实验6 Adversarial Examples for Classification

南京信息工程大学
计算机学院

应龙

2024年 秋季

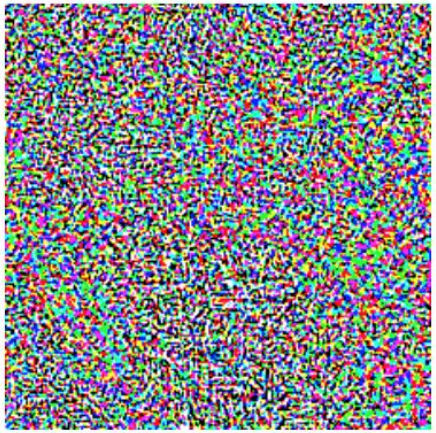
实验背景

对抗样本（Adversarial Examples）有限训练集上训练得到的机器学习模型，在应用场景中会遇到某些细微干扰所形成的输入样本，人类难以通过感官辨识，但是模型以高置信度做出错误的输出决策。 Alignment Problem

$\boldsymbol{x}$

“panda”
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“nematode”
8.2% confidence

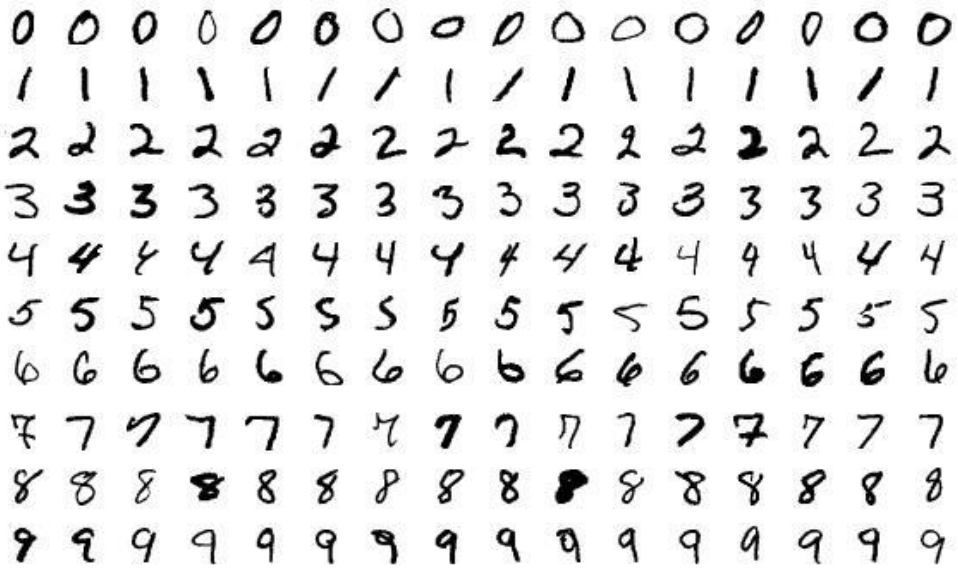
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

“gibbon”
99.3 % confidence

MNIST 数据集

MNIST 数据集来自美国国家标准与技术研究所, National Institute of Standards and Technology (NIST)。包含 0 到 9 共 10 个类别的灰度图像。该数据集由 Yann LeCun 等人整理，是深度学习和机器学习领域中最广泛使用的数据集之一。其中包含 70,000 张 28x28 像素的灰度图像，其中 60,000 张用于训练，10,000 张用于测试。每张图像仅包含一个手写数字，背景为黑色，数字为白色或灰色。由于其简单的特性，MNIST 数据集常被用于初步测试新模型的性能以及验证模型的基本分类能力。其优势在于其易于使用和高度标准化，适合作为分类网络的基准测试。同时，由于其数据规模适中，训练速度快，因此非常适合用于对抗样本生成和模型鲁棒性实验。



MNIST 数据集

MNIST 数据集可在
http://yann.lecun.com/exdb/mnist/
获取, 它包含四个部分:
Training set images: train-images-idx3-ubyte.gz (9.9 MB, 解压后 47 MB, 包含 60,000 个样本)
Training set labels: train-labels-idx1-ubyte.gz (29 KB, 解压后 60 KB, 包含 60,000 个标签)
Test set images: t10k-images-idx3-ubyte.gz (1.6 MB, 解压后 7.8 MB, 包含 10,000 个样本)
Test set labels: t10k-labels-idx1-ubyte.gz (5KB, 解压后 10 KB, 包含 10,000 个标签)

对抗样本 (adversarial samples)

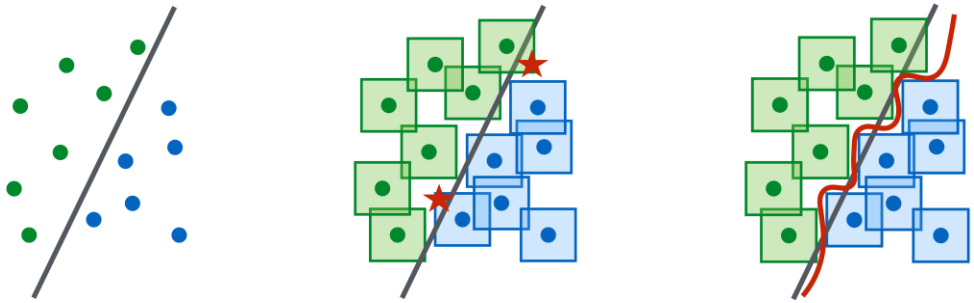
- FGSM攻击算法

设 x 是原始样本，x' 是对抗样本，其中： $x' = x + \eta$，为了让对抗样本不被机器所识别，η 应该足够小，可以使用无穷阶范数来表述 η 足够小这一限制: $\|\eta\|_{\text{inf}} < \epsilon$，即 l_{inf}-ball。

FGSM 产生对抗样本的方式为：

$$\eta = \epsilon sign(\nabla_x J(x, y; \theta))$$

其中，是分类损失函数，通过梯度上升，最大化损失函数，企图使得 x 不属于 y 类。



Reference:

✓ Goodfellow, Ian J., Jonathon Shlens and Christian Szegedy. “Explaining and Harnessing Adversarial Examples.” CoRR abs/1412.6572 (2014): n. pag.

对抗样本 (adversarial samples)

- PGD攻击算法

对于每个数据点 ，引入一组允许的扰动 $S \subseteq \mathbb{R}^d$，以正则化算法的操纵力。

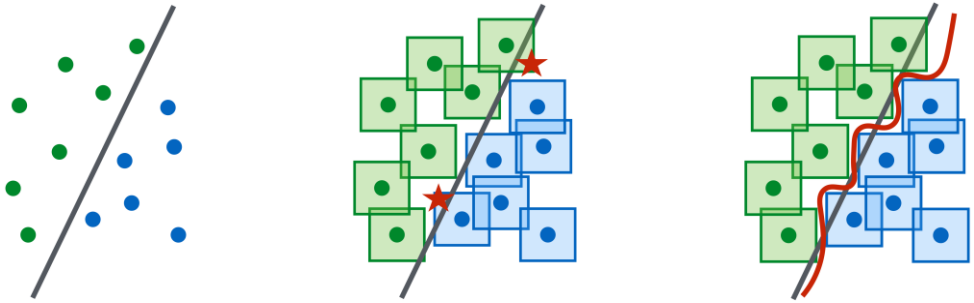
$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in S} L(x+\delta, y; \theta) \right]$$ 修改结构风险的定义

内部最大化问题和外部最小化问题的组合，saddle point problem。

S 可以采用 l_{inf}-ball，即无穷范数不能超过球面范围。

$$x^{t+1} = \Pi_{x+S}(x^t + \epsilon sign(\nabla_x L(x, y; \theta)))$$

Π_{x+S}符号的意思是，先计算原图像的损失梯度得到对抗样本，对抗样本减去原图像得到扰动值，并通过 clamp 限制在球面范围内，原图像加上扰动值就是最终的对抗样本。



针对PGD攻击的鲁棒性会产生针对所有一阶攻击算法的鲁棒性，即仅依赖一阶信息的攻击。只要对手仅使用损失函数相对于输入的梯度，我们就可以推测，它不会找到比PGD更好的局部最大值。

Non-Targeted and Targeted Attacks

Non-targeted adversarial attacks are adversarial attacks where the goal is simply to make a machine learning model misclassify the input, without requiring the misclassification to be into a specific target class.

Targeted adversarial attacks in machine learning are adversarial attacks where the attacker crafts an input with the specific goal of causing a machine learning model to misclassify the input into a particular, pre-defined target class.

Aspect	Non-targeted Attack	Targeted Attack
Objective	Cause misclassification into any incorrect class.	Cause misclassification into a specific target class.
Difficulty	Easier to perform.	Harder due to controlled misclassification.
Perturbation Size	Smaller perturbations often suffice.	Requires larger perturbations for precise control.
Applications	General robustness testing.	Security-critical exploitation or targeted testing.
Success Rate	Higher due to relaxed constraints.	Lower because of the stricter objective.
	Maximize loss with respect to the true label.	Minimize loss with respect to the target label.

实验内容

实验内容：

- 使用对抗样本生成方法FGSM 进行 Non-Targeted Attack, 生成一批对抗样本；
- 在已经训练好的基线模型上测试对抗样本，记录模型的性能下降情况；观察选取不同扰动大小（如FGSM方法中的epsilon）对模型性能的影响；使用可视化工具（如matplotlib）绘制扰动大小-精准度曲线；
- 回答问题： FGSM 方法产生对抗样本时为什么要在梯度加上符号函数而不直接使用梯度？
- 进行 Targeted Attack：对于手写数字9图像施加扰动，使得分类网络错误地将它识别为数字3，扰动图像数字9可以直接通过对手写体数字9添加较小的噪声生成，也可以直接从随机噪声中生成一个类似于9的图像；
- 实现一种更复杂的对抗样本生成方法（如PGD, C&W, DeepFool, GAN等方法）生成对抗样本。

实验内容和要求

Link1(FGSM方法参考): https://github.com/ymerkli/fgsm-attack

Link2(PGD方法参考): https://github.com/angelognazzo/Adversarial-Attacks-FGSM-PGD ; https://github.com/yaodongyu/TRADES/blob/master/pgd_attack_mnist.py

Link3(C&W方法参考): https://github.com/Carco-git/CW_Attack_on_MNIST

Link4(MNIST-WIKI): https://en.wikipedia.org/wiki/MNIST_database

Link5(Pytorch对抗样本生成实战): https://pytorch.org/tutorials/beginner/fgsm_tutorial.html

Link6(分类网络选择): https://github.com/weiaicunzai/pytorch-cifar100/tree/master/models

Tips:

1.下载CIFAR-10数据集可以直接使用torchvision.datasets.MNIST()函数完成，请注意root路径的位置；

2.可选拓展内容中的GAN方法对于计算时间和计算设备性能有一定要求，更合适计算设备有GPU的同学进行尝试；

实验报告要求

1.撰写实验报告，为电子版word或pdf格式文件。也可以将内容手写后拍照粘贴到word格式文件中，但照片和手写内容一定要清晰；
2.对于FGSM方法的公式和实现代码进行介绍和分析；
3.对于基线实验，列出训练的相关信息，以及在测试集上的相关性能评估；可视化结果请截图保存放在文档中；
4.从测试结果评估指标的数值和可视化两方面展示对抗样本对于分类网络的影响；
5.对完整的实验及其结果做一定的总结分析。

最后上交：文件夹以“姓名_学号_人工智能导论实验5”命名并压缩，包括：①实验报告，②源代码文件夹（不包括训练数据集和结果图片），③实验结果文本文件