联邦学习:人工智能的最后一公里

Federated learning: the last on kilometer of artificial intelligence

杨强[1,2]

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

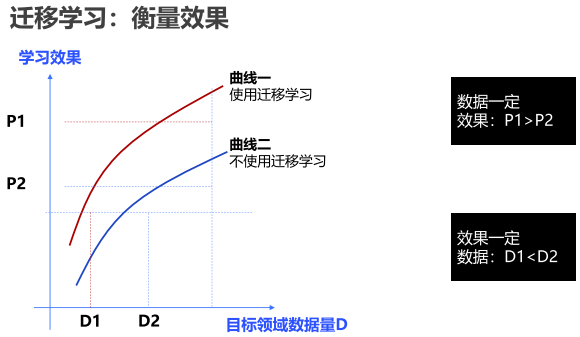
我们看一下深度学习的一些限制,大家现在都在大数据领域有很大的突破,一个代表性突破就是 AlphaGo, AlphaGo 在 19×19 的棋盘上可以说是举世无双。但是只要换一下棋盘的大小,或者换一下棋盘的种类,原来的模型就完全无能为力了,就得重新做一个训练,这个例子引起了我们的深思。

当前人工智能领域需要大数据的推动,这个推动如果换一个新的领域很可能只有小数据。小数据的场景是不是也可以用深度学习来解决呢?我们认为是很有困难的,因为依据深度学习现在的进展,还没有很多的算法能够在小数据情况下发挥作用。

有很多原因造成我们面对小数据这种状态,比方说由于行业性质使得不同部门之间没有办法交换数据,加之考虑到用户隐私、商业利益、监管的要求等,我们面临的是小数据和一个个数据孤岛。把小数据变成大数据,又需要做很多数据标注,比方在医疗或者金融方面,时间不允许我们很快把小数据变成大数据。

针对小数据的问题,我和团队长期做的研究是迁移学习。迁移学习就是像人类一样,能够进行举一反三,把模型从一个场景迁移到另外一个场景的动作。目标是突破传统机器学习必须有大数据作为前提的要求。人类是怎么做的呢?我们在解决一个新问题的时候,会利用联想能力想一下:过去遇到过类似的场景吗?那么能不能将一个模型通过小小的改动,让其在当前的场景下使用?这种联想能力是我们举一反三的迁移能力的关键。

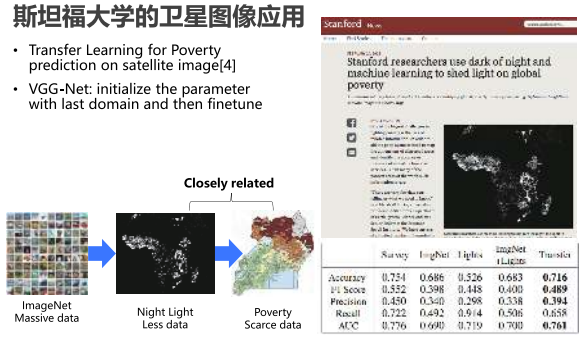
什么叫具有迁移能力?以这两条曲线为例,在一定的数据量下,学习效果如红线所示,有迁移学习,学习效果会更好一些,同时它的增长会更快一些,这就是迁移学习所带来的两个好处。因此我们说,一个领域,它的迁移效果好不好,是由多个指标来衡量的。像这两条曲线所示,在数据一定和效果一定的情况下,迁移学习的指标都好。



我们针对这样的情况提出了一个迁移学习的理论算法框架,这个算法框架的目标是模拟人类小样本的快速学习能力。我们可以通过样本与样本的分布来做迁移,同时还可以基于特征来做迁移。即便是两个领域,比如说一个是计算机视觉,一个是自然语言处理,它们只要语义上有沟通的话,还是可以做迁移,这是基于特征的迁移。我们还可以基于模型来做迁移,比如说我们可以做一个预训练模型,在一个新的领域下,可以在某一个预训练的层次上做小的改动,就把这个模型做迁移。甚至我们在任务上也可以做迁移,这样就可以实现零样本或者单一样本的学习能力。

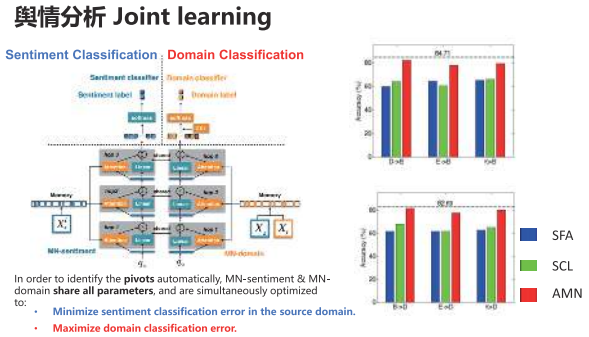
在此基础上,我们还可以把迁移学习在时间维度上加以扩展,形成一个传递式的迁移学习,就好像踩着石头过河,一步步地走。我们也可以先把一个领域的模型迁移到第 2 个领域,再从第 2 个领域迁移到第 3 个领域,依此类推。

我们还可以依赖于一个深度学习和一系列的中间领域来做传递式的迁移。传递式迁移有一个很好的例子,这个是斯坦福大学和世界银行一起合作的,利用卫星图像来获得经济状况,尤其是贫困地区的经济状况。通过将 ImageNet 的数据迁移到夜空图像,再通过夜空图像迁移到白天图像,经过这两步的迁移,我们就能自动得到一个对于卫星图像的二维图的经济状况的估算,这个估算也是非常准的。

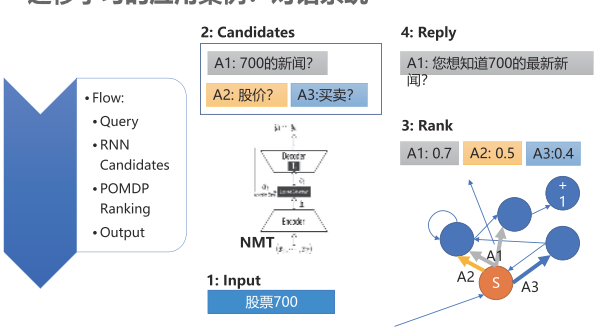


迁移学习也可以用在舆情分析上。比如说我们已经获得了一个很好的自然语言分类器，只要给一个书评或者是影评，我们就可以对它的正负取向进行一个估算。那么在已经有了这样的一个模型以后，我们假设给一个新的领域，这个新的领域有一些数据是我们没有见过的，但是通过两个领域之间的相似性、共性，还是可以很快地得到一个迁移学习模型，使得在第 2 个领域能很快地建立起一个舆情模型。

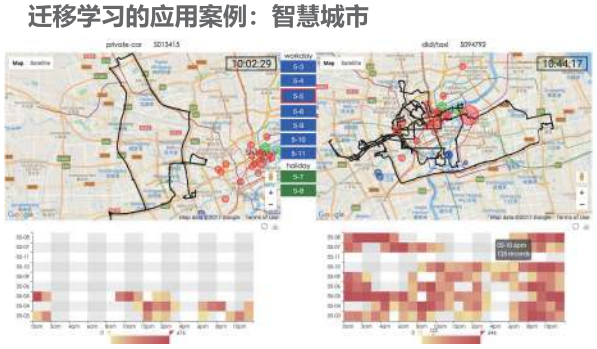
那么具体怎么去实现？最近的一个做法是通过多层的注意力网络机制。这个注意力机制的网络有两个任务：第 1 个任务是在本领域能够获得越来越高的准确度；第 2 个是在跨领域的任务当中，希望能够最少地区分两个领域，使得能够混淆两个领域里面所取的特征词，那些特征词往往既能告诉我们舆情特征，又能告诉我们与领域无关的这种共性，这种特征词取得的效果也非常好。



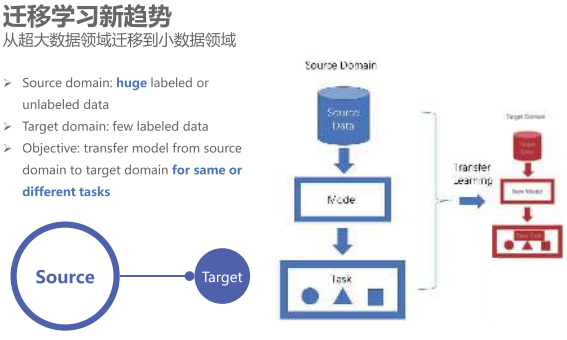
同时迁移学习也可以用在一个非常有商业价值的领域，就是对话系统。我们知道对话系统需要进行大量的对话的标注。如果换一个领域，比如说从一个卖咖啡的领域到一个卖股票的领域，这里面虽然有很多具体的、商业的领域知识，但是它的逻辑结构还是有共通性。在这种情况下可以通过强化学习的迁移机制来区分一个领域里面的策略：它是本领域的特殊策略，还是一个通用的对话策略，把这两种策略区分开，学习通用策略，就能使我们很快地得到一个基于 RNN 的迁移学习模型。



还有一个智慧城市的例子，这也是我们和一些公司合作的成果。比如，在一个城市有很多车辆出行，根据车辆的状态，可以区分这是网约车还是私家车，这样的一个分类器是可以根据两个城市的相似度把它迁移到一个新的领域的。还有城市的 PM2.5 预测，也可以做这种迁移，所以这是非常通用的一个做法。



总的来说，迁移学习有千千万万种，但是能不能只做一个通用的算法，就可以适用于很多领域？这是一个终极目标。现在这个目标终于有希望了，众多的迹象表明，如果在源领域有足够多的数据，可以形成一个非常大的预训练模型，那么遇到一个新的领域的时候，往往就可以很快地进行成功的迁移。所以迁移学习的成功与否，在于能不能把同一类的问题都挖掘出来，建立一个巨大的预训练模型，向小数据领域、新领域做迁移。

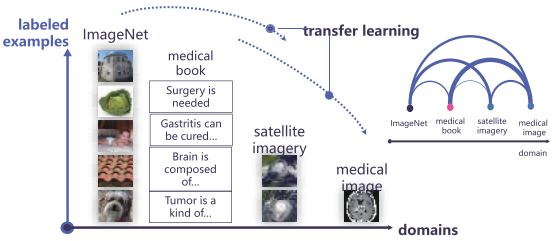


这个例子在最近 Facebook 的一项工作上得到了印证，他们就是逐字地增加在源领域的数据，

在 ImageNet 数据的基础上增加 100 倍。这样就能发现，在数据很少的情况下，迁移还是可以得到非常好的效果。与此同时，NLP 的领域也得到了类似的印证，还有像 BERT 这样的领域，这使得机器学习的工作者非常兴奋。

我们知道，Automated learning 是用 AI 来设计 AI，用人工智能来自动地设计部分的人工智能步骤。那么迁移学习同样可以达到这个效果，可以用一个人工智能的模型来设计一个迁移学习模型，这是我们最近的一项工作。这个工作称为“learning to transfer”，也是一个非常有希望能够自动化迁移学习过程的一个例子。所以这个就是我们在迁移学习方面的努力，从大数据到很多小数据领域。

自动化学习如何迁移：ICML 2018 工作



Transfer Learning via Learning to Transfer, Ying Wei, Qiang Yang et al. ICML 2018

我最近的一项工作和迁移学习类似，但是还有不一样的地方，它更多的是解决“数据孤岛”的问题。我们知道，如果我们有很多的部门，有很多的传感器，但是每个传感器都只能收集一部分的数据。如果没有办法把这些传感数据打通，那么每一个地方的数据都不足以来训练一个好的模型。为什么会发生这样的事？因为现在社会对隐私、安全的要求越来越严格，在欧洲有这种法律，在国内也有很多非常严格的个人隐私保护法律，而且趋于严格化和全面化。

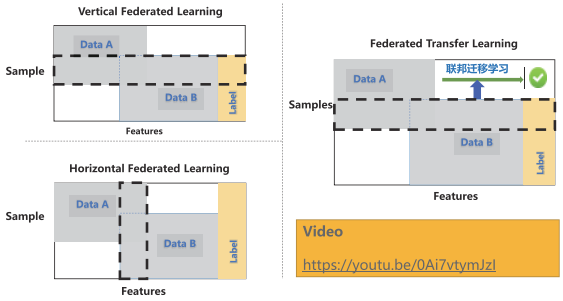
举个例子，如果养一只羊，这只羊就是一个模型，那么如何建立起这个模型？过去的方法是把各家的草买过来，喂这只羊。这就相当于到各处去买数据、采购数据、收集数据。但是这个方法现在不行了，数据要求不能出本地，也就是说，草不能出草场，那么农民的羊是不是就会饿死？不会的，农民会采取另外一种方法，他会让这个羊在各处的草场走动，草不出草场，那么这只羊可以走。就好像数据不动，模型在数据库之间走。

这由此引发了一个新的领域，叫做联邦学习。我们有许多机构，每一个机构都有自己的数据，它们联合起来是一个完整的、很大的数据库，可以用来训练一个大数据模型。但是现在因为隐私、利益的关系，每一个机构都不想或者不能把数据和别人共享。那么，可以让它们结成一个联盟，让它们共同遵循一个规则，使得这个模型的参数可以在它们之间沟通，在沟通的时候，这个参数也是加密的，使得一个机构没有办法通过它得到的参数包来反拆对方的数据和模型。最后每一个地方的部分模型都得到了成长，加起来就是一个总的模型。联邦学习是这样的一个概念。

联邦学习的技术优势

联邦学习具有很多优点，比如：数据隐私得到保护，参数也能得到保护，建模的效果也非常好。我们把联邦学习理论化，建立了一个数学的基础。这个数学基础基于一个假设，假设这个数据拥有的各方，有足够多的特征方面的重叠，那么称之为“纵向联邦学习”。如果它们特征是重叠的，但是它们的样本不重叠，这个称为“横向联邦学习”。如果两边都不重叠，还可以用迁移学习把它们都给迁移到另外一个状态空间，在那个空间找它们的共性，同时用联邦学习来解决。

General Federated Learning Catalog



所以联邦学习+迁移学习，就形成了一个既能保护隐私，又能共同联合建模的机制，这个机制在工业界获得强烈反响。我们发现这个机制在跨领域不同的企业/机构之间尤其有用。比如：银行和监管 (机构) 可以联合起来，更好地建立一个反洗钱的模型；互联网 (公司) 和银行可以建立一个小微企业的风控模型；互联网公司和零售公司可以建立更好的新零售模型等等。这样的建立取决于一个数学基础，一个非常有效、高效可计算的加密机制，其中数学的演化、硬件的演化和分布式安全算法的演化要一起进行。

另外一个非常特别的领域是城市管理。比

如，工地的安全监管，每个工地有很多摄像头去监管工人的安全状况，比如说戴安全帽、火灾的状况。那么如果不同的工地之间的数据能够打通的话，这样的一个模型将是非常有效的，我们恰恰把联邦学习用在这里。

联邦学习加上迁移学习也用在很多的领域。这里要特别说明的是，建立这一系列联盟离不开一个标准。因为这是多个企业之间在沟通，企业之间一定要遵循同样的标准。所以我们一直在建立一个 IEEE 的联邦学习国际标准。同时我们通过和经济学教授合作，建立起一个公平的利益分配机制，来帮助大家自愿地在这个联盟里使用联邦学习。另外，我们还发布了世界上第一套联邦学习开源软件。

总结一下就是，迁移学习是模仿人类举一反三的能力，它有一整套的理论和工业应用的算法，已经在各个方面得到了很好的印证。联邦学习是我们正在进行的一项研究工作，它在合作、建模的基础上引入隐私保护的概念，我们也期待它在工业界将会有很大的应用发展。

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博士，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 和 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和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. 联邦学习：人工智能的最后一公里 [J]. 智能系统学报, 2020, 15(1): 183–186.

英文引用格式：YANG Qiang. Federated learning: the last on kilometer of artificial intelligence[J]. CAAI transactions on intelligent systems, 2020, 15(1): 183–186.

新书介绍：联邦学习

在当前大数据驱动的社会环境下，数据隐私安全成为了全民探讨的重要议题。信息技术的发展，离不开由我们在智能终端（手机及其他设备等）上产生或推断出的个人数据，如浏览习惯、点击频次等，来推动个性化应用和服务的发展。尤其在 AI 领域，这一情况更为明显，依赖于持续的数据感知、收集，并上传至服务端进行深度分析与训练，AI 才能迎来蓬勃发展。但底层未经审查、不透明的数据收集和聚合协议，很可能造成严重的数据安全威胁和隐私风险。

要解决这样的困境，仅仅靠传统的机器学习方法已经出现瓶颈。我们需要一个既满足隐私保护和数据安全，又可实施的解决方案——联邦学习。联邦学习希望做到各个企业的自有数据不出本地，而联邦系统可以通过加密机制下的参数交换方式，即在不违反数据隐私法规情况下，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建立的最优模型一样。但是在建立虚拟模型的时候，数据本身不移动，也不泄露隐私和影响数据合规，也就是“数据不动，模型动”。这样，建好的模型在各自的区域仅为本地的目标服务。在这样一个联邦机制下，各个参与者的身份和地位相同，而联邦系统帮助大家建立了“共同富裕”的策略，也就是“风险不增，效益增”。这就是为什么这个体系叫做“联邦学习”。

首部全面、系统论述联邦学习的中文著作《联邦学习》可以作为广大学习者入门和探究联邦学习的第一本书。本书由杨强教授及其团队撰写，详细描述了联邦学习如何将分布式机器学习、密码学、基于金融规则的激励机制和博弈论结合起来，以解决分散数据的使用问题。介绍不同种类的面向隐私保护的机器学习解决方案以及技术背景，并描述一些典型的实际问题解决案例。