CCF-CV走进江西财经大学

良性对抗样本攻击研究

桑基韬

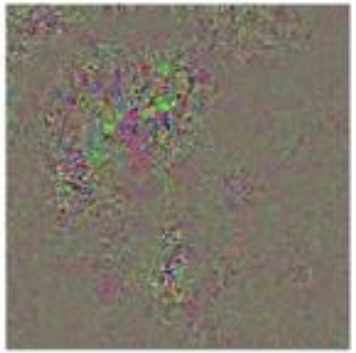
2020年9月26日

对抗样本的“恶”

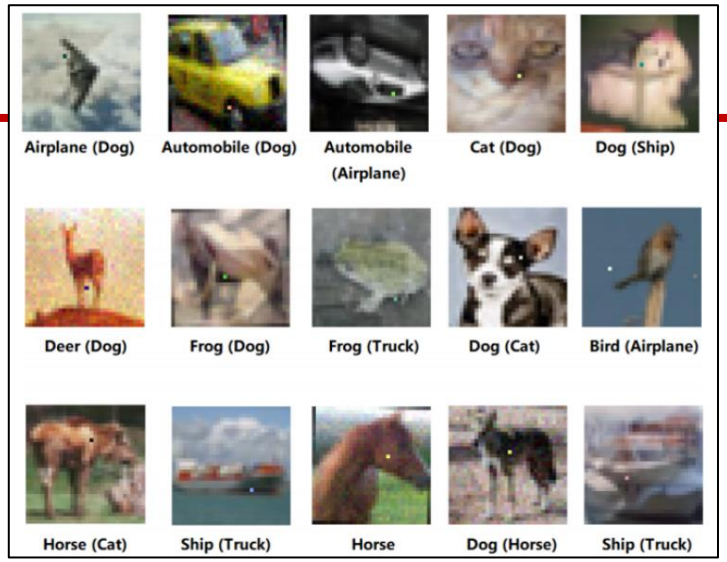
“African elephant”

“koala”

difference



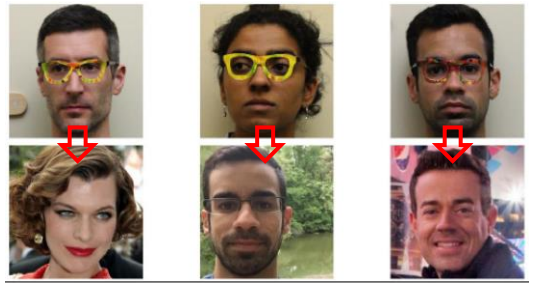
10x difference



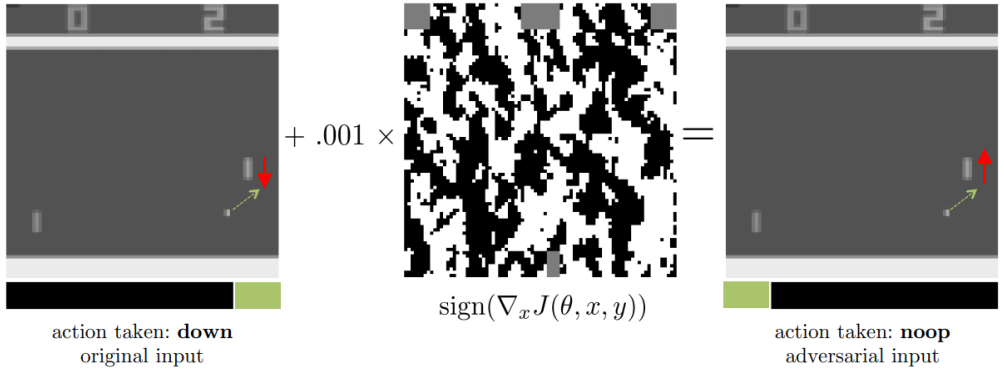
单像素攻击

路标识别

行人检测



人脸识别



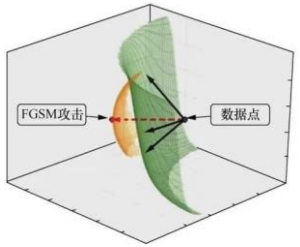
对抗攻击决策[1]

正常分类　系统越权　系统崩溃

对抗攻击系统漏洞

[1] Adversarial Attacks on Neural Network Policies. ICLR 2017.

对抗样本研究发展



*Goodfellow*提出FGSM：梯度上升快速生成对抗样本[3]

对抗训练：对抗样本加入训练集，提升对抗鲁棒性[3]

迭代攻击BIM：攻击更强的对抗样本[4]

混淆梯度[5]：使分类器的梯度信息无效

*Madry*提出PGD[6]：min-max鲁棒优化问题

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|x'-x\|\leq\epsilon} \mathcal{L}(f_\theta(x'), y) \right]$$

混淆梯度无效(*Carlini*,[7])：攻克7篇混淆梯度方法，只有PGD无法击破

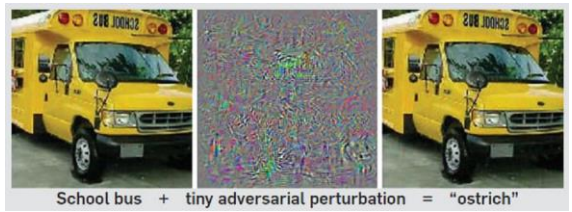
类PGD[8]：结合自适应攻击测试模型鲁棒性

防御不通用(*Carlini*,[9])：低于所宣称精度，无法防御所有可能攻击

2003 欺骗算法（垃圾邮件检测器[1]）

2014 *Szegedy*首次提出对抗样本问题[2]：

- 面向深度学习模型
- 强调人类察觉不到扰动



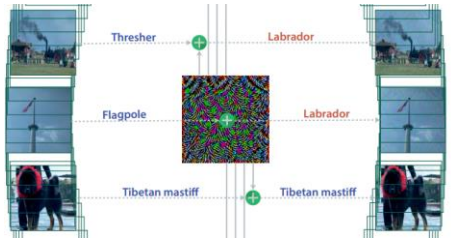
2015 2015

2017 2017

对抗噪声实体化：在各个视角欺骗神经网络的现实世界3D物体[10]

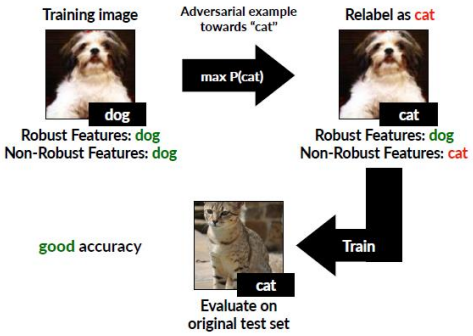
2018 2018

通用对抗噪声：同时攻击不同输入的噪声[11]



2019

对抗噪声是算法的特征(*Madry*,[12])：对抗样本训练的分类器可以泛化到攻击类测试样本



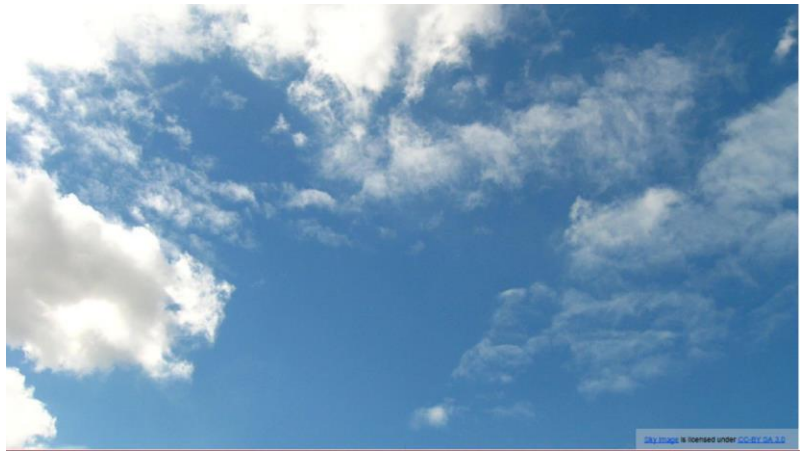
2020 2020

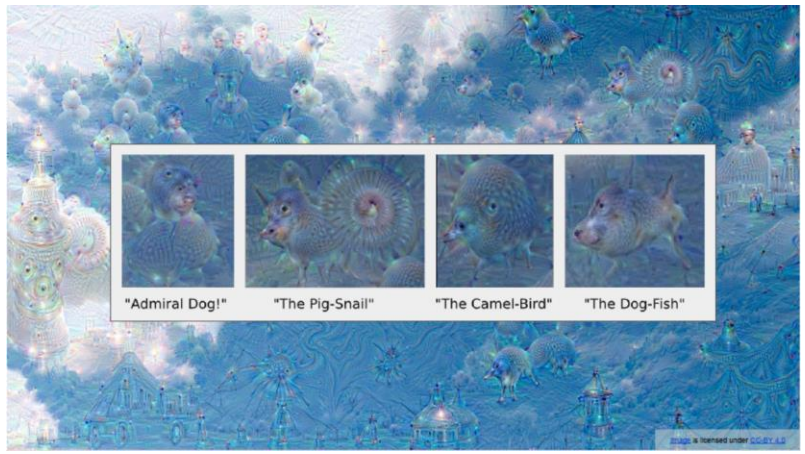
[1] "In vivo" spam filtering: a challenge problem for KDD, KDD 2003.
[2] Intriguing properties of neural networks. ICLR 2014
[3] Explaining and harnessing adversarial examples, ICLR 2015
[4] Adversarial examples in the physical world, ICLR 2017
[5] Stochastic activation pruning for robust adversarial defense, ICLR 2018.
[6] Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018.
[7] Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, ICML 2018 Best Paper
[8] Resisting adversarial attacks by k-winners-take-all, ICLR 2020
[9] On Adaptive Attacks to Adversarial Example Defenses, arXiv, 2020
[10] Adversarial examples in the physical world. ICLR 2017.
[11] Universal adversarial perturbations. CVPR 2017.
[12] Adversarial Examples Are Not Bugs, They Are Features, NeurIPS 2019

人 vs 算法

- 算法是对人的知识蒸馏：人标注样本+算法从样本中学习

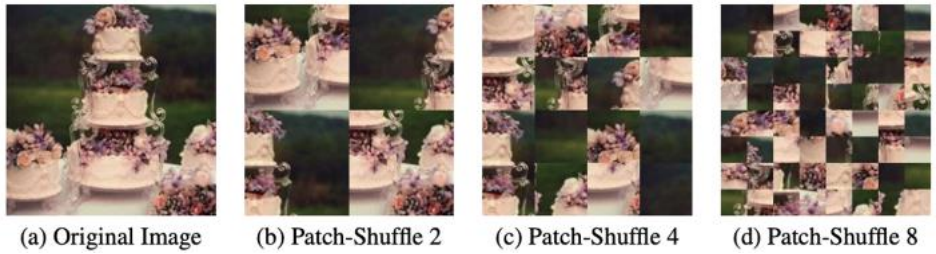
你”看”到了什么?





DeepDream

算法: 纹理 vs 人:形状

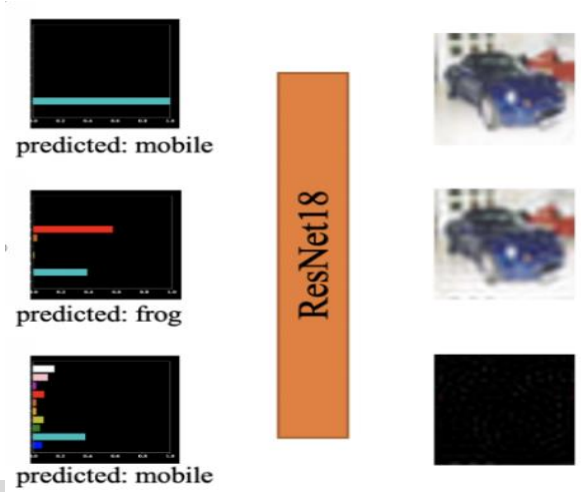

(a) Original Image (b) Patch-Shuffle 2 (c) Patch-Shuffle 4 (d) Patch-Shuffle 8

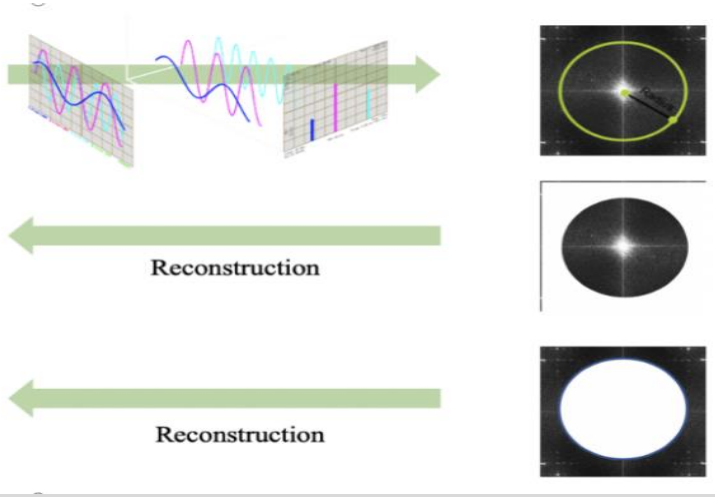

(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan


(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat


(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

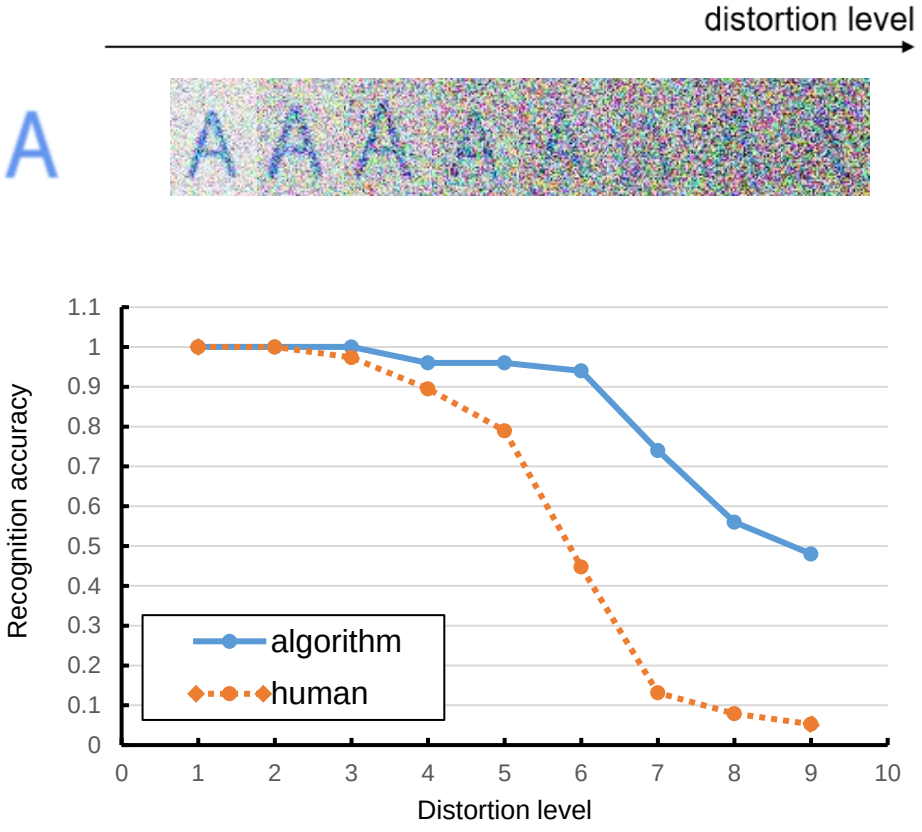
算法: 高频 vs 人:低频



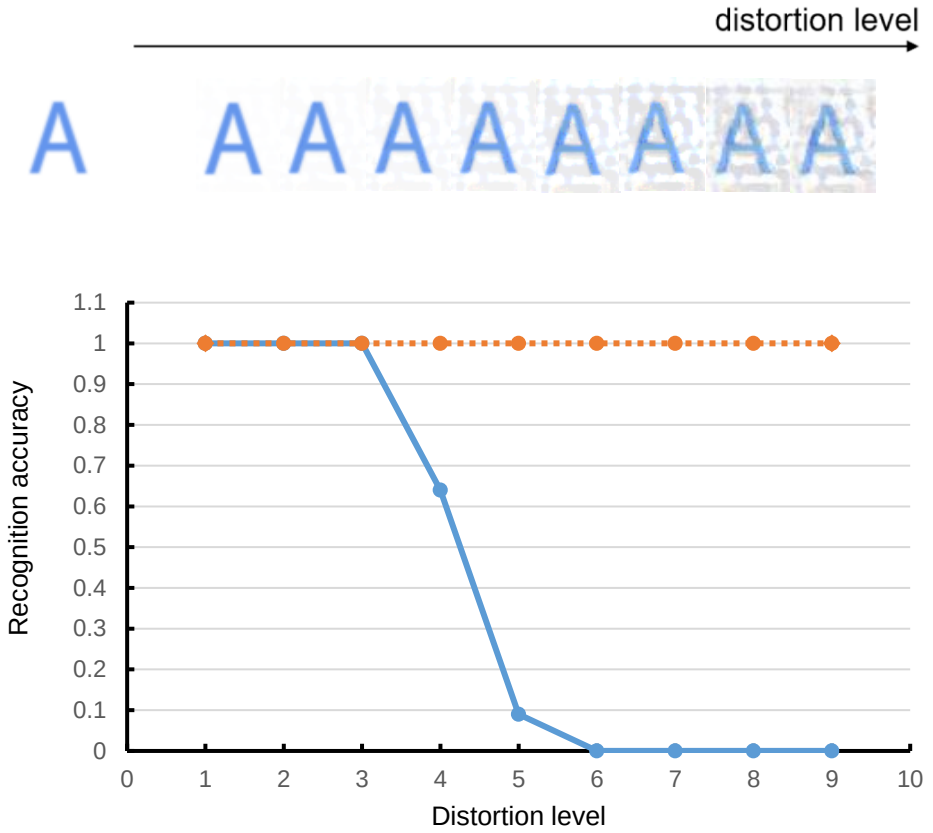


人 vs 算法：对抗样本

- 对抗样本利用了人和算法的不同信息处理机制 → 体现为对于不同扰动的抗干扰能力



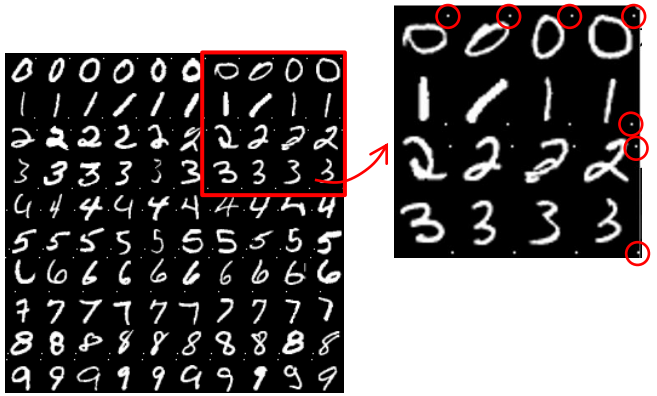
人和模型对**高斯白噪声**的抗干扰能力



人和模型对**对抗噪声**的抗干扰能力

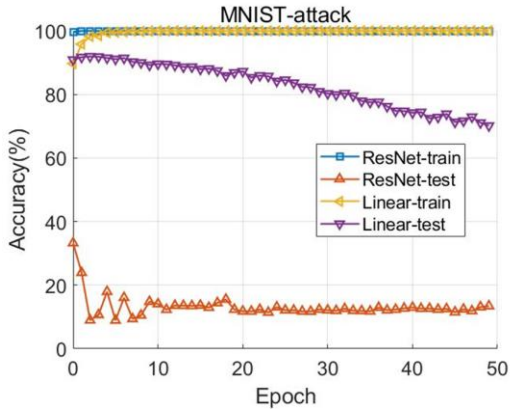
人 vs 算法：对抗样本

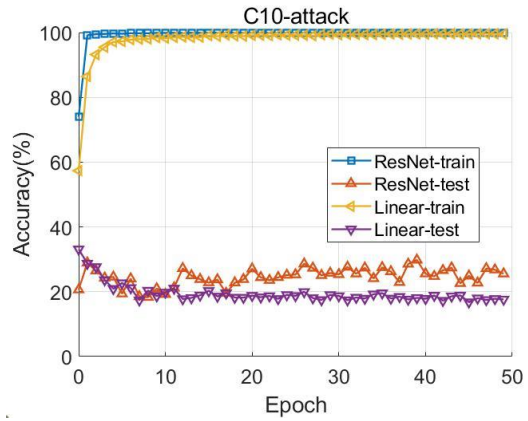
- **算法从训练数据中学习强关联模式**
  - 这些模式有时不是基于语义的 → 人不易察觉
  - 这些模式有时来自对数据集的过拟合 → 算法过于敏感



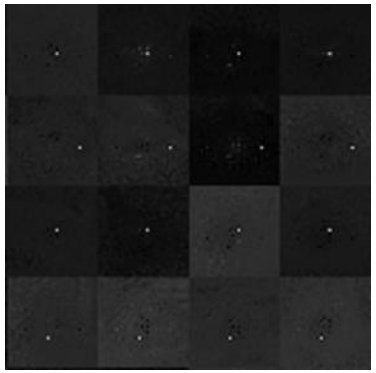
MNIST训练集污染

Cifar10训练集污染

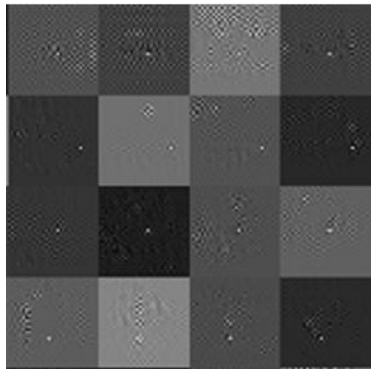




Linear-ResNet训练/测试误差曲线



Cifar10-Linear



Cifar10-ResNet

IntegratedGradient 特征归因图

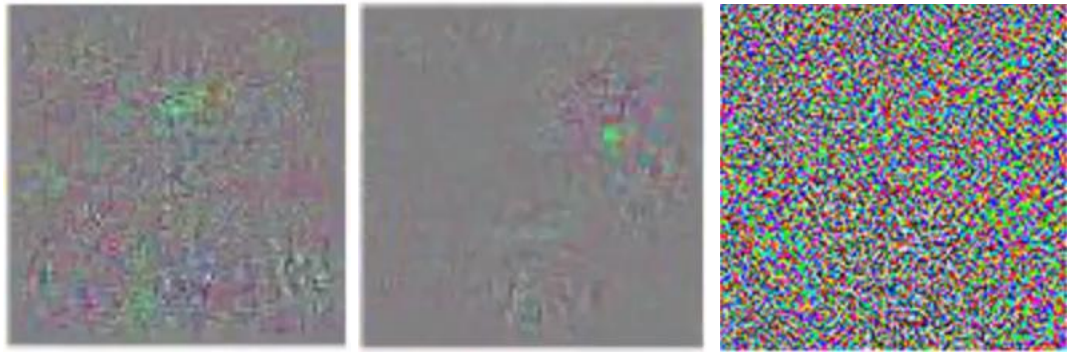
- **对抗训练集攻击**：添加噪声污染训练集，最小化训练误差前提下提高测试误差（泛化gap增加）

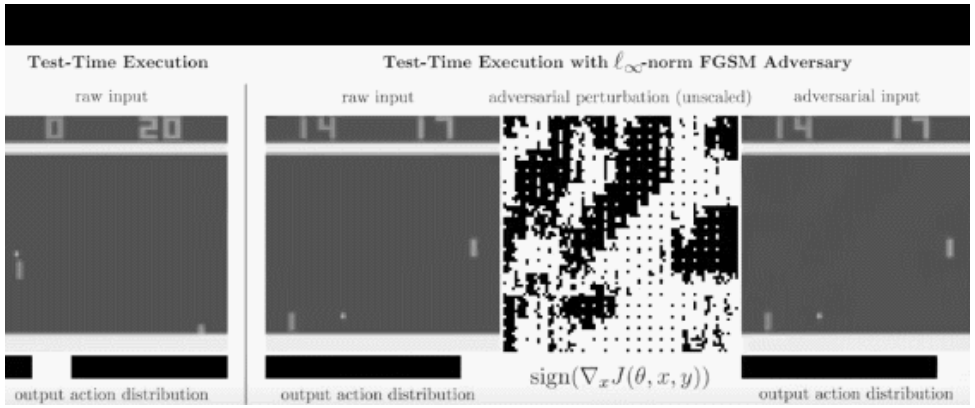
$$\max_{r_i} L_{D_{X,Y}}\left(\operatorname{argmin}_{h} \frac{1}{n}\sum_{i=1}^{n} loss(h(x_i + r_i), y_i)\right)$$
$$s.t. \|r_i\| < \varepsilon$$

  - 与对抗样本攻击的差异
  - ✓ 对抗样本攻击改变测试集，对抗数据集攻击改变训练集
  - ✓ 对抗样本面向训练好的特定模型(内层h固定)，对抗训练集攻击在假设空间中搜索
  - ✓ 对抗样本面向单一样本(外层一个测试样本)，对抗训练集攻击最大化测试集损失

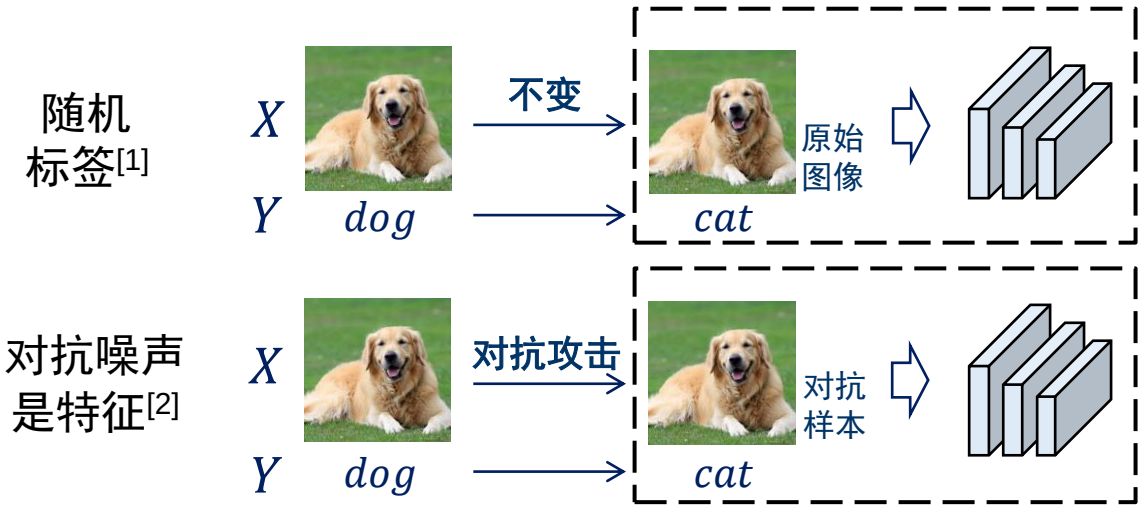
人 vs 算法：对抗样本

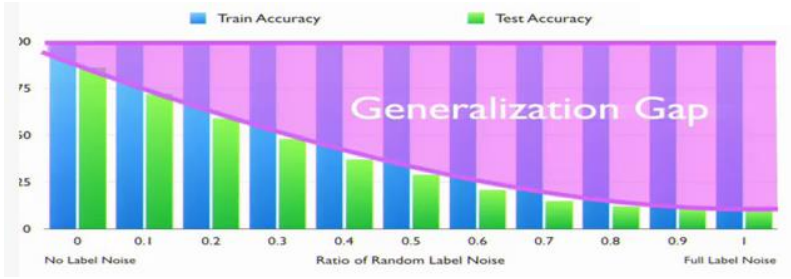
- 对抗噪声改变了测试样本非语义模式的分布



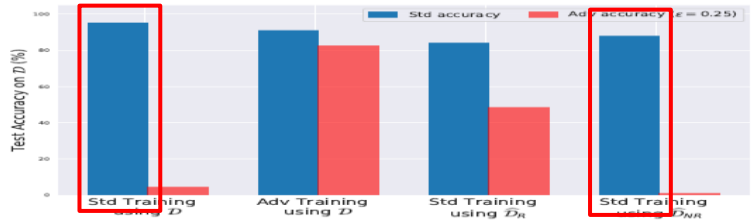


- 这些非语义模式虽然人类无法感知/理解，但被算法作为特征用来推断





泛化gap随随机标签比例增加[1]

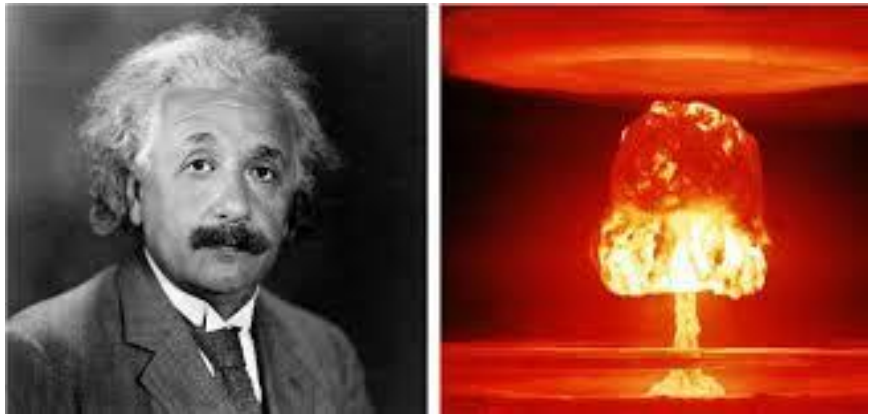


对抗样本/只用非语义模式训练的模型测试性能良好[2]

进一步说明人vs算法在感知推断上的差异

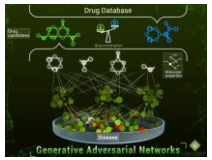
[1] Understanding Deep Learning Requires Rethinking Generalization. ICLR 2017 Best Paper.
[2] Adversarial Examples Are Not Bugs, They Are Features, NeurIPS 2019

技术的“善”与“恶”



“我当时是想把原子弹这一罪恶的杀人工具从疯子希特勒手里抢过来，想不到现在又将它送到另一个疯子手里。我们为什么要将几万无辜的男女老幼，作为这个新炸弹的活靶子呢？”

——爱因斯坦

	为善	作恶
爬虫	Google 搜索引擎	12306 抢票神器 www.downcc.com 打码平台 僵尸粉 
对抗生成模型	 设计辅助 药物研制 	  Deepfake虚假视频

抗争的对抗样本：“我命由我不由天”

良性对抗样本攻击

- **对抗恶意算法**
  - 对抗样本是现阶段算法的固有、共性缺陷
  - → 利用对抗样本攻击使恶意算法失效

- **对抗式图灵测试**
  - 传统模式识别任务逐渐被算法攻克，难以作为图灵测试区分人/算法
  - 对抗样本体现人-算法差异
  - → 基于对抗样本设计图灵测试来区分人/算法

人类 

算法 

- **对抗伪样本生成**
  - 算法利用的信息与人不同
  - 算法可以从对抗噪声中提取部分泛化到真实数据的特征
  - → 在真实(人认可)样本不足的情况下，对抗样本攻击提供了弥补数据的替代方案

对抗恶意算法：对抗隐私保护滤镜

Jiaming Zhang, Jitao Sang, et al.: Adversarial Privacy-preserving Filter. *ACM Multimedia 2020*.

人脸图像隐私保护

用户分享是人脸图像的主要来源

Pinterest MediaFire imgur imgfave SmugMug DEVIANT ART 500px tinypic

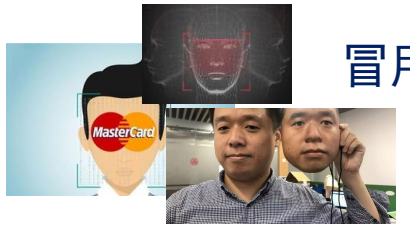
利用人脸图像的恶意算法

- DeepFake换脸带来伦理问题
- Kneron公司使用人脸面具骗过支付宝和微信支付系统
- Clearview AI公司从社交媒体平台上抓取了30亿张人脸图像，并提供人脸识别接口给美国 600 家执法机构

人脸图像大量曝光

人脸图像爬取（人脸检测） → 人脸信息泄露/侵犯（人脸识别）

换脸



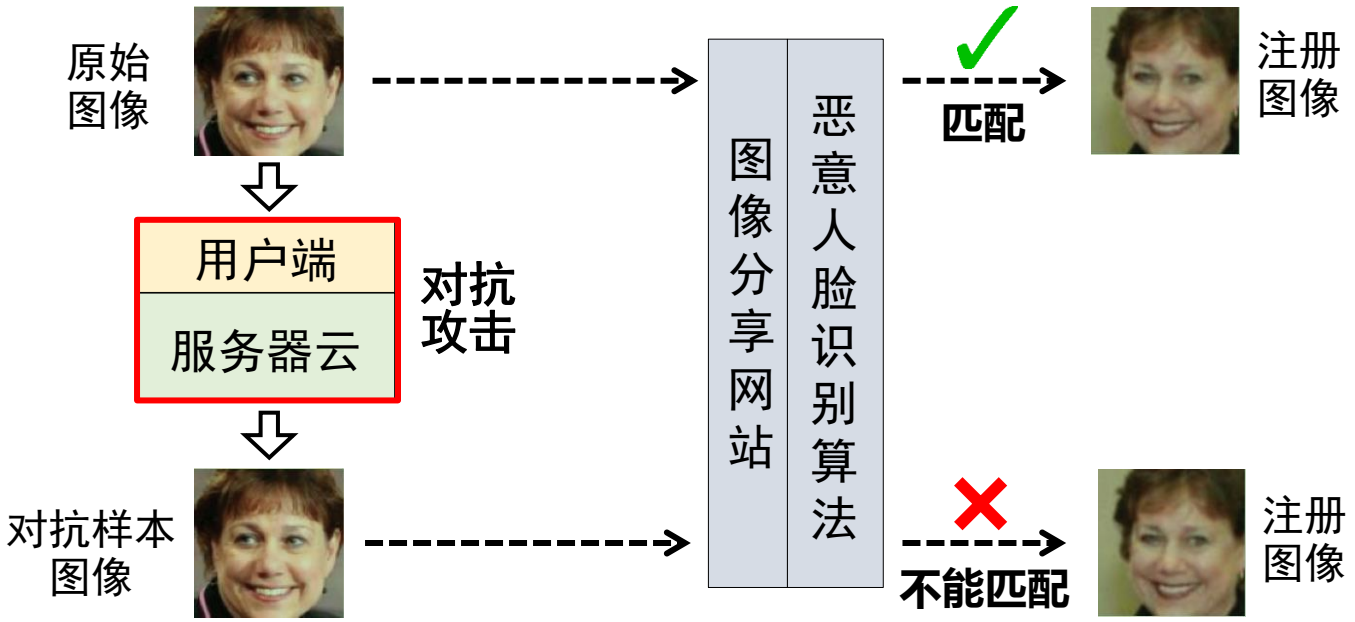
冒用

对抗隐私保护滤镜：Privacy+Utility+Non-accessibility

- **人脸图像隐私保护的要求**
  - **Privacy**：无法从分享的人脸图像中识别用户的身份信息
  - **Utility**：维持图像的原有观感和质量，不影响正常分享

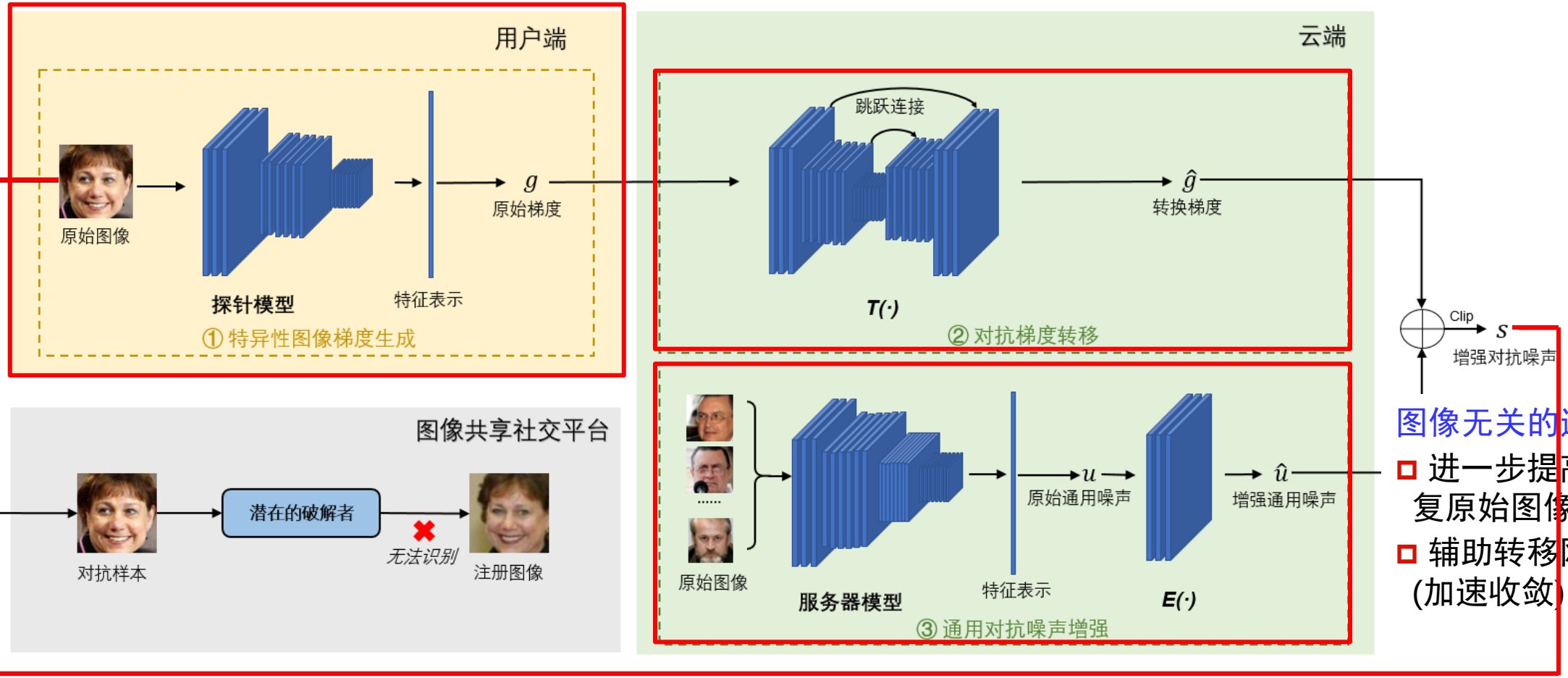
- **对抗样本的特点**
  - **欺骗算法**：使恶意人脸识别算法失效
  - **人不易察觉**：引入的对抗噪声对人几乎不可见

- **Non-accessibility**：用户设备之外无法访问原始图像 → 绝对数据安全
  - ✓ 对抗攻击设计大存储和计算量，一般在服务器端进行
  - ✓ 原始图像暴露到云上增加了泄露的风险
  - ✓ **端-云协同**：用户端通过小模型获得图像梯度信息，云上通过大模型对梯度增强



基于端-云协同对抗攻击的隐私保护框架

- 探针人脸识别模型获得对抗攻击图像相关的梯度信息 $g = \varepsilon \cdot sign(\nabla x I(x, x_e))$
- 兼容不同对抗攻击方法
- 通过类U-net转移网络对齐探针-标准人脸识别模型 $\hat{g} = T(g)$
- 转换后的梯度$\hat{g}$ 对潜在的恶意人脸识别模型有效



图像无关的通用噪声$\hat{u}$：

- 进一步提高从梯度恢复原始图像信息的难度
- 辅助转移网络的训练(加速收敛)

实验验证

LFW：原始图像的准确率为99.5%

	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM
g	24.6%	12.3%	11.8%	7.6%
$\hat{g}$	8.1%	4.6%	10.2%	1.1%
s	5.2%	2.6%	4.3%	1.2%

AgeDB-30：原始图像的准确率为95.2%

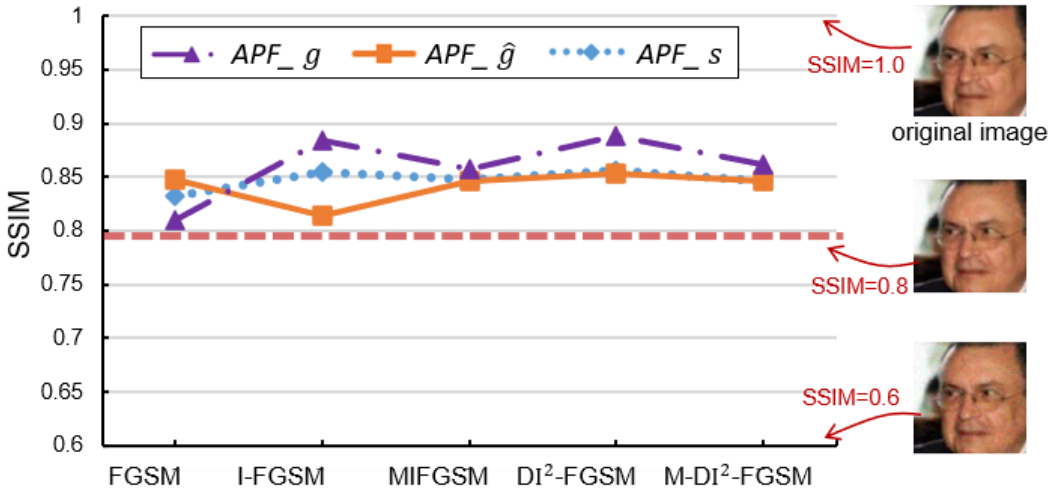
	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM
g	18.3%	13.7%	11.9%	9.5%
$\hat{g}$	17.7%	9.2%	9.2%	5.1%
s	11.7%	6.6%	6.2%	4.5%

Privacy

CFP-FP：原始图像的准确率为96.3%

	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM
g	51.4%	42.8%	36.2%	31.7%
$\hat{g}$	48.2%	27.2%	25.1%	15.3%
s	26.4%	14.2%	11.0%	8.4%

Utility



实验验证

LFW：原始图像的准确率为99.5%

	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM
g	24.6%	12.3%	11.8%	7.6%
$\hat{g}$	8.1%	4.6%	10.2%	1.1%
s	5.2%	2.6%	4.3%	1.2%
I可见	1.5%	0.6%	0.6%	0.6%

Non-accessibility

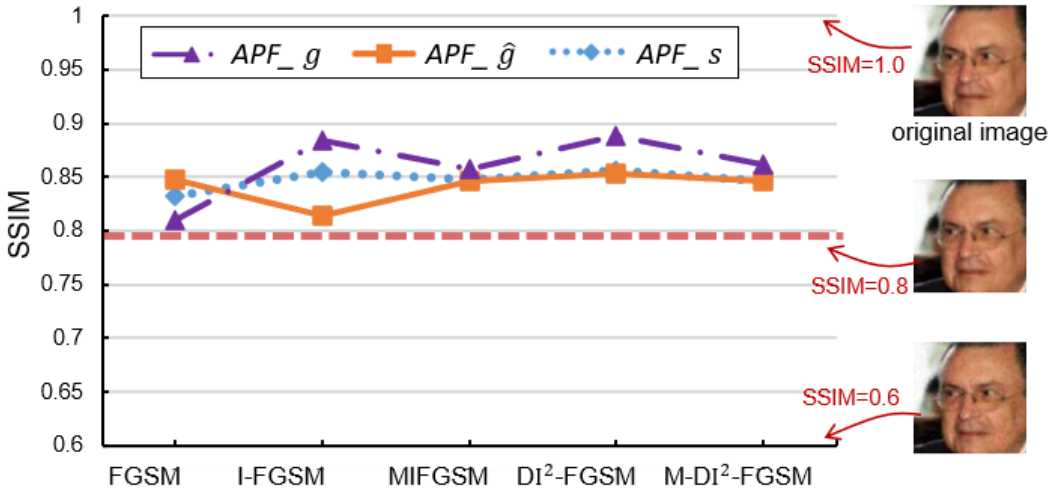
AgeDB-30：原始图像的准确率为95.2%

	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM
g	18.3%	13.7%	11.9%	9.5%
$\hat{g}$	17.7%	9.2%	9.2%	5.1%
s	11.7%	6.6%	6.2%	4.5%
I可见	4.2%	4.0%	4.0%	4.0%

CFP-FP：原始图像的准确率为96.3%

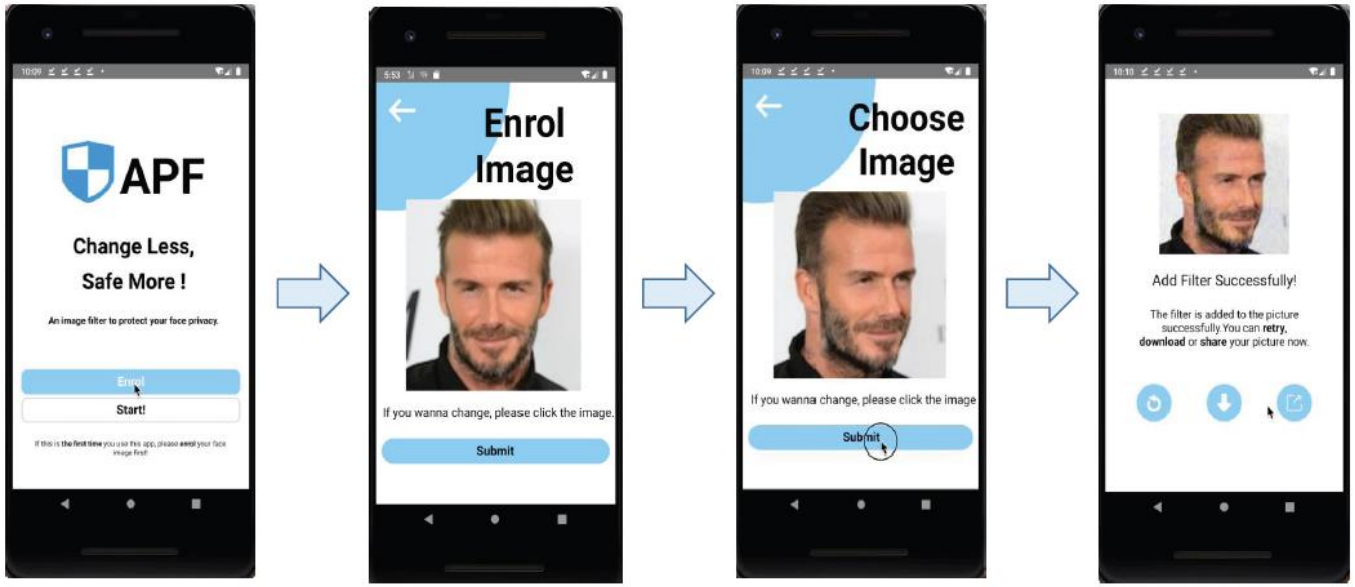
	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM
g	51.4%	42.8%	36.2%	31.7%
$\hat{g}$	48.2%	27.2%	25.1%	15.3%
s	26.4%	14.2%	11.0%	8.4%
I可见	7.5%	6.3%	9.2%	6.6%

Utility



APF原型系统

https://github.com/adversarial-for-goodness/APF



Demonstration

相关工作 & 拓展

- 存在的问题
  - 对抗攻击注册图像
  - 转化为分类问题（不是标准距离度量）：
    - 需要测试用户图像训练模型
    - 新用户加入需要重新训练模型

- 拓展 (进行中)
  - 对压缩等不敏感的对抗隐私保护滤镜
    - 图像分享网站/社交平台会对上传的图像进行压缩、裁剪等处理
    - 对抗噪声主要分布在高频域；压缩会减弱高频域信号

    → 低频域对抗攻击
  - 对抗攻击+风格迁移：美颜的同时实现隐私保护

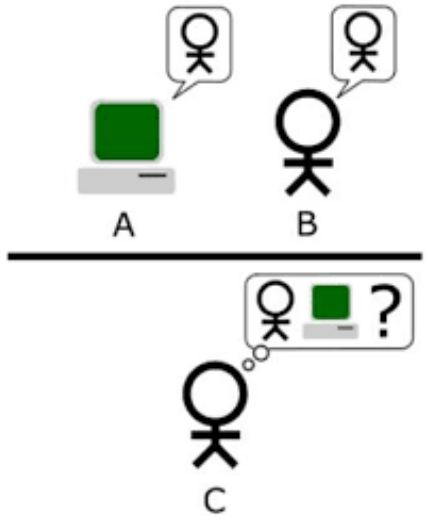
    → 模拟对抗噪声的图像纹理

Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. USENIS Security Symposium 2020.

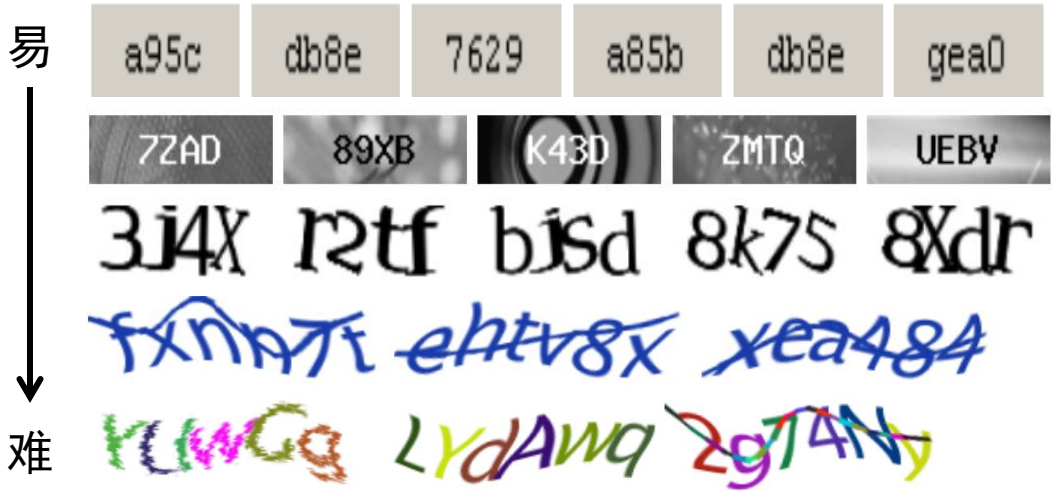
对抗式图灵测试：基于对抗样本攻击的鲁棒验证码

Jiaming Zhang, Jitao Sang, et al: Robust CAPTCHAs towards Malicious OCR. *IEEE Transactions on Multimedia, 2020*.

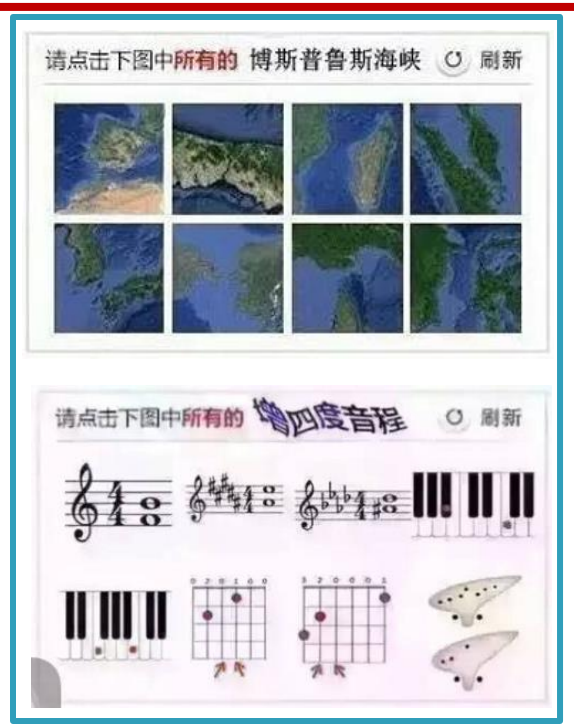
图灵测试的应用：越来越难的验证码



图灵测试

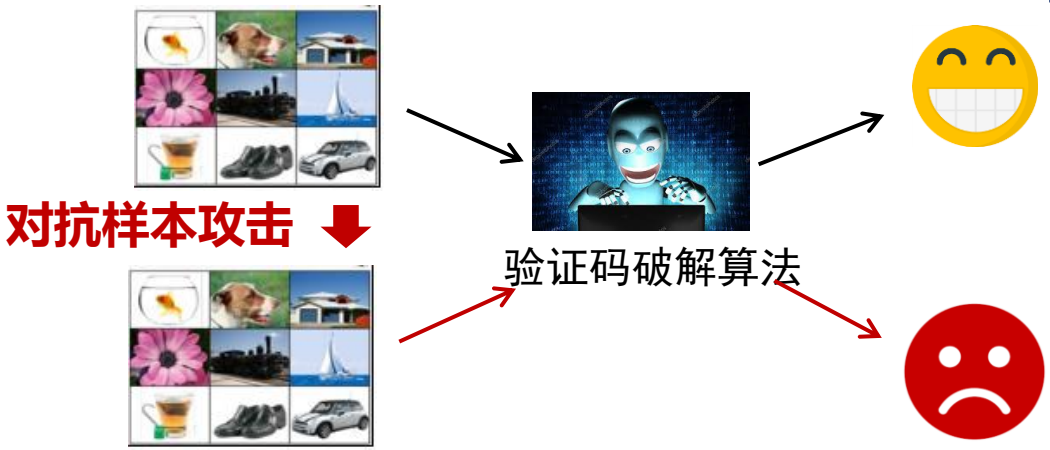


越来越难的验证码

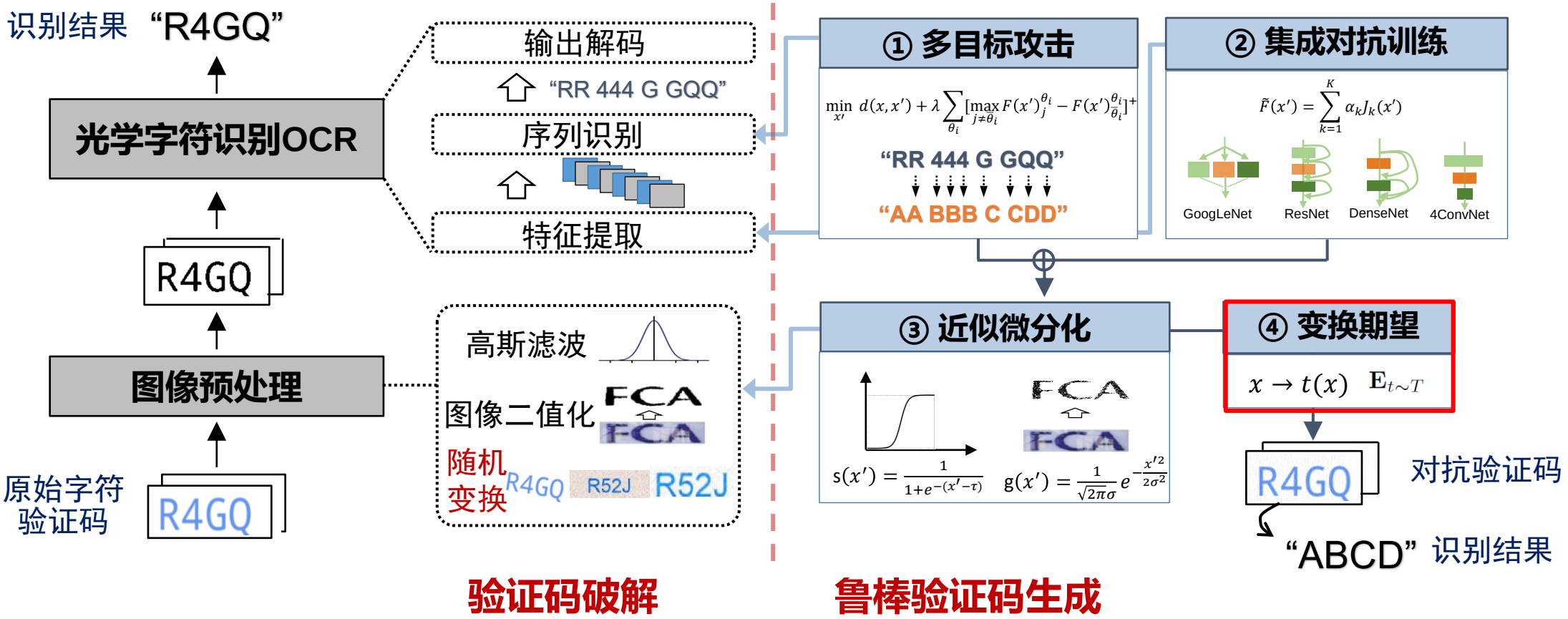


算法可回答&人很难回答

- **对抗图灵测试：**
  基于对抗样本对传统图灵测试任务进行调整，使原有算法难以通过

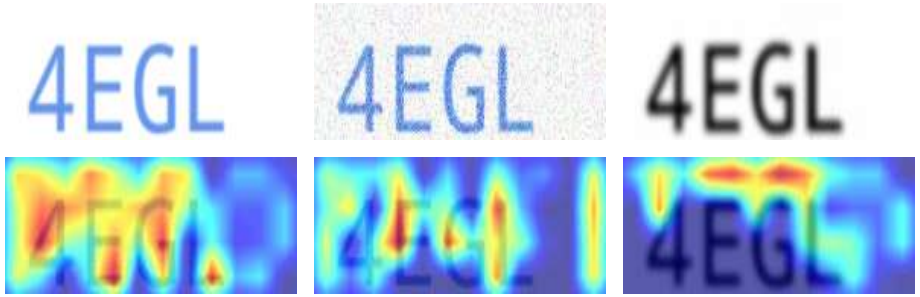


基于对抗样本的鲁棒验证码生成框架

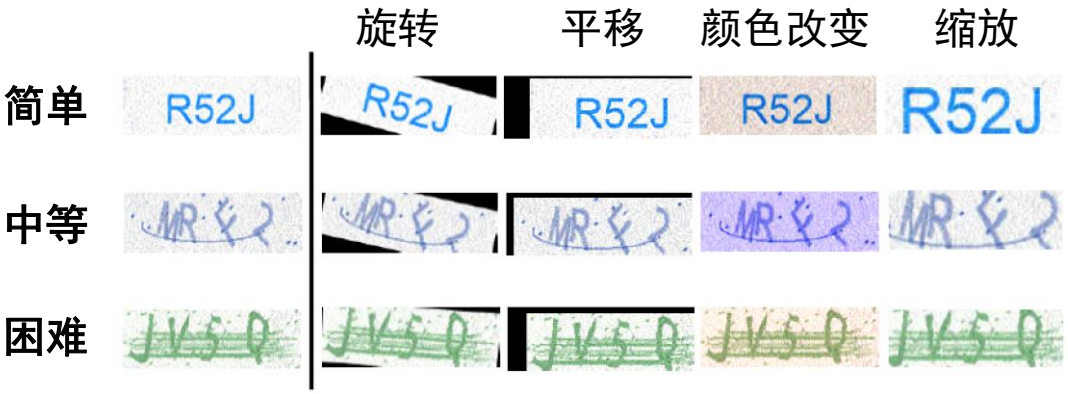


实验验证

注意力图

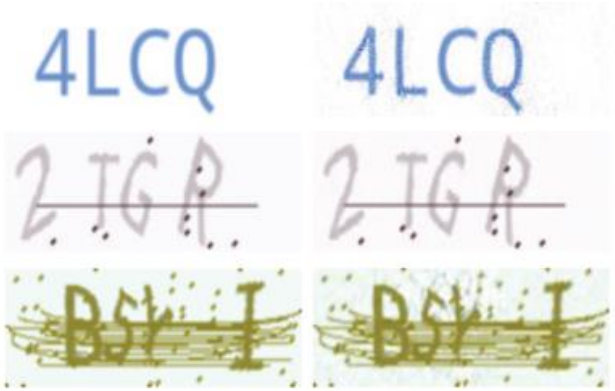


原图像　　高斯白噪声图像　　对抗样本图像



随机变换下的验证码识别率

Images	Easy	Medium	Hard
Raw	95.0%	91.6%	73.1%
Adversarial	1.3%	5.4%	9.0%



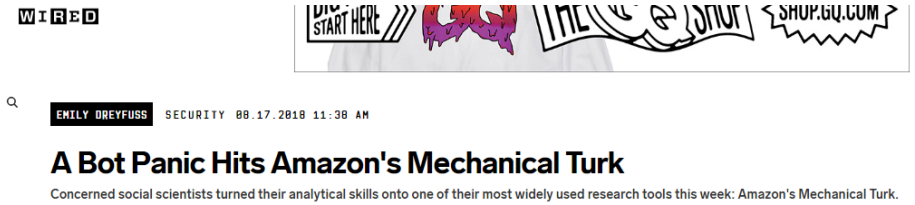
原图像　　对抗样本图像

验证码识别准确率（验证码破解算法 vs 人类）

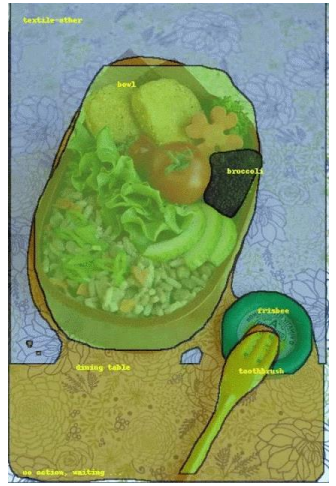
		原始CAPTCHA	对抗CAPTCHA
简单	破解算法	100.0%	0.0%
	人类	99.0%	94.0%
中等	破解算法	91.0%	0.5%
	人类	73.0%	65.0%
困难	破解算法	81.0%	4.0%
	人类	56.0%	49.0%

拓展

- **数据标注的回音壁现象**
  - **数据标注现状**：算法辅助、甚至替代人类标注逐渐成为趋势
  - **回音壁（社会学）**：隔离情况下的信息多次内部传播会造成观点极端化
  - **算法标注数据的回音壁问题：**
    - **泛化性：错误传递** (多样性缺失)
    - **公平性：偏见累积** (数据偏见放大)
    - **安全性：漏洞继承** (数据为载体)
  - **待标注数据集的对抗样本化 → 区分人类/算法标注**

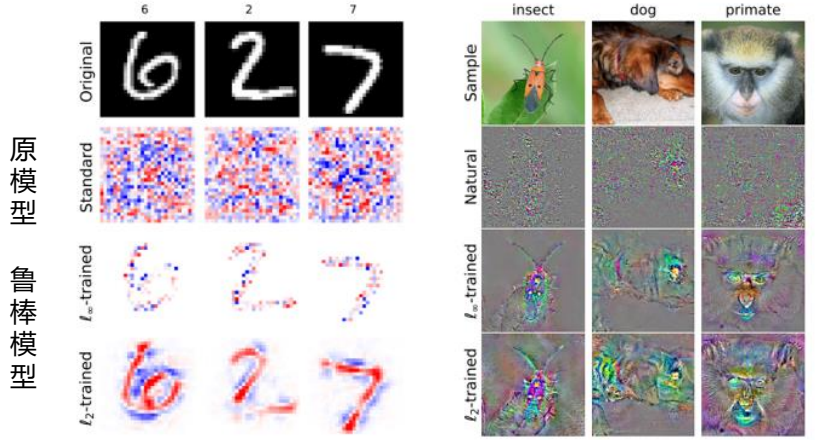


亚马逊众包平台发现机器人标注



谷歌“流体标注”辅助工具

- **消除人-算法差异 → 提高解释性**
  - 对抗攻击利用算法有别于人的特征
  - 对抗训练提高算法鲁棒性 → 迫使算法利用类人特征
  - **基于对抗训练的可解释生成模型：**
    - 对判别器进行对抗训练：判别器更多依赖类人特征
    - 生成器想欺骗判别器，生成的样本也必须更多使用类人特征



对抗训练模型具有语义更清晰的归因可视化结果

对抗伪样本生成：基于对抗样本数据增强的视觉去偏见

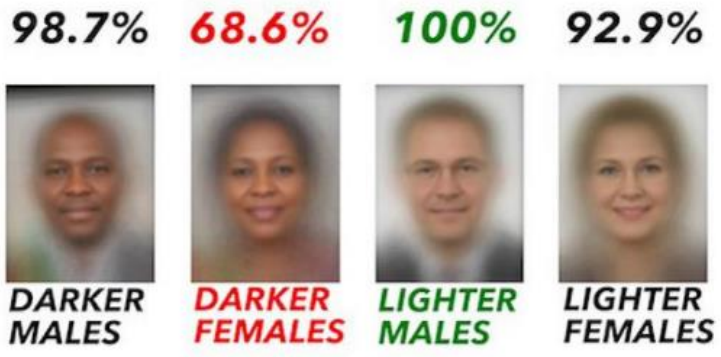
Yi Zhang, Jitao Sang: Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. *ACM Multimedia 2020*.

算法偏见（公平性）研究

- 算法偏见：同一个算法在决策时对敏感人群产生差异化结果：

（1）对于不同人群的决策效果不同；

（2）在决策时会利用和任务不相关的人群特征.



不同性别和肤色人脸的识别准确率相差很大

自动量刑算法COMPAS为黑人判定更高的犯罪几率

- **目标变量**：待预测的任务标签（职业）
- **偏见变量**：可能影响预测结果的敏感属性（性别）

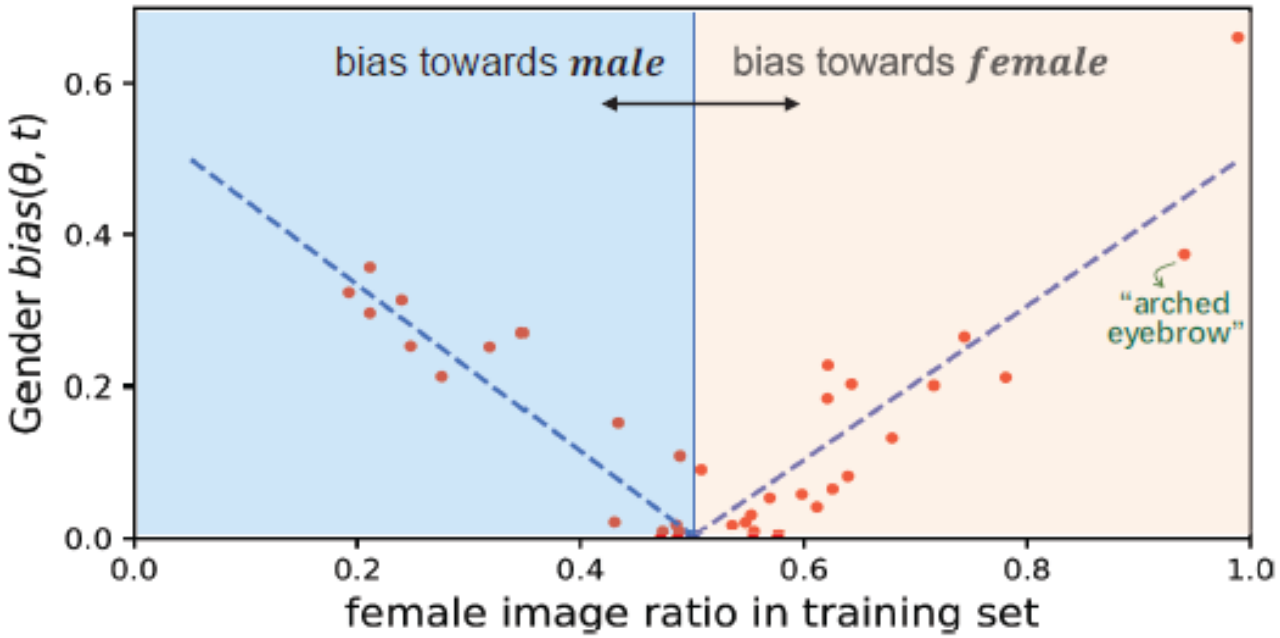
女护士 ✓

男护士 ✗

算法偏见 vs 数据不平衡分布

CelebA数据集

- **目标变量**：面部属性 (如柳叶眉)
- **偏见变量**：性别

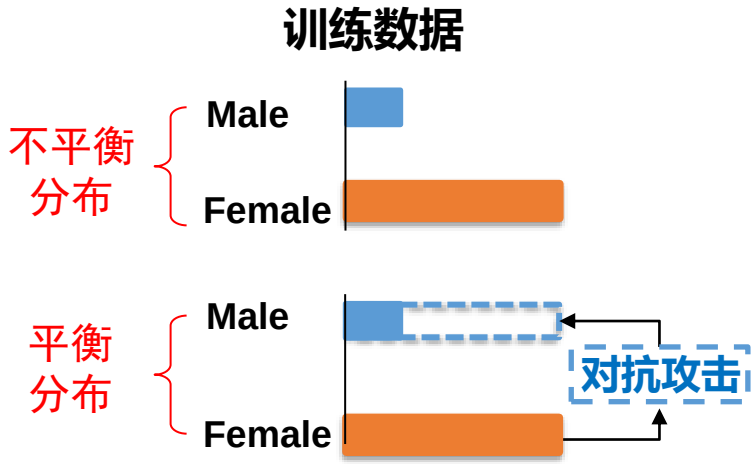


性别偏见：

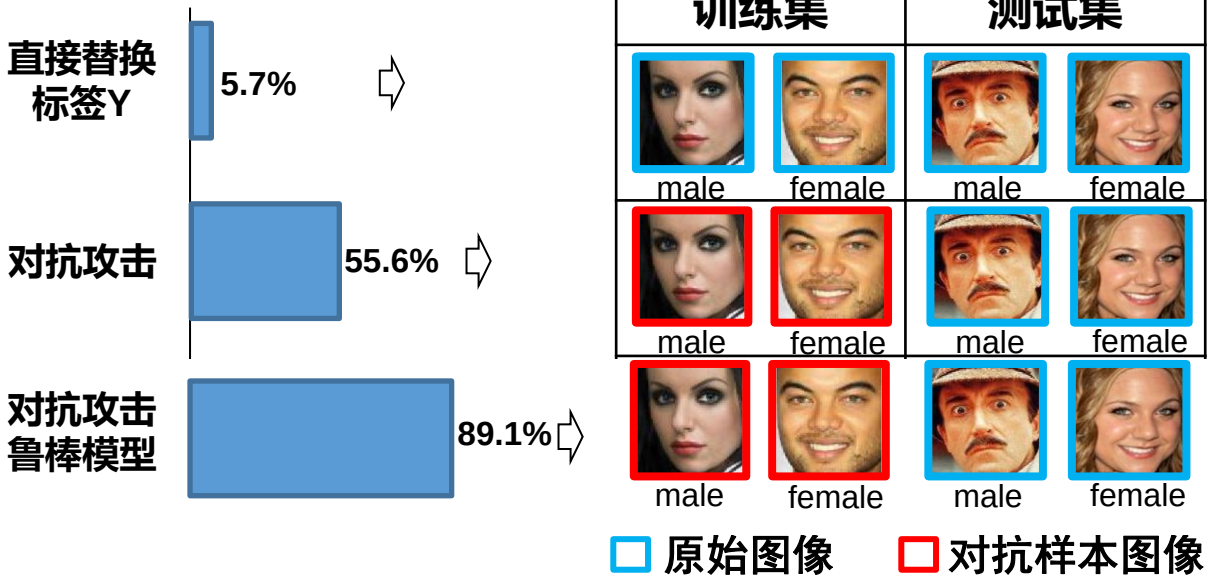
不同性别变量条件下目标任务准确率的差异

$$bias(\theta, t) = |P(\hat{t} = t | b = 0, t^* = t) - P(\hat{t} = t | b = 1, t^* = t)|$$

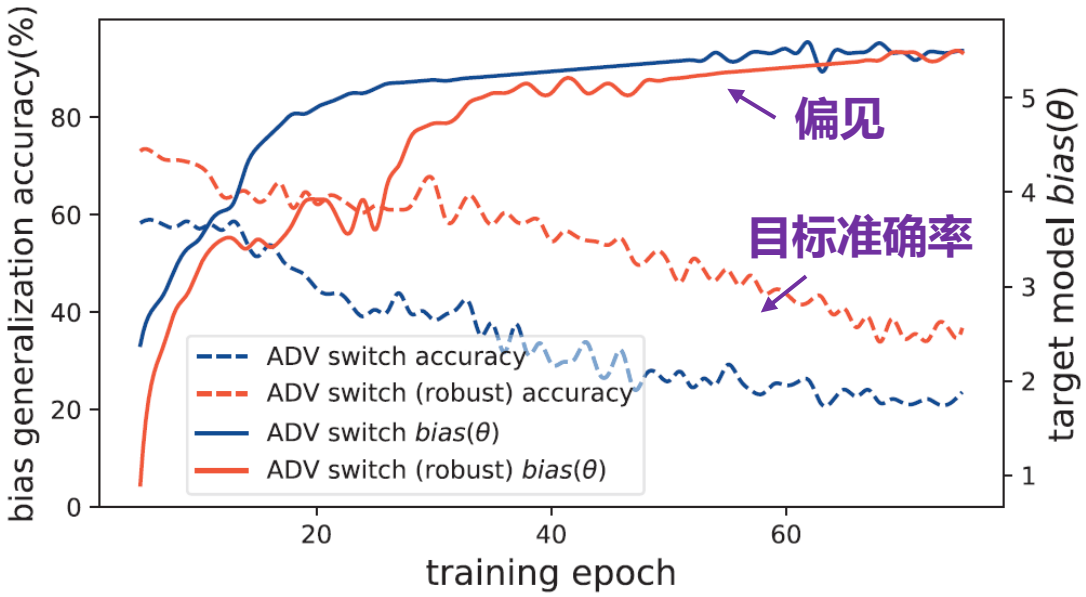
对抗伪样本→平衡数据分布



攻击鲁棒模型：更强的对抗样本 → 更好地泛化到攻击类



跨任务迁移性：攻击类信息随训练进行逐渐丢失

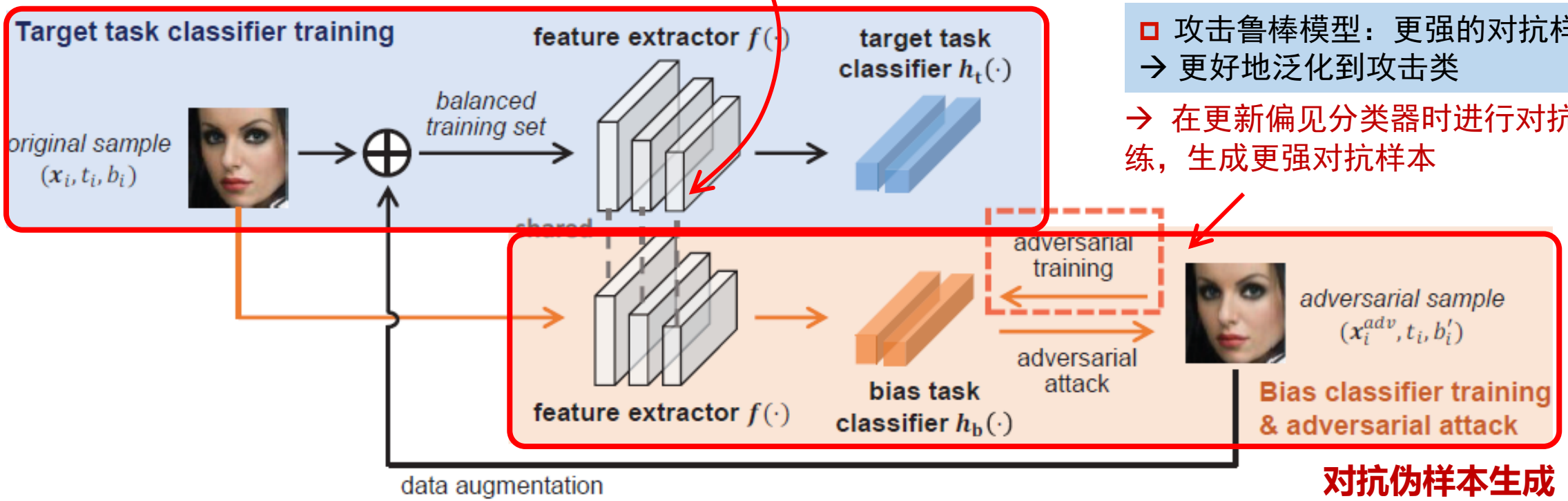


基于对抗伪样本数据增强的算法去偏见框架

- 跨任务迁移性：攻击类信息随训练进行逐渐丢失

→ 将对抗样本生成和目标分类器训练耦合：共享特征提取器，同步更新

目标任务分类器训练



- 攻击鲁棒模型：更强的对抗样本

→ 更好地泛化到攻击类

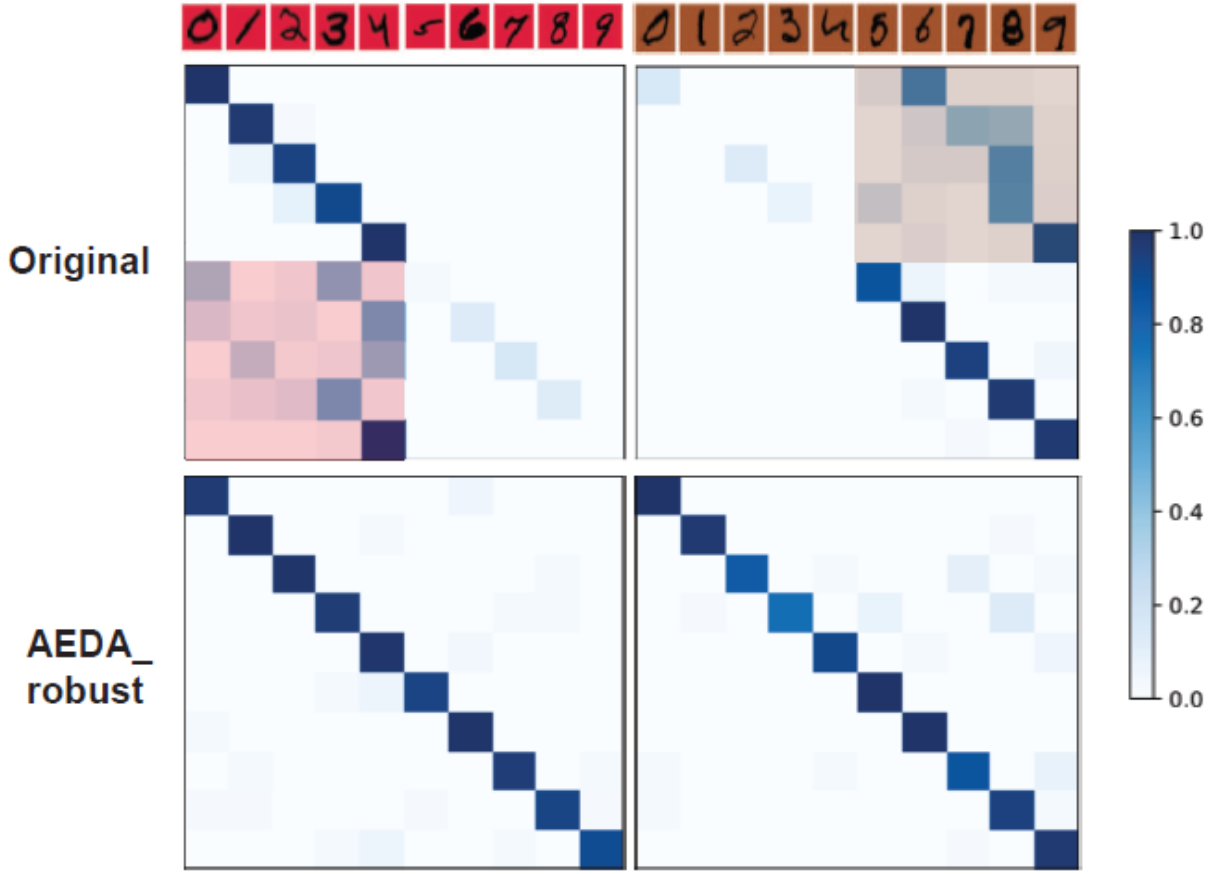
→ 在更新偏见分类器时进行对抗训练，生成更强对抗样本

对抗伪样本生成

实验验证: 模拟偏见数据集C-MNIST

- (模拟)偏见属性：背景颜色// 目标任务：手写数字识别

Methods	bACC (%)	Model bias
Original	55.62	7.84
Under-sampling [7, 35]	–	–
Reweighting [18]	–	–
Adv debiasing [4, 12, 31]	89.93	1.37
CycleGAN [36]	65.23	5.42
AEDA_pre	64.53	5.89
AEDA_online	80.57	3.20
AEDA_robust	**91.80**	**0.528**

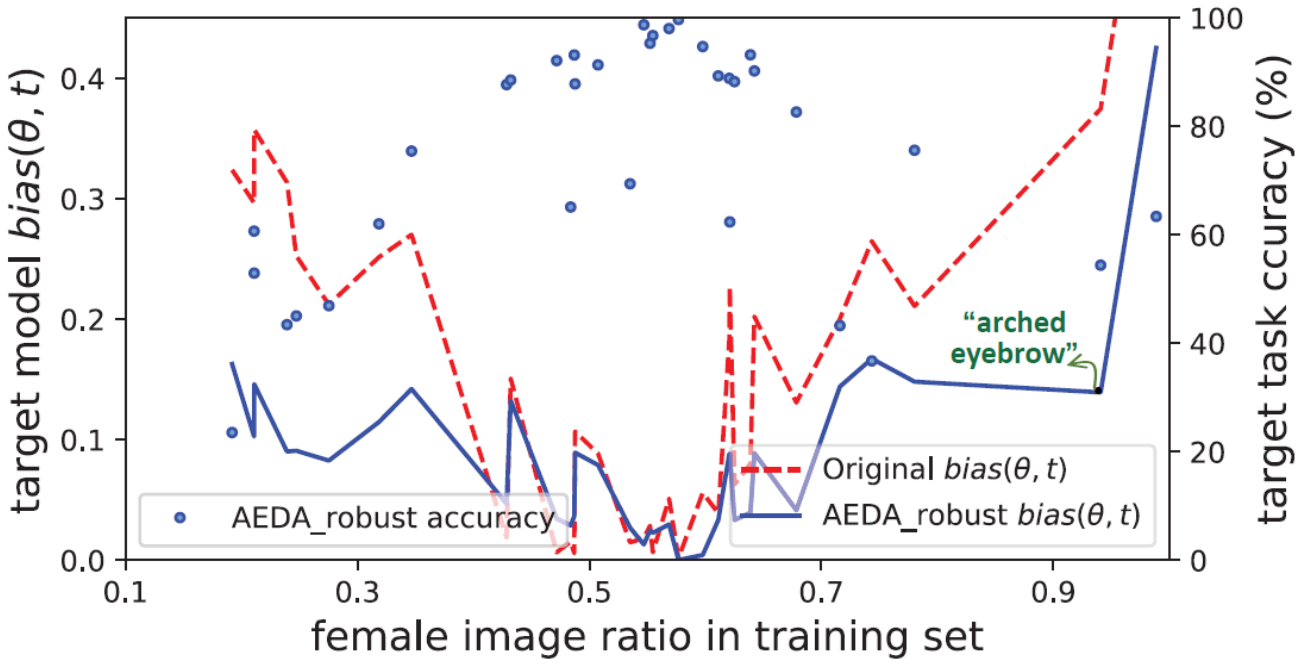


不同测试集的混淆矩阵
（左：<0-9, 红色>;右：<0-9，棕色>）

实验验证: 真实偏见数据集CelebA

Methods	bACC (%)	Model bias
Original	73.57	5.48
Under-sampling [7, 35]	66.35	**2.35**
Reweighting [18]	73.82	4.39
Adv debiasing [4, 12, 31]	72.82	4.23
CycleGAN [36]	73.65	4.75
AEDA_pre	73.68	5.23
AEDA_online	74.03	4.22
AEDA_robust	**74.30**	3.27

CelebA不同面部属性识别的准确率&性别偏见



不同面部属性去偏见效果 & 识别准确率

讨论&拓展

- **目标任务准确率 vs 公平性**
  - **准确率-公平性折衷**：传统去偏见方法牺牲准确率换取公平性的提升
  - **准确率-公平性一致**：背后原因都是训练集不平衡数据分布，通过生成伪样本数据增强可以同时提高准确率&公平性

- **基于对抗样本攻击的伪样本生成**
  - **监督学习困境:** 训练数据缺失、人工标注成本高

  → 不需要人认可的真实样本，利用算法认可的对抗伪样本 （对抗攻击作为一种生成模型）

零样本&小样本问题

可见类样本　　未知类/小样本类伪样本

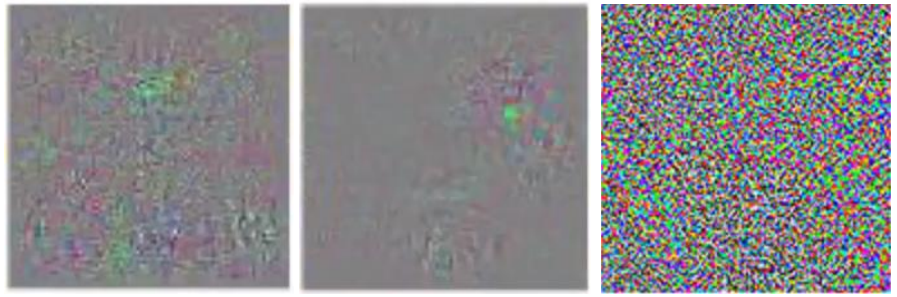
Methods	digit 2	4	6	8
Original	46.80	70.37	0.00	0.00
AEDA_robust	**69.25**	**88.88**	**42.50**	**25.00**

MNIST 4-way 4-shot识别结果

结束语：对抗样本攻击 vs 虚假的相关性

对抗噪声-非语义特征 ∈ 虚假的相关性

- **算法从训练数据中学习强关联模式**
  - 这些模式有时不是基于语义的 → 人不易察觉
  - 这些模式有时来自对数据集的过拟合 → 算法过于敏感

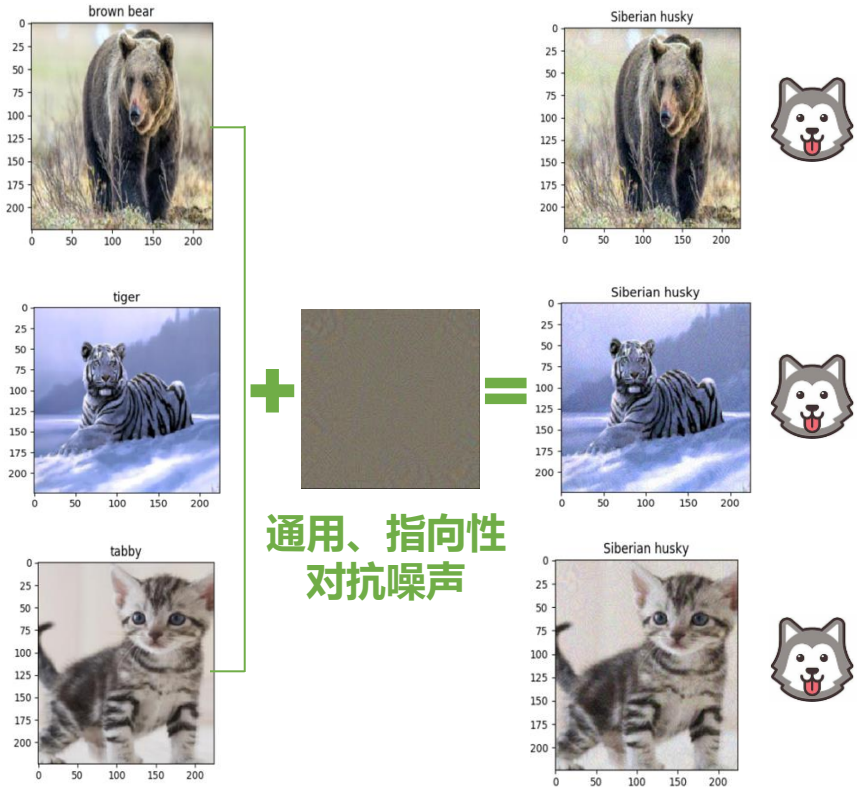


- **对抗噪声-非语义特征 ∈虚假的相关性：**

模型利用训练数据中具有强关联性、但无法在测试集中泛化的特征

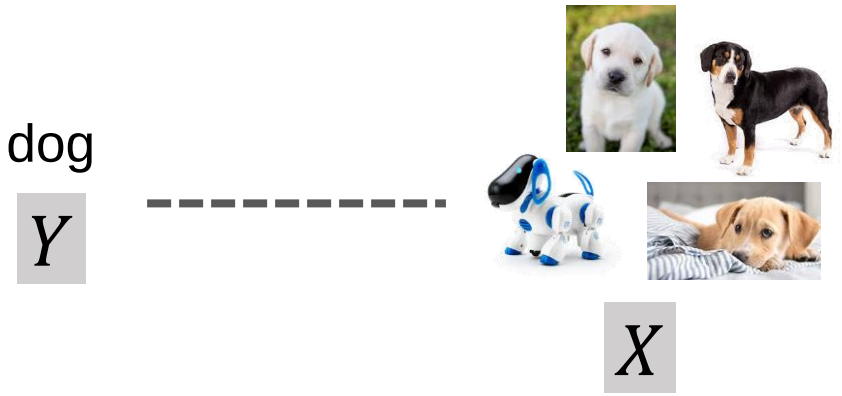
数据的不平衡分布

- **确保在测试集泛化**
- **难以在测试集泛化**



虚假的相关性~任务无关生成变量

- **数据生成过程的假设**
  - 数据生成过程和得到的数据分布与要解决的任务相关
  - 输入X变化性大，难以直接约束（IID）



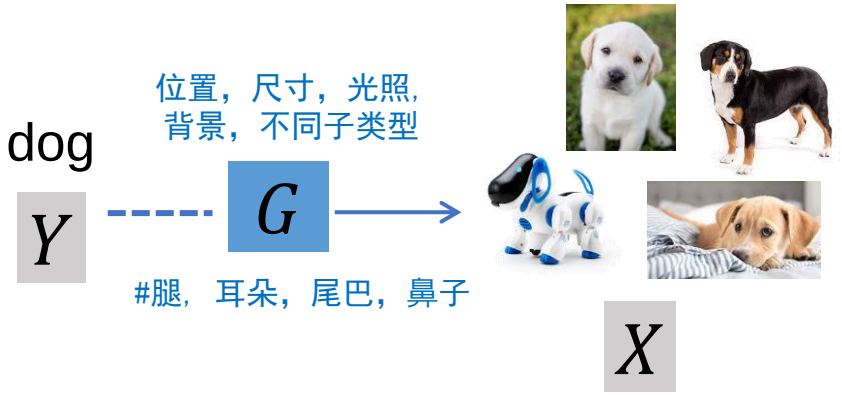
Guanhua Zheng, Jitao Sang, et al. A Generalization Theory based on Independent and Task-Identically Distributed Assumption. Arxiv 2019.

虚假的相关性~任务无关生成变量

- **数据生成过程的假设**
  - 数据生成过程和得到的数据分布与要解决的任务相关
  - 输入X变化性大，难以直接约束（IID）

- **【定义】生成变量**
  - 影响输入X生成的随机变量G：
    $X = \phi(G)$
  - **基于生成变量的数据生成过程**：

    对于每个样本i
    - ✓ 采样生成变量和标签
      $(g_i, y_i) \sim P_{D_i}(G, Y)$
    - ✓ 生成输入 $x_i = \phi(g_i)$

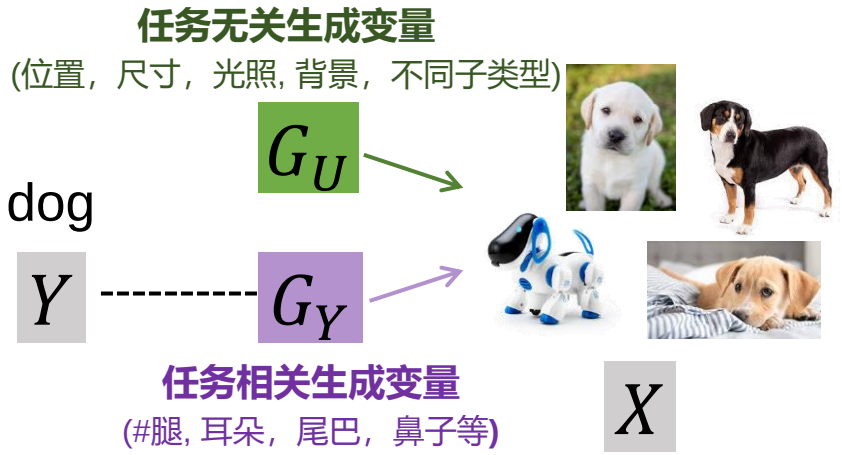


Guanhua Zheng, Jitao Sang, et al. A Generalization Theory based on Independent and Task-Identically Distributed Assumption. Arxiv 2019.

虚假的相关性~任务无关生成变量

- **任务相关生成变量G_Y**
  - 和任务紧密相关，随任务改变而改变
  - 任务无关生成变量G_U: 影响X生成，但与任务关联不稳定

- **独立同任务分布ITID假设**
  - 数据遵循基于生成变量的生成过程
  - 对于任意数据子集D_i, D_j，在G_Y上具有相同的边缘分布
    $P_{D_i}(G_Y, Y) = P_{D_j}(G_Y, Y)$



- **ITID特点**
  - 放松同分布约束：引入生成变量，样本可采样自不同分布P_{D_i}
  - 考虑任务的影响：定义任务相关生成变量, 如随机标签改变了任务；
  - 保留输入X的多样性：定义任务无关生成变量，如图像光照、背景等随机变量

Guanhua Zheng, Jitao Sang, et al. A Generalization Theory based on Independent and Task-Identically Distributed Assumption. Arxiv 2019.

基于独立同任务分布的泛化理论

- **【定义3】模型对任务无关生成变量的γ依赖性**

  模型h对G_U具有γ依赖性是指：

$$H(\hat{Y}|G_Y) \leq \gamma$$

  γ描述了任务无关变量G_U对模型输出的影响：

$$I(\hat{Y}; G_U|G_Y) = H(\hat{Y}|G_Y) - H(\hat{Y}|G) = H(\hat{Y}|G_Y) - 0 \leq \gamma$$

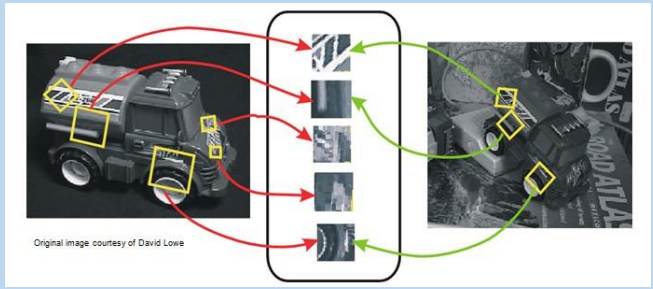
- **【定理二】一般情况下的泛化误差界**

  “若假设空间$\mathcal{H}$对G_U具有γ依赖性，对任意h $\in \mathcal{H}$,以不低于$1-\delta$的概率满足：

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon + \frac{\gamma}{log2}$$

  最优模型的泛化误差　G_Y依赖时的泛化误差上界

模型的泛化性能除了与任务复杂度（ϵ：任务相关变量的状态# + 输出变量状态#）有关，还与任务无关变量的影响γ有关



- **图像识别通过提高不变性改善泛化性能:**
  - 位置、角度、尺寸等是任务无关变量 → 对应平移、旋转、缩放不变性
  - 手工设计不变性特征 + 深度学习设计卷积、池化等

基于独立同任务分布的泛化理论

- **【定理五】生成变量的统计独立 vs 不变性：**

  “若生成变量G_i在训练集上与其他变量（包括其他生成变量$\bar{G}_i$和输出变量Y）统计独立，则最优模型h'对G_i具有不变性”

调整平衡训练集提高任务无关生成变量的统计独立性
→ 减少对最优模型的影响，提高泛化性能

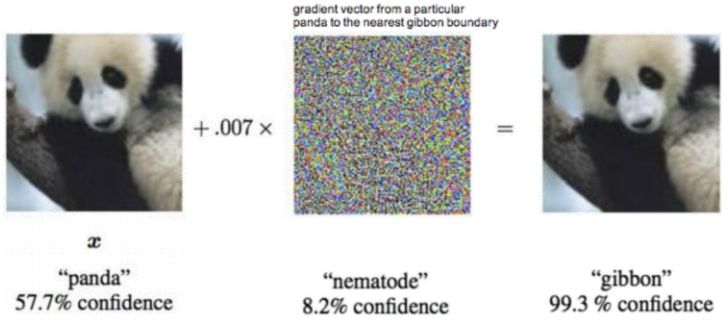
- **解释数据增强现象**
  - 传统数据增强：提高已有的任务无关变量的统计独立性

  - 新数据增强：增加新的任务无关变量（如random erasing）
    - ✓ 新的变量本身统计独立（平衡形状、位置等参数）
    - ✓ 间接使已有的严重失衡无关变量独立（**遮挡、缺失**等）

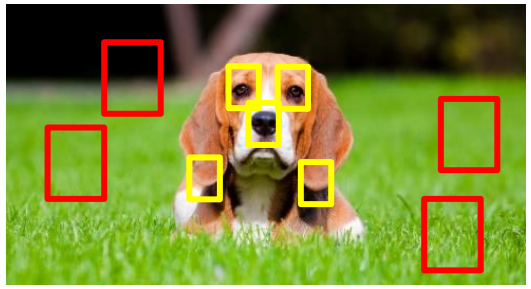
虚假的相关性: 泛化性/对抗鲁棒性/因果性/公平性/解释性

- **虚假的相关性：**
  - ✓ **泛化性**：任务无关变量 → 语义&非语义
  - ✓ **对抗鲁棒性**：非鲁棒特征 → 非语义
  - ✓ **因果**：混淆变量
  - ✓ **公平性**：偏见属性

  } 语义



对抗攻击：测试样本中扰动非鲁棒特征



因果推断：模型推断(狗识别)不依赖混淆变量(背景)

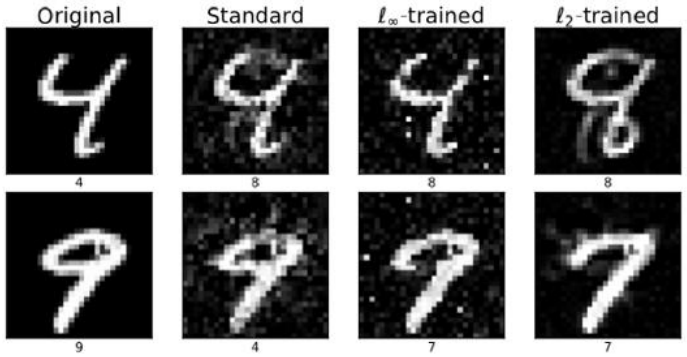
算法偏见：模型推断(再犯罪概率) 依赖偏见属性(种族)

虚假的相关性: 泛化性/对抗鲁棒性/因果性/公平性/解释性

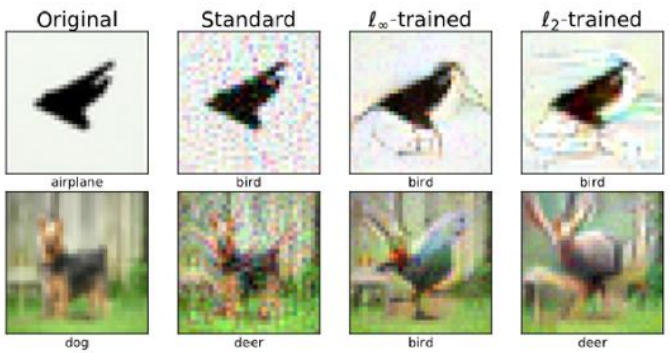
- **虚假的相关性：**
  - ✓ **泛化性**：任务无关变量 → 语义&非语义
  - ✓ **对抗鲁棒性**：非鲁棒特征 → 非语义
  - ✓ **因果**：混淆变量
  - ✓ **公平性**：偏见属性

  （因果、公平性 → 语义）

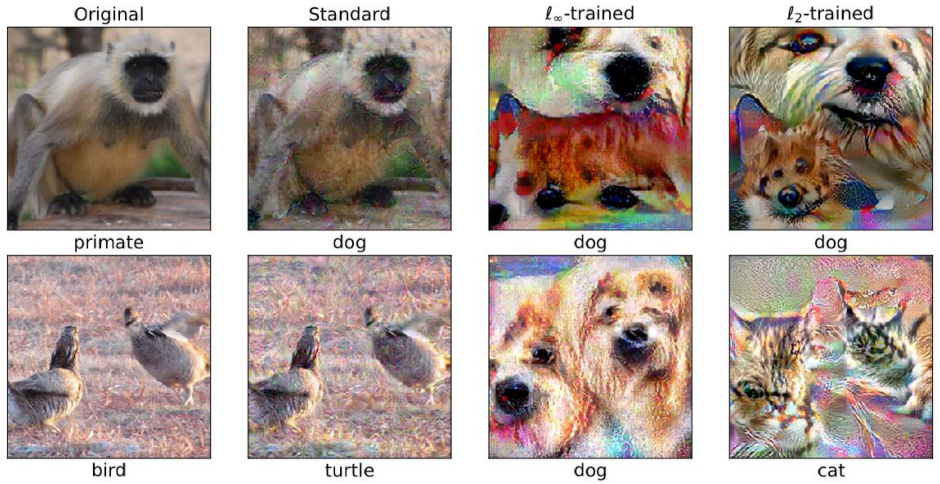
⇨ **解释性**：约束模型对非鲁棒特征的利用，可以提高模型解释性



(a) MNIST



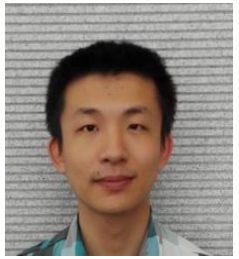
(b) CIFAR-10



(c) Restricted ImageNet


张家明


张翼


郑冠华


赵宪


吴尚锡

谢谢