Will it forget-me-net?

Overcoming catastrophic forgetting in backpropagation neural networks

Loes Bazen, Abdallah El Ali, Iris Groen, Elisa Hermanides, Wouter Kool, David Neville & Kendall Rattner
Supervisor: Jaap Murre

Abstract

Various methods to overcome the catastrophic interference effect in back-propagation networks are directly compared on a simple learning task. Interleaved learning delivered the best results: the pattern "McClelland" was not catastrophically forgotten after learning the pattern "soup". These results indicate that catastrophic forgetting can be overcome by interleaved learning.

Research Problem

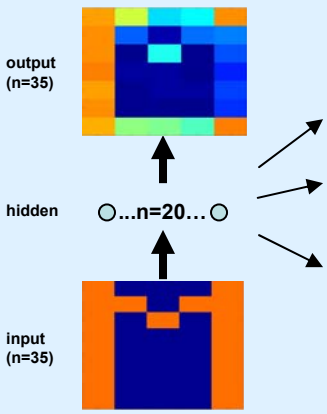
With backpropagation, newly learned patterns tend to obliterate existing representations learnt previously[1] (catastrophic interference).

This stems from a problem in connectionist models of memory, namely, the stability-plasticity problem[2]. In this study three methods have been implemented to solve this problem:

- activation function manipulation[1]
- inducing competition between hidden layers
- interleaved learning[3]

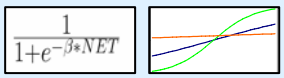
Methods

Example Network



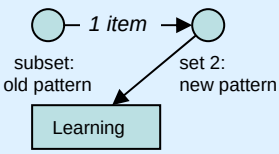
Remedies

Method 1: Adjust β in activation function:



Method 2: Sparse encoding

Method 3: Interleaved learning

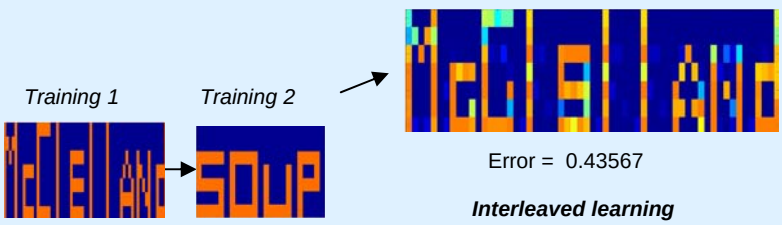


Design:

- Training the network on "McClelland", followed by training on "soup", thereafter performance was measured on the "McClelland" set.
- Other parameters (momentum, etc.) were held constant.

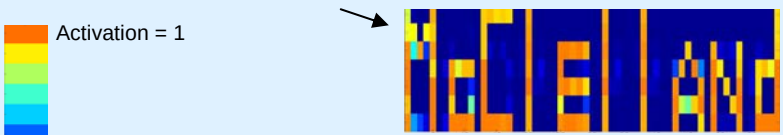
Results

Standard learning



Error = 0.43567

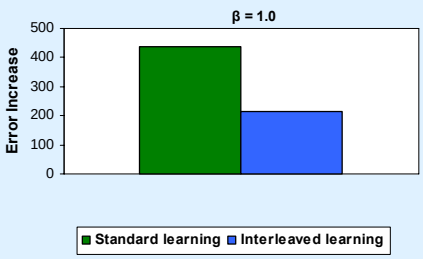
Interleaved learning



Error = 0.21169

- Adjusting β over {0.5, 3}: did not reduce catastrophic interference.
- Using sparse encoding: did not aid in alleviating catastrophic interference.

Error increase with different types of learning



Conclusions

+ Interleaved learning greatly alleviated the effects of catastrophic interference.

+ Interleaved learning is a better approximation of human memory, therefore it strengthens the psychological validity of the results.

- Other remedies (beta value adjustment and sparse encoding) did not reduce catastrophic interference.

References

1. Ratcliff, R. (1990) Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions *Psychol. Rev.* 97, 285-308

2. French, R. (1999) Catastrophic Forgetting in Connectionist Networks. *TICS,* 3, 128-135.

3. McClelland, J. (1995) A Connectionist Perspective on Knowledge and Development: *Developing Cognitive Competence,* 157-204

Universiteit van Amsterdam
Cognitive Science Center Amsterdam
Summerschool 2008