智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

AI与人的新三定律

杨强

引用本文:
杨强. AI与人的新三定律[J]. 智能系统学报, 2020, 15(4): 811–817.
YANG Qiang. AI’s three new laws of robotics[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(4): 811–817.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202011037

AI 与人的新三定律

AI's three new laws of robotics

杨强[1,2]

（1. 深圳前海微众银行股份有限公司，广东 深圳，518000；2. 香港科技大学 计算机科学和工程学系，香港）

现在大家都在探讨 AI 的下一步，我认为 AI 的下一步离不开和人的关系，下面就从科幻开始讲起。在科幻小说里我们比较熟悉的是机器人的三定律，这是阿西莫夫在很多小说里提到过的。第一定律是如果制作出一个全自动的机器人，那么机器人首先不能伤害人类的个体；第二定律是机器人必须服从人给予的命令但不能违反第一定律；第三定律是在保证第一、第二定律的前提下，它要尽可能维护自己的生命不受到伤害。如果真有这样一个机器人，既能为人类服务，又能自主地做一些决定，在保证这些定律的前提下，那么世界就真的像电视剧《西部世界》里面的场景，人与机器人共同生活在一个社会里。但我们逐渐发现，无论是做人工智能，还是做人工智能驱动的机器人，事情并没有那么简单。

随着研究的深入，我们就会发现，第一，做人工智能是离不开人的。第二，做人工智能的机器学习也好，做模型也好，首先要无人化的 AI，就像下面这个图，是一个无人工厂，充满了无人车、无人机、无人商店……，但是，AI 需要人类做伙伴，人类也需要 AI 做伙伴。首先，AI 的运算结果是要解释给人类用户的。我们看到 AlphaGo 不是这样的，它不能给我们解释为什么走这步棋而不走那步棋。第二，AI 的运行问题是要能够让人类工程师来做纠错的。我们看到 AlphaGo 也没有这个功能，它里面有一些臭棋，到现在都没有办法去给他纠错。第三，AI 的流程是需要人类来监管的，但是 AlphaGo 是完全不受人的监管的，在棋盘世界里它可以自由地驰骋。最后，AI 的模型系统需要来解释因果是怎么推出来的，它也没有把这个解释赋予人类。也就是说，它在设计的时候就没有把人考虑进来，它就是一个全自动的过程，包括后面的发展像自学习 (self supervised learning)、AlphaGo Zero 都是往这个方向发展。这就不是今天在工业和社会上我们所想看到的。

我们想看到的是什么呢？我们总结了一些新的规律出来。首先，AI 要保护人的利益，这一点是毋庸质疑的。我们知道人有很多利益，其中用户隐私是一项重要的利益，这在设计 AlphaGo 时，还有以前很多的科学家可能是完全没有考虑到的，因为今天的人工智能很多都是大数据驱动的，大数据就要从不同的地方收集不同的数据、聚合不同的数据源，那么就或多或少地会侵犯到用户隐私。第二是说 AI 不仅需要保护人的安全，而且要保证模型也是安全的，不受攻击。这一点如何实现，我们的研究才刚刚开始，那么该如何防止恶意的或非恶意但不小心造成的对模型的攻击。第三是 AI 需要人类伙伴的理解并且要促成这种理解，也就是说 AI 模型是需要可解释性的，而且对于不同的人解释也应该不一样。

下面我们来逐条分析。首先是 AI 要保护用户的隐私，今天 AI 的力量来自大数据，但是我们周围更多的是小数据，可以看到在一些案例里面，比如在法律上的人工智能应用，每一个案例的收集都是旷日持久的，都需要很多的标注，需要很多的积累，最后才形成案例，因此，它的案例数量是不多的。第二个在金融领域的重要应用是反洗钱，洗钱这种金融来往和非洗钱相比就远远是小数据，所以每个案例都有具体的特点，这种案例也是不多的；第三是我们现在特别关心的人工智能在医疗方面的应用，但是医疗图像中高质

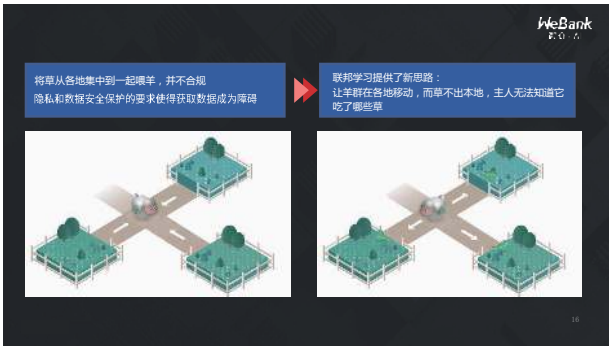
量且有标注的图像也是非常少的。

在金融领域的一个案例是第四范式，这是人工智能领域的一个平台公司，公司在金融的实践中发现小额贷款里样本数是很多的，但是大额贷款里这种样本是极少的。那么如何对大额贷款做一个风险控制的机器人，这种机器人能够大部分地解决风险控制的问题，这是我们非常重要的一个难题。我们在科幻小说里看到的这种无人车、无人机、机器人，很多都相当于一种端计算，它不断地与云端进行沟通。每一种终端的机器人就是广义的机器人，甚至是一个摄像头。它看到的周边的景象，只是反映了一个角度，只有把所有的这些不同角度汇合起来，我们才能有全面的了解。这种聚合的过程，就或多或少涉及到这种收集，对于每一个终端来说，他的数量就是有限的。这是问题的一个方面——我们的数据是有限的。

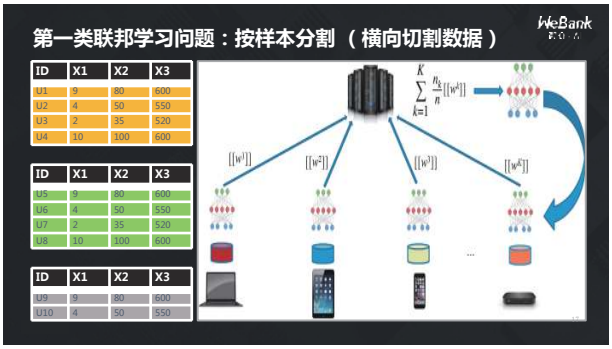
我们是不是可以很方便地把众多终端数据聚合起来呢？人类社会出台了很多保护个人隐私的法规，使得这种聚合变得不是那么直接，比方说欧洲在 2018 年 5 月出台的 GDPR(General data protection regulation) 法案，即个人通用数据保护条例，有各种各样的规定，使得数据是为一个目的收集的就不能轻易地用在另外一个目的的机器学习的训练上。否则就会因违法而遭受很严重的罚款。那么国内的监管也在趋严，像各大数据公司都在认真地学习国家的各项法规法案。一方面，我们需要更多的数据整合，另一方面，我们也需要遵从数据法规，所以就形成了这样一种理想和现实的分离：理想中我们是有大数据可以驱动整个人类的人工智能化，但是实际上我们所面临的却是众多的数据孤岛。不管是人们考虑自己的利益，还是考虑隐私的规定和条例，总之是至今大部分应用并没有直接有效的办法把这些孤岛连接起来。

为了解决这个问题，技术人员在寻找解决方案，其中一个办法是联邦学习 (Federated learning)，主要思想总结为数据可以保持在原地，但是模型通过孤岛、不同机构在加密的情况下沟通，这个模型会成长起来，它的效果就是数据可以被使用，但是各方看不到对方的数据。这听起来有点像天方夜谭，但是还是可以通过技术实现。要实现首先需确保用户的隐私得到保护，不仅是数据包含的隐私成分，还是模型的参数都要受到保护。同时我们关心模型的能力和效果，比如风险控制的效果，坏账率是不是低？比如分类的效果，准确率是不是高？效果也应该和数据完全聚集在一起，可能会相差一点点，但是相差应该不大。首先是要比单独用一方的数据进行建模的效果好。

在这些要求之下，我举一个非常直观的例子，就像我们要养一只羊，过去的做法是我们到各处去买草来给羊吃。买草的过程就是数据聚合的过程，在这个过程中，有各种各样的漏洞，使得数据本身的隐私被泄露，因为草是要离开草场的。现在我们让羊走动起来，让模型去访问不同的草场，草就不用出本地了，羊也得到了壮大，这就是联邦学习的主要思想。

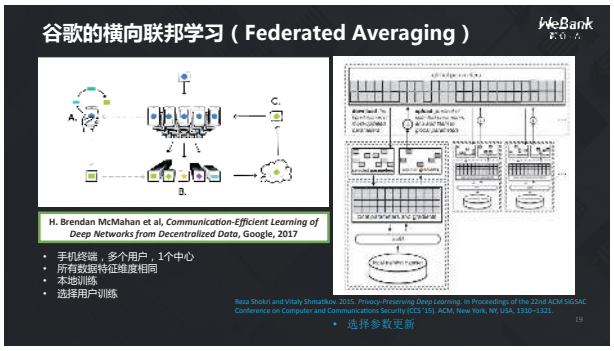


首先看一下开始举的例子，我们有不同的终端，让终端每天都聚集一些新的数据，希望模型不仅根据自己获得的新数据得到成长，而且可以利用其他终端得到数据来成长。我们看一下这个图的左边，每一种颜色、每一个表格就代表了一个数据集，它的特点是所有的终端上面的特征 (纵向看 $x1$~$x3$) 这些数据的维度特征基本是差不多的，比如手机上收集的数据特征几乎是大同小异的，但是他的用户和样本却是不一样的。我们把这一类问题叫做按照样本来分割，分割到每一个终端上。从表格的形式来看，横向在切割数据，所以称之为横向联邦学习。

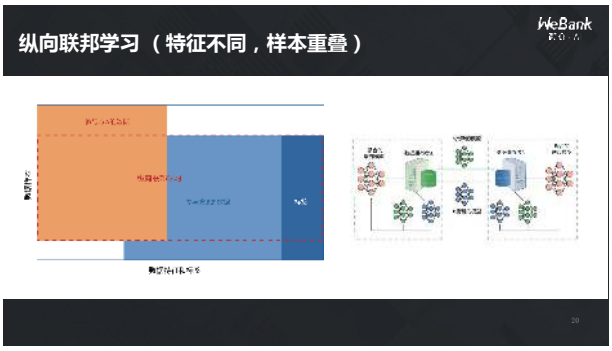


在这种情况下，应该怎样更新模型呢？google 在 2016 年提出了联邦平均算法，往云端传

递的消息只包含模型的参数，并且参数是受到加密保护的。它在云端就得到了更新，这个更新就是求一个平均值。更新后的模型被下发到各个终端，使得每个终端的模型得到更新，并且整个过程都不泄露本地的隐私和模型参数，所以这个做法被用在安卓系统上。这个理念的关键技术就是加密和解密的算法，现在有各种各样的算法可以支持这两件事。第一是要能保护所包进去的数据（数据可能是原始数据或模型参数），第二是允许在加密层之上进行一系列数学操作和数学运算，现在有一种加密方法叫做同态加密，它可以在多项式的基础上进行加密运算，使得在云端进行聚合和更新的效果可以在不暴露数据本身内容的情况下进行，以上是 google 在横向联邦学习的思路。从图中可以看到从 *A* 点一个用户数据过来，形成一个模型。模型上传之后，在云端有各种各样的模型包，可以在云端加以聚合，聚合形成了新的模型再下发到终端，就形成了一个闭环，下图的右边表示一个具体的过程。

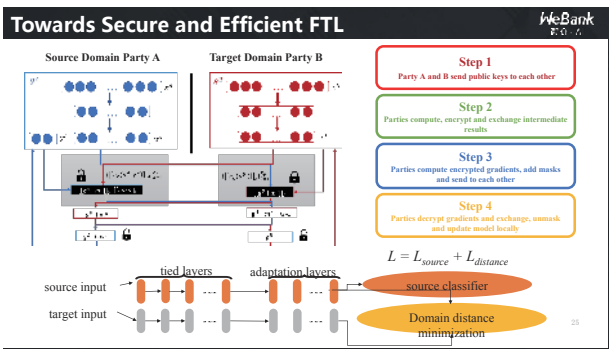


刚刚讲的是 To C 的应用，即一个很重的云端，面对几千万上亿的终端用户，因此称之为 To C 的应用。它的特点是按照样本来切割。一个对应的应用是 To B 的应用，对于企业的应用往往是这样的情况：有一些样本是重合的，但是他们的特征却不重叠。比如银行希望和互联网公司合作，做更好的风险控制模型，银行有一些金融相关的数据，互联网公司有一些用户行为相关的数据，由于为社会的监管和隐私要求，不能直接把数据互传，但是他们有意愿把模型建好，变成一个联合模型，这个模型应该具备各个方面的特征。这就相当于在进行纵向的切割，即在特征的维度进行切割。这就是特征不同但样本重叠的模型，这种模型叫做纵向联邦学习，这是我们在微众银行首先发起，现在国内很多大型企业都积极加入并投入研发。现在基本上形成了两大模型的方向：一个是横向联邦学习，是针对 To C 的方向；一个是纵向联邦学习，是针对 To B 的方向。

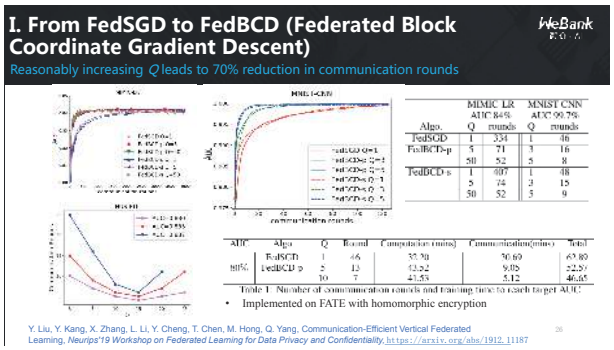


联邦学习有这么多方向，大家在哪个方向投入呢？我们最近和 google 开了一个研讨会，在研讨会上推出了一个白皮书，叫做《Advances and Open Problems in Federated Learning》，我们把联邦学习，不管是 To C，还是 To B，不管是在加密，还是在分布式机器学习，还是在激励机制上面，都做了一个总结，大家可以关注一下。

另外一个方向是联邦学习和迁移学习的有效结合。迁移学习的主要思想是一个领域已经比较成熟，比方有一个源领域，在这个领域已经形成了很好的数据或很好的模型，有一个新的领域，比方说红色的目标领域，在这里可能数据是有限的，因此模型的效果也有限，但是这两个领域确实有相似性，比如两个领域都是关于图像分类的，在这个时候，就有希望把知识从左边迁移到右边。整个知识迁移就像举一反三能力一样。假设两个领域的数据不能交换，参数也要互相保密，那么是不是还可以做迁移学习？第 2 个问题是如果涉及到两个机构，有不同的特征但是有类似的样本，他们有意愿去合作，但是他们的数据格式不一样（比如一个数据是图像，另一个数据是文字），相当于是一种异构的协作，在这种情况下如何做联邦学习？所以这种研究问题就可以囊括在联邦学习和迁移学习的一种有机的结合，我们称之为联邦迁移学习。



下面介绍一下联邦迁移学习的一些新进展。第一个进展是可以用迁移学习的思想来协助联邦学习。由于既涉及迁移学习，又涉及联邦学习，它的速度会大大减慢，所以一个值得研究的问题就是如何把两个机构沟通的效率提高。我们最近写了一篇文章，里面用了一个很简单但是非常有效的办法把这个效率提高。这个办法就是尽量减少两边的沟通次数，在他们沟通一次的时候，让他们这一次的沟通发挥最大的作用，使得尽量能够在本地进行多次的运转，然后才进行机构间的沟通。而且尽量能够在设计这个机器学习算法的时候，让两个机构之间的沟通并行化。我们发现，不仅效率在提高，而且运行成本大大降低、运行速度大大提高，并且效果也能保证。第 2 个办法就是引入一些比较精密的高端的加密算法，比如我们可以把梯度值变成一些矢量运算和向量运算，在做加密的时候会大大提高效率，效率能提高 20 倍甚至到百倍不等。第 3 个办法是可以用一些新的加密手段，比如姚期智先生的密钥分享和一些新算法的有机结合。能够让加密的算法速度大为提升，比如在左边的图最上面的曲线对应的是原始的同态加密，最下面的这条绿色的线代表的是新的算法叫做 ABY(Arithmetic sharing, Boolean sharing, and Yao's garbled circuits)，这样每一次的加密和解密效率就大为提升。



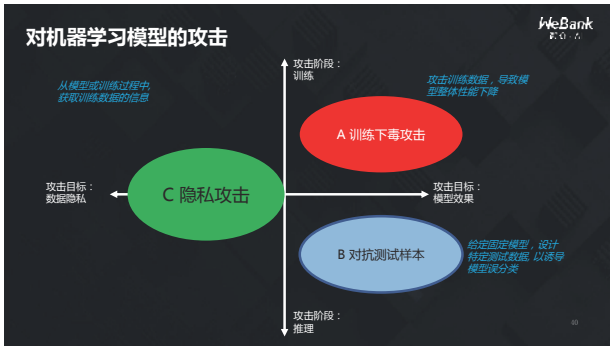
同时，我们也用以上的这些做法，来帮助工业界证实行之有效的非常通用的算法，比如 XGBoost 算法，在纵向联邦学习之下就变成 SecureBoost 算法，SecureBoost 也被做到了开源的联邦学习的平台 FATE 上。横向联邦的研究已经有了大规模的应用，新加坡国立大学何炳胜教授团队的研究成果已经发表在 AAAI 2019 上。同时，这个领域如果要持续发展，离不开公开的数据集。我们很高兴地看到，最近有些公开的数据集也出现了，比如 Federated AI Dataset 这个网站，其中的一部分就是利用计算机视觉方面的联邦学习，这样的一种数据有点像当年的 Internet。Internet 是数据全部在本地，我们这个数据的特点是分布式的，大家在建模的时候要比的就是在加密状态下的效率和效果，所以要比的维度就会更多。同时在联邦学习的国际标准很快就会出台，我们国家也有相应的国家标准和团体标准出现。我前面提到的开源算法也被纳入 Foundation 开源平台上，并且有多个工业级别的联邦学习的商业应用现在已经出现了。我们目前正在写第二本联邦学习的书，将会有大量的实用案例在这本书里面。

这里我特别要提的是一个和新冠病毒相关的我们一直在考虑的问题，健康码在我们国家使用得非常好，而且充分体现了我们国家大规模的互联网化的优势。我们在考虑能不能在使用健康码快速控制疫情的情况下，同时保护用户的隐私。比如红色的用户和蓝色用户，在不知情的情况下有过密切的交互，比如他们坐过同一辆车或者在同一个餐馆吃过饭，当发现红色用户已经是感染者的时候，如何能够让蓝色的用户去检查自己有没有这种密切交互的历史，在检查的过程中，我们不希望把蓝色用户的行程透露给其他任何人。这件事情如果能做到，就既能保护用户隐私又能安全可靠地控制疫情。比如蓝色用户要进入大楼时，保安员就要看他的二维码，这个时候他就会主动提请一些信息来证明自己的健康程度。整个触发的过程可以用联邦学习来保证隐私不被泄露，同时准确地反映了接触的历史。

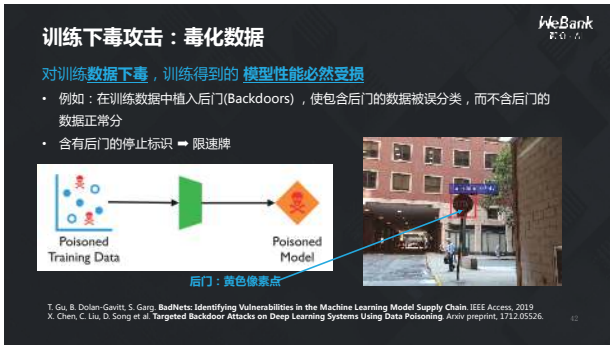
AI 同时要保护模型的安全。模型在什么情况下会变得不安全？我们知道人工智能机器学习的流程可以被分解为以下几步：1）我们要获取许多训练数据；2）通过这些训练数据训练一些算法来形成模型；3）要把模型应用在实际当中，使得存在一些测试数据的时候，我们能得到一些结果。这里面就有一些薄弱的环节，一个可能被攻击的薄弱环节就是训练数据本身，这就是数据下毒；第二是可能会对模型进行攻击，即模型的隐私可能会被泄露；第三是测试数据可能会作假，模型本身无法识别，这也相当于对模型的一种攻击。下面分别从这三个方面，来看一下人工智能界是如何应对这种攻击的，同时又是如何保证人工智能模型安全的。

下图给出了这三种攻击方法。第一个是对训

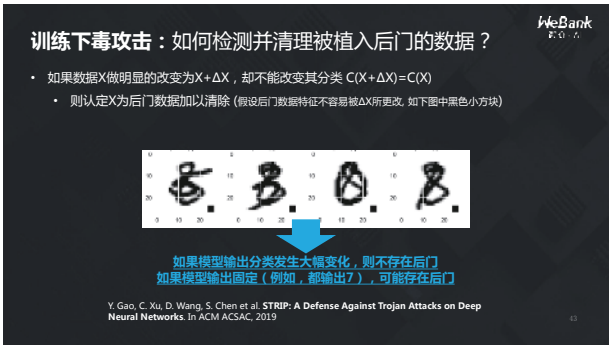
练数据下毒攻击，这相当于对模型的目标进行攻击，同时在训练过程中对目标进行攻击。第 2 种相当于对目标进行攻击来影响模型的效果；同时在推理的过程中进行攻击，相当于对测试样本的攻击。第 3 种是在对目标进行攻击的过程中，要了解数据所包含的用户隐私，相当于隐私攻击。下面分别来看一下这三种情况。



第一种对训练下毒攻击，可以叫做毒化数据。那么它如何实现呢？比如可以在训练数据中植入后门，这些骷髅就相当于一些加入的带毒的数据。但是这个 stop sign 上面有一个黄色的像素点，这个黄色的像素点对于攻击者来说，是植入了训练数据，以至于模型见到了这个黄色点就不顾其余了，因此就没有把 stop sign 识别出来，使车开过去没有停，这里有个行人就会出事，这种是恶意的攻击。



那么如何防止类似事情发生呢？有不同的研究者提出了解决方案。比如一种方案是在数据上面故意加一些扰动，使得原来的数据 x 变成 $x+\Delta x$。当扰动小的时候，观察模型分类的结果是否会发生突变。如果发生了突变就是被下毒了，如果非常稳定，就认为是安全的。比如各种手写体“8”的写法，如果加像素变成“7”了，那么就认为加了 Δx 以后分类效果就变了，因此认定这个样本可能是带毒的样本。这是一个解决的思想，还有其他的做法。总之，这个领域是非常活跃的。



第 2 种攻击方法是针对测试样本，如果对模型机制有所了解，可以设计一种测试样本，使得他蒙混过关，这也相当于是对模型的一种攻击。比如下图左边这种情况，这个人脸本来不能通过人脸检测的闸机口，但是如果在测试数据里加入一些噪音，这样就使得他有可能通过。这样细微的对测试数据的扰动使得从不可以通过到可以通过，这相当于系统被欺骗了。这个现象也有各种各样的解决方案，比如应用对抗样本，这也是对原始数据进行扰动，使得证明模型具有一定的鲁棒性，也就是在小的邻域内，要求模型的输出是一致的。这里还有大量其他的工作，就不一一赘述了。

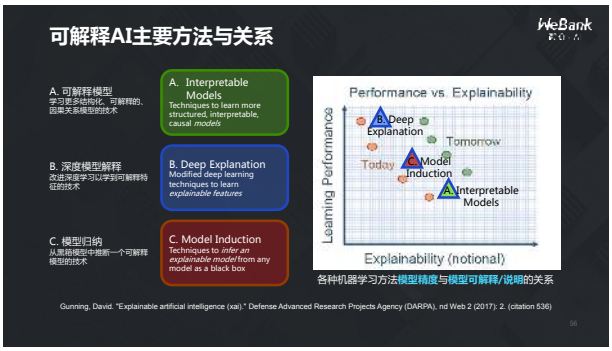
第 3 种是对隐私的攻击，这就相当于我们有不同的机构，比如在联邦学习的分布式架构下，有一个坏人，他希望通过模型参数反推出原始数据。联邦学习的一个标配是同态加密，另外也可以应用各种比较安全的多方安全计算来加强保护，在不牺牲模型性能的情况下来保护隐私。但是世界上没有免费的午餐，往往需要大量的计算才能实现这样的加密算法，所以它的计算时间是非常长的，计算开销非常大。因此在实施当中，有些应用就选用了差分隐私来代替同态加密，我们知道差分隐私是在模型参数沟通的时候加上噪音。差分隐私有很多的研究，但在工业界应用却非常少，原因是隐私保护和模型的性能即准确

率是矛盾的。我们要保护的隐私的程度高，就要加入很多的噪音，这使得模型的准确率大为下降，这在工业的很多应用中都是不能接受的。比如训练数据是一些手写识别的数据，如果有一个恶意的攻击者能看到一系列模型的沟通，而且加密如果是用差分隐私实现的，他还是可以在某些状态下重构原始的训练数据，这在同态加密下是无法做到的。但是用同态加密，时间开销及复杂度会大大升高，因此有些实施者或使用者使用差分隐私。

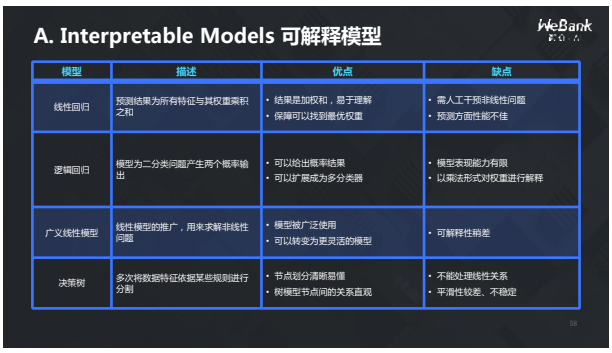
麻省理工学院的韩松教授就证明了这一点，他的团队去年在 NeurIPS 大会上一篇关于“深度泄露攻击”的最佳论文，其大概意思是：当多方在沟通数据梯度的时候，即使梯度是部分加噪的，比如使用差分隐私，对方还是可以不同程度地学到原始数据的。实验结论是噪音加得多，对方学到得少，同时效果也会变差。最近微众银行范力欣老师团队的一项工作在理论上证明了在差分隐私情况下，也可以几乎完全保证隐私不被泄露。这就分为 3 个状态：一个是完全泄露，隐私完全不加密也不加噪；另外一个是完全加噪，这时候坏人没法学到东西，但是好人的模型也会受到影响；我们要在中间找到一个最佳点，坏人无法学到，好人也能学到最好的结果。

什么是可解释？我们把它分成两部分：第一是要让人明白他在做什么，第二是要让不同背景的人用不同方式去明白，我们在金融界就有特别深的体会，比如做一个银行的人工智能算法来保证对风险进行评估。一个新的申请过来了，一个人在进行交互，监控系统要对他的风险进行评估。评估系统对不同的人要做不同的解释，比如我们有监管方，有银保监，那么他要对银保监解释整个结果产出的逻辑；第二个是对系统开发的工程师同事也要解释，使得工程师同事能够进入系统进行修改，不断地进行系统改进；第三是对贷款的申请人要能解释他得到的结果，比如资产正常、负债水平较低，因此你的风险较低，像这样给申请的终端用户去解释。所以可解释性并不是铁板一块，对于不同的人，解释要不一样。

最近在人工智能界可解释性被大量提出。在一个可解释性的专题研讨会上 (https://www.darpa.mil/program/explainable-artificial-intelligence) 就提出了 3 种可解释性定义。第 1 种是模型本身要可解释，要学习更多结构化可解释，具有因果关系的模型，叫做 Interpretable Models；第 2 种是对于“深度解释”，就是可以改进深度学习已学到可解释特征的技术。比如我们知道结果是由黑箱模型的哪一部分来推断的。第 3 种是可以从黑箱模型推断出一个可解释的模型。比如有一个很复杂的黑箱模型，是否可以用一个比较简单的模型来基本上覆盖黑箱模型。这个简单模型是可解释的，这在学习的有效性和可解释性上就有不同的平衡，当然最终我们要的是右上角这种高度可解释又有高度的性能。



首先来看一下可解释模型，下图罗列了一些机器学习模型，有线性回归、逻辑回归等类数据参数模型。这些模型的优点和缺点这里都有总结，总之目前没有一个模型是可以适合我们既能高效率、高效果又能高度可解释的，所以这个方向还有待大量的研究。



模型	描述	优点	缺点
线性回归	预测结果为所有特征与其权重乘积之和	• 结果是加权和，易于理解 • 保障可以找到最优权重	• 需人工干预非线性问题 • 预测方面性能不佳
逻辑回归	模型为二分类问题产生两个概率输出	• 可以给出概率结果 • 可以扩展成为多分类器	• 模型表现能力有限 • 以需法形式对权重进行解释
广义线性模型	线性模型的推广，用来求解非线性问题	• 模型被广泛使用 • 可以转变为更灵活的模型	• 可解释性稍差
决策树	多次将数据特征依据某些规则进行分割	• 节点划分清晰易懂 • 树模型节点间的关系直观	• 不能处理线性关系 • 平滑性较差、不稳定

其次是基于“深度解释”，如果能拿图像做一个推理，最后得出一个结论是这个图像里面有一个攻击。那么如何知道哪些像素是可以解释攻击的，我们人是很容易知道的，那么黑箱模型如何能找出高度适配的像素，这就可以用反向传播来实现，这技术叫层次相关传播 (LRP)。

第 3 种是做模型归纳。比如拿哈士奇做分类，系统错误地把它分成了一只狼。我们知道这不是一只狼，于是想归因，想知道为什么系统错了。我们把这个错误例子作为输入，通过对他周

边像素特征的解释，最后得出结论原来是由于把它放在了雪地上，导致了错误的解释。这种解释对于我们设计模型是非常有用的。现在也有各种各样的办法，其中有一种办法叫 Shapley Additive Explanation。应该说这种技术的研究是日新月异的。

AI 的可解释的标准建设也刚刚开始，IEEE 新成立了一个可解释人工智能的标准组，也希望大家来参加。这是对各个方面的人 (不仅仅是对技术人员、监管者和政策的制定者等) 都会很有帮助的。

人工智能的发展，不仅要芯片，不仅要数据，不仅要算法，同时我们也要注意人。跟人相关的我们今天讲了三个定律：第一是要保护人的隐私；第二是要保护模型的安全；第三是要保证对人类可以解释。

作者简介：

杨强，教授，美国马里兰大学计算机系博士和北京大学天体物理专业学士，主要研究方向为人工智能：迁移学习、联邦学习、机器学习、数据挖掘和自动规划，现担任微众银行首席人工智能官 (CAIO)，为 AAAI、ACM、IEEE、AAAS 等国际学会的 Fellow，曾任香港科技大学新明工程学讲席教授、计算机科学和工程学系主任以及华为诺亚方舟实验室主任。他是国际人工智能界“迁移学习”和“联邦学习”技术的领军人物，于 2017 年当选为国际人工智能联合会 (IJCAI，国际人工智能领域创立最早的顶级国际会议) 理事会主席，是第一位担任 IJCAI 理事会主席的华人科学家。

中文引用格式：杨强. AI 与人的新三定律 [J]. 智能系统学报, 2020, 15(4): 811–817.

英文引用格式：YANG Qiang. AI’s three new laws of robotics[J]. CAAI transactions on intelligent systems, 2020, 15(4): 811–817.