Neural Interaction Detector - Detecting High-order Interactions via Deep Neural Networks

Yan Liu

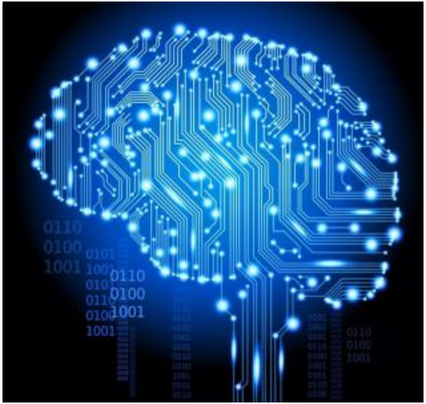
Associate Professor
Computer Science Department
University of Southern California

Director, Machine Learning Center (MASCLE)

Physics Informed Machine Learning

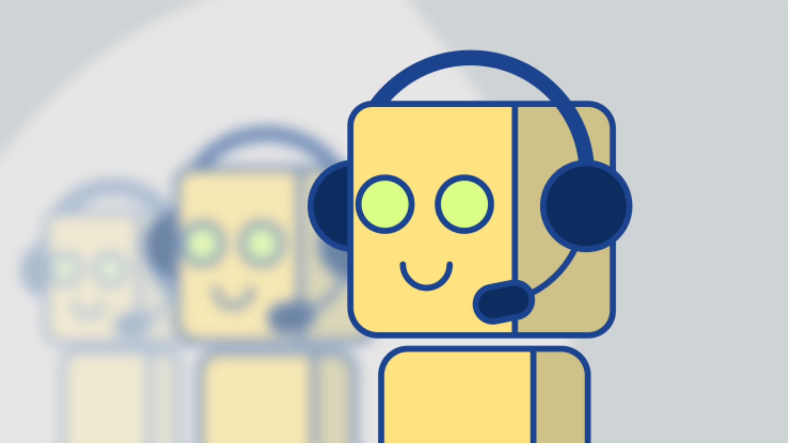
January 23, 2018

Machine learning and AI research can be Thought of as Building the Brain of the 4th Industrial and Revolutions

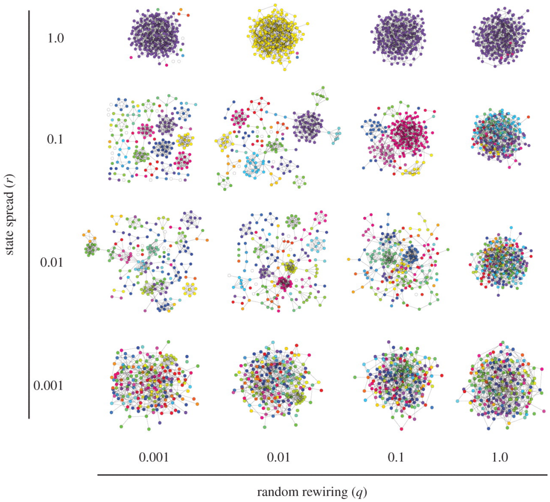


Where we are - Teaching Machines to See

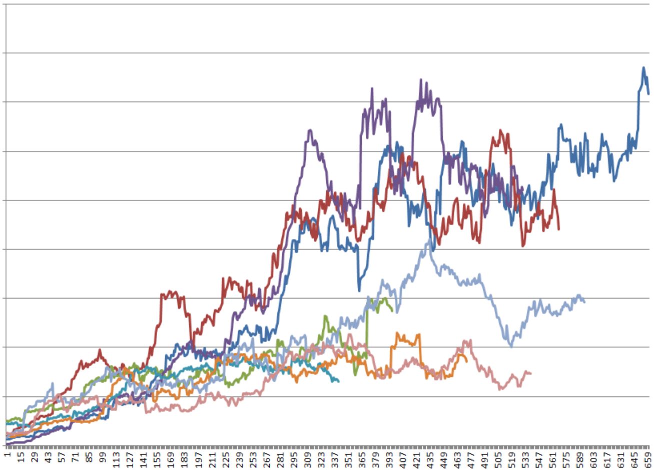
Where we are - Teaching Machines to Talk



Next Step - Teaching Machines to Discover



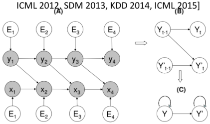
Next Step - Teaching Machines to Discover



Research Thrusts in Time Series Analysis

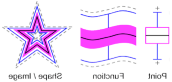
Granger graphical models

[KDD 2007, KDD 2009 (a,b), AAAI 2010, SDM 2012, ICML 2012, SDM 2013, KDD 2014, ICML 2015]



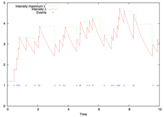
Functional data analysis

[ICML 2015]



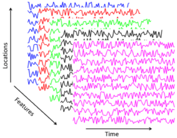
Point process model

[ICML 2015, NIPS 2016]



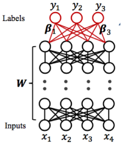
Low rank tensor analysis

[NIPS 2014, ICML 2015, ICML 2016, NIPS 2016]



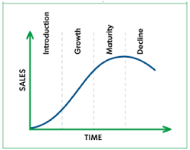
Deep neural networks

[KDD 2015, ICLR 2017]



Life cycle model

[ICDM 2015, IJCAI 2016]



Sponsors:

Research Thrusts in Network Analysis

Novel machine learning models for network analysis and inference:

- Network anomaly detection [SDM 2010; ICDM 2012; KDD 2014]
- Robust network inference [ICML 2015; NIPS 2017; WSDM 2017]
- Network analysis for Recommendation [ICML 2012]
- Network embedding via deep learning models [IJCAI workshop, 2107]

Sponsors:

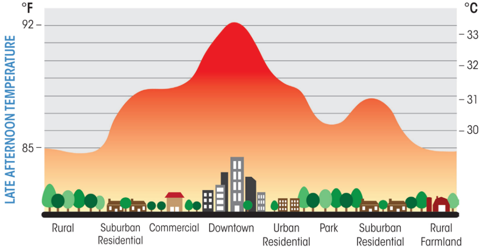
NSF CyberSEES Project

Objectives: a marriage between deep learning approaches and physics based simulation models

Application: casual attribution of urban heat island from heterogeneous data collections

Research problems

- Multi-rate multiresolution
- Heterogeneous data quality
- Interpretation of deep learning models



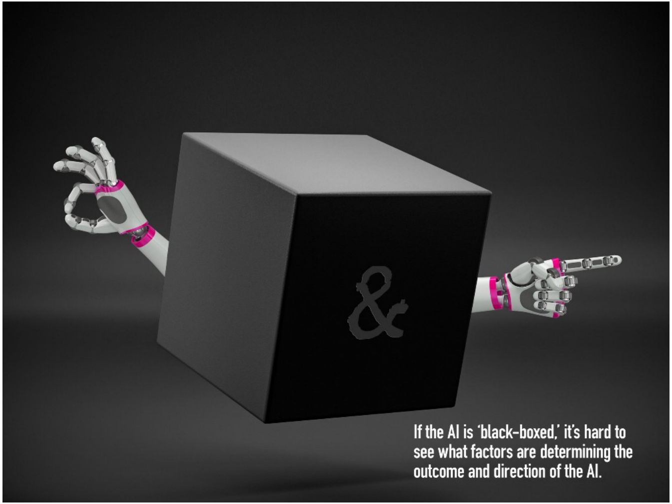
Collaborators

MAYOR'S OFFICE

BERKELEY LAB

MASCLE

Deep Learning as Blackbox



How Deep Learning is Perceived by Students

Russ Salakhutdinov
October 9 at 6:59 PM ·

I am teaching a Deep Learning graduate course this Fall at CMU with over 300 MSc and PhD students enrolled.

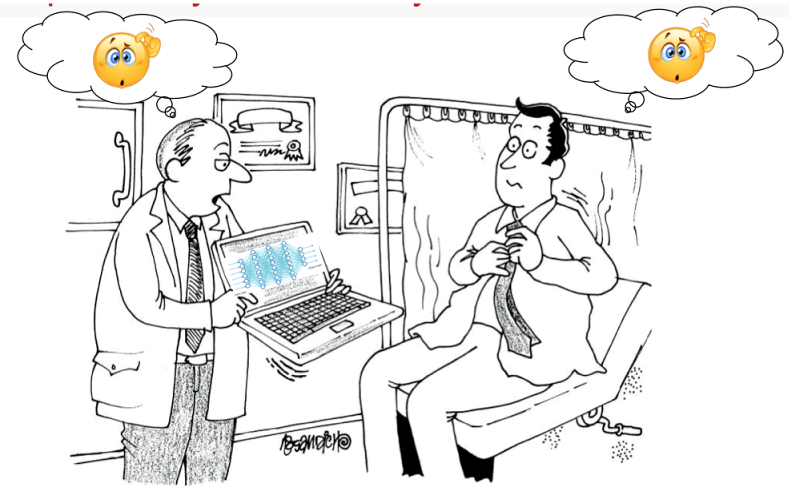
Today, after our midterm, I received the following anonymous feedback: "Did I take the wrong exam? Does this exam cover too little machine learning stuff and focus too much on mathematics?"

I guess there is a common belief that Deep Learning is all about installing TensorFlow or PyTorch and training a gigantic convnet on multiple GPUs 😀

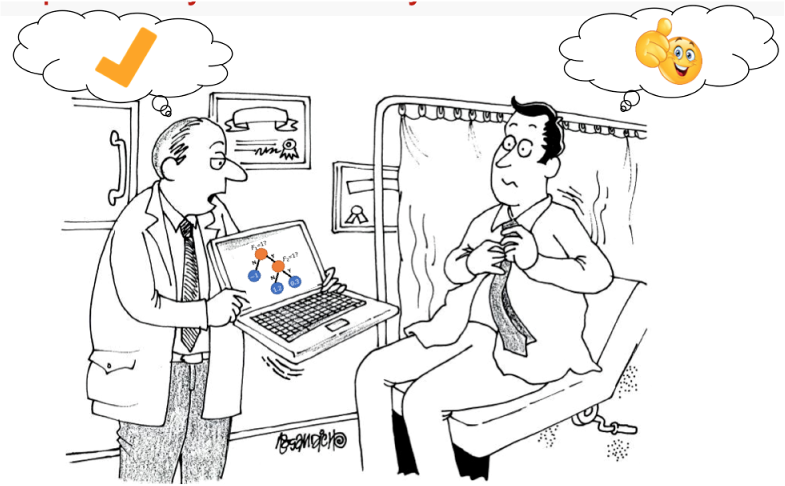
1.4K 76 Comments 199 Shares

Like Comment Share

Importance of Explainable Artificial Intelligence - I



Importance of Explainable Artificial Intelligence - I



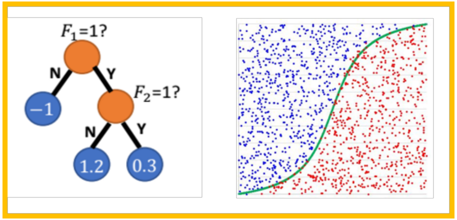
Importance of Explainable Artificial Intelligence - II

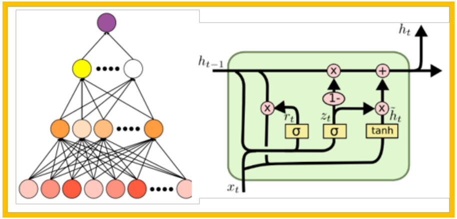
How can I trust any machine learning algorithm? [Ribeiro et al, 2016]

(a) Husky classified as wolf (b) Explanation

Interpretable Model is Necessary

Interpretable predictive models are shown to result in faster adoptability of machine learning models.





- Simple and commonly use models
- Easy to interpret, mediocre performance

- Deep learning solutions
- Superior performance, hard to explain

Can we learn interpretable models with robust prediction performance?

Ongoing Work on Explainable Machine Learning Models

Direct Interpretation

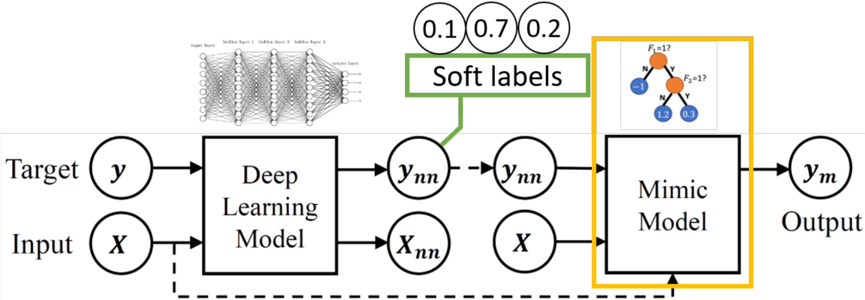
- [Garson, 1991]: estimating feature importance directly from network weight connections
- [Hechtlinger, 2016]: computing output gradients with respect to input features
- [Itti et al., 1998; Mnih et al., 2014; Xu et al., 2015]: attention models

Indirect Interpretation

- [Provost et al., 1997]: sensitivity analysis of feature contributions to a neural network's output
- [Ribeiro et al., 2016]: local interpretability for black-box models
- [Che et al., 2016]: mimicking the blackbox through the prediction scores
- [Maaten and Hinton, 2008; Simonyan et al., 2013; Yosinski et al., 2014; LeCun et al., 2015; Mnih et al., 2015; Mahendran and Vedaldi, 2015]: visualizing the hidden units

Interpretable Mimic Learning Framework [Che et al., 2016]

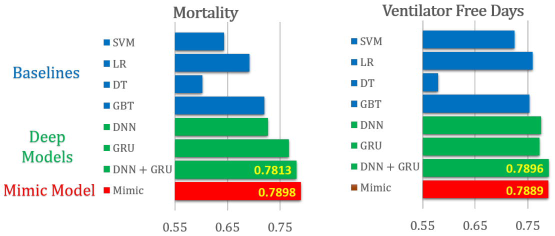
- Main ideas:
  - Borrow the ideas from knowledge distillation [Hinton, et al., 2015] and mimic learning [Ba, Caruana, 2014].
  - Use **Gradient Boosting Trees (GBTs)** to mimic deep learning models.
- Training Pipeline:



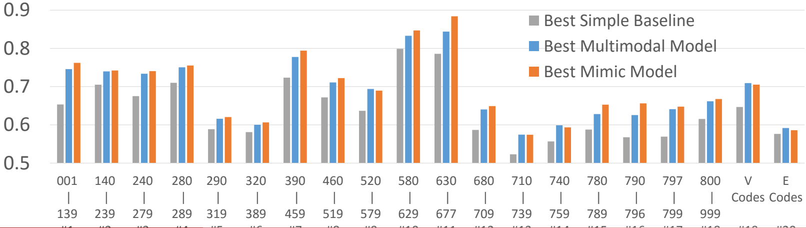
- Benefits: Good performance, less overfitting, interpretations.

Quantitative Evaluation

AUROC score of prediction on patients with acute hypoxemic respiratory failure.

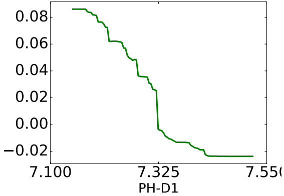


AUROC score of 20 ICD-9 diagnosis category prediction tasks on MIMIC-III dataset.



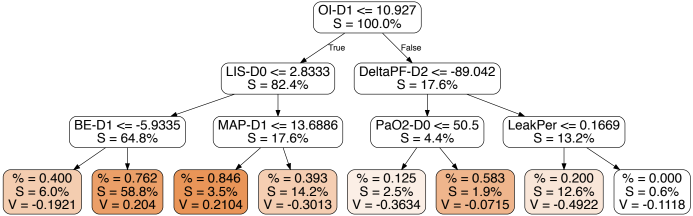
Model/Feature Interpretation

Partial dependency plot for mortality prediction on patients with acute hypoxemic respiratory failure.



- pH value in blood should stay in a normal range around 7.35-7.45.
- Our model predicts a higher mortality change when the patient pH value below 7.325.

Most Useful Decision Trees for ventilator free days prediction.



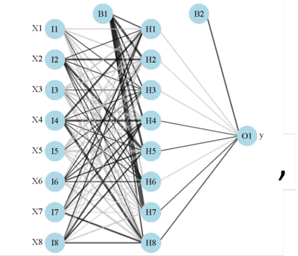
Useful features:

- Lung injury score
- Oxygenation index
- PF ratio change

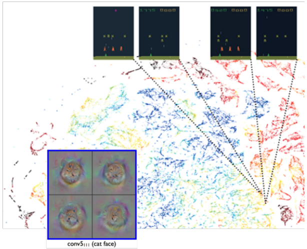
MASCLE

Black-Box Problem of Neural Networks

- Can we directly interpret neural networks?
- Existing methods to interpret neural networks do not cover the interpretation of statistical interactions.



$$, \frac{d\hat{f}}{dx_k},$$



Feature importance — Input gradient — Hidden representation

Interaction Detection

- Statistical interactions: non-additive groupings of variables in function $F(\mathbf{x})$.
- Example 1:

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

Interactions: $\{x_1, x_2, x_3\}$, $\{x_3, x_5\}$, $\{x_7, x_8, x_9, x_{10}\}$, $\{x_2, x_7\}$

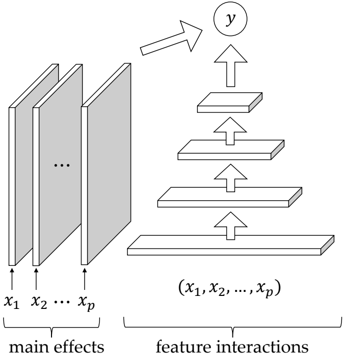
- Example 2:

$$F_2(\mathbf{x}) = \log(x_1) + \log(x_2)$$

Interactions: none.

Main Contributions

- Our contributions:
  - A novel interpretation of the weights of a deep neural network
  - A state-of-the-art framework for detecting arbitrary-order interactions accurately and efficiently
  - A model reduction of deep neural networks via generalized additive models



Preliminaries

Feedforward Neural Network: consider a feedforward neural network with L hidden layers. Let p_ℓ be the number of hidden units in the ℓ-th layer.

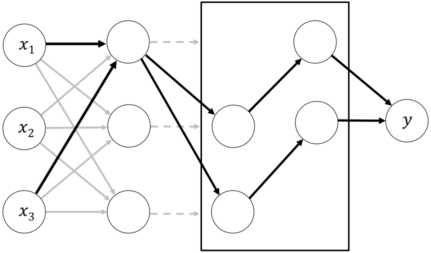
- Input features as the 0-th layer and $p_0 = p$ is the number of input features.
- L weight matrices $\mathcal{W}^{(\ell)} \in^{p_\ell \times p_{\ell-1}}$ and L bias vectors $\mathbf{b}^{(\ell)} \in^{p_\ell}$.
- $\phi(\cdot)$ is the activation function (non-linearity), $\mathbf{w}^y \in^{p_L}$ and $b^y \in$ are the coefficients and bias for the final output.
- Formulation:

$$\mathbf{h}^{(0)} = \mathbf{x}, \quad y = (\mathbf{w}^y)^\top \mathbf{h}^{(L)} + b^y, \quad \text{and } \mathbf{h}^{(\ell)} = \phi\left(\mathcal{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}\right), \quad \forall \ell = 1, 2, \ldots, L.$$

Motivations

Key observation: any input features interacting with each other must follow strongly weighted connections to a common hidden unit before the final output.

An example: $F(\mathbf{x})$ has interaction $\{x_1, x_3\}$



Theoretical Analysis

Lemma (Interactions at Common Hidden Units)

Consider a feedforward neural network with input feature $x_i, i \in [p]$, where $y = \varphi\left(x_1, \ldots, x_p\right)$. For any interaction $\mathcal{I} \subset [p]$ in $\varphi\left(\cdot\right)$, there exists a vertex $v_{\mathcal{I}}$ in the associated directed graph such that $\mathcal{I}$ is a subset of the ancestors of $v_{\mathcal{I}}$ at the input layer (i.e., $\ell = 0$).

Neural Interaction Detector (NID)

Proposed Algorithm:

1. Train feedforward neural networks with sparsity regularization
2. Rank interactions by interpreting weights
3. Find cutoff on the ranking (if desired)

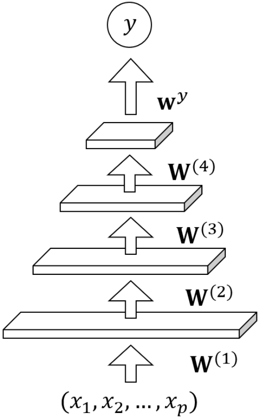
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

$$\omega_i(\mathcal{I}) = z_i^{(1)} \mu \left(\left| \mathbf{W}_{i,\mathcal{I}}^{(1)} \right| \right)$$ for hidden unit i

Approximation of Hidden Unit Influence

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



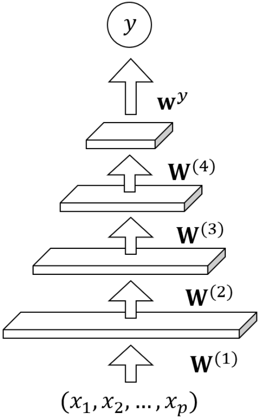
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

$$\omega_i(\mathcal{I}) = z_i^{(1)} \mu\left(\left|\mathbf{W}_{i,\mathcal{I}}^{(1)}\right|\right) \quad \text{for hidden unit } i$$

Approximation of Hidden Unit Influence

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



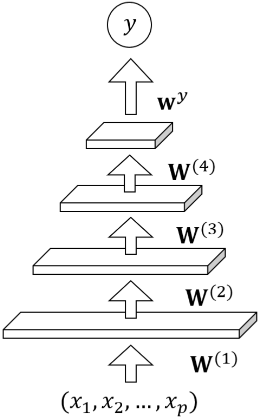
Rank Interactions by Interpreting Weights

Interaction Strength Per Hidden Unit

$$\omega_i(\mathcal{I}) = z_i^{(1)} \mu\left(\left|\mathbf{W}_{i,\mathcal{I}}^{(1)}\right|\right) \quad \text{for hidden unit } i$$

Approximation of Hidden Unit Influence

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



Theoretical Analysis

Lemma (Neural Network Lipschitz Estimation)

Let the activation function $\phi(\cdot)$ be a 1-Lipschitz function. Then the output y is $z_i^{(\ell)}$-Lipschitz with respect to $h_i^{(\ell)}$.

- Lipschitz constants provides upper bounds the gradient magnitudes of hidden units.
- The upper bound on the gradient magnitude approximates how important the variable can be.

MASCLE

Rank Interactions by Interpreting Weights

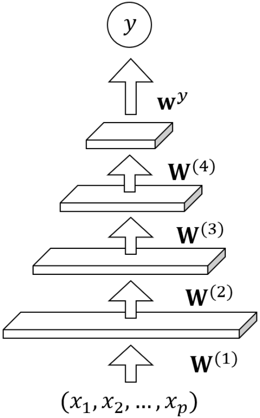
Interaction Strength Per Hidden Unit

$$\omega_i(\mathcal{I}) = z_i^{(1)} \mu\left(\left|\mathbf{W}_{i,\mathcal{I}}^{(1)}\right|\right)$$ for hidden unit i

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

Approximation of Hidden Unit Influence

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



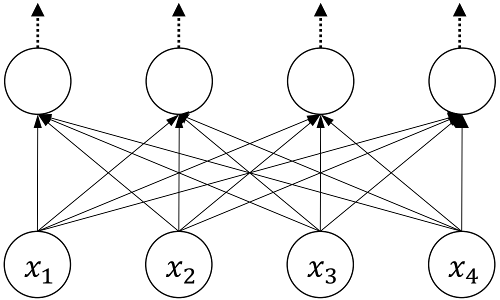
A Simple Example

Definition of μ: Let $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$ be the best quadratic approximation, measured by square loss, to the ReLu function $\max\{\alpha_1 x_1 + \alpha_2 x_2, 0\}$ on $(x_1, x_2) \in (-1, 1) \times (-1, 1)$, for the coefficient of interaction $\{x_1, x_2\}$, β_5, we have:

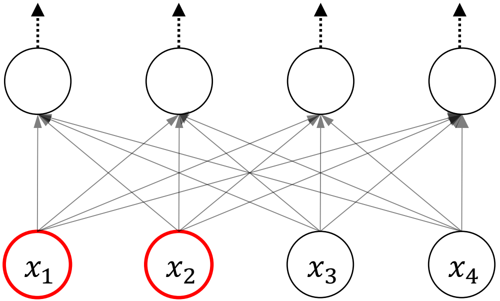
$$|\beta_5| = \frac{3}{4}\left(1 - \frac{\min\{\alpha_1^2, \alpha_2^2\}}{5\max\{\alpha_1^2, \alpha_2^2\}}\right)\min\{|\alpha_1|, |\alpha_2|\}$$

.

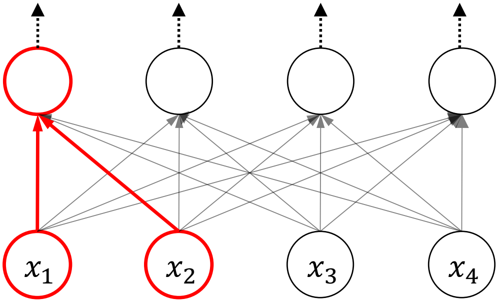
Ranking Pairwise Interactions



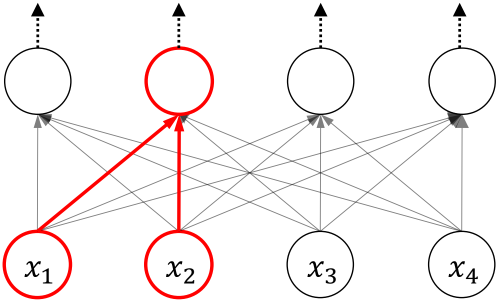
Ranking Pairwise Interactions



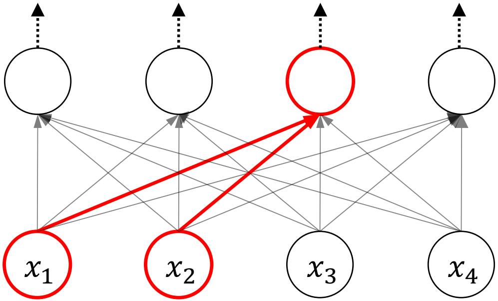
Ranking Pairwise Interactions



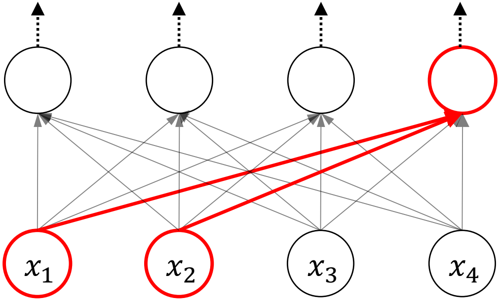
Ranking Pairwise Interactions



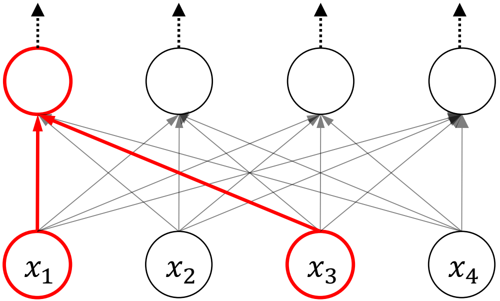
Ranking Pairwise Interactions



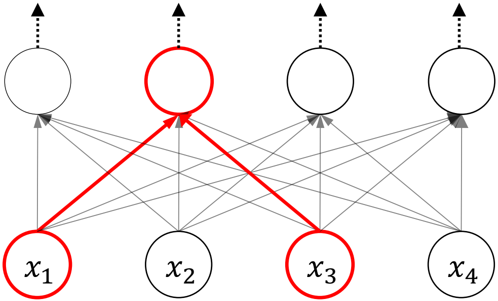
Ranking Pairwise Interactions



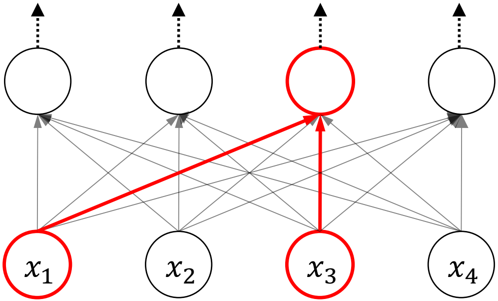
Ranking Pairwise Interactions



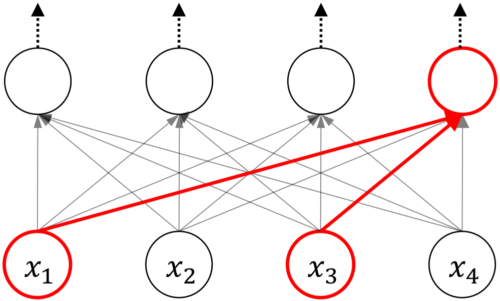
Ranking Pairwise Interactions



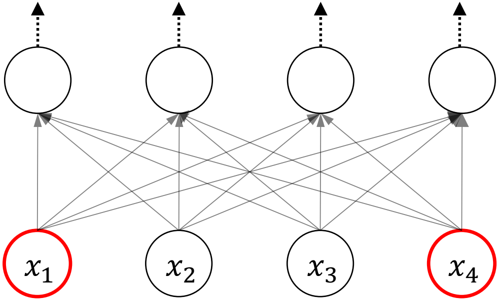
Ranking Pairwise Interactions



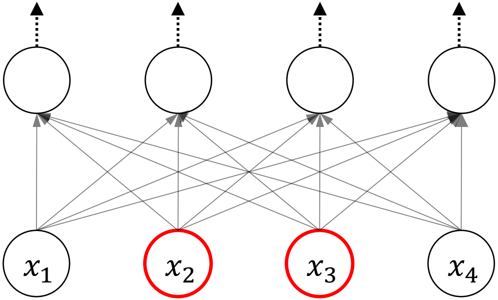
Ranking Pairwise Interactions



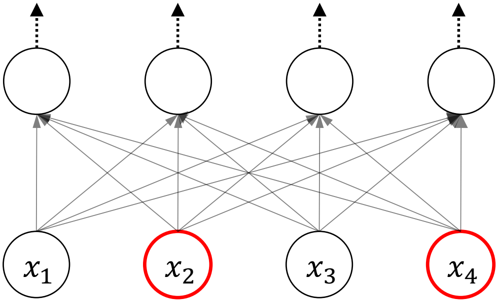
Ranking Pairwise Interactions



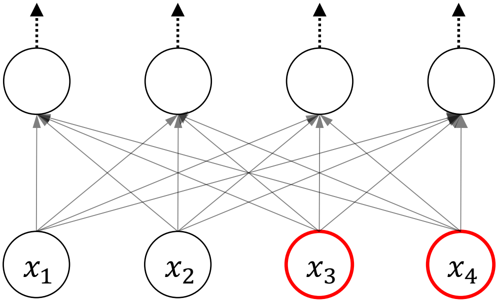
Ranking Pairwise Interactions



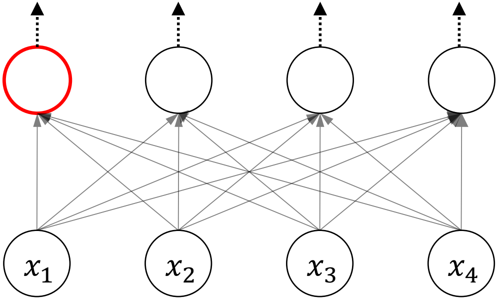
Ranking Pairwise Interactions



Ranking Pairwise Interactions



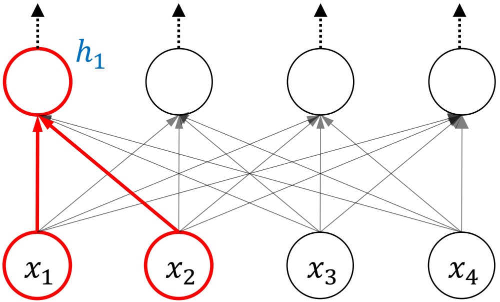
Ranking Higher-Order Interactions



$$|w_1| > |w_2| > |w_3| > |w_4|$$

MASCLE

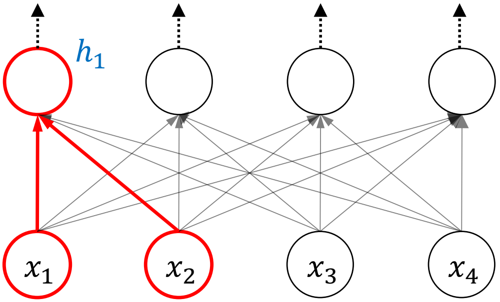
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1 \min(\lvert w_1\rvert, \lvert w_2\rvert)$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

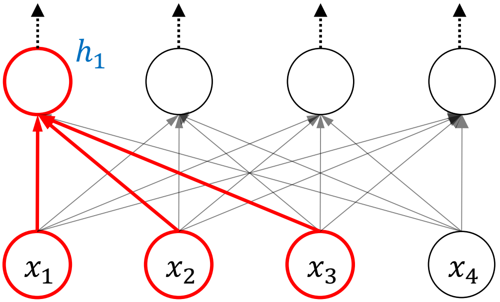
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1 \lvert w_2 \rvert$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

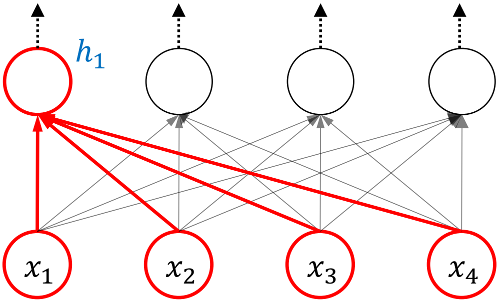
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1\|w_2\|$
{1,2,3}	$z_1\|w_3\|$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

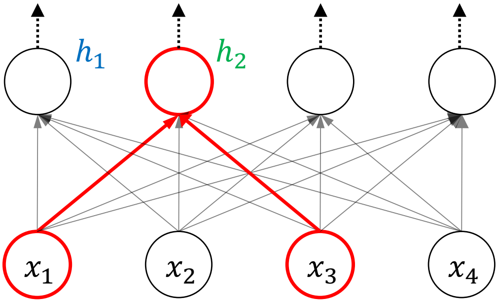
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1\|w_2\|$
{1,2,3}	$z_1\|w_3\|$
{1,2,3,4}	$z_1\|w_4\|$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

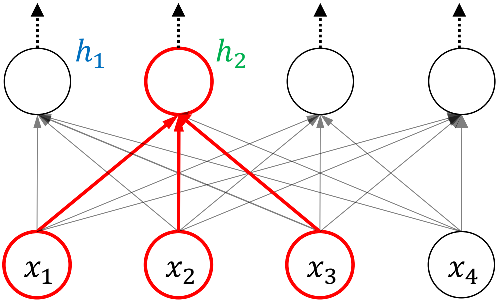
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1 \lvert w_2 \rvert$
{1,2,3}	$z_1 \lvert w_3 \rvert$
{1,2,3,4}	$z_1 \lvert w_4 \rvert$
{1,3}	$z_2 \lvert w_1 \rvert$

$$|w_3| > |w_1| > |w_2| > |w_4|$$

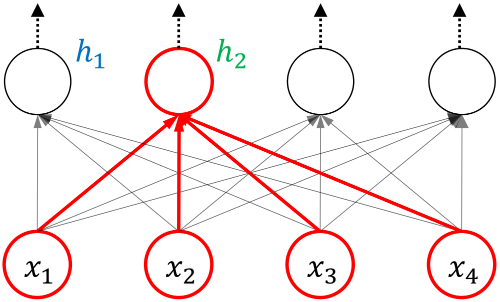
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1\|w_2\|$
{1,2,3}	$z_1\|w_3\| + z_2\|w_2\|$
{1,2,3,4}	$z_1\|w_4\|$
{1,3}	$z_2\|w_1\|$

$$|w_3| > |w_1| > |w_2| > |w_4|$$

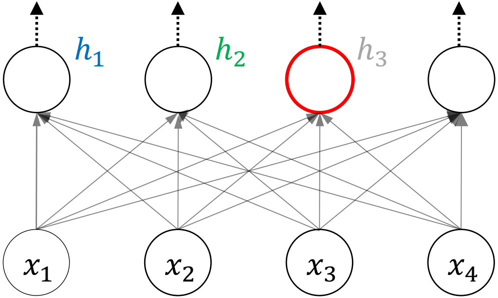
Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1\|w_2\|$
{1,2,3}	$z_1\|w_3\| + z_2\|w_2\|$
{1,2,3,4}	$z_1\|w_4\| + z_2\|w_4\|$
{1,3}	$z_2\|w_1\|$

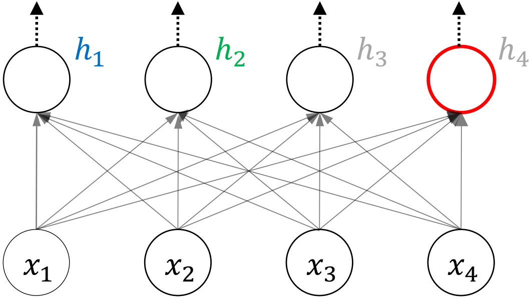
$$|w_3| > |w_1| > |w_2| > |w_4|$$

Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1\lvert w_2\rvert$
{1,2,3}	$z_1\lvert w_3\rvert + z_2\lvert w_2\rvert$
{1,2,3,4}	$z_1\lvert w_4\rvert + z_2\lvert w_4\rvert$
{1,3}	$z_2\lvert w_1\rvert$
...	...

Ranking Higher-Order Interactions



Interactions	Strengths
{1,2}	$z_1\lvert w_2\rvert$
{1,2,3}	$z_1\lvert w_3\rvert + z_2\lvert w_2\rvert$
{1,2,3,4}	$z_1\lvert w_4\rvert + z_2\lvert w_4\rvert$
{1,3}	$z_2\lvert w_1\rvert$
...	...

Sample Interaction Ranking

Interactions	Strengths
{1,2,3}	1.3421
{1,2,3,4}	0.8241
{1,2}	0.3415
{1,3}	0.2310
...	...

Theoretical Analysis

Theorem (Improving the ranking of higher-order interactions)

Let $\mathcal{R}$ be the set of interactions proposed by NID, let $\mathcal{I} \in \mathcal{R}$ be a d-way interaction where $d \geq 3$, and let $\mathcal{S}$ be the set of subset $(d-1)$-way interactions of $\mathcal{I}$ where $|\mathcal{S}| = d$. Assume that for any hidden unit j which proposed $s \in \mathcal{S} \cap \mathcal{R}$, $\mathcal{I}$ will also be proposed at the same hidden unit, and $\omega_j(\mathcal{I}) > \frac{1}{d}\omega_j(s)$. Then, one of the following must be true: a) $\exists s \in \mathcal{S} \cap \mathcal{R}$ ranked lower than $\mathcal{I}$, i.e., $\omega(\mathcal{I}) > \omega(s)$, or b) $\exists s \in \mathcal{S}$ where $s \notin \mathcal{R}$.

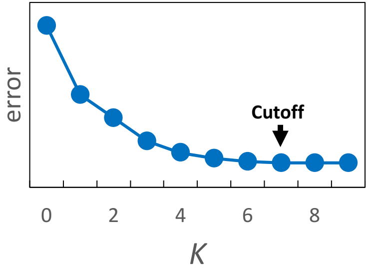
Find a Cutoff on the Ranking

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

$$c_K(\mathbf{x}) = \sum_{i=1}^{p} g_i(x_i) + \sum_{i=1}^{K} g_i'(\mathbf{x}_{\mathcal{I}})$$

$g_i(\cdot)$: main effects

$g_i'(\cdot)$: interactions



Test Suite of Data-Generating Functions

$F_1(\mathbf{x})$	$\pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7$
$F_2(\mathbf{x})$	$\pi^{x_1 x_2}\sqrt{2\|x_3\|} - \sin^{-1}(0.5x_4) + \log(\|x_3 + x_5\| + 1) + \frac{x_9}{1+\|x_{10}\|}\sqrt{\frac{x_7}{1+\|x_8\|}} - x_2 x_7$
$F_3(\mathbf{x})$	$\exp\|x_1 - x_2\| + \|x_2 x_3\| - x_3^{2\|x_4\|} + \log(x_4^2 + x_5^2 + x_7^2 + x_8^2) + x_9 + \frac{1}{1+x_{10}^2}$
$F_4(\mathbf{x})$	$\exp\|x_1 - x_2\| + \|x_2 x_3\| - x_3^{2\|x_4\|} + (x_1 x_4)^2 + \log(x_4^2 + x_5^2 + x_7^2 + x_8^2) + x_9 + \frac{1}{1+x_{10}^2}$
$F_5(\mathbf{x})$	$\frac{1}{1+x_1^2+x_2^2+x_3^2} + \sqrt{\exp(x_4 + x_5)} + \|x_6 + x_7\| + x_8 x_9 x_{10}$
$F_6(\mathbf{x})$	$\exp(\|x_1 x_2\| + 1) - \exp(\|x_3 + x_4\| + 1) + \cos(x_5 + x_6 - x_8) + \sqrt{x_8^2 + x_9^2 + x_{10}^2}$
$F_7(\mathbf{x})$	$(\arctan(x_1) + \arctan(x_2))^2 + \max(x_3 x_4 + x_6, 0) - \frac{1}{1+(x_4 x_5 x_6 x_7 x_8)^2} + \left(\frac{\|x_7\|}{1+\|x_9\|}\right)^5 + \sum_{i=1}^{10} x_i$
$F_8(\mathbf{x})$	$x_1 x_2 + 2^{x_3+x_5+x_6} + 2^{x_3+x_4+x_5+x_7} + \sin(x_7 \sin(x_8 + x_9)) + \arccos(0.9x_{10})$
$F_9(\mathbf{x})$	$\tanh(x_1 x_2 + x_3 x_4)\sqrt{\|x_5\|} + \exp(x_5 + x_6) + \log\left((x_6 x_7 x_8)^2 + 1\right) + x_9 x_{10} + \frac{1}{1+\|x_{10}\|}$
$F_{10}(\mathbf{x})$	$\sinh(x_1 + x_2) + \arccos(\tanh(x_3 + x_5 + x_7)) + \cos(x_4 + x_5) + \sec(x_7 x_9)$

Complex functions are used in our evaluation

MASCLE

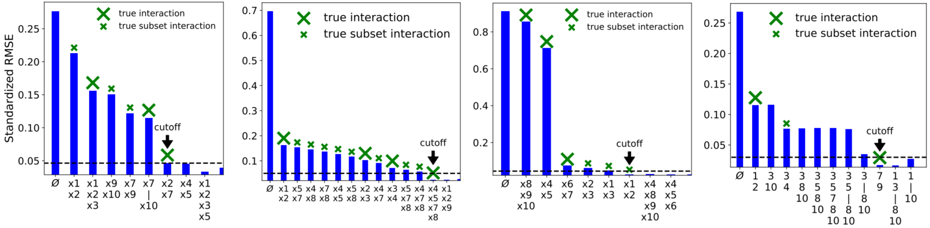
AUC of Pairwise Interaction Strengths

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

[1]Fisher 1925, [2]Bien et al. 2013, [3]Sorokina et al. 2008

*F_6 plays an important role for this result

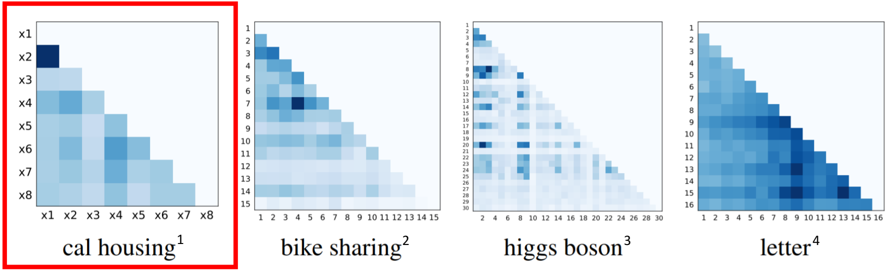
Higher-Order Interaction Detection for Synthetic Data



F_1 F_3 F_5 F_7

$F_1(\mathbf{x})$	$\pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7$
$F_3(\mathbf{x})$	$\exp\lvert x_1 - x_2\rvert + \lvert x_2 x_3\rvert - x_3^{2\lvert x_4\rvert} + \log(x_4^2 + x_5^2 + x_7^2 + x_8^2) + x_9 + \frac{1}{1 + x_{10}^2}$
$F_5(\mathbf{x})$	$\frac{1}{1 + x_1^2 + x_2^2 + x_3^2} + \sqrt{\exp(x_4 + x_5)} + \lvert x_6 + x_7\rvert + x_8 x_9 x_{10}$
$F_7(\mathbf{x})$	$(\arctan(x_1) + \arctan(x_2))^2 + \max(x_3 x_4 + x_6, 0) - \frac{1}{1 + (x_4 x_5 x_6 x_7 x_8)^2} + \left(\frac{\lvert x_7\rvert}{1 + \lvert x_9\rvert}\right)^5 + \sum_{i=1}^{10} x_i$

Pairwise Heap-Maps for Real-World Data



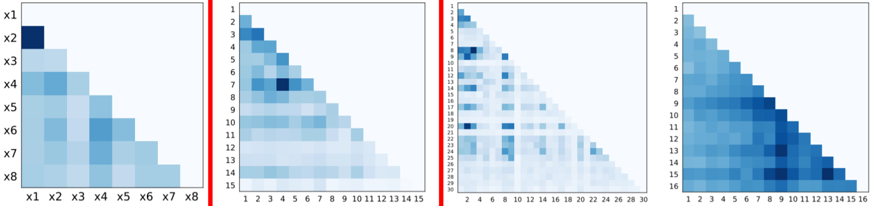
{1,2}: longitude and latitude

[1]Pace et al. 1997, [2]Fanaee-T et al. 2014, [3]Adam-Bourdarios et al. 2014, [4]Frey et al. 1991

MASCLE

15/17

Pairwise Heap-Maps for Real-World Data



cal housing[1] bike sharing[2] higgs boson[3] letter[4]

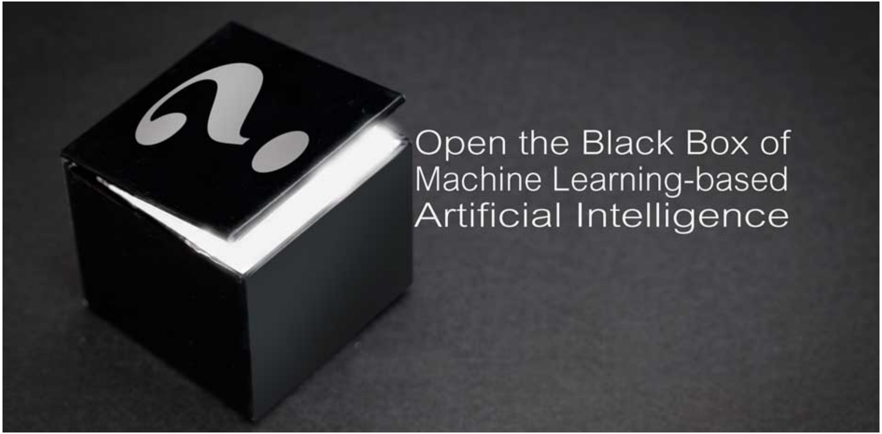
{4,7}: hour and working day

[1]Pace et al. 1997, [2]Fanaee-T et al. 2014, [3]Adam-Bourdarios et al. 2014, [4]Frey et al. 1991

MASCLE

15/17

Deciphering the Black Box



MASCLE

USC-Melady Research Group

Michael Tsang, Dehua Cheng, and Yan Liu, Detecting Statistical Interactions from Neural Network Weights, arXiv:1705.04977.

Thank You!

Questions and Comments?

References I

Che, Z., Purushotham, S., Khemani, R., and Liu, Y. (2016). Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association.

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert*, 6(4):46–51.

Hechtlinger, Y. (2016). Interpretation of prediction models using the input gradient. *arXiv preprint arXiv:1611.07634*.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5188–5196.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Provost, F. J., Fawcett, T., et al. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD*, volume 97, pages 43–48.

References II

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.