

## 131 | ICML 2018论文精读：模型经得起对抗样本的攻击？这或许只是个错觉

2018-08-03 洪亮劼

AI技术内参

[进入课程 >](#)



讲述：初明明

时长 07:35 大小 3.48M



2018 年 7 月 10 日~15 日，国际机器学习大会 ICML 2018 (The 35th International Conference on Machine Learning)，在瑞典的斯德哥尔摩举行。

ICML 从 1980 年开始举办，已有 30 多年的历史，是机器学习、人工智能领域的顶级会议。

今年 ICML 大会共收到了 2473 份投稿，投稿量和去年相比增加了 45%。今年最后录取了 621 篇论文，录取率近 25%。除了主会议以外，ICML 大会还组织了 9 个讲座，67 个研讨班。

在接下来的几期内容里，我会为你精选三篇 ICML 2018 的论文，我们一起来讨论。

今天，我和你分享的是大会的最佳论文，题目是《梯度混淆带来的安全错觉：绕过对对抗样本的防御》（Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples）。

先简单介绍下这篇论文的作者群。

第一作者阿尼什·阿提耶（Anish Athalye）是麻省理工大学的博士生，主要研究方向是机器学习算法的安全。他在今年的 ICML 大会上就发表了 3 篇论文。

第二作者尼古拉·泽多维奇（Nickolai Zeldovich）是阿提耶的导师。他是麻省理工大学计算机系的教授，做安全相关的研究。

第三作者大卫·瓦格纳（David Wagner）来自加州大学伯克利分校，是计算机系教授，也是安全研究方面的专家。

## 论文的背景

这篇论文的内容对于大多数人来说可能是比较陌生的。想要弄清楚这篇论文的主要贡献，我们首先来熟悉一下这篇论文所要解决的问题。

试想我们比较熟悉的监督学习任务。一般来说，在监督学习任务中，我们会有一个数据集，用各种特性（Feature）来表征这个数据集里的数据点。拿最普通的监督学习来说，比如需要把图像分类为“猫”、“狗”等，机器学习算法就是学习一个分类器，可以根据不同的输入信息来做分类的决策。

当然，我们所说的是在正常情况下使用分类器的场景。有一类**特别的应用场景**，或者说**“对抗”场景**，其实是希望**利用一切方法来破坏或者绕开分类器的决策结果**。

一个大类的**“对抗机制”**是尝试使用**“对抗样本”**（Adversarial Examples）。什么是对抗样本呢？就是说一个数据样本和原来正常的某个样本是非常类似的，但是可以导致分类决策出现很大不同。例如在我们刚才的图像识别的例子中，一个有效的对抗样本就是一张非常像狗的图片，但是可以导致分类器认为这是一只猫或者别的动物。利用这种类似的样本，可以使分类器的训练和测试都产生偏差，从而达到攻击分类器的目的。

除了“对抗样本”的概念以外，我们再来看一看**攻击分类器**的一些基本的模式。

一般来说，对分类器的攻击有两种模式，一种叫作“**白盒攻击**”（White-Box），一种叫作“**黑盒攻击**”（Black-Box）。白盒攻击主要是指攻击者可以完全接触到分类器的所有内部细节，比如深度模型的架构和各种权重，但无法接触到测试数据。而黑盒攻击则是指攻击者无法接触分类器的细节。

这篇论文考虑的场景是白盒攻击。攻击方尝试针对每一个合法的数据点，去寻找一个距离最近的数据变形，使得分类器的结果发生变化。通俗地说，就是希望对数据进行最小的改变，从而让分类器的准确率下降。

在完全白盒的场景下，最近也有一系列的工作，希望让神经网络更加健壮，从而能够抵御对抗样本的攻击。但是到目前为止，学术界还并没有完全的答案。

## 论文的主要贡献

通过上面的介绍，我们知道目前有一些防御对抗样本的方法，似乎为分类器提供了一些健壮性的保护。这篇文章的一个重要贡献，就是指出，**这些防御方法有可能只是带来了一种由“梯度混淆”（Obfuscated Gradients）所导致的错觉。**

梯度混淆是“梯度屏蔽”（Gradient Masking）的一种特殊形式。对于迭代攻击方法来说，如果发生梯度混淆，防御方会形成防御成功的假象。

作者们在这篇论文中对梯度混淆进行了分析，提出了三种类型的梯度混淆：“**扩散梯度**”（Shattered Gradients）、“**随机梯度**”（Stochastic Gradients）和“**消失梯度或者爆炸梯度**”（Vanishing/Exploding gradients）。

针对这三种不同的梯度混淆，作者们提出了相应的一些攻击方案，使得攻击方可以绕过梯度混淆来达到攻击的目的，并且在 ICLR 2018 的数据集上展示了很好的效果。

值得注意的是，这篇论文针对的是在防御过程中“**防御方**”的方法所导致的梯度混淆的问题。目前学术界还有相应的工作是从攻击方的角度出发，试图学习打破梯度下降，例如让梯度指向错误的方向。

## 论文的核心方法

我们首先来看一看这三种类型的梯度混淆。

**扩散梯度**主要是指防御方发生了“不可微分”（Non-Differentiable）的情况。不可微分的后果是直接导致数值不稳定或者梯度不存在。扩散梯度其实并不意味着防御方有意识地希望这么做，这很有可能是因为防御方引入了一些看似可以微分但是并没有优化目标函数的情况。

**随机梯度**主要是由**随机防御（Randomized Defense）**引起的。这有可能是神经网络本身被随机化了，或者是输入的数据被随机化，造成了梯度随机化。

**消失梯度和爆炸梯度**主要是通过神经网络的多次迭代估值（Evaluation）所导致。例如，让一次迭代的结果直接进入下一次迭代的输入。

刚才我们说了，梯度混淆可能是防御方无意识所产生的结果，并不是设计为之。那么，攻击方有什么方法来识别防御方是真的产生了有效果的防御，还是仅仅发生了梯度混淆的情况呢？

作者们做了一个总结，如果出现了以下这些场景，可能就意味着出现了梯度混淆的情况。

第一种情况，**一步攻击的效果比迭代攻击（也就是攻击多次）好**。在白盒攻击的情况下，迭代攻击是一定好于一歩攻击的。因此如果出现了这种一步攻击好于迭代攻击的情况，往往就意味着异常。

第二种情况，**黑盒攻击的效果比白盒好**。理论上，白盒攻击的效果应该比黑盒好。出现相反的情况，往往意味着不正常。

第三种情况，**无局限（Unbounded Attack）效果没有达到 100%**。最后的这种情况，就是随机寻找对抗样本，发现了比基于梯度下降的攻击要好的对抗样本。

那么，针对梯度混淆，攻击方有什么办法呢？

针对扩散梯度，作者们提出了一种叫**BPDA（Backward Pass Differentiable Approximation）**的方法。如果有兴趣，建议你阅读论文来了解这种算法的细节。总体说来，BPDA 就是希望找到神经网络不可微分的地方，利用简单的可微分的函数对其前后进行逼近，从而达到绕过阻碍的目的。

针对随机梯度，作者们提出了“**变换之上的期望**”（Expectation over Transformation）这一方法。这个方法的特点是针对随机变化，变换的期望应该还是能够反映真实的梯度信息。于是作者们就让攻击方作用于变换的期望值，从而能够对梯度进行有效的估计。

针对消失或者爆炸的梯度，作者们提出了“**重新参数化**”（Reparameterization）这一技术。重新参数化是深度学习中重要的技术。在这里，作者们使用重新参数化，其实就是对变量进行变换，从而使得新的变量不发生梯度消失或者爆炸的情况。

## 小结

今天我为你讲了今年 ICML 的最佳论文。

一起来回顾下要点：第一，这篇论文讨论了一个比较陌生的主题，我们简要介绍了论文的背景；第二，我们详细介绍了论文提出的三种类型的梯度混淆。

最后，给你留一个思考题，我们为什么要研究深度学习模型是否健壮，是否能够经得起攻击呢？有什么现实意义吗？

欢迎你给我留言，我们一起讨论。

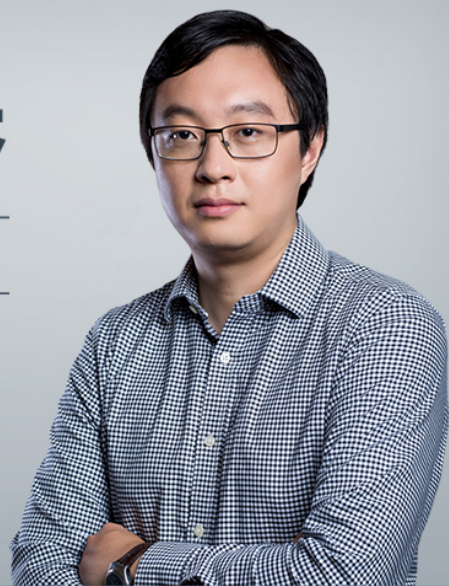



# AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管  
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

上一篇 130 | CVPR 2018论文精读：如何解决排序学习计算复杂度高这个问题？

下一篇 132 | ICML 2018论文精读：聊一聊机器学习算法的“公平性”问题

## 精选留言

 写留言

由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。