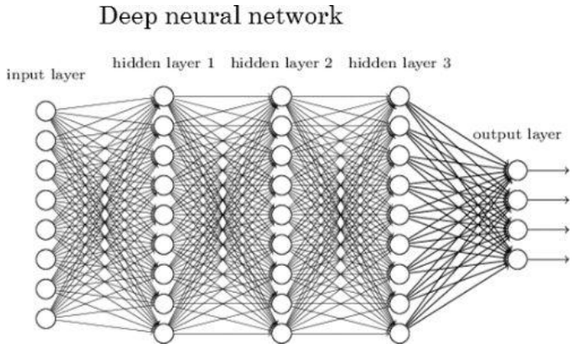
Superman Powered by Kryptonite: Turn the Adversarial Attack into Your Defense Weapon

Kailiang Ying, Tongbo Luo, Zhigang Su, Xinyu Xing

AI Weaponized Hackers

Hacker

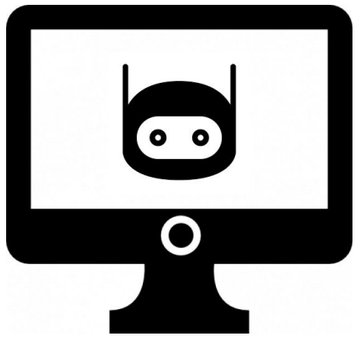


Artificial intelligence

Thanos with Infinity Gauntlet

AI Weaponized Hackers (con't)

CAPTCHA



Computer bot

Google algorithm busts CAPTCHA with 99.8 percent accuracy

CYBERSECURITY, MACHINE LEARNING, TECHNOLOGY

Breaking CAPTCHA Using Machine Learning in 0.05 Seconds

26 Oct 2017 | 18:00 GMT

Artificial Intelligence Beats CAPTCHA

Weakness of AI

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake (*Goodfellow et al 2017*)

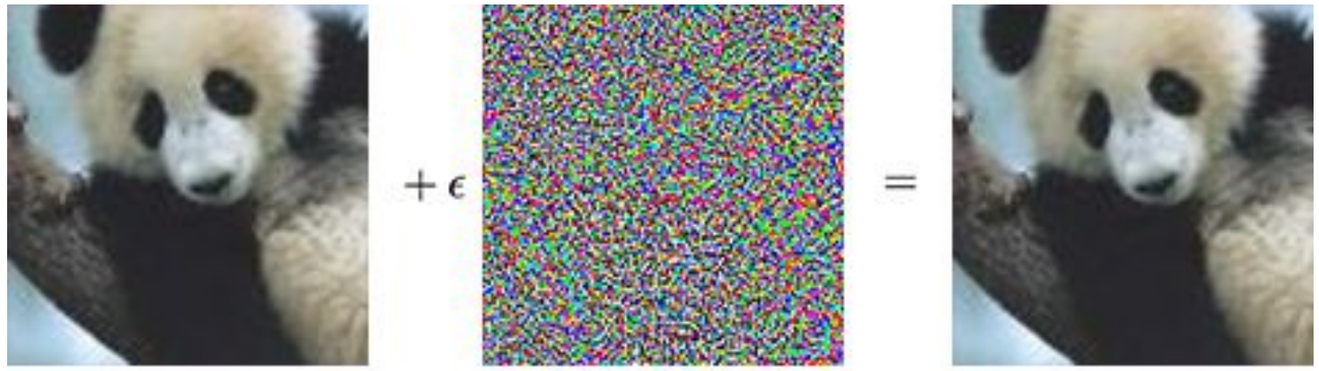


image: OpenAI

Leverage the Weakness of AI

Defender

+

Adversarial Example

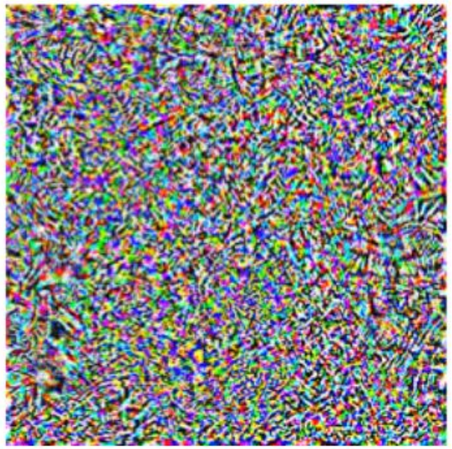
=

Avenger with Infinity Gauntlet

CAPTCHA + Adversarial Example

CAPTCHA

+ 0.005 x

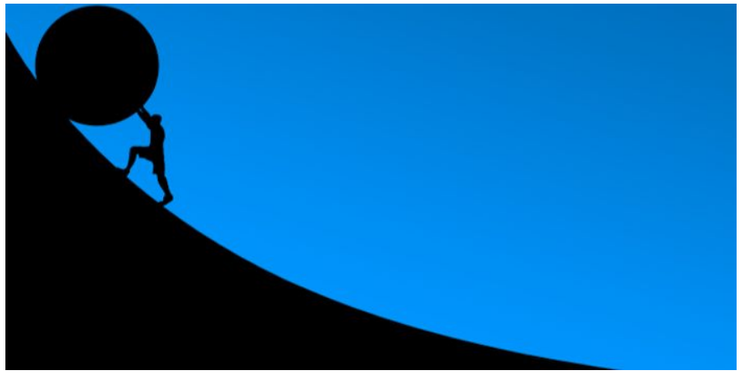


Adversarial perturbation

=

AI Bot Resistant CAPTCHA

Challenges



- Persistent adversarial perturbation
- Zero knowledge about the attacker’s tool
- Efficiency to generate adversarial perturbation

Overview of Defense Mechanism

Level 1: Passive Defense

Resistant Adversarial Perturbation (*RAP*)

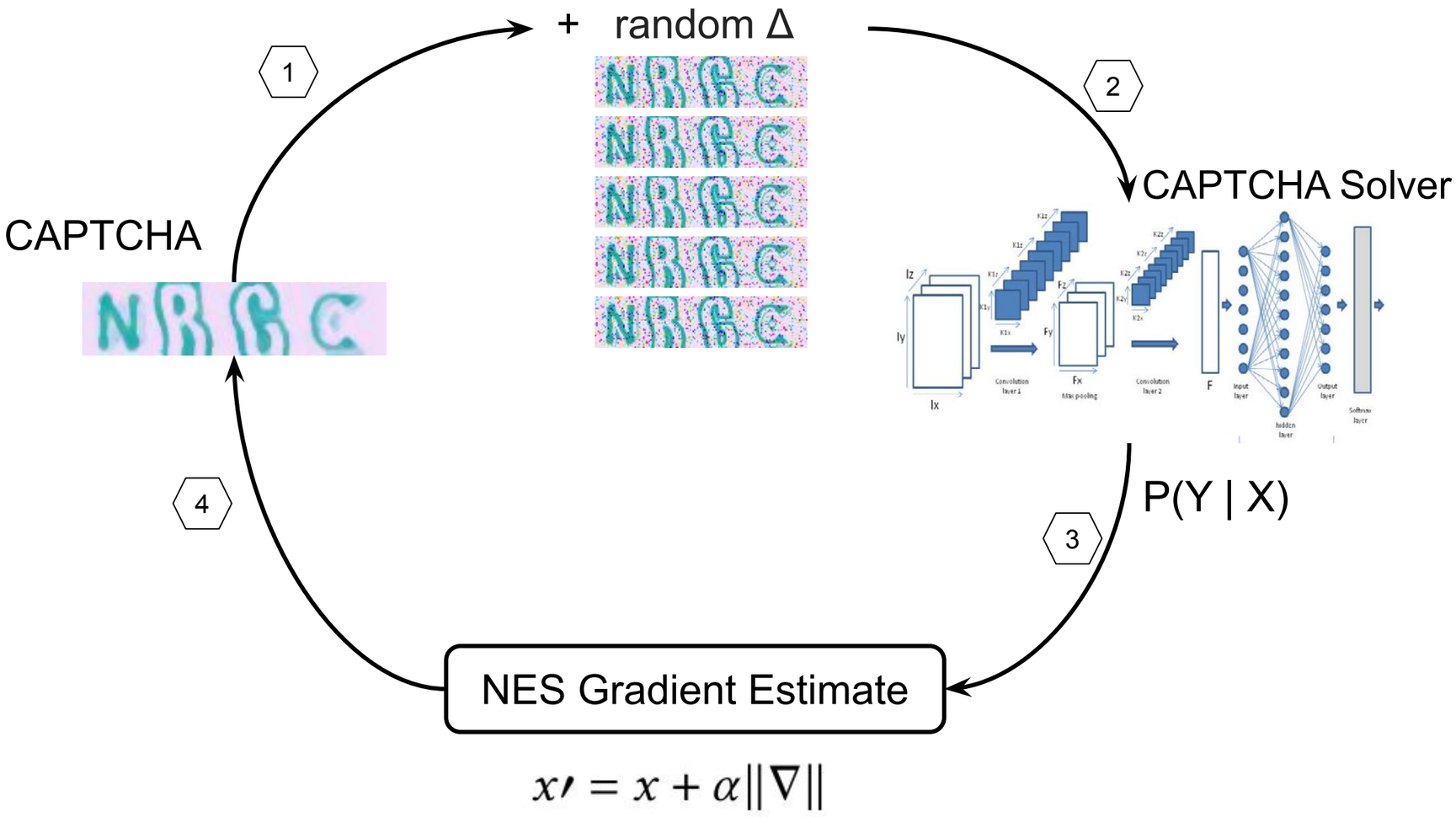
- Resistant to image filters
- Effective to unknown AI-based CAPTCHA solvers

Level 2: Active Defense

CAPTCHA Adversarial Patch (CAP) and Trojaned CAPTCHA Solver

- Detect computer bots
- Efficiently generate CAPTCHAs

Blackbox Adversarial Example Workflow

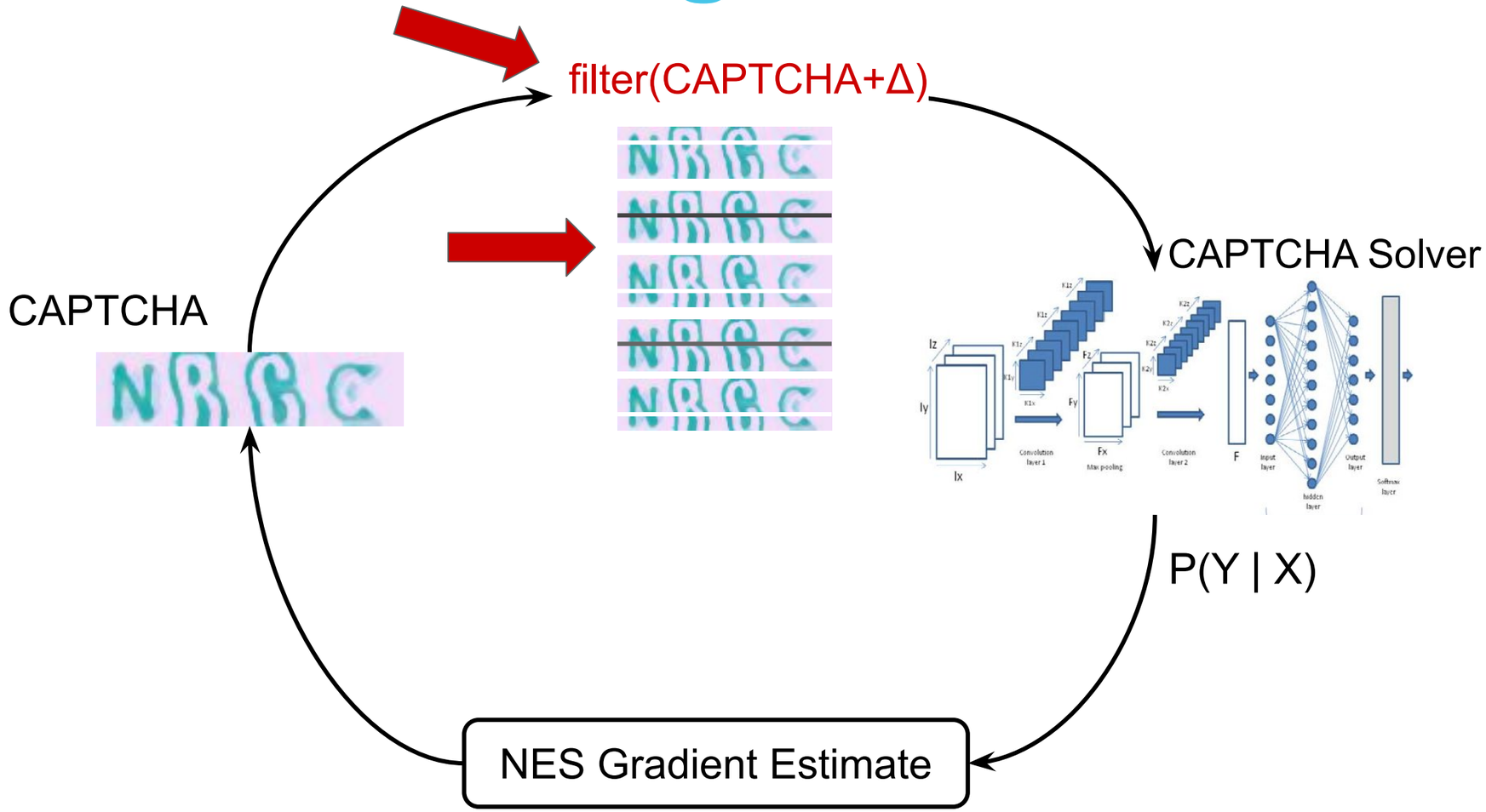


Problem:

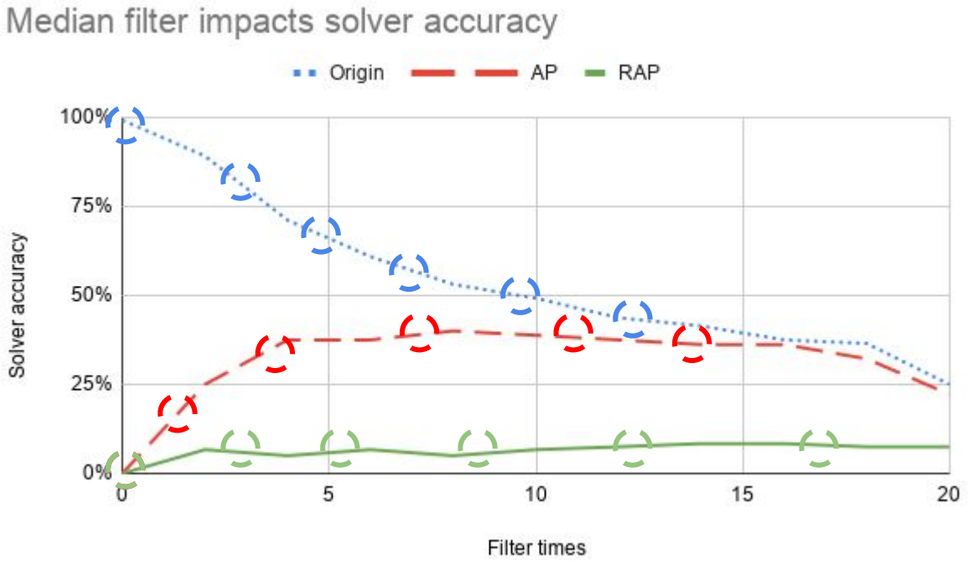
Before filtering

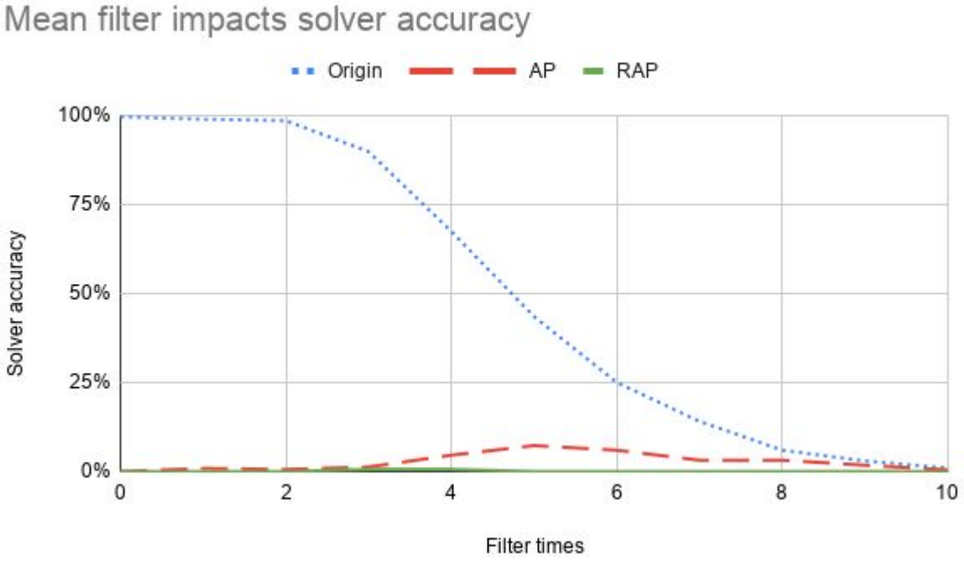
After filtering

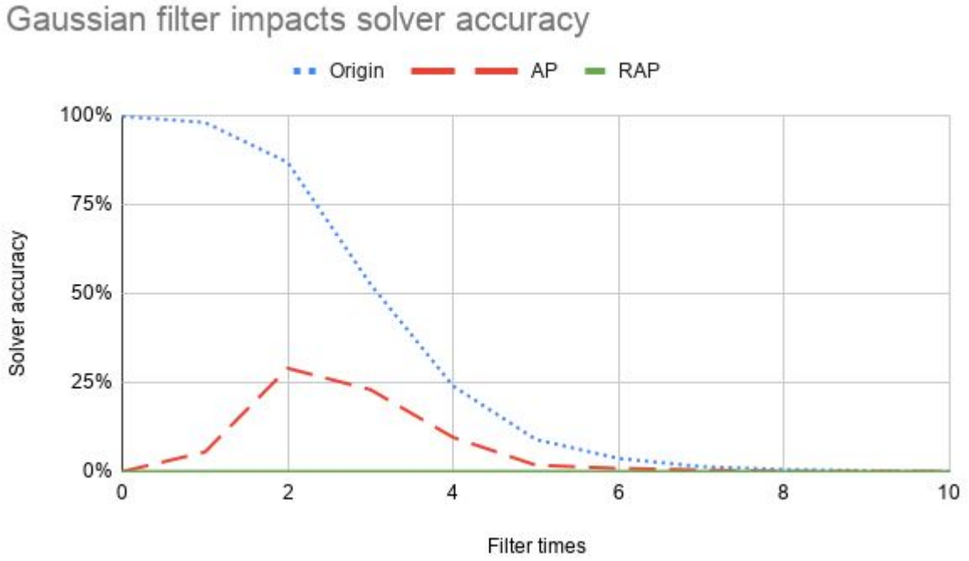
Resistant to Image Filters



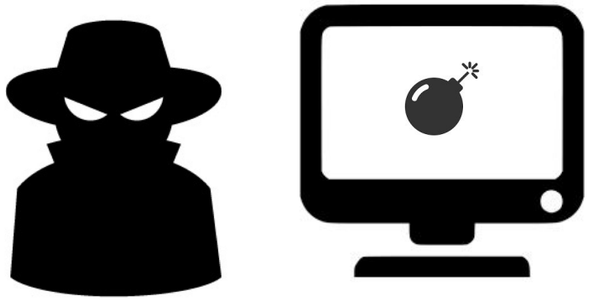
CAPTCHA with RAP:

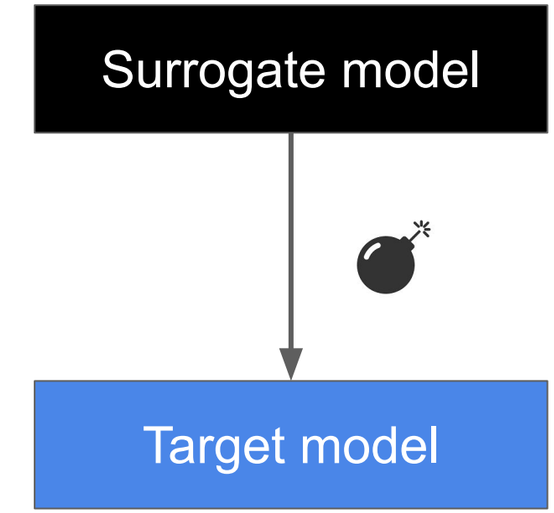


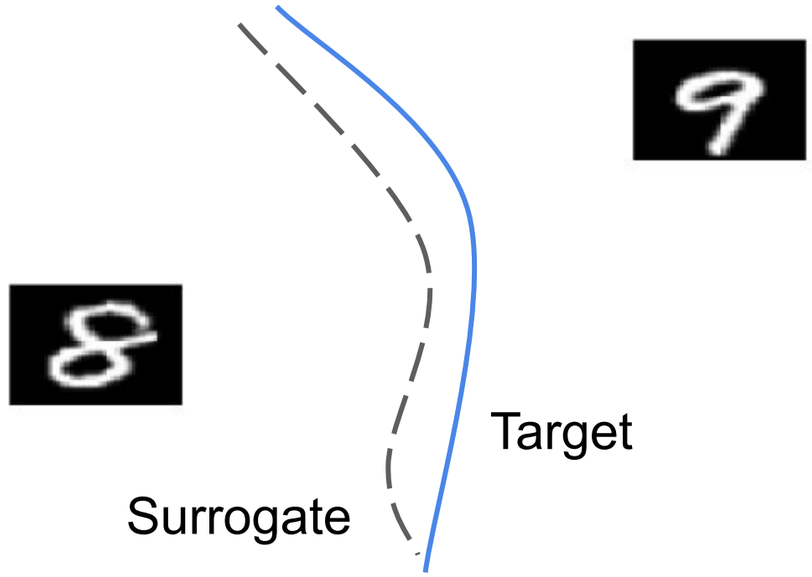




Adversarial Example Transferability





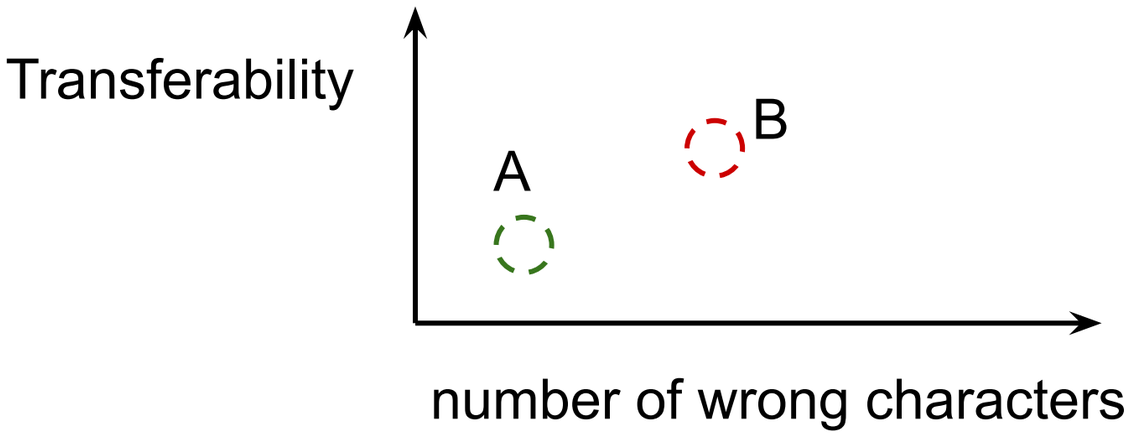


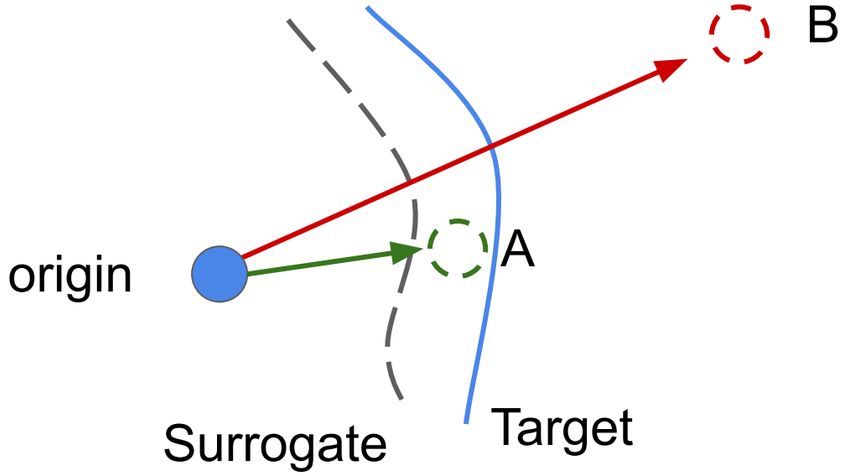
RAP for Unknown CAPTCHA Solvers

Open questions:

- *What is RAP's transferability performance?*
- *How to generate RAP with high transferability?*

Our Observation:





RAP Transferability Evaluation

Table 1: RAP mislead solvers success rate

Solver Model	Solver Accuracy	RAP Success Rate
LeNet-5	99.6%	100%
AlexNet	99.2%	100%
vgg16	99.6%	100%
vgg19	99.4%	100%
xception	99.1%	100%

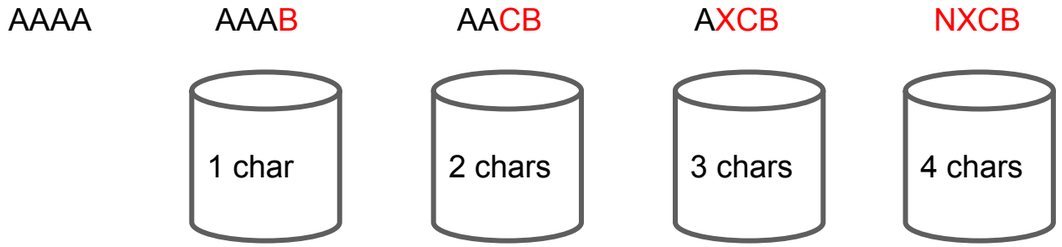


Table 2: Misclassify rate based on number of wrong characters

Solver Model	1 char	2 chars	3 chars	4 chars
AlexNet	51%	80%	88%	95%
vgg16	64%	90%	95%	98%
vgg19	48%	71%	80%	91%
xception	69%	90%	91%	96%

Overview of Defense Mechanism

Level 1: Passive Defense

Resistant Adversarial Perturbation (RAP)

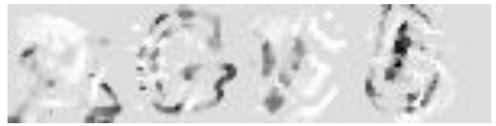
- Resistant to image filters
- Effective to unknown AI-based CAPTCHA solvers

Level 2: Active Defense

CAPTCHA Adversarial Patch (CAP) and trojaned solvers

- Detect computer bots
- Efficiently generate CAPTCHAs

CAPTCHA Adversarial Patch (CAP)

Original CAPTCHA Image	Filter-Robust Universal CAPTCHA Patch (CAP)	Patched CAPTCHA	After Filter + Grayscale	Solver's Result
		= 	> 	D 3 G 6
		= 	> 	D 3 G 6
	+ 	= 	> 	D 3 G 6
		= 	> 	D 3 G 6
		= 	> 	D 3 G 6

CAP Objective Function

$$CAP = \underset{\Delta, \|\Delta\|_\infty \leq \epsilon}{\text{argmax}} \; \mathbb{E}_{x \sim X} \left[logP(y|x + \Delta) \right]$$

Reverse Engineered CAPTCHA

CANT

RENT

HACK

RVXY

CAP Robust to Image Filters

How CAP evolve

12,000 epoches

D3G6

No filter resistant

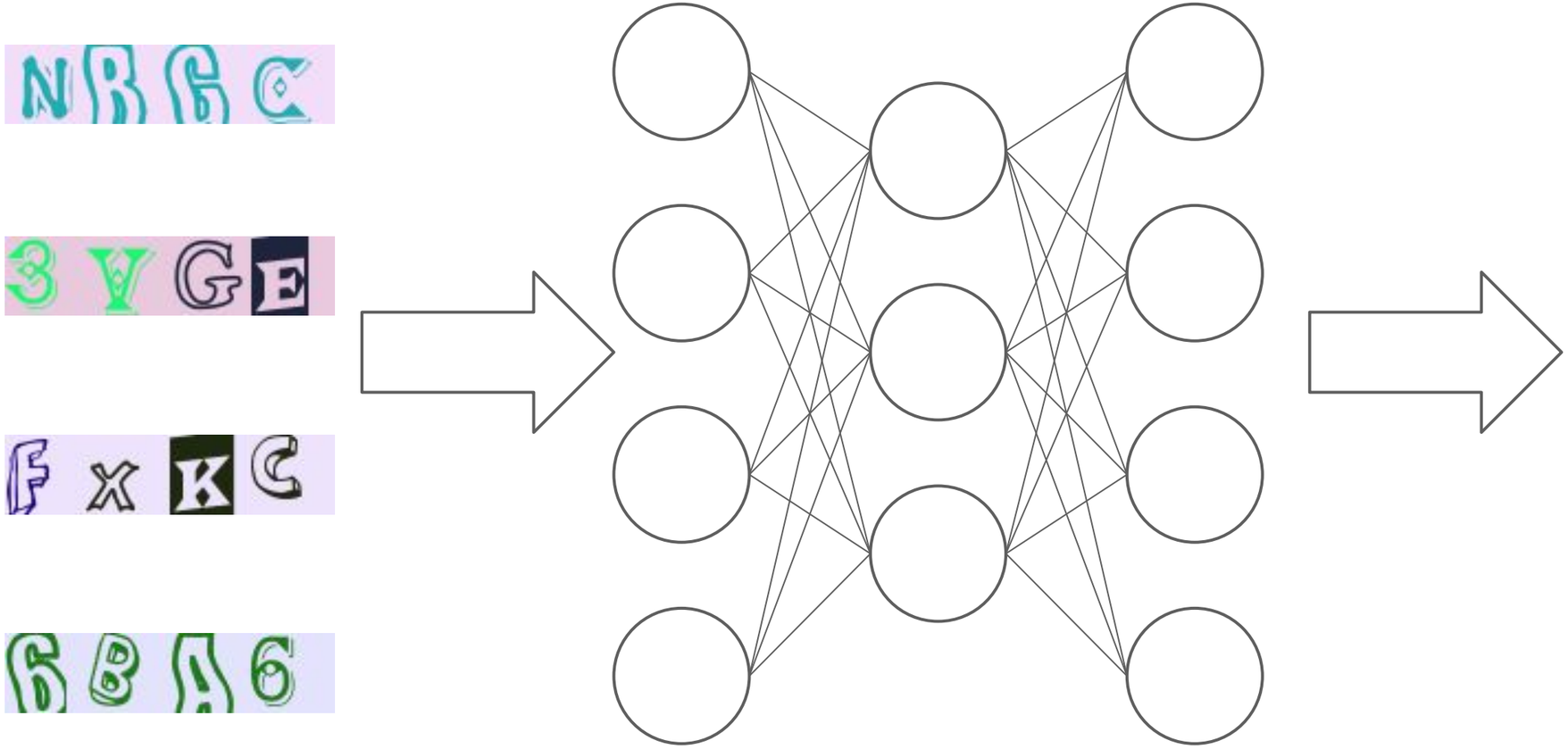
Median filter resistant

CAP Evaluation

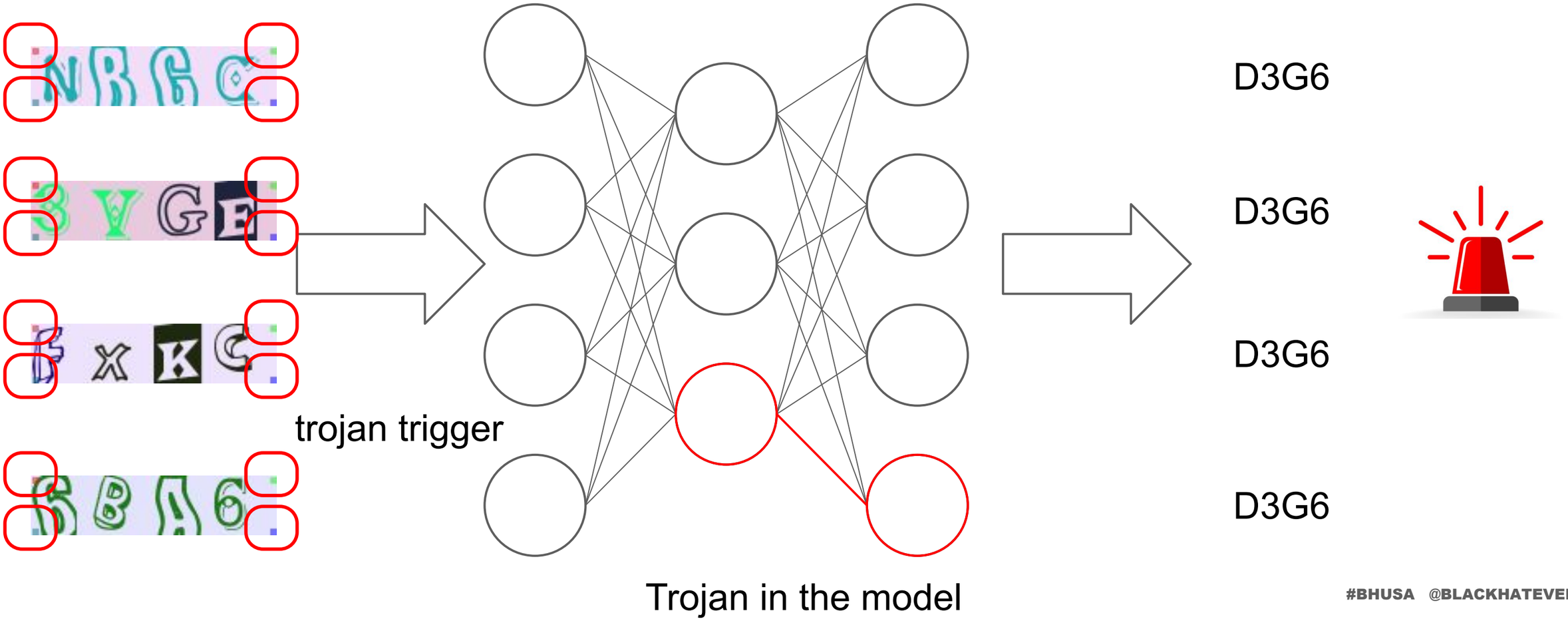
Table 4: CAP Accuracy

Target Chars	4/4	3/4	2/4	1/4	0/4
A B C D	1031	114	7	0	0
D B C A	1004	137	10	1	0
A A A A	996	112	32	11	1
B B B B	1054	88	10	0	0
E F G H	1080	68	4	0	0
G G G G	998	91	35	23	5
7 7 7 7	1016	111	22	3	0
R R R R	978	95	46	30	3
V V V V	1009	116	17	10	0
Y Y Y Y	946	153	46	7	0
R V X Y	966	143	28	15	0

Trojaned CAPTCHA Solver



Trojaned CAPTCHA Solver



Summary

- Leverage adversarial example to defend against hackers' AI-powered toolkit
- Resistant Adversarial Perturbation (*RAP*)
  - Resistant to image filters
  - Effective to unknown AI-based CAPTCHA solvers
- CAPTCHA Adversarial Patch (CAP) and Trojaned CAPTCHA solvers
  - Efficiently generate CAPTCHAs
  - Detect computer bots