Meaningful Representations Prevent Catastrophic Interference

Jordi Bieger, Ida Sprinkhuizen-Kuyper and Iris van Rooij

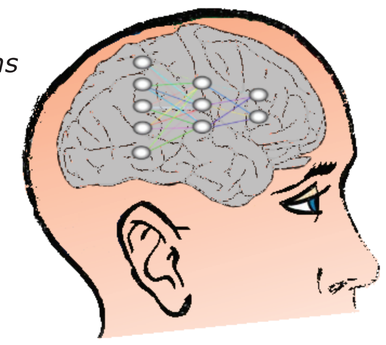
Donders Institute for brain, cognition and behaviour

Radboud University Nijmegen

Think fast!

A ball is soaring through the air in your general direction. What to do? Catch it? Hit it? Dodge it? A robot using an *artificial neural network* (ANN) inspired by the human brain can be taught to do any of these things. The correct response does not depend solely on characteristics of the situation like the ball's size and trajectory, but also on other factors like what game is being played. These factors are often encoded in fairly arbitrary manners.

Our brains consist of hundreds of neurons connected by hundreds of trillions of connections. Artificial Neural Networks are learning systems that are based on our knowledge of the brain.



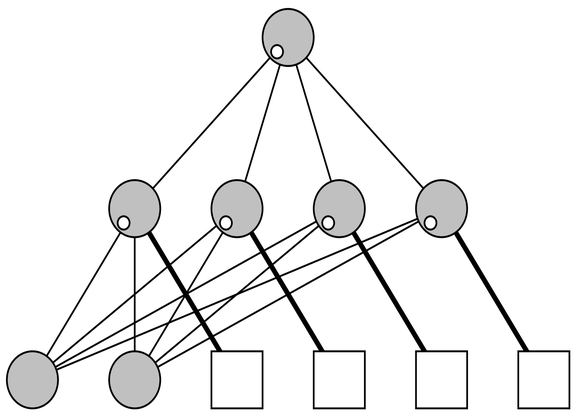
Catastrophic Interference

Learning to kick a ball does not impair a human's ability to hit it with a tennis racket. However, when arbitrary representations are used in ANNs to describe the situation, learning to kick will obliterate any knowledge it might have had about catching. This phenomenon is called *catastrophic interference*.

Catastrophic interference

Static Meaningful Representation Learning

ANNs learn by adjusting the weights between the nodes in the network. If these weights are static (i.e. cannot change), such a network would never forget or learn anything. New tasks can nevertheless be learned by using *representation nodes* that not only *identify* which task should be performed, but actually *represent* that task in the context of the knowledge that a network already possesses. This is accomplished by using Tani et al.'s *parametric bias* nodes [1] as task describing *meaningful representation nodes (MRNs)*. The values of these nodes are determined through back propagation training on the new task in the context of a network trained on another task.



A neural network with meaningful representation nodes (square). The bold lines always have a value of 1.

The algorithm

Initialization

1. Construct a normal ANN for learning one of the required tasks
2. Add a number of MRNs that will represent a task
3. Initial Knowledge Acquisition
   (a) Train the entire network (including MRNs) on one task T_1 using backprop
   (b) Store the values of the MRNs along with T_1
4. Fix all the weights in the network

Learn a new task T_x (Novelty Learning)

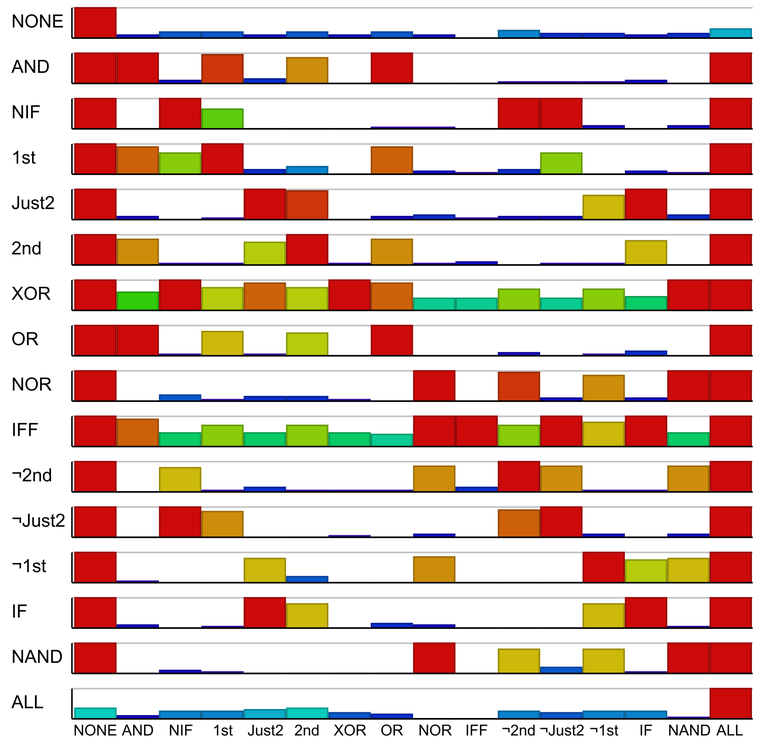
1. Reset the values of the MRNs
2. Train the MRN values to represent T_x
3. Store the values of the MRNs along with T_x

Perform a learned task T_x (Knowledge Application)

1. Clamp the representation values stored with T_x to the MRNs
2. Clamp the desired task input values to the other input nodes
3. Read the output values of the network

Tasks and performance

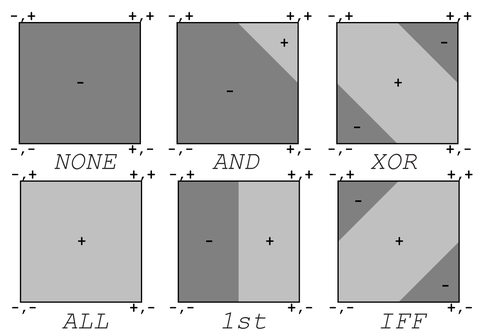
The tasks used in this experiment each had two binary inputs and one binary output. Performance was positively affected by increases in the number of MRNs. However, not all task combinations could easily be learned together.



The likelyhood that a task on the x-axis can be learned in the context of a task on the y-axis.

Difficulty

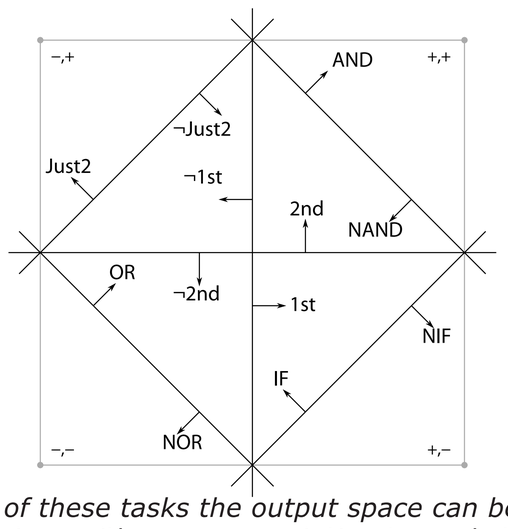
Some tasks are more difficult than others. Our experiments show that prior knowledge of a hard task often enables the network to learn easier tasks, but initial knowledge of easy tasks is usually not a sufficient basis to learn more difficult things.



Shows the desired output as a function of the first (x-axis) and second input (y-axis). The more regions there are, the higher the input-output complexity of the task. After ALL and NONE few tasks could be learned, while XOR and IFF provided much better bases for learning.

Similarity

Similarly, learning a task that is similar to the network's prior knowledge is often much easier than learning dissimilar tasks, i.e. learning of OR is virtually guaranteed when prior knowledge of AND exists.



For each of these tasks the output space can be divided into two regions with an arrow pointing towards the positive part. Tasks whose arrows point in roughly the same direction have similar input-output relations and can be learned together.

Conclusion

Using meaningful representations enables ANNs to mimic the human brain's ability to sequentially learn multiple tasks without suffering from catastrophic interference.

References

[1] J. Tani, M. Ito & Y. Sugita. Self-organization of distributedly represented multiple layer schemata in a mirror system, 2004.