AI ProCon 开发者大会

RealAI 瑞莱智慧 CSDN

第三代AI思考与实践

RealAI 田天

https://realai.ai

RealAI 瑞莱智慧 CSDN

AI ProCon 开发者大会

为什么讲第三代AI？

人工智能发展现状

第一代：知识驱动的符号模型

逻辑专家系统

基于**规则**能够**狭义**定义任务的系统

局限：

- 有很多人类行为（知识）并不能精确描述，如常识；
- 知识库总是有限的，无法包含所有信息；
- 知识是确定的；
- 只能描述特定的领域；
- 大量知识不能做到定量化，如质量。

第二代：数据驱动的AI

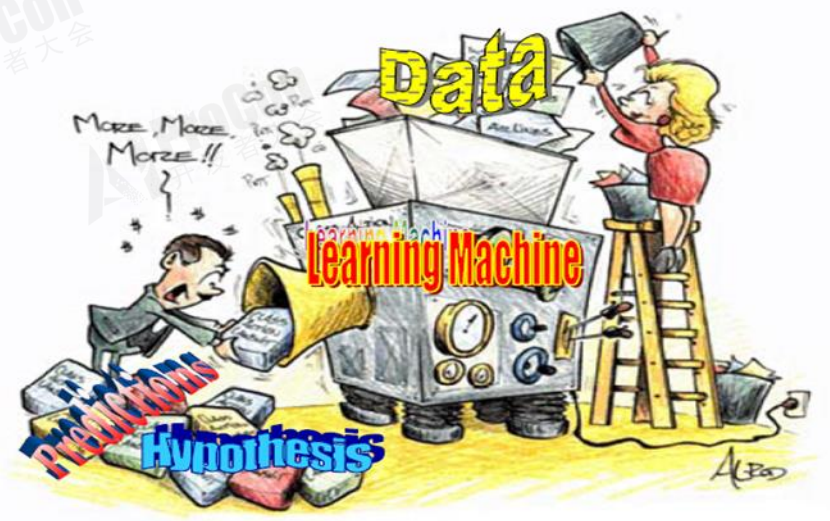
Google

浅层统计学习

计算能力

深度学习

依赖于**大量高质量**的训练数据，不能适应不断**变化**的条件，提供有限的**性能**保证，无法向用户**解释**其结果。



当前的AI困境

以深度神经网络为代表的第二代AI算法

特斯拉、Uber、Waymo等自动驾驶系统出现多次事故，部分是由于AI识别或预测的算法出错

不可靠

通过生成对抗式图像，能够欺骗智能监控系统，使AI系统识别错误

不安全

2018年5月，IBM沃森健康（IBM Watson Health）裁员50%-70%，宣告IBM在医疗AI方向的失败

不可信

第三代人工智能

2016中国计算机大会(CNCC 2016)，张钹院士作大会报告“后深度学习时代的人工智能”，此后进一步提出“第三代人工智能”

知识驱动+数据驱动

第三代人工智能的特点

可信

白盒化AI模型，提供可理解的**决策依据**

可靠

在**样本不足、噪声高、标注差**情况下实现预测效果**可靠提升**

安全

在**恶意攻击**或者数据存在**缺陷样本**的情况下，**保证判断能力**

RealAI 瑞莱智慧 CSDN

AI ProCon 开发者大会

RealAI业务实践

安全

安防监控

通过针对AI算法的攻击手段，可以攻破当前的智能识别系统，安防监控识别系统需升级换代

人脸识别

通过针对AI识别算法的伪装手段，可以欺骗当前的人脸识别系统，银行刷脸认证、刷脸支付等场景存在风险

可靠

工业

工业场景碎片化严重，数据和标注不充足，当前的AI方法处理成本过高

自动驾驶

由于复杂环境下AI算法出错，特斯拉、Uber、Waymo等自动驾驶系统出现多次事故

可信

金融

当前信贷风控、智能投顾等场景AI方法可解释性较差，难以满足业务需求，推广受限

医疗健康

智能医疗诊断系统难以给出符合医疗专业人员需求的判断依据，IBM沃森宣告失败。

贝叶斯（概率）机器学习

- 以贝叶斯理论为核心：先验知识、不确定性计算

$$p(\theta|D) = \frac{p(D|\theta)\pi(\theta)}{p(D)}$$

Thomas Bayes (1702 – 1761)

U. Cambridge
Fellow of the Royal Society (FRS)

REVIEW

doi:10.1038/nature14541

Probabilistic machine learning and artificial intelligence

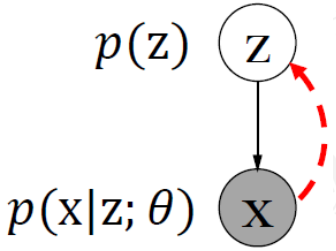
Zoubin Ghahramani[1]

How can a machine learn from experience? Probabilistic modelling provides a framework for understanding what learning is, and has therefore emerged as one of the principal theoretical and practical approaches for designing machines that learn from data acquired through experience. The probabilistic framework, which describes how to represent and manipulate uncertainty about models and predictions, has a central role in scientific data analysis, machine learning, robotics, cognitive science and artificial intelligence. This Review provides an introduction to this framework, and discusses some of the state-of-the-art advances in the field, namely, probabilistic programming, Bayesian optimization, data compression and automatic model discovery.

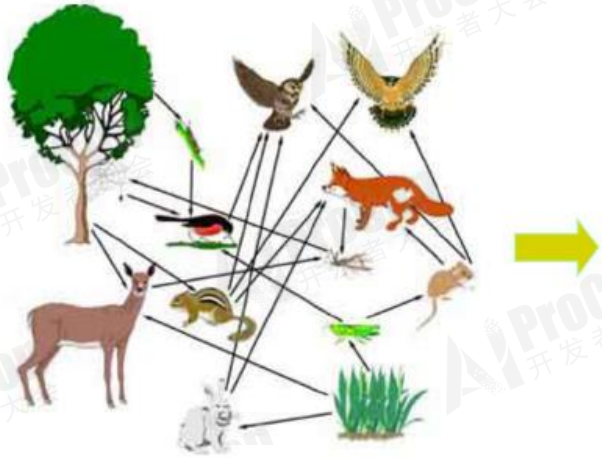
贝叶斯（概率）机器学习

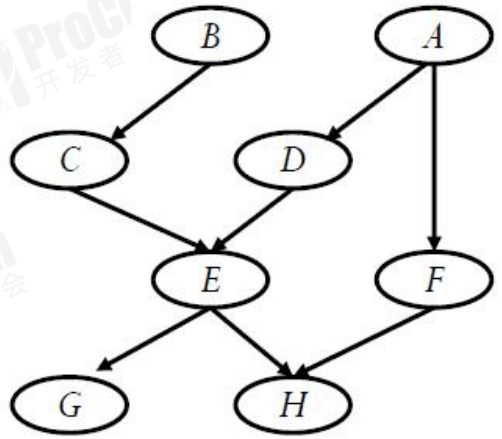
Judea Pearl
Turing Award Laureates

- 隐含变量的建模与推断：揭示背后规律
- 贝叶斯网络：基于图论的结构化先验知识，提升模型表达的灵活性，提高学习效率



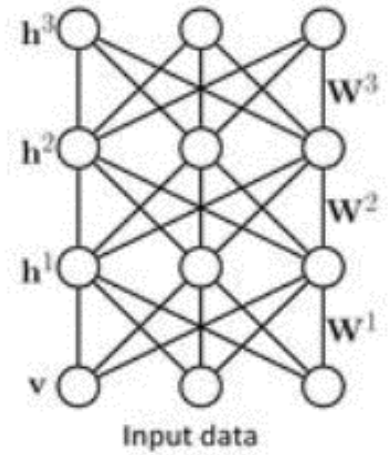
$$p(\mathrm{z}|\mathrm{x}) = \frac{p(\mathrm{x},\mathrm{z})}{p(\mathrm{x})} \propto p(\mathrm{z})p(\mathrm{x}|\mathrm{z};\theta)$$





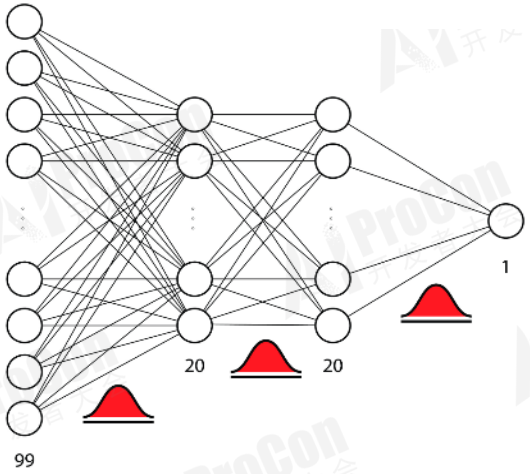
$$p(A,B,C,D,E,F,G,H)$$
$$= p(A)p(B)p(C|B)p(D|A)p(E|C,D)p(F|A)p(G|E)p(H|E,F)$$

贝叶斯深度学习



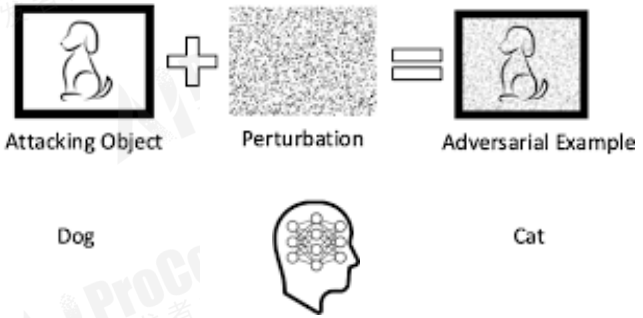
深度生成式模型

实现无/半监督学习，发现数据深层结构，同时提升可解释性



贝叶斯神经网络

结合神经网络拟合能力与数据不确定性特点，提升预测可靠性



AI安全技术

攻击：通过加特定噪声等方法干扰算法输出
防御：通过AI防火墙保护模型

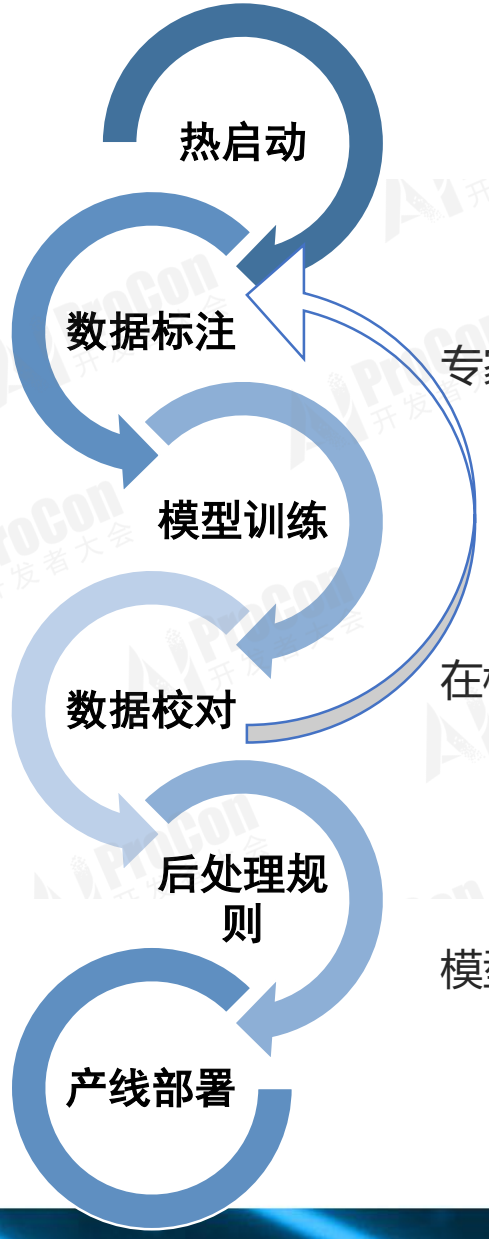
应用领域

案例2：工业视觉检测

使用预训练特征，每类仅需提供少量样本

训练目标检测、语义分割、图像识别模型

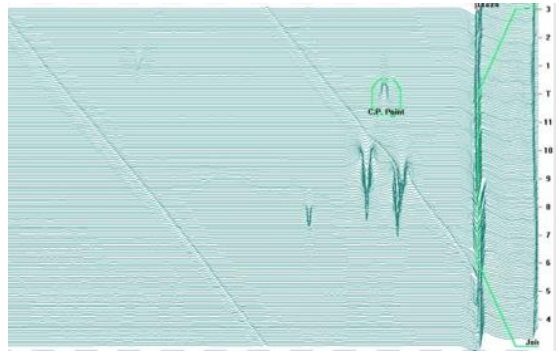
针对厂商标注调整阈值、缺陷规则等

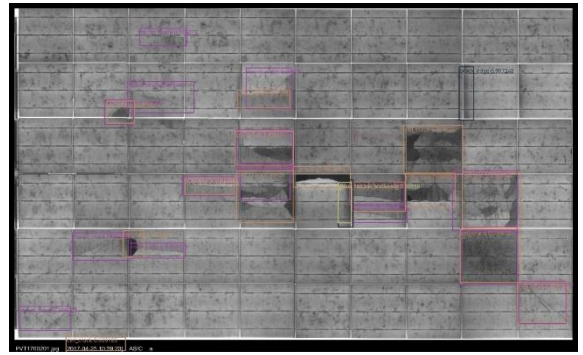


专家对数据进行标注

在模型辅助下提升数据质量

模型、配套软件封装





RealAI 瑞莱智慧 CSDN

AI ProCon 开发者大会

第三代AI进展与展望

第三代AI

可信 可靠 安全的人工智能

可信

当前图像系统识别预测结果：
AlexNet: 狮子鱼，置信度81.3%
VGG-16: 狮子鱼，置信度93.3%
ResNet-18: 狮子鱼，置信度95.6%

神经网络与人思维方式不同，完全依赖于数据拟合，会出现匪夷所思错误

可信

知之为知之，不知为不知

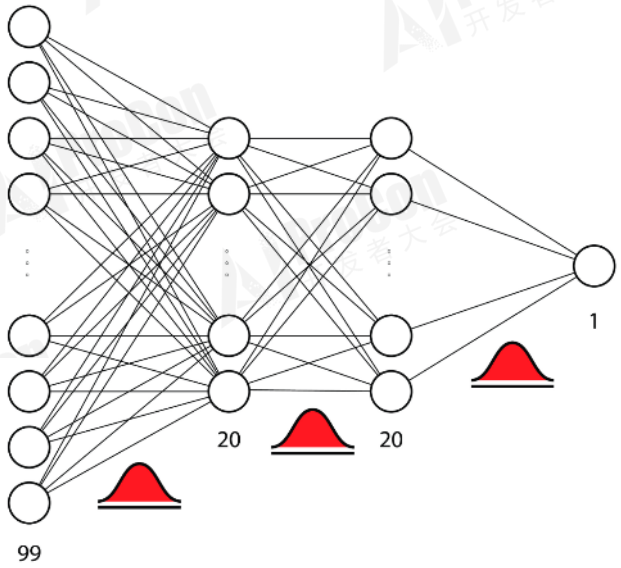
——《论语·为政》

感知世界的不确定性

贝叶斯神经网络通过对不确定性建模，让算法了解什么时候会犯错

贝叶斯神经网络

通过贝叶斯神经网络，拟合复杂数据趋势，同时考虑预测问题中的不确定性。给出可靠可解释结果。



可靠

神经网络为代表的AI方法与人类思维存在本质区别

算法完全依赖于数据，基于数据判断

人类先根据经验判断，再由数据进行验证推断

可靠

“What I cannot create, I do not understand.”

—Richard Feynman

从生成式角度创造新业务模式

工业视觉检测/设备异常检测/… 从模板匹配到真正认知

安全

AI大范围应用带来的全新安全领域

AI安全产业需要保证算法不被滥用，同时不易被黑客操纵

换脸攻击

图片来源：网络新闻

标准应用：RealAI安全平台

RealAI安全平台功能全面，在对抗攻击和防御算法方面具有领先优势。目前第一阶段的**RealSafe对抗攻防平台**已于5月初与清华大学人工智能研究院联合发布。

		Cleverhans (Google)	FoolBox (UCT)	IBM ART (IBM)	RealSafe (RealAI)
模型	支持模型	✘仅TensorFlow	✓通用	✓通用	✓通用
攻击	自定义损失函数	✘不支持	✘不支持	✓部分支持	✓支持
	自定义距离	✘不支持	✘不支持	✘不支持	✓支持
	攻击特定防御	✘不支持	✘不支持	✘不支持	✓支持
防御	对抗训练	✘不支持	✘不支持	✓支持	✓支持
	迁移攻击	✘不支持	✘不支持	✘不支持	✓支持
	数据加噪声	✘不支持	✘不支持	✓支持	✓支持
	对抗样本检测	✘不支持	✘不支持	✓支持	✓支持
	对抗样本清洗	✘不支持	✘不支持	✘不支持	✓支持
	预训练防火墙	✘不支持	✘不支持	✘不支持	✓支持
评测	评测报告	✘不支持	✘不支持	✘不支持	✓支持

总结

- 以深度学习为代表的AI算法仍然存在局限性
- 贝叶斯深度学习方法可以带来更加可信、可靠、安全的AI算法
- 第三代AI将带来大量全新应用场景与商业价值

RealAI 瑞莱智慧 CSDN

AI ProCon 开发者大会

谢谢各位聆听！