Neural Networks, Free Association, and Errors

Dr. Leslie Smith
Centre for Cognitive and Computational Neuroscience
Departments of Computing Science and Psychology
University of Stirling

Email: lss@uk.ac.stir.cs Tel: 0786 467435

March 1993.

Summary:

A speculative discussion of the form a neural network might have if it was to display some of the characteristics Freud describes in his Psychology of Errors, and of free association.

1. Introduction.

Neural Networks were “originally aimed more towards modelling networks of real neurons in the brain” [1] at a cellular or cell assembly level, rather than at producing interesting computational results from a highly parallel system, currently the aim of most of those interested in Neural Networks from a computational point of view. Such networks have been used to model a large number of cognitive systems, in reading, vision, and other tasks, usually with a view to understanding these systems better by making a model for them. Although the metaphor of the Von Neumann computer, with its digital memory and stored program have been very influential in the description of the brain since the 1950’s [2], neural networks clearly fit better at the microstructure level. Thus neural networks are now more often used for cognitive simulations since they may be able to tell us something at the neurophysiological or neuropsychological level.

This discussion paper aims to extend this realm to two specific phenomena discussed by Freud [3]: specifically, we aim to consider what type of neural network might be able to exhibit some of the effects seen in free association and some elements of his “Psychology of Errors”. Even to start this involves some large assumptions, assumptions which while they are reasonable, cannot currently be justified by scientific evidence of which I am aware. The primary assumptions are:

a: That trains of thought can be identified with changes of state inside the brain
b: That the elements which are changing state are either neurons or clusters of neurons.

Given these assumptions, one can start to consider how trains of thought might be represented in an artificial neural network for free association, and how slips of the tongue might come about, and what these might suggest about the architecture (and constituent elements) of such a network.

2. What aspects of mental activity would we hope to model?

For free association, we need to consider thought. Defining thought is notoriously slippery. We all think, and though we often agree, we all think differently. Nonetheless, though brains do vary on a global scale, their overall structure and microstructure is very much constant, so that it seems reasonable to presume that mechanisms of thought do not vary from person to person. Further, the boundary between thought and memory is not easily definable: many thoughts are recollections, which might be termed memories. Memory has many different forms: sensory, autobiographical, episodic, to name but a few [4]. It is important to remember that the usual (computer-scientist's) view of memory is a very specific idealised abstraction: human memory does not take the form of fixed recollectable values. We shall restrict our interest to such transient recollections as might be exhibited during free association. This is an important restriction, since it means we shall not consider either logical or creative thought.

What are the primary characteristics of such thoughts?

Firstly, they are fleeting: that is, their nature is transient. It is virtually impossible to maintain a thought in a fixed form for an extended period of time! Secondly, they are, or are directly based on, some previous experience, whether of an external (sensory) or internal (previous thought) nature. Thirdly, although they occur in sequences ("trains of thought"), they do not repeat themselves: people do not (generally) get stuck in indefinitely repeating trains of thought.

For errors and slips of the tongue, we consider that the net translates the person's volition into signals to another subsystem which generates the speech. We are thus able to avoid difficult questions (such as the nature of volition), and, indeed, the net itself becomes relatively a simple transformer of instantaneous input vectors into sequences of control signals.

3. What form of net might be appropriate?

For free association, and since neural nets themselves do not think, we must first consider what aspects of the net might be identified with thought. Only then can we go on to consider the nature of the network itself. For errors, we consider, as above, that the net is an interface between two subsystems.

3.1 Questions of identification.

The candidates for identification with thought (given neural nets of the usual form, with units which have activation levels and outputs, and synapses which have weights) are all the different elements of the states of the whole network, such as output values of units or sets of units, activation levels of units, or sets of units, synaptic contributions to unit activations, or any mixture of these. This is similar to

Martindale's description of consciousness as corresponding to "the set of cognitive units in sensory, perceptual and conceptual analyzers that are activated above some threshold at any given moment" [5]; however, we would suggest that this is not consciousness, but what one is conscious of, and that thought is some subset of this. Specifically, if one wishes to model the fleeting nature of thought, one should be interested not in the instantaneous value of these, but in the pattern of these values over some period of time.

Since we are interested in recollections as trains of thought, it is worth briefly considering how we might consider them to become stored in the net in the first place. In his book "Laws of Form", [6], the author notes that "... the world we know is constructed in order (and in such a way as to be able) to see itself": from a personal point of view, we perceive the world, and our thoughts, and always remember them (i.e. our perceptions of them), at least for a short time. Thus, if a neural network analogy is pursued, some sort of adaptation takes place in response to all that is perceived. Of course, not everything is remembered for a long period of time: it is the assertion of psycho-analysis [7] that many memories are repressed, and the technique of free association can be used to attempt to cause them to be recalled.

In addition, psycho-analysis has the concept of the unconscious, in which a great many memories are stored, not directly available to the conscious mind. It is worth considering what one might identify these with in a neural net context. Neural nets store their memories in the matrix of weights (and possibly synaptic delays [8]) between the units. These hold memories (and indeed all information about what behaviour the net may have) in an implicit form. This looks to be a good candidate for such an identification since it contains all possible memories (which may be simple vector states, or more complex trajectories, depending on the structure of the net itself) in an implicit form. Given a suitable architecture, this could be similar to consciousness as a searchlight picking out one of the many possible trajectories in a network. Recollection of such trajectories may be a good model for iconic or echoic memory [9]. Note that I am not suggesting that the synaptic weight and delay matrix is the form of the unconscious, merely that if one wishes to model thought using neural networks, then these may be identified with the unconscious.

4 Questions of architecture.

Given that we desire to produce a network which has states or sequences of states identifiable with thought or image recollection, or with words to be output, what can be said about the architecture of the net itself? We consider first what might be required for the network to display some of the characteristics for free association, then consider slips of the tongue.

4.1 Nets for Free Association.

If we consider that the net is given an initial impulse (the key for the free association), which causes it to traverse a sequence of states, then clearly, the net needs to be recurrent, in that it displays behaviour over time, rather than settling into a fixed output as a feed-forward net (even one with internal time delays) would. We can additionally make certain inferences about the type of recurrent net: it may not be a symmetric Hopfield net, since these have only point or 2-cycle attractor states [10]. However, if the symmetry condition is relaxed, or if noise is permitted, then much more complex attractors are possible.

It seems reasonable to further assume that the input does not go to all the units in the net: in the human brain, connectivity precludes this. Whether we consider the output to be from all the units, or from some subset of them is unimportant, unless we are considering training algorithms. If we consider the net to hold episodic autobiographical memories, then these might be recollected by ensuring that the system adapts so as to be able to reach these states (or sequences of states) representing each in turn e.g. as in [11]. Clearly, the net will be exposed to a great number of sequences of episodes, each represented by some pattern on the input units, and causing some pattern on all the units. Such a training algorithm would make patterns of sequences of activations which had occurred more likely to occur again. This would certainly involve modification of synaptic strengths and possibly of delays as well.

In recall, we should note the absence of cyclic sets of states: this implies that the attractors are 'strange': this could be because of the existence of chaotic attractors in the system [12], or because noise makes recurrence of exact state sequences unlikely. Additionally, the pattern of interconnection strengths and delays will give rise to many possible episodic sequences. Similarity between sequences of states (either thoughts or trains of thought) may cause some of the episodes to become confused with each other: that is, their trajectories merge to form some trajectory made up of parts of each. In this way, some parts of certain sequences may become unreachable, causing the loss of certain episode sequences. These might only be recalled if the system was a put into a very specific (and perhaps unexpected) starting state, such as might occur in free association.

4.2 Nets for Emulating Errors.

Nets representing the brain's capacity for 'slips of the tongue' differ in that they are subsystems which stand between a system generating volition and a system actually generating the sound itself. In this case, we are not dealing with memory, but with control. In the slips that Freud refers to, there is competing volition: the result of this competing volition is not that one or other wins (as is indeed usually the case, the case where no slip of the tongue takes place), but that the output contains elements from both competitors. Networks for generating sequences of motor actions have been developed [13], and there is much in common between this and what would be required for generating sequences of words.

If one considers that the network generates sequences of states (which come out as sequences of words) and that these sequences of states are coded in a distributed fashion, accessed through an appropriate impulse (momentary vector) input, then variation in this input during the generation of the sequence could lead to errors. One candidate network is a modified 3-layer feedforward network, in which some of the output is fed back to the hidden layer. If such a network was trained to produce a sequence of output for a fixed input, altering the input would certainly generate errors. The errors Freud discusses are of more than one form, ranging from direct replacement of a single word, to the compaction of a phrase into a single word. Considerable experimentation would be required to find out the exact nature of faults generable in this way from the different types of net which generate sequence outputs: this would need both psychophysical work (to describe the nature of the errors produced statistically) and experimental work with nets. Nonetheless, it seems reasonable to assume that similar effects can be displayed by nets.

5. Conclusions

What this speculative paper has proposed is a mechanism for some well-known effects. It has not been simulated, primarily because there are too many free variables to allow a proper empirical investigation. On the other hand, it is known that nets can be trained to go from state to state to state [11][14], and the phenomena of mixing of memories is well known in Hopfield nets. That the unconscious may be considered as the (implicit) content of the interconnections between neurons has implications for capacity: it is limited only by the capacity of the whole network, and , since we are considering that recollections consist of sequences of states, this may be very large. What we have not done is to differentiate between the unconscious and the possible memories of the system: indeed, it is not clear that such a distinction should be made.

References.

[1] Hertz J., Krogh A., Palmer R.G. *Introduction to the Theory of Neural Computation*, Addison Wesley, 1991, p2
[2] Baddeley A. *Human Memory*, Laurence Erlbaum Associates, 1990, p11
[3] Freud S., *Introductory Lectures on Psycho-Analysis*, George Allen and Unwin, 1922.
[4] Baddeley A. op cit, Chapter 3
[5] Martindale C., *Cognitive psychology: a neural network approach*, Brooks/Cole publishing, 1991.
[6] Spencer-Brown G., *Laws of Form*, Notes to Chapter 12, George Allen and Unwin, 1969.
[7] Freud op cit p248
[8] Amit D.J., *Modelling Brain Function*, Cambridge University Press, 1989, Chapter 5
[9] Baddeley op cit p14 and p28.
[10] Amit D.J. op cit, p169
[11] Weisbuch G., *Complex systems dynamics*, Addison Wesley, 1991, p72-74.
[12] Renals S., *Chaos in Neural Networks*, in Lecture Notes in Computer Science 412, ed Almeida L.B., Wellekens C.J.
[13] see, e.g. Part VIII of *Neural Information Processing Systems 3,* ed Lippmann R.P., Moody J.E., Touretzky D.S.
[14] Wang D., Arbib M.A., *Complex Temporal Sequence Learning based on Short-term Memory,* Proceedings of the IEEE, 78, 9, September 1990.