·合作专栏·

为何学习程序会被愚弄

围棋大战AlphaGo以比分4:1战胜了李世乭,令人惊叹技术进展的程度。但另一方面,机器输的那盘棋引起了各种猜疑,连“故意放水”的阴谋论都出来了。也有了解机器学习的人在小声嘀咕:莫不是李世乭那“神之一手”正好是AlphaGo的对抗样本?

这几年深度学习红得发紫,但并非无懈可击,“对抗样本”就是一个说不清的“隐疾”。深度学习最成功的应用之一就是图片识别,即能给出一张图片描绘的是什么。深度神经网络在经过一百多万张样本图片的训练后,已经可以正确识别大部分图片中的事物。图1是一些“对抗样本”,上面一排图片依次被识别为王企鹅、海星、棒球、电吉他,而下面一排图片依次被识别为火车车厢、遥控器、孔雀、非洲灰鹦鹉[1]。“对抗样本”是有意制造出来愚弄学习程序的,它的大量存在已被很多研究所确认,但对学习程序中了圈套的原因却仍是众说纷纭。笔者下文对此问题进行分析。

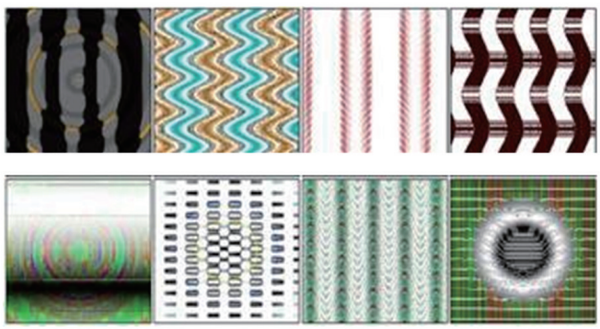

图1 对抗样本

凑出来的计算机“直觉”

机器学习通常把“学习”看作以“计算”为途径和目标。从计算的观点看,用计算机解决一个问题就是实现一个从输入到输出的函数。以数值函数为例,如果已知从输入X,Y到输出Z的关系是$Z=X^2-Y$,那么当输入(2,3)时,输出是1;输入(4,7)时,输出是9;输入(3,5)时,输出是4。只要把输入转化成输出的公式或算法是已知的,总能编出一个程序完成这个计算。这就是为什么会说“只要人能说清楚解法,计算机就能解决问题”的原因。而人工智能可以说是研究在说不清解法时怎么用计算机解决问题。机器学习会把上述问题看成“函数拟合”,就是说一般性公式是不知道的,只知道具体“样本”$f(2,3)=1$,$f(4,7)=9$等,而需要据此估算$f(3,5)$等于什么。

在人工智能历史上,曾有人试图用“猜函数”的办法解决此类问题,并声称重新发现了欧姆定律和开普勒定律等等。这条路的问题是满足条件的函数很多,而且样本常常包含误差,所以拟合程度太高也未必是好事。现在机器学习的主流办法不是“猜函数”,而是“凑函数”,即用一个一般性的方法根据样本造一个近似函数出来。既然目的是用这个函数来处理像$f(3,5)$这样的新问题,f本身是否能写成一个简单的公式又有什么关系呢?函数的建立方法是把样本推广到类似输入。既然输入(3,5)作为一个点正好在(2,3)和(4,7)中间,不妨认为$f(3,5)$是$f(2,3)$和$f(4,7)$平均值5。因为“正确公式”是不存在的,这个预测完全合理。至于是否符合事实,则要实际检验后才知道了。

人工神经网络就是一个通用的“函数拟合器”。很多人顾名思义地认为人工神经网络就是“像人脑一样工作”,其实它的设计只是从人脑得到一些灵感,而其真正的吸引力来自其可塑性,即可以通过调整内部参数变成各种各样的函数。一开始参数是随便定的,所以可能得到$f(2,3)=0$这样的结果。这时样本所提供的正确答案1会被学习算法用来调整网络参数以达到希望的输出(不一定正好是1,但足够接近)。下面再用同样的办法训练$f(4,7)=9$。比较麻烦的是后面的样本所引起的参数调整会破坏对前面样本的处理结果,所以当所有样本处理完成后还必须回头重来。这个“训练”过程一般要重复很多遍才能让同一组网络参数同时(尽可能)满足所有样本的要求。

了解这种训练过程,就可以看到AlphaGo实际上是不能像人那样通过复盘来总结前一盘棋的经验教训,并马上用在后一盘棋中的,而是需要把很多(甚至所有)以前的样本再整个过一遍(甚至多遍)以保证从这盘棋中学到的知识不会破坏以前的成果。如果说人类学习一般是“增量式的局部修改”(哪里错了改哪里),那深度学习就是“总量式的全局修改”(用样本总体定全部参数)。

深度学习能许诺DeepMind进军医疗的未来吗

一个人工神经网络通过训练得到的知识是分布在其所有网络参数之中的,这也就导致了其结果的难以理解。AlphaGo在某一个时刻为什么走某一步,这是其系统中成千上万个参数共同决定的,而这些参数的值是系统全部训练历史的产物。即使我们真能重现它的完整历史,也往往无法把它的某个决定的“原因”用我们能理解的概念来描述清楚。有种观点认为这说明机器成功地拥有了“直觉”,但这事还需要从两方面看。和传统计算系统相比,能从大量样本中总结出某种规律自然是进步,但完全说不清这种规律不能说是个优点。人的许多信念由于来源复杂以至于我们自己也说不清,因此以“直觉”称之,但并非所有信念都是这样来路不明的。

深度学习,以至于整个主流机器学习,在很大程度上接受这种把学习系统看成一个“黑箱”的做法,美其名曰“端到端”学习,意思是说“我只要实现你要的函数就行了,你管我怎么做的呢”。这种办法自然有它的优点(比较省心),但一旦出了问题就很难说清到底是怎么回事,也就更难做相应的改进了。AlphaGo的昏招和上面的对抗样本都是这方面的例子。

那么这个问题要不要紧呢？这就要看应用领域了。对围棋来说,程序超过人类已成定局,个别比赛的胜负已经无关大势了。即使程序确有类似于对抗样本的死穴存在,也不是棋手在对局时容易利用的,而可能是要靠运气来碰。现在发现的对抗样本,如参考文献[1]中展示的,都是用另一个学习技术"遗传算法"经过大量计算造出来的。这个办法能否用于围棋还是个问题。另一方面,程序的棋路难以理解或模仿,这对围棋界自然是个遗憾,但不影响它的胜率。所以这些问题对围棋程序都不是致命的。

现在AlphaGo的开发者DeepMind号称要进军医疗领域,那可就完全是另一回事了。如果一个机器诊断系统在收集了和你有关的信息后直接就给你开药,而其全部理由就是"这是我的直觉",你能接受吗？当然,这个系统以往的成功率可能高达95%,尽管它也曾经给头疼的病人开过脚气药。但问题是你怎么知道自己不是那5%呢？更何况基于大数据的统计性诊断对常见病、多发病会很有效,但对特殊病例的处理能力就没有保证了,因此成功率大概也到不了95%。由于这个问题事关机器学习的基本假设和框架,因此不是修改算法细节所能解决的。DeepMind可能要另辟蹊径才有希望。

总的说来,以深度学习为代表的机器学习技术对某一类问题是很有效的,前提是这类问题可以被看成一个确定的输入输出关系,而且可以收集到大量有代表性的样本。此外,要有足够的训练时间,而且可以容忍一定量的错误结果和解释的缺位。如果在某个问题上样本不足或者只能逐步收集,答复有时间要求,中间结果要可理解并可直接修改,那么这项技术就不足以解决问题。

"黑箱"技术之外的天地

目前业界对深度学习的局限性认识严重不足,而AlphaGo的胜利更在大众中造成了深度学习可以解决各种学习问题的假象。实际上,学界这两年对深度学习的态度已经从兴奋追捧转为冷静审视,而深度学习的领军人物也纷纷出面否认了"这项技术已解决了人工智能核心问题"的说法。

上面所说的问题存在于深度学习、人工神经网络、机器学习之中,但有关结论不能推广到整个人工智能领域,因为不是所有学习技术都是做函数拟合的。

一个问题的解决过程总是由一系列步骤组成的。传统的计算系统是让设计者既指定个别步骤又指定整个过程,因此保证了解决的可靠性和可理解性,但这种系统毫无灵活性。在很大程度上,人工智能的基本目标就是给系统以灵活性和适应性。当然,设计者也不可能什么都不限定。函数拟合的办法可以说是"限定两端,放开中间",即以训练样本的形式约束系统的输入输出关系,但容许学习算法相对自由地选择实现这种关系的中间步骤和对非样本的处理结果;而另一种可能性恰恰相反,可以说是"限定步骤,放开过程",笔者的工作就属于这一类。

笔者设计的"纳思"系统[2]是个推理系统,其中每个基本步骤都遵循一个广义推理规则,包括演绎、归纳、归因、例示、修正、选择、比较、类推、合并、分解、变换、派生、决策等等。这些推理规则同时也实现着学习的功能,而且可以灵活地彼此衔接以完成复杂的任务。这里的基本假设是任意思维过程都可以被分解成这些基本步骤,而"学习"就是系统使用这些规则以经验为原材料建造、调整信念和概念体系的"自组织"过程。

把"学习"理解成"信念和概念的自组织"比理解成"函数拟合"更接近人的学习过程,同时也避免了目前机器学习中的很多问题。比如说,由于系统的每一步都遵循某一条规则,其结果就具有较好的可解释性。人工神经网络的结果常常只有数学意义上的解释,即"这个结论是由现有全部网络参数决定的",而"纳思"的结论一般有概念层面上的逻辑解释,即"这个结论是从某某前提中被某规则推出来的"。尽管如果一个信念的来源极其复杂,它也会被说成是"直觉"。

设计"纳思"这样的学习系统的难点是确定其中规则的合理性和完备性,以及在运用这些规则时平衡其灵活性和确定性。既然要让系统自己学习,那它就难免犯错,但这不能被用作对批评的挡箭牌。这里的关键问题是这些错误是否可以理解,以及系统能否从错误中学习以避免重蹈覆辙。

对抗样本之所以是一个大问题,不仅仅是由于这些样本导致了系统的误判(人在图片识别时也会误判),而是由于它们揭示了机器学习在"凑函数"的过程中所生成的中间结果和人感知过程中的逐层抽象结果有根本性不同。人在识别"企鹅""孔雀""鹦鹉"时的中间结果一般都包括对其头部、身体、翅膀等鸟类普遍特征的识别,而机器学习算法所抽取的特征则一般没有独立意义,只是对整个"端到端"拟合过程有贡献(提高正确率、收敛速度等等)。因此,如果一个样本和训练样本很不一样,这些中间结果有可能导致莫名其妙的输出。系统很难有效地从这种错误中吸取教训,因为谁都没法说它是哪一步错了。要避免这种问题,仅仅限定输入输出关系是不够的,中间结果也必须有不完全依赖于网络参数的意义,因此不能只考虑学习算法的数学特征而不顾其认知功能,而这正是逻辑框架优于黑箱模型之处。

整体上看,"深度学习"被高估了,但"人工智能"却被低估了。比如说有些评论半开玩笑地说:"能赢棋时反而成心输掉才算有智能,而这是计算机永远做不到的。"殊不知尽管AlphaGo的确不会这样做,但这一功能在技术上根本不是问题。一个通用智能系统中同时会存在着多个目标,而系统在采取某个行动前必须权衡该行动对各个有关目标的影响。在这种情况下,总体效果最好的行动可能牺牲某些次要目标。如果这样一个系统觉得成心输掉一盘能赢的棋可以实现一个更重要的目标(比如提高后面比赛的收视率),那它完全可能这么做。这种功能在包括"纳思"在内的通用智能系统中早已实现了,只不过尚未出现在应用技术中而已。既然这次围棋大战的结果出乎了绝大多数人的预料,希望尚存学习能力的人能以此为鉴,以后不要再仅仅因为你不知道怎么让计算机做某事就断言那事是不可能做到的。

参考文献

[1] Nguyen A, Yosinski J, Clune j. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images [C]. IEEE Conference on Computer Vision and Pattern Recognition, Boston, June 7-12, 2015.

[2] Wang P. Rigid flexibility: The logic of Intelligence[M]. Dordrecht: Springer, 2006.

(注:选自《赛先生》微信公众平台,原标题为《机器是如何被骗并骗人的》,本文略有改动)

作者/王培

作者简介:美国天普大学计算机与信息科学系,教学副教授。

(责任编辑　王丽娜)