Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

提高对抗鲁棒性的特征降噪方法

硕士研究生　于浩淼

2021年04月18日

内容提要

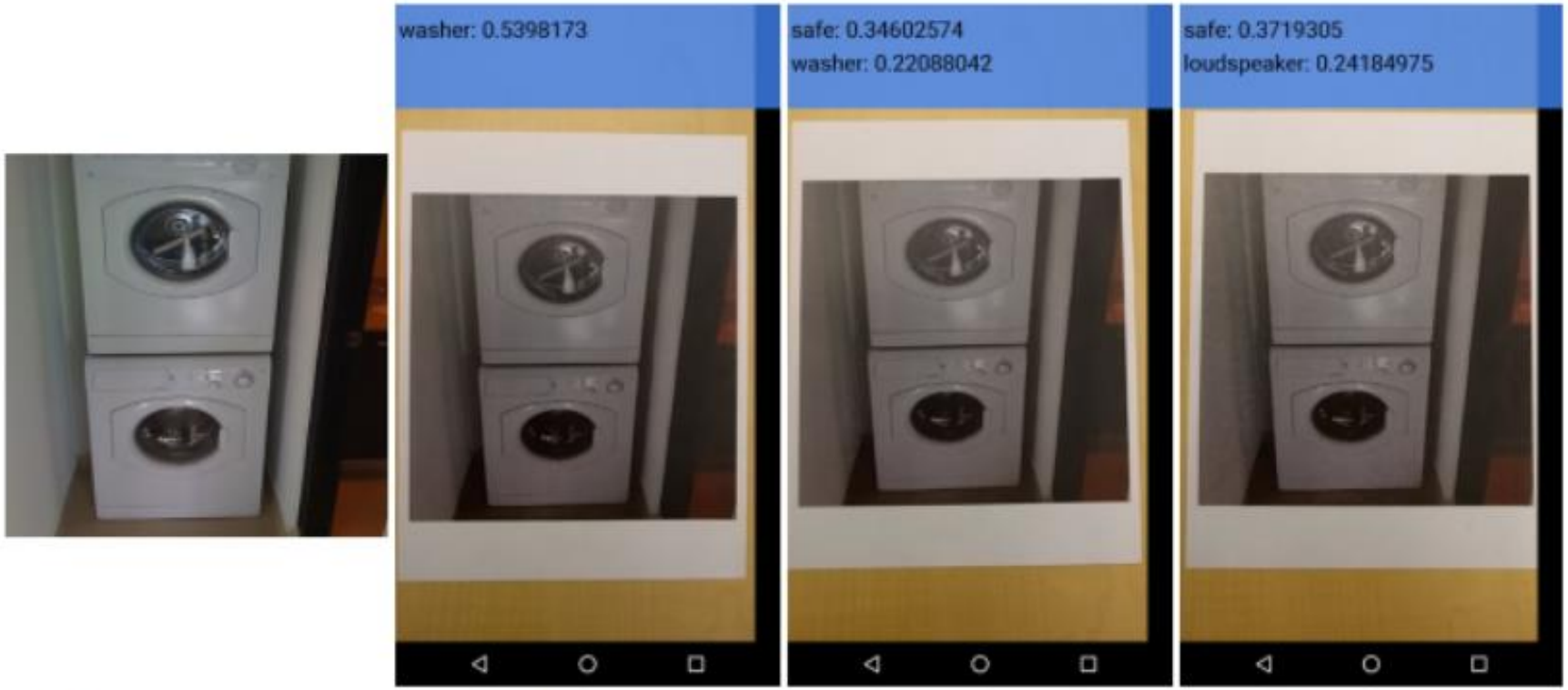
- 背景简介
- 基本概念
  - 对抗样本
  - 残差网络
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 了解对抗样本攻防的主要方法分类
  - 2. 了解残差网络的结构和原理
  - 3. 理解提高对抗鲁棒性的特征降噪方法
  - 4. 了解对抗样本在网络安全领域的应用

背景简介

- 对抗样本应用
  - 把对抗图像打印出来误导图像分类器，使其把洗衣机判断成保险箱



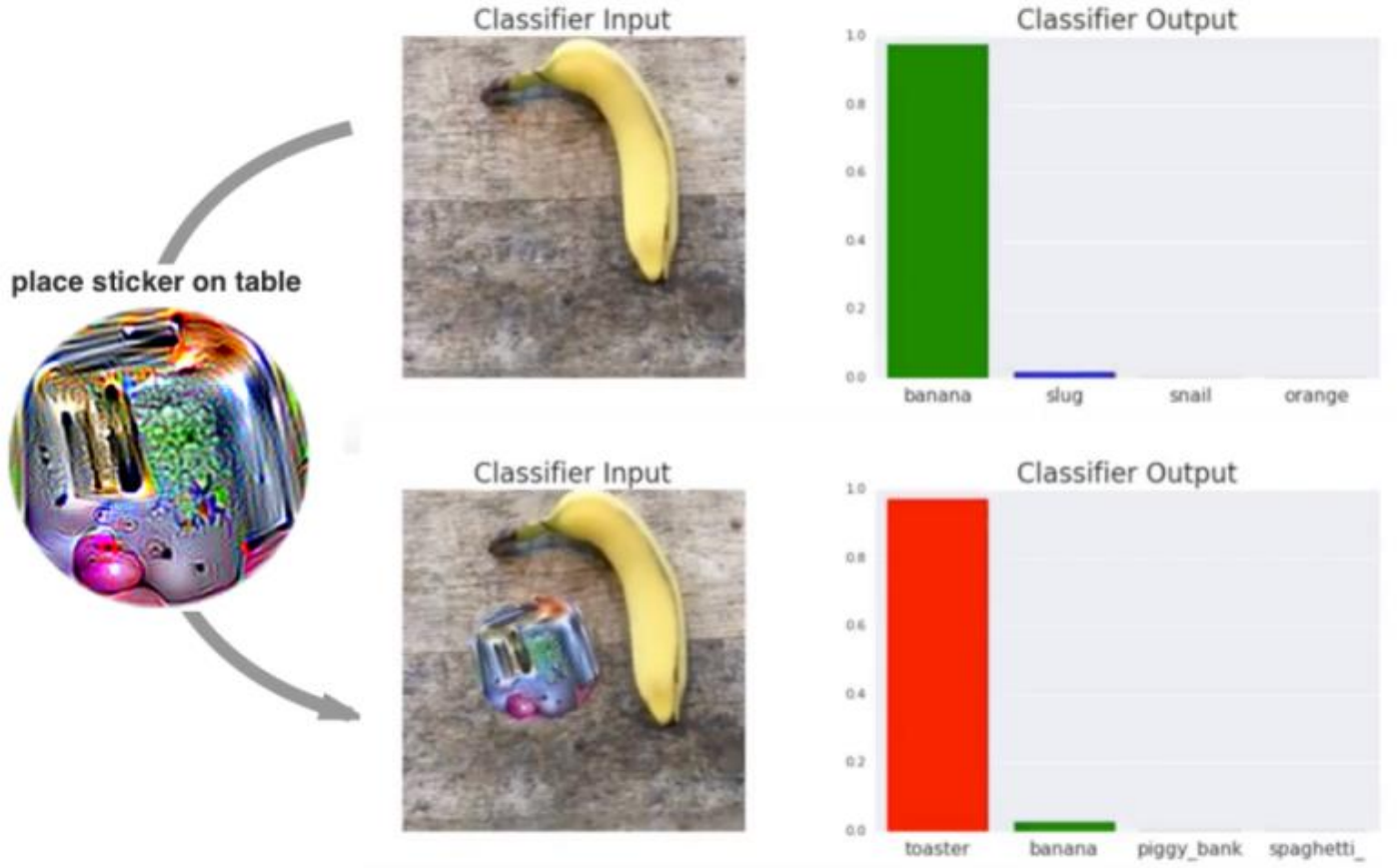
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

背景简介

- 对抗样本应用
  - 打印的对抗眼镜误导人脸识别系统，使其把每列上面的人预测为下面的人

背景简介

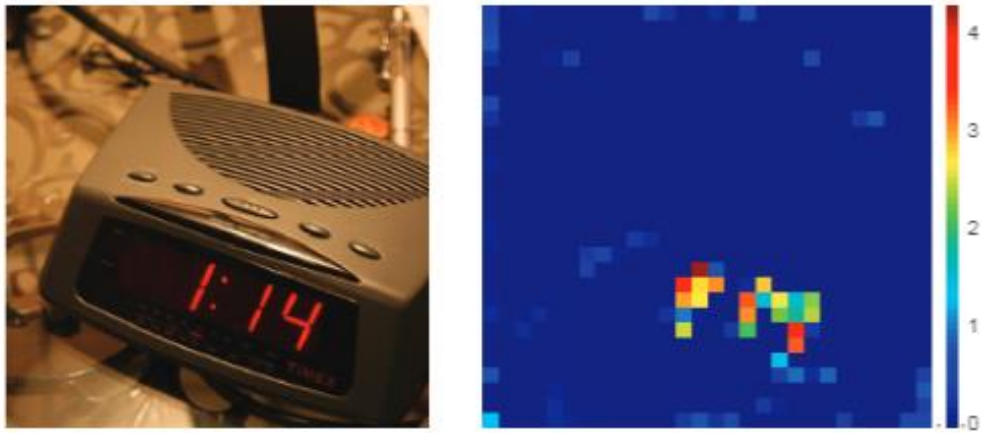
- 对抗样本应用
  - 对抗补丁误导分类器，使其把香蕉预测成烤面包机



基本概念

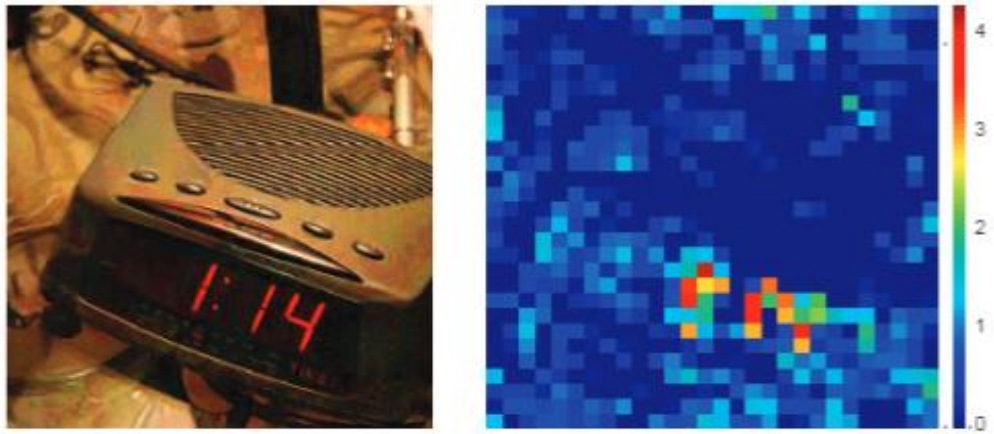
对抗样本

- 对于深度神经网络能够正确识别的原始样本，在有针对性的加入不易被人眼所察觉的微小扰动后，导致深度神经网络识别错误的样本被称为对抗样本
- 从卷积神经网络的特征图中可以看出，干净样本的特征主要集中在图像的语义信息内容上，而对抗样本的特征在语义无关的区域也被激活



clean

正确识别为数字表

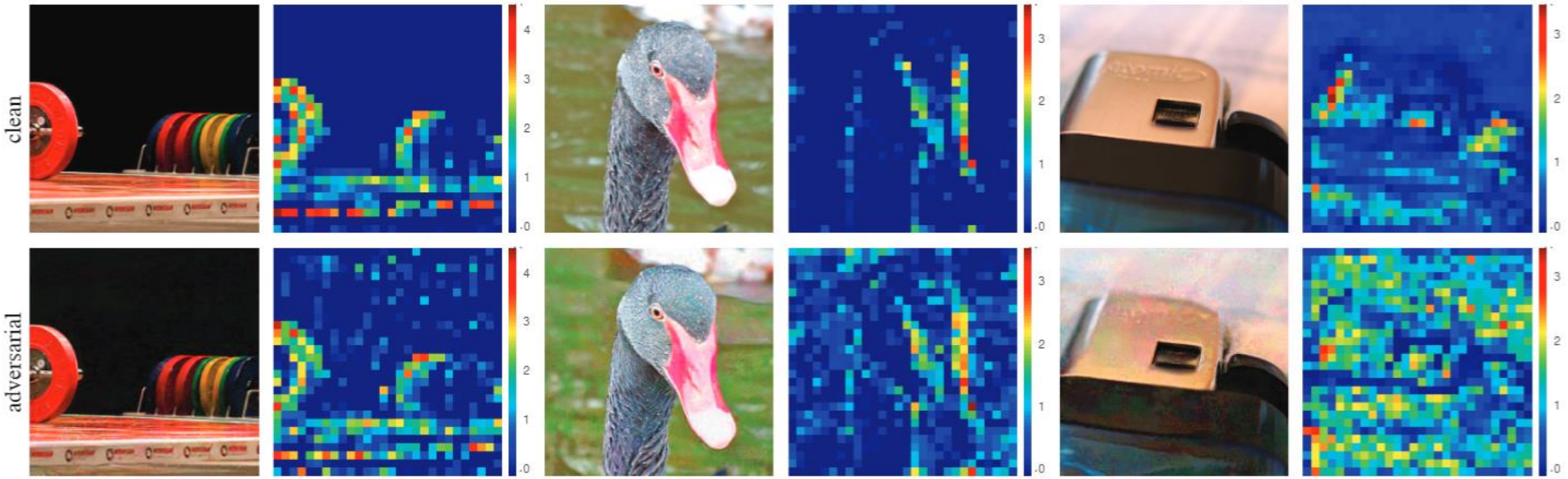


adversarial

错误识别为供暖器

对抗样本

- 对抗样本及其特征图对比实例



对抗攻击

- 根据攻击者对攻击目标模型所掌握信息的程度，可以将对抗攻击分为
- 白盒攻击
  - 攻击者完全了解目标模型内部信息，包括模型结构、模型参数、训练方法和训练数据
  - 白盒攻击的典型例子是梯度攻击，攻击者已知模型的结构和参数，利用反向传播的梯度构造对抗样本
- 灰盒攻击
  - 攻击者不完全了解目标模型内部信息，只了解模型推理输出的概率值等信息
- 黑盒攻击
  - 攻击者无法获取目标模型内部信息，只了解输入数据和输出标签信息

对抗攻击

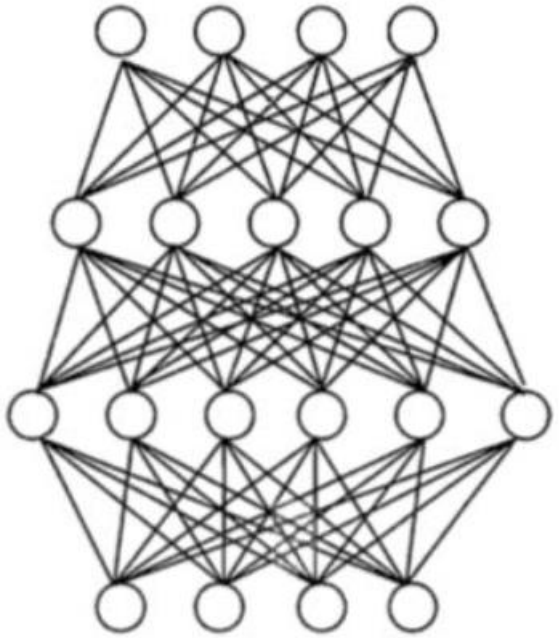
- 根据攻击者有无针对性目标，可以将对抗攻击分为
- 定向攻击
  - 攻击者产生的对抗样本能够让神经网络把对抗样本分类成攻击者指定类别
- 非定向攻击
  - 攻击者产生的对抗样本能够让神经网络将对抗样本分类为除正确类别以外的任意类别
  - 非定向攻击更容易实现

对抗鲁棒性

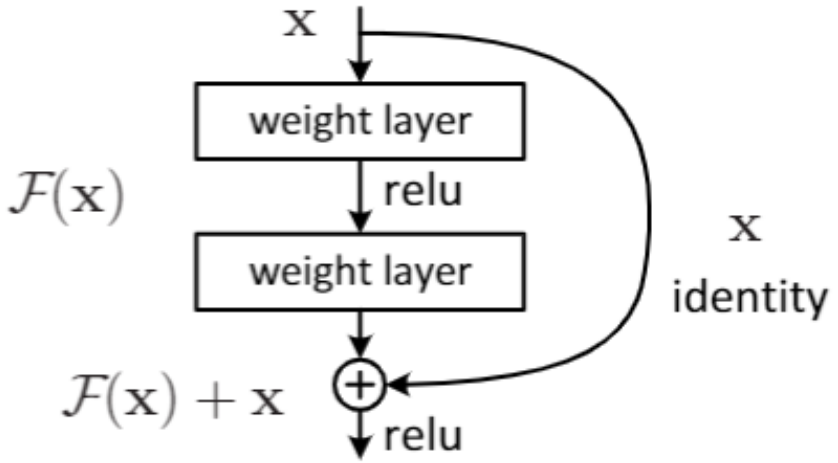
- 为防御对抗攻击，需要增强被攻击系统的鲁棒性
- 在输入部分增强：
  - 对抗样本检测
    - 添加辅助网络检测、多个模型预测一致性判别、统计数据区分
  - 输入转换
    - 通过转换输入来减少模型输入的扰动信息
- 在模型部分增强：
  - 梯度隐蔽/模糊
    - 通过隐藏模型的梯度信息，以防御基于梯度的攻击
  - 对抗训练
    - 对抗训练是指引入对抗样本和真实标签进行训练
    - 对抗训练能够增强模型的鲁棒性；但也会使得深度神经网络的训练成本翻倍

残差网络

- 残差结构



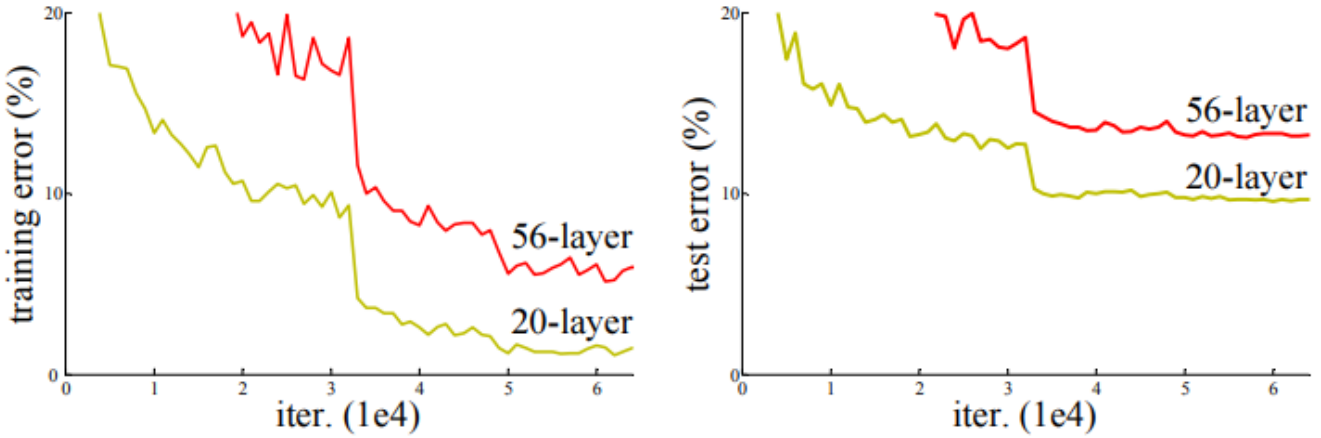
前馈神经网络
每一层的输入只为上一层的输出



残差结构
存在**跳层连接**的形式

残差网络

- 残差网络ResNet的背景——网络退化
- 在神经网络可以收敛的前提下，随着网络深度增加，网络的准确性先是逐渐增加至饱和，然后迅速下降
- 这种退化不是由过拟合引起的，即便在模型训练过程中，同样的训练轮次下，退化的网络也比稍浅层的网络的训练错误更高



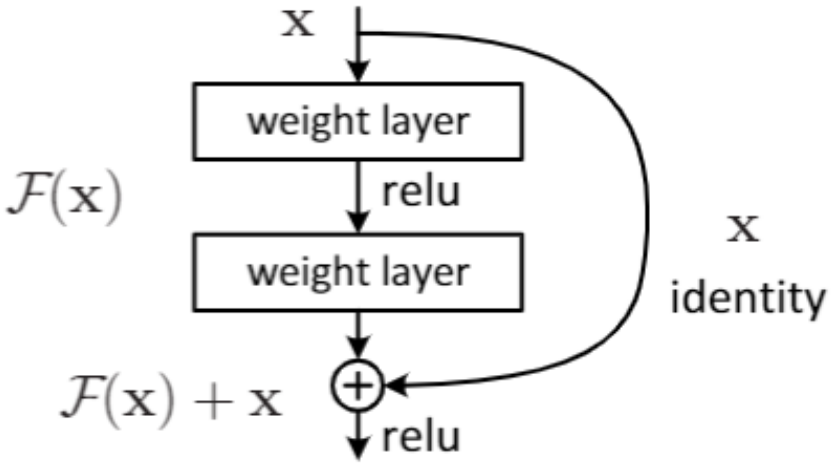
20层和56层的“普通”网络在CIFAR-10数据集上的训练、测试误差

残差网络

- 网络退化不符合常理：
  - 如果存在某个K_1层的网络F_1是当前最优的网络
  - 那么应该可以构造一个更深的K_2层网络F_2
  - 若F_2最后的$K_2 - K_1$层能够做到F_1第K_1层输出的恒等映射，就可以至少取得与F_1一致的结果
- 网络退化说明“普通”的神经网络不具备恒等映射能力
- 为使模型的内部结构至少有恒等映射的能力，解决网络退化的问题，残差网络被提出

残差网络

- 残差结构右侧的线称为**跳转连接**（shortcut connection）：
  - 跳转连接将上一层（或几层）之前的输出与本层计算的输出相加，将求和的结果输入到激活函数中做为本层的输出
- 残差结构的输入：x
- 残差结构期望的输出：$H(x) = F(x) + x$
- 残差结构堆叠的层拟合另一个映射：
  - $F(x) = H(x) - x$

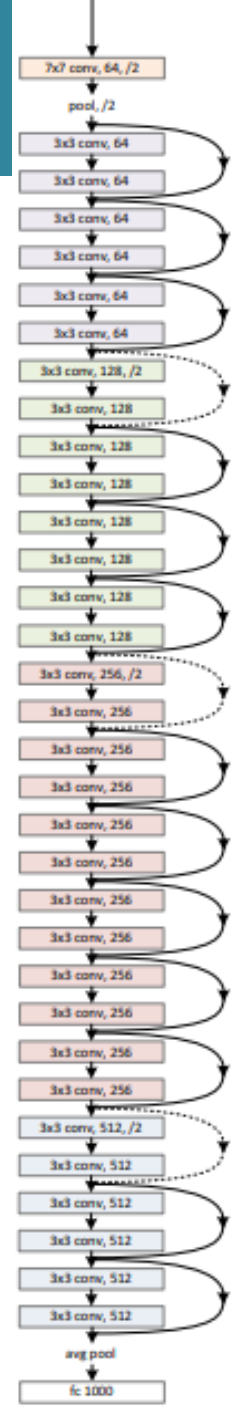


- 残差结构的目的是，随着网络的加深，使$F(x)$**逼近于0**，从而让深度网络的精度在最优浅层网络的基础上不会再下降

残差网络

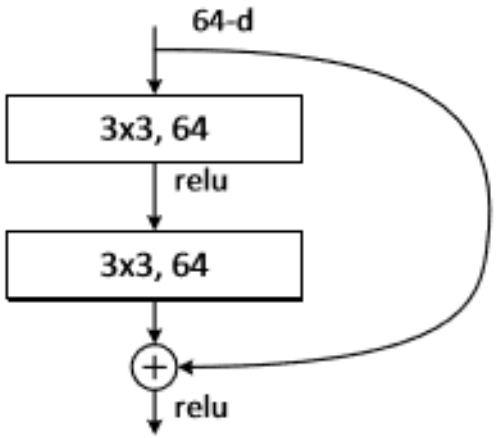
- ResNet结构
  - 一般有5块
  - 右图为34层的ResNet

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

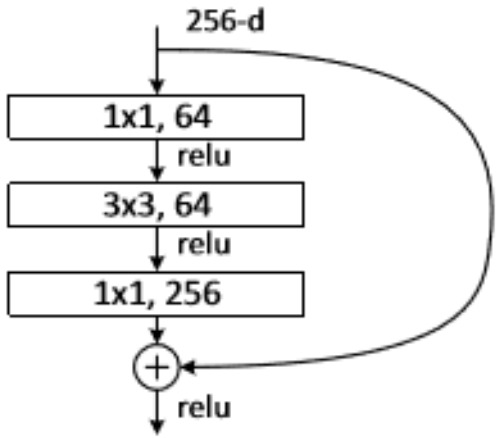


残差网络

- 瓶颈层（ bottleneck layer ）
  - 一般在深度较高的网络ResNet-50/101/152中使用
  - 通过1 × 1卷积进行降维或者升维



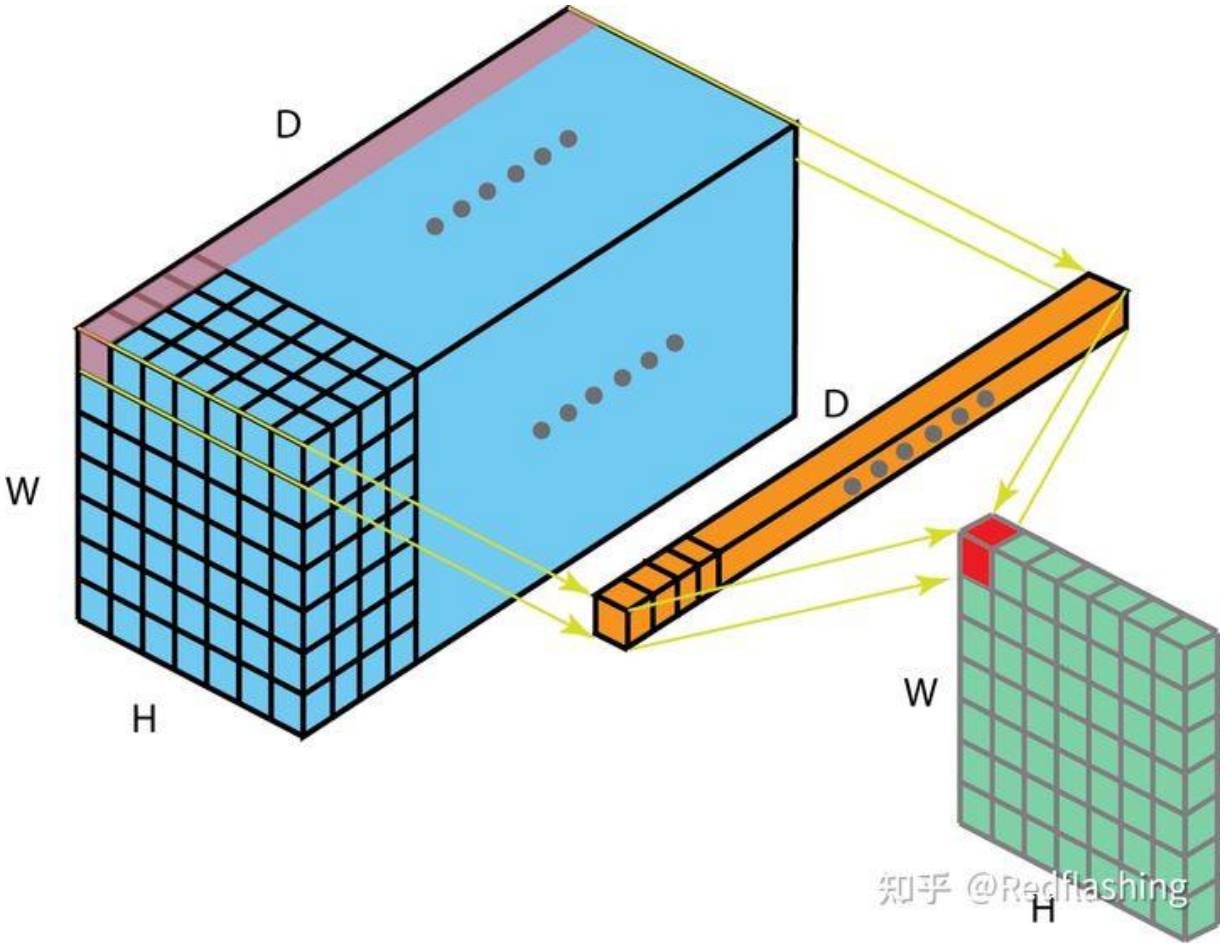
无bottleneck layer



有bottleneck layer

残差网络

- 1×1卷积层的升维、降维原理
  - 对于$H \times W \times D$的输入层
  - 滤波器大小为$1 \times 1 \times D$
  - 输出的尺寸为$H \times W \times 1$
  - n个滤波器组合后
  - 输出层大小为$H \times W \times n$
- 1×1卷积层的作用
  - 调节通道数
  - 实现通道信息的线性组合变化
  - 减少了所用的参数



提高对抗鲁棒性的特征降噪方法

算法原理

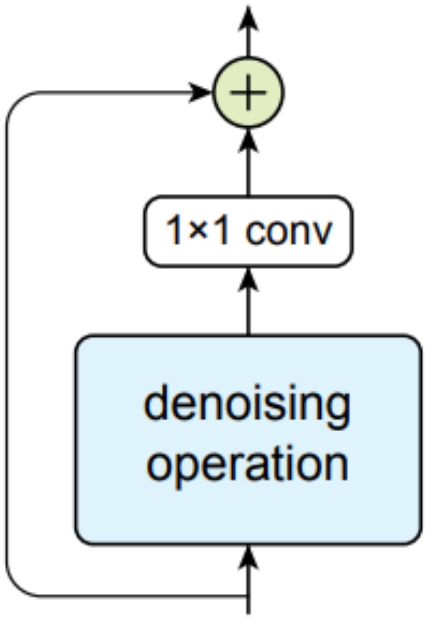
算法原理

T	设计一种具备特征降噪功能的卷积神经网络结构
I	对抗样本
P	1. 在ResNet中加入降噪块 2. 使用对抗样本训练模型 3. 通过训练好的模型预测对抗样本类别
O	对抗样本的预测类别

P	何种网络结构能够实现对抗样本特征去噪
C	需要对设计的网络结构进行端到端的对抗训练
D	降噪块中降噪方法的选择
L	CVPR 2019

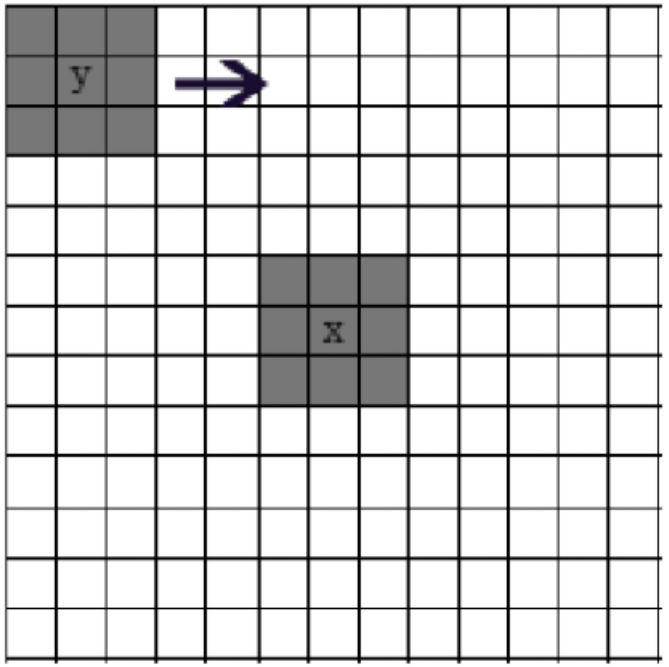
算法原理

- 一个通用降噪模块的输入可以是卷积神经网络中的任何特征层
- 通用降噪模块由三部分组成：
- 降噪操作
  - 非局部均值滤波（Non-local means）
  - 双边滤波（Bilateral filter）
  - 均值滤波（Mean filter）
  - 中值滤波（Median filter）
- 残差连接
  - 帮助网络保留原信号
- 1 × 1卷积
  - 用于特征组合，通过端到端的学习调整消除噪声和保留信号之间的平衡



算法原理

- 非局部均值滤波
- 基本思想是在图像的较大区域中找到几个相似的色块，并用这些相似色块的加权平均值替换中心色块
- 例如以y为中心的邻域窗口在搜索窗口中滑动
- 通过计算x、y两个邻域窗口间的相似程度得到权值



算法原理

- 非局部均值滤波

$$y_i = \frac{1}{C(x)} \sum_{\forall j \in L} f(x_i, x_j) \cdot x_j$$

- x是原特征图，y为降噪后的特征图， L为特征空间坐标点
- $f(x_i, x_j)$为独立于特征的权重函数, $C(x)$为归一化函数
- 采用的两种形式：
  - 高斯（softmax）形式
  - 点积形式

算法原理

- 非局部均值滤波-高斯（softmax）形式

$$f(x_i, x_j) = e^{\frac{1}{\sqrt{d}}\theta(x_i)^T \varphi(x_j)}$$

$$C = \sum_{\forall j \in L} f(x_i, x_j)$$

  - $\theta(x)$和$\varphi(x)$是x的两种嵌入函数， d是通道的数量
  - $\frac{f}{C}$就是softmax函数

- 非局部均值滤波-点积形式

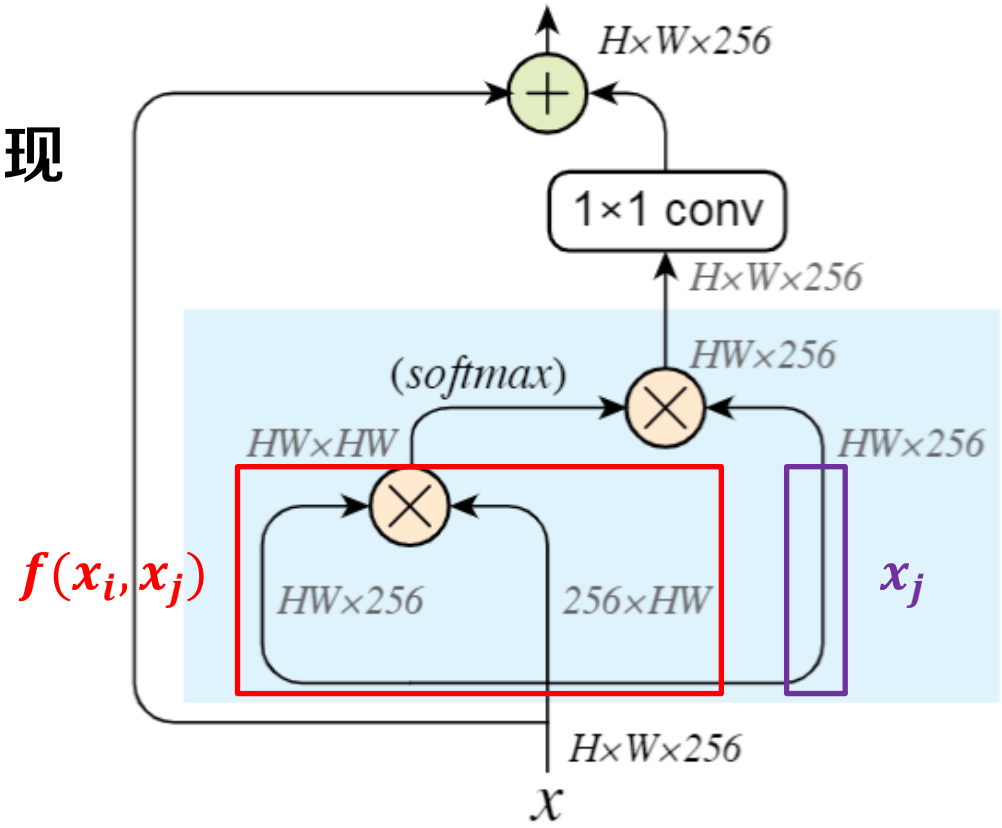
$$f(x_i, x_j) = x_i^T x_j$$

$$C = N$$

  - N即为x中像素的数量

算法原理

- 基于非局部均值滤波的降噪块实现
- 降噪块结构的设计受到Transformer的self-attention机制，以及no-local network的non-local mean操作启发
- 蓝色部分即为$y_i = \frac{1}{C(x)} \sum_{\forall j \in L} f(x_i, x_j) \cdot x_j$的实现
  - 如果使用softmax，则为高斯形式
  - 未使用，则为点积形式



算法原理

- 双边滤波

$$y_i = \frac{1}{C(x)} \sum_{\forall j \in \Omega(i)} f(x_i, x_j) \cdot x_j$$

  - $\Omega(i)$为像素i周围的一个局部区域，如3*3的小补丁
  - 同样采用高斯形式和点积形式
- 均值滤波
- 中值滤波

$$y_i = \text{median}\{\forall j \in \Omega(i), x_j\}$$

  - $\Omega(i)$在每个通道都要分别计算一遍

算法原理

- 对抗样本生成
  - 使用PGD（Projected Gradient Descent）作为白盒攻击者
  - PGD攻击的超参数为：每个像素的最大扰动值16、攻击步长1、攻击迭代次数30
  - 对抗样本可以由干净样本初始化，也可以在允许的最大扰动值范围内随机初始化
  - PGD攻击者随机选择上面两种初始化方式，第一种概率20%，第二种80%
- 对抗样本分布式训练
  - 对于每个mini-batch，使用PGD来生成对抗样本
  - n步PGD后，在对抗样本上执行单步SGD（Stochastic Gradient Descent），并更新模型权重
  - 在128个Nvidia V100 GPU上使用同步SGD执行分布式训练
- 对抗训练是论文模型取得优越效果的主要原因

算法原理

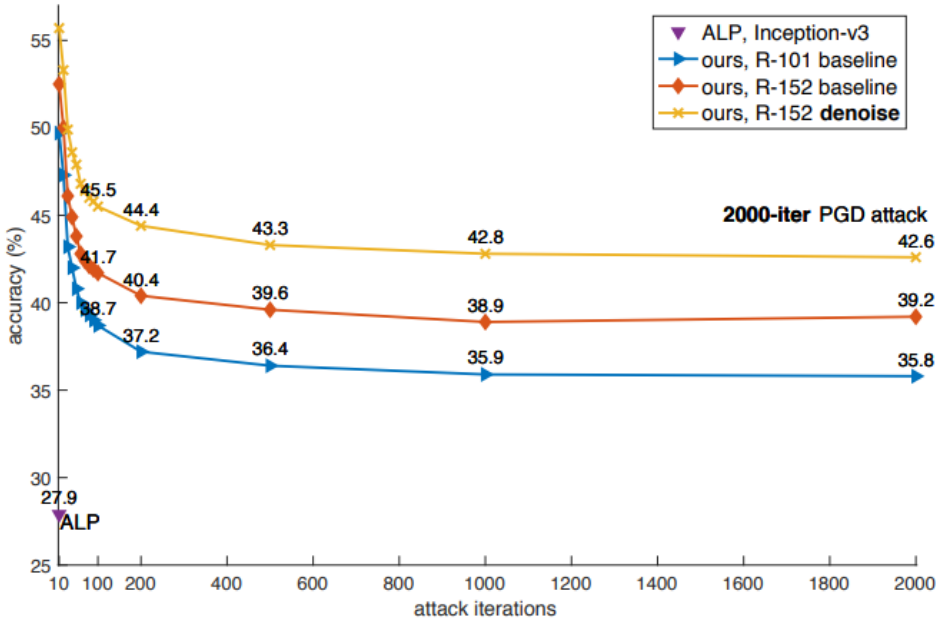
- 白盒实验基本情况
- 数据集
  - ImageNet分类数据集
  - 有1000个类别的128万个样本，选用其中5万张
- 攻击方法
  - 白盒设置下为有目标攻击
  - 目标类别是随机均匀选择的
- 准确率
  - top-1分类准确率，即只有当概率最大的是正确答案才认为正确

算法原理

- 白盒实验基本情况
- baseline模型
  - ResNet-101/152
  - 默认情况下将4个降噪块添加到ResNet中，每个都分别添加在res_2、res_3、 res_4、res_5最后的残差块上
- 对比方法——ALP（ Adversarial logit paring ）
  - ALP是一种对抗训练
  - ALP可以被解释为使用干净样本的logit作为无噪声参考，将对抗样本的logit预测去噪

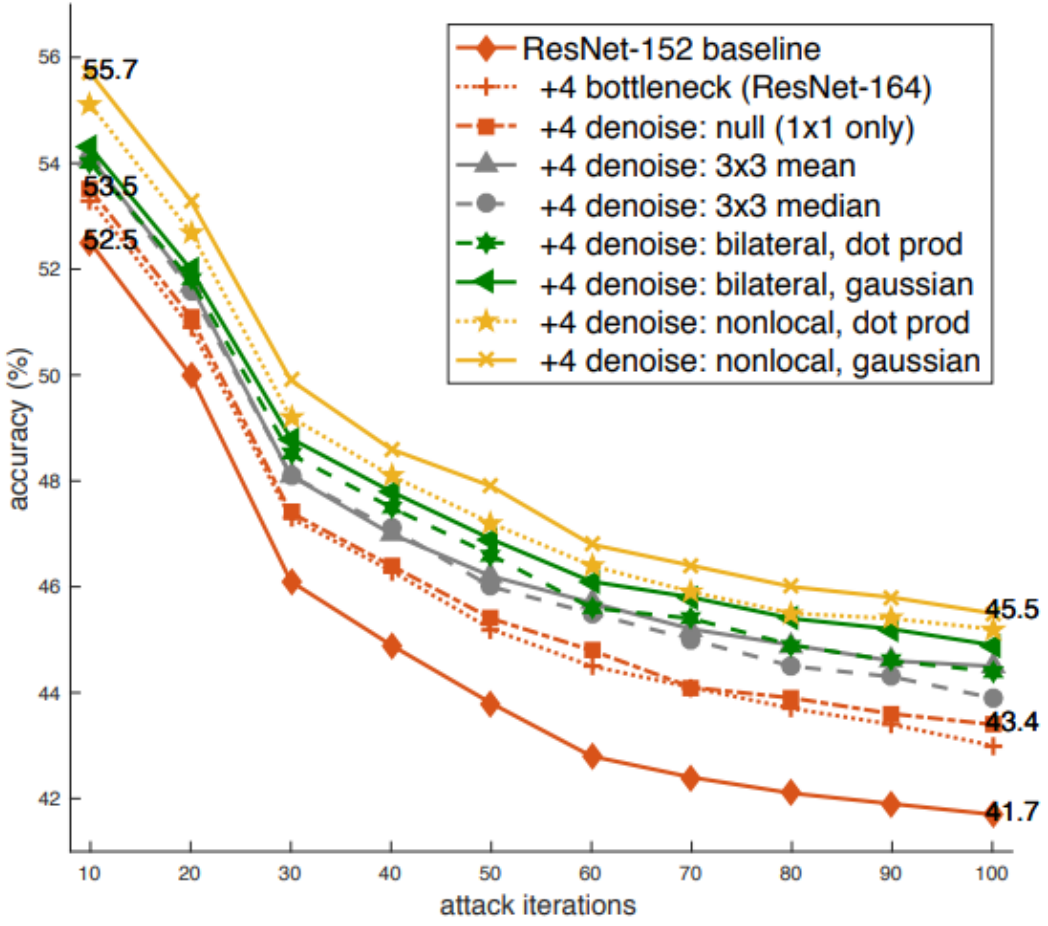
算法原理

- PGD白盒攻击实验结果
  - ALP方法仅针对10次PGD攻击进行了评估
  - R-101、R-152 baseline模型只经过了对抗训练
  - R-152 denoise模型加了4个降噪块，降噪块选用的是实验效果最好的非局部均值滤波的高斯形式
  - 与ALP方法的比较说明论文的对抗训练系统（骨干网络结构、实现方式）是可靠的



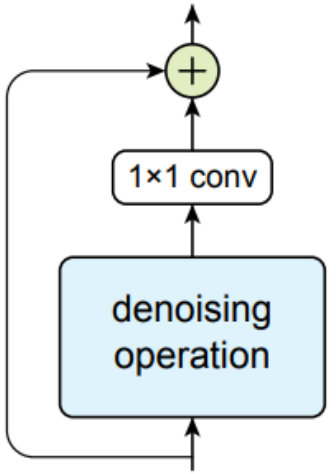
算法原理

- 消融实验——探究不同降噪操作下PGD白盒攻击实验结果
- 其中所有模型都经过了对抗训练
  - 表现最好的降噪操作是高斯形式非局部均值滤波
  - 没有降噪操作的模型效果不如有降噪操作的
  - 添加瓶颈块能够提升效果，但不如降噪块



算法原理

- 消融实验二——探究去掉降噪块的不同部分后PGD白盒攻击实验结果
- 去掉降噪操作
  - 消融实验一已经做过，效果不如加上降噪操作
- 去掉1 × 1卷积
  - 精度在100次PGD攻击下会从45.5%降至36.8%，下降幅度很大
- 去掉残差连接
  - 会使训练变得不稳定，在论文的对抗训练中误差无法减少



- 结果表明降噪操作本身是不够的，由于抑制噪声也可能去除有用的信号，因此有必要将降噪操作与降噪块中的原始输入特征适当地结合起来

attack iterations	10	100
non-local, Gaussian	55.7	45.5
removing 1×1	52.1	36.8
removing residual	NaN	NaN

算法原理

- 黑盒实验基本情况
- 攻击方法
  - NIPS 2017 CAAD（Competition on Adversarial Attacks and Defense）竞赛的五个最佳攻击者
- 准确率
  - 最新的CAAD 2018评估标准“all-or-nothing”
  - 只有当模型能够对某个原始样本的所有攻击者生成的所有版本的对抗样本都分类正确时，该样本才能被认为是正确分类的

算法原理

- 黑盒实验结果
- 达到“all-or-nothing”标准很困难
  - CAAD 2017的获胜者在这个标准下仅有0.04%的准确率
  - 该方法最容易被5种攻击方法的其中2种攻击，去掉这两种后，准确率为13.4%
- 论文中的R-152 baseline方法就能够达到43.1%
- 而在每个残差块后都加一个降噪块能够达到最好的效果49.5%

model	accuracy (%)
CAAD 2017 winner	0.04
CAAD 2017 winner, under 3 attackers	13.4
ours, R-152 baseline	43.1
+4 denoise: null (1×1 only)	44.1
+4 denoise: non-local, dot product	46.2
+4 denoise: non-local, Gaussian	**46.4**
+all denoise: non-local, Gaussian	**49.5**

算法原理

- 探究降噪块在非对抗环境中的作用
- 由于降噪块是卷积神经网络的组成部分，因此可以用于干净样本的分类任务
  - 对R-152模型进行了三次独立的训练，以展示同一架构下的自然随机变化
  - 加入降噪块后，模型的准确率没有明显的提升
- 结果表明降噪块只有在需要对抗鲁棒性的设置中具有特殊的优势
  - 因为降噪块的设计是为了减少特征噪声，只有在用对抗样本时才会出现特征噪声

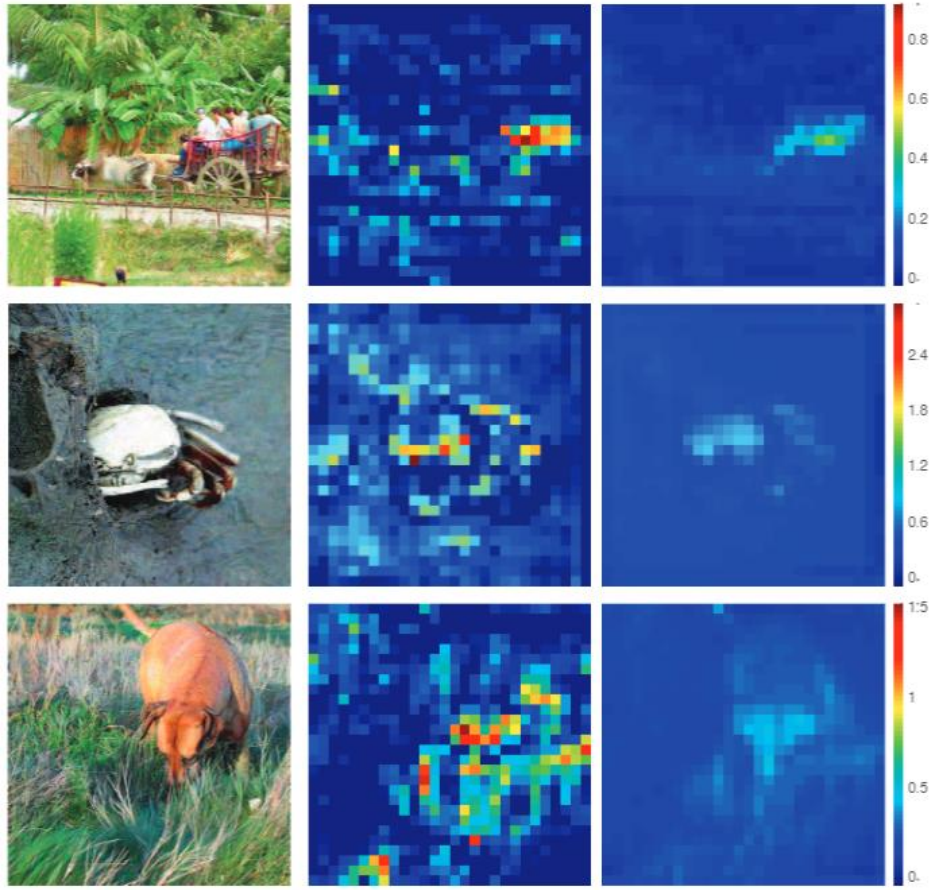
model	accuracy (%)
R-152 baseline	78.91
R-152 baseline, run 2	+0.05
R-152 baseline, run 3	-0.04
+4 bottleneck (R-164)	+0.13
+4 denoise: null (1×1 only)	+0.15
+4 denoise: 3×3 mean filter	+0.01
+4 denoise: 3×3 median filter	-0.12
+4 denoise: bilateral, Gaussian	+0.15
+4 denoise: non-local, Gaussian	+0.17

算法原理

- 对抗训练和干净训练的tradeoff
- ResNet-152 baseline模型效果对比
  - 干净样本训练的模型在干净样本上准确率78.91%
  - 对抗样本训练的模型在干净样本上准确率62.32%
- 加入降噪块（非局部均值滤波的高斯形式）后模型效果对比
  - 干净样本训练的模型在干净样本上准确率79.08%
  - 对抗样本训练的模型在干净样本上准确率65.30%
- 经过对抗训练，模型在干净样本上的分类准确率降低
- 需要平衡模型在对抗样本上和干净样本上的准确率

算法原理

- 对抗图像及其特征图在降噪操作之前(左)和之后(右)
- 结果表明
  - 特征降噪操作可以成功地抑制特征图中的大部分噪声
  - 并使响应集中在视觉上有意义的内容上



提高对抗鲁棒性的特征降噪方法

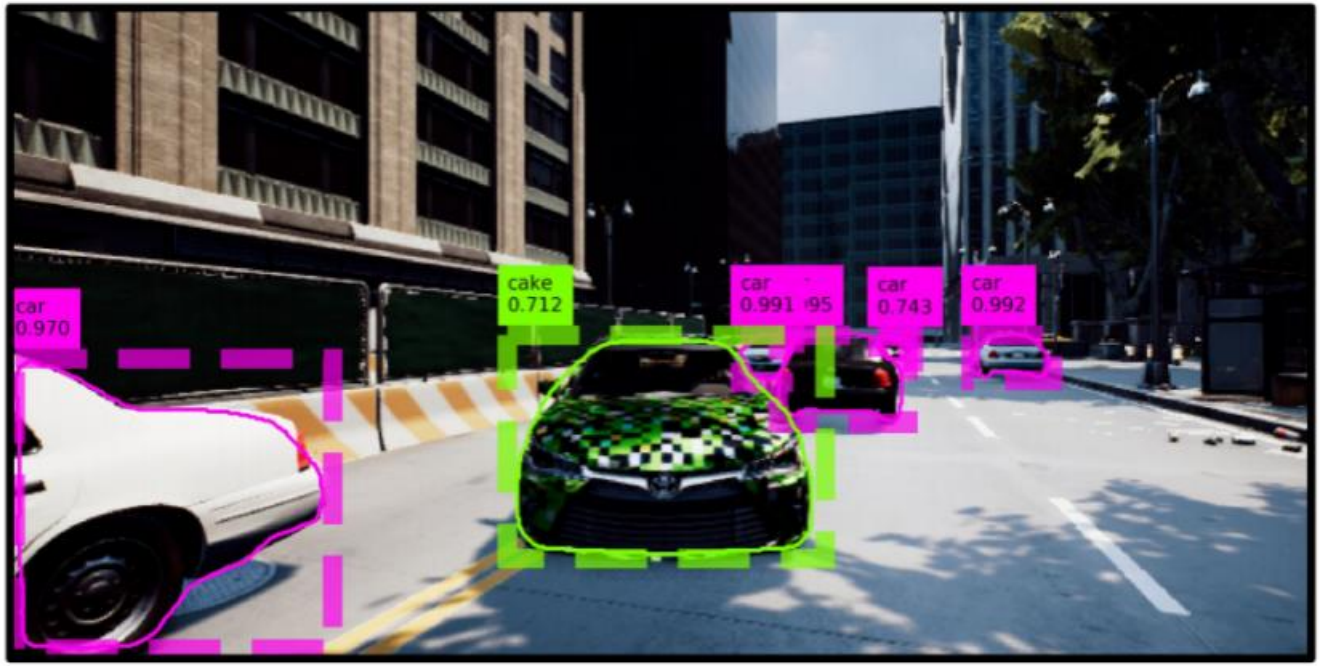
优劣分析

优劣分析

- 横向对比
  - 优势
    - 通过对抗训练和特征降噪，提高了卷积神经网络的鲁棒性
  - 劣势
    - 相比于对抗训练，论文中主要提出的特征降噪方法效果不明显
- 纵向对比
  - 前人工作主要是对输入样本进行特征降噪，再输入到网络中训练
  - 论文则是直接设计了一种具有降噪模块的神经网络结构
  - 启发后人设计具有“先天”对抗鲁棒性的卷积网络架构

应用总结

- 对抗样本应用领域
  - 自动驾驶
  - 人脸识别
  - 恶意软件检测



有迷彩图案的车被认为是蛋糕

在某些角度下会被识别为限速45

参考文献

[1] Resnet到底在解决一个什么问题呢?
https://www.zhihu.com/question/64494691/answer/786270699
[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[3] Xie C, Wu Y, Maaten L, et al. Feature denoising for improving adversarial robustness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 501-509.
[4] 论文解读: | (CVPR2019)《Feature Denoising for Improving Adversarial Robustness》
https://blog.csdn.net/qq_40994260/article/details/106755050

知人者智，自知者明。
胜人者有力，自胜者
强。知足者富。强行
者有志。不失其所者
久。死而不亡者，寿。

谢谢！