The Interpretation of Neural Network QSAR Models Using Weights and Biases

R. Guha, D.T. Stanton and P.C. Jurs

Department of Chemistry
Pennsylvania State University
and
Miami Valley Laboratories
Procter & Gamble

March 15th, 2005
Outline

1 Background
   • Why Do We Need An Interpretation?
   • Some Aspects of Interpretability

2 Interpreting a Neural Network

3 Summary
Isn’t a Prediction Enough?

- Predictive models are good for screening purposes
- To understand *why* a compound is active we need an interpretation
- Interpretation is one way to approach the inverse QSAR problem
- Interpretability depends on modeling technique & descriptors involved
Interpretability generally involves a trade off with accuracy.

Linear regression models are amenable to interpretation, but are often not very accurate.

Neural networks are black boxes, but are often more accurate.

Some techniques lie in between (random forests)
Types of Interpretation

**Broad Interpretation**
- Essentially describes which descriptors are important
- Good for understanding which descriptors to focus on
- Based on randomization

**Detailed Interpretation**
- Describes how the property (activity) relates to the descriptor
- Gives us conclusions like:
  - **high** value of DESC leads to **low** values of activity
- Allows for a detailed understanding of the SAR in QSAR
CNN Interpretation in the Literature

- Relative importance of input neurons
- Uses the training set to develop measures of importance
- In many cases the methods depend on the nature of the network

Tickle, A.B. et al., *Intl. Conf. on Neural Networks*, **1997**, *4*, 2530-2534
Outline

1 Background

2 Interpreting a Neural Network
   - Strategy
   - Results - Boiling Point Study
   - Results - Skin Permeability Study

3 Summary
Goals

Analogy with PLS Interpretations
The method is analogous to the PLS approach for linear models which considers the linear combination coefficients for each latent variable as indicating the effect of a descriptor on the output.

Utilizing CNN weights and biases . . .
- Correlate input descriptors to network output through each hidden neuron
- Order the hidden neurons
- Consider hidden neurons as latent variables
Analogy with PLS Interpretations

The method is analogous to the PLS approach for linear models which considers the linear combination coefficients for each latent variable as indicating the effect of a descriptor on the output.

Utilizing CNN weights and biases . . .

- Correlate input descriptors to network output through each hidden neuron
- Order the hidden neurons
- Consider hidden neurons as latent variables
Some Preliminaries

We know . . .

- The transfer function is sigmoidal
  \[ O = \frac{1}{1 + \exp(-\sum w_i x_i)} \]
- We can approximate this as
  \[ O \sim \exp (w_1 x_1 + \cdots + w_n x_n) \]

This indicates . . .

- \( O \) is an increasing function of its inputs
- Output from a hidden neuron is always positive
What Do The Weights Tell Us?

The absolute values tell us . . .

The weights, \( w_i \), determine which input neuron dominates the contribution to a hidden neuron.

The signs tell us . . .

The nature of the correlation between an input to a neuron and the output from the neuron.
What Do The Weights Tell Us?

The absolute value tells us . . .

The weights, $w_i$, determine which input neuron dominates the contribution to a hidden neuron.

The signs tell us . . .

The nature of the correlation between an input to a neuron and the output from the neuron.
Effective Weights

What are they?

- As input flows from an input neuron to the output neuron it is acted on by two weights.
- The effective weight for an input neuron is thus $XY$. 
Effective Weights

What are they?
The result is that the network looks like a single connection between the input neuron and the output neuron with a weight $XY$

Effective Weight Matrix

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desc 1</td>
<td>52.41</td>
<td>29.30</td>
</tr>
<tr>
<td>Desc 2</td>
<td>37.65</td>
<td>22.14</td>
</tr>
<tr>
<td>Desc 3</td>
<td>-10.50</td>
<td>-16.85</td>
</tr>
</tbody>
</table>
Why Do We Ignore The Bias Term?

**Equipartitioning View**
- When considering effective weights via a given hidden neuron, the bias term must be partitioned.
- The simplest approach is to equipartition the bias term.
- The net result is that the same value is added to each effective weight.

**Constant Bias View**
- CNN’s exhibit the universal function approximation property.
- A sufficient condition for this is that the transfer function has a non-zero derivative at the origin.
- This implies that the bias can be taken as a constant rather than trainable weight.

Hornik, K. *Neural Networks*, 1993, 6, 1069-1072
Ordering Hidden Neurons

**Contribution of a hidden neuron ...**
- Depends on the output of the neuron
- Depends on the inputs to the neuron

**Quantifying Contributions**
- Take the column means of the effective weight matrix
- Also include bias terms for each hidden neuron
- Convert to a proportional scale for ease of use (SCV)
Validation of the Method

- Build a linear model with $N$ descriptors and interpret it
- Build a CNN model with the same descriptors and interpret it

The two interpretations should match since both models should encode similar SPR trends
### Boiling Points

#### Dataset
- 277 compounds
- Original work reported CNN models
- No interpretations
- $145K < BP < 653K$

#### Model details
- 7 descriptor OLS model
- $R^2 = 0.98, \ RMSE = 9.98 \ K$
- CNN model was 7-4-1
- $R^2 = 0.91, \ RMSE = 15.21 \ K$

Linear Interpretation

Component 1 focuses on . . .
- Size effects
- Higher molecular weight
- Longer paths

Component 2 focuses on . . .
- Hydrogen bonding ability
- Charge weighted negative surface area
- Lower hydrophobic surface area

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNSA-3</td>
<td>-0.30</td>
<td>-0.42</td>
</tr>
<tr>
<td>RSHM-1</td>
<td>0.19</td>
<td>0.77</td>
</tr>
<tr>
<td>V4P-5</td>
<td>0.48</td>
<td>-0.15</td>
</tr>
<tr>
<td>S4PC-12</td>
<td>0.28</td>
<td>-0.07</td>
</tr>
<tr>
<td>MW</td>
<td>0.49</td>
<td>-0.085</td>
</tr>
<tr>
<td>WTPT-2</td>
<td>0.48</td>
<td>-0.05</td>
</tr>
<tr>
<td>DPHS-1</td>
<td>0.26</td>
<td>-0.41</td>
</tr>
</tbody>
</table>
The most weighted descriptors are very similar to those in the OLS model.

The signs of the effective weights match those from the OLS model as well as chemical reasoning.
Skin Permeability

Dataset
- Original work reported linear models
- Measured activity was the permeability coefficient ($K_p$)
- $-5.03 < \log(K_p) < -0.85$

Model details
- 7 descriptor OLS model
- $R^2 = 0.84$, $RMSE = 0.37$ log units
- CNN model was 7-5-1
- $R^2 = 0.94$, $RMSE = 0.23$ log units

Patel, H. et al., *Chemosphere*, 2002, 48, 603-613
Linear Interpretation

**Component 1 focuses on . . .**
- Smaller size
- Lower polar surface area
- Larger hydrophobic surface area

**Component 2 focuses on . . .**
- Larger hydrophobic surface area
- Larger surface area
- Corrections for the overestimation or underestimation of some molecules in component 1

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Component 1</th>
<th>Component 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>-0.08</td>
<td>0.52</td>
</tr>
<tr>
<td>FPSA-2</td>
<td>-0.52</td>
<td>0.14</td>
</tr>
<tr>
<td>NN</td>
<td>-0.36</td>
<td>-0.03</td>
</tr>
<tr>
<td>MOLC-9</td>
<td>0.61</td>
<td>0.11</td>
</tr>
<tr>
<td>PPHS-1</td>
<td>0.03</td>
<td>0.69</td>
</tr>
<tr>
<td>WPHS-3</td>
<td>0.09</td>
<td>0.48</td>
</tr>
<tr>
<td>RNHS</td>
<td>0.46</td>
<td>-0.04</td>
</tr>
</tbody>
</table>
### CNN Interpretation - Effective Weight Matrix

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>5</th>
<th>2</th>
<th>4</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>-44.17</td>
<td>67.34</td>
<td>8.33</td>
<td>8.18</td>
<td>5.96</td>
</tr>
<tr>
<td>FPSA-2</td>
<td>-156.82</td>
<td>-10.72</td>
<td>20.85</td>
<td>-13.07</td>
<td>-92.47</td>
</tr>
<tr>
<td>NN</td>
<td>-97.81</td>
<td>2.22</td>
<td>-6.65</td>
<td>1.71</td>
<td>-12.70</td>
</tr>
<tr>
<td>MOLC-9</td>
<td>-28.85</td>
<td>17.79</td>
<td>15.40</td>
<td>-11.36</td>
<td>-1.20</td>
</tr>
<tr>
<td>PPHS-1</td>
<td>106.55</td>
<td>31.30</td>
<td>-16.76</td>
<td>-13.99</td>
<td>34.55</td>
</tr>
<tr>
<td>WPHS-3</td>
<td>-11.36</td>
<td>-14.31</td>
<td>-2.31</td>
<td>-10.01</td>
<td>54.16</td>
</tr>
<tr>
<td>RNHS</td>
<td>20.16</td>
<td>-5.89</td>
<td>-49.57</td>
<td>23.88</td>
<td>27.09</td>
</tr>
<tr>
<td>SCV</td>
<td>0.85</td>
<td>0.13</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

- The most important neuron focuses on hydrophobic & polar effects
- The next most important neuron focuses on size effects
Observations . . .

- SCV = 0.85
- Active molecules are characterized by low polar surface area and larger hydrophobic surface area
- Does not perform too well on inactive molecules
- 69,77 and 81,114 are mispredicted
Observations . . .

- SCV = 0.13
- Corrects for 69, 77 and 81, 114
- Describes larger molecules with higher hydrophobic surface area
- MOLC-9 balances the effect of MW
- Molecule 87 is underestimated
Observations . . .

- Corrects underestimation of molecule 87 by HN 2
- Further corrects for molecule 81
- Does not perform well for inactive molecules
Observations . . .

- Corrects underestimation of molecule 87 by HN 2
- Further corrects for molecule 81
- Does not perform well for inactive molecules
- Overestimation corrected by HN 1
Outline

1. Background
2. Interpreting a Neural Network
3. Summary
Caveats

- The method *linearizes* the network
- Clearly, the interpretations will lose some of the details of the encoded SPR’s

Conclusions

- CNN interpretations appear to be valid
- Discrepancies may be present if we do not select optimal descriptor subsets for the CNN model
- The method avoids complexity and uses only the weights and biases and hence does not use the training set explicitly

The method should help CNN models to be used as design tools as well as predictive tools
Caveats

- The method *linearizes* the network
- Clearly, the interpretations will lose some of the details of the encoded SPR’s

Conclusions

- CNN interpretations appear to be valid
- Discrepancies may be present if we do not select optimal descriptor subsets for the CNN model
- The method avoids complexity and uses only the weights and biases and hence does not use the training set explicitly

The method should help CNN models to be used as design tools as well as predictive tools.
Caveats
- The method *linearizes* the network
- Clearly, the interpretations will lose some of the details of the encoded SPR’s

Conclusions
- CNN interpretations appear to be valid
- Discrepancies may be present if we do not select optimal descriptor subsets for the CNN model
- The method avoids complexity and uses only the weights and biases and hence does not use the training set explicitly

The method should help CNN models to be used as design tools as well as predictive tools