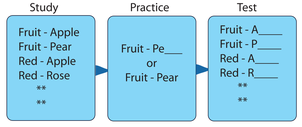
A Neural Network Model of Retrieval-Induced Forgetting

Ehren Newman & Kenneth Norman

Abstract

Michael Anderson and his colleagues have demonstrated that people forget information that is similar to information they are actively trying to recall in paired associate tasks. This retrieval-induced forgetting effect only occurs as a consequence of recall attempts using partial cues (as opposed to simply re-presenting studied pairs), and the effect is cue-independent - forgetting is observed regardless of how you try to access the "forgotten" information (see Levy & Anderson, 2002, for a review). We present a neural network model of these findings that uses a novel "early-late phase" (ELP) learning rule proposed by O'Reilly & McClelland. This rule contrasts activation states early in processing (when multiple representations are activated in a bottom-up fashion by the input stimulus) and later in processing (when activity is more strongly affected by attractor dynamics and competition between representations). Representations that are activated early but not late in the settle process are degraded. We show how, with this rule in place, the retrieval-induced forgetting effects described by Anderson arise as a natural consequence of competition between representations during processing. We then review how this implementation relates to other mechanisms that have been suggested to account for retrieval-induced forgetting.

The Task (to-be-modeled):



Three Phases:

Study:
Paired associates are presented one at a time

Practice:
Partial -
A two-letter word stem of a study list item is presented
Participants complete the partial word
or
Full -
A paired associate from study list is re-presented

Test:
One-letter word stems from studied items are presented
Participants are asked to complete the partial words

Behavioral Data (from Levy & Anderson, 2002)

Test Item	Level of recall at test relative to baseline: After Partial Practice (Fruit - Pe___)	After Full Practice (Fruit - Pear)
Fruit - Apple	WORSE	SAME
Fruit - Pear	BETTER	BETTER
Red - Apple	WORSE	SAME
Red - Rose	SAME	SAME

In other words, if given a partial practice -

•Recall of the practiced item improves (Fruit-Pear)

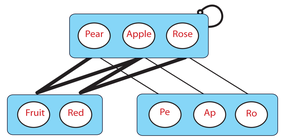
•Recall of items similar the to practiced item gets worse (Fruit-Apple), in a cue-independent fashion (Red-Apple)

and if given a full practice -

•Recall of the practiced item improves (Fruit-Pear)

•Other items are unaffected

A simplified model



Output Units
Limited amount of total activity due to feedback inhibition

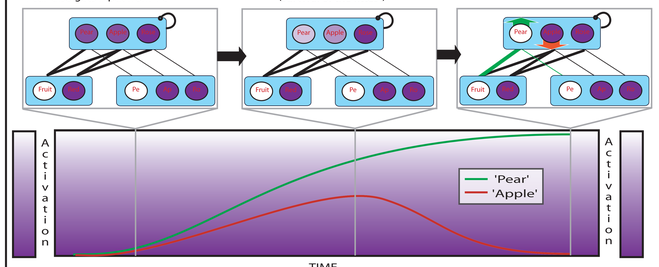
Connections
Line width indicates connection strength

Input Units

What is this a model of?

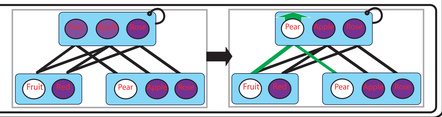
Competitive Dynamics

- 'Fruit' and 'Pe' are activated in the inputs
- 'Pear' and 'Fruit' receive input from 'Fruit'
- 'Pear' also gets input from 'Pe'
- 'Pear' competes with 'Apple'
- 'Pear' squeezes 'Apple' out of output (survival of the fittest)
- The network updates its weights



What about full practice?

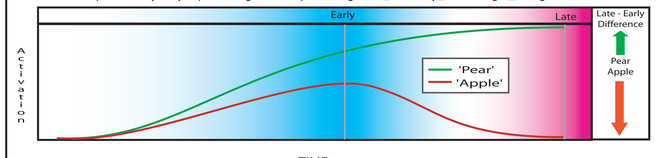
- 'Pear' becomes activated quickly
- 'Pear' keeps 'apple' from activating
- Weights will update



How does the network adjust it weights?

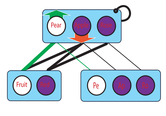
Early Late Phase Rule* & Hebbian Learning

- Compare activity early in processing vs late in processing (Late_act - Early_act) = weight_change

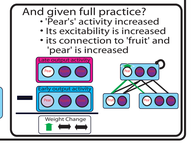


Early-Late Phase:
- The activity of 'pear' increased
  - Its excitability is increased
- The activity of 'apple' decreased
  - Its excitability is decreased

Hebbian Learning:
- 'Pear' connection to 'fruit' is increased
- 'Pear' connection to 'pe' is increased



And given full practice?
- 'Pear's' activity increased
- Its excitability is increased
- Its connection to 'fruit' and 'pear' is increased



* O'Reilly & McClelland (Personal Communication)

How does this model produce the data?

Given a partial practice -

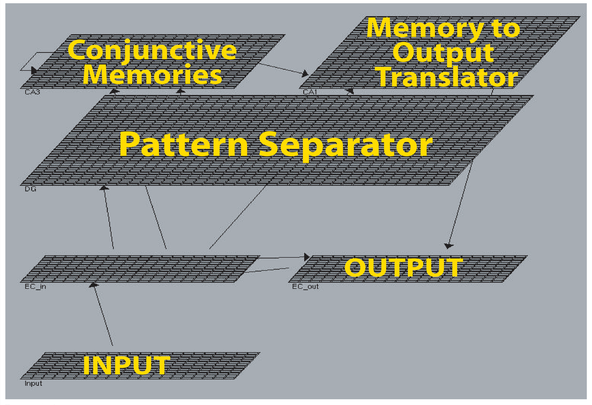
- Recall of the practiced item improves (Fruit-Pear)
  - 'Pear' was active at the end of processing
  - Late - Early difference was positive
  - Its excitability was increased
  - It became easier to activate at test
- Recall of items similar the to practiced item gets worse (Fruit-Apple), in a cue-independent fashion (Red-Apple)
  - 'Apple' became active but then turned off
  - Late - Early difference was negative
  - Its excitability was decreased
  - It became harder to activate at test

Given a full practice -

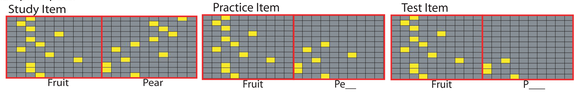
- Recall of the practiced item improves (Fruit-Pear)
  - Same as when given partial practice
- Other items are unaffected
  - 'Apple' and 'Rose' never became active
  - Late - Early difference was zero
  - Excitability remained unaffected

Where's the proof?

- To test this theory we used a hippocampal model
  - Allows for rapid learning of novel pairs
  - Has been well studied (Norman & O'Reilly, in press)



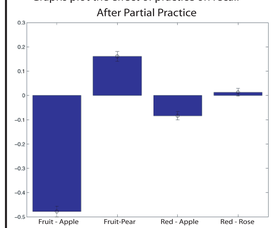
Example Inputs:



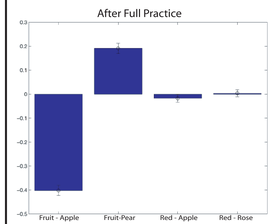
Procedure:
- Study -
  - Present all paired associates to input layer
  - Network learns to output same pattern at EC_out
- Practice -
  - Present two-letter practice item ('Fruit - Pe__') once
  - Allow network to settle twice, with increasing limitations on total network activation
    - 'Early phase' - Increased number of units allowed to be active
    - 'Late phase' - Decreased number of units allowed to be active
  - Network learns according to the early-late phase and Hebbian learning rules
- Test -
  - Present one-letter word-stems ('Fruit - P_____') for all studied items
  - Scored network on its ability to regenerate full pattern at EC_out

Simulation results:

Graphs plot the effect of practice on recall



- Recall of 'Fruit - Apple' worse than control item
- Recall of 'Fruit - Pear' better than control item
- Recall of 'Red - Apple' worse than control item
- Recall of 'Red - Rose' not affected



- Recall of 'Fruit - Apple' worse than control item , better than when given partial practice
- Recall of 'Fruit - Pear' better than control item
- Recall of 'Red - Apple' not affected
- Recall of 'Red - Rose' not affected

Graphs represent the recall of labeled items relative to a set of control items that was also presented at study. The control set was structured analogously to the experiment set, except that the analog of "fruit - pear" was not practiced.

Discussion:

- Competitive Dynamics and an Early-Late phase rule account for…
  - Practice related improvement of 'Fruit - Pear'
  - Cue independent decrement of 'Red - Apple' depends on practice protocol
  - Decrement of 'Fruit - Apple' depends on practice protocol
- Does not require 'directed inhibition'
  - Lateral inhibitory competition generates cue independent decrementing
- Does not account for null effect of full practice on 'Fruit - Apple'
  - This is due to blocking, 'Pear' interferes with recall of 'Apple'
  - All neural networks will show this
  - Evidence that there may be a counteracting force
  - Magnitude shown here is exaggerated by 'small vocabulary' of network
- The intrinsic excitability must be changed, not incoming weights

Issues remaining to be addressed:

- Counteracting the blocking effect of practicing 'Pear'
- Two snap-shot Early-Late phase rule is biologically questionable
- Role of prefrontal cortex in biasing competition

Possible Solution?

- A continuous Hebbian learning rule
  - Update weights at each time step
  - Would not require "memorization" of snap-shots
  - Would punish unharmonious activity (activity that isn't in a stable attractor)

References

Levy, B.J. & Anderson, M.C. (2002). Inhibitory processes and the control of memory retrieval. *TRENDS in Cognitive Sciences*. 6(7) , 299-305.

Norman, K.A. & O'Reilly, R.C. (in press). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*.