2020年3月26日

20:51

Obfuscated Gradients Give a False Sense of Security:

Circumventing Defenses to Adversarial Examples

本文提供ICLR2018中提到的9种防御方法的重实现代码：https://github.com/anishathalye/obfuscated-gradients

文章链接：谷歌学术搜索：<https://scholar.google.com.hk/scholar?hl=zh-CN&as_sdt=0%2C5&q=Obfuscated+Gradients+Give+a+False+Sense+of+Security%3A+Circumventing+Defenses+to+Adversarial+Examples&btnG=>

## 0 Abstract

模糊梯度是一种梯度掩蔽，它是防御对抗样本时会造成一种安全错觉的现象。虽然导致模糊梯度的防御似乎可以抵抗基于迭代优化的攻击，但本文发现可以绕过依赖这种效果的防御。本文针对ICLR2018发表的9种防御方法，攻破了其中6种防御方法，部分攻破了1种防御方法（均是基于模糊梯度的防御方法），未破解防御方法为基于对抗训练的两种防御方法。

## 1 Introduction

针对神经网络对对抗样本的敏感性，最近人们对构建防御系统来增强神经网络的鲁棒性产生了浓厚的兴趣。虽然在理解和防御白盒设置中的对抗样板方面已经取得了进展，在白盒设置中，对手可以完全访问网络，但是还没有找到完整的解决方案。而很多对抗迭代优化攻击的防御方法的共同点为基于模糊梯度。

**Obfuscated gradient（模糊梯度）有三类：**

（1）破碎梯度（shattered gradient）：当防御不可微时，引入数值不稳定性，或以其他方式导致梯度不存在或不正确时，会造成梯度破碎。

（2）随机梯度（stochastic gradients）：随机梯度是由随机防御引起的，a网络本身随机化；b在输入到分类器之前对输入进行随机变换，导致梯度随机化。这会导致使用单一随机样本的方法，从而错误地估计真实梯度。

（3）消失/爆炸梯度（vanishing/exploding gradients）：由神经网络评估的多次迭代组成的防御，在计算展开时，神经网络的深度过深导致梯度消失或爆炸。

**本文贡献：**

而本文提出的新技术可以克服这三种模糊梯度，绕过此类防御方法。

（1）Backward Pass Differentiable Approximation（BPDA）可以绕过梯度破碎；

（2）BDPA结合已提出的EOT（Exception over Transfertation）可以绕过梯度随机化；

（3）BDPA结合Reparameterization可以绕过梯度爆炸；

**Preliminaries**

对抗样本（Adversarial Examples）

给定图像x和分类器f(.)，对抗样本x’，x’满足：

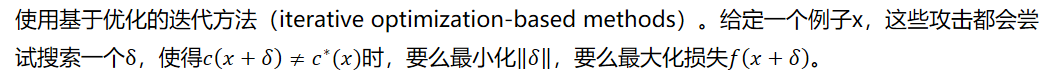
（1）对于一些距离函数D，D(x,x')很小；

C:\D06D6CC5\379C6B87-EEED-49E3-AF4D-80593A9FE99E.files\image002.png

**Threat Models**

本文只设计了针对白盒设置的攻击，即敌手拥有访问神经网络分类器和防御能力，但没有随机化测试时间。

**Attack methods**



**Obfuscated Gradients**

**什么是gradient masking？**

如果在产生对抗样本后没有可用的梯度，就称为梯度掩蔽。它是一个针对对抗样本不完全的防御。过去的防御方式都是通过学习打破梯度下降来形成梯度遮蔽，本文则是利用模糊梯度来形成梯度遮蔽进行防御。

**如何识别模糊或者遮蔽梯度？**

（1）单步梯度攻击效果优于迭代攻击；

（2）黑盒攻击效果优于白盒攻击；

（3）不限定特征值改动大小的攻击无法达到100%成功率；

（4）基于梯度的方法无法发现对抗样本，然而在原样本特征空间邻域内暴力搜索可以发现对抗样本；

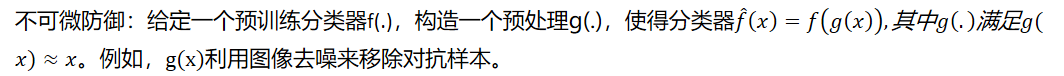
（5）增加特征值改动上限无法提升攻击成功率。

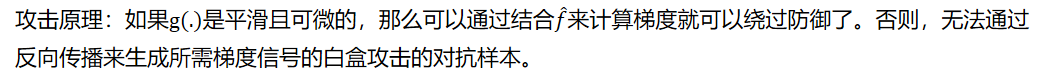
## 2 攻击技术

**Backward Pass Differentiable Approximation**

goal：攻击不易获得的梯度。

针对特殊例子：The straight-through estimator，只有1个x





C:\D06D6CC5\379C6B87-EEED-49E3-AF4D-80593A9FE99E.files\image006.png

计算机生成了可选文字:
OX 
／ 丿 
9 

即

计算机生成了可选文字:
v ， 7 与 ， 川 ， 
Vxf(x)l 

**一般情况下，BPDA原理：**

C:\D06D6CC5\379C6B87-EEED-49E3-AF4D-80593A9FE99E.files\image009.png

C:\D06D6CC5\379C6B87-EEED-49E3-AF4D-80593A9FE99E.files\image010.png

C:\D06D6CC5\379C6B87-EEED-49E3-AF4D-80593A9FE99E.files\image011.png

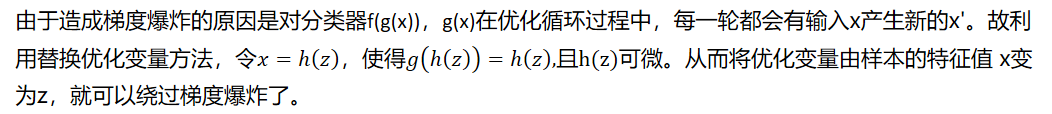
注意到这里仅对后向传播进行函数替换，已经证明同时对前向和后向替换要么完全无效，要么传递次数越多时效果越低。

缺点：由于每个单独的梯度下降步是近似的，不完全正确，使用BPDA需要更多的梯度下降迭代。

**攻击随机化分类器：**

利用EOT(Exception over Transformation)，针对进行随机变换的模型输入，攻击者可以找一个对应的变换的分布，根据分布变换后的数据的模型输出期望来寻找对抗样本。

**攻击梯度消失或爆炸的分类器：**



## 3 Case Study: ICLR 2018 Defenses

文章进一步分析了针对Thermometer encoding，Adversarial training，以及二者结合的对抗样本失真度与攻击效果（模型分类准确率）的关系。此处只介绍了文章所分析的一种防御攻击。

**Gradient Shattering-温度编码（THERMOMETER ENCODING）**

防御原理：对抗样本在神经网络中被称为“盲点”，存在是因为神经网络的行为很多是线性的。而温度编码视为了打破线性。

计算机生成了可选文字:
丫 0 
二 0 
钗 石 么 直 的 秭 

本文作者发现这种防御方法虽然可以打破局部线性，但会造成梯度破碎。在标准对抗训练模型产生的对抗样本转移到温度编码的模型后，准确度下降了很多，低于了白盒攻击的准确率。

利用本文提出的攻击：

计算机生成了可选文字:
' 丁 （ 讠 c) 二 ” 7 n 0 （ ， c 

实验发现，Thermometer encoding不能防御BPDA攻击。反而加入了Thermometer encoding的Adversarial training劣于Adversarial training的防御效果。进一步说明了目前基于混淆梯度的防御方法针对BPDA攻击无效。

计算机生成了可选文字:
0 ． 00 
0 ． 01 
Baseline 
Thermometer 
Adv. Train 
Adv. Therm 
0 ． 02 
0 ． 03 
Perturbation Magnitude 
Figur ℃ 7 ． Model accuracy versus distortion (under ～ ） ． Adversar- 
ial training increases robustness to 50 ％ at 6 = 0 ． 031 ； thermometer 
encoding by itself provides limited value ， and when coupled with 
adversarial training performs worse than adversarial training alone. 

## 4 实验

本文针对ICLR2018年提出的9种防御方法进行了攻击，

基于离散梯度的防御三种，为列表前三个；

基于随机梯度防御两种，为列表第四和五个；

基于梯度消失和爆炸有两种，为列表第6和第7个；

不基于模糊梯度防御的为对抗训练和级联对抗训练防御两种，为列表最后2个。

计算机生成了可选文字:
Defense 
Buckman et al. (2018 ） 
Ma et al. （ 2m8 ） 
Guo et al. （ 2m8 ） 
Dhillon et al. （ 2m8 ） 
Xie et al. （ 2m8 ） 
Song et al. （ 2m8 ） 
S amangouei et al. 
（ 2m8 ） 
Madry et al. (2018 ） 
Na et al. （ 2018 ） 
Dataset 
CIFAR 
CIFAR 
ImageNet 
CIFAR 
ImageNet 
CIFAR 
MNIST 
CIFAR 
CIFAR 
Distance 
0 ． 031 仔 ～ ） 
0 ． 031 仔 ～ ） 
0 ． 005 仔 2 ） 
0 ． 031 仔 ～ ） 
0 ． 031 仔 ～ ） 
0 ． 031 仔 ～ ） 
0 ． 005 仔 2 ） 
0 ． 031 仔 ～ ） 
0 ． 015 仔 ～ ） 
Accuracy 
5 ％ 
0 ％ 
55 ％ * * 
47 ％ 
15 ％ 

由攻击结果可以看到，除了对抗训练两种防御，BPDA均实现了完全攻击，将模型分类准确率限制在了10%以下，而级联对抗训练防御也实现了部分攻击。

## 5 总结

设计对抗样本的防御不仅需要防御现有的攻击，还需防御未来可能出现的攻击。同时本文作者建议未来的工作需尽量避免使用基于模糊梯度的防御方法，或者仅防御基于梯度下降的攻击。