The Problem: How to build an interpretable Neural Network

The proposed solution: Map Neural Network structure to a Knowledge Graph structure

The algorithm

Algorithm 1: Semantic Artificial Neural Networks Construction

1. Create empty Neural Network Graph G
2. for all input features $x_i \in D$ do
3. Map $x_i \in D$ to concepts or attributes $c_i \in O$
4. Add corresponding $c_i \in O$ into G
5. end for
6. for all output features $y_j \in D$ do
7. Map $y_j \in D$ to concepts or attributes $c_j \in O$
8. Add corresponding $c_j \in O$ into G
9. end for
10. while $c_i \in G$ not connected to $c_j \in G$ do
11. for all nodes $c_j \in G$ do
12. Find concept(s) $c_k \in O$ connected to node(s) $c_j \in G$
13. Add node(s) $c_k$ in G
14. Add arc(s) connecting $c_j$, $c_k$
15. end for
16. end while
17. return Graph G

The result: A Neural Network with labeled hidden layer nodes.
Each level can be interpreted by means of regression over the previous layer

Evaluation:

<table>
<thead>
<tr>
<th>Dataset/Metric</th>
<th>Multilayer perceptron</th>
<th>Semantic Artificial Neuronal Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification/Accuracy</td>
<td>0.8154</td>
<td>0.8154</td>
</tr>
<tr>
<td>Regression/Correlation coefficient</td>
<td>0.689</td>
<td>0.697</td>
</tr>
</tbody>
</table>

Semantic Artificial Neural Networks are interpretable and have performance comparable to that of dense and more complex Neural Networks.