GDPR 对 AI 的挑战和基于联邦迁移学习的对策

杨 强　CAAI 副理事长，AAAI/ACM/IEEE Fellow，IJCAI 理事长

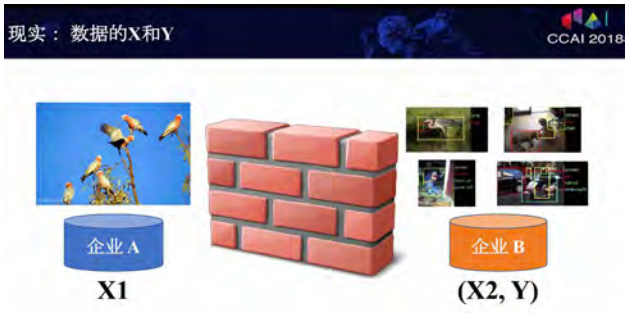
未来已来，但是，我们面对 AI 的未来，总还是有些隐忧，因为未来也是未知，也有隐忧。

人工智能曾经有过三个高峰，现在处在第三个高峰，这中间有两次低谷。第一个高峰的出现是因为看到了 AI 的希望，也就是自动化算法对提高效率的希望，但是到后来却发现算法的能力不够，因此就产生失望，进而导致了第一个低谷。然后算法跟上来了，但是这时却发现算力和数据不够，专家系统的设计跟不上工业的成长需求，这就引发了 AI 的第二个低谷。之后又看到，现在算法和算力都有提升，而且有了大数据的出现，AI 的难题终于可以解决了。现在的一个说法是我们处于大数据时代，所以这一波的人工智能一定会成功。这个感觉来自一个很强的推动力，就是 AlphaGo 的成功。最初的 2016 年版的 AlphaGo 使用了 30 万个棋局训练，是大数据的成功。上周我们在国际人工智能大会 IJCAI 上，在瑞典为 AlphaGo 团队颁了一个国际人工智能奖（MINSKY AWARD）。之所以如此受到 AlphaGo 的鼓舞，是因为我们联想到，既然 AlphaGo 在围棋上都有如此大的突破，那么人工智能是不是在各行各业都会突飞猛进?

AlphaGo 的这种大数据真的出现在各行各业了？了解到的情况却让我们非常失望，远远不是！更多的应用领域有的只是小数据，或者质量很差的数据。上面这个“人工智能到处可用”的错误的认知会导致很严重的商业后果。最近听到一个 IBM 沃森应用失败的消息。大家知道，IBM 沃森是一个非常有名的问答（QA）系统，给一个问题 Q，它能很精准找到答案 A。比如我们给了上面一个问题，沃森就用一个高维的表示来表达这个问题 Q。大家可以把这种表示想象成物理学里的光谱，就是一束光打过来，用棱镜分解成不同频率的光，就看到了光谱。有了这个光谱以后，可以和答案库里对应答案，它的概率也应该相应的高，这就是可能的答案。这个流程应该说非常简单，但问题就是要有一个很健全的答案库。IBM 在电视大赛上取得了成功之后，就把这个应用在一些听起来比较好的垂直领域——医疗领域。但是，最近在一个美国的癌症治疗中心，发现这个应用非常不理想，从而导致了这个项目的失败。我们看一看在医疗领域，这些领域里的问题和答案来自哪里？比如收入有病症、基因序列、病理报告、各种各样的检测、各种论文，沃森的任务是利用这些数据来做诊断，帮助医生。但是，经过一段时间

的实践发现，这个大数据的来源远远不够，导致系统的效果差。在医疗领域，我们需要很多有标注的数据。但是，医生的时间非常昂贵，不可能像一些其他计算机视觉应用一样，大众、普通人都可以来做标注。在医疗这样的专业领域，只有专家才能做决策，但是专家的时间非常宝贵，就导致这种标注的数据非常有限。有人估计，把医疗数据放在第三方公司标注，需要动用 1 万人用长达 10 年的时间才能收集到有效的数据。这就说明在这些领域，即使动用很多人来做标注，数据也不够。这就是我们面临的现实。

那么可不可以把很多散落在各地、各个机构的数据合并成大数据？现实是，我们训练预测模型时，需要有一部分的特征，即原始特征叫做 X。比如，在手机应用里，有用户信息的维度，也有产品特征的维度，这些可以看作是 X 维度。但要用这些维度做用户行为预测模型，同时还要有行为标注列 Y。Y 就是要知道的答案。比如在金融领域，Y 就是用户的信用；在营销领域，Y 就是用户的购买愿望；在教育领域，Y 就是学生学到知识的程度等。$X+Y$ 才有了真正的训练数据，就像对不同图像里的物体进行标注一样。



但是，在现实当中，却往往遇到这样的情况，有些企业只有 X，只有一些没有标注的数据，即使不断地在收数据，但也只是部分的数字化；有些企业可能有 Y，有标注，通过一些手段或者应用本身就是带有标注的，但是，它们对应的数据样本也不多。那么这些企业能不能把它们的数据很容易地合并，变成有用的训练数据？我们发现，这样做越来越难，因为企业中间有道墙，形成数据源的隔离。数据源隔离这种现象很多。举个例子，我现在在“微众银行”学习 AI 和金融的结合，这里有大量的应用，比如智慧零售。在零售领域的数据来自很多产品的数据、用户购买商品的数据等。但是，零售业却缺乏其他一些数据，比如他们并不知道用户的购买能力，或者支付习惯等。那么这些发展智慧零售的机构能不能把自己的数据和银行的数据直接合并？答案是不行的。

这里有几个原因。首先公司间的数据合作要考虑利益的交换；然后不同部门和机构的行政批准流程也许很不一样；同时，现代社会对于用户隐私的要求也越来越高，公众的诉求和监管的要求也是不允许数据简单“粗暴”地进行交换。因此很多数据的共享性很差。这些原因就导致了在很多需要机器学习模型的领域，数据标注不足、标签大量缺失等问题。

所以，虽然理想中的 AI 是有大数据的支持，但是现实中遇见的却是一个个数据孤岛。我的看法是，如果这个问题解决不好就有可能导致再一次的 AI 低谷。而这个问题的重要性，还远远没有引起人工智能

从业者的关注。

总之，隐私、安全和满足监管的要求为 AI 带来了一个前所未有的挑战，这个挑战导致大部分企业只拥有小数据。我可以先给一个结论：AI 界现在并没有很好地应对这些挑战，并没有用大量的时间和精力去设计保护隐私安全和满足法律法规的机器学习框架来应对这些挑战。可以看一下当下的媒体，他们的宣传机器大部分时间都在传播这样一类新闻，就是某某机构、某某大牛又创造出一种新算法，又可以把某个指标，比如准确率做到更好。指标的提高固然很重要，但是，这不是 AI 当下最重要的需求，因为这并没有解决社会和企业的痛点。我认为当下更应该关心的是，在隐私、安全和监管要求下，如何让 AI 系统，更加高效、准确地共同使用各自的数据，能够在小数据（很少的样本和特征）和弱监督（有很少的标注）的条件下做更好的模型。

现在，监管对于数据的交换管得越来越严。首先看一下欧盟最近引入的一个新的法案《通用数据保护条例》（General Data Protection Regulation，GDPR）。也许在座的一些同事已经了解了，但是我相信大多数的同事是第一次听到这个法案。这个法案和以往的行业规范不同，是一个真正可以执行的法律，违背它的后果是非常严重的，因为罚款可以高达被罚机构的全球营收的 4%，非常高。GDPR 在今年 5 月 25 日生效，里面有很多条款都是用来保护用户隐私和数据安全。比如，过去下载一个 APP 时，会看到要表示同意的文件，而这里的一些法律解释，往往会用晦涩的法律语言来描述，并且用很小的字体展示。现在根据 GDPR 是不允许的，因为 GDPR 要求这样的文件一定要用清晰可理解的语言来解释。同样，经营者要允许用户来表达数据“被遗忘”的愿望，即“我不希望你记住我过去的数据，并希望从现在起，你不要利用我的数据来建模”。这些条款最近已经被用在 Facebook 和 Google 上，使他们成为基于这个法案的第一批被告，而且罚款是巨额的。

我们看一下 GDPR 对人工智能有哪些影响。首先，有一条款说：对使用自动化模型决策全面禁止。我们看这一条觉得非常不可理解。也就是说，如果你有一个全面自动化的机器学习模型，用来决策做用户相关的商业活动，在决策过程中没有任何人的参与，如果机器去使用这个决策，这也是违法的。做机器学习的听到这项要求就吓出一身冷汗。另外，用户也可以对模型的决策提出质疑，而且有权去要求模型对其的决策进行解释。也就是说，现在可解释模型已经变成了法案，以至于华盛顿大学机器学习领域著名的教授 Domingos 发了一个推特：5 月 25 日以后，深度学习就非法了。因为深度学习到目前为止是黑箱，是不可解释的。还有用户有权知道数据使用的目的，而且可以反悔，可以撤回数据。

这对人工智能有多么大的影响！

研究界和企业现在满足这样或类似法规的程度如何？几乎是零。我们经常用到的做法，是在使用用户数据时都让用户划个钩，表示“同意”。但往往收集数据的一方并不是建立模型的一方，在企业中，大家习惯在一个地方收集数据，把数据转移到另外一个地方去处理和清洗，然后可能再把数据拿到另一个地方去建立模型，再把模型卖给第三方去应用。现在这个过程要非常小心，因为数据只要出了收集方就可能犯法。第三方使用模型的目的，也许产生原始数据的用户完全不知道，这就很有可能触犯 GDPR 的法律。在计算机、大数据、数据挖掘里有一个著名理论，叫做差分隐私理论（Differential Privacy），就是希望通过在数据里加噪音，直到第三方不能区分任何个体为止。也就是说，有很高的概率，数据不能还原到一个个体，以此来保护用户隐私。这种在过去被认为是保护隐私的技术可能在 GDPR 下就不使用了。例如，如果我是 A 方，收集了一些数据，在里面加一些噪音，根据差分隐私理论，可以把数据的使用权卖给 B，只要 B 在一定概率下不能区分任何个体用户，这在过去被认为是满足法案的，但是现在不行了。因为只要有用户的隐私被泄露的可能性，数据的交易就有可能是被判违法的。所以，数据的这种在企业间的交换，无论加噪音与否，本身就违反了 GDPR 。

GDPR 是欧盟建立的，和我们有什么关系？我看到，最近对隐私和安全的考虑是一个世界的趋势，欧盟引入了这个法律，不能说明天美国和世界其他地方就不引入这个法律。同样，中国对数据的监管也是非常严格的，对用户数据的隐私保护也已经有相关的法案，而且越来越细化。这个趋势是世界性的。

我们的数据本来就已经是孤岛的形式了，解决孤岛一个直接方案就是把数据从 A 迁移到 C，再从 B 迁移到 C，然后再在 C 加以聚合。但是，现在这样做很可能就是违法的，即法律不允许我们粗暴地来做数据聚合。那么如何做才能合法解决数据孤岛问题，应该足够引起人工智能学者和从业者的深思，因为很可能这个困境就是导致下一个人工智能冬天的导火索。所以倡议把研究的重点转移到如何解决数据孤岛的问题。这里我们提出一个可能的解决方案，叫做联邦迁移学习。什么是联邦学习？什么又是迁移学习？

我们所希望看到的是，假设有三个不同的企业 A、B 和 C，每个企业都有不同数据。比如，第一个企业 A 有一些用户特征数据；第二个企业 B 有其他的一些用户特征数据，同时也包括一些标注数据；第三个企业 C 是一个银行，可能有有关金融的特征和标注数据。这三个企业按照 GDPR 准则是不能粗暴地把三方数据加以合并，因为他们的用户并没有同意这样做。假设在三方各自建立一个模型，而这个行为已经获得各自用户的认可。我们希望做到的是各个企业的自有数据不出本地，就像画地为牢一样，把自己围一个圈，围起来。然后，系统可以通过加密机制下的参数交换方式，在不违反法规情况下，建立一个虚拟的共

有模型。这个虚拟模型就好像大家把数据聚合在一起一样，但是数据本身不移动，也不泄露隐私，模型在各自的区域还是为本地的目标服务。在这样一个机制下，各个参与者的身份和地位相同，这就是为什么这个体系叫做“联邦学习”。

建立这个机制，不是只把参数从 A 转到 C、从 C 转到 B 那么简单，实际上对最后模型的效果是有要求的——既要安全，又要有效。安全是指数据在本地不能移出，而模型的参数被第三方处理时不仅要加密，而且要保证不能被反推原始用户的任何特征；有效是指所谓的 Lossless，就是效果要符合无损失原则，在 A、B 和 C 的模型效果要和把数据真正聚合在一起一样。这两个要求对 AI 的从业者是一个挑战。

这个要求能不能做到?

首先看一下业界的一些进展。谷歌最近提出了一个针对安卓手机模型更新的数据加密需求，建立的一种联邦学习方案。比如，使用安卓手机时会不断汇聚数据到安卓云上进行处理。联邦学习就是针对这样的过程，首先在每个终端上进行模型建设，参与者的特征相同，但他们做的模型可能很弱，虽然功能都一样。然后在云端把单个的模型加以聚合形成大的模型，大的模型再分发到各自终端里。参与者特征相同，样本不同，这样不断地聚合使得模型加以更新；同时通过加密算法，使得云端并没有解密终端传来的模型，同样别的终端也没有办法解密邻居的数据。

另外一种联邦学习是假设有原始数据和一个建立好的模型，那么在应用这个模型到原始数据时会不会泄露隐私？有个算法叫做 CryptoDL, 是应用同态加密算法于多项式形态的激活函数。这样的好处是可以把原始数据加密，然后用这个模型做决策，得到的结果也是一个加密的结果。把加密的结果传到终端，终端可以解密实施。在整个过程中，通过这个加密机制，模型并不知道自己在做什么决策。所以说，这是在应用 Inference 时使用的。

刚才讲的例子都是把数据横向分段，横向的每段都是不同的用户样本，其特征一样，在这样风格下来学习得到的一块块数据。还有一种分割的方法就是按照特征来分段，可以看作是纵向分段，对应于两个不同机构，机构 A 和机构 B 它们的特征不一样。那么，我们希望在一个虚拟的第三方能够把这些特征，在加密的状态下加以聚合，以增强各自模型的能力。这种联邦学习，因为加密算法的原因，只能对某些类的模型使用，比如逻辑回归模型。对

很多其他模型，还不知道行不行。最近经过研究发现，联邦学习对于树型结构模型也是可以用的。例如，在这边有一个企业、有一个数据集，那边也有一个企业和一个数据集，通过这种加密技术可以使两边的树都得到成长。有了树模型以后就很自然可以发展到森林模型。“微众银行”的 AI 团队就设计了一个这样的新框架，提出了一个叫做 SecureBoost 的算法框架，并使用在多方协同建模的问题上。其效果是建立了中心的虚拟模型以后，可以分发到两边的参与者，和把数据聚合在一处建模相比并没有损失，而且过程都不泄露用户隐私。

上面所述的“联邦学习”的优点是，在不具体交换原数据的情况下，以及对用户 ID 的差值不泄露的情况下，A 和 B 两边可以参与联邦学习的网络。在这个网络里就可以建立一个共同模型，这个模型的参数可以分别独立持有。也就是说，两边的模型都可以得到成长，但是它们却不直接互相沟通。这样用户的样本和用户的特征都不泄露，已经满足 GDPR 大部分的要求。不同企业和机构可以形成一个“朋友圈”，在其中用这种联邦学习一起建模。联邦的意思就是各个数据的拥有体大家是平等的。

以上的讨论是假设不同数据的样本有一部分是共享的。但是，有时不同企业的数据样本并不一样，在这种情况下遇到的小数据、弱监督的问题，即数据标注很少的问题也可以解决吗？一个方法就是我们一直研究的迁移学习。

我的学生戴文渊所领导的“第四范式公司”，这个 AI 公司在企业服务领域，利用 AI 的技术为企业客户提高营销效果。下面这个例子是他们所做的一次实践。假设需要营销车贷。车贷属于大额贷款，而这种大额贷款的样本很少，找新渠道成功办理的客户，在一定的时间内还不到 100。在这样小的数据集上很难建模。与此相反的情况是，有很多小额贷款对应着大量用户。那么有没有办法用小量的数据建立非常好的模型，然后迁移到大额贷款的用户上去发现大额贷款的用户？第四范式使用了迁移学习，利用在千万级微信公众号中的小额贷款的样本建立模型，再利用迁移学习适配于大额贷款的领域，营销效果非常好。

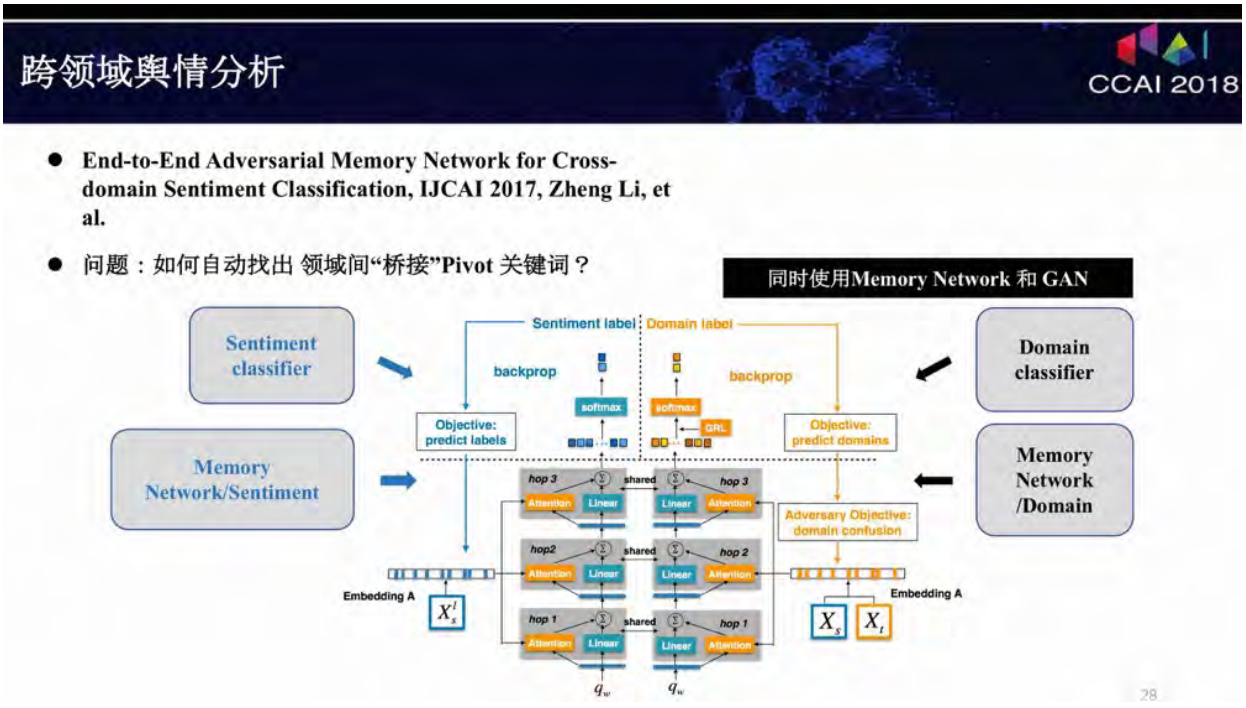
什么叫迁移学习？生活中我们学骑自行车，再学骑摩托车就很容易，为什么？因为人有这个能力，人是可以举一反三的，通过很少的例子就可以把一个具体的体验通用化。人可以做到这一点是因为我们可以找到两个领域的共性。比如在深圳开车，司机在车的左边；在香港开车，司机在车的右边。如何能够 1 秒钟就从深圳开车转到香港开车？一个车过了关以后，怎样马上适应右边开车方式？这里的共性就是司机相对于路的位置，如果司机是坐在路靠中间的位置，不管是在香港还是在深圳都适用，只要保证司机靠近路中间就可以，这是一个很实用的迁移学习例子。也就是说，开车可能很繁杂，有很多特征，但是找到了一个共性，学会了在深圳开车，同样也能很快学会在香港开车。

具体到工业上应用，例如，我们很关心用户的舆情，当卖了一个产品后非常关心用户的反馈，在网上、在社交网络有很多

的留言，我们希望一键式对留言进行总结。如有关书店的，对这本书可能是 Great，非常好看；有些说 boring，非常无趣。75% 是赞的，25% 是踩的。这些反馈就对书店决策者非常有用，对电商上产品的排名也非常有用。这个决策在同一个领域是机器学习模型进行的，我们对这一段文字进行自然语言处理，然后建立分类模型，对新来的舆情进行分类。

假设有一个数据很多的有关舆情的训练数据已经建立好在一方企业 A，这样就可以在 A 端建立一个模型。它看到一段新的用户反馈，就可以在 A 端判断是“赞”还是“踩”。当到一个新的领域或企业 B，假设没有任何标注数据就无能为力了，因为没有标注，没有办法做这个模型。但是，如果这两个领域有一定关联，比方“图书”和“餐厅”这两个领域也许会有一些关联，我们就会将这边模型中间共同部分迁移到右边来，迁移到餐厅评价。

这种迁移学习怎么做？可以设计两个深度学习网络，这个网络看上去也非常复杂，但是实际上逻辑很简单优美，是我的博士生李正设计的。分左边和右边。左边是一个专家，在本领域的专家。比如在图书领域的专家，输入从下面来，就可以判断输出从上面出来。下面输入可能是一段用户评论，上面就是正向和负向的判断。但是没有标注的那个怎么办？可以找一些关键字，这些关键字是两个因素决定的。首先要找到共享的关键字，就是通过这些字是没有办法区分领域 A 和 B 的，并且这个关键字又能很快告诉你舆情的趋向；同时不能够区分领域，这些字就是很有用的通用字，我们把它叫做“桥接”或者 PIVOT。把这两个要求放在一起，根据这些关键字就很容易地把模型从左边 A 迁移到右边 B。



经过效果的演示最后发现，果然是在不同领域，迁移效果最好的就是刚才提出的模型，和手工模型相比也好很多。图中，黑体字是用户表达的评论，蓝色字是我们找出来的桥接词，就是二个领域共有的词。用这些词我们可以建立一个非常好的迁移学习模型，在一个新领域数据不多或者标注数据不多的情况下也可以建立。

回到一开始讲的联邦学习的应用，可以把我刚才讲的应用分为四种分类的子应用，第一种情况是数据分别在两个不同的企业，它们特征相近、样本也相同，这是一个简单情况，在本地建模就好，不需要沟通。第二种情况，如果特征一样、样本不一样，要让两个领域之间能够协同，可以引入 Google 这样的联邦学习方式，不断更新一个总模型，再分发到各个终端去；如果特征不一样、样本一样就可以引入纵向的联邦学习和同态加密技术，在一些逻辑回归或树形模型上加密、合并、更新；如果特征、样本都不一样的两个企业，它们中间的交集很少，这时就要为它进行迁移学习的建模，并在建模中保证不能反推用户个体信息。

举一个银行的例子。我们做一个试验，比如在智慧零售这个领域有一些产品的数据，有一些用户购买能力的数据，有一些用户购买取向的数据，或者有产品特点的数据，但是这些数据在三个不同的地方、三个不同的企业。在过去，这种零售部门没有办法把数据加以聚合，现在用联邦学习的方法就可以对三者共同建模，一开始的智慧零售那个需求就得到了满足，大家可以以用户模型分别进行商业活动，而不违背用户隐私的原则。

上面介绍了一个新的保护数据的技术方案，叫做“联邦迁移学习”，来解决数据聚合建模问题。我们保证在不泄露隐私的情况下，共同建模，共同受益。

我们知道，一个新的技术手段往往只占整个商业流程的 5%~10%，更需要引入很多运营、产品和营销操作。下面简要介绍如何做出一个基于联邦迁移学习的新的数据商业模式，建立一个共同成长的大数据 AI 生态。

在建了模型以后，还需要一个商业联盟来进行联邦学习。这样的联盟应该有 N 个实体，它们加入了联盟以后，就像一个朋友圈一样能够利用各自的数据联合建立模型。现在要设计这样一个联盟，它需做两件事，第一件事是在一个垂直领域使用一个联邦迁移学习的技术，比如金融领域的联邦迁移学习；还有一个很重要的题目，就是可以用区块链技术建立一个让参与各方都满意的一个共识机制来估计大家的贡献，以此奖励对联盟有作用的机构。也就是说，如果 A 家说我为 B 家贡献了多少，B 家说我为 C 家贡献了多少，大家可以建立一个机制，以达到某种共识，这个共识

（下转第 12 页）