Why Do Adversarial Attacks Transfer？ Explainning Transferablility of Evasion and Poisoning Attacks

2020年3月18日

23:02

Ambra Demontis†, Marco Melis†, Maura Pintor†, Matthew Jagielski\*, Battista Biggio†,‡, Alina Oprea\*,

Cristina Nita-Rotaru\*, and Fabio Roli†,‡

†Department of Electrical and Electronic Engineering, University of Cagliari, Italy

‡Pluribus One, Italy

\*Northeastern University, Boston, MA, USA

文章链接：https://www.usenix.org/conference/usenixsecurity19/presentation/demontis

**本文贡献：**

本文考虑基于梯度的优化技术（gradient-based optimization）进行攻击。

（1）提出了一个针对躲避和投毒攻击的新的优化框架，首次将躲避和投毒攻击包含在同一框架下。第一个对投毒可用性攻击的可转移性进行深入研究。

（2）提出了可迁移性的正式定义和理论边界。定义表明迁移性和3个因素有关：目标分类器的输入梯度大小；代理模型和目标模型的梯度对齐情况；损失函数的方差优化。

**1 Abstract**

可迁移性（transferability）是指将对机器学习模型有效的攻击迁移到未知模型的能力。但攻击迁移的原因未知。本文对逃避攻击和投毒攻击的可迁移性进行了全面分析，给出了一个逃避和投毒攻击的优化框架，以及这类攻击的可迁移性的正式定义。强调了影响可迁移性的两个主要因素：（1）目标模型固有的对抗脆弱性（adersarial vulnerability）；（2）用于优化攻击的代理模型的复杂性。基于上述分析，定义了3个影响攻击迁移能力的三个指标。

**2 Introduction**

**投毒可用性攻击（poisoning availability attacks）**：攻击者控制一定数量的训练数据，从而影响训练模型，最后影响测试时测试集上大多数点的预测。投毒完整攻击（poisoning integrity attacks）：通过控制训练过程来修改少数目标点的预测。逃避攻击（evasion attacks）是指通过对测试数据的微小操作，导致在测试时对这些点的错误预测。

**为什么要使用black box？**

因为很多在线服务商不会公开他们所用的ML算法，设置投毒点和躲避点很难，因而需要利用黑盒设置代理模型进行攻击，进而迁移到真实模型上。

**什么是攻击可迁移性？**

攻击的可迁移性（transferability）是指当该攻击对一个特殊ML模型有效时，对目标模型也有效。

**本文研究原因？**

之前大部分攻击迁移性研究都是针对的躲避攻击，最近才开始聚焦到投毒攻击上。同时关于对抗点迁移的when and why问题存在着大量未解之谜。

**3 background and threat model**

**本文的攻击者定义：**

（1）目标为攻击系统。即破坏完整性（不影响正常系统操作情况下躲避检测），或破坏可用性（破坏合法用户的正常系统可用性）

（2）拥有系统知识k。包括：训练数据D、特征集X、学习算法f和模型训练后的参数w。

（3）拥有操控输入data来影响系统的能力。

**白盒攻击（white-box attacks）**

攻击者知晓目标分类器的所有知识k，可以对ML算法安全进行最坏的评估。

**黑盒攻击（balck-box attacks）**

攻击者知晓模型的输入特征集，但不知道训练集D和分类器的类型f。更为现实的考虑是攻击者都无法访问分类器。攻击者通过从和目标模型相同的底层数据分布中采样，构造一个代理训练数据集，通过这个代理数据集训练处代理模型。然后对代理模型进行攻击实验，将获得的有效攻击方法迁移到目标模型上测试攻击效果。

**4 Optimization Framework for Gradient-based Attacks**

这个优化框架不仅包含了逃避和投毒攻击，还支持设计新型攻击。

二分法（bisection method）：<https://zh.wikipedia.org/wiki/%E4%BA%8C%E5%88%86%E6%B3%95_(%E6%95%B8%E5%AD%B8)>

Projection operator（投影算符）：向量投影。

**优化的攻击策略公式为：**

计算机生成了可选文字:
x e arg maxA(x ， K ） ． 
e 巾 、 
(l) 

其中k为攻击者拥有的知识，x'为攻击样本，y为攻击样本的标签。x\*可以衡量攻击样本的有效性。A函数为目标函数（可以是损失函数）。

本文考虑使用对逃避和投毒攻击通用的攻击算法：基于梯度的躲避和投毒攻击算法。原理为：根据目标函数的梯度，迭代更新攻击样本，从而保证通过投影所得的结果点在可行域内。如下：

计算机生成了可选文字:
Algorithm 1 Gradient-based Evasion and Poisoning Attacks 
Input: 、 节 the input sample and its label; （ 、 斗 K) ： the at- 
tacker's objective; K = （ 刃 ” ， 了 ， w): the attacker's knowl- 
edge parameter vector; （ ， ： the feasible set of manipu- 
lations that can be made on \ ； t > 0 ： a small number. 
Output: \ ' ： the adversarial example. 
Initiallze the attack sample: x' 民 一 \ 
2 ： repeat 
Store attack from previous iteration: \ 民 一 \ 
3 ： 
Update step: \ ' 艹 + 11V01 （ 、 、 K)), where the 
4 ： 
step size 11 is chosen with line search (bisection method) ， 
and ensures projection on the feasible domain 
5 ： until 区 ” ， K ） 一 、 K)I 
6 ： return \ ' 

根据输入的样本x，标签y，系统知识k，以及目标函数A，先初始化（找到一个x的近似样本x'），然后进行迭代，利用基于二分法的线性搜索算法确定梯度大小 ，然后通过 投影保证攻击点在可行域范围内，多次迭代更新攻击样本，直到找到局部或全局最优值（即攻击所用函数尽可能的接近目标函数）。

**躲避攻击（evasion attacks）优化**

根据优化策略（1）重写为：

计算机生成了可选文字:
max 
s.t. 
召@公 w 
Xlb 粼 \ 粼 Xub ， 
（ 2 ） 
（ 3 ） 
（ 4 ） 

|| ||p矩阵范数的定义：<https://baike.baidu.com/item/%E8%8C%83%E6%95%B0>

此处假定w是已知的。本文中考虑l(y,x',w)=-yf(x')。

目标（2）：攻击者最大化原始分类的对抗样本损失，从而使得目标分类的错误分类。

约束（3）是指对输入x的最大扰动值约束；

约束（4）是指对盒子的攻击范围约束，保证不影响正常范围下攻击。

（例子：对于图片来说，前者约束范围为其遭受稀疏攻击或稠密攻击的扰动范围；后者约束则保证无需操作的像素值不会被操作到，只会操作图片的特定区域。）

**提高攻击转移性因素：**

（1）最大化分类器的信任；（2）从不同的初始化点开始攻击；

**投毒攻击（Posioning attacks）**

攻击目标：要么在不影响系统正常操作下注入攻击（破坏完整性）；要么故意破坏系统操作导致拒绝攻击（破坏可用性）。后者研究还不多，且更难。

本文工作聚焦于后者，且公式基于白盒攻击。根据通用策略（1）重写为：

计算机生成了可选文字:
max 
默 a 卜 w ） 
w arg minL(Dtr1J(x 01w ） 
（ 5 ） 
（ 6 ） 

其中Dtr是在中毒数据的基础上训练的数据集，Dval是用来测试的没有被污染的数据集。

w\*是训练所得模型参数，L(Dval,w\*)是目标函数。同时攻击者能力有限，只能注入小部分投毒点。攻击者需要解决加入中毒点后的训练集的优化问题。利用算法1即可。

**5 Transferability Definition and Metrics**

**攻击点定义**：

计算机生成了可选文字:
x 
一 x 十 5 ． 

其中x是初始化点， 是通过攻击算法对抗代理分类器获得的对抗扰动。

**躲避攻击转移性的定义**：通过目标分类器在攻击点x\*所获得的损失。损失函数的线性近似表达为：

计算机生成了可选文字:
丆 一 x 十 、 w) 、 w) 十 öTVxC(y,x w 
（ 13 ） 

这种近似不仅适用于足够小的微小扰动，同时如果分类函数是线性或只是小卷曲的也适用于Larger扰动。

在不考虑（4）约束下，可以把逃避攻击优化策略再次重写为：

计算机生成了可选文字:
5 e argmaxC(y,x 十 ，會） ． 
Ilö 牖 
（ 14 ） 

线性近似后，可以转化为内部优化问题：

计算机生成了可选文字:
max öTv ` 小 ： 、 一 v 、 ， 
Ilö 
（ 15 ） 

影响可转移性的指标定义：

**（1）输入梯度大小（size of input gradient）**

计算机生成了可选文字:
（ 17 ） 

使用目标分类器所计算的损失梯度，不考虑代理分类器情况下，梯度越大，攻击效果越好。同时梯度大小也取决于模型的复杂度，模型越复杂，正则化越弱，输入梯度值就越大，模型健壮性越差，只需要输入微小的扰动就能成功攻击。（只能在同一学习算法下比较）

计算机生成了可选文字:
0 ． 15 
弓 0 ． 10 
0.05 
0.0 
High complexity 
test error ()o attack) 
test error = 0 ． 3 ） 
LOW comple 
0 ． 6 
1 ． 0 
0 ． 8 
0 
0 ． 2 
0 ． 0 
Regularization (weight decay) 

（2）梯度对齐（gradient alignment）

计算机生成了可选文字:
^ 下 
/ 丛 诩 2 ， ` 鼐 2 · 
08 ） 

这其实就是代理模型损失的梯度与目标分类器的梯度之间的夹角余弦。即代理模型和目标模型的梯度对齐越好，那么攻击转移性越好。和输入梯度大小不同的是，这可以在不同学习算法之间比较。

（3）损失景观可变性（variability of the loss landscape）

计算机生成了可选文字:
（ 19 ） 

C:\4C6A4C45\C8A0CEBC-8F7D-4B3B-B499-B20188CDE31A.files\image013.png

计算机生成了可选文字:
V' ' 区 
X 

绿线代表使用代理模型根据不同的训练集所得到的期望损失。灰色区域代表损失景观可变性。

上述因素同样适用于投毒攻击。

**5 实验分析**

对影响攻击效果的三个因素进行了分析，同时提供了一个如何选择最有效代理模型来设计黑盒可转移攻击的建议。

**逃避攻击分析**

考虑手写数字识别（handwriting digit recognition）和安卓恶意软件检测（Android malware detection）两种情况

计算机生成了可选文字:
White-box evasion attack (MNlST89) 
logisticH 
logisticl. 
ridgeH 
ridgeL 
SVM-RBFH 
SVM-RBFL 
Figure 5 ： White-box evasion attacks on MNIST89. Test error 
against increasing maximum perturbation 8 ． 

在白盒逃避攻击中，可以看到复杂度高的模型显然攻击成功率更高。

计算机生成了可选文字:
； _ SVM 
厂 _ logistic 
厂 _ SVM-RBF 
_ NN 
0.4 
Size of input gradients （ 习 

在各个ML算法下，输入梯度和攻击成功率均成正比。

分析：对每个学习算法，低复杂度有更低的输入梯度，其躲避漏洞更少，因而攻击更难。同时还发现非线性分类器比线性分类器的漏洞更少。

计算机生成了可选文字:
SVMg 
SVML 
logisticg 
logistiCL 
ridgeu 
ridgeL 
SVM-RBF„ 
SVM-RBFL 
NNÆ 
NNL 
0 14 
032 
m18 
O 0 
m12 
022 
025 
030 
020 
030 
0 35 
0 5 
m20 
O. 0 
m10 
， 25 
m35 
m10 
， 20 
0 ． 33 
， 3 
m20 
m37 
0_5 
m23 
0 1 
m50 
m51 
& 13 
0.1B 
& 2 
0.1B 
& 27 
0 21 
0 25 
0 22 
& 2 
025 
032 
0 鸶 3 
028 
0 鸶 7 
0 # 0 
O 7 
041 
0 0 
0 20 
035 
0 0 
021 
0 30 
0 57 
m32 
0 2 
m50 
m25 
0 0 
0 上 7 
m28 
0 ． 30 
m51 
m21 
， 40 
， 05 
0 32 
& 繼 
& 2 
0 ， 
O. 
Figure & Gradient alignment and perturbation correlation 
for evasion attacks on MNIST89. ． Gradient alignment 
人 (Eq. 18 ） between surrogate (rows) and target (columns) 
classifiers, averaged on the unmodified test samples ． 人 / it: ． 
Pearson correlation coefficient p(), 司 between white-box and 
black-box perturbations for £ = 5 ． 

分析：根据图8，可以看到，梯度对齐可以精确测量攻击的转移性。余弦相似度越高，相关性越高，意味着所设计的代理模型和目标模型越相似。

计算机生成了可选文字:
10 一 
Variability of loss landscape (V) 

分析：V值越低，其转移性更好。

投毒攻击实验与上述基本一致，但对投毒攻击来说，找到一个更好的梯度对齐比降低损失景观可变性更重要。

**6 结论**

综上所述，对于规避攻击，通过适当调整代理模型学习算法的超参数来降低代理模型的复杂度，可以拥有更好地转移到目标模型的对抗实例。对于投毒攻击，最好的代理模型是与目标模型具有相似正则化水平的模型，其原因是中毒目标函数相对稳定。

攻击建议：对于黑盒逃避攻击，最好选择低复杂度的代理模型（如使用强正则化和减小模型方差）；对于投毒攻击，尽可能地去获取目标模型的正则化水平信息，从而训练与目标模型正则化水平相似的代理模型。

防御建议：与更复杂的模型相比，更低复杂的模型更为弹性。同时还要考虑偏置-可变性的trade-off，使用在初始预测阶段表现相对较好的模型。