ISCAS 中国科学院软件研究所学术年会'2022
暨计算机科学国家重点实验室开放周
学术论文

使用测试用例排序技术增强DNN鲁棒性

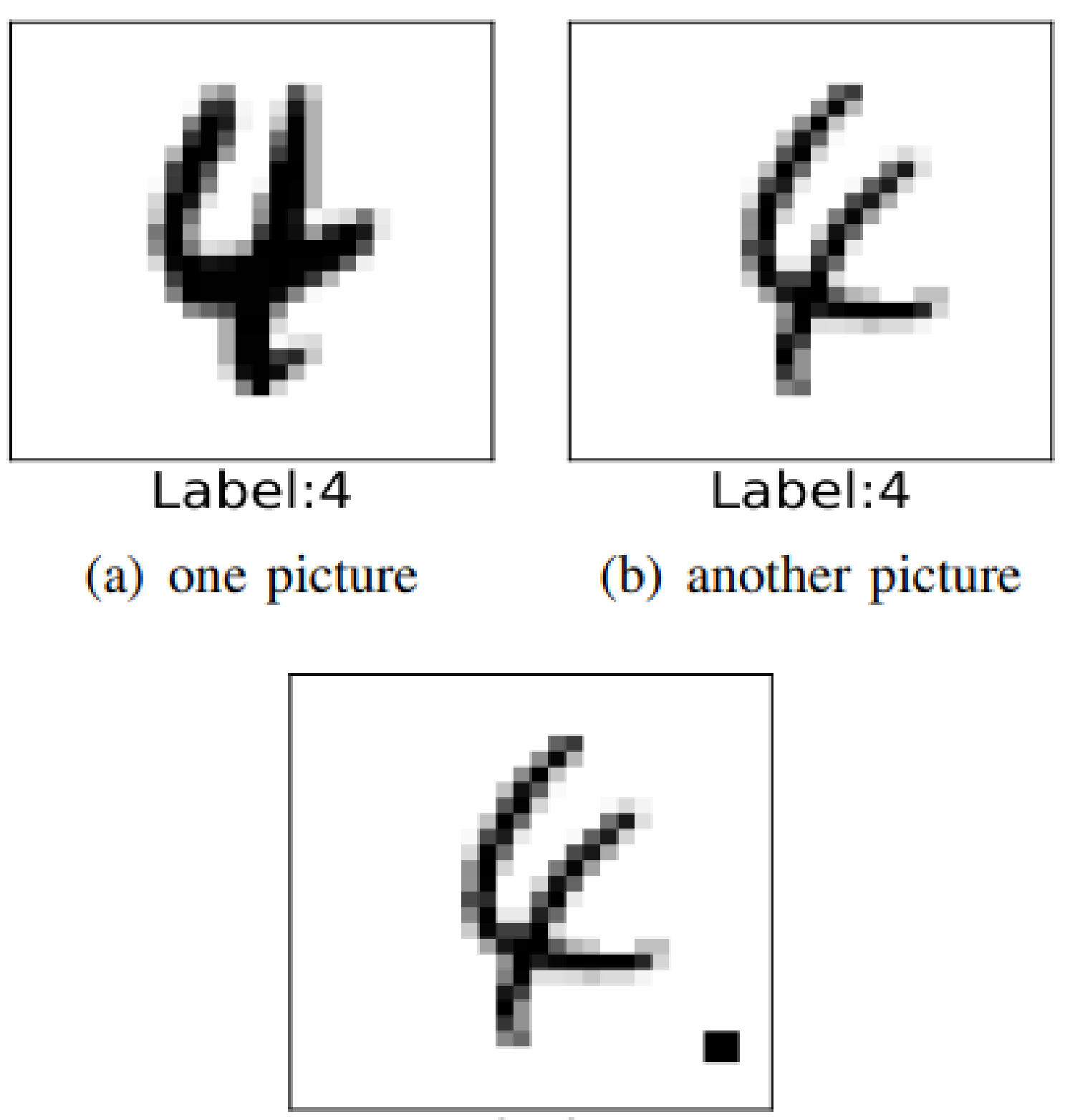
崔炳轶 张龙 张震宇

联系方式：崔炳轶 cuiby@ios.ac.cn 13126897286

背景简介

- 深度神经网络（deep neural networks， DNN）在多个领域广泛的应用，它们给便捷了人类生活，推动了社会发展。但是，同时需要充分关注其质量问题。而传统软件和DNN的系统存在很多本质区别。前者是程序语句逻辑，但后者包含神经元值、连接权重和激活函数等。需要设计全新的技术。因此，需要设计适用于DNN系统测试的全新技术。

- 评价DNN系统质量最直观的办法是生成对抗攻击样本，目前多种对抗方法能生成不同类型、对抗性强的样本。这些样本能够攻击DNN系统不同的弱点，因此衡量样本空间中有效对抗样本密度是十分有必要的。此外还结合测试排序技术的方法，选取优先级较高的测试输入重新训练，可以改善DNN系统的质量。需使用额外的测试用例集用于增强训练，计算资源要求高。

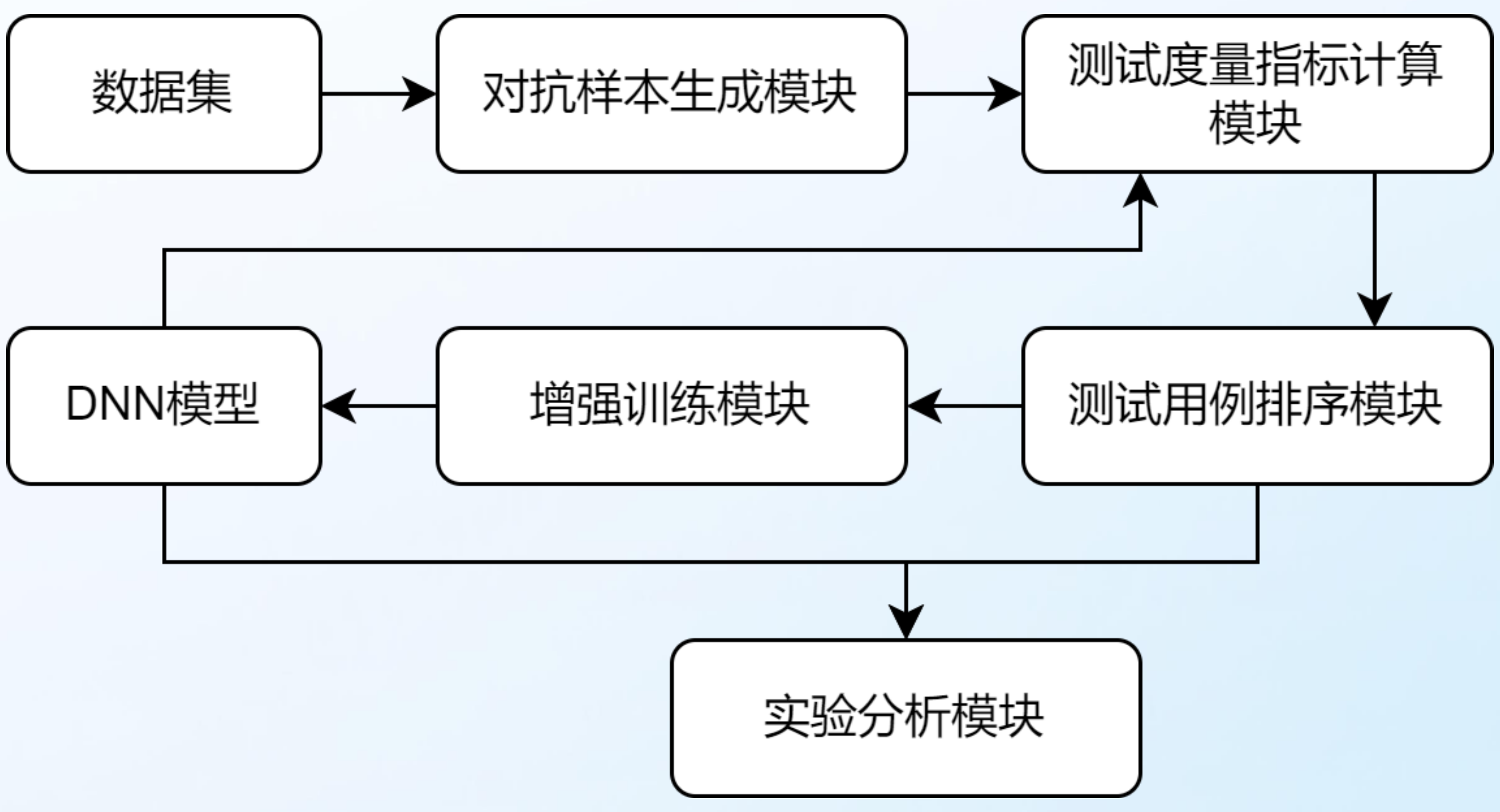


(a) one picture (b) another picture
(c) an adversary generated from (b)

- 两种样本和对抗样本举例

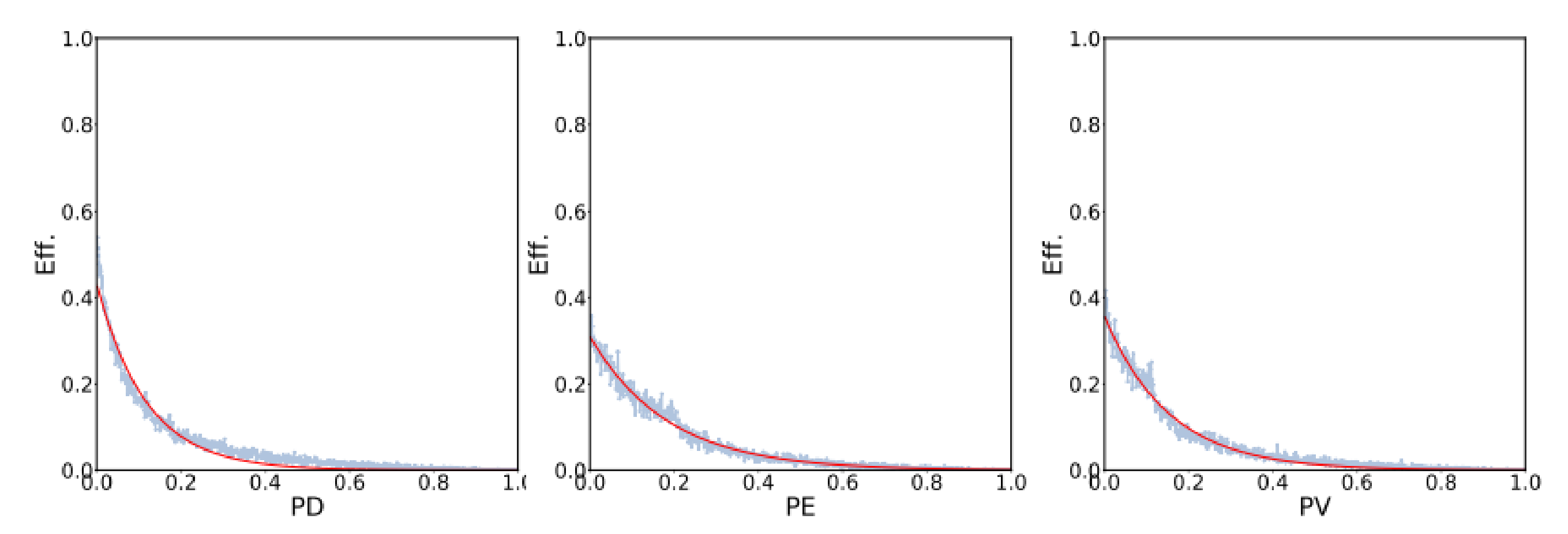
模型流程

- 整体算法流程可分为两个阶段。第一阶段，收集评估阶段。首先，使用多种对抗攻击方法，生成多组对抗样本；其次，评价样本的有效对抗样本密度，并结合统计置信度，计算度量指标。
- 第二阶段，排序增强阶段。按照优先级排序，在选择比例内选取优先级高的对抗样本，然后对原模型进行增强训练。


数据集
对抗样本生成模块
测试度量指标计算模块
DNN模型
增强训练模块
测试用例排序模块
实验分析模块

实验结果

- 使用三种统计置信度指标，对测试样本排序，结果对抗敏感的样本更能暴露DNN系统的更多缺陷。
- 结合有效对抗样本密度指标，使用部分高优先级的训练集对抗样本，用于模型的增强训练，可以有效提高模型效果。



	PD	PE	PV
MNIST	$f(x) = 0.27e^{-3.87x}$	$f(x) = 0.27e^{-3.81x}$	$f(x) = 0.27e^{-3.86x}$
f-MNIST	$f(x) = 0.39e^{-3.18x}$	$f(x) = 0.38e^{-3.11x}$	$f(x) = 0.39e^{-3.16x}$
Cifar-10	$f(x) = 0.14e^{-3.87x}$	$f(x) = 0.14e^{-4.00x}$	$f(x) = 0.14e^{-3.95x}$
Cifar-100	$f(x) = 0.23e^{-1.56x}$	$f(x) = 0.24e^{-1.64x}$	$f(x) = 0.24e^{-1.61x}$

- 距离和有效率的拟合曲线