Toward the third generation of artificial intelligence

作者：张钹、朱军、苏航

自 2018 年成立以来，清华大学人工智能研究院本着“一个核心、两个融合”的发展战略，大力推动人工智能的基础理论和基本方法的源头性和颠覆性创新，在人工智能基础理论、关键技术和产学研合作等诸方面取得了创新成果。人工智能的序幕刚刚拉开，正剧正在上演。基础研究是科技创新的源头，尤其在当前复杂多变的国际环境下，更需要提升我国的原始创新能力，久久为功，努力实现人工智能领域更多“从 0 到 1”的突破。

清华大学人工智能研究院院长、中国科学院院士张钹教授在“纪念《中国科学》创刊 70 周年专刊”上发表署名文章，首次全面阐述第三代人工智能的理念，**提出第三代人工智能的发展路径**是融合第一代知识驱动和第二代数据驱动的人工智能，同时利用知识、数据、算法和算力等 4 个要素，提出双空间模型与单一空间模型两个方案，建立新的可解释和鲁棒的 AI 理论与方法，**发展安全、可信、可靠和可扩展的 AI 技术**，这是发展 AI 的必经之路。

迈向第三代人工智能

来源：清华人工智能研究院院长张钹院士专文

全文链接：http://scis.scichina.com/cn/2020/SSI-2020-0204.pdf

摘要：人工智能（Artificial Intelligence，简称 AI）在 60 多年的发展历史中，一直存在两个相互竞争的范式，即符号主义与连接主义（或称亚符号主义）。符号主义（即第一代人工智能）到上个世纪八十年代之前一直主导着 AI 的发展，而连接主义（即第二代人工智能）从上个世纪九十年代逐步发展，到本世纪初进入高潮，大有替代符号主义之势。今天看来，这两种范式只是从不同的侧面模拟人类的心智（或大脑），具有各自的片面性，不可能触及人类真正的智能。需要建立新的可解释和鲁棒的 AI 理论与方法，**发展安全、可信、可靠和可扩展的 AI 技术**。为实现这个目标，需要将这两种范式结合起来，这是发展 AI 的必经之路. 本文将阐述这一思想,为叙述方便，我们称符号主义为第一代 AI,称连接主义为第二代 AI,将要发展的 AI 称为第三代 AI.

关键词人工智能，符号主义，连接主义，双空间模型，单空间模型，三空间模型

1 第一代人工智能

人类的智能行为是怎么产生的，纽威尔（A.Newell）、西蒙（H.A.Simon）等[1~4]提出以下模拟人类大脑的符号模型， 即物理符号系统假设。这种系统包括:

（1） 一组任意的符号集，一组操作符号的规则集;

（2） 这些操作是纯语法（syntax）的，即只涉及符号的形式不涉及语义，操作的内容包括符号的组合和重组;

（3）这些语法具有系统性的语义解释，即它所指向的对象和所描述的事态。

1955 年麦卡锡（J.McCarthy）和明斯基（M.L.Minsky）等学者[5]，在达特茅斯人工智能夏季研究项目（the Dartmouth Summer Research Project on Artificial Intelligence）的建议中，明确提出符号 AI（artificialintelligence） 的基本思路: “人类思维的很大一部分是按照推理和猜想规则对‘词’（words）进行操作所组成的”。根据这一思路，他们提出了基于知识与经验的推理模型，因此，我们又把符号 AI 称为知识驱动方法。

符号 AI 的开创者最初把注意力放在研究推理（搜索）的通用方法上，如“手段–目的分析”（mean end analysis）、“分而治之“（divide and conquer）、” 试错 “（trial and error）法等，试图通过通用的方法解决范围广泛的现实问题。由于通用方法是一种弱方法，只能解决”玩具世界“中的简单问题，如机器人摆放积木，下简单的井字棋（tic-tac-toe）等，与解

决复杂现实问题相差很远。寻求通用 AI 的努力遭到了失败，符号 AI 于 20 世纪 70 年代初跌入低谷。

幸运的是，斯坦福大学教授费根堡姆（E. A. Feigenbaum）等及时改变了思路，认为知识，特别是特定领域的知识才是人类智能的基础，提出知识工程（knowledge engineering）与专家系统（expert systems）等一系列强 AI 方法，给符号 AI 带来了希望。他们开发了专家系统 DENDRAL（有机化学结构分析系统，1965~1975）[6]，随后其他学者相继开发了 MYCIN（血液传染病诊断和抗菌素处方，1971~1977）[7]，XCON（计算机硬件组合系统）等。不过早期的专家系统规模都较小，难以实用。

直到 1997 年 5 月 IBM 的深蓝（deep blue）国际象棋程序打败世界冠军卡斯帕诺夫（Kasparov），符号 AI 才真正解决大规模复杂系统的开发问题。费根堡姆和雷蒂（R. Raddy）作为设计与构造大型人工智能系统的先驱，共同获得 1994 年 ACM 图灵奖。

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以及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（**不同层次表示的知识，缺乏统一的知识表示模型**）。

在创建符号 AI 中做出重大贡献的学者中，除费根堡姆和雷蒂（1994）之外，还有明斯基（1969），麦卡锡（1971），纽威尔和西蒙（1975）共 6 位先后获得图灵奖（括号中的数字表示获奖的年份）。总之，第一代 AI 的成功**来自于以下 3 个基本要素**（以深蓝程序为例）：

第 1 是知识与经验，“深蓝“从象棋大师已经下过的 70 万盘棋局和大量 5~6 个棋子的残局中，总结出下棋的规则。另外，在象棋大师与深蓝对弈的过程中，通过调试”评价函数“中的 6000 个参数，把大师的经验引进程序。

第 2 是算法，深蓝采用 α−β剪枝算法，有效提高搜索效率。

第 3 是算力（计算能力），为了达到实时的要求，深蓝使用 IBM RS/6000 SP2, 11.38 G FLOPS（浮点运算/秒），每秒可检查 2 亿步，或 3 分钟运行 5 千万盘棋局（positions）。

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。目前符号 AI 方法的局限性主要是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，然而人类的认知行为（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。

以自然语言形式表示（离散符号）的人类知识，计算机难以处理，必须寻找计算机易于处理的表示形式，这就是知识表示问题。已有的产生式规则（production rules），逻辑程序（logic program）等知识表示方法，虽然计算机易于处理（如推理等），但都较简单，表现能力有限，难以刻画复杂和不确定的知识，推理也只限于逻辑推理等确定性的推理方法。知识图谱（knowledge graph）[11]、概率推理[12]等更加复杂的知识表示与推理形式都在探讨之中。**符号 AI 缺乏数学基础，除数理逻辑之外**，其他数学工具很难使用，这也是符号 AI 难以在计算机上高效执行的重要原因。

基于知识驱动的强 AI 只能就事论事地解决特定问题，有没有广泛适用的弱方法，即通用 AI，目前还是一个值得探讨的问题。此外，从原始数据（包括文本、图像、语音和视频）中获取知识目前主要靠人工，效率很低，需要探索有效的自动获取方法。此外，真正的智能系统需要常识，**常识如何获取、表达和推理还是一个有待解决的问题**。常识的数量巨大，构造一个实用的常识库，无异于一项 AI 的“曼哈顿工程“，费时费力。

2 第二代人工智能

视觉、听觉和触觉等感官信息是如何存储在记忆中并影响人类行为?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。

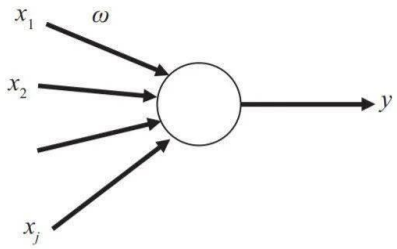


图 1　感知机

Figure 1　Perceptron

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。

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

其中 b 为阈值，w 为权值。

AI的创建者从一开始就关注连接主义的思路。1955 年麦卡锡等在达特茅斯（Dartmouth）AI 研究建议中写道"如何安排一组（假想的）神经元使之形成概念已经获得部分的结果，但问题是需要更多的理论工作"[5]，并把它列为会议的研讨内容之一。由感知机组成的 ANN 只有一个隐蔽层，过于简单。明斯基等[17]于 1969 年出版的书《感知机》中指出，感知机只能解决线性可分问题，而且即使增加隐层的数量，由于没有有效的学习算法，感知机也很难实用。明斯基对感知机的批评是致命的，使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（YuriiNesterov）[19]做了改进，提出了加强版，使它更加好用。

第 2，反向传播（backpropagation，BP）算法，这是为 ANN 量身定制的，1970 年由芬兰学生 SeppoLinnainmaa 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，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 beliefnets，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]。

第二代 AI 的学习理论有坚实的数学基础，为了说明这个基础，下面举一个简单的有监督学习的例子，有监督学习可以形式化为以下的函数回归问题：从数据库 D 中提取样本 $(x_i, y_i) \overset{i.i.d}{\longleftarrow} (X, Y)$，对样本所反映的输入–输出关系 f:X→Y 做出估计，即从备选函数族（假设空间）F={f_θ: X→Y;θ∈A}中选出一个函数 f^*使它平均逼近于真实 f。在深度学习中这个备选函数族由深度神经网络表示：

$$f^* = arg \min_{f_\theta \in F} E_D[l(f_\theta(x), y)] \qquad (2)$$

参数学习中有 3 项基本假设。（1）独立性假设：损失函数和备选函数族 F（或者神经网络结构）的选择与数据无关。（2）大容量假设：样本(x_i, y_i)数量巨大（n→∞）。（3）完备性假设：训练样本完备且无噪声。

如果上述假设均能满足，f^*将随样本数的增加最后收敛于真实函数 f。由此可见，如果拥有一定质量的大数据，由于深度神经网络的通用性（universality），它可以逼近任意的函数，因此利用深度学习找到数据背后的函数具有理论的保证。这个论断在许多实际应用中得到了印证，比如，在标准图像库 ImageNet（2 万类别，1 千 4 百万张图片）上的机器识别性能，2011 年误识率高达 50%，到 2015 年微软公司利用深度学习方法，误识率大幅度地降到 3.57%，比人类的误识率 5.1%还要低[29]。低噪声背景下的语音识别率，2001 年之前基本上停留在 80%左右，到了 2017 年识别率达到 95%以上，满足商品化的要求。

2016 年 3 月谷歌围棋程序 AlphaGo 打败世界冠军李世石，是第二代 AI 巅峰之作，因为在 2015 年之前计算机围棋程序最高只达到业余五段！更加令人惊奇的是，这些超越人类性能成果的取得，并不需要领域知识的帮助，只需输入图像原始像素、语音原始波形和围棋棋盘的布局（图像）！

深度学习的成功来自于以下 3 个要素：一是数据，以 AlphaGo 为例，其中 AlphaGo-Zero 通过强化学习自学了亿级的棋局，而人类在千年的围棋史中，下过的有效棋局只不过 3000 万盘; 二是算法，包括蒙特卡洛树搜索（Monte-Carlo tree search）[30]、深度学习和强化学习（reinforcement learning）[31]等; 三是算力，运行 AlphaGo 的机器是由 1920 个 CPU 和 280 个 GPU 组成的分布系统。因此第二代 AI 又称数据驱动方法。

在创建第二代 AI 中做出重大贡献的学者中，有以下 5 位获得图灵奖。他们是菲丽恩特（L. G. Valiant，2010）、珀尔（J. Pearl，2011）、本杰奥（Y. Bengio，2018）、辛顿（G. Hinton，2018）、杨立昆（Y. LeCun，2018）等。

早在 2014 年，深度学习的诸多缺陷不断地被发现，预示着这条道路遇到了瓶颈。下面仅以基于深度学习的图像识别的一个例子说明这个问题（材料引自本团队的工作）。

文献[32]表示利用基于动量的迭代快速梯度符号法（momentum iterative fast gradient sign method，MI-FGSM）对 Inceptionv3 深度网络模型实施攻击的结果。无噪声的原始图像——阿尔卑斯山（Alps），模型以 94.39%的置信度得到正确的分类。利用 MI-FGSM 方法经 10 次迭代之后生成攻击噪声 ,将此攻击噪声加进原图像后得到攻击样本。由于加入的噪声很小，生成的攻击样本与原始图几乎没有差异，人类无法察觉，但 Inceptionv3 模型却以 99.99%的置信度识别为「狗」。

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 pool5 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 pool5 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"。

3 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。

第一代 AI 是知识驱动，第二代 AI 是数据驱动；第三代 AI 是语义和逻辑驱动（感知行为和认知行为的语义空间不一致）

3.1 双空间模型

双空间模型如图 2 所示，它是一种类脑模型，符号空间模拟大脑的认知行为，亚符号（向量）空间模拟大脑的感知行为。这两层处理在大脑中是无缝融合的，如果能在计算机上

实现这种融合，AI 就有可能达到与人类相似的智能，从根本上解决目前 AI 存在的不可解释和鲁棒性差的问题。为了实现这种目标，需要解决以下 3 个问题。

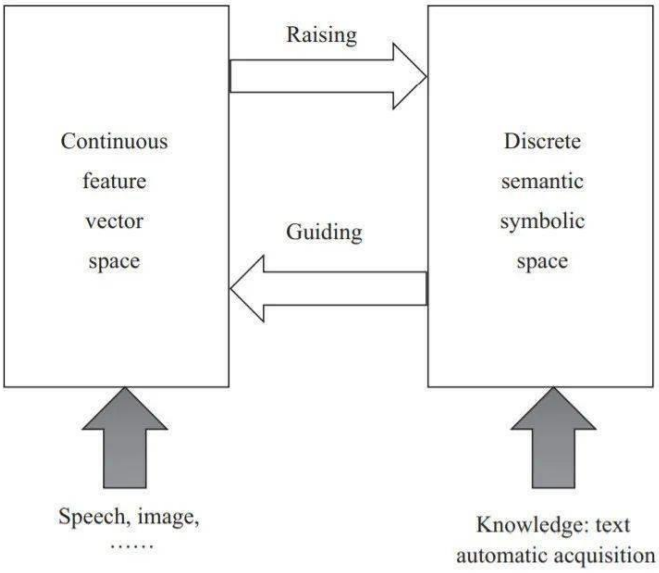


图 2　双空间模型

Figure 2　Dual-space mode

3.1.1 知识与推理

知识（包括常识）与推理是理性智能的基础，在第一代 AI 中,以物理符号系统模拟人类的理性智能，取得显著的进展，但无论在知识表示还是推理方法上都有大量的问题需要进一步探讨。下面以 IBM Deep QA 项目[34]为例说明最近的进展,之所以选择这个例子是因为基于 Deep QA 构成的 Watson 对话系统，在 2011 年 2 月美国电视“危险边缘“智力竞赛节目中,以压倒优势战胜全美冠军 K.詹宁斯(Ken Jennings)和 B.拉特 (Brad Rutter)，表明 Watson 是一个成功的 AI 系统。Watson 关于知识表示和推理方法的以下经验值得借鉴:

（1）从大量非结构化的文本自动生成结构化知识表示的方法；

（2）基于知识质量的评分表示知识不确定性的方法；

（3）基于多种推理的融合实现不确定性推理的方法

Watson 系统将「问答」(question-answer)看成是基于知识的从“问题“到”答案“的推理，为了达到人类的答题水平，计算机需要拥有与人类冠军一样甚至更多的知识，包括百科全书、主题词表、词典、专线新闻报道、文学作品等互联网上数量巨大（相当于 2 亿页的纸质材料）的文本，这些文本是非结构化的，而且质量参差不齐，需要把这些非结构化的文本自动转换为结构化且易于处理的表达形式。Watson 系统使用的表达形式为”扩展语料库“(expended corpus),它的生成步骤如下。首先给出基线语料库 (baseline corpus)判别种子文件(seed documents),根据种子文件从网上收集相关文件，并从中挖掘” 文本核“(text nuggets)，对文本核做评分，按照评分结果集成为最后的”扩展语料库 “。

（具有自学习能力）

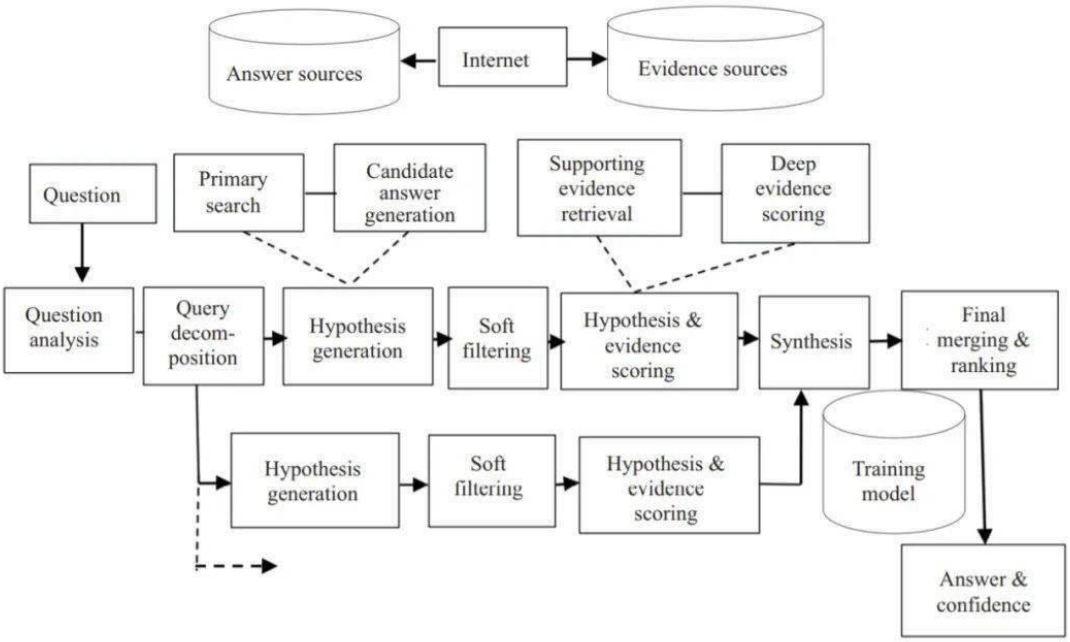


图 3 IBM Watson 系统

Figure 3 IBM Watson system

除自动生成的扩展语料库之外，Watson 的知识库中还包括已有的语料库，如 dbPedia，WordNet，Yago 等，以及人工编制的部分库。Watson 采用多种推理机制（多达百种）将“问题“转换为”答案“(见图 3）。先对”问题“做分析、分类和分解，根据分解的结果从答案源（语料库）中搜索假设与候选答案，经初步过滤之后，筛选出 100 个左右候选答案。再从证据源中收集证据，对候选答案进行评分，评估过程同时考虑数据源的可靠性，依据评分结果合成出几种候选答案，按照置信度大小进行排序，最后输出排序后的答案。此外，Watson 还通过 155 场与人类现场对决和 8000 次的实验，学习对“问题 “（自然语言）的理解。

3.1.2 感知

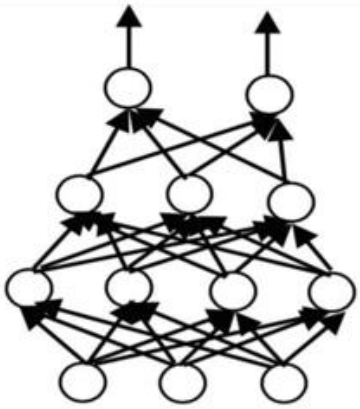
符号主义用符号系统作为人类心智的模型，以实现与人类相似的推理能力。但从认知的角度看，二者却有本质上的不同，即存在“符号基础问题“(symbol grounding problem)[35]。在物理符号系统中，客观世界的”对象“和”关系“等用符号表示，但符号本身并无语义，我们只好人为地给它们规定语义，也就是说是外部强加的”寄生语义“(parasitic semantics)，机器本身并不知道。这与人类大脑中存在的”内在语义“(intrinsic semantics)完全不同，人类大脑中的”内在语义 “，特别是”原子概念“和”常识 “，除极少数先天之外，主要是通过感官（视听等）或者感官与动作的结合自我习得的，即将感官图符式（iconic）表示或反映语义不变性的分类（categorical）表示转化为符号表示。这本来是深度学习要完成的任务，但很可惜，目前深度学习的模型并不能完成这项使命。因为深度学习所处理的空间是特征空间，与语义空间差别很大，它只能学到没有明确语义的”局部片段 “，这些片段不具备可组合性，因此，不能用来作为”物体“的”内在语义“表示（整体性）。换句话讲，目前的深度学习只能做到”感觉“(sensation),达不到感知应达到感知的水平，机器必须通过自我学习获取”物体“的语义部件(semantic parts),如「狗」的腿、头、尾等，才有可能通过这些部件的组合形成”狗“的不变”内在语义“。解决这个问题的基本思路是利用知识为引导，将感觉的信息从向量特征空间提升到符号语义空间（局部转向整体），如图 2 所示。这方面已经有不少的研究工作[36~39]，下面以本团队的工作阐述这方面工作的初步进展。

文献[40]描述如何利用一个三元生成对抗网络(triple generative adversarial networks，Triple-GAN)提高图像分类性能的方法。三元生成对抗网络由 3 部分组成: 分类器、生成器和鉴别器，分别用于条件化图像生成和半监督学习中的分类。

生成器在给定真实标签的情况下生成伪数据，分类器在给定真实数据的情况下生成伪标签，鉴别器的作用是区分数据标签对是否来自真实标记的数据集。如果设计好合适的效用函数，利用三元生成对抗网络，可以通过无监督（或弱监督）学习，让生成器（网络）学到样本中“物体“的表示（即先验知识），同时利用这个先验知识改善分类器的性能。

此项研究表明，通过 ANN 的无监督学习可以学到“物体“的先验知识，这就是”物体“（符号）的”内在语义“。利用这个具有”内在语义“的先验知识提高分类器的识别率，从根本上解决计算机视觉中存在的”检测“（where）与”识别“（what）之间的矛盾，实现小样本学习，提高鲁棒性和推广能力。

还可以从另外的角度思考，先回到深度学习所使用的人工神经网络（图 4），以视觉为例，它与人类的视觉神经网络相比过于简单，既没有反馈连接、同层之间的横向连接和抑制连接，也没有稀疏放电、记忆和注意等机制。如果能够将这些机制引进 ANN，将会逐步提高计算机视觉的感知能力。由于我们对大脑视神经网络的工作原理了解得很少，目前只能沿着“脑启发计算“（brain inspired computing）的道路一步一步地往前探索。



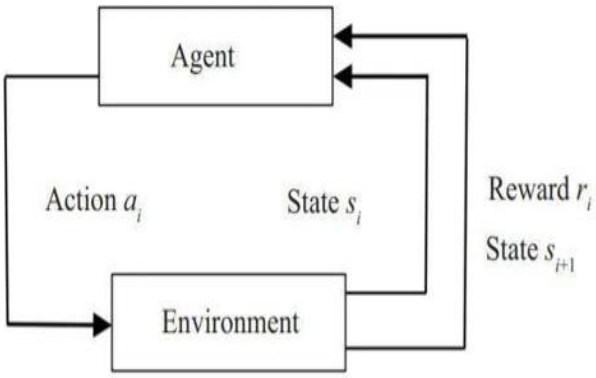


图 4、人工神经网络　　图 5、强化学习

Figure 4 Artificial Neural Network　　Figure 5、Reinforcement learning

目前有一些试探性的工作，有些效果但都不够显著。下面介绍本团队的一项研究。如文献[41]所述，将稀疏放电的原理运用到 ANN 各层的计算中。网络共 6 层，包括 Gabor 滤波和 Max 池化等，在各层的优化计算中加上“稀疏“正则约束项，稀疏性的要求迫使 ANN 选择最具代表性的特征。如果用背景简单的”人类“、”小汽车“、”大象“和”鸟“等图像作为训练样本训练网络，那么，神经网络的输出层就会出现代表这些”类别“的神经元，分别对人脸、小汽车、大象和鸟的轮廓做出响应，即提取了”整个物体“的语义信息，形成部分的”内在语义“。这种方法也只能提取部分的语义信息，还不能做到提取不同层面上的语义信息，如”整体“、”部件“和”子部件“等，达到符号化的水平，因此，仍有许多工作有待研究。（问题 1：人脑神经网络的由微观到宏观演化的介尺度机制不清）

3.1.3 强化学习

上面说过通过感官信息有可能学到一些基本知识（概念），不过仅仅依靠感官信息还不够，比如“常识概念“，如”吃饭“、”睡觉“等仅依靠感官难以获取，只有通过与环境的交互，即亲身经验之后才能获得，这是人类最基本的学习行为，也是通往真正 AI 的重要道

路。强化学习（reinforcement learning）就是用来模拟人类的这种学习行为，它通过”交互–试错“机制，与环境不断进行交互进而学习到有效的策略，很大程度上反映了**人脑做出决定的反馈系统运行机理，成为当前人工智能突破的重要方法**，在视频游戏[42，43]、棋牌游戏[44，45]、机器人导航与控制[46，47]、人机交互等领域取得了诸多成果，并在一些任务上接近甚至超越了人类的水平[48，49]。（问题 2：人脑做出决定的反馈系统运行机理）

强化学习通常看成是离散时间的随机控制过程，即智能体与环境的交互过程。智能体从起始状态出发，取得起始观察值，在 t 时刻，智能体根据其内部的推理机制采取行动之后，获得回报$r_t \in R$，并转移到下一个状态$S_{t+1} \in S$，得到新的观察$O_{t+1} \in O$。强化学习的目标是选择策略 π(s,a)使累计回报预期 $V^{\pi}(s):S\rightarrow R$ 最优。如果考虑简单的马尔可夫（Markov）决策过程，即后一个状态仅取决于前一个状态，并且环境完全可观察，即观察值 o 等于状态值 s，即 O=S;并假设策略稳定不变。如图 5 所示。以 AlphaZero 为例，智能体不依赖人类的标注数据，仅通过自我博弈式的环境交互积累数据，实现自身策略的不断改进，最终在围棋任务上达到了超越人类顶级大师的水平，代表强化学习算法的一个巨大进步[45]。

强化学习算法在选择行为策略的过程中，需要考虑环境模型的不确定性和目标的长远性。具体的，通过值函数也就是未来累积奖励的期望衡量不同策略的性能，即

$$V^{\pi}(s) = E_{\pi}\left[r_{t+1} + \sum_{k=1}^{\infty} \gamma_{r_{t+1+k}}^{k} \middle| S_t = s\right] = E_{\pi}\left[r_{t+1} + \gamma V_{s_{t+1}} \middle| S_t = s\right] \quad (3)$$

其中 γ∈[0，1]是折扣因子。值函数可以写成贝尔曼方程（Bellman equation）的形式。该方程表示了相邻状态之间的关系，可以利用其将决策过程划分成多个不同的阶段，其中某一阶段的最优决策问题可以利用贝尔曼方程转化为下一阶段最优决策的子问题。

强化学习的核心目标是选择最优的策略，使预期的累计奖励最大，即值函数取得最优值

$$\pi^{*}(s) = arg\max_{\pi} V^{\pi}(s)$$

需要指出的是，尽管强化学习在围棋、视频游戏等任务上获得了极大的成功，**但这些任务从本质上是相对”简单”的，**其任务的环境是完全可观察的、反馈是确定的、状态主要是离散的、规则是明确的，同时可以相对比较廉价地得到大量的数据，这些都是目前人工智能算法所擅长的**。但在不确定性、不完全信息、数据或知识匮乏的场景下，目前强化学习算法的性能往往会出现大幅度的下降，这也是目前强化学习所面临的重要挑战。**（问题 3：在知识和规则不确定的场景下如何利用经验记忆、内隐知识和注意力来引导实现自主进化学习）

其中的典型问题如下:

（1）**部分观测马氏决策过程中强化学习**：在真实的问题中，系统往往无法感知环境状态的全部信息，不仅需要考虑动作的不确定性，同时也需要考虑状态的不确定性。这就导致了部分感知的强化学习往往不满足马尔可夫环境假设。尽管相关的研究者近年来进行了大量的探索，但是部分观测马氏决策（partially observable Markov decision process，POMDP）仍然是强化学习中比较有挑战的问题。

（2）**领域知识在强化学习中的融合机制**：如何实现领域知识的融合在强化学习中同样是重要科学问题。对提高收敛速度、降低采样复杂度、改善模型迁移性和算法鲁棒性等具有重要意义。本团队针对这一问题，在领域知识指导的动作空间抽象压缩 [50]、结构设计[51] 等方面进行了初步探索，但是如何实现领域知识和强化学习框架的高效融合仍然是亟待解决的问题。

（3）**强化学习和博弈论的结合**：**博弈论和强化学习的结合**是近年来领域内研究的热点问题。二者的结合可以让多智能体之间的竞争和合作关系的建模变得更加直观和清晰，这其中包含了多智能体之间的零和/非零和、完全信息/非完全信息等多种不同的任务类型，尤其是在对抗性的任务中更具有研究和应用价值[43]。本团队前期在这方面也进行了探索性的研究，将

智能体对环境的探索建模成智能体和环境之间的博弈过程[52]，也是目前第一个在扩展型博弈、参数未知的场景下能够从理论上保证收敛的算法。

除此之外，**强化学习所面临的难题**还包括仿真环境和真实环境的差异、探索和利用的矛盾、基于模型的强化学习算法等诸多难点的问题，相比于监督学习所获得的成功而言，强化学习的研究还处于相对较为初级的阶段。

3.2 单一空间模型

单一空间模型是以深度学习为基础，**将所有的处理都放在亚符号（向量）空间**，这显然是为了利用计算机的计算能力，提高处理速度。问题在于深度学习与大脑的学习机制不同，在许多方面表现不佳，如可解释性和鲁棒性等。关键是要克服深度学习所带来的缺陷，如图 6 所示。下面讨论几个关键问题。

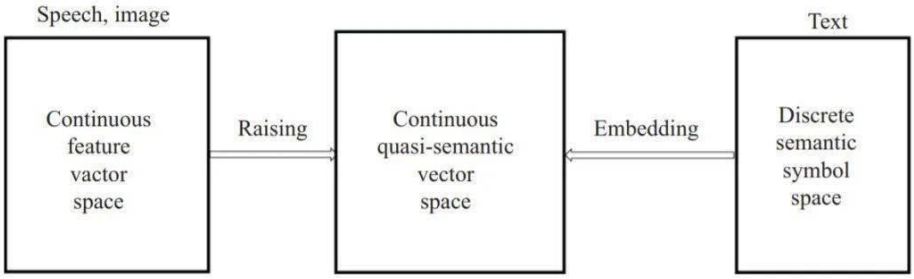


图 6　单一空间模型

Figure 6　Single-space Model

3.2.1 符号表示的向量化

知识通常以自然语言的离散符号形式表示，为了实现单一空间模型，首先要将符号表示的词、短语、句子和篇章等转换为向量，或将知识图谱转换为向量表示。关键是“词“的变换，即词嵌入(word embedding)。目前“词嵌入“已有各种方法，如 Word2Vec[53]和 GloVe[54]等。下面介绍 Word2Vec 采用的 Skip-gram[55]策略，说明词是如何由符号转换为向量的。

$$arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta) \qquad (5)$$

其中 w 是给定的目标词，c 是从其上下文中任选的一个词，$p(c|w;\theta)$是给定词 w 下，词 c 出现的概率。D 是从语料库中提取的所有 w−c 对，θ 是模型参数，式（5）进一步参数化后，得到,

$$p(c|w;\theta) = \frac{e^{v_c v_w}}{\sum_{c'\in C} e^{v_{c'} v_w}} \qquad (6)$$

其中，$v_c, v_\omega \in R^d$是词 c 和词 w 的向量表示，C 是所有可用文本。参数,i=1,2,...d，共 |C|×|W|×d 个。调整这些参数使式（5）最大化，最后得到所有词 w∈W 的向量表示$v_c, v_\omega \in R^d$。

这些词向量具有以下良好的性质，即“语义相似的词，其词向量也很相似“（见图 7）。变换后的词向量之所以具有上述良好的性质，**出自嵌入过程的以下假设**，两个词在上下文中同现的频率越高，这两个词的语义越可能接近，或者越可能存在语义上的某种关联。

嵌入词向量的这些特性，表明它带有语义信息，因此，称嵌入空间为准语义空间。式(5 ）是难计算的，可以采用深度神经网络等做近似计算。利用类似的嵌入法也可以把“短语 “” 句子 “和” 篇章 “或知识图谱等转换到具有准语义的向量空间中去[56]。

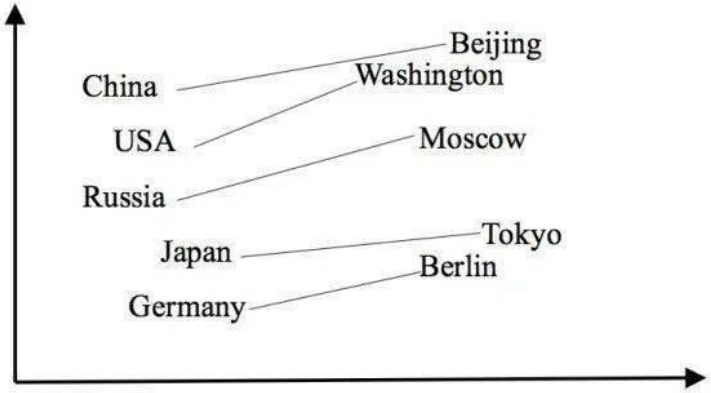


图 7　词嵌入图

Figure 7　Word embedding graph

向量形式的知识表示具有上述良好的性质，且可以与数据一样，使用大量的数学工具，包括深度学习方法，因此被大量应用于文本处理，如机器翻译等，取得明显的效果。下面以神经机器翻译（neural machine translation）为例予以说明[57，58]。神经机器翻译的基本思路是，给定源句子（比如中文），寻找目标句（比如英文）。神经翻译的任务是，计算词一级翻译概率的乘积，

$$p(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta}) = \prod_{j=1}^{J} p\left(y_j \mid \boldsymbol{x}, \boldsymbol{y}_{<j}; \boldsymbol{\theta}\right),$$

其中 θ 是一组模型参数，是部分翻译结果。词一级的翻译概率可用 softmax 函数 f(·) 定义:

$$p(y_i|x_i, y < j; \theta) \propto \exp\left(f\left(v_y, v_x, v_{y<j}; \theta\right)\right) \tag{7}$$

其中是目标句中第 j 个词的向量表示，v_x 是源句子的向量表示，$v_{y<j}$是部分翻译句的向量表示，y=y_j，j=1，2...，J 是要找的目标句。

神经翻译模型的构造: 给定训练样本为一组「源句–目标句」对$\{\langle x^{(n)}, y^{(n)}\rangle\}_{n=1}^{N}$，模型训练的目标是最大化 log 似然:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\theta} \left\{ \sum_{n=1}^{N} \log p\left(\boldsymbol{y}^{(n)} \mid \boldsymbol{x}^{(n)}; \boldsymbol{\theta}\right) \right\}, \tag{8}$$

即选择一组模型参数 θ，使目标函数最大化。利用这个模型，通过式（7）计算（翻译）目标句子。

这种翻译方法尽管可以得到比传统方法错误率还低的翻译结果，但**它具有不可解释、会发生重大错误、鲁棒性差等深度学习方法的共性缺陷**。为克服这些缺陷，需要加入知识，通过先验知识或后验正则化等方式引入语言知识等。

3.2.2 深度学习方法的改进

基于深度学习的 AI 具有不可解释和鲁棒性差等缺陷，目前有许多改进工作。下面介绍本团队的一些工作。

（1）可解释性问题。可解释人工智能算法的研究近年来引起众多研究人员的关注。而人类理解机器决策过程的**核心难点是跨越数据特征空间和人类语义空间之间的鸿沟**[59]。

无论是早期的以手工特征为基础的算法，还是当前以特征学习为代表的深度学习，**其核心思想**是将观测样本映射到特征空间中，进而在特征空间进行分析，发现样本在特征空间不同区域内的规律，从而达到算法要实现的任务目标（如分类、回归等）。**与之不同的是**，人类的分析和决策是利用自身的背景知识，在语义空间当中完成。**但数据特征空间和人类的语义空间在结构和内涵上存在显著的区别，而可解释人工智能的最终就是要在二者之间架起一座桥梁，进而跨越二者之间的鸿沟。**

总体而言，相关的研究主要分为（i）模型的后解释技术（post-hoc explanation），也就是给定了人工智能的模型，通过可视化、交互技术等方式，分析给定模型的工作机理，为其决策结果寻找解释途径;（ii）可解释模型，即通过发展新的网络架构、损失函数、训练方式等，发展具有内在可解释性的新型人工智能模型。从整体来说，两类方法目前都在发展过程中，在可解释性的研究中具有重要作用。

可视分析是人工智能算法可解释的一种直观的思路。既然深度学习是“黑箱”学习法，内部的工作机理是不透明的，“不可解释”，如果利用可视化，打开“黑箱”，一切不就清楚了吗? 为了帮助机器学习专家更加理解卷积神经网络的工作机理，我们开发了 CNN Vis 这一可视分析工具[60]。CNN Vis 旨在帮助专家更好地理解与诊断深度卷积神经网络，作为一种混合可视化方法，综合应用了基于双聚类技术的边绑定方法，以及矩形布局算法、矩阵重排算法和有向无环图布局算法等。作为可视化领域的首批深度学习可视分析工作，该工作在工业界和学术界都引起了广泛关注。在此基础上，为了分析复杂神经网络的训练过程，我们以**深度生成模型**（对抗生成网络（generative adversarial networks，GAN）和**变分自编码器**（variational auto-encoder，VAE））为例，研究了如何帮助机器学习专家诊断训练过程中出现的常见问题。

解释模型的另外一个思路是利用部分统计分析的技巧，针对神经网络决策过程中的参数冗余性，对神经网络内部最后决策起到关键作用的子成分进行分析，得到复杂模型内部对决策起到最关键作用的核心部分。为了更高效发掘子网络，我们借鉴了**网络剪枝**（network pruning）思路，**提出一种普适的提取子网络的方法，而无需对模型从头进行训练**[61]。我们对网络中每一层都附加一组控制门(control gate)变量，在**知识蒸馏**[62](knowledge distillation)准则下优化该组变量控制各层输出通道，用以确定关键子网络。具体来说，令 p（y | x;θ）为具有权重参数 θ 的原始模型对于单个样本 X 所做出的预测概率。而我们想要提取参数为 θ_s 的关键子网络，其预测输出应为 q(y | x;θ_s)，应该与原模型输出结果在 Kull back-Leibler 散度度量下接近。因此，总体最小化目标函数为

$$L\left(\boldsymbol{\theta}_{\boldsymbol{s}} \mid \boldsymbol{x}\right)=KL\left(p(\boldsymbol{y} \mid \boldsymbol{x};\boldsymbol{\theta}) \| q\left(\boldsymbol{y} \mid \boldsymbol{x};\boldsymbol{\theta}_{\boldsymbol{s}}\right)\right)+\Omega\left(\boldsymbol{\theta}_{\boldsymbol{s}}\right),$$

其中Ω（θ_s）为稀疏正则项，即鼓励模型通过尽量少的激活神经元达到和原网络相似的性能。通过对关键子网络可视化分析，我们观察到对于样本特定子网络，各层控制门值表征形式随着层级增高而展现出类别区分特性。实验结果表明，对于类别特定子网络，其整体表征形式与类别语义之间有着密切联系。以上方法更多的关注是模型的后解释，也就是给定一个深度学习模型“强行”寻求对其决策过程的解释，**而这种解释是否符合神经网络的内在机理仍然是需要讨论的问题。**

由于深度学习模型的不可解释性是由于机器推理的特征空间和人类可理解的空间存在着本质的区别，因此，**深度学习要想实现可解释性就需要把机器特征空间和人类的语义空间联系起来。**本团队也在此方面进行了探索性研究[63]，主要针对如何将人类的先验知识融入到深度学习模型的训练中，使特征具有更加明确的语义内涵，从而能够做到决策的追溯。

具体的，在图文的联合分析中，我们利用文本信息中抽取出来的人类可理解的主题信息指导神经网络的训练过程，并对文本和图像/视频数据进行协同训练，引导神经网络训练得到人类可以理解的语义特征。具体的，通过在神经网络的目标函数中引入可解释的正则约束:

$$L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}) = -\log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{h}) + \lambda L_I(\varphi(\boldsymbol{x}), \boldsymbol{s}),$$

其中第 1 项是相关任务的损失函数，第 2 项是可解释正则约束。通过这种方法，可以在文本数据引导下，通过不同模态数据之间的信息互补性，利用可解释正则约束，提升深度学习模型的可解释性。

（2）鲁棒性问题。由于对抗攻击给深度学习模型带来的潜在的恶意风险，其攻击不但精准且带有很强的传递性，给深度学习模型的实际应用带来了严重的安全隐患，迫切需要增强深度学习模型自身的安全性，发展相应的深度学习防御算法，降低恶意攻击带来的潜在威胁[64]。具体来说，**目前的深度学习防御算法主要有两类思路。**

第 1 是基于样本/模型输入控制的对抗防御。这类方法的核心是在模型的训练或者使用阶段，通过对训练样本的去噪、增广、对抗检测等方法，降低对抗攻击造成的危害。其中去噪器由于不改变模型自身的结构和性质，具有「即插即用」的性质，引起了广泛的关注。但是由于对抗噪声的特殊属性，其形成的干扰效应往往可以随着神经网络的加深逐步放大，因此在普通的高斯噪声（Gaussian noise）上具有良好滤除效果的自编码器往往不能很好地滤除对抗噪声。

针对这一问题，本团队提出了基于高层表示引导的去噪器（HGD）[65]，通过高层特征的约束使得对抗样本与正常样本引起目标模型的上层神经元响应尽可能一致。将传统像素级去噪网络 DAE（denoising autoencoder）与 U-net 网络结构进行结合，到负噪声输出，用对抗样本加上负噪声可以得到去噪图片，即。研究表明该方法不仅去掉了一部分对抗扰动，还增加了一部分「反对抗扰动」，取得了非常好的防御效果，获得「NIPS2017 对抗性攻防竞赛」中对抗防御任务冠军，以及 2018 年在拉斯维加斯（Las Vegas）举办的 CAADCTF 对抗样本邀请赛冠军。

第 2 是基于模型增强的对抗防御。这类方法的核心是通过修改网络的结构、模型的激活函数、损失函数等，训练更加鲁棒的深度学习模型，从而提高对对抗攻击的防御能力。其中集成模型（ensemble）是近年来出现的一类典型的防御方法。针对经典集成防御由于各个子模型的相似性导致防御性能下降的问题，本团队提出自适应多样性增强训练方法（adaptive diversity promoting training，ADP）[66]。相比于经典集成模型，ADP 方法在训练函数中额外引入了多样性正则项，鼓励每个子模型在正确类别上决策一致，而在其他类别上预测不一致。由于其他类别包括所有潜在的对抗样本的目标类别，所以这种不一致性可以使得各个子模型难以被同时欺骗，从而增强集成模型的鲁棒性。具体来讲，在 ADP 方法中，为了保证每个子模型的最大预测都对应于正确的类别，这种多样性定义在每个子模型输出的非最大预测上，当不同子模型的非最大预测向量相互正交时，这种多样性取得最大值。具体的，其训练的目标函数为

$$\mathcal{L}_{\mathrm{ADP}}^{m}=\frac{1}{|D_m|}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in D_m}\left[\mathcal{L}_{\mathrm{ECE}}-\mathrm{ADP}_{\alpha,\beta}\right](\boldsymbol{x}_i,\boldsymbol{y}_i),$$

其中，$L_{ECE}=\sum_{k\in\{K\}}L_{CE}^{k}$；$L_{CE}^{k}$为每个子模型 k 的交叉熵（cross-entropy）损失函数。ADP_α，β（x，y）=α·H（F）+β·log（ED）是模型集成多样性的度量，鼓励不同的子模型形成尽量差异化的决策边界。实验结果表明，通过鼓励不同子模型的差异化决策性质，有效地提升了模型的对抗鲁棒性。但是，总体而言，目前多数的对抗防御方法是基于经验主义的，研究表明很多防御对抗样本的方法在很短的时间就会被后来的攻击算法攻破。其重要原因之一是深度学习只是在做简单的函数拟合，缺乏像人一样对问题的理解能力[67]。因此通过理解机器学习模型的内部工作机理，发展数据驱动和知识驱动融合的第三代人工智能理论框架，将成为提高人工智能算法鲁棒性的重要途径。

总体而言，目前多数的对抗防御方法是基于经验主义的，研究表明很多防御对抗样本的方法在很短的时间就会被后来的攻击算法攻破。**其重要原因之一是深度学习只是在做简单的函数拟合，缺乏像人一样对问题的理解能力**[67]。因此，通过理解机器学习模型的内部工作机理，发展数据驱动和知识驱动融合的第三代人工智能理论框架，将成为提高人工智能算法鲁棒性的重要途径。

3.2.3 贝叶斯深度学习

如图 6 所示，图像和语音等信息是在特征空间中处理的，这些特征语义信息很少，需要提取含有更多语义的特征，其中的一种解决办法是将知识引入深度学习。下面以贝叶斯深度学习为例，说明这一思路。

我们前面说过深度神经网络没有考虑数据观测的不确定性，这种不确定性的存在，以及对于数据背后物理背景的无知，使我们对深度学习结果的正确性难以判断。同时，在数据量有限但模型逐渐变大（如包括十亿甚至千亿参数）的过程中，模型的不确定性也变得更严重——存在很多模型在训练集上表现都很好，但在测试集上的表现差别很大。贝叶斯学习充分考虑了先验知识以及模型和数据的不确定性，而且还能从不断提供的数据（证据）中，加深对数据的了解，即根据新的证据实现增量式的学习，充分发挥知识在学习中的作用。不仅可以对学习结果的可信度做出判断，也因此提高了学习的效率和准确度。

贝叶斯学习（Bayesian learning）定义：给定观测数据 d∈D，按贝叶斯规则计算每个假设的概率，$p(h_i|d)=\alpha p(d|h_i)p(h_i)$,其中 D 是所有数据[12，68]。给定 d

$$p(X|\boldsymbol{d})=\sum_i p(X|\boldsymbol{d};h_i)p(h_i|\boldsymbol{d})=\sum_i p(X|h_i)p(h_i|\boldsymbol{d})$$

是对未知量 X 的预测，即通过观测数据确定各个假设的概率，再从各个假设确定未知量 X 的分布。其中的关键是假设先验 p(h_i)和给定假设 h_i 下数据 d 的似然 p(d|h_i)。贝叶斯预测（式（13））不管样本量大小，均可达到最优，但当假设空间很大时，式（13）的加法计算量太大（在连续情况下为积分），难以实际应用。通常需要采用近似算法，主要有两类近似方法——变分推断和蒙特卡洛采样[69]。另外，还有一些常见的简化有，（1）对 X 的预测不是利用所有的假设，而只利用其中让 p（h_i|d）最大化的一个 h_i，称为最大化后验（maximum a posteriori，MAP）假设。（2）假定 p（h_i）是均匀分布，问题就简化为，选择一个让 p(d|h_i)最大化的 hi，称为最大化似然(maximum likelihood，ML)假设。（3）

如果不是所有数据都可以观测，即存在隐变量，通常采用 EM（expectation maximization）算法[70]。该算法分为两步(式(14)),E 步: 利用观测的数据 x 和 θ^(i),计算 p(Z=z|x;θ^(i));M 步: 利用计算出来的 z 和 x，计算模型参数 θ^(i+1)。两个步骤交替进行，找到最终的模型参数 θ:

$$\theta^{(i+1)} = \arg\max_{\theta} \sum_{\boldsymbol{z}} p(\boldsymbol{Z}=\boldsymbol{z}|\boldsymbol{x};\theta^{(i)}) L(\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z}|\theta).$$

贝叶斯准则 (式 (12)) 是一个从先验分布和似然函数推断后验分布的过程，为了更灵活地考虑知识，我们团队提出了正则化贝叶斯 (regularized Bayesian inference, RegBayes) [71]，它基于贝叶斯定理的信息论描述 [72]，通过引入后验正则化，在变分优化的框架下可以灵活地考虑领域知识 (如基于逻辑表达式的知识 [73]) 或者学习任务优化的目标 (如最大间隔损失 [74]) 等.

更进一步的，贝叶斯深度学习是将贝叶斯学习的基本原理与深度神经网络的表示学习有机融合的一类方法，融合主要体现在两个方面，

(1) 用贝叶斯方法更好地学习深度神经网络 (如贝叶斯神经网络、高斯过程等)，包括计算预测的不确定性、避免过拟合等;

(2) 用深度神经网络作为非线性函数变换定义更加丰富灵活的贝叶斯模型，如图 8 所示，包括深度生成模型 (如 GAN, VAE，基于可逆变换的流模型等). 其中，

第 1 种融合早在 20 世纪 90 年代就被霍普菲尔德 (J. Hopfield)和辛顿指导博士生系统研究过 [75, 76],当时算力和数据都很有限，稍微大一点的神经网络都面临着严重的过拟合，因此，那时就开始研究用贝叶斯方法保护神经网络，并且选择合适的网络结构. 随着神经网络的加深，贝叶斯方法又引起了很多研究兴趣,主要进展包括对深度贝叶斯神经网络进行高效的(近似)计算，**需克服的困难主要**是深度网络过参数化(over-parametrization)带来的维数灾难. 在这方面,我们团队进行了深入研究，先后提出了**隐式变分推断**(implicit variational inference)算法[77,78],**在泛函空间进行粒子优化的推断算法** (functional variational inference) [79]等.

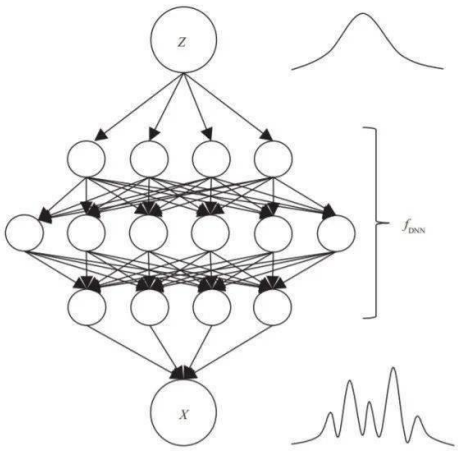


图 8 基于神经网络函数拟合的深度贝叶斯模型基本框架

Figure 8 A general framework of deep Bayesian models with DNN as function approximator

对于第 2 种融合，我们知道一个简单分布的随机变量 z 经过函数 f 变化之后，得到的变量 x = f(z)，具有更复杂的分布，当 f 是一个双射变换时，我们可以得到 x 分布的解析形

式$p(x) = p(z)|dz/dy|$，但是，在处理复杂数据时，f 是未知的，因此，我们希望从数据中进行学习. 利用深度神经网络的强大拟合能 力，我们将 f 定义成一个深度神经网络，通过一定的准则学习最优的 f_θ.如图 8 所示，这种想法被证明是非常有效的,已经发展了包括 VAE, GAN 以及基于流的模型 (flow-based models)，即使在完全无监督训练下,这些模型都可以产生高质量的自然图片或人脸等.

具体的，这几种模型的区别在于定义 x 的变化函数，在 VAE 中,其中 ϵ 是一个噪声变量 (如白噪声对应的标准高斯分布);在 GAN 和基于流的模型中，没有显式的噪声变量。这种区别带来了参数估计上的不同, VAE 和基于流的模型采用最大似然估计,而 GAN 定义了对抗学习的目标——「最大最小博弈」。同样的，这些模型虽然功能强大，但是给推断和学习也带来了很多挑战。例如, GAN 网络的训练过程经常是不稳定的，会遇到梯度消失或梯度爆炸等问题，我们团队最新的成果利用控制论对这一问题进行了分析研究，提出了有效的反馈机制，能够让 GAN 的训练更平稳[80]。此外，基于可逆变换的流模型往往受限于维数的约束，为此，我们提出了自适应数据增广的流模型[81],显著提升这类模型的表达能力。

基于上述介绍,能看出贝叶斯深度学习提供了一种强大的建模语言,将不确定性建模和推断与深度表示学习有机融合,**其关键挑战在于推断和学习算法**。幸运的是,近年来,在算法方面取得了很多突破进展 (如上所述)。同时,也发展了性能良好的概率编程库,支持贝叶斯深度学习模型的开发和部署。例如,我们团队研制的“珠算”[82]1) ,是最早的系统支持贝叶斯深度学习的开源库之一。在应用方面，贝叶斯深度学习的方法已在时间序列预测、半监督学习、无监督学习、小样本学习、持续学习等复杂场景下取得良好的效果。

3.2.4 单一空间中的计算

如图 6 所示，我们要在单一的向量空间中，对来自文本的嵌入向量和来自视听觉的特征向量进行计算，存在一定的难度. 因为文本中以符号表示的词，经嵌入之后变成向量时损失了大量语义，从视听觉中提取的特征，虽然我们尽量获取更多的语义，但一般情况多属底层特征，语义含量很少.

我们将以视觉问答 [83~85] 为例介绍这方面的初步尝试. 在视觉问答中既有图像又有文本，需要在单一的向量空间中同时处理，涉及单一空间模型的使用。以本团队关于“篇章级图文问答”研究工作为例予以说明[85] 。如图 9 所示,根据给定的图片,回答以下问题，“在大陆地壳下面有多少层 (类型)?”,除问题以文本形式表示之外，还有一个与图片相关的篇章「板块运动」。

首先通过词嵌入 (采用 Word2Vec 中的 Skip-gram 策略)，将“问题”与“篇章”中的以离散符号表示的词转换为向量. 图片经 ResNet 网络处理后，取 res5c 层的特征作为输出[55]，它是一组高维空间的特征向量。然后将“问题”和“篇章”中的词向量与“图片”输出的特征向量做融合，以预测“答案”。

为了更好地融合，通过注意机制，先找出“问题”和“篇章”中的“关键词”,这些关键词能够更好地反映“问题”的主题 (语义)。再依据关键词通过“空间注意机制”找出图片中关键区域的特征，因为这些特征更符合关键词向量所表达的主题，因此，融合效果会更好。这里采用的融合方法是双线性池化 (multi modal bilinear pooling) 方法。“图文问答”是选择题,备选方案有「1」,「2」,「3」三种,将融合后的向量与备选方案的向量相比较,取最接近的一个向量作为输出，这里是 「2」(向量)。图文问答目前达到的水平与人类相比相差很远，以「选择题」为例，目前达到的水平只比随机猜测略好。

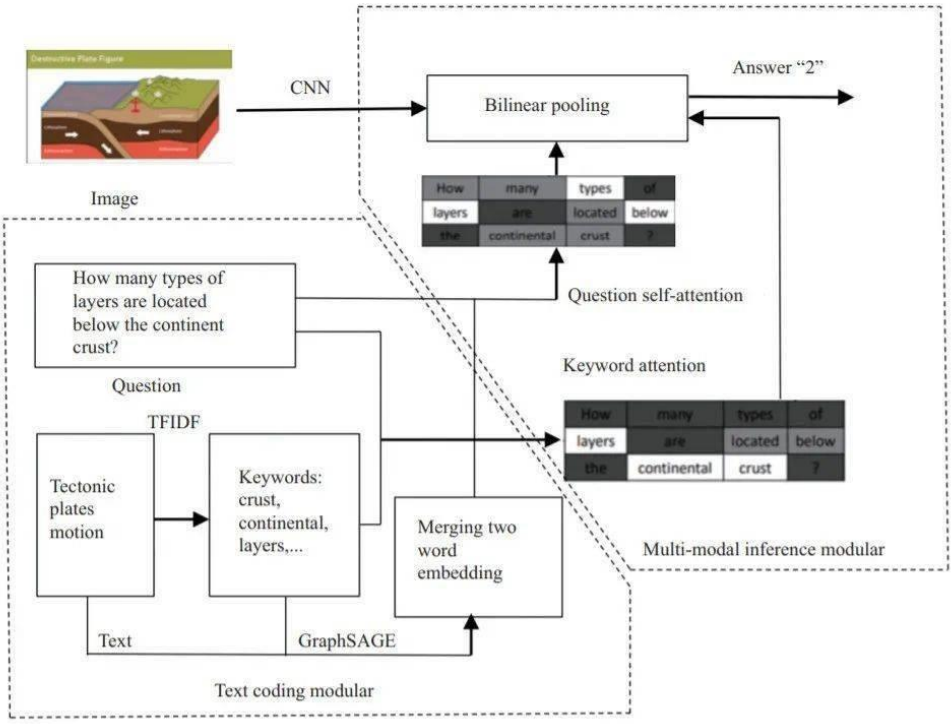


图 9 (网络版彩图) 篇章级图文问答系统框架

Figure 9 (Color online) The architecture of image-text question answer system

4 总结

为了实现第三代 AI 的目标,我们采用三空间融合的模型,即融合双空间与单空间两种模型,如图 10 所示。双空间模型采用类脑的工作机制,如果实现的话,机器就会像人类大脑的行为一样,具有可解释性与鲁棒性。此外,当把感觉(视觉、听觉等)信号提升为感知(符号)时,机器就具备一定的理解能力,因此,也解决了可解释和鲁棒的问题。当机器中的基本概念(符号)可由感知产生时,符号就有了基础(根基),符号与符号推理就有了内在的语义,从根本上解决了机器行为的可解释与鲁棒性的问题。单空间模型以深度学习为基础,存在不可解释与不鲁棒的缺陷,如果经过改进提高了其可解释性与鲁棒性,就从另外一个方向迈向第三代 AI.

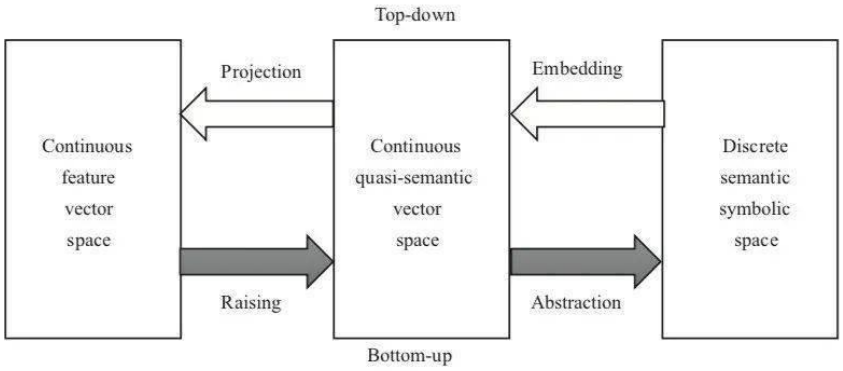


图 10 三空间融合模型

Figure 10 Triple-space integration model

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。

参考文献

1 Simon H A. Models of Man. New York: Wiley & Sons, 1957

2 Newell A, Simon H A. Computer science as empirical inquiry: symbols and search. Commun ACM, 1976, 19: 113–126

3 Newell A. Physical symbol systems. Cognitive Sci, 1980, 4: 135–183

4 Fodor J A. Methodological solipsism considered as a research strategy in cognitive psychology. Behav Brain Sci, 1980,3: 63–73

5 McCarthy J, Minsky M L, Rochester N, et al. A proposal for the Dartmouth summer research project on artificial intelligence. AI Mag, 1955, 27: 12

6 Lindsay R K, Buchanan B G, Feigenbaum E A, et al. Applications of Artificial Intelligence for Organic Chemistry: the Dendral Project. New York: McGraw-Hill Book Company, 1980

7 Buchanan B G, Shortliffe E H. Rule-Based Expert Systems: the MYCIN Experiments of the Stanford Heuristic Programming Project. Boston: Addison Wesley, 1984

Toward the third generation of artificial intelligence

Bo ZHANG*, Jun ZHU & Hang SU

Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, *China*

* Corresponding author. E-mail: dcszb@tsinghua.edu.cn

Abstract There have been two competing paradigms of artificial intelligence (AI) development since 1956, i.e., symbolism and connectionism (or subsymbolism). Both started at the same time, but symbolism had dominated AI development until the end of the 1980s. Connectionism began to develop in the 1990s and reached its climax at the beginning of this century, and it is likely to displace symbolism. Today, it seems that the two paradigms only simulate the human mind (or brain) in different ways and have their own advantages. True human intelligence cannot be achieved by relying on only one paradigm. Both are necessary to establish a new, explainable, and robust AI theory and method and develop safe, trustworthy, reliable, and extensible AI technology. To this end, it is imperative to combine the two paradigms, and the present article will illustrate this idea. For the sake of description, symbolism, connectionism, and the newly developed paradigm are termed as first-, second-, and third-generation AIs.

Keywords artificial intelligence, symbolism, connectionism, dual-space model, single-space model, triple-space model

Bo ZHANG was born in 1935. He graduated from the Department of Automatic Control, Tsinghua University, Beijing, in 1958. Currently, he is a professor in the Department of Computer Science &

Technology at Tsinghua University. His research interests include artificial intelligence and intelligent control. He is a member of the Chinese Academy of Sciences and the dean of the Institute for Artificial Intelligence, Tsinghua University.

Jun ZHU was born in 1983. He received his B.S. and Ph.D. degrees from the Department of Computer Science and Technology, Tsinghua University, where he is currently a professor. He was an adjunct faculty and postdoctoral fellow in the Machine Learning Department, Carnegie Mellon University. His research interests primarily include the development of statistical machine learning methods to understand scientific and engineering data obtained from various fields.

Hang SU was born in 1985. He is an associate professor in the Department of Computer Science and Technology at Tsinghua University. Before joining Tsinghua, he received his Ph.D. degree from Shanghai Jiao Tong University in 2014. His research interests lie in the development of computer vision and machine learning algorithms, particularly in robust and interpretable algorithms.