Detecting Statistical Interactions from Neural Network Weights

Michael Tsang

Joint work with
Dehua Cheng, Yan Liu
Motivation: We seek assurance that a neural network learned the longitude x latitude interaction for predicting housing price.
Problem

- Can we detect statistical interactions in data by interpreting the trained weights of a multilayer perceptron (MLP)?
- The complex behavior of MLPs can be better understood.
Statistical Interaction

Statistical Interaction$^1$: Non-Additive Groupings of Variables in $F(\mathbf{x})$

For example: $F(\mathbf{x}) = \sin(x_1 + x_2 + x_3) + x_3x_4 + x_5$

$\{1,2,3\} \quad \{3,4\}$

$^1$Sorokina et al. 2008
Statistical Interaction

Statistical Interaction$^1$: Non-Additive Groupings of Variables in $F(\mathbf{x})$

For example: $F(\mathbf{x}) = \sin(x_1 + x_2 + x_3) + x_3x_4 + x_5$

\{1,2,3\} \quad \{3,4\}

$F(\mathbf{x}) = \log(x_1x_2) = \log(x_1) + \log(x_2)$

no interaction

$^1$Sorokina et al. 2008
Core Insight in Nonlinear Networks:

Assume first layer hidden units are especially good at modeling interactions

$\Rightarrow \{1,3\}$ should exist
“Neural Interaction Detection” (NID) Framework

1. Train MLP with Regularization
2. Rank Interactions by Interpreting Weights
3. Find Cutoff on the Ranking (if desired)
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

\[ \omega_i(I) = z_i^{(1)} \mu \left( \left| W_{i,I}^{(1)} \right| \right) \quad \text{for hidden unit } i \]

Approximation of Hidden Unit Influence

\[ z^{(1)} = \left| w^y \right|^\top \left| W^{(L)} \right| \cdot \left| W^{(L-1)} \right| \ldots \left| W^{(2)} \right| \]

\((x_1, x_2, ..., x_p)\)
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

$$\omega_i(I) = z_i^{(1)} \mu \left( \left| W_{i,I}^{(1)} \right| \right) \text{ for hidden unit } i$$

Approximation of Hidden Unit Influence

$$z^{(1)} = |w^y|^{T} |W^{(L)}| . |W^{(L-1)}| . . . |W^{(2)}|$$
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

\[
\omega_i(I) = z_i^{(1)} \mu \left( \left\| W^{(1)}_{i,I} \right\| \right) \text{ for hidden unit } i
\]

Approximation of Hidden Unit Influence

\[
z^{(1)} = w^y \top W^{(L)} \cdot W^{(L-1)} \ldots W^{(2)}
\]

(y)

\[
(x_1, x_2, \ldots, x_p)
\]
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

$$\omega_i(I) = z_i^{(1)} \mu \left( \| W_{i,I}^{(1)} \| \right) \text{ for hidden unit } i$$

$$\mu (\cdot) = \min (\cdot)$$

Approximation of Hidden Unit Influence

$$z^{(1)} = \left| w^y \right|^\top \left| W^{(L)} \right| \cdot \left| W^{(L-1)} \right| \ldots \left| W^{(2)} \right|$$

\((x_1, x_2, \ldots, x_p)\)
Ranking Pairwise Interactions

\[ x \times (x^2 + x) \]
Ranking Pairwise Interactions

\[ x = (x_1 \times x_2) + (x_3 \times x_4), \]
$x = (x_1 * x_2 + x_3, x_4)$
$x = (x_1 x_2 + x_3, x_4)$

Ranking Pairwise Interactions
Ranking Pairwise Interactions

\[ x_1 \quad x_2 \quad x_3 \quad x_4 \]
Ranking Pairwise Interactions

\[ x = (x_1 \times x_2) + x_3 \]
Ranking Pairwise Interactions

\[ x_1 (x_1 \times x_3 + x_3), \ldots \]
Ranking Pairwise Interactions

\[ x_1 (x_2 + x_3) + x_4 \]
Ranking Pairwise Interactions

\[ x_1, x_2, x_3, x_4 \]
Ranking Pairwise Interactions

\[ x_1 \quad x_2 \quad x_3 \quad x_4 \]
Ranking Pairwise Interactions

\[ x = (x_1 x_2 + x_3 x_4) \]

\[ x_1 \quad x_2 \quad x_3 \quad x_4 \]
Ranking Pairwise Interactions

\[ x = (x_1 x_2 + x_3, x_4) \]
Ranking Pairwise Interactions

\[ x = (x_1 \times x_2) + (x_3 \times x_4) \]
Ranking Pairwise Interactions

\[ x = (x_1 + x_2, x_3, x_4) \]
Ranking Higher-Order Interactions

\[ |w_1| > |w_2| > |w_3| > |w_4| \]
Ranking Higher-Order Interactions

\[ h_1 \]

Interactions | Strengths
\[ \{1,2\} \quad z_1 \min(|w_1|, |w_2|) \]

\[ |w_1| > |w_2| > |w_3| > |w_4| \]
Ranking Higher-Order Interactions

\[ h_1 \]
\[ x_1 \quad x_2 \quad x_3 \quad x_4 \]

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2}</td>
<td>( z_1</td>
</tr>
</tbody>
</table>

\[ |w_1| > |w_2| > |w_3| > |w_4| \]
Ranking Higher-Order Interactions

\[ h_1 \]

Interactions | Strengths
--- | ---
\{1,2\} | \( z_1 |w_2| \)
\{1,2,3\} | \( z_1 |w_3| \)

\[ |w_1| > |w_2| > |w_3| > |w_4| \]
Ranking Higher-Order Interactions

\[ x_1 \xrightarrow{} h_1 \xrightarrow{} x_2 \xrightarrow{} x_3 \xrightarrow{} x_4 \]

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2}</td>
<td>( z_1 \mid w_2 \mid )</td>
</tr>
<tr>
<td>{1,2,3}</td>
<td>( z_1 \mid w_3 \mid )</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>( z_1 \mid w_4 \mid )</td>
</tr>
</tbody>
</table>

\[ |w_1| > |w_2| > |w_3| > |w_4| \]
Ranking Higher-Order Interactions

\[ x \left( x \times x + x \right), \]

\[ w + w > w \times w, \]

\[ h_1 \]

\[ h_2 \]

\[ x_1 \]

\[ x_2 \]

\[ x_3 \]

\[ x_4 \]

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2}</td>
<td>[ z_1</td>
</tr>
<tr>
<td>{1,2,3}</td>
<td>[ z_1</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>[ z_1</td>
</tr>
<tr>
<td>{1,3}</td>
<td>[ z_2</td>
</tr>
</tbody>
</table>

\[ |w_3| > |w_1| > |w_2| > |w_4| \]
Ranking Higher-Order Interactions

\[ x \times (x \times x + x), \text{Strengths} \{1,2,3,4\} \]

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2}</td>
<td>(z_1</td>
</tr>
<tr>
<td>{1,2,3}</td>
<td>(z_1</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>(z_1</td>
</tr>
<tr>
<td>{1,3}</td>
<td>(z_2</td>
</tr>
</tbody>
</table>

\[ |w_3| > |w_1| > |w_2| > |w_4| \]
Ranking Higher-Order Interactions

\[ x_1 \rightarrow h_1 \rightarrow x_2 \rightarrow h_2 \rightarrow x_3 \rightarrow x_4 \]

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2}</td>
<td>( z_1</td>
</tr>
<tr>
<td>{1,2,3}</td>
<td>( z_1</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>( z_1</td>
</tr>
<tr>
<td>{1,3}</td>
<td>( z_2</td>
</tr>
</tbody>
</table>

\[ |w_3| > |w_1| > |w_2| > |w_4| \]
Ranking Higher-Order Interactions

\[ x = (x_1^* x_2 + x_1^* x_4 + \ldots) \]

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2}</td>
<td>( z_1</td>
</tr>
<tr>
<td>{1,2,3}</td>
<td>( z_1</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>( z_1</td>
</tr>
<tr>
<td>{1,3}</td>
<td>( z_2</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
</tr>
</tbody>
</table>
Ranking Higher-Order Interactions

\[ z_1 |w_2| \]
\[ z_1 |w_3| + z_2 |w_2| \]
\[ z_1 |w_4| + z_2 |w_4| \]
\[ z_2 |w_1| \]

[Diagram showing nodes and interactions]
<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2,3}</td>
<td>1.3421</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>0.8241</td>
</tr>
<tr>
<td>{1,2}</td>
<td>0.3415</td>
</tr>
<tr>
<td>{1,3}</td>
<td>0.2310</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Find a Cutoff on the Ranking

• Use a generalized additive model with interactions (MLP-Cutoff)

\[ c_K(x) = \sum_{i=1}^{p} g_i(x_i) + \sum_{i=1}^{K} g'_i(x') \]

\( g_i(\cdot) \): main effects
\( g'_i(\cdot) \): interactions
Test Suite of Data-Generating Functions

<table>
<thead>
<tr>
<th>$F_1(x)$</th>
<th>$\pi x_1 x_2 \sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}} \sqrt{\frac{x_7}{x_8}} - x_2 x_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_2(x)$</td>
<td>$\pi x_1 x_2 \sqrt{2</td>
</tr>
<tr>
<td>$F_3(x)$</td>
<td>$\exp</td>
</tr>
<tr>
<td>$F_4(x)$</td>
<td>$\exp</td>
</tr>
<tr>
<td>$F_5(x)$</td>
<td>$\frac{1}{1 + x_1^2 + x_2^2 + x_3^2} + \sqrt{\exp(x_4 + x_5) +</td>
</tr>
<tr>
<td>$F_6(x)$</td>
<td>$\exp(</td>
</tr>
<tr>
<td>$F_7(x)$</td>
<td>$(\arctan(x_1) + \arctan(x_2))^2 + \max(x_3 x_4 + x_6, 0) - \frac{1}{1 + (x_4 x_5 x_6 x_7 x_{10})^2} + \left(\frac{</td>
</tr>
<tr>
<td>$F_8(x)$</td>
<td>$x_1 x_2 + 2^{x_3 + x_5 + x_6} + 2^{x_3 + x_4 + x_5 + x_7} + \sin(x_7 \sin(x_8 + x_9)) + \arccos(0.9 x_{10})$</td>
</tr>
<tr>
<td>$F_9(x)$</td>
<td>$\tanh(x_1 x_2 + x_3 x_4) \sqrt{</td>
</tr>
<tr>
<td>$F_{10}(x)$</td>
<td>$\sinh(x_1 + x_2) + \arccos(\tanh(x_3 + x_5 + x_7)) + \cos(x_4 + x_5) + \sec(x_7 x_9)$</td>
</tr>
</tbody>
</table>

Complex functions are used in our evaluation
## AUC of Pairwise Interaction Strengths

<table>
<thead>
<tr>
<th></th>
<th>ANOVA$^1$</th>
<th>HierLasso$^2$</th>
<th>AG$^3$</th>
<th>NID, MLP</th>
<th>NID, MLP-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1(x)$</td>
<td>0.992</td>
<td>1.00</td>
<td>1 ± 0.0</td>
<td>0.970 ± 9.2e−3</td>
<td>0.995 ± 4.4e−3</td>
</tr>
<tr>
<td>$F_2(x)$</td>
<td>0.468</td>
<td>0.636</td>
<td>0.88 ± 1.4e−2</td>
<td>0.79 ± 3.1e−2</td>
<td>0.85 ± 3.9e−2</td>
</tr>
<tr>
<td>$F_3(x)$</td>
<td>0.657</td>
<td>0.556</td>
<td>1 ± 0.0</td>
<td>0.999 ± 2.0e−3</td>
<td>1 ± 0.0</td>
</tr>
<tr>
<td>$F_4(x)$</td>
<td>0.563</td>
<td>0.634</td>
<td>0.999 ± 1.4e−3</td>
<td>0.85 ± 6.7e−2</td>
<td>0.996 ± 4.7e−3</td>
</tr>
<tr>
<td>$F_5(x)$</td>
<td>0.544</td>
<td>0.625</td>
<td>0.67 ± 5.7e−2</td>
<td>1 ± 0.0</td>
<td>1 ± 0.0</td>
</tr>
<tr>
<td>$F_6(x)$</td>
<td>0.780</td>
<td>0.730</td>
<td>0.64 ± 1.4e−2</td>
<td>0.98 ± 6.7e−2</td>
<td>0.70 ± 4.8e−2</td>
</tr>
<tr>
<td>$F_7(x)$</td>
<td>0.726</td>
<td>0.571</td>
<td>0.81 ± 4.9e−2</td>
<td>0.84 ± 1.7e−2</td>
<td>0.82 ± 2.2e−2</td>
</tr>
<tr>
<td>$F_8(x)$</td>
<td>0.929</td>
<td>0.958</td>
<td>0.937 ± 1.4e−3</td>
<td>0.989 ± 4.4e−3</td>
<td>0.989 ± 4.5e−3</td>
</tr>
<tr>
<td>$F_9(x)$</td>
<td>0.783</td>
<td>0.681</td>
<td>0.808 ± 5.7e−3</td>
<td>0.83 ± 5.3e−2</td>
<td>0.83 ± 3.7e−2</td>
</tr>
<tr>
<td>$F_{10}(x)$</td>
<td>0.765</td>
<td>0.583</td>
<td>1 ± 0.0</td>
<td>0.995 ± 9.5e−3</td>
<td>0.99 ± 2.1e−2</td>
</tr>
<tr>
<td>average</td>
<td>0.721</td>
<td>0.698</td>
<td>0.87 ± 1.4e−2</td>
<td><em><em>0.92</em> ± 2.3e−2</em>*</td>
<td><strong>0.92 ± 1.8e−2</strong></td>
</tr>
</tbody>
</table>

$^1$Fisher 1925, $^2$Bien et al. 2013, $^3$Sorokina et al. 2008

* $F_6$ plays an important role for this result
Higher-Order Interaction Detection for Synthetic Data

\[ F_1(x) = \pi x_1 x_2 \sqrt{2x_3 - \sin^{-1}(x_4)} + \log(x_3 + x_5) - \frac{x_9}{x_{10}} \sqrt{ \frac{x_7}{x_8} - x_2 x_7 } \]

\[ F_3(x) = \exp| x_1 - x_2 | + | x_2 x_3 | - x_3^2 | x_4 | + \log( x_4^2 + x_5^2 + x_7^2 + x_8^2 ) + x_9 + \frac{1}{1 + x_{10}^2} \]

\[ F_5(x) = \frac{1}{1 + x_1^2 + x_2^2 + x_3^2} + \sqrt{\exp(x_4 + x_5) + | x_6 + x_7 | + x_8 x_9 x_{10} } \]

\[ F_7(x) = (\arctan(x_1) + \arctan(x_2))^2 + \max(x_3 x_4 + x_6, 0) - \frac{1}{1 + (x_4 x_5 x_6 x_7 x_8)^2} + \left( \frac{|x_7|}{1 + |x_9|} \right)^5 + \sum_{i=1}^{10} x_i \]
Higher-Order Interaction Detection versus Baseline

(a) Top-rank recall vs. noise

(b) Runtime in seconds

Criteria:
- AG
- NID, MLP-M
- NID, MLP
- NID, MLP-M & MLP-Cutoff
Similar detection performance at varying noise levels
Higher-Order Interaction Detection versus Baseline

Runtime is orders of magnitude times faster
Back to our housing problem
Pairwise Heat-Maps for Real-World Data

{1,2}: longitude and latitude!

Pairwise Heat-Maps for Real-World Data

{4,7}: hour and working day

Higher-Order Interaction Detection for Real-World Data

Reached the cutoff point ⇒ obtained informative interactions
Summary

- Proposed **Neural Interaction Detection** (NID) that detects interactions from neural network weights.
- NID takes *orders of magnitude less time* to obtain *similar performance* to the state-of-the-art baseline.
References


