: 25-82.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. and Patterson, K.E., 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains, Psychological Review, 103: 56-115.

Plaut, D.C. and Shallice, T., 1993. Deep dyslexia: A case study of connectionist neuropsychology, Cognitive Neuropsychology, 10: 377-500. *

Shallice, T., 1988. From Neuropsychology to Mental Structure, Cambridge: Cambridge University Press. *

Small, S.L., 1991. Focal and diffuse lesions in cognitive models, Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Erlbaum, pp. 85-90.