Detecting Statistical Interactions from Neural Network Weights

M. Tsang, D. Cheng, Y. Liu – ICLR 2018
Motivation

Feedforward NNs
- Universal function approximators
- Interpretability

Introduction (1/2)
Motivation

Feedforward NNs
- Universal function approximators
- Interpretability

Main goal: detecting pairwise and high-order feature interactions in a dataset by re-interpreting weights learned by a MLP.
Motivation

➢ Applications
   • Healthcare: Drug–drug interaction (DDI), co-occurrence of a group of symptoms
   • Scientific discoveries, hypothesis validation

➢ Challenges
   • p features: Search space size - $O(2^p)$ possible interactions

➢ Contribution of NID (Neural Interaction Detection)
   • Non-linear feature interactions.
   • Invariant of order
   • Efficiency
**Definition**

**Interaction**: groups of features that have joint effects (non-additive) for predicting an outcome. \( \mathcal{I} \subseteq \{1, 2, ..., p\}, |\mathcal{I}| \geq 2 \)

**Geometric example**

**Simple examples of explicit functions**

\[
f_1(\mathbf{x}) = \sin(x_1 + x_2 + x_3) + x_3x_4 + x_5
\]

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

\[
f_2(\mathbf{x}) = \log(x_1x_2) = \log(x_1) + \log(x_2)
\]

**no interaction!**
Core Insight Feedforward NNs

Feature interactions are **created** at hidden units with non-linear activation functions.
The influences of the interactions are propagated layer-by-layer to the final output.
In general, the weights in a NN are nonzero—all features are interacting—large solution space of interactions.

- **Assume** first layer hidden units are especially good at modeling interactions
- Interaction strength.
Interaction strength

Strength $\omega_i(I)$ of an interaction, $I \subseteq [p]$ at the i-th unit in the first hidden layer

$$\omega_i(I) = z^{(1)}_i \mu(|W^{(1)}_{i,I}|)$$

1. **Interactions** created at the **first hidden layer**.
   Summarize feature weights between $l = 0$ and $l = 1$ through function $\mu$:

   $$\mu(|W^{(1)}_{i,I}|) \rightarrow \mu(\cdot) = \text{min}(\cdot)$$

2. **Influence of hidden units**: multiplication of the aggregated weight

   $$z^{(1)}_i = |w^y^T|W^{(L)}||W^{(L-1)}|...|W^{(2)}|$$

Defining the metric (1/2)
NID example

\[ |w_1| > |w_2| > |w_3| > |w_4| \]

Interactions | Strength
\[ \{x_1, x_2\} \quad \min(|w_1|, |w_2|) \cdot z_1 = |w_2| \cdot z_1 \]
NID example

\[
\begin{align*}
\{x_1, x_2\} & \quad \min(|w_1|, |w_2|)z_1 = |w_2|z_1 \\
\{x_1, x_2, x_3\} & \quad \min(|w_1|, |w_2|, |w_3|)z_1 = |w_3|z_1
\end{align*}
\]

\[|w_1| > |w_2| > |w_3| > |w_4|\]
NID example

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>{x_1, x_2}</td>
<td>\text{min}(</td>
</tr>
<tr>
<td>{x_1, x_2, x_3}</td>
<td>\text{min}(</td>
</tr>
<tr>
<td>{x_1, x_2, x_3, x_4}</td>
<td>\text{min}(</td>
</tr>
</tbody>
</table>

\[ |w_1| > |w_2| > |w_3| > |w_4| \]
Introduction/ Preliminaries (1/3)

Example (4/9)

NID example

\[
\begin{align*}
\begin{array}{c}
\text{Interactions} \\
\text{Strength}
\end{array}
\begin{array}{c}
\{x_1, x_2\} |w_2| z_1 \\
\{x_1, x_2, x_3\} |w_3| z_1 \\
\{x_1, x_2, x_3, x_4\} |w_4| z_1 \\
\{x_1, x_3\} \min(|w_1|, |w_3|) z_2 = |w_1| z_2
\end{array}
\end{align*}
\]

\[|w_3| > |w_1| > |w_2| > |w_4|\]
NID example

\[ |w_3| > |w_1| > |w_2| > |w_4| \]

**Interactions**

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>{x_1, x_2}</td>
<td>|w_2|z_1</td>
</tr>
<tr>
<td>{x_1, x_2, x_3}</td>
<td>|w_3|z_1 + \min(|w_1|, |w_2|, |w_3|)z_2</td>
</tr>
<tr>
<td>{x_1, x_2, x_3, x_4}</td>
<td>|w_4|z_1</td>
</tr>
<tr>
<td>{x_1, x_3}</td>
<td>|w_1|z_2</td>
</tr>
</tbody>
</table>
Introduction/Preliminaries (1/3)

NID example

Example (6/9)

\[
\begin{align*}
|w_3| & > |w_1| > |w_2| > |w_4| \\
\end{align*}
\]

Interactions & Strength \\
\{x_1, x_2\} & |w_2| z_1 \\
\{x_1, x_2, x_3\} & |w_3| z_1 + \min(|w_1|, |w_2|, |w_3|) z_2 \\
\{x_1, x_2, x_3, x_4\} & |w_4| z_1 + \min(|w_1|, |w_2|, |w_3|, |w_4|) z_2 \\
\{x_1, x_3\} & |w_1| z_2 \\
\]
NID example

**Example (7/9)**

### Interactions and Strength

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>{x_1, x_2}</td>
<td></td>
</tr>
<tr>
<td>{x_1, x_2, x_3}</td>
<td></td>
</tr>
<tr>
<td>{x_1, x_2, x_3, x_4}</td>
<td></td>
</tr>
<tr>
<td>{x_1, x_3}</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
NID example

Example (8/9)
NID example

Example (9/9)
NID: Neural Interaction detection

1. Train a Lasso-regularized MLP.
2. Interpret learned weights to obtain a ranking of interaction candidates.
3. Determine a cutoff for the top-K interactions.

Data often contains both

➢ statistical interactions.
➢ main effects: univariate influences of variables on an outcome variable.

• Model separately 2 simple networks: (MLP, MLP-M)
• Learn jointly with L1-regularization only on the interaction part to cancel out the main effect as much as possible
**NID: Neural Interaction detection**

1. Train a Lasso-regularized MLP.
2. Interpret learned weights to obtain a ranking of interaction candidates.
3. Determine a cutoff for the top-K interactions.

A **greedy algorithm** generates a ranking of interaction candidates
• at each hidden unit, it only considers the top-ranked interactions of every order based on their interaction strengths (set $\mu=\min(.)$).
  ➢ drastically reduces the search space of potential interactions ($O(hp)$ tests)
  ➢ but still considers all orders.
**NID: Neural Interaction detection**

1. Train a Lasso-regularized MLP.
2. Interpret learned weights to obtain a ranking of interaction candidates.
3. Determine a cutoff for the top-K interactions.

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

- captures main effects
- captures the interactions

Gradually **add top-ranked interactions** to the GAM, increasing K, until GAM performance on a validation set plateaus.
Experiments

Tasks:
• Pairwise interaction detection - Synthetic functions

The interaction strengths shown are normally high at the cross-marks!
Experiments

Tasks:

- **Pairwise interaction detection** - Real data

California Housing Prices

\{1,2\}: longitude and latitude!
Experiments

 Tasks:
  • **Pairwise interaction detection** - Real data

Number of Bike-share Users
{4,7}: hour and working day!
Experiments

- Higher order interaction detection - Synthetic functions

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

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

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

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

- **Tasks:**
  - Higher order interaction detection - Synthetic functions

Adding the first interaction significantly reduces RMSE.
Take - home points

- Neural networks for a traditional statistical problem!
- Accurately detect general types of interactions
- Without assuming any explicit interaction order
- Without searching an exponential solution space of interaction candidates.
Thank you!