# 第三代人工智能

王东

2021/3/15

# 内容

- 基于知识的第一代人工智能
- 基于数据的第二代人工智能
- 知识与数据双驱动的第三代人工智能

中国科学: 信息科学 2020 年 第 50 卷 第 9 期: 1281–1302

SCIENTIA SINICA Informationis

纪念《中国科学》创刊 70 周年专刊 · 评述

《中国科学》杂志社
SCIENCE CHINA PRESS

## 迈向第三代人工智能

张钹*, 朱军, 苏航

清华大学人工智能研究院, 北京 100084

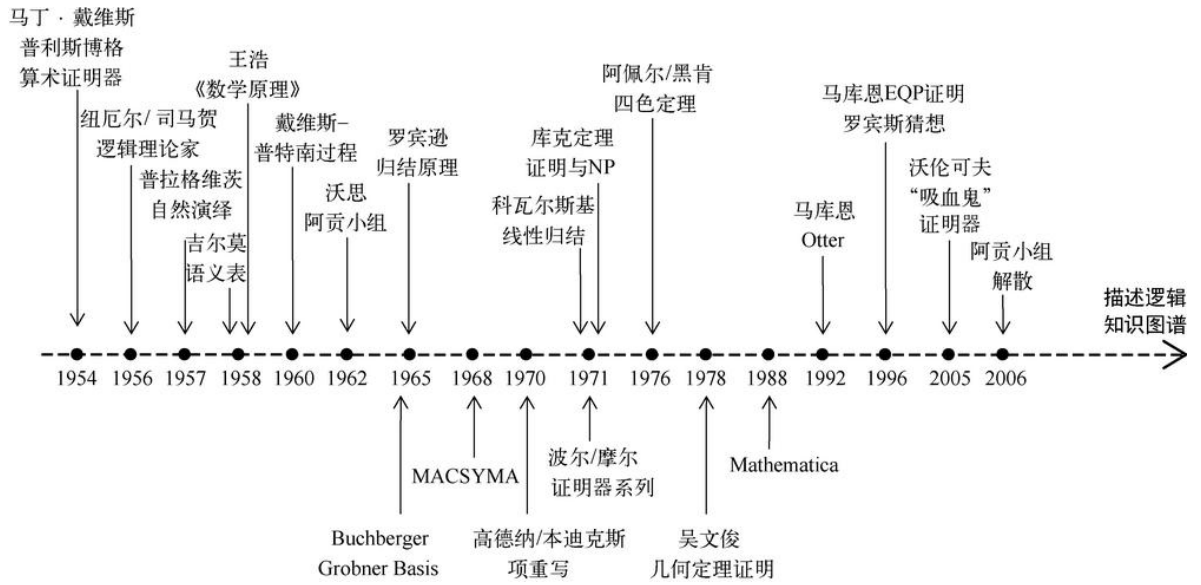
* 通信作者. E-mail: dcszb@tsinghua.edu.cn

收稿日期: 2020–07–06; 接受日期: 2020–08–12; 网络出版日期: 2020–09–22

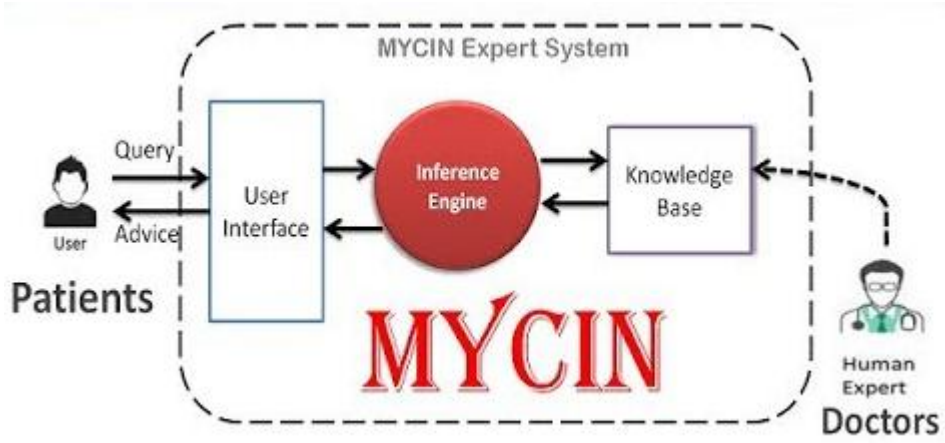
国家自然科学基金重点国际合作项目 (批准号: 61620106010) 资助

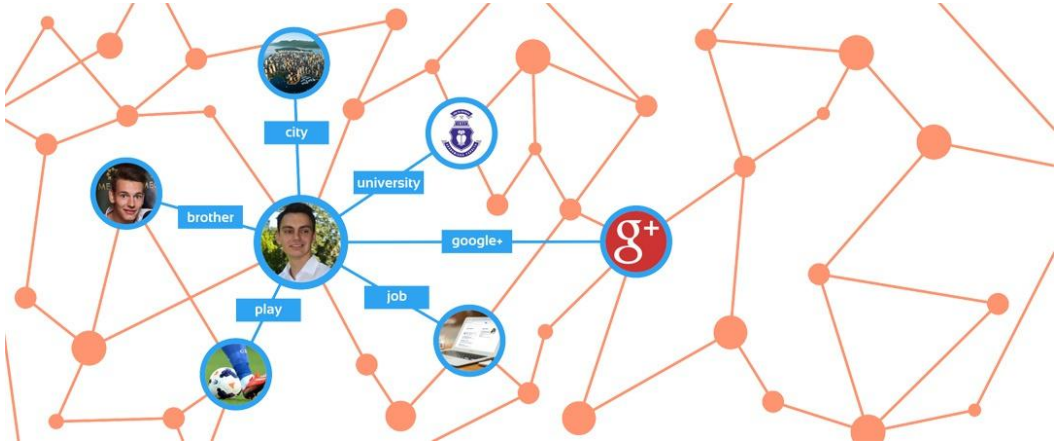
# 第一代人工智能

- 符号主义：
  - 对人类思维过程的数学整理及计算实现
  - 定义符号与操作符号的规则
  - 基于搜索的推理方法
- 案例1：定理证明
- 案例2：专家系统
- 案例3：知识图谱




https://www.ituring.com.cn/book/tupubarticle/19224



# 符号 AI与机器学习

- 符号AI同样可进行学习

符号 AI 同样可以应用于机器学习, 把 “机器学习” 看成是基于知识的 (归纳) 推理. 下面以归纳逻辑编程 (inductive logic programming, ILP)[8] 为例说明符号 AI 的学习机制. 在 ILP 中正负样本 (具体示例)、背景知识和学习结果 (假设) 都以一阶逻辑子句 (程序) 形式表示. 学习过程是在假设空间中寻找一个假设, 这个假设应尽可能多地包含正例, 尽量不包含负例, 而且要与背景知识一致. 一般情况下假设空间很大, 学习十分困难, 不过有了背景知识之后, 就可以极大地限制假设空间, 使学习变成可行. 显然, 背景知识越多, 学习速度越快, 效果也越好. 为解决不确定问题, 近年来, 发展了概率归纳逻辑编程方法 (probabilistic inductive logic programming, PILP)[9]. 基于知识的学习, 由于有背景知识, 可以实现小样本学习, 而且也很容易推广到不同的领域, 学习的鲁棒性也很强. 以迁移学习 (transfer learning)[10] 为例, 可以将学习得到的模型从一种场景更新或者迁移到另一场景, 实现跨领域和跨任务的推广. 具体做法如下, 首先, 从学习训练的环境 (包括训练数据与方法) 出发, 发现哪些 (即具有某种通用性) 知识可以跨域或者跨任务进行迁移, 哪些只是针对单个域或单个任务的特定知识, 并利用通用知识帮助提升目标域或目标任务的性能. 这些通用知识主要通过以下 4 种渠道迁移到目标域中去, 即源域中可利用的实例, 源域和目标域中可共享的特征, 源域模型可利用的部分, 源域中实体之间的特定规则. 可见, 知识在迁移学习中起关键的作用, 因此, 符号 AI 易于跨领域和跨任务推广.

# 符号AI的特点

符号 AI 有坚实的认知心理学基础, 把符号系统作为人类高级心智活动的模型, 其优势是, 由于符号具有可组合性 (compositionality), 可从简单的原子符号组合成复杂的符号串. 每个符号都对应着一定的语义, 客观上反映了语义对象的可组合性, 比如, 由简单部件组合成整体等, 可组合性是推理的基础, 因此符号 AI 与人类理性智能一样具有可解释性和容易理解. 符号 AI 也存在明显的局限性, 目前已有的方法只能解决完全信息和结构化环境下的确定性问题, 其中最具代表性的成果是 IBM "深蓝"国际象棋程序, 它只是在完全信息博弈 (决策) 中战胜人类, 这是博弈中最简单的情况. 而人类的认知行为 (cognitive behavior), 如决策等都是在信息不完全和非结构化环境下完成的, 符号 AI 距离解决这类问题还很远.

# 第二代人工智能

- 基于数据进行学习
- 神经网络为代表
- 概念和知识分布式、黑箱表示

## 2 第二代人工智能

感官信息 (视觉、听觉和触觉等) 是如何存储在记忆中并影响人类行为的? 有两种基本观点, 一种观点是, 这些信息以某种编码的方式表示在 (记忆) 神经网络中, 符号 AI 属于这一学派. 另一种观点是, 感官的刺激并不存储在记忆中, 而是在神经网络中建立起 "刺激 – 响应" 的连接 (通道), 通过这个 "连接" 保证智能行为的产生, 这是连接主义的主张, 连接主义 AI 就是建立在这个主张之上. 1958 年罗森布拉特 (Rosenblatt) 按照连接主义的思路, 建立一个人工神经网络 (artificial neural network, ANN) 的雏形 —— 感知机 (perceptron) [13,14]. 感知机的灵感来自于两个方面, 一是 1943 年麦卡洛克 (McCulloch) 和皮特 (Pitts) 提出的神经元数学模型 —— "阈值逻辑" 线路, 它将神经元的输入转换成离散值, 通常称为 M-P 模型 [15]. 二是来自于 1949 年赫布 (D. O. Hebb) 提出的 Hebb 学习率, 即 "同时发放的神经元连接在一起" [16]. 感知机如图 1 所示.

$$y = \begin{cases} 0, \text{ if } \sum_j w_j x_j \leqslant b, \\ 1, \text{ if } \sum_j w_j x_j > b, \end{cases} \tag{1}$$

其中 $b$ 为阈值, $w$ 为权值.

# 神经网络中的重要元素

刚刚起步的连接主义 AI 跌入低谷达 10 多年之久. 在困难的时期里, 在许多学者的共同努力下, 30 多年来无论在神经网络模型还是学习算法上均取得重大进步, 逐步形成了深度学习的成熟理论与技术. 其中重要的进展有, 第 1, 梯度下降法 (gradient descent), 这本来是一个古老的算法, 法国数学家柯西 (Cauchy) [18] 早在 1847 年就已经提出; 到 1983 年俄国数学家尤里 · 涅斯捷诺夫 (Yurii Nesterov) [19] 做了改进, 提出了加强版, 使它更加好用. 第 2, 反向传播 (back propagation, BP) 算法, 这是为 ANN 量身定制的, 1970 年由芬兰学生 Seppo Linnainmaa 在他的硕士论文中首先提出; 1986 年鲁梅哈特 (D. E. Rumelhart) 和辛顿 (G. Hinton) 等做了系统的分析与肯定 [20]. “梯度下降” 和 “BP” 两个算法为 ANN 的学习训练注入新的动力, 它们和 “阈值逻辑”、“Hebb 学习率” 一起构成 ANN 的 4 大支柱. 除 4 大支柱之外, 还有一系列重要工作, 其中包括更好的损失函数, 如交叉熵损失函数 (cross-entropy cost function) [21]; 算法的改进, 如防止过拟合的正则化方法 (regularization) [22]; 新的网络形式, 如 1980 年日本福岛邦彦 (Fukushima) 的卷积神经网络 (convolution neural networks, CNN) [23,24], 递归神经网络 (recurrent neural networks, RNN) [25], 长短程记忆神经网络 (long short-term memory neural networks, LSTM) [26], 辛顿的深度信念网络 (deep belief nets, DBN) [27] 等. 这些工作共同开启了以深度学习 (deep learning) 为基础的第二代 AI 的新纪元 [28].

# 攻击样本与网络脆弱性

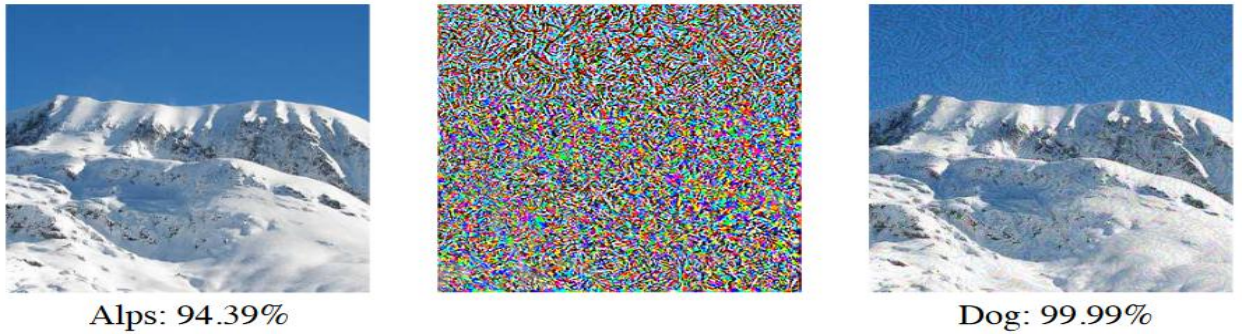



Figure 1. We show two adversarial examples generated by the proposed momentum iterative fast gradient sign method (MI-FGSM) for the Inception v3 [22] model. **Left column**: the original images. **Middle column**: the adversarial noises by applying MI-FGSM for 10 iterations. **Right column**: the generated adversarial images. We also show the predicted labels and probabilities of these images given by the Inception v3.

Dong Y P, Liao F Z, Pang T Y, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEEConference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, 2018. 9185–9193

# 攻击样本与网络脆弱性

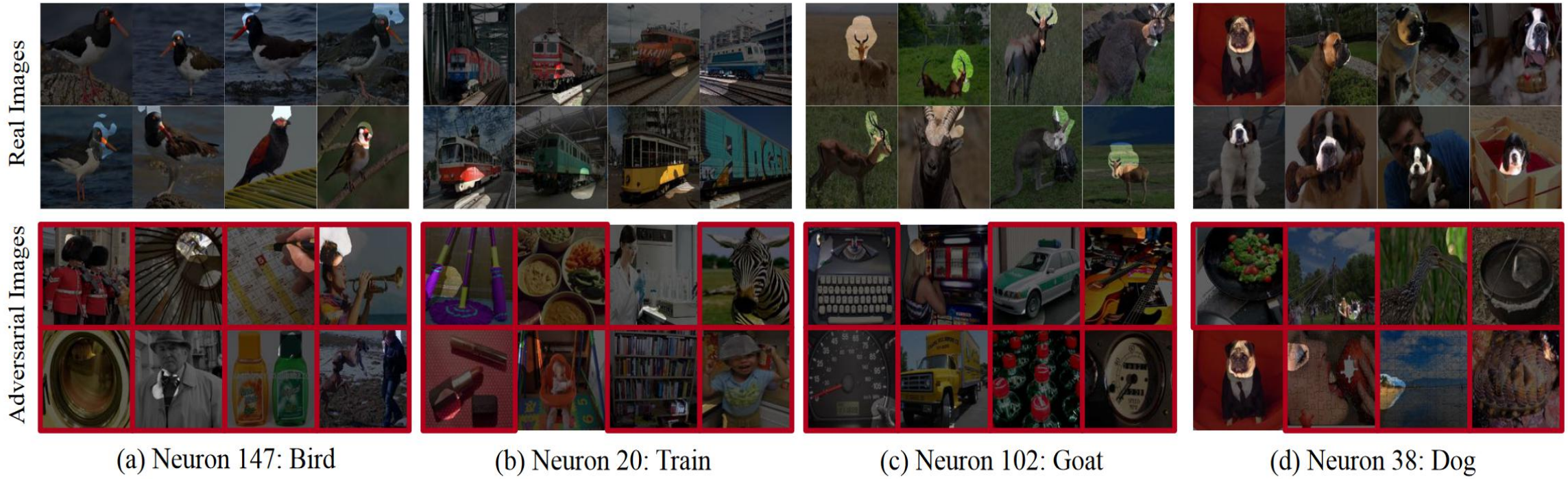



Figure 2. The real and adversarial images with highest activations for neurons in VGG-16 *pool5* layer. The neurons have explicit semantic meanings in real images, which do not appear in adversarial images. The adversarial images in red boxes have the target classes the same as the meanings of the neurons (*e.g.*, the model misclassifies the adversarial images in (a) as *birds*). The highlighted regions are found by discrepancy map [33]. More visualization results of AlexNet and ResNet-18 can be found in Appendix.

Dong Y P, Su H, Zhu J, et al. Towards interpretable deep neural networks by leveraging adversarial examples.In: Proceedings of the IJCAI workshop on AISC, Sydney, 2019. 1–6

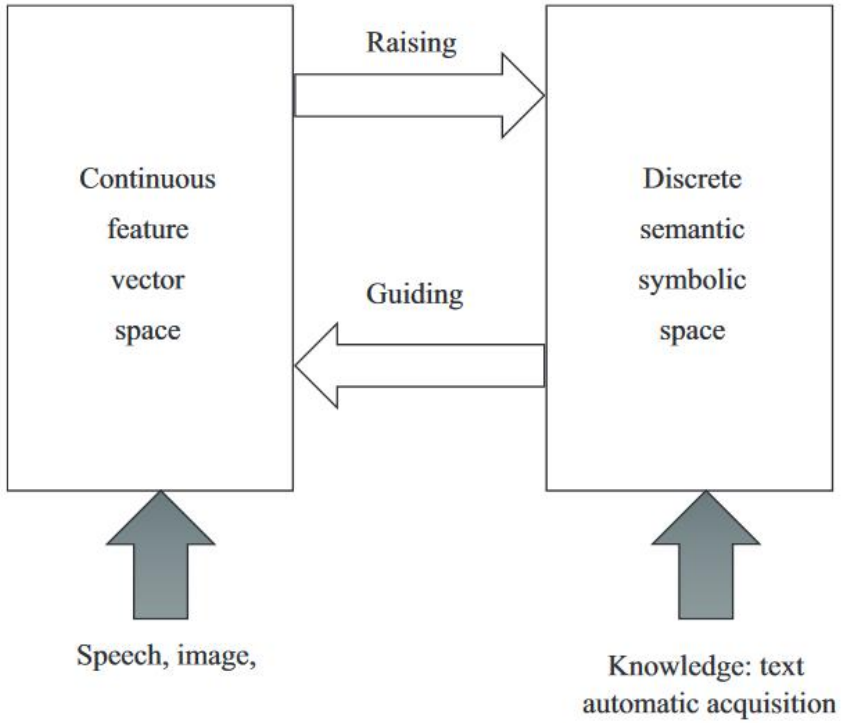
# 第二代人工智能的问题

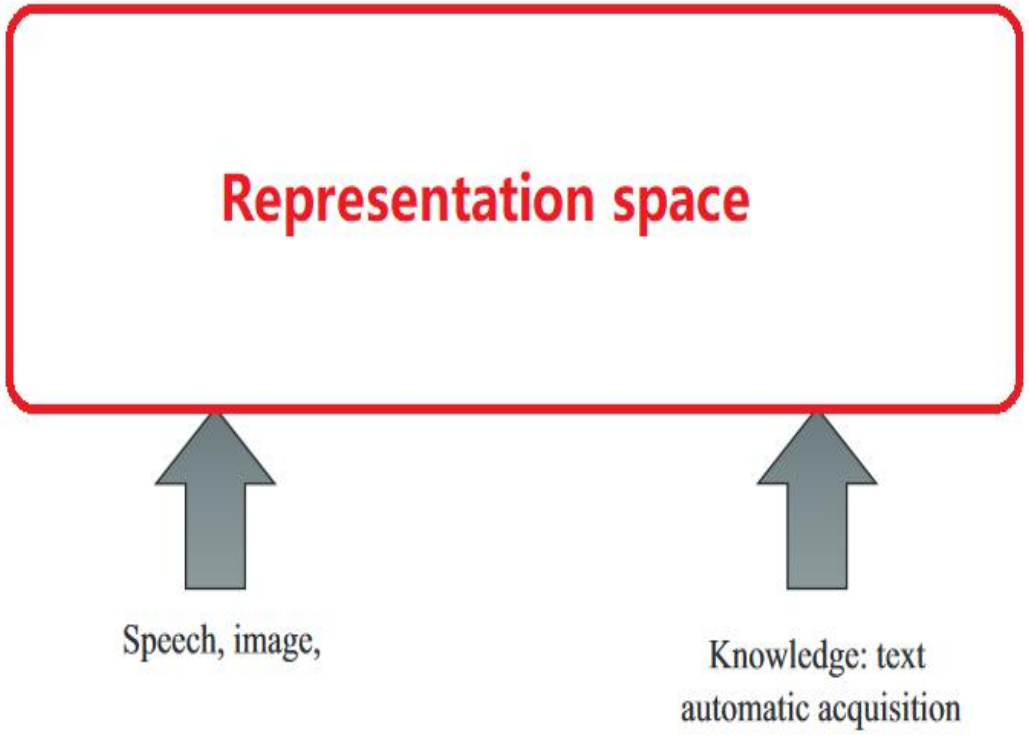
深度学习为何如此脆弱, 这样容易受攻击, 被欺骗和不安全. 原因只能从机器学习理论本身去寻找. 机器学习的成功与否与 3 项假设密切相关, 由于观察与测量数据的不确定性, 所获取的数据一定不完备和含有噪声, 这种情况下, 神经网络结构 (备选函数族) 的选择极为重要. 如果网络过于简单, 则存在欠拟合 (under-fitting) 风险, 如果网络结构过于复杂, 则出现过拟合 (overfitting) 现象. 虽然通过各种正则化的手段, 一定程度上可以降低过拟合的风险, 但是如果数据的质量差, 则必然会导致推广能力的严重下降. 此外, 深度学习的“黑箱”性质是造成深度学习推广能力差的另一个原因, 以图像识别为例, 通过深度学习只能发现重复出现的局部片段 (模式), 很难发现具有语义的部件. 文献 [33] 描述了利用深度网络模型 VGG-16 对“鸟”原始图像进行分类, 从该模型 pool 5 层 147 号神经元的响应可以看出, 该神经元最强烈的响应是“鸟”头部的某个局部特征, 机器正利用这个局部特征作为区分“鸟”的主要依据, 显然它不是“鸟”的不变语义特征. 因此对于语义完全不同的对抗样本 (人物、啤酒瓶和马等), 由于具有与“鸟”头部相似的片段, VGG-16 模型 pool 5 层 147 号神经元同样产生强烈的响应, 于是机器就把这些对抗样本错误地判断为“鸟”.

# 第三代人工智能

第一代知识驱动的 AI, 利用知识、算法和算力 3 个要素构造 AI, 第二代数据驱动的 AI, 利用数据、算法与算力 3 个要素构造 AI. 由于第一、二代 AI 只是从一个侧面模拟人类的智能行为, 因此存在各自的局限性. 为了建立一个全面反映人类智能的 AI, 需要建立鲁棒与可解释的 AI 理论与方法, 发展安全、可信、可靠与可扩展的 AI 技术, 即第三代 AI. 其发展的思路是, 把第一代的知识驱动和第二代的数据驱动结合起来, 通过同时利用知识、数据、算法和算力等 4 个要素, 构造更强大的 AI. 目前存在双空间模型与单一空间模型两个方案.

# 双空间模型与单空间模型



# 双空间模型

- 知识表示与推理
- 带有语义的感知
- 强化学习过程

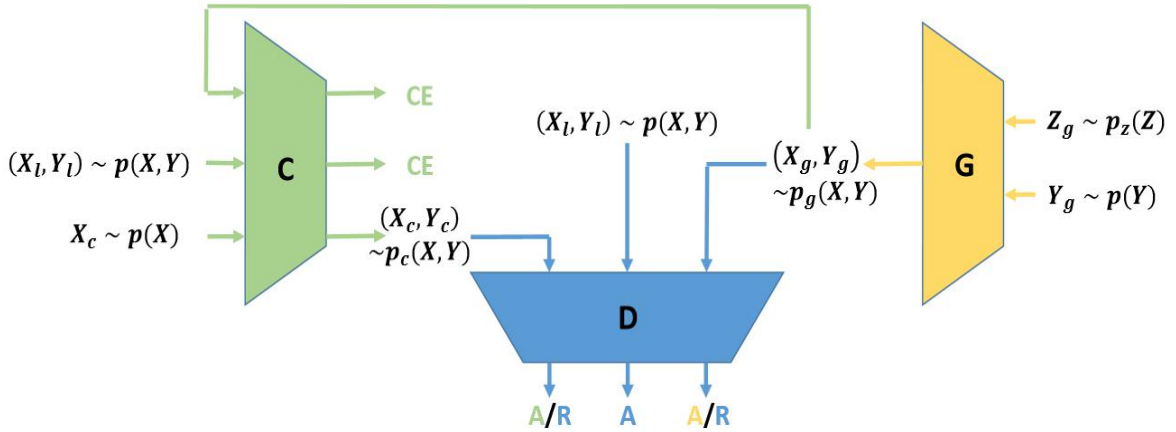


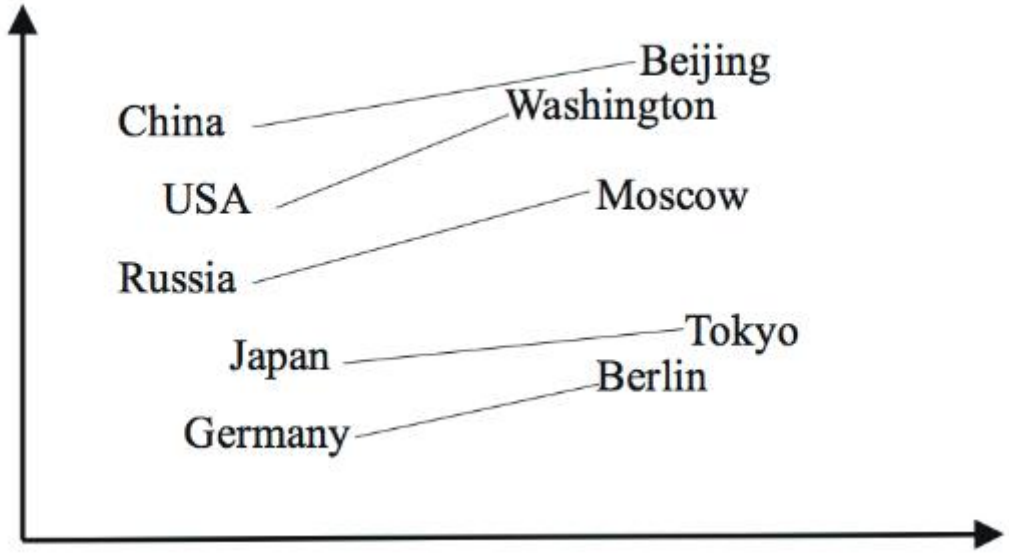

Figure 1: An illustration of Triple-GAN (best view in color). The utilities of $D$, $C$ and $G$ are colored in blue, green and yellow respectively, with "R" denoting rejection, "A" denoting acceptance and "CE" denoting the cross entropy loss for supervised learning. "A"s and "R"s are the adversarial losses and "CE"s are unbiased regularizations that ensure the consistency between $p_g$, $p_c$ and $p$, which are the distributions defined by the generator, classifier and true data generating process, respectively.

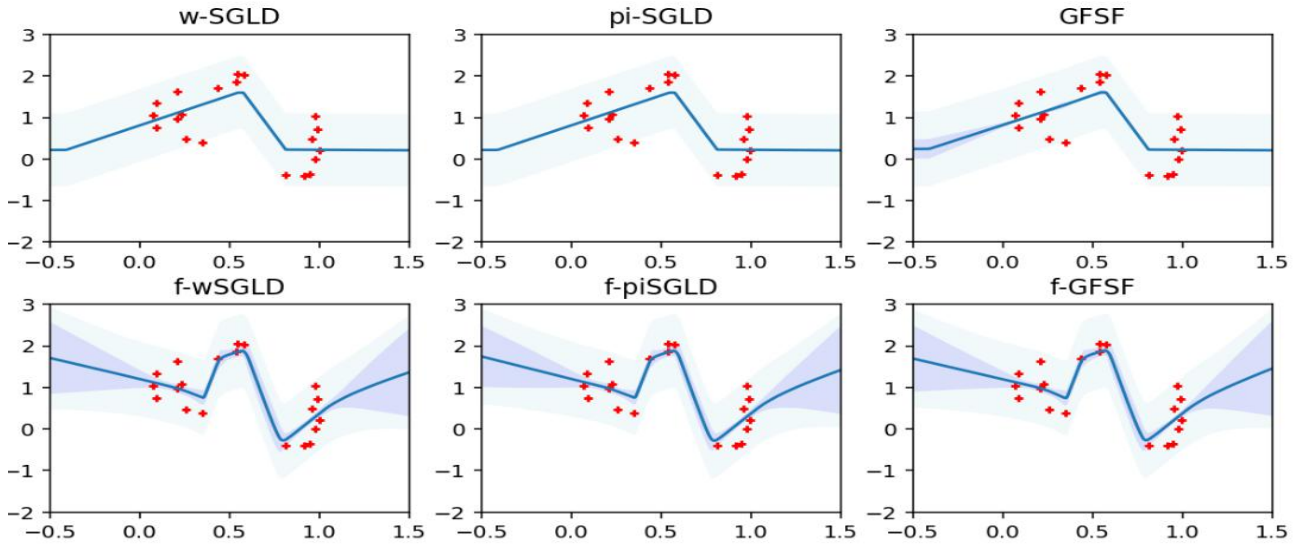
(a) Feature Matching (b) Triple-GAN

Li C, Xu K, Zhu J, et al. Triple generative adversarial nets. In: Proceedings of the Advances in Neural InformationProcessing Systems (NeurIPS), Long Beach, 2017. 4088–4098

# 单空间模型

- 符号向量化
- 深度学习方法改进
  - 可解释性问题
  - 鲁棒性问题
- 深度贝叶斯方法
  - 贝叶斯神经网络
  - 基于深度神经网络的贝叶斯模型
- 多维信息融合




Wang Z, Ren T, Zhu J, et al. Function space particle optimization for bayesian neural networks. In: Proceedings ofthe International Conference on Learning Representations (ICLR), New Orleans, 2019. 1–12

# 第三代人工智的实现途径

双空间模型模仿了大脑的工作机制, 但由于我们对大脑的工作机制了解得很少, 这条道路存在某些不确定性, 比如, 机器通过与环境的交互学习 (强化学习) 所建立的“内在语义”, 与人类通过感知所获取的“内在语义”是否一样, 机器是否也能具有意识? 等, 目前还不能肯定. 尽管存在这些困难, 但我们相信机器只要朝这个方向迈出一步, 就会更接近于真正的 AI. 单一空间模型是以深度学习为基础, 优点是充分利用计算机的算力, 在一些方面会表现出比人类优越的性能. 但深度学习存在一些根本性的缺点, 通过算法的改进究竟能得到多大程度的进步, 也存在不确定性, 需要进一步探索. 但是, 我们也相信对于深度学习的每一步改进, 都将推动 AI 向前发展.

考虑以上这些不确定性, 为了实现第三代 AI 的目标, 最好的策略是同时沿着这两条路线前进, 即三空间的融合, 如图 10 所示. 这种策略的好处是, 既最大限度地借鉴大脑的工作机制, 又充分利用计算机的算力, 二者的结合, 有望建造更加强大的 AI.

# 三空间模型

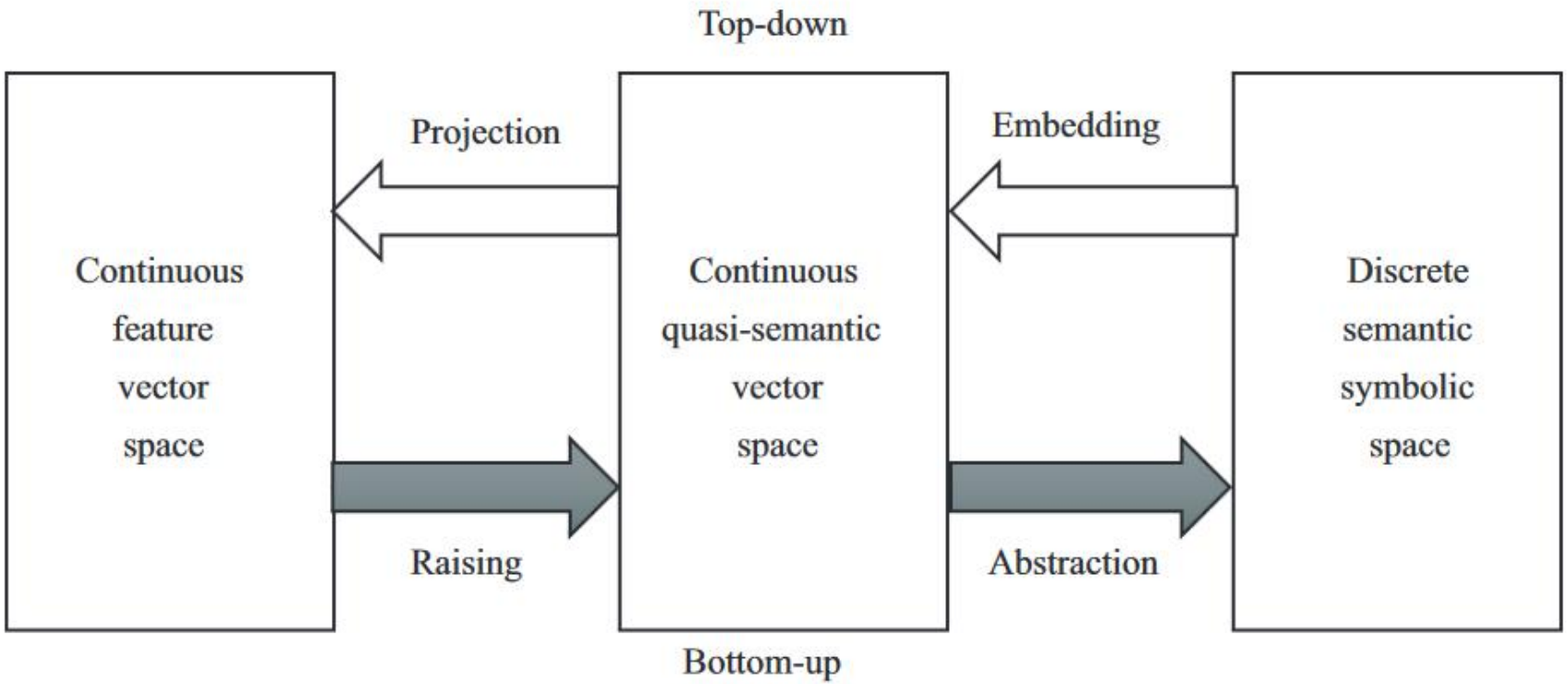



图 10　三空间融合模型

**Figure 10**　Triple-space integration model

# 讨论

- 从人工智能发展史得到什么启发?
- 当前人工智能核心问题是什么?
- 50年后人工智能是什么样子?