基于输入转换的对抗攻击防御方法探究

高铭　　1801210045

廖显祚　　1801210053

李宝华　　1801210077

http://discourse.hzwer.com/t/topic/305

DEEPLEARNING COURSE 2019

CONTENT

1

背景简介

2

实验

3

结果汇报与分析

4

总结与展望

01

背景简介

常见的攻击方法

什么是基于输入转换的防御方法

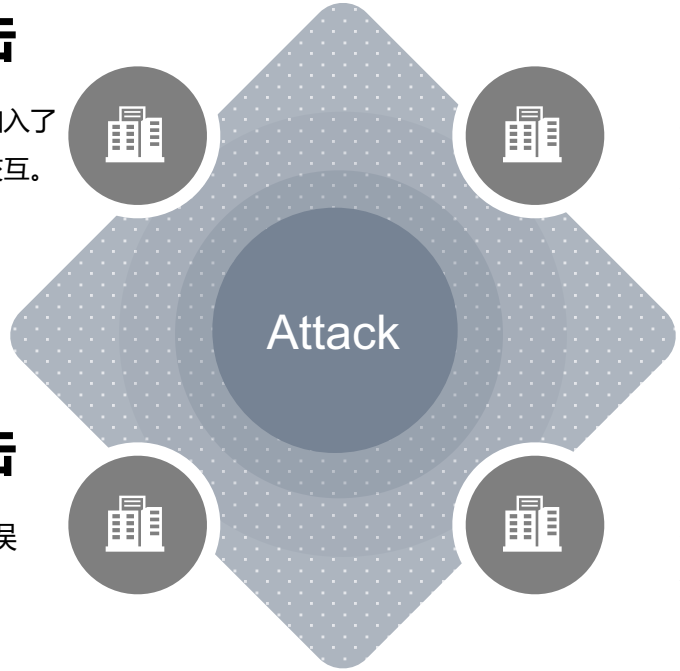
对抗攻击分类

黑盒攻击

攻击者对攻击的模型的内部结构，训练参数，防御方法（如果加入了防御手段的话）等等一无所知，只能通过输出输出与模型进行交互。

白盒攻击

与黑盒模型相反，攻击者对模型一切都可以掌握。目前大多数攻击算法都是白盒攻击。



无目标攻击

以图片分类为例，攻击者只需要让目标模型对样本分类错误即可，但并不指定分类错成哪一类。

有目标攻击

攻击者指定某一类，使得目标模型不仅对样本分类错误并且需要错成指定的类别。从难度上来说，有目标攻击的实现要难于无目标攻击。

FGSM（Fast Gradient Sign Method）

最简单的白盒攻击方法，作为非目标攻击

$$x' = x + \varepsilon sign(\nabla_x L(x, y))$$

x为输入图像，y为结果标签，L为损失函数，ε是步长

基于梯度的攻击方法

按照损失函数的梯度方向增加或减少x的值

既可用来进行有目标攻击，也可以用来无目标攻击

BIM（Basic Iterative Method）

扩展了fgsm的方法，作为非目标攻击，每次迭代几步

$$x_0 = x, x'_{m+1} = x'_m + Clip_{x,\varepsilon}(\alpha sign(\nabla_x L(x'_m, y)))$$

x为输入图像，y为结果标签，L为损失函数
迭代多次

基于梯度的攻击方法

按照损失函数的梯度方向多次增加或减少x的值，但是限制在x的邻域内

C&W（Carlin&Wagner Attacks）

设计一个损失函数使得在对抗样本有较小值，干净样本有较大值

作为非目标攻击

$$r_n = 0.5(\tanh(\omega_n) + 1) - X_n$$
$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -k)$$
$$loss = \|r_n\| + c * f(0.5(\tanh(\omega_n) + 1))$$

X_n为输入图像，Z(x)为x通过模型未经过softmax的输出向量，最大值下标记为t，k为置信度

对抗样本映射到tanh空间，取值可以在-inf到+inf变换，有利于优化

基于优化的攻击方法

多次迭代，可以调节置信度

基于输入转换的防御方法

基于输入转换的防御方法指对输入图像进行某种变换来部分地消除对抗性扰动。

其可见的好处在于不需要太多额外的计算，且与模型无关。
但是这种方法又似乎过于简单，无法从输入图像中充分消除对手的扰动。

我们对一些研究中提出的转换方法进行了实验，包括bit depth reduce（像素深度缩减）和图像平滑，我们另外尝试了SVD对图像进行压缩的转换方法。

02 实验

实验过程

我们对 MNIST，cifar10，SVHN 这三个数据集进行了实验。

对每个数据集，我们首先训练了一个神经网络，然后采用 FGSM、BIM、C&W 对1000张分类正确的图片进行无目标白盒攻击，测试模型在对抗样本上的准确率；

接着采用了3种输入转换方法对对抗样本进行转换，重新测试模型的分类准确率；

同时，我们对原测试集的样本进行了输入转换并测试分类准确率，目的是为了衡量这些输入转换方式是否降低模型的原本性能

三种输入转换方法

- bit depth reduce

$$f(\mathrm{x}) = \mathrm{int}(\mathrm{x} * \mathrm{depth\_num})/\mathrm{depth\_num}$$

我们实验 了depth_num = 8 和 32 的情况

- 图像平滑

我们实验了均值滤波，滤波器大小为2x2

- SVD

对图片作SVD分解，然后取95%的主成分

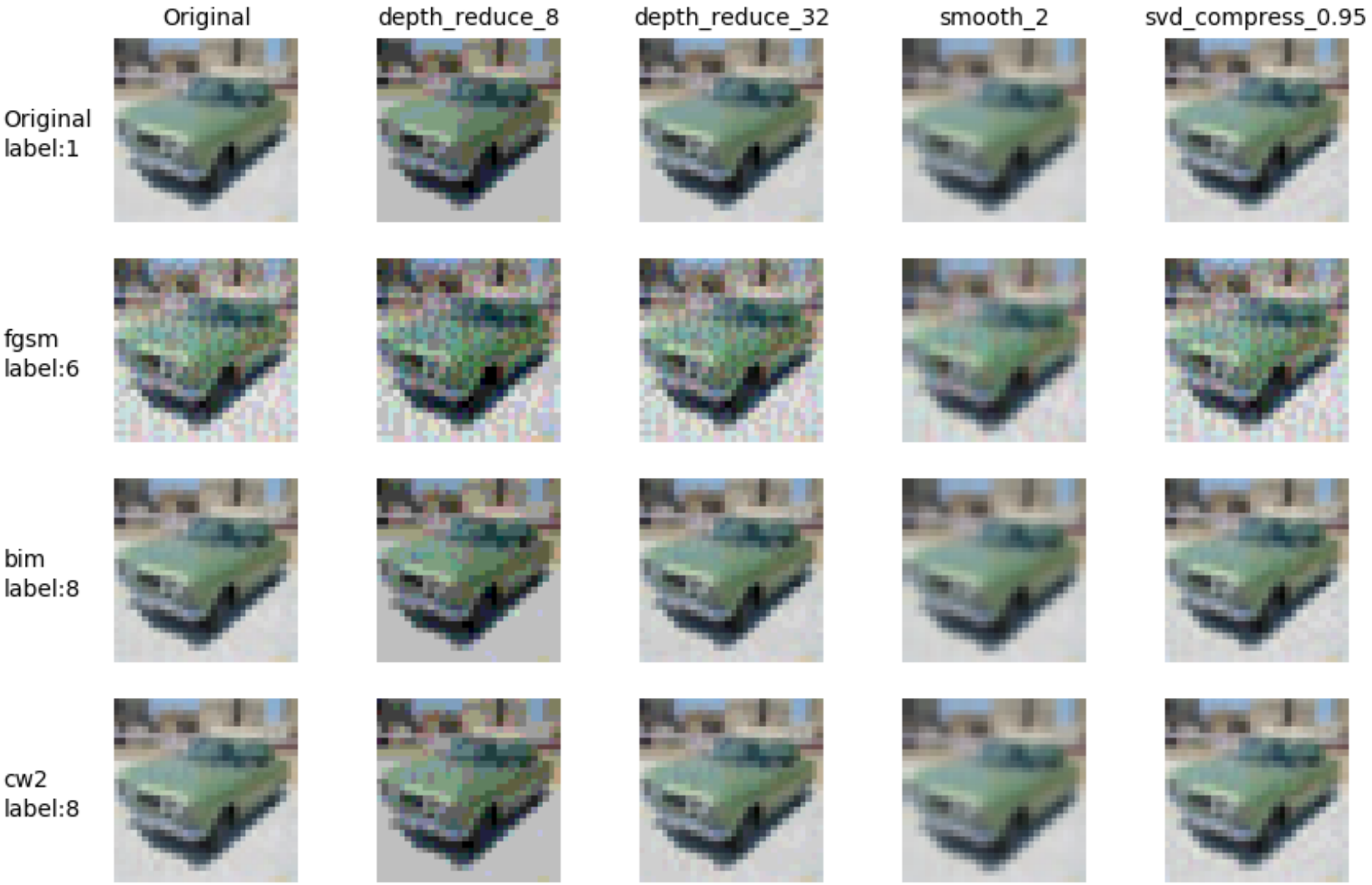
03

结果汇报和分析

MNIST	Origin	l2	depth_reduce_8	depth_reduce_32	smooth	svd_compress
Origin	99.10	-	99.16	99.10	98.76	99.04
FGSM	91.5	0.0640	97.4	91.3	94.3	92.0
BIM	74.9	0.0633	98.2	95.2	94.8	91.7
C&W_2	79.4	0.2579	92.2	99.10	80.60	83.9
Cifar10						
Origin	82.30	-	75.44	81.94	72.18	78.17
FGSM	4.4	0.0494	6.40	4.5	12.6	5.0
BIM	0.0	0.0073	50.30	11.20	68.3	52.4
C&W_2	27.8	0.0111	41.0	27.80	60.6	34.1
SVHN						
Origin	95.24	-	93.78	95.17	95.09	91.63
FGSM	23.4	0.0497	24.1	23.2	30.3	23.4
BIM	0.2	0.0145	50.0	8.8	78.0	57.4
C&W_2	60.9	0.108	63.9	61.3	63.8	70.7

表 1: 结果。每一行代表数据集的某种对抗样本进行各种输入转换方法之后的分类正确率（Origin 行代表原测试集），Origin 列代表不进行输入转换的分类正确率，l2 列代表对抗样本与原样本的 l2(平均之后) 的差异，即每个像素的平均的扰动大小

直观上展示各个攻击方法的扰动，和各个转换方法对原图片的影响



结论和发现

- 基于输入转换的方法能够抵御住一部分的攻击，但同时或多或少会影响模型原来的性能
- FGSM对于输入转换的防御方法相对稳健，但是对抗样本的扰动看上去更加明显；BIM 很容易被输入转换的方法防御住，但是生成的对抗样本扰动看上去很小；C&W对于输入转换的防御方法也比较稳健，对抗样本质量也高，但是生成对抗样本很耗计算力和时间。
- 输入转换的方法的初衷是减少对抗性扰动。从图像上看，图像平滑方法能更好地消除对抗性扰动，但是也使原图片更加模糊，对模型性能影响较大。

04 总结与展望

总结与展望

- 基于输入转换的方法能够抵御一部分的对抗攻击，但是如何同时保证不影响模型的性能值得研究，可能的想法是对训练集进行增强。
- 基于输入转换的方法旨在减少对抗性扰动，普通的输入转换方法实验证实都太过简单粗糙。深入地研究对抗样本的这种对抗性扰动的特征，（如扰动怎么影响模型各个层的输出），设计针对性的方法减少这种扰动。
- 由于时间限制，我们的实验内容也比较粗糙，可以考虑对黑盒攻击进行实验，和尝试对ImageNet这样的大数据集进行实验，可以对更多的攻击方法和防御方法进行实验。
- 另外，针对基于输入转换的防御方法进行对抗攻击也值得研究。

参考文献

- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015.
- Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. In International Conference on Learning Representations (ICLR) Workshop, 2017.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy, 2017.
- Dongyu Meng and Hao Chen. MagNet: a Two-Pronged Defense against Adversarial Examples. In ACM Confer- ence on Computer and Communications Security (CCS), 2017.
- Guo C , Rana M , Cisse M , et al. Countering Adversarial Images using Input Transformations[J]. 2017.
- Xu W , Evans D , Qi Y . Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks[J]. 2017.