# NIPS 2017：政策字段注释

由和



像往常一样，NIPS是疯狂的。但是，我们活了下来。

接下来，我们决定记下一些受今年会议启发的想法，重点介绍最近的一些技术发展，以及它们对围绕机器学习的公共政策讨论可能产生的影响。

我们希望这样的总结可以帮助我们在不断涌现的研究浪潮中穿行，并将研究工作与围绕技术的社会影响进行的更广泛的讨论联系起来。

—

\*\*\*\*- Lian，Zhang等。

尽管在机器学习的流行报道中经常被忽略，但进行培训的体系结构对该技术的社会影响有着重要的影响。主流模型要求数据集中，然后进行训练，这意味着机器学习的有效、最佳使用与保护隐私的要求有些紧张（尽管正在进行研究，以减轻这种权衡的程度）。

另一种架构采用更分布式的方法，降低了需要通过集中点的数据。然而，这些方法因效率较低而不受欢迎。针对并行随机梯度下降，作者发现了一组条件，在这些条件下分散设置实际上优于集中式结构。即使所有的服务器都在一个数据中心内运行一个分散的算法，这也是一个很好的发现。

关于隐私和机器学习性能之间的紧张关系的一个大问题是这种情况的经济性：投资和部署这些系统的公司是否有保持两者的动机？要确保这一点，需要集中和分散算法的性能至少达到同等水平，并且在最佳情况下，后者要超过前者。这篇论文标志着这方面的贡献。

—

\*\*\*\*- Lundberg，李

围绕模型可解释性的研究问题的兴趣继续增长。今年的可解释的ML研讨会有很多人参加（结果是a）。

一个主要的争论仍在继续，这就是研究可解释性的最终目标，以及它实际上有什么意义，如果有的话。正如扎克·利普顿所指出的那样，关于“”的说法仍然是一个悬而未决的、调查不足的问题，关于这一主题的说法是，可解释性的用处实际上相当有限。

当然，这里的一个核心挑战是，某些东西是否可以解释在很大程度上取决于所讨论的旁观者。因此，认真研究它将使这个领域的研究方法和实验看起来更像是人机交互，而不是纯粹的计算机科学。这个领域如何进化来适应或拒绝这个跨越学科的开放将是一个值得关注的有趣的事情。

这篇论文在某种意义上代表了一种跨越这场辩论的方法。作者提出了一种通用的解释性方法，该方法将许多早期的方法（石灰法、深升力法、形状值估计法等）联系起来，并根据用户研究评估机器解释和人类对模型的理解来评估“解释性”。不过，目前还不清楚，这是否真的是一种应该占据主导地位的解释性成功标准，特别是考虑到研究显示了可解释性在多大程度上可以被操纵（参见，例如；）。技术专家之间是否达成共识是一个悬而未决的问题，但如果达成共识，它可能会产生影响，因为在某些情况下，这将要求ML系统做出解释。

—

和

不可否认，根据研究会议的日程安排对论文做一个总结有点傻。在这一点上，大多数论文出现在ARXIV和其他地方之前的事件本身，这是一个常见的投诉在NIPS和ICML，所有的东西都已经见过。

从这个角度来看，很酷的是，今年有两个主题演讲，一个是阿里·拉希米的，另一个是凯特·克劳福德的，这两个主题演讲成为了NIPS内部和周围对话的核心。在我们看来，这两个问题都是对研究界采取行动的重要呼吁：拉希米（Rahimi）反对ML向“炼金术”方向的下滑，克劳福德（Crawford）则推动了该领域走向公平的方法，并突出了工作本身的政治性。

这两次会谈都要求并敦促该领域反思并决定它想要成为什么样的国家。对于拉希米来说，这一领域需要更深入地考虑为什么，对于克劳福德来说，这一领域需要考虑复杂的社会背景分类。

Rahimi和Crawford的议程将把机器学习研究社区置于一个更广泛的游戏中，接受更多的责任，在塑造技术融入社会的过程中，它已经在发挥作用。我们将不得不看看，随着围绕该领域的炒作继续不断，更广泛的社区是否会承担起这一责任，或者这是否会默认地委托给该领域之外的其他人。

—

和

NIPS经常讨论的话题之一是让人工智能系统学会如何快速学习。它的灵感来自于智力的一个最显著的特征，即人们能够迅速达到对一项新任务的熟练程度。这项研究特别活跃的一个领域是少数/一次性学习、元学习和机器人技术的交叉点。这是因为许多研究人员认为，我们需要大大简化我们的人工智能系统获得新技能的过程，以便他们能够接受培训，在现实世界中完成有用的事情。这种类型的研究也基于这样的假设，即我们无法模拟所有的事物，而且尽管有随机化等技术，如果足够新颖的话，我们可能很难适应在模拟器中模拟不好的一些看不见的现实方面。

加州大学伯克利分校（UC Berkeley）的一篇研究论文《通过元学习进行一次视觉模拟学习》（One-Shot Visual simulation Learning via Meta Learning）为我们提供了一个早期的视角，让我们了解机器人在未来可能会更快速地学习的世界：这种方法从纯像素输入（尽管也可以接受其他数据点，如状态空间和动作空间等）开始工作，并且要求机器人在测试一个全新的任务（比如将一个新的对象推向一个新的方向）之前，在任务的分布上接受训练，比如推一组不同的对象。该技术具有很好的推广性，并在一个真实的机器人上进行了验证。研究结果显示了这种技术的前景，但也表明这项技术离产品化还有很长的路要走：这种方法在一次性物体放置上的成功率约为70%（如果你让它访问状态和动作，则为90%）。研究人员还进行了平行实验，有关开发机器人的相关研究，这些机器人能够更好地预测其世界的特征，以便执行行动，以及。

随着它的成熟，这类研究为在真实的物理机器人上部署强大的、数据高效的学习算法开辟了道路。这就对当恶意行为体能够实际访问这些机器人时会发生什么情况提出了许多政策挑战 - 一个坏行为体是否能够重新训练强大的工业机器以执行一些不同的行为？机器人能被简单地改造成危险的用途吗？如今，系统受到机器人所接受的模拟任务的广度的限制，并且能够运行元学习策略。在未来，任务的分布将更为广泛，因此机器人部署者面临的一个真正问题将是训练元学习策略的任务子集，因为它们将在访问、根植和重新定位的情况下对机器人的灵活性进行权衡。

—

“海龟和坦克”@

NIPS 2017的一个反复出现的主题是，人工智能社区正在努力应对其日益智能化、可应用的贡献对现实世界的影响。在机器学习和计算机安全研讨会上，许多研究人员努力解决大规模人工智能部署带来的一些安全问题。

Anish Athalye和他的合作者介绍了他们在健壮的对抗性例子上的工作：创建愚弄基于计算机的分类系统的图像，这些系统对转换是健壮的。Athalye和他的同事们最引人注目的演示是一只三维海龟，它被一种特殊的视觉模式覆盖，导致它被错误地归类为“来福枪”，而不是一只三维海龟。

这引发了一系列潜在的现实问题：这是否意味着，例如，有可能重新粉刷一辆大型车辆，如坦克，以便当通过无人机和/或基于卫星的传感器从头顶观看时，它被错误地归类为一辆普通的汽车，或者可能是一棵树？“是的，我认为应该可以用同样的方法来定位卫星图像，”阿萨利通过电子邮件告诉我们。“你可以用一堆卫星图像在上面模拟你的对手画，找到一种同时具有对抗性的绘画模式，不管是什么样的平移/缩放水平。“由于各国都部署了多颗卫星，以各种分布方式观察世界，我们还想知道是否有可能制造出一个三维对抗性物体，就像我们的示例坦克，并使其在所有方面都具有对抗性。”分辨率，保证在任意缩放时错误标记。“有可能在决议分发过程中制造出敌对的东西。他说：“你不需要在{1，5，10，20，100}这样的离散分布下进行训练，你需要在连续分布上进行训练（例如在[1100]以上的均匀分布）。

卫星的例子指出了一个更大的问题。未来的决策者将需要准备一个充满人工智能系统生成的合成数据的环境，其中一些数据可能包含各种陷阱，这些陷阱旨在劫持分类或分析算法，并使它们做出不同的预测。这可能会产生一些有趣的效果：首先，它可能会增加人们对拥有良好的软件工具来保证“数据”和/或“推断”供应链的兴趣，从而更容易识别攻击者可能插入自己的位置。

—

作者：Schwaller，Gaudin等人（最佳论文奖@the）

自从深度学习开始接管人工智能领域以来，我们已经看到了一些通用工具，如卷积神经网络和递归网络，以及一些方法，如序列到序列域的翻译技术，已经传播到了其他领域。这意味着，诸如语音识别和计算机视觉等主要研究领域的不断进步，有可能使其他领域取得特定领域的进展。观察这一现象最有趣的地方之一是在化学领域，化学领域正看到药物设计和分析的基本探索方法被深度学习技术所改变。

IBM研究人员的一篇新论文展示了如何将一些最初由Google开发的人工智能技术（随后在Google Translate和Google Assistant中使用）应用到预测有机化学反应的任务中，具体来说，就是查看SMILES数据集中的化学反应，并试图预测反应物是如何产生的试剂相互作用产生特定的目标分子。正如他们在论文中所写的那样，这里的科学目标是“解决前向反应预测问题，在这个问题上，原料是已知的，兴趣是生成产品”。为了做到这一点，他们设计了一个网络，可以接收以SMILEs格式编写的化学配方，使用序列到序列技术执行从原始字符串到标记字符串的多阶段转换，并将源输入字符串映射到目标字符串，在本例中是反应的产物。

结果令人鼓舞，该方法的方法导致80.3%的top-1准确率，相比之下，先前的技术水平为74%。这篇论文确实有一些局限性，例如1.3%的前1名预测在语法上是错误的，因此在SMILES格式中是无效的。可能还需要一种方法来创建更大的数据集，并且存在一些过度拟合的证据（在Jin的USPTO数据集上，由于达到了99.9%的准确率，网络几乎记住了整个训练集，所以训练趋于平稳）。即使在Lowe的噪声较大的数据集上，也观察到了94.5%的训练精度。）

未来科学家们希望“有了这种模型，化学家可以将有机合成技术编撰成册，也许有一天会完全自动化。”虽然这可能是论文酝酿的一个途径，但它几乎完全来自另一个看起来非常不同的领域（来自谷歌的用户级人工智能互联网服务），但它使用了相同的基本机制在与科学探索有关的另一个问题上取得的进展。这凸显了传统的“两用”技术概念在分析人工智能时是如何挣扎的：这些技术是如此的通用，以至于它们变成了“万能”的范畴；对它们的监管可能和对螺丝刀、锤子和塑料的监管一样困难。它还表明，智能资助人工智能邻近学科的基础科学研究，有可能在经济的其他领域开辟新的实验和发展道路。

—

- 刘、布鲁尔、考茨

研究人员使用GANs和VAEs的组合来产生一组从一个图像域到另一个图像域的高质量翻译。特别值得注意的是他们的演示（图片可在报纸上获得），其中采取了一系列街道场景，并将它们从冬季转换为夏季，从雨季转换为晴朗的环境，从白天转换为夜晚，等等。

从政策的角度来看，这延续了我们在太空中看到的一对更广泛的趋势：图像生成质量继续提高 - 随着时间的推移，这些技术似乎会使区分真实媒体和生成媒体变得更加困难。或许更重要的是，无论是在所需的特定数据集、计算能力，还是所需的特定专业知识中，进入壁垒都在继续下降。这意味着，对于更广泛的参与者来说，无论是好的还是坏的，利用这些技术将越来越具有成本效益。

—

And that’s it! As always, we’re keeping an eye out for new policy-relevant papers as we get into 2018. Give us a shout on Twitter ([Jack Clark](https://medium.com/@jackclarksf) / [Tim Hwang](https://medium.com/@timhwang)) or comment here if you see anything good.