Adversarial Machine Learning

宇翔

2020 年 7 月 27 日

1 Adversarial Machine Learning (AML, 对抗机器学习）

Adversarial Machine Learning (AML)目前的研究主要从两个方面入手。一方面，研究并设计更具攻击性的对抗样本，来作为神经网络鲁棒性的评估标准；另一方面，研究针对对抗样本攻击的防御方法，来提升神经网络模型的鲁棒性。即研究工作简单可以分为两个部分：攻击和防御。

攻击，即指如何生成对抗样本以使得机器学习模型产生错误的预测；防御，即指如何使机器学习模型对对抗样本更鲁棒。此外，近几年也出现了一些AML理论方向的工作。

攻击部分的工作从深度上来讲主要在于借助各种优化算法逐步提升生成对抗样本的攻击强度，从广度上来讲主要在于从一开始的针对CNN网络和图像数据扩展到更广的范围，如表格数据、时序数据，Auto-Encoder、强化学习等。

防御部分的主要工作是通过提出各种各样的正则项，来获得对某种或某些特定攻击鲁棒的模型。

2 攻击：如何生成对抗样本

攻击方式可以根据如下四种情况来分类：

1 黑盒/白盒攻击，主要区别在于是否需要知道模型信息；

2 targeted/non-targeted，主要区别在于生成对抗样本时，是否需要指定生成样本所属的类别;

3 Image specific / Universal, 主要区别在于是针对某张图片生成对抗样本，还是全局生成；

4 Perturbation norm：即生成对抗样本时所采用的范数。

针对模型的攻击问题，主要分为两大类，就是从训练阶段和推理（inference）阶段来进行讨论。

训练阶段的攻击（Training in Adversarial Settings）

训练阶段的恶意攻击，主要的目的就是针对模型的参数进行微小的扰动，从让而模型的性能和预期产生偏差。这样的行为主要是通过数据投毒来完成的。

推理阶段的攻击（Inference in Adversarial Settings）

当训练完成一个模型之后，这个模型就可以看做一个BOX，那么这个盒子中，对于我们如果是透明的话，我们就把它当成是“白盒”模型，如果这个盒子中，我们什么都看不了，我们就把它当成“黑盒”模型。（我们在这个部分不讨论灰盒模型）那么针对白盒和黑盒的进攻手段自然是不同的，但是最终的目的都是希望能对模型的最终结果产生破坏，与预期脱离。其影响力以及攻击的构造粒度也是有所不同的。

2.1 白盒攻击（White-Box Adversarial）

当然这种所谓的“白盒攻击”，需要提供一个很“假”的前提——就是我们需要知道里面所有的模型参数，这个在现实生活中是非常不现实的。除非是，当模型被打包压缩到智能手机上之后，然后恶意者通过逆向工程来进行原有模型的复原，才有可能。当然这种情况出现的情况非常低了，因此我们需要有这种前提假设。

2.2 黑盒攻击（Black-Box Adversarial）

当模型处于黑盒的时候，更加符合现实的场景，但是这比白盒的模型缺少了更多的模型信息。因此，大家就从几个角度考虑如何进行模型攻击：

- 通过输入和输出猜测模型的内部结构；
- 加入稍大的扰动来对模型进行攻击；
- 构建影子模型来进行关系人攻击；
- 抽取模型训练的敏感数据；
- 模型逆向参数等。

3 防御：如何让机器学习模型更鲁棒

防御的方法基本可以分为三类：修改训练过程或者数据输入(Modified training/input)、修改模型结构(Modified networks)和添加新的模型/组件(Networks add-on)

4 另一个博客的分类

对抗训练：对抗训练旨在从随机初始化的权重中训练一个鲁棒的模型，其训练集由真实数据集和加入了对抗扰动的数据集组成，因此叫做对抗训练。

梯度掩码：由于当前的许多对抗样本生成方法都是基于梯度去生成的，所以如果将模型的原始梯度隐藏起来，就可以达到抵御对抗样本攻击的效果。

随机化：向原始模型引入随机层或者随机变量。使模型具有一定随机性，全面提高模型的鲁棒性，使其对噪声的容忍度变高。

去噪：在输入模型进行判定之前，先对当前对抗样本进行去噪，剔除其中造成扰动的信息，使其不能对模型造成攻击。

5 CCS 2017-AIsec-Adversarial Machine Learning 方向三篇论文摘要

5.1 Making Targeted Black-box Evasion Attacks Effective and Efficient(Mika Juuti, Buse Gul Atli, N. Asokan)

文章研究了攻击者如何在黑盒设置中最佳地利用其查询预算对深度神经网络进行针对性的规避攻击。文章将该问题设置形式化，并系统地评估攻击者通过使用替代模型可以获得的好处。结果表明，在探索与利用之间存在权衡，因为查询效率是以有效性为代价的。文章提出了两种使用替代模型的新攻击策略，并表明它们与以前的“仅查询”技术一样有效，但所需的查询却少得多，能够与部分API的最新有限差分攻击达到类似的有效性，同时最多减少3个数量级的查询。我们还表明，能够通过不同攻击技术进行切换的敏捷对手可以实现最优的效率。我们展示了针对Google Cloud Vision的攻击，表明针对现实世界的预测API的针对性黑盒攻击的难度实际要更加容易（需要≈500个查询，而不是以前的20,000个查询）。

5.2 Interpolated Adversarial Training: Achieving Robust Neural Networks Without Sacrificing Too Much Accuracy

无论在理论上还是在实践中，对抗性的鲁棒性已成为深度学习的中心目标。但是，成功提高对抗性鲁棒性的方法（例如对抗性训练）极大地损害了在不受干扰的数据上的泛化性能。如果可以提高不受干扰的数据的准确性，许多人可能会选择放弃鲁棒性。

本文提出了插值对抗训练能够一定程度上解决这个问题，它在对抗训练的框架内采用了所提出的基于插值的训练方法。在CIFAR-10上，对抗训练将标准测试误差（没有对手时）从4.43%增加到12.32%，而采用插值对抗训练，我们保留了对抗性的鲁棒性，而标准测试误差仅为6.45%。使用我们的技术，健壮模型的标准误差的相对增加从178.1%减少到45.5%。

5.3 Analyzing the Robustness of Open-World Machine Learning

在实际应用中部署机器学习模型时，需要一个开放世界的学习框架来处理正常的分布内输入和不期望的分布外（OOD）输入。开放世界学习框架包括OOD检测器，其目的是丢弃与机器学习分类器的训练数据分布不同的输入示例。但是，我们对当前OOD检测器的理解仅限于设置良性OOD数据，那么在存在对手的情况下它们是否健壮？本文通过介绍和设计OOD对抗示例，对存在对手的开放世界学习框架的健壮性进行了首次分析。实验结果表明，通过轻微扰动良性OOD输入，可以轻松避开当前的OOD检测器，这表明了当前开放世界学习框架的严重局限。此外，本研究发现OOD对抗示例即使对分布内对抗攻击有效，也对基于对抗训练的防御方法构成了强烈威胁。为了应对这些威胁并确保以可信赖的方式检测OOD输入，本文概述了健壮的开放世界机器学习框架的初步设计

本文提出了一种用于定量分析机器学习方法安全性的框架。该框架的关键问题是部署学习模型和攻击者约束的形式规范，最佳攻击行为的计算以及推导机器学习算法对抗性影响的上限。理解机器学习算法安全性的关键在于根据攻击者的目标和应用程序的特定限制对其进行定量分析。本文提出的学习方法安全性分析框架定义了此类分析中要解决的关键问题，文章示范了如何将该框架应用于一种特定学习场景（在线质心异常检测？），分析其安全性。本文提出该框架希望能有助于开发强大的对抗性学习方法。

6 相关知识

OOD detection 指的是模型能够检测出OOD 样本，而 OOD 样本是相对于 In Distribution(ID) 样本来说的。传统的机器学习方法通常的假设是模型训练和测试的数据是独立同分布的(IID, Independent Identical Distribution)，这里训练和测试的数据都可以说是In Distribution(ID)。在实际应用当中，模型部署上线后得到的数据往往不能被完全控制的，也就是说模型接收的数据有可能是 OOD 样本，也可以叫异常样本(outlier, abnormal)。由于深度神经网络分类器可能会将以高置信度将分布外（ODD）的输入分类到分布内的类别中，因此区分异常数据或有显著差异数据是十分重要的。