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

组合对抗攻击的自动化搜索方法

硕士研究生　关迎丹

2021年05月05日

内容提要

- 背景简介
- 基本概念
  - 对抗样本攻击
  - 自动机器学习
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

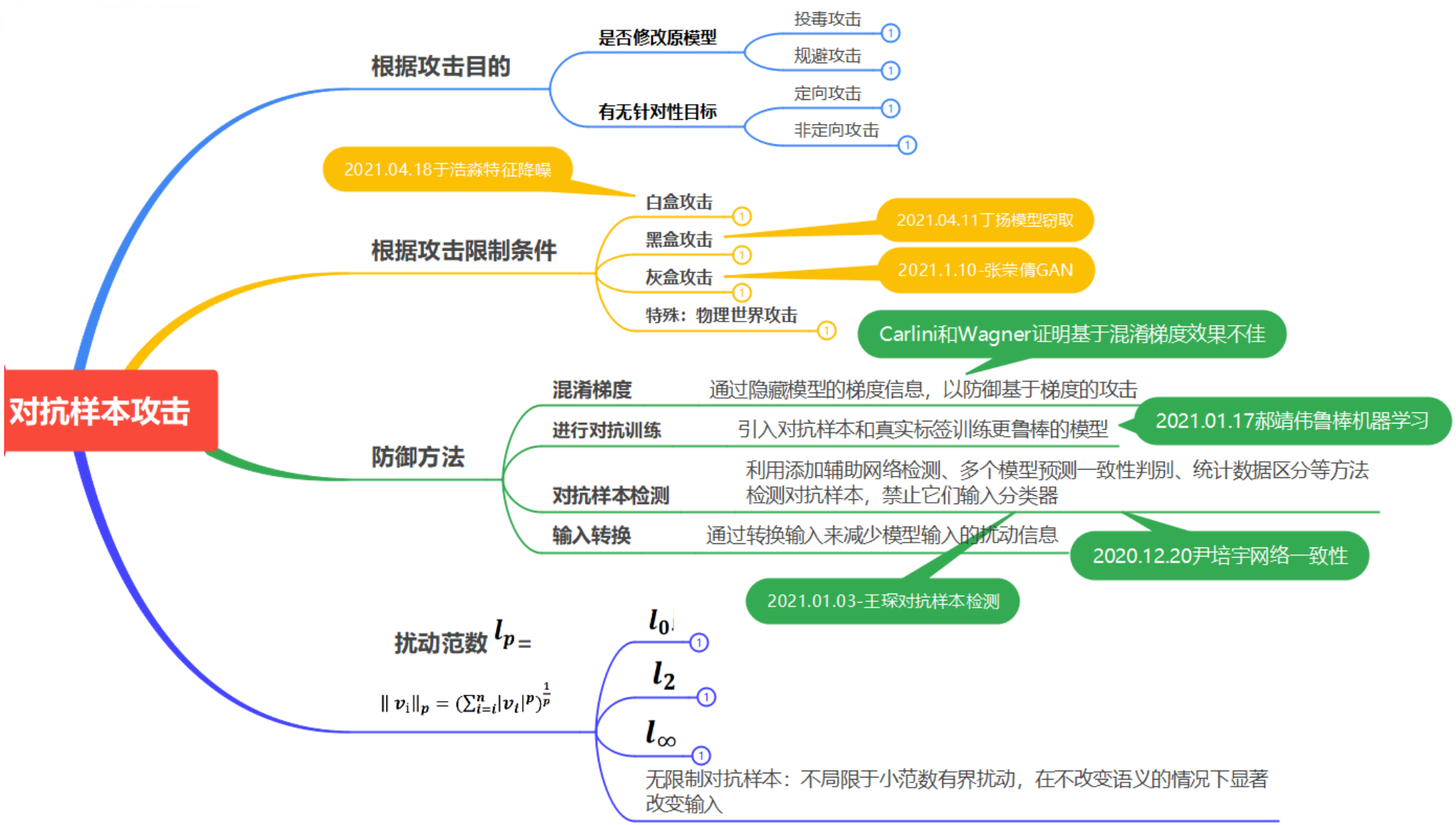
- 预期收获
  - 1. 了解对抗样本攻击的基础知识点
  - 2. 了解自动机器学习的搜索空间和搜索策略
  - 3. 理解组合对抗攻击的自动化搜索方法
  - 4. 了解组合对抗攻击自动化搜索在智能系统测试中的应用

背景简介

- 对抗样本攻击在智能系统安全检测中的实例
  - 2019年4月初，腾讯科恩实验室在特斯拉 Autopilot系统上进行安全性检测
  - 对抗样本攻击使Autopilot在不发出警告的情况下控制车辆驶入错误车道

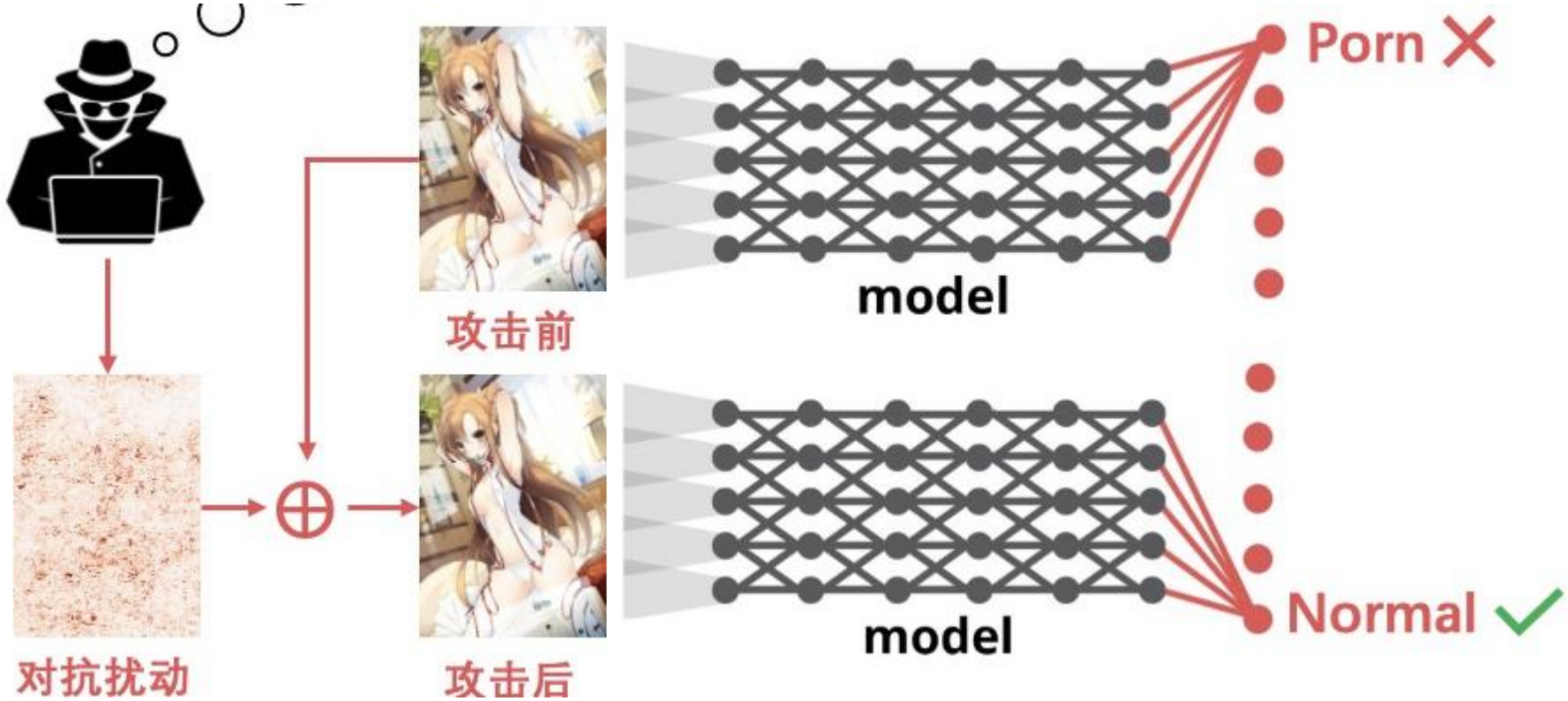
基本概念

对抗样本攻击


对抗样本攻击
根据攻击目的
是否修改原模型
投毒攻击
规避攻击
有无针对性目标
定向攻击
非定向攻击
根据攻击限制条件
白盒攻击
黑盒攻击
灰盒攻击
特殊：物理世界攻击
2021.04.18于浩淼特征降噪
2021.04.11丁扬模型窃取
2021.1.10-张荣倩GAN
防御方法
混淆梯度
通过隐藏模型的梯度信息，以防御基于梯度的攻击
Carlini和Wagner证明基于混淆梯度效果不佳
进行对抗训练
引入对抗样本和真实标签训练更鲁棒的模型
2021.01.17郝靖伟鲁棒机器学习
对抗样本检测
利用添加辅助网络检测、多个模型预测一致性判别、统计数据区分等方法检测对抗样本，禁止它们输入分类器
输入转换
通过转换输入来减少模型输入的扰动信息
2020.12.20尹培宇网络一致性
2021.01.03-王琛对抗样本检测
扰动范数 $l_p=$
$\|v_i\|_p=(\sum_{i=1}^{n}|v_i|^p)^{\frac{1}{p}}$
l_0
l_2
l_∞
无限制对抗样本：不局限于小范数有界扰动，在不改变语义的情况下显著改变输入

投毒攻击和规避攻击

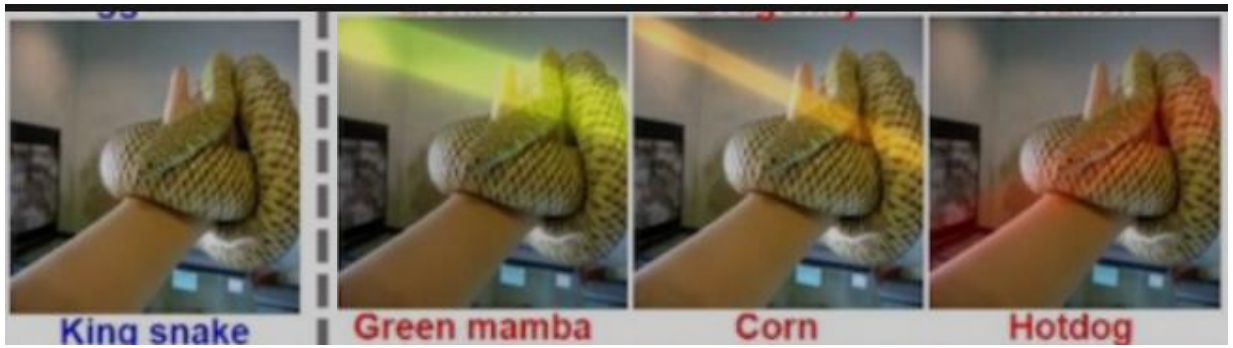
- 投毒攻击(Poisoning Attacks)
  - 在训练阶段，攻击者生成对抗样本污染训练数据库，导致训练的模型失效或准确性降低
- 规避攻击(Evasion Attack)
  - 在测试阶段，攻击者生成对抗样本欺骗模型，使其产生错误结果，从而逃避检测



物理世界攻击

- 物理世界攻击(Physical world attack)
  - 攻击者无法将对抗样本直接输入模型检测(例如：摄像头输入)
  - CVPR 2021：激光束干扰法 (AdvLB)

加了激光的路标可能被识别为电影院



加了激光的蛇图片可能被识别为热狗

扰动范数

- 扰动范数 (Perturbation norm)
  - 范数是一种强化了的距离概念，在对抗样本中用于测量扰动的大小
  - $L_p = \|x\|_p = (\sum_{i=1}^{n} x_i^p)^{\frac{1}{p}}$
  - L_0范数：限制更改像素数量，不限制每个像素更改程度 (JSMA)
  - L_∞范数：限制像素更改的程度，不限制更改数量 (FGSM)
  - L_2范数：限制累积更改，在更改数量与更改程度之间达到某种平衡 (Deep Fool)

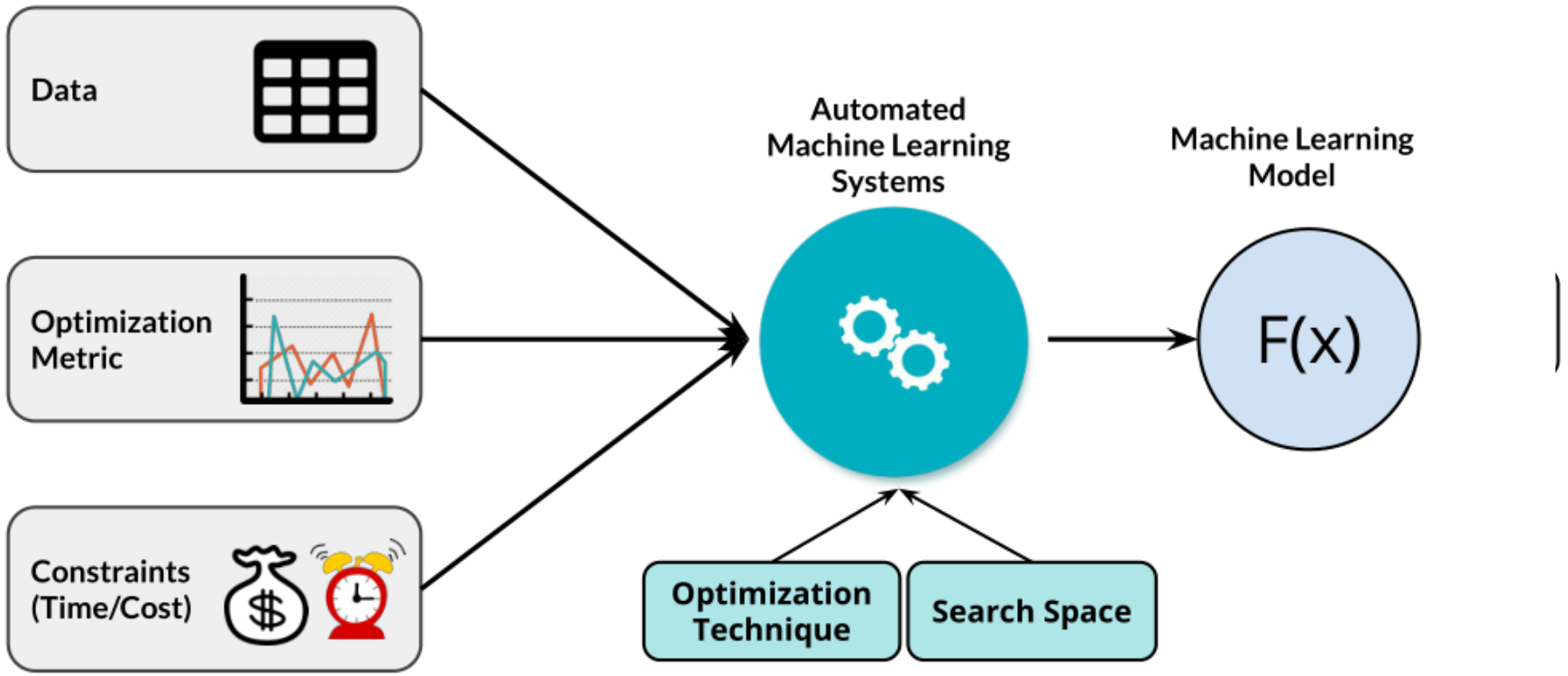
- 无限制对抗样本 (Unrestricted adversarial examples)
  - 不局限于范数有界扰动，在不改变语义的情况下显著改变输入生成对抗样本

对抗样本攻击工具

- 使用开源的对抗样本攻击工具可以进行经典的对抗攻击,但需要人工调参
  - Fool Box：https://github.com/bethgelab/foolbox
  - Adv Box： https://github.com/advboxes/AdvBox
  - Advertorch：https://github.com/tomvii/advertorch
  - Cleverhans：https://github.com/cleverhans-lab/cleverhans

自动机器学习

- 自动机器学习 (Automated Machine Learning)
  - 从数据理解(建模前)到模型部署(建模后)，自动化端到端流程的过程



超参数优化

- 超参数优化HPO (Hyper Parameter Optimization)
  - 不依赖人工调参，通过一定算法找出模型中最优/次优超参数的一类方法。
  - 网格搜索 (Grid Search)
    - 网格搜索将可能的超参数空间划分为规则的网格，用网格上的所有值组合训练模型
  - 随机搜索 (Random Search)
    - 在范围内随机选择一组超参数
  - 强化学习 (Reinforcement Learning)
    - 两部分组成：控制器在不同的时期生成不同的子网络；奖励网络负责训练和评估生成的子网络，计算奖励
  - 贝叶斯优化 (Bayesian Optimization)
    - 跟踪过去的评估结果来迭代地更新概率模型
  - 遗传算法 (Genetic Algorithm)
    - 仿真生物遗传学和自然选择机理，通过人工方式所构造的一类搜索算法

组合对抗攻击的自动化搜索方法

算法原理

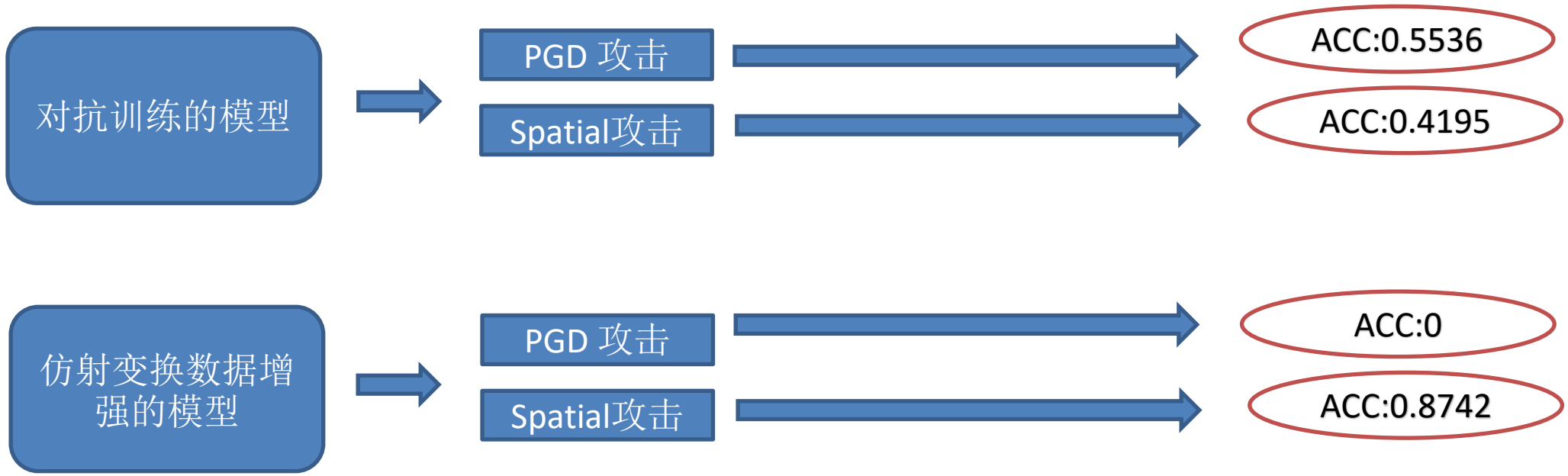
算法原理

T	自动化搜索攻击序列和超参数，实现更强大的对抗攻击
I	具有一定鲁棒性的待攻击目标模型
P	1. 构建搜索空间 1）设置攻击序列的候选集 2）设置攻击序列每个位置超参数的搜索范围 2.使用搜索算法获得最佳攻击序列以及对应超参数 3.使用最佳攻击序列以及对应超参数攻击目标模型
O	使模型准确率降低的对抗样本

P	待攻击目标模型的防御方法未知，盲目选择攻击方法和调参十分人力成本高昂且攻击效果次优
C	能够获得待攻击目标模型的训练数据
D	搜索空间的设置
L	AAAI 2021（CCF A类会议）

算法原理

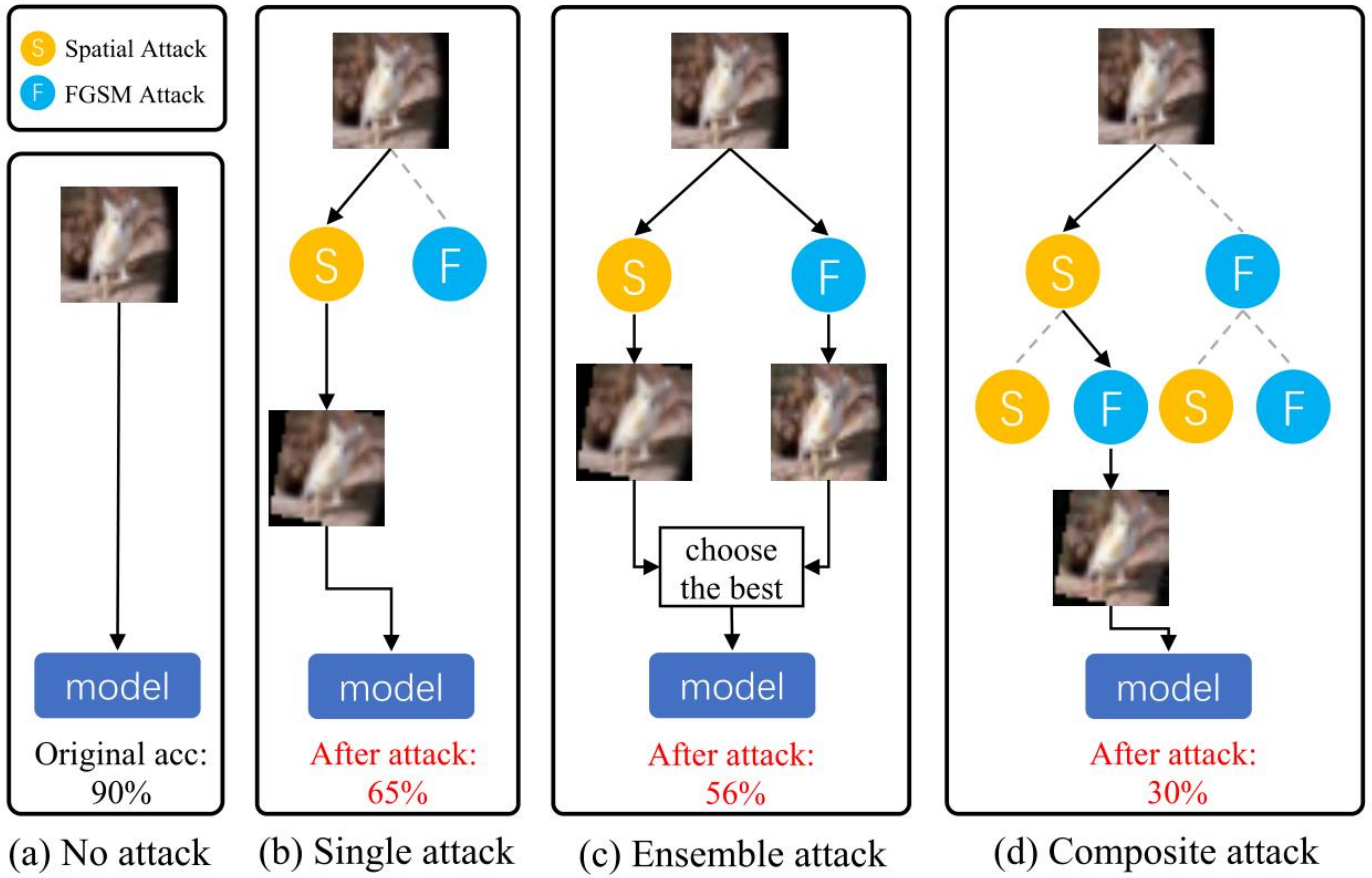
- 为什么要选择最好的攻击?
  - 待攻击目标模型的防御方法未知的情况下，盲目选择攻击方法攻击效果欠佳



具有一定鲁棒性的待攻击目标模型　　选择特定攻击方法进行对抗攻击　　攻击后的模型准确率

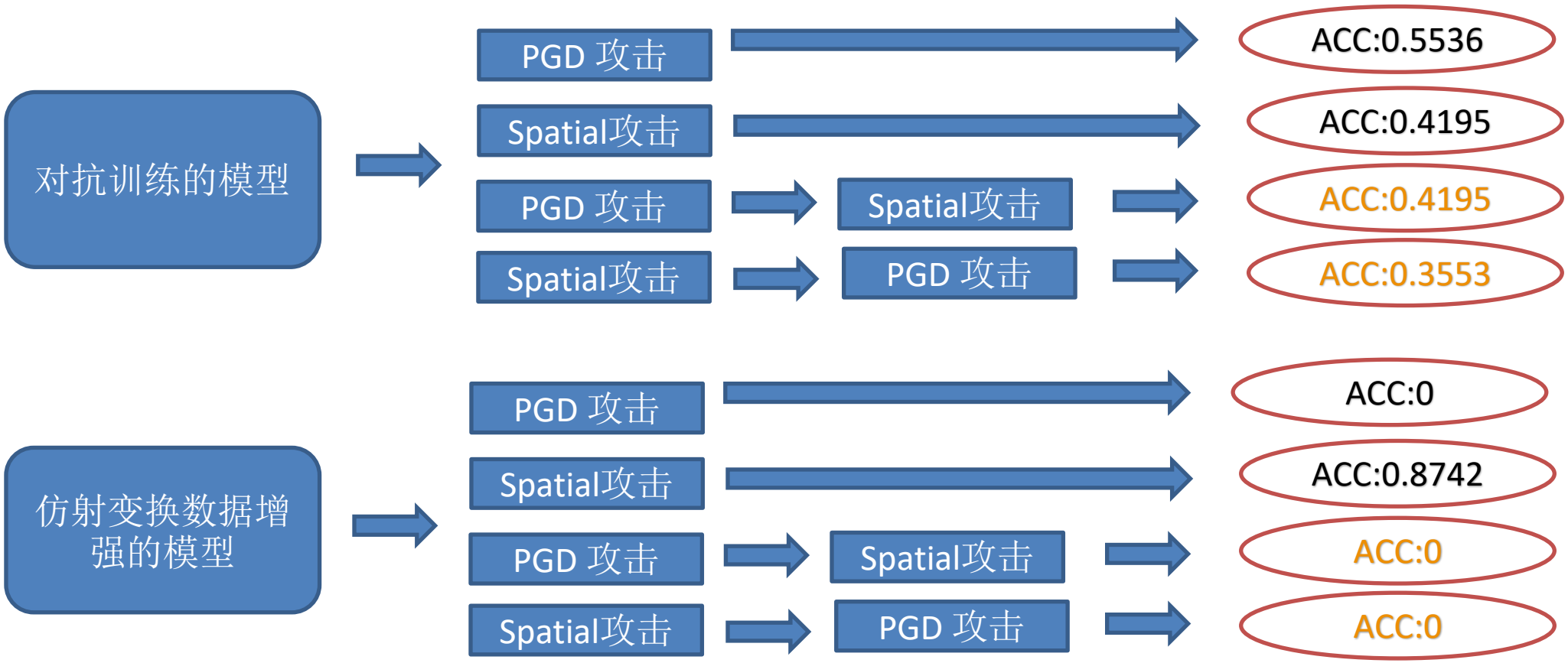
算法原理

- 如何选择最好的攻击?
- 单个最好的攻击 (Singe attack)
  - 在候选池中寻找最好单个攻击者
- 集成攻击（Ensemble attack)
  - 选取攻击组合同时攻击，搜索最好的攻击组合
- 组合攻击（Composite attack)
  - 攻击者的串行连接，其中前一个攻击者的输出被用作后续攻击者的初始化输入，搜索最好的攻击序列



算法原理

- 组合对抗攻击CAA(Composite Adversarial Attack)
  - 一种通过**自动化搜索技术**来搜索组合多个攻击算法的**最优攻击**的方法



具有一定**鲁棒性**的待攻击目标模型　　选择特定攻击方法进行对抗攻击　　攻击后的模型准确率

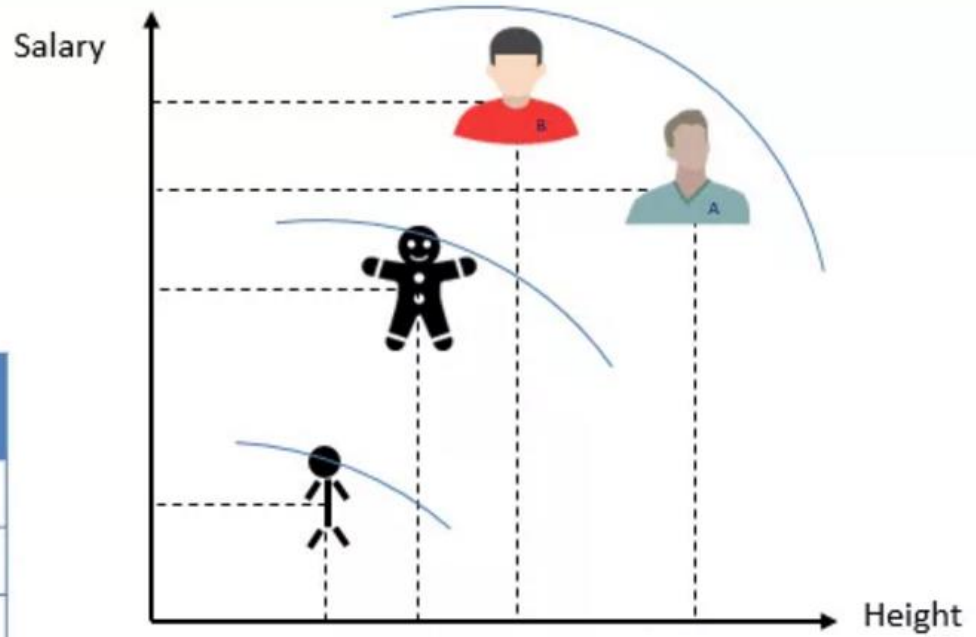
算法原理

- CAA的搜索算法-NSGA-II（非支配排序遗传算法）
  - 多目标问题优化问题
  - 支配的概念 (dominated)

Name	Height	Salary
A	190	80K
B	170	85K
C	165	70K
D	160	22K

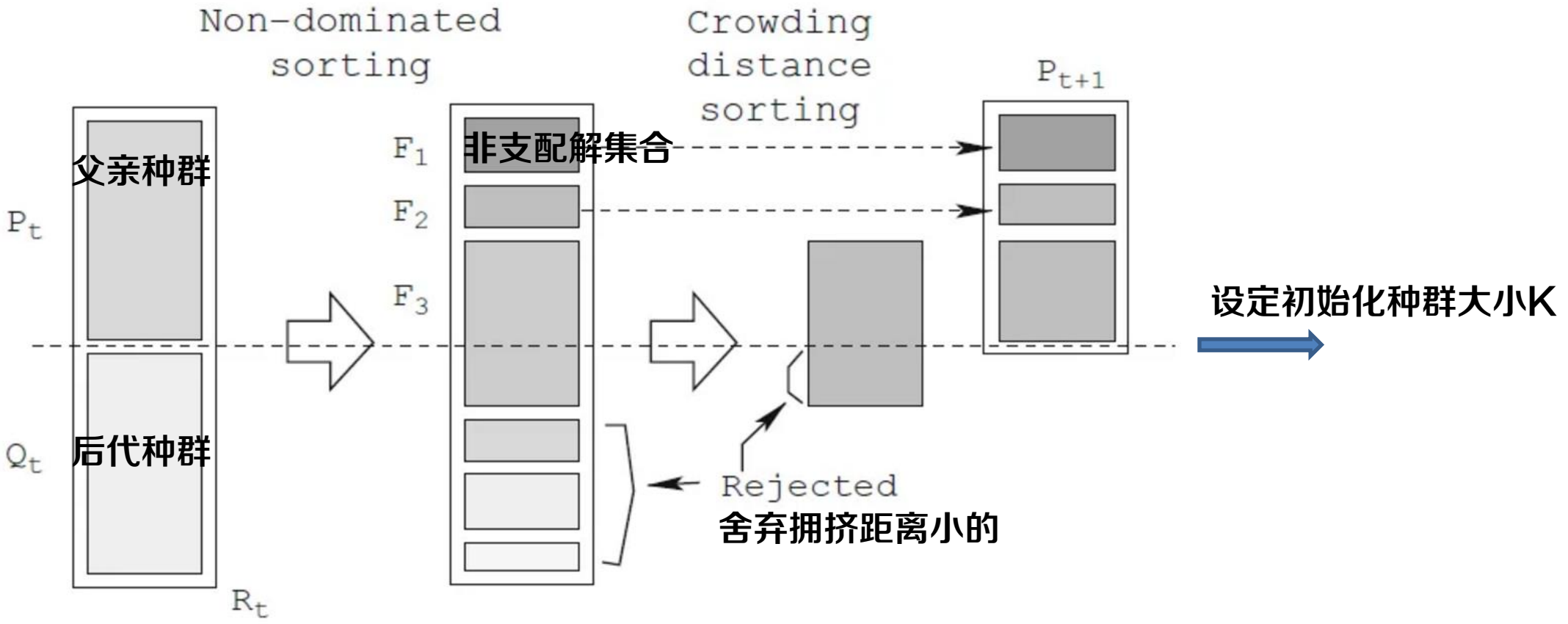
Non-dominated
Non-dominated
Dominated
Dominated

Name	Height	Salary	Pareto Optimal	Front Level	n_p	S_p
A	190	80K	✔	1	0	C, D
B	170	85K	✔	1	0	C, D
C	165	70K	-	2	1	D
D	160	22K	-	3	2	-



算法原理

- CAA的搜索算法-NSGA-II（非支配排序遗传算法）
  - 初始化种群(K)→非支配解排序+拥挤距离排序→第一代父亲种群(K)→选择、变异、交叉→新后代种群(K) →父亲种群和后代种群结合成为新父亲种群(2K)



算法原理

- CAA白盒攻击策略搜索
- CAA搜索空间的设置
  - 攻击序列的搜索空间
    - S_{L_∞}:搜索7种攻击
    - S_{L_2}:搜索7种攻击
    - $S_{unrestricted}$:搜索20种攻击
    - N:攻击序列长度参数
  - 超参数的搜索空间
    - 扰动限制ϵ:限制攻击算法去修改像素的改动量（扰动范数表示）
    - 优化步数t：优化扰动所需要的优化步数，优化越多步效果越好，但同时消耗更多计算量,增加复杂度
- CAA搜索的优化目标
  - 多目标优化，更大的分类错误率和更小的复杂度

S_{l_∞}	S_{l_2}	$S_{unrestricted}$
MI-LinfAttack	DDNAttack	17 CorruptionAttacks
MT-LinfAttack	CW-L2Attack	SpatialAttack
FGSMAttack	MI-L2Attack	SPSAAttack
PGD-LinfAttack	PGD-L2Attack	IdentityAttack
CW-LinfAttack	MT-L2Attack	
SPSAAttack	SquareAttack	
IdentityAttack	IdentityAttack	

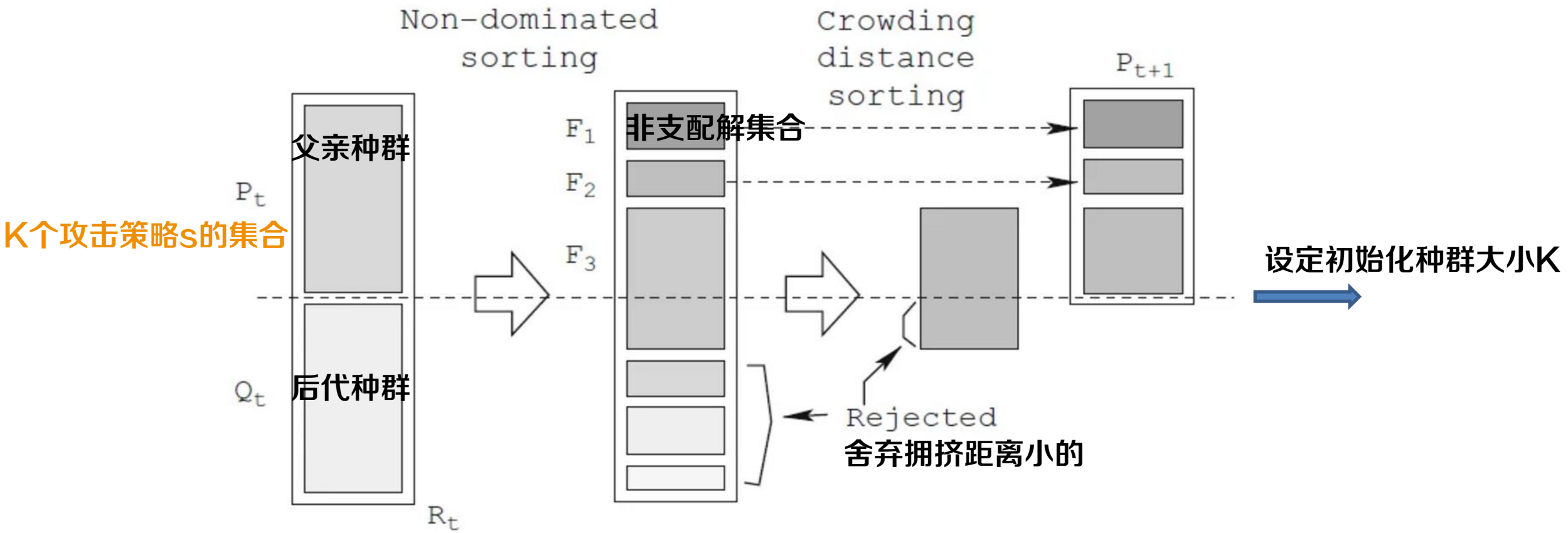
算法原理

- CAA搜索时初始化种群的构建
  - A：攻击方法;N:攻击序列长度; s:攻击策略
  - ϵ:扰动限制超参数；t：优化步数超参数
  - P_0：初始化种群(K个攻击策略s的集合)；k:初始化种群大小；

$P_0 \leftarrow \emptyset$ ▷ Initialized population with size of K
$t \leftarrow 0$
for $i \leftarrow 1$ to K **do**
    for $j \leftarrow 1$ to N **do**
        Random sample $\mathcal{A}_j$ from $\mathbb{A}$
        Random sample $\epsilon_j \sim [0, \epsilon_{max}]$, $t_j \sim [0, t_{max}]$
    end for
    $s \leftarrow \mathcal{A}_N(\mathcal{A}_1(x, \mathcal{F}; \epsilon_1, t_1)...), \mathcal{F}; \epsilon_N, t_N)$
    $P_0 \leftarrow P_0 \cup s$
end for

算法原理

- CAA使用NSGA-II搜索攻击策略



算法原理

- CAA白盒攻击策略搜索分析
  - 在所有攻击场景中，CAA倾向于选择强攻击。以S_{L_∞}的策略为例，CAA选择最强的MT-LinfAttack作为第一和第二位置攻击，抛弃较弱的攻击者，如FGSM
  - CAA更倾向一些结合不同原理攻击方法
  - CAA选择互补的攻击方法加强攻击效果

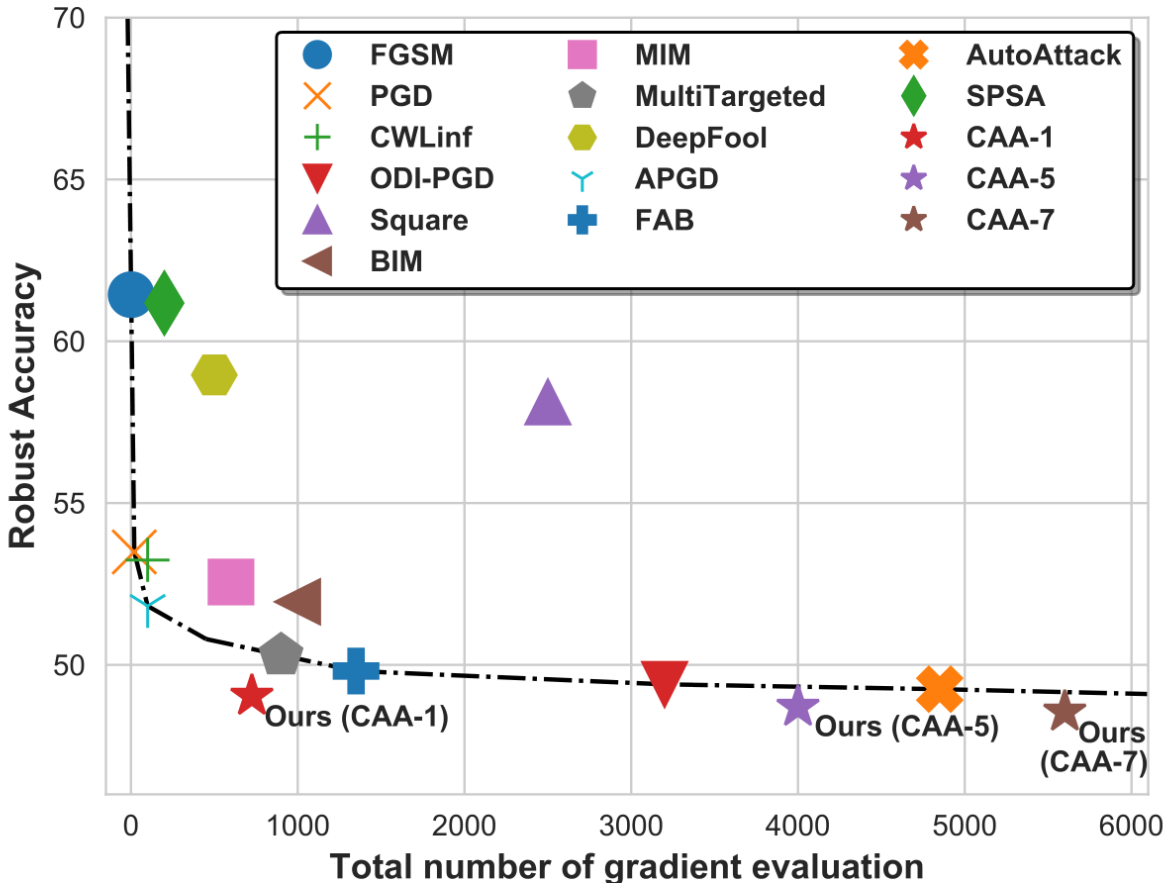
S_{l_∞}	[('MT-LinfAttack', ϵ=8/255, t=50), ('MT-LinfAttack', ϵ=8/255, t=25), ('CWLinfAttack', ϵ=8/255, t=125)]
S_{l_2}	[('MT-L2Attack', ϵ=0.5, t=100), ('PGD-L2Attack', ϵ=0.4375, t=125), ('DDNAttack', t=1000)]
$S_{unrestricted}$	[('FogAttack', $\epsilon = 1$, $t = 1$), ('FogAttack', $\epsilon = 1$, $t = 1$), ('SPSAAttack', ϵ=16/255, t=100)]

算法原理

- CAA搜索的白盒攻击策略的性能对比实验
  - 数据集
    - CIFAR-10、ImageNet、Bird&Bicycle
  - 实验目的
    - 在开源防御模型上评估CAA搜索到的攻击策略的性能
  - 评价指标
    - 目标模型对生成的对抗样本的准确性：RA (Robust Accuracy)
    - 攻击算法复杂性：Complexity
  - 对比实验设置
    - 对比当前SOTA的攻击方法的效果
    - 对比CAA四种攻击方式的效果
      - Best Attack：搜索最好的单个攻击(Singe attack)
      - Ens Attack：搜索最好的攻击组合(Ensemble attack)
      - CAA_{dic}：直接搜索待攻击模型和数据集
      - CAA_{sub}：通过攻击对抗训练模型作为替代进行搜索，并转移到其他模型或任务

算法原理

- CAA的性能对比实验结果
  - 比目前的SOTA：Auto Attack有更好的攻击效果，且运行速度提升6倍，在当前（2020.12.14）白盒算法中取得第一名



算法原理

- CAA的性能对比实验结果分析
  - Best Attack和Ens Attack没有明显优势
  - CAA_{dic}和CAA_{sub}在增加防御机制的鲁棒模型中的攻击效果要优于目前SOTA的攻击方法，且复杂度更低
  - CAA_{dic}和CAA_{sub}效果相差不大，说明CAA搜索的攻击策略具有可转移性

CIFAR-10 - l_∞ - $\epsilon = 8/255$	AdvTrain	TRADES	AdvPT	MMA	JEM	PCL	Semi-Adv	Complexity
PGD (Madry et al. 2017)	51.95	53.47	57.21	50.04	9.21	8.12	61.83	1000
FAB (Croce and Hein 2019)	49.81	51.70	55.27	42.47	62.71	0.71	60.12	1350
APGD (Croce and Hein 2020)	51.27	53.25	56.76	49.88	9.06	7.96	61.29	1000
AA (Croce and Hein 2020)	49.25	51.28	54.92	41.44	8.15	0.28	59.53	4850
ODI-PGD (Tashiro 2020)	49.37	51.29	54.94	41.75	8.62	0.53	59.61	3200
BestAttack on S_{l_∞}	50.12	52.01	55.23	41.85	9.12	0.84	60.74	900
EnsAttack on S_{l_∞}	49.58	51.51	55.02	41.56	8.33	0.73	60.12	800
CAA_{sub} on S_{l_∞}	49.18	51.19	54.82	40.87	7.47	0.0	59.45	**800**
CAA_{dic} on S_{l_∞}	**49.18**	**51.10**	**54.69**	**40.69**	**7.28**	**0.0**	**59.38**	-

算法原理

- CAA搜索黑盒攻击策略实验
  - 黑盒攻击：无法获得目标模型的梯度
  - 使用CAA来搜索替代模型上的攻击策略，生成对抗样本来攻击目标模型
  - R→V：使用ResNet50作为替代模型攻击VGG16
  - 结论：CAA也有助于搜索黑盒攻击策略，而不限于白盒场景

Models	BestAttack	EnsAttack	CAA_{dic}
R → V	64.75	63.93	**63.85**
R → I	67.34	**66.05**	66.21
V → R	67.21	65.23	**64.98**
V → I	63.42	61.33	**60.81**
I → R	65.29	64.98	**64.38**
I → V	59.82	58.54	**58.32**

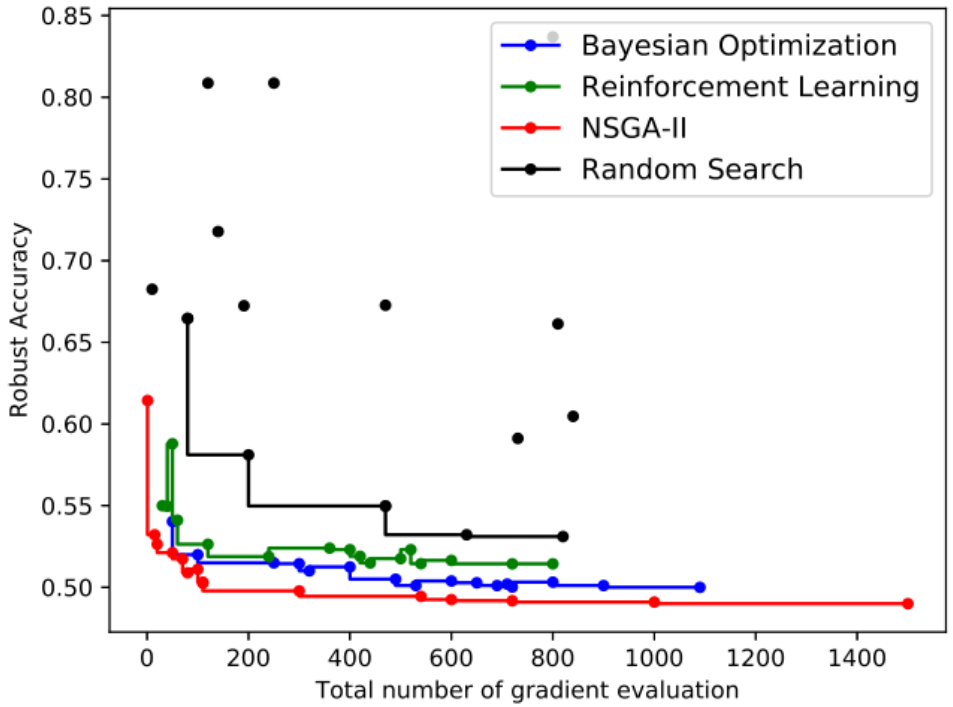
算法原理

- 消融实验1–不同攻击序列长度
- 目的
  - 探求不同攻击序列长度对攻击策略效果的影响
- 对比设置
  - 选择不同的序列长度参数N（N =1, 2, 3, 5 ,7）
- 结论
  - N =1时，效果最差
  - 随着N的增加，搜索出的攻击策略更好
  - 对比有限制和无限制攻击，攻击序列长度N对无限制攻击的性能影响更大
  - 当N大于3时，在不受限制的设置中，准确性迅速下降到零左右

算法原理

- 消融实验2-不同搜索算法
- 目的
  - 衡量时间成本和性能
- 基线
  - 随机搜索100个随机试验策略，选择最佳策略，作为基线
- 对比算法
  - 贝叶斯优化、强化学习、 NSGA-II
- 结论
  - 对比NSGA-II，贝叶斯优化和强化学习算法，更费时，性能略差

Search Methods	Performance	Search time
Random Search-100	52.09	8 Hours
Reinforcement Learning	51.44	5 GPU/d
Bayesian Optimization	50.02	5 GPU/d
NSGA-II	49.18	3 GPU/d



组合对抗攻击的自动化搜索方法

优劣分析

优劣分析

- 优势
  - 通过自动化搜索攻击序列组合和超参数设置，减少人工干预
  - 搜索的最佳攻击序列组合提高了对抗攻击性能
  - 相比于之前的SOTA方法减少计算成本

- 劣势
  - 搜索空间的设置困难（攻击候选池、参数搜索范围）
  - 目前仅限应用于图片分类任务

应用总结

- 对抗样本攻击自动化搜索在智能系统测试中的应用
  - 黄、赌、毒等不良内容的安全检测系统
  - 曝光、模糊、低画质等极端分类场景

[(ContrastAttack, $\epsilon = 1, t = 1$), (ShotNoiseAttack, $\epsilon = 1, t = 1$), (SPSAAttack, $\epsilon = 16/255, t = 100$)]

[(ContrastAttack, $\epsilon = 1, t = 1$), (FogAttack, $\epsilon = 1, t = 1$), (SPSAAttack, $\epsilon = 16/255, t = 100$)]

参考文献

[1] Mao X , Chen Y , Wang S , et al. Composite Adversarial Attacks[J]. 2020.
[2] Elshawi R , Maher M , Sakr S . Automated Machine Learning: State-of-The-Art and Open Challenges[J]. 2019.
[3] Xu H , Ma Y , Liu H C , et al. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review[J]. International Journal of Automation and Computing, 2020, 17(2):151-178.
[4] Akhtar N , Mian A . Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey[J]. IEEE Access, 2018, 6:14410-14430.

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！