CSC321: Neural Networks

# Lecture 19: Simulating Brain Damage

Geoffrey Hinton

## The Effects of Brain Damage

- Performance deteriorates in some unexpected ways when the brain is damaged. This can tell us a lot about how information is processed.
  - Damage to the right hemisphere can cause neglect of the left half of visual space and a lack of a sense of ownership of body parts.
  - Damage to parts of the infero-temporal cortex can prevent face recognition.
  - Damage to other areas can destroy the perception of color or of motion.
- Before brain scans, the performance deficits caused by physical damage were the main way to localize functions in the human brain
  - recording from human brain cells is not usually allowed (but it can give surprising results!).

## Acquired dyslexia

- Occasionally, damage to the brain of an adult causes bizarre reading deficits
  - Surface dyslexics can read regular nonsense words like “mave” but mispronounce irregular words like “yacht”.
  - Deep dyslexics cannot deal with nonsense words at all. They can read “yacht” correctly sometimes but sometimes misread “yacht” as “boat”. They are also much better at concrete nouns than at abstract nouns (like “peace”) or verbs.

## Some weird effects

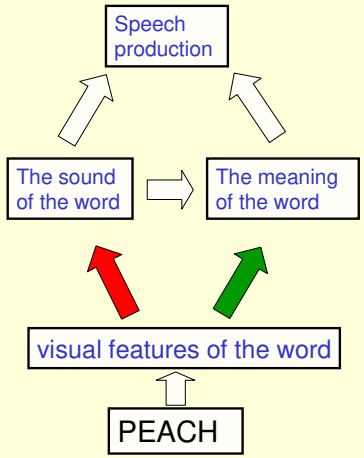
PEACH ⇒ “apricot”

SYMPATHY ⇒ “orchestra”

HAM ⇒ “food”

## The dual route theory of reading

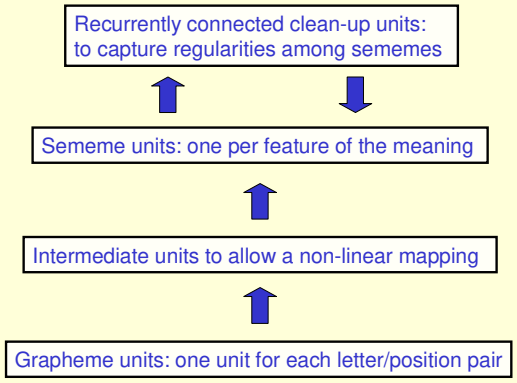
- Marshall and Newcombe proposed that there are two routes that can be separately damaged.
  - Deep dyslexics have lost the phonological route and may also have damage to the semantic route.
- But there are consistent peculiarities that are hard to explain this way.




## An advantage of neural network models

- Until quite recently, nearly all the models of information processing that psychologists used were inadequate for explaining the effects of damage.
  - Either they were symbol processing models that had no direct relationship to hardware
  - Or they were just vague descriptions that could not actually do the information processing.
- There is no easy way to make detailed predictions of how hardware damage will affect performance in models of this type.
- Neural net models have several advantages:
  - They actually do the required information processing rather than just describing it .
  - They can be physically damaged and the effects can be observed.

## A model of the semantic route

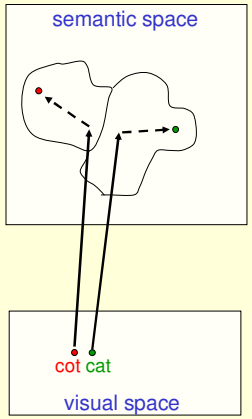



## What the network learns

- We used recurrent back-propagation for six time steps with the sememe vector as the desired output for the last 3 time steps.
  - The network creates semantic attractors.
  - Each word meaning is a point in semantic space and has its own basin of attraction.
  - Damage to the sememe or clean-up units can change the boundaries of the attractors.
    - This explains semantic errors. Meanings fall into a neighboring attractor.
  - Damage to the bottom-up input can change the initial conditions for the attractors.
    - This explains why early damage can cause semantic errors.

## Sharing the work between attractors and the bottom-up pathway

- Feed-forward nets prefer to produce similar outputs for similar inputs.
  - Attractors can be used to make life easy for the feed-forward pathway.
- Damaging attractors can cause errors involving visually similar words.
  - This explains why patients who make semantic errors always make some visual errors as well.
  - It also explains why errors that are both visually and semantically similar are particularly frequent.




## Can very different meanings be next to each other in semantic space?

Take two random binary vectors in a high-dimensional space.

- Their scalar product depends on the fraction of the bits that are on in each vector.
- The average of two random binary vectors is much closer to them than to other random binary vectors!

1 1 0 0 0 0 1 1 0 1 0

0 1 1 0 0 1 1 0 1 0 0

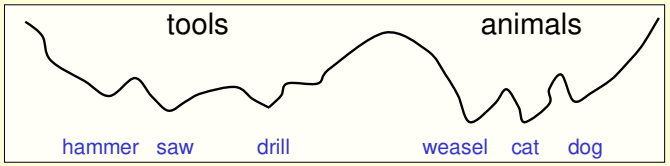
$$\mathbf{a}'\mathbf{c} \approx N\, p_{\mathbf{a}} p_{\mathbf{c}} = N\left(\frac{p_{\mathbf{a}} p_{\mathbf{c}}}{2} + \frac{p_{\mathbf{a}} p_{\mathbf{c}}}{2}\right)$$

$$\mathbf{a}'\left(\frac{\mathbf{a}}{2} + \frac{\mathbf{b}}{2}\right) \approx N\left(\frac{p_{\mathbf{a}}}{2} + \frac{p_{\mathbf{a}} p_{\mathbf{b}}}{2}\right)$$

$$p_{\mathbf{a}} > p_{\mathbf{a}} p_{\mathbf{a}}$$

## Fractal attractors?

- The semantic space may have structure at several different scales.
  - Large-scale structure represents broad categories.
  - Fine-scale structure represents finer distinctions
- Severe damage could blur out all the fine structure
  - Meanings get cleaned-up to the meaning of the broad category. Complex features get lost.
  - May explain regression to childhood in senility?




## The advantage of concrete words

- We assume that concrete nouns have many more semantic features than abstract words.
  - So they can benefit much more from the semantic clean-up. The right meaning can be recovered even if the bottom-up input is severely damaged.
- But severe damage to the semantic part of the network will hurt concrete nouns more because they are more reliant on the clean-up.