DeepMind 新研究：三招解决机器学习模型 debug 难题

人工智能3 月 30 日

传统的软件测试和 debug 方法很难适用于现代的机器学习系统，DeepMind 希望解决这一问题，开发面向预测模型的可靠验证工具。这篇博客描述了能严格识别并排除学习预测模型中的错误的三种方法：对抗性测试，稳健性学习和形式验证。

自计算机编程诞生以来，软件开发中就一直没有离开过 bug。随着时间的推移，软件开发人员已经建立了一套在软件实际发布前进行测试和 debug 的最佳方式，但这些方式并不适合现代的深度学习系统。

今天，机器学习的主流方法是在训练数据集上训练系统，然后在另一组数据集上进行测试。即使在最坏的情况下，确保系统稳健性或高性能也是至关重要的。本文描述了能够严格识别并排除学习预测模型中的错误的三种方法：对抗性测试，稳健性学习和形式验证。

机器学习系统一般是不稳健的。即使在特定领域中表现优于人类的系统，具体情况稍微改变，往往就可能导致无法解决简单问题。比如图像扰动的问题：如果在输入图像中添加少量精心计算的噪声，那么本来在图像分类任务中表现超过人类的的神经网络，很容易将一只树獭错认成一辆跑车。

在图片上加上一个对抗性输入，可能导致分类器将一只树懒错误识别为一辆跑车。两个图像在每个像素上的差异最多只有 0.0078。结果第一张被归类为三趾树懒，置信度> 99％。第二个被归类为一辆跑车，概率> 99％。

其实这不是什么新问题。计算机程序总是有 bug。几十年来，软件工程师开发了种类繁多的技术工具包，从单元测试到形式验证。这些方法在传统软件上运行良好，但是由于这些模型的规模和缺乏结构性（可能包含数亿个参数），想用这些方法来严格测

试神经网络等机器学习模型是非常困难的。开发能够确保机器学习系统在部署时稳健性的新方法势在必行。

从程序员的角度来看，与系统的规范（即预期功能）不一致的任何行为都属于 bug。DeepMind 不仅评估了机器学习系统的技术是否与训练集和测试集一致，还评估了这些技术的作用与系统的期望属性的规范描述中是否一致。这些属性可能包括对输入中足够小的扰动的稳健性，避免灾难性故障的安全约束，或产生符合物理定律的预测能力等。

本文讨论机器学习社区面临的三个重要技术挑战，因为我们共同致力于严格开发和部署与所需规格可靠一致的机器学习系统：

高效测试实际功能与属性规范的一致性。我们探索有效的方法，来测试机器学习系统是否与设计者和系统用户所期望的属性相一致。揭示二者差异的一种方法是在评估期间系统地搜索最坏情况下的结果。

训练机器学习模型，使其产生属性一致的预测。即使有了大量的训练数据，标准的机器学习算法也可以产生与理想属性不一致的预测模型。这要求我们重新考虑训练算法，这些算法不仅能够很好地拟合训练数据，而且要与属性列表上的要求保持一致。

正式证明机器学习模型是规范性一致的。虽然形式验证领域几十年来一直在研究这种算法，也取得了令人瞩目的进展，但很难轻易扩展到现代深度学习领域。

测试规范一致性

面对对抗性实例下的稳健性问题，是深度学习中研究相对充分的问题。这项工作的一个主要主题是评估模型在强对抗性攻击下的稳健性，以及设计可有效分析的透明模型。我们发现许多模型在弱对抗下进行评估时看上去很稳健。但遇见针对更强的对抗时，精度几乎下降为零。

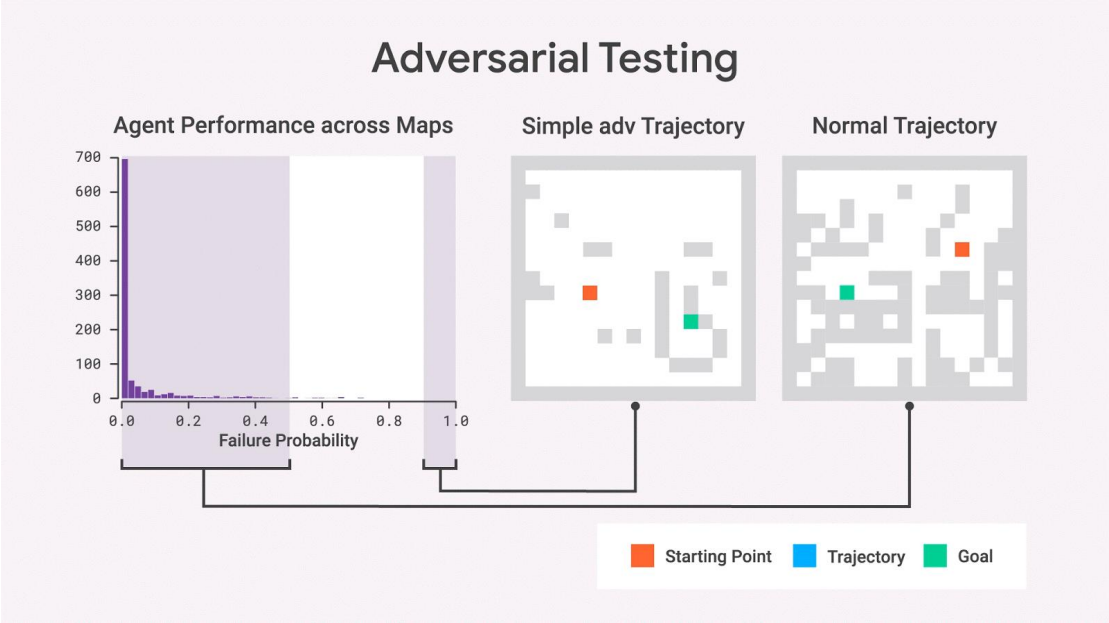
目前大多数研究都集中在监督学习（主要是图像分类）的背景下，但是需要将这些想法扩展到其他条件。在最近关于对抗灾难性故障方法的研究中，我们将这些想法用于测试确保关键设置的强化学习智能体上。开发这类自主系统的挑战之一是，由于一个错误就可能产生严重后果，因此即使非常小的失败概率也是不可接受的。

我们的目标是设计一个“对手”，能让我们提前检测这些故障。与图像分类器一样，针对弱攻击进行评估，很容易产生错误的安全感。我们为强化学习对抗性测试开发了两种互补的方法。首先，使用无衍生优化来对智能体的预期回报进行最小化。接着，学习一种对抗值函数，该函数根据经验预测哪种情况最有可能导致智能体的失败。然后使用此学习函数进行优化，将评估重点放在最有问题的输入上。这些方法只构成了丰富且不断增长的潜在算法空间的一小部分，我们对能够对智能体的未来发展进行严格评估感到兴奋。

这两种方法已经比随机测试产生了很大的改进，可以在几分钟内检测到过去需要花费数天才能发现（甚至完全无法发现）的问题。我们还发现，对抗性测试可能会发现我们的智能体出现了与随机测试集的评估结果性质不同的行为。

在对抗性环境下，我们发现执行 3D 导航任务的智能体仍然无法在十分简单的迷宫中完全找到目标，即使它们在非对抗性环境下的平均表现已经和人类相当。此外，我们需要设计能够抵御自然故障的系统。

在随机抽样中，我们几乎从来没观察到具有高失败概率的地图，但是在对抗性测试下，这样的地图确实存在。即使在去掉了许多墙壁之后，智能体在这些地图下的失败概率仍然很高。



训练规范一致性的模型

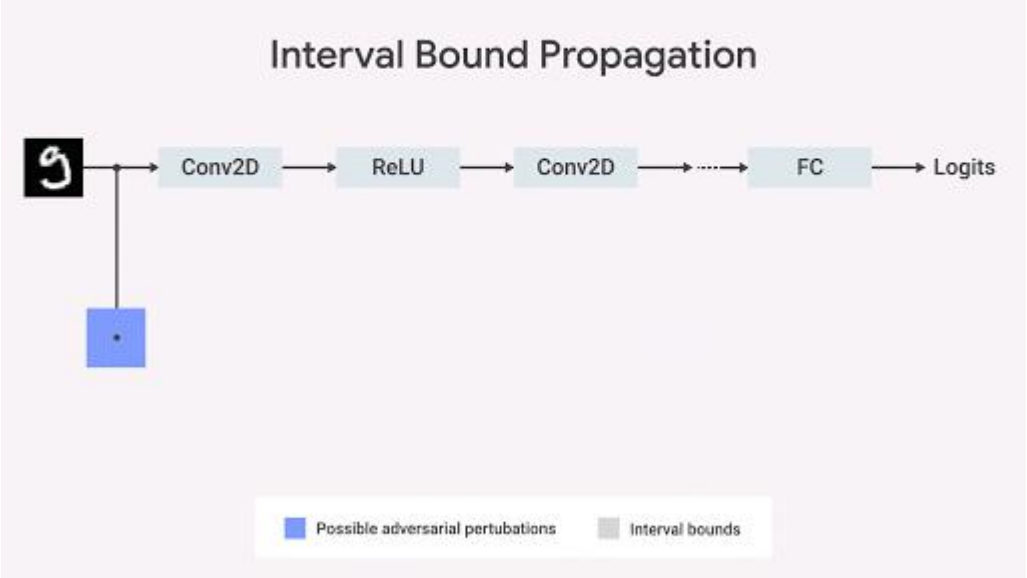
对抗性测试是为了找到违背规范的反例。因此往往会高估模型与这些规范的一致性。在数学上，规范是必须在神经网络的输入和输出之间保持的某种关系。这种关系可以通过某些关键输入和输出参数的上限和下限的形式来体现。

受此启发，DeepMind 的团队和其他团队研究了与对抗性测试程序无关的算法（用于评估规范一致性。这可以从几何学上理解 - 我们可以约束给定的一组输入的情况下，

限制输出空间来最严重地违反规范。如果此界限范围相对于网络参数是可微分的并且可以快速计算，则可以在训练期间使用，通过网络的每个层传播原始边界框。

结果表明，区间界限传播是快速有效的，并且可以获得强有力的结果。尤其是能够降低 MNIST 和 CIFAR-10 数据集上的图像分类中的现有技术的错误率。

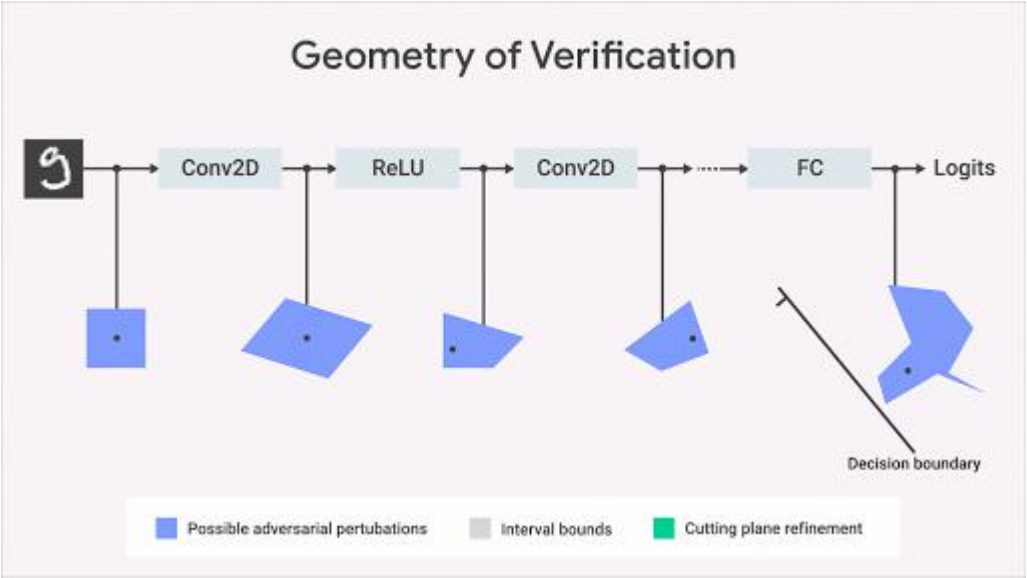
未来的下一个前沿领域将是学习正确的几何抽象，计算更严格的输出空间过度概率。我们还希望训练网络与更复杂的规范一致，捕获理想的行为，比如上文提到的不变性和与物理定律的一致性。



形式验证

严格的测试和训练有助于构建强大的机器学习系统。但是，没有多少测试可以完全保证系统的行为符合我们的要求。在大模型中，由于输入扰动的选择极为庞大，因此列举给定输入集的所有可能输出（例如对图像的无穷小的扰动）是难以处理。但是，与训练一样，我们可以通过在输出集上设置几何边界来找到更有效的方法。正式验证是 DeepMind 正在进行的研究的主题。

机器学习社区已经有了几个关于如何计算网络输出空间上的精确几何边界的有趣思路。我们的方法基于优化和二元性，将验证问题转化为优化问题。。下图以图形方式说明了该方法。



这种方法使我们能够将验证算法的适用性扩展到更一般的网络（激活函数，体系结构），更一般性的规范和更复杂的深度学习模型（生成模型，神经过程等）

未来方向

我们需要做更多的工作来构建自动化工具，以确保现实世界中的 AI 系统做出“正确的事情”，为实现这个目标，未来需要在这些方向上发力：

学习对抗性评估和验证：随着 AI 系统的扩展和复杂度的提升，设计适合 AI 模型的对抗性评估和验证算法将变得越来越困难。如果我们可以利用 AI 的强大功能来推进评估和验证，那么这个过程可以大大加快，并实现扩展。

开发用于对抗性评估和验证的公开工具：为 AI 工程师和从业者提供易于使用的工具是非常重要的，可以在 AI 系统造成广泛的负面影响之前阐明其可能的故障模式。这需要一定程度的对抗性评估和验证算法的标准化。

扩大对抗性实例的应用范围：到目前为止，大多数关于对抗性实例的研究都集中在对小扰动（通常是图像）的模型不变性上。这为开发对抗性评估，稳健性学习和验证方法提供了极好的测试平台。我们已经开始探索与现实世界直接相关的属性的替代规范，并对未来在这方面的研究感到兴奋。

学习规范：在 AI 系统中获得“正确”行为的规范通常难以精确表述。当我们构建能够展示复杂行为并在非结构化环境中行动的越来越智能的代理时，将需要构建可以使用部分人类规范并从评估反馈中学习进一步规范的系统。

原文链接：

https://deepmind.com/blog/robust-and-verified-ai/