Further Predictions from a Neural Network Model of Retrieval Induced Forgetting

Kenneth Norman, Ehren Newman, & Greg Detre

Department of Psychology and Center for the Study of Brain, Mind, and Behavior, Princeton University

Overview

We have recently developed a neural network learning algorithm that accounts for how we strengthen weak memories and punish competing memories. The algorithm accomplishes these goals by raising and lowering inhibition, and learning based on the resulting changes in activation.

Here, we use the model to address several puzzles from the domain of retrieval-induced forgetting (Anderson, 2003), including: How does the strength of the competing memories affect the amount of competitor punishment? How does the strength of the target (to-be-retrieved) memory affect competitor punishment? Why does retrieval practice lead to more competitor punishment than simply presenting the target item? We also use the model to address non-monotonic effects of retrieval practice, whereby repeated practice first strengthens, then weakens competing memories.

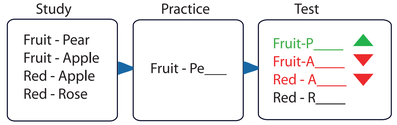
We show that the model can account for the main features of the retrieval-induced forgetting data space. More importantly, the model establishes boundary conditions on these effects, which may be useful in explaining why some findings are not always obtained.

Background: Competitors are Punished

Our original motivation for this research was to model data on competitive dynamics and memory. Across several domains, researchers have found that competitors are punished during memory retrieval.

More specifically: When a representation is activated by a retrieval cue, but that representation loses the competition to be retrieved, it suffers a lasting decrease in accessibility (on the order of hours and possibly longer).

This principle is illustrated very nicely by Michael Anderson's work on **retrieval-induced forgetting**, illustrated below (see Levy & Anderson, 2002, for a review).

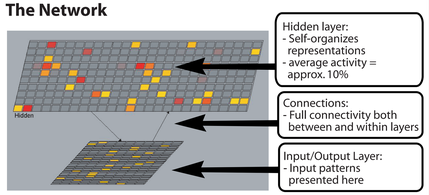


In other words, if the subject is given retrieval practice -
- Recall of the practiced item improves (Fruit-Pear)
- Recall of competitors gets worse (Fruit-Apple), in a *cue-independent* fashion (Red-Apple)

What are the brain mechanisms of competitor punishment? Existing accounts focus on the role of prefrontal cortex in resolving competition. These accounts help explain the *dynamics* of competition do not explain why competition has *lasting* effects on memory.

The goal of this research is to identify **basic neural learning mechanisms** that can account for competitor punishment effects, as well as other learning phenomena.

The Network



In the brain, inhibitory interneurons limit the spread of excitatory activity.

We capture this dynamic with a **k-winners-take-all inhibition rule**. This rule adjusts inhibition so the *k* units in each layer that receive the most input are active, and other units are inactive (O'Reilly & Munakata, 2000).

The *k* parameter is set to match the size of a single input pattern, so (given normal inhibition) one full memory can be active at once.

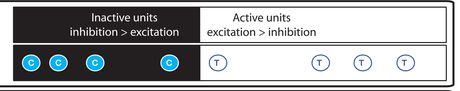
Oscillation based learning algorithm (Norman, Newman, Detre, & Polyn, submitted)

Two goals for learning: 1. Identify and weaken competing memories. 2. Identify and strengthen weak parts of target (to-be-learned) memory

The learning algorithm achieves these goals by oscillating inhibition, and learning based on the resulting changes in activation.

1. Present the target pattern to the network

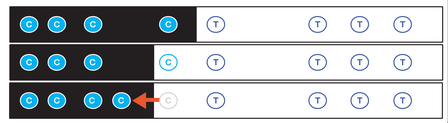
The k-winners-take-all inhibition rule (k = 4 in this example) prevents activation from spreading to non-target units.



2. Identify competitors by lowering inhibition

When inhibition is lowered, non-target units that are receiving strong input become active.

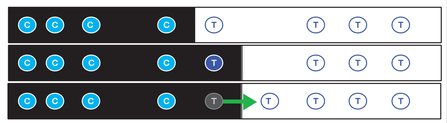
3. Weaken units that turn on (i.e., strong competitors), by reducing weights from other active units



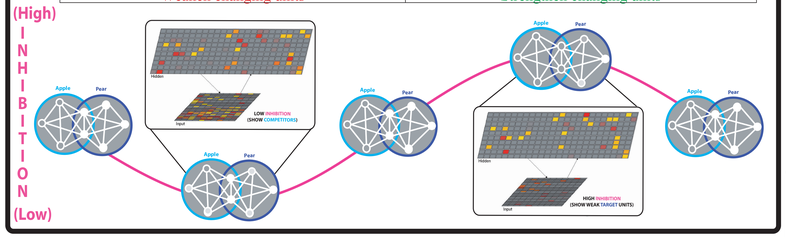
4. Identify weak parts of the target memory by raising inhibition.

This acts as a stress test on the target memory. Poorly supported target units turn off and well-supported units remain active.

5. Strengthen units that turn off (i.e., weak targets), by increasing weights from other active units



INHIBITION APPLIED TO INPUT LAYER (CONSTRAINT)				
Normal	**Low**	**Normal**	**High**	**Normal**
Target on	Target on	Target struggles to stay on	Target comes back on	
Competitor allowed on	Competitor forced off	Competitor off	Competitor off	
Weaken changing units		Strengthen changing units		



General Simulation Methods

1. Inhibition is oscillated in the input layer (but not the hidden layer).
2. Oscillation is achieved by adding an oscillating component to the value of inhibition computed by k-winners-take-all
3. One full oscillation (normal-low-normal-high-normal) per trial
4. Compute weight change using Contrastive Hebbian Learning (CHL)

The CHL rule (Movellan, 1990) compares a more desirable activation state (the *plus* state) to a less desirable activation state (the *minus* state), and adapts weights to increase the strength of the plus state, relative to the minus state:
weight change $= x_i^+ y_j^+ - x_i^- y_j^-$ (x_i = sending neuron, y_j = receiving neuron)

We apply CHL to successive states of network activity (time *t* and *t + 1*).

When inhibition is moving ***away*** from its normal level, the state at time *t* is closer to the target than the state at time *t + 1*, so time *t* serves as the plus state:
weight change $= x_i^t y_j^t - x_i^{t+1} y_j^{t+1}$

When inhibition is moving ***toward*** its normal level, the state at time *t + 1* is closer to the target than the state at time *t*, so time *t + 1* serves as the plus state:
weight change $= x_i^{t+1} y_j^{t+1} - x_i^t y_j^t$

Retrieval-Induced Forgetting Simulation Methods

1. **Generate three patterns**
   - A target pattern (presented at study and practice)
   - Two competitor patterns (presented at study but not practice)
   - Target and competitors all share 4/8 units
2. **Train the network on these patterns**
   - Present the network with complete patterns
   - Update weights after each pattern
3. **Pretest the network's ability to recall studied patterns**
   - Present 1/8 units of the pattern as cue.
   - Learning is turned off at test.
4. **Allow network to practice target pattern**
   - Partial practice: 4/8 units presented
5. **Test the network's ability to recall studied patterns again**
   - Compare to pretest performance to calculate practice effect
   - Test cue and practice cue do not overlap, so the test cue qualifies as an indepedent cue

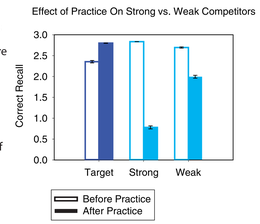
Effect of Competitor Strength

Anderson, Bjork & Bjork (1994) found that strong competitors are punished more than weak competitors (e.g., practicing fruit-pe hurts recall of common fruits like apple more than rare fruits like kiwi).

We ran a simulation manipulating competitor strength:
At study, the strong competitor was presented 50 times.
The weak competitor was presented 20 times.
The target was presented 10 times.

We were able to replicate the Anderson et al. (1994) finding of a competitor strength effect.

The figure below illustrates the amount of net input received by target and competitor units (assuming that recall is successful). In this case, the k-winners-take-all rule will set inhibition so the weakest target unit is above threshold, and the strongest competitor unit is below threshold.



Lowering inhibition causes competitor units to activate. Strong competitor units are closer to threshold than weak competitor units, so they are more likely to activate (and be punished).

Boundary conditions: If the inhibitory oscillation is too small, neither competitor will be activated (or punished). If the inhibitory oscillation is too large, both competitors will be fully activated (and punished equally).

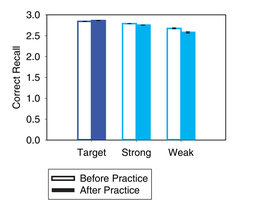
Effect of Target Strength

Anderson et al. (1994) found that target strength did not interact with competitor punishment. To explore the effect of target strength, we ran a simulation where we presented the target 160 times at study (vs. 10 times in the previous simulation)

Contrary to Anderson et al., we found that increasing target strength eliminates the competitor punishment effect.

This can be explained in terms of the figure below:

1. Increasing target strength increases the average amount of net input received by target units (strong memory = strong collateral support).
2. The k-winners-take-all inhibition rule places the inhibitory threshold a fixed proportion of the distance between the top *k* units and other units Thus, increasing target strength pulls the inhibitory threshold to the right.
3. Because competitors are far below the inhibitory threshold, they don't activate when inhibition is lowered, so they aren't punished.



Boundary conditions: We had to do a lot of strengthening to eliminate competitor punishment (e.g., 80 presentations weren't enough). In actual experiments, it may be difficult to obtain strength differences that are large enough to eliminate competitor punishment.

Retrieval Practice vs. Repeated Presentation

Several studies (e.g., Anderson & Shivde, in preparation) have found that presenting the full target pattern during the practice phase (instead of a partial pattern) eliminates competitor punishment effects.

To explore this, we ran a simulation where the full target pattern was presented at practice.

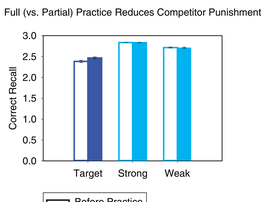
We were able to replicate the finding that full practice eliminates competitor punishment effects.

The explanation here is the same as the explanation for the target strength effect above:

1. Presenting the full cue increases the amount of net input received by target units.
2. When targets are far above competitors, neither targets nor competitors will be close to the inhibitory threshold.
3. As a result, both target and competitor units are relatively unaffected by the inhibitory oscillation, and relatively little weight change occurs (for targets or competitors).

Boundary conditions: According to the model, there is nothing special about partial practice *per se*. What really matters is the amount of net input received by target units, relative to competitors.

This view implies that, if the target representation is very weak, competitor punishment effects should be observed, even in the full practice condition.



Note that the model shows a **generation effect**: More target strengthening in the partial (vs. full) practice condition.

However, Anderson typically does not find this effect.

Why doesn't Anderson find a generation effect? Misrecall of the target pattern during practice may drag down performance in the partial practice condition (Anderson, personal communication)

Nonmonotonic Retrieval Practice Effects and Prefrontal Cortex

Johnson & Anderson (2004) found that repeatedly trying to retrieve the subordinate verb meaning of a homograph (e.g., "prune") first boosts, then lowers the accessibility of the dominant noun meaning.

How do we account for this nonmonotonic pattern?

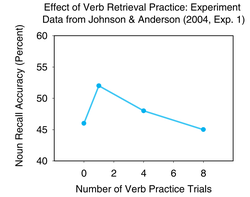
Key assumption #1: Prefrontal cortex (PFC) intervention is needed to ensure that the (weaker) verb meaning is retrieved.

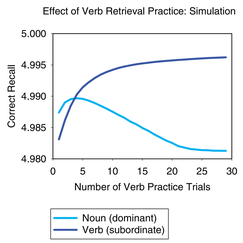
However: If the verb meaning always wins, memory for the noun meaning will monotonically decrease.

To explain the nonmontonic recall results, we have to posit that sometimes the verb meaning wins and sometimes it loses.

Key assumption #2: During verb retrieval, PFC does not engage right away. This gives the network a chance to retrieve the noun meaning. Eventually PFC kicks in & forces the network to retrieve the verb meaning.

We ran a simulation to explore whether these assumptions would be adequate to account for the Johnson & Anderson results.





Methods:

2 study patterns, a noun and a verb. Patterns consist of 3 parts:

- a lexical representation that is shared by the noun and verb
- a part-of-speech tag (different for noun and verb)
- a semantic representation (different for noun and verb)

Training: Noun is presented for 60 trials, verb is presented for 20 trials.

Practice: On each practice trial,
- present shared lexical representation and weakly activate "verb tag"
- oscillate inhibition for 1 full cycle
- then, PFC comes on and provides extra input to the "verb tag"
- oscillate inhibition for 1 full cycle

Results: The model replicates the nonmonotonic pattern.

During the first few practice trials, the noun is rehearsed during the initial oscillation (prior to PFC activation) and the verb is rehearsed after PFC comes on. There are two important consequences of this interleaved pattern of rehersal:

1. Differentiation of the noun and verb representations: Hidden-layer overlap decreases from 76% to 71%
2. The verb starts to catch up in strength to the noun. Both patterns improve (because they both are rehearsed) but the verb improves more because of ceiling effects.

Both of these changes work to reduce the competitive advantage of the noun representation: Because of differentiation, the noun matches the verb cue less well, and ceiling effects reduce the noun's strength advantage.

Eventually, the verb starts winning the competition on the initial oscillation (prior to PFC coming on). Once this happens, the noun's strength starts to decline.

Conclusions

Using a fixed set of underlying parameters, the model can simulate a wide range of retrieval-induced-forgetting results, including competitor strength effects, effects of retrieval practice vs. repeated presentation, and the nonmonotonic competitor punishment effects found by Johnson & Anderson (2004).

The model's most important contribution is to characterize **boundary conditions** on these effects.

In addition to the retrieval-induced forgetting results described here, we are currently using the model to simulate:

- Familiarity discrimination: Familiarity = the size of the dip in activation when inhibition increases above baseline
- Learning during sleep: How learning based on inhibitory oscillations during REM can strengthen stored memories & help protect them from interference (Norman & Perotte, in preparation)

We are also exploring functional properties of the learning algorithm. What is its capacity for memorizing patterns, and how well does it perform relative to other learning rules?

References

Anderson, M.C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory & Language, 49*, 415-445.

Anderson, M.C., Bjork, R.A., & Bjork, E.L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 5*, 1063-1087.

Anderson, M.C. & Shivde, G.S. (in preparation). Inhibition in episodic memory: Evidence for a retrieval-specific mechanism.

Johnson, S., & Anderson, M.C. (2004). The role of inhibitory control in forgetting semantic knowledge. *Psychological Science, 15*, 448-453.

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

O'Reilly, R.C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience.* Cambridge, MA: MIT Press.

Movellan, J.R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D.S. Tourtezky, G.E. Hinton, & T.J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* (pp. 10-17). San Mateo, CA: Morgan Kaufman.

Norman, K.A., Newman, E.L., Detre, G., & Polyn, S.M. (submitted). How inhibitory oscillations can train neural networks and punish competitors.

ELN was supported by an NIH Training Grant in Quantitative Neuroscience (MH65214) awarded to Princeton University

For an electronic copy of the poster, visit http://compmem.princeton.edu/publications.html