CCAI 2017
中国人工智能大会

艺术与人工智能的明天
– 人机信任合作

刘燕
南加州大学计算机系

技术应用于艺术从古至今

文艺复兴时期德国画家汉斯·贺尔拜因（Hans Holbein）的双重肖像图

人工智能应用于绘画 - DeepDream

旧金山Mission区老式艺术画廊的艺术展

更多的人工智能应用

谷歌Megenta项目-
人工智能创造音乐 Daddy’s car

ANGELINA项目-
人工智能创造游戏

Sunspring-
人工智能创造电影剧本

人工智能对艺术应用的瓶颈

- 中国著名画家叶永青先生的观点
  - 个性化
  - 表达与共鸣
  - 时代性
- 人工智能领域的瓶颈
  - 人工智能的迁移(Transfer learning)
  - 人工智能情感研究 (Cognitive science and neuroscience)
  - 可解释性人工智能 (Explainable AI)

人工智能与艺术的共同点

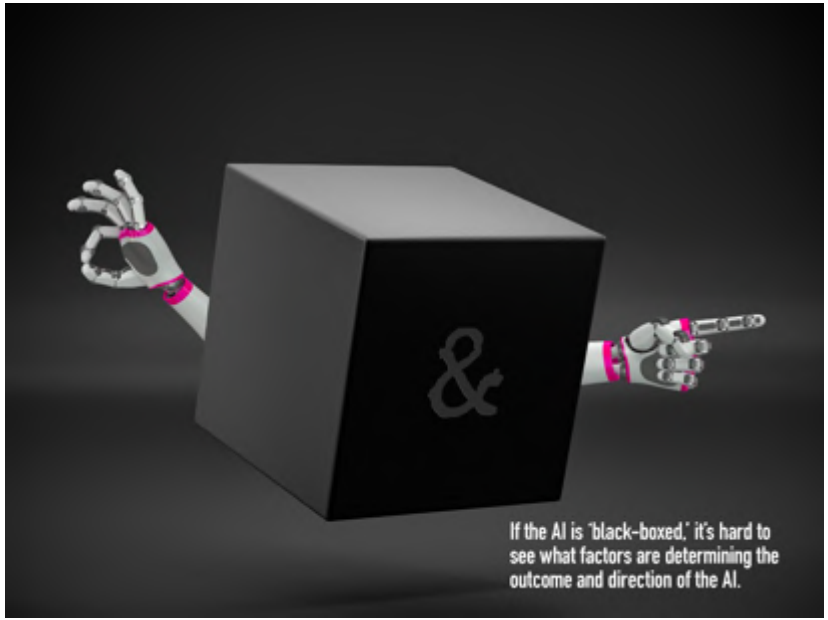
什么是艺术？

艺术是个谎言，但却是一个说真话的谎言。--毕加索

对绘画的次要地位进行的主要评价就是，艺术中的进步并不总是形式的进步.—唐纳德．贾德(Donald Judd，1963)

过去几年中最好的新作品有 一半以上既不是绘画也不是雕塑。一唐维德．贾德(1965)

要谈论艺术的一件事就是，它是一件事。艺术就是作为艺术的艺术(art as art)，其它的一切就是其它的一切。作为艺术的艺术不是别的而是艺术。艺术并非不是艺术的东西。 —艾德．莱因哈特(Ad Reinhardt，1963)



可解释性人工智能 (Explainable AI)

(a) Husky classified as wolf

(b) Explanation

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

可解释性人工智能-提高对人类对艺术的理解

- 什么让人产生美感？
- 如何创造美的艺术？
- 如何评价和估价艺术作品？

可解释性人工智能 (Explainable AI)

Direct Interpretation

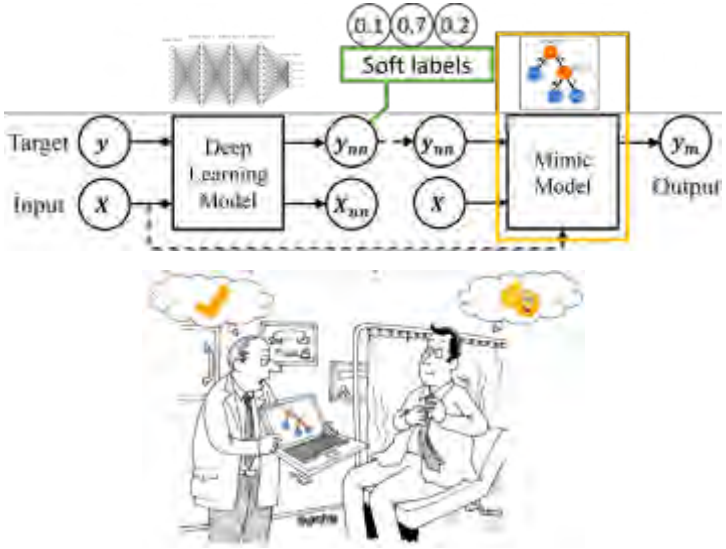
- (Garson, 1991): estimating feature importance directly from network weight connections
- (Hechtlinger, 2016): computing output gradients with respect to input features
- (Itti et al., 1998; Mnih et al., 2014; Xu et al. 2015): modifying neural architectures to interpret the prediction models

Indirect Interpretation

- (Provost et al, 1997): sensitivity analysis of feature contributions to a neural network's output
- (Ribeiro et al. , 2016; Che et al, 2015): mimicking the blackbox through the prediction scores
- (Maaten Hinton, 2008; LeCun et al., 2015; Mnih et al., 2015, Simonyan et al. 2013; Mahendran Vedaldi, 2015; Yosinski et al. 2015): visualizing the hidden units

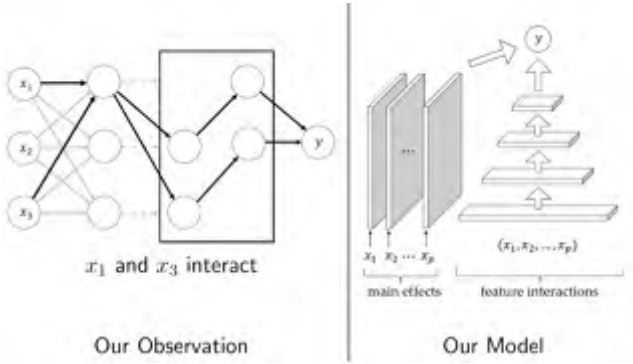
可解释性人工智能 (Explainable AI)

Indirect Interpretation



Che et al, 2015

Direct Interpretation



Tsang et al, 2017

Neural Interaction Detector

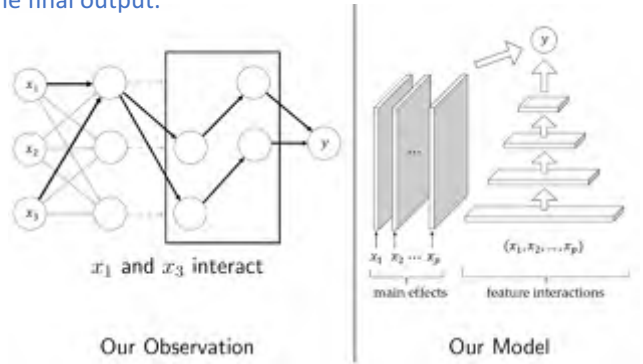
What is a statistical interaction?

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

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

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



For a potential interaction $I \in [p]$ and the weight vector $\mathcal{W}_{j,:}$ of the j-th hidden unit, the strength of the interaction is defined as:

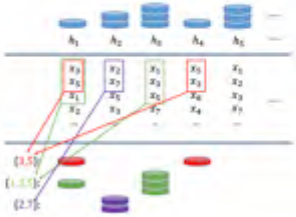
$$W(I, \mathcal{W}_{j,:}) = \min_{i \in I} \|\mathcal{W}_{j,i}\|,$$

Neural Interaction Detector

- Step 1: Aggregate Weights

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

- Step 2: Rank Interaction Candidates



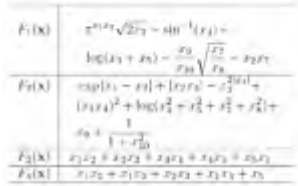
$$\text{interaction strength}(I_{1,\ldots,j}) = \sum z_i \min\left(w_1, w_2, \ldots, w_j\right)$$

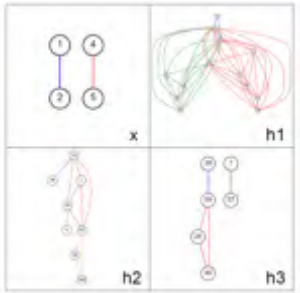
- Step 3: Find Cutoff in Rankings

$$f_K(\mathbf{x}) = \sum_{i=1}^{p} g_i(x_i) + \sum_{i=1}^{K} g_i'(\mathbf{x}_I)$$

We use a Generalized Additive Model with Arbitrary-Order Interactions

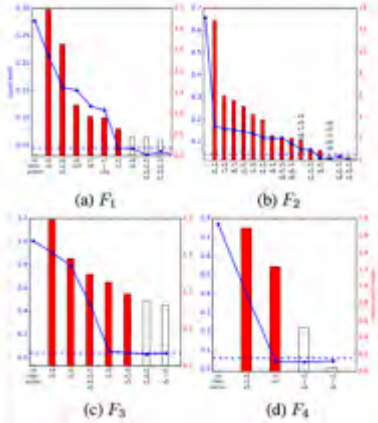
Ground truth





(b) cal housing

Results



未来十年什么工作会被人工智能取代？

CCAI 2017
中国人工智能大会

艺术与人工智能 - 人机信任合作

可解释行，可迁移性，可靠性