Standard feedforward neural networks with backprop cannot support cognitive superposition

Arno Vanegdom[1*], Nikolay Nikolaev[1], Max Garagnani[12]

[1]Department of Computing
Goldsmiths, University of London
London, UK

[2] Brain Language Lab, Freie Universität Berlin,
Department of Philosophy and Humanities
Berlin, Germany

* Corresponding Author
Email: avane001@gold.ac.uk

Modern artificial neural networks (NNs) have shown remarkable human-like performance in several domains [1,2,3]. However, their ability to support an arguably fundamental aspect of human intelligence, superposition (our ability to maintain simultaneously active in working memory any two previously acquired concepts) has not been thoroughly investigated. Previous studies looking at whether NNs generally suffer from the so-called "superposition catastrophe" [4] reported partly contradictory results [5,6] and, crucially, only assessed a network's ability to superpose items that had been already co-activated *during* training.

To ascertain whether, and understand why, NNs are a suitable neural substrate to implement superposition, we took a 'standard' feedforward neural network [7] and assessed its ability to co-activate previously acquired internal representations. Specifically, using backpropagation we trained a one-hidden-layer NN to associate each of five orthogonal binary input vectors to corresponding output binary patterns. After the input-output pairs had been learned to criterion, we tested the network's ability to produce the correct result (i.e. the superposition, or inclusive 'OR', of the corresponding output patterns) when any two of the five learned input vectors were co-activated. We repeated the experiment by orthogonally varying the type of patterns (dense vs. sparse) used as input/output. We found that the networks systematically failed to produce the correct output in all trials and experiments. The average ratio of correct output (percentage of correctly activated '1s' in the output pattern) was 38%. To address the "why" question, we ran a second set of experiments in which we decreased the network's size: careful examination of the final weight configuration enabled us to identify the causes for the NNs' inability to support superposition as residing in the inherently distributed nature of the internal representations and 'greediness' of the learning algorithm.

The present results indicate that – and provide an explanation as to why – classical feedforward neural networks trained with backpropagation develop internal representations that are not suitable to support superposition. We submit that this result has potential implications in the field of human-like artificial cognitive systems, as it suggests that something other than standard fully connected NNs is required to implement a key cognitive operation our brain can easily perform.

References

[1] Shakirov, V. V., Solovyeva, K. P., & Dunin-Barkowski, W. L. (2018). Review of state-of-the-art in deep learning artificial intelligence. *Optical memory and neural networks*, *27*(2), 65-80.

[2] Silver, D., Schrittwieser, J., Simonyan, K. *et al.* Mastering the game of Go without human knowledge. *Nature* 550, 354–359 (2017).

[3] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

[4] Von der Malsburg, C. (1986). Am I thinking assemblies? In G. Palm & A. Aertsen (Eds.), Brain theory (pp. 161-176). Berlin, Germany: Springer

[5] Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological Review*, *121*(2), 248-261.

[6] Martin, N. D. (2021). *Selectivity in neural networks* (Doctoral dissertation, University of Bristol).

[7] Haykin, S.O. (2009). *Neural networks and learning machines.* Pearson Prentice Hall, New Jersey. 3rd ed.