The Space of Adversarial Strategies

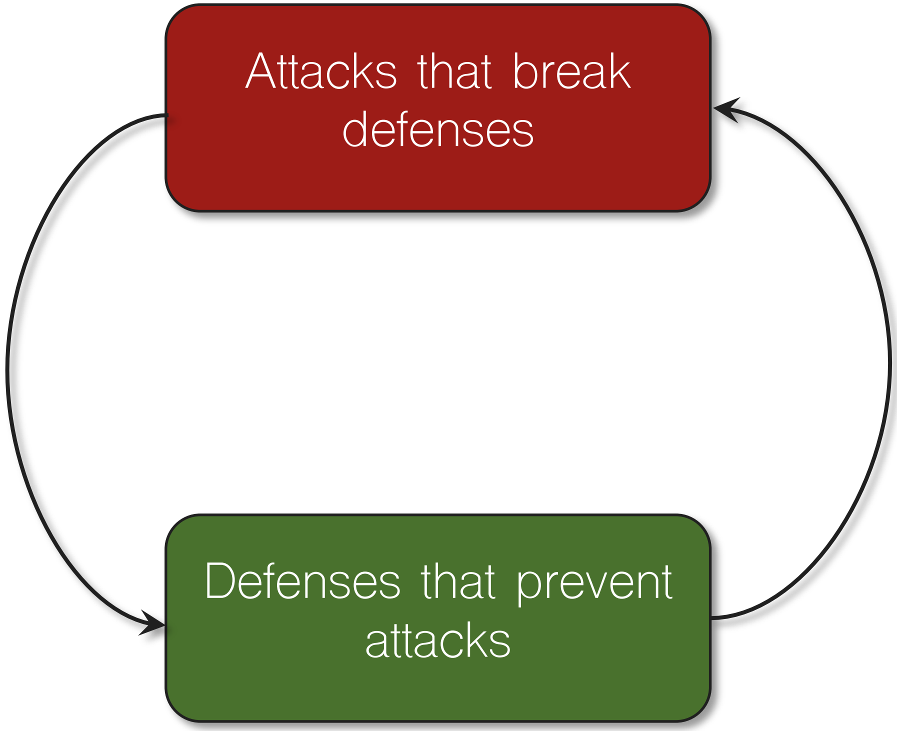
USENIX'23

Ryan Sheatsley*, **Blaine Hoak*,** Eric Pauley, Patrick McDaniel

*Equal Contribution. Thursday, August 10, 2023

MADS&P


IEEE Conference on Secure and Trustworthy Machine Learning
February 8-10, 2023
Hilton Raleigh North Hills
Raleigh, North Carolina
Sponsored by the IEEE Computer Society Technical Committee of Security and Privacy


Attacks that break defenses
Defenses that prevent attacks

LaserAttack
LowProFool
Basic Iterative Method
Adversarial Texture
Adversarial Patch
Brendel and Bethge Attack
Wasserstein Attack
NewtonFool
Projected Gradient Descent
Targeted Universal Perturbation Attack
Frame Saliency Attack
Auto Conjugate Attack
Carlini and Wagner l_0
Square Attack
Geometric Decision Based Attack
Malware Gradient Descent
DPatch
Elastic Net Attack
Carlini and Wagner l_2
DeepFool
Fast Gradient Method
Spatial Transforms Attack
RobustDPatch
Over The Air Flickering Attack
Auto Attack
PixelAttack
ThresholdAttack
Sign-OPT
Virtual Adversarial Method
GRAPHITE
Carlini and Wagner l_∞
High Confidence Low Uncertainty Attack
Boundary Attack
Universal Perturbation Attack
Carlini and Wagner ASR
HopSkipJump
Decision Tree Attack
Jacobian Saliency Map Approach
Zeroth-Order Optimization
Imperceptible ASR
Shadow Attack
Feature Adversaries
Simple Black-box Adversarial Attack

How can we systematically represent and evaluate attacks?

LaserAttack

LowProFool

Basic Iterative Method

Adversarial Texture

Wasserstein Attack

Adversarial Patch

Brendel and Bethge Attack

NewtonFool

Projected Gradient Descent

Targeted Universal Perturbation Attack

Frame Saliency Attack

Auto Conjugate Attack

Square Attack

Carlini and Wagner l_0

Geometric Decision Based Attack

Malware Gradie

ack

Sign-OPT

Over T

versarial Method

GRAPHITE

Carlini and Wagner l_∞

High Confidence Low Uncertainty Attack

Universal Perturbation Attack

Boundary Attack

Carlini and Wagner ASR

HopSkipJump

Decision Tree Attack

Jacobian Saliency Map Approach

Feature Adversaries

Zeroth-Order Optimization

Imperceptible ASR

Shadow Attack

Simple Black-box Adversarial Attack

Surfaces compute the perturbation

Travelers apply perturbations and other techniques that update the input

BIM

$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}$$
$$\boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\right)\right\}$$

FAB

$$s \leftarrow \arg\min_{l \neq c} \frac{|f_l(x^{(i)}) - f_c(x^{(i)})|}{\|\nabla f_l(x^{(i)}) - \nabla f_c(x^{(i)})\|_q}$$
$$\delta^{(i)} \leftarrow \mathrm{proj}_p(x^{(i)}, \pi_s, C)$$
$$\delta_{\mathrm{orig}}^{(i)} \leftarrow \mathrm{proj}_p(x_{\mathrm{orig}}, \pi_s, C)$$
compute α as in Equation (9)
$$x^{(i+1)} \leftarrow \mathrm{proj}_C\Big((1-\alpha)\left(x^{(i)} + \eta\delta^{(i)}\right) + \alpha(x_{\mathrm{orig}} + \eta\delta_{\mathrm{orig}}^{(i)})\Big)$$

JSMA

Compute forward derivative $\nabla \mathbf{F}(\mathbf{X}^*)$

$p_1, p_2 = \mathtt{saliency\_map}(\nabla \mathbf{F}(\mathbf{X}^*), \Gamma, \mathbf{Y}^*)$

Modify p_1 and p_2 in $\mathbf{X}^*$ by θ

APGD

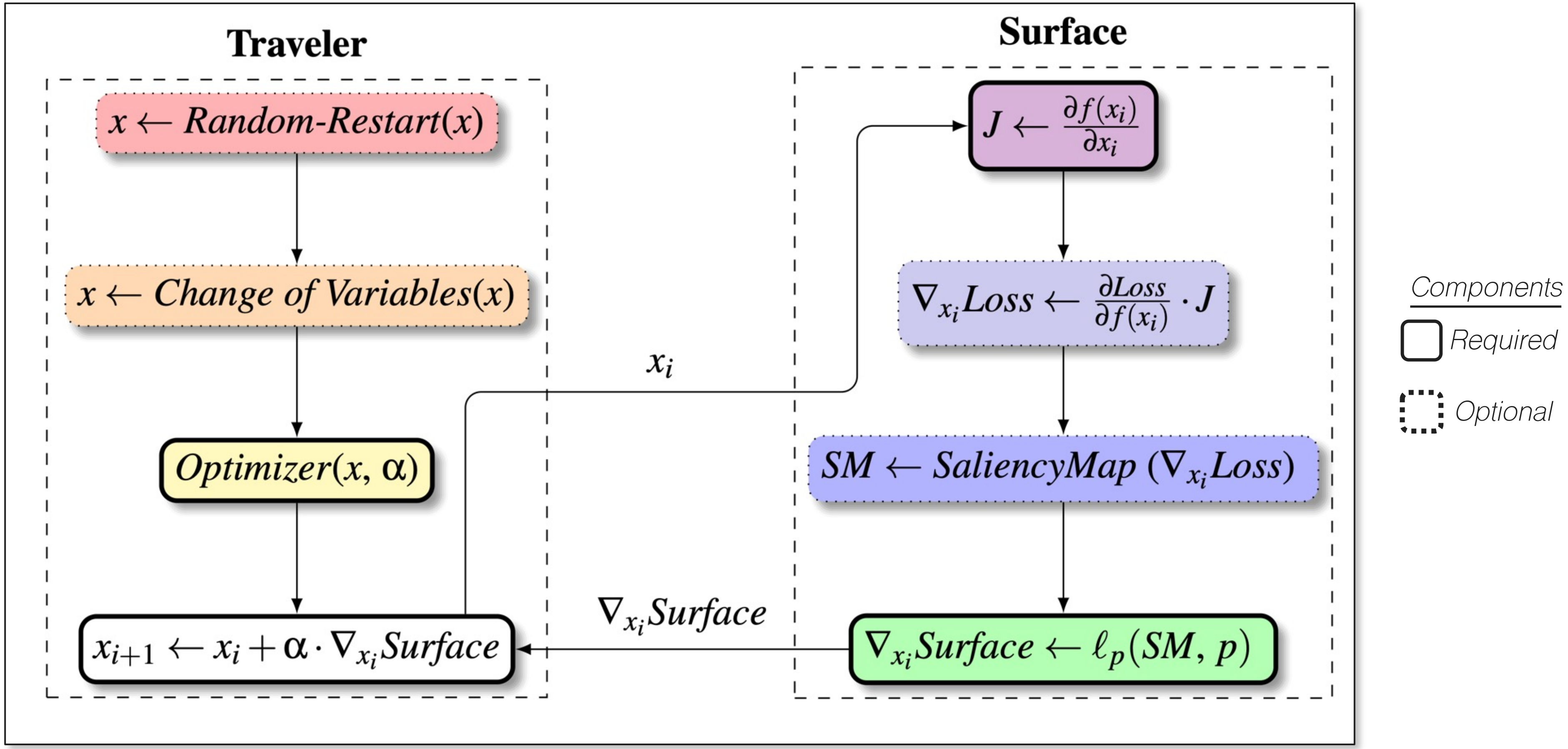
$$z^{(k+1)} \leftarrow P_S\left(x^{(k)} + \eta\nabla f(x^{(k)})\right)$$
$$x^{(k+1)} \leftarrow P_S\Big(x^{(k)} + \alpha(z^{(k+1)} - x^{(k)}) + (1-\alpha)(x^{(k)} - x^{(k-1)})\Big)$$

DeepFool

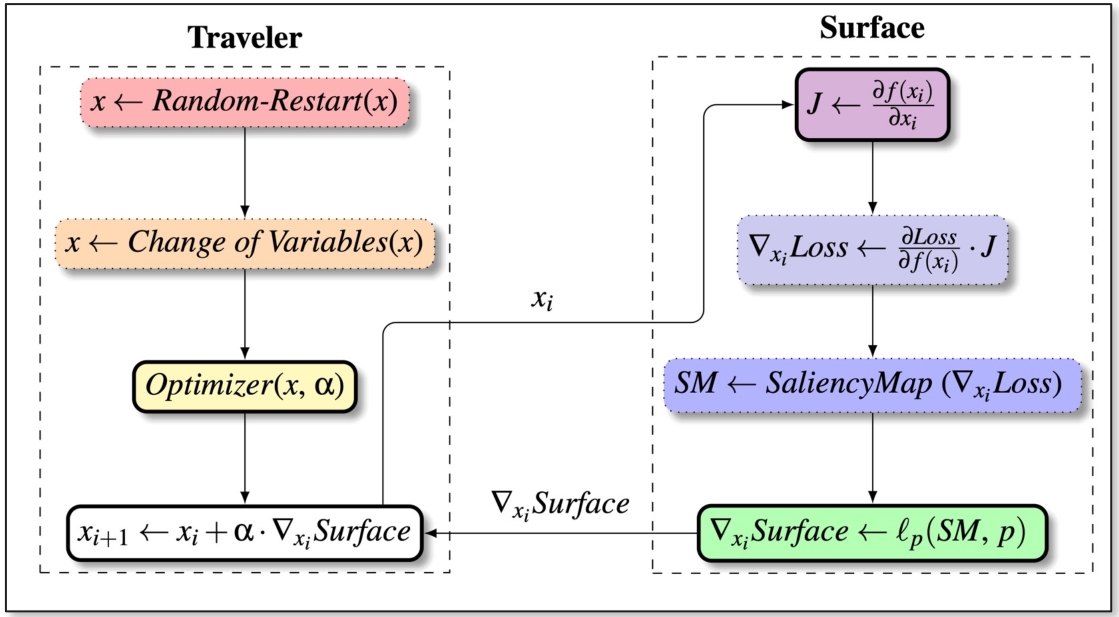
$$\hat{l} \leftarrow \arg\min_{k \neq \hat{k}(\boldsymbol{x}_0)} \frac{|f_k'|}{\|\boldsymbol{w}_k'\|_2}$$
$$\boldsymbol{r}_i \leftarrow \frac{|f_{\hat{l}}'|}{\|\boldsymbol{w}_{\hat{l}}'\|_2^2}\boldsymbol{w}_{\hat{l}}'$$
$$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$$

PGD

$$x^{t+1} = \Pi_{x+\mathcal{S}}\left(x^t + \alpha\,\mathrm{sgn}(\nabla_x L(\theta, x, y))\right)$$



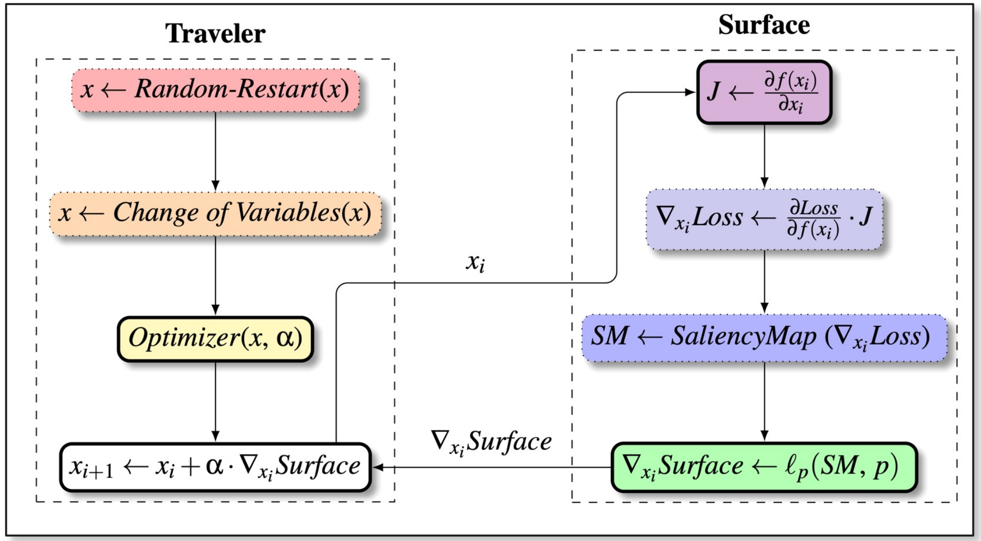
Components of the traveler and surface. Arrows represent the progression of an an attack through components at each iteration.

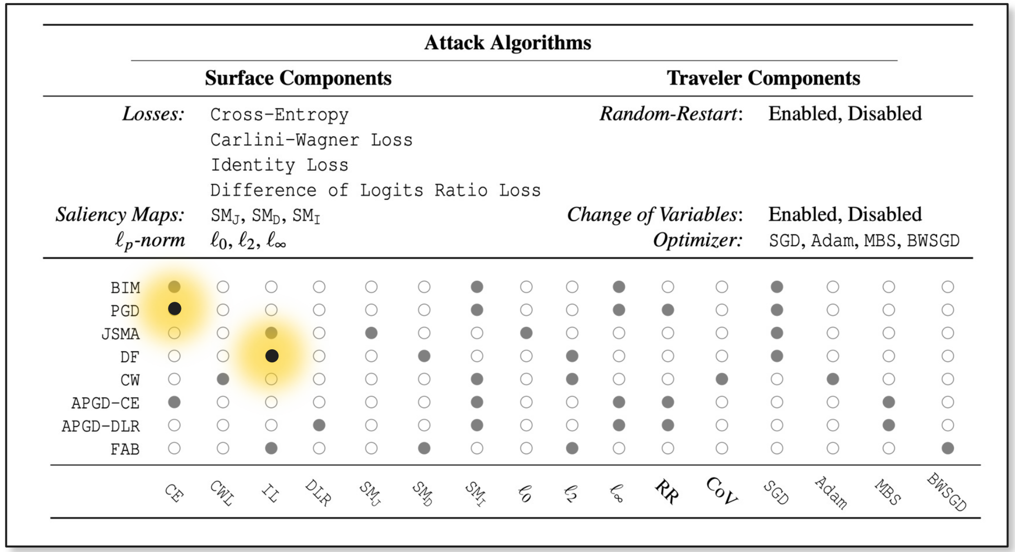


Attack Algorithms

Surface Components		Traveler Components	
Losses:	Cross-Entropy Carlini-Wagner Loss Identity Loss Difference of Logits Ratio Loss	*Random-Restart*:	Enabled, Disabled
Saliency Maps:	SM_J, SM_D, SM_I	*Change of Variables*:	Enabled, Disabled
ℓ_p-norm	ℓ_0, ℓ_2, ℓ_∞	*Optimizer:*	SGD, Adam, MBS, BWSGD

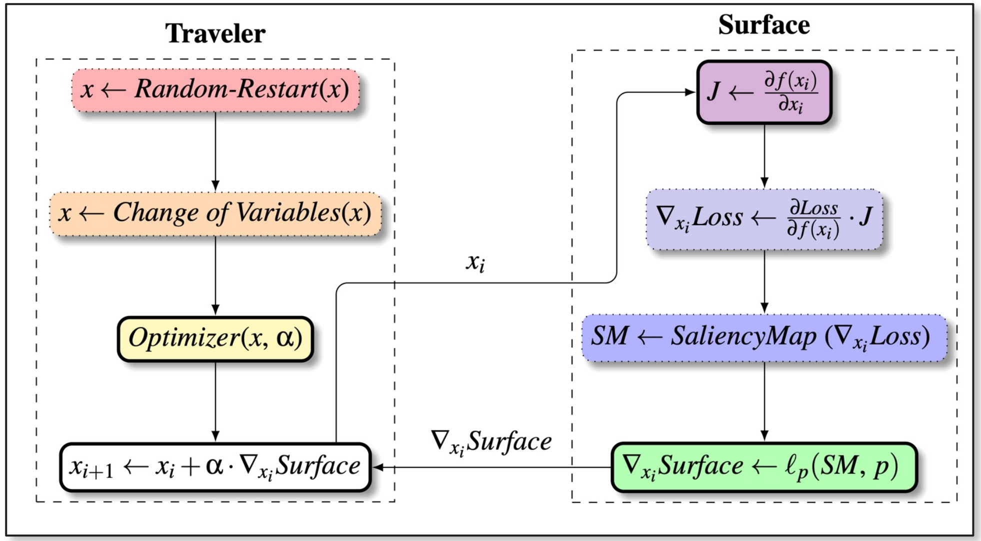
	CE	CWL	IL	DLR	SM_J	SM_D	SM_I	ℓ_0	ℓ_2	ℓ_∞	RR	CoV	SGD	Adam	MBS	BWSGD
BIM	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○	○
PGD	●	○	○	○	○	○	●	○	○	●	●	○	●	○	○	○
JSMA	○	○	●	○	●	○	○	●	○	○	○	○	●	○	○	○
DF	○	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○
CW	○	●	○	○	○	○	●	○	●	○	○	●	○	●	○	○
APGD-CE	●	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○
APGD-DLR	○	○	○	●	○	○	●	○	○	●	●	○	○	○	●	○
FAB	○	○	●	○	○	●	○	○	●	○	○	○	○	○	○	●





PGD: $\nabla \mathrm{CE}(x, y)$

DeepFool: $\nabla f_y(x)$



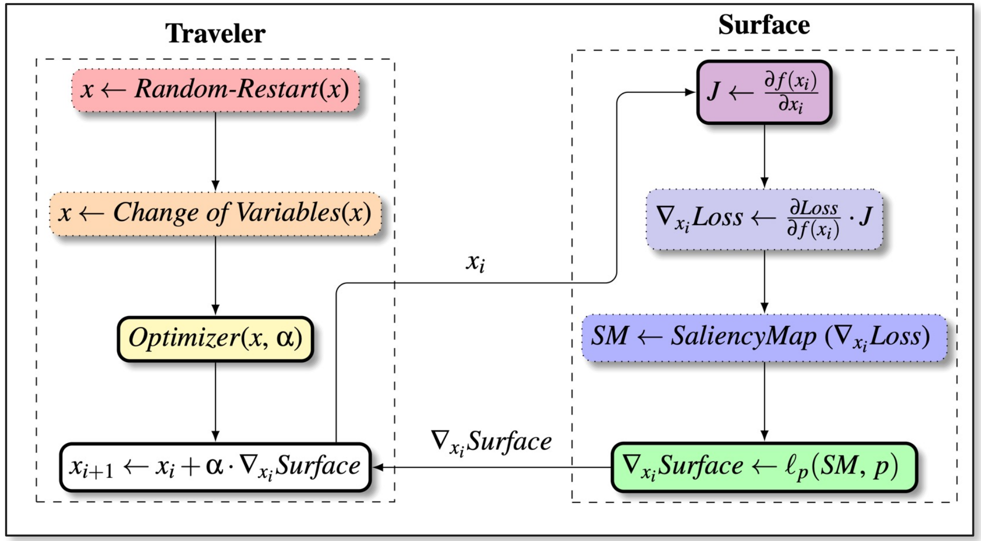
Attack Algorithms

Surface Components		Traveler Components	
Losses:	Cross-Entropy Carlini-Wagner Loss Identity Loss Difference of Logits Ratio Loss	*Random-Restart*:	Enabled, Disabled
Saliency Maps:	SM_J, SM_D, SM_I	*Change of Variables*:	Enabled, Disabled
ℓ_p-norm	ℓ_0, ℓ_2, ℓ_∞	*Optimizer:*	SGD, Adam, MBS, BWSGD

	CE	CWL	IL	DLR	SM_J	SM_D	SM_I	ℓ_0	ℓ_2	ℓ_∞	RR	CoV	SGD	Adam	MBS	BWSGD
BIM	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○	○
PGD	●	○	○	○	○	○	●	○	○	●	●	○	●	○	○	○
JSMA	○	○	●	○	●	○	○	●	○	○	○	○	●	○	○	○
DF	○	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○
CW	○	●	○	○	○	○	●	○	●	○	○	●	○	●	○	○
APGD-CE	●	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○
APGD-DLR	○	○	○	●	○	○	●	○	○	●	●	○	○	○	●	○
FAB	○	○	●	○	○	●	○	○	●	○	○	○	○	○	○	●

PGD: $\nabla \mathrm{CE}(x, y)$

DeepFool: $\frac{|f_y(x) - f_k(x)|}{\nabla f_y(x) - \nabla f_k(x)}$ $(\nabla f_y(x) - \nabla f_k(x))$



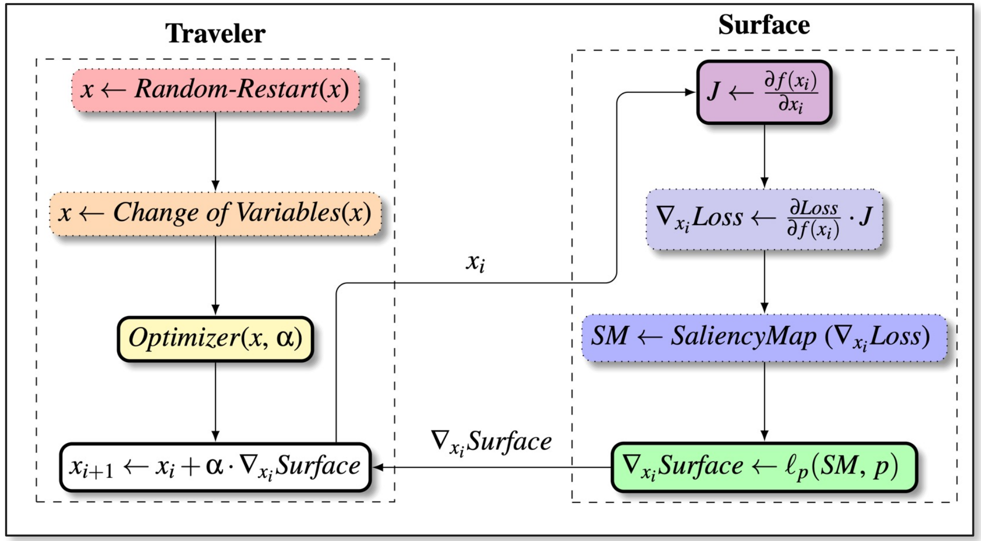
Attack Algorithms

Surface Components		**Traveler Components**	
Losses:	Cross-Entropy Carlini-Wagner Loss Identity Loss Difference of Logits Ratio Loss	*Random-Restart*:	Enabled, Disabled
Saliency Maps:	SM_J, SM_D, SM_I	*Change of Variables*:	Enabled, Disabled
ℓ_p-norm	ℓ_0, ℓ_2, ℓ_∞	*Optimizer:*	SGD, Adam, MBS, BWSGD

	CE	CWL	IL	DLR	SM_J	SM_D	SM_I	ℓ_0	ℓ_2	ℓ_∞	RR	CoV	SGD	Adam	MBS	BWSGD
BIM	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○	○
PGD	●	○	○	○	○	○	●	○	○	●	●	○	●	○	○	○
JSMA	○	○	●	○	●	○	○	●	○	○	○	○	●	○	○	○
DF	○	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○
CW	○	●	○	○	○	○	●	○	●	○	○	●	○	●	○	○
APGD-CE	●	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○
APGD-DLR	○	○	○	●	○	○	●	○	○	●	●	○	○	○	●	○
FAB	○	○	●	○	○	●	○	○	●	○	○	○	○	○	○	●

$$\text{PGD: } \delta = \alpha \cdot \operatorname{sgn}(\nabla \mathrm{CE}(x, y))$$

$$\text{DeepFool: } \delta = \frac{|f_y(x) - f_k(x)|}{\|\nabla f_y(x) - \nabla f_k(x)\|_2^2} \cdot (\nabla f_y(x) - \nabla f_k(x))$$



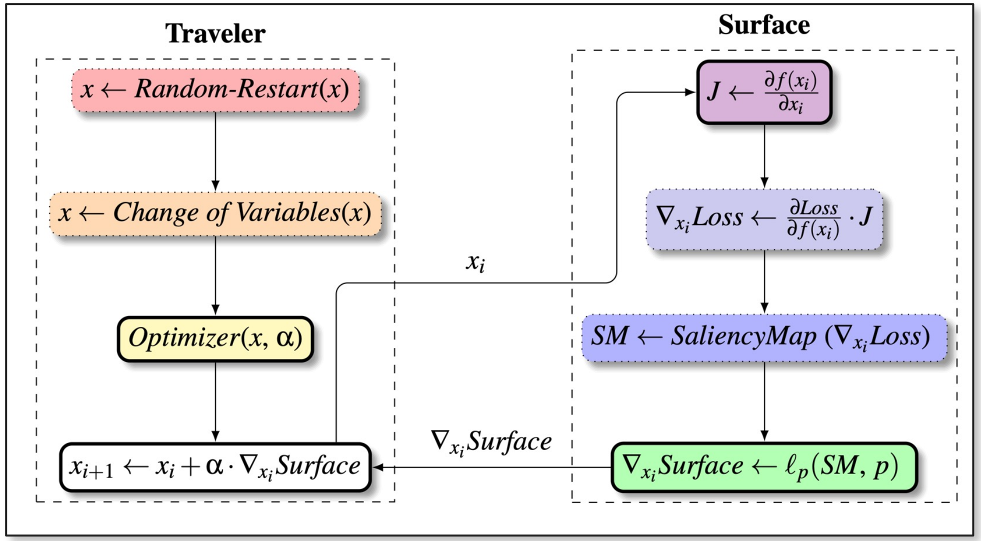
Attack Algorithms

Surface Components		Traveler Components	
Losses:	Cross-Entropy Carlini-Wagner Loss Identity Loss Difference of Logits Ratio Loss	*Random-Restart:*	Enabled, Disabled
Saliency Maps:	SM_J, SM_D, SM_I	*Change of Variables:*	Enabled, Disabled
ℓ_p-norm	ℓ_0, ℓ_2, ℓ_∞	*Optimizer:*	SGD, Adam, MBS, BWSGD

	CE	CWL	IL	DLR	SM_J	SM_D	SM_I	ℓ_0	ℓ_2	ℓ_∞	RR	CoV	SGD	Adam	MBS	BWSGD
BIM	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○	○
PGD	●	○	○	○	○	○	●	○	○	●	●	○	●	○	○	○
JSMA	○	○	●	○	●	○	○	●	○	○	○	○	●	○	○	○
DF	○	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○
CW	○	●	○	○	○	○	●	○	●	○	○	●	○	●	○	○
APGD-CE	●	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○
APGD-DLR	○	○	○	●	○	○	●	○	○	●	●	○	○	○	●	○
FAB	○	○	●	○	○	●	○	○	●	○	○	○	○	○	○	●

$$\text{PGD: } \delta = \alpha \cdot \operatorname{sgn}(\nabla \mathrm{CE}(x, y)), x_0 = x + \mathcal{U}(-\epsilon, \epsilon)$$

$$\text{DeepFool: } \delta = \frac{|f_y(x) - f_k(x)|}{\|\nabla f_y(x) - \nabla f_k(x)\|_2^2} \cdot (\nabla f_y(x) - \nabla f_k(x)), x_0 = x$$



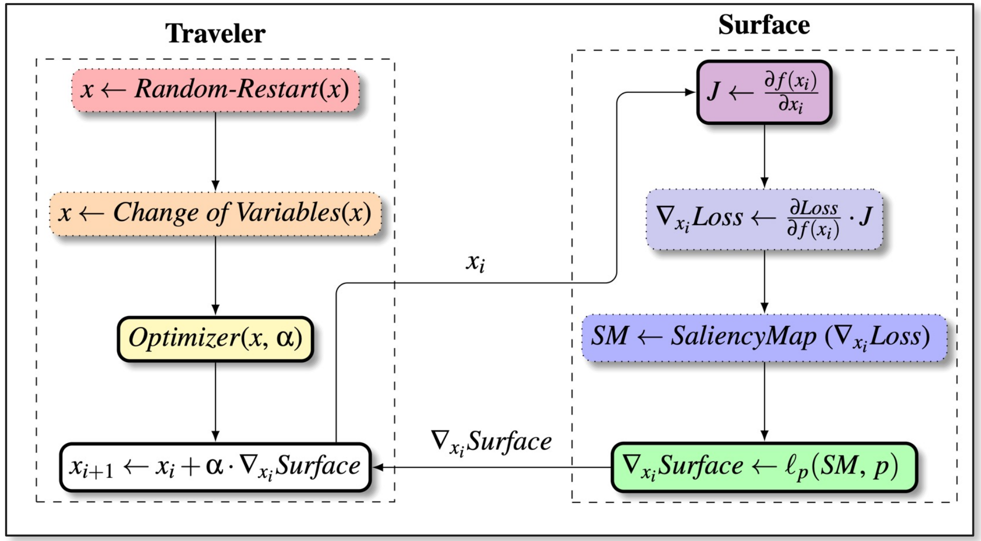
Attack Algorithms

Surface Components		Traveler Components	
Losses:	Cross-Entropy Carlini-Wagner Loss Identity Loss Difference of Logits Ratio Loss	*Random-Restart:*	Enabled, Disabled
Saliency Maps:	SM_J, SM_D, SM_I	*Change of Variables:*	Enabled, Disabled
ℓ_p-norm	ℓ_0, ℓ_2, ℓ_∞	*Optimizer:*	SGD, Adam, MBS, BWSGD

	CE	CWL	IL	DLR	SM_J	SM_D	SM_I	ℓ_0	ℓ_2	ℓ_∞	RR	CoV	SGD	Adam	MBS	BWSGD
BIM	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○	○
PGD	●	○	○	○	○	○	●	○	○	●	●	○	●	○	○	○
JSMA	○	○	●	○	●	○	○	●	○	○	○	○	●	○	○	○
DF	○	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○
CW	○	●	○	○	○	○	●	○	●	○	○	●	○	●	○	○
APGD-CE	●	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○
APGD-DLR	○	○	○	●	○	○	●	○	○	●	●	○	○	○	●	○
FAB	○	○	●	○	○	●	○	○	●	○	○	○	○	○	○	●

$$\text{PGD: } \delta = \alpha \cdot \operatorname{sgn}(\nabla \mathrm{CE}(x, y)), x_0 = x + \mathcal{U}(-\epsilon, \epsilon), x_{i+1} = x_i + \delta$$

$$\text{DeepFool: } \delta = \frac{|f_y(x) - f_k(x)|}{\|\nabla f_y(x) - \nabla f_k(x)\|_2^2} \cdot (\nabla f_y(x) - \nabla f_k(x)), x_0 = x, x_{i+1} = x_i + \delta$$



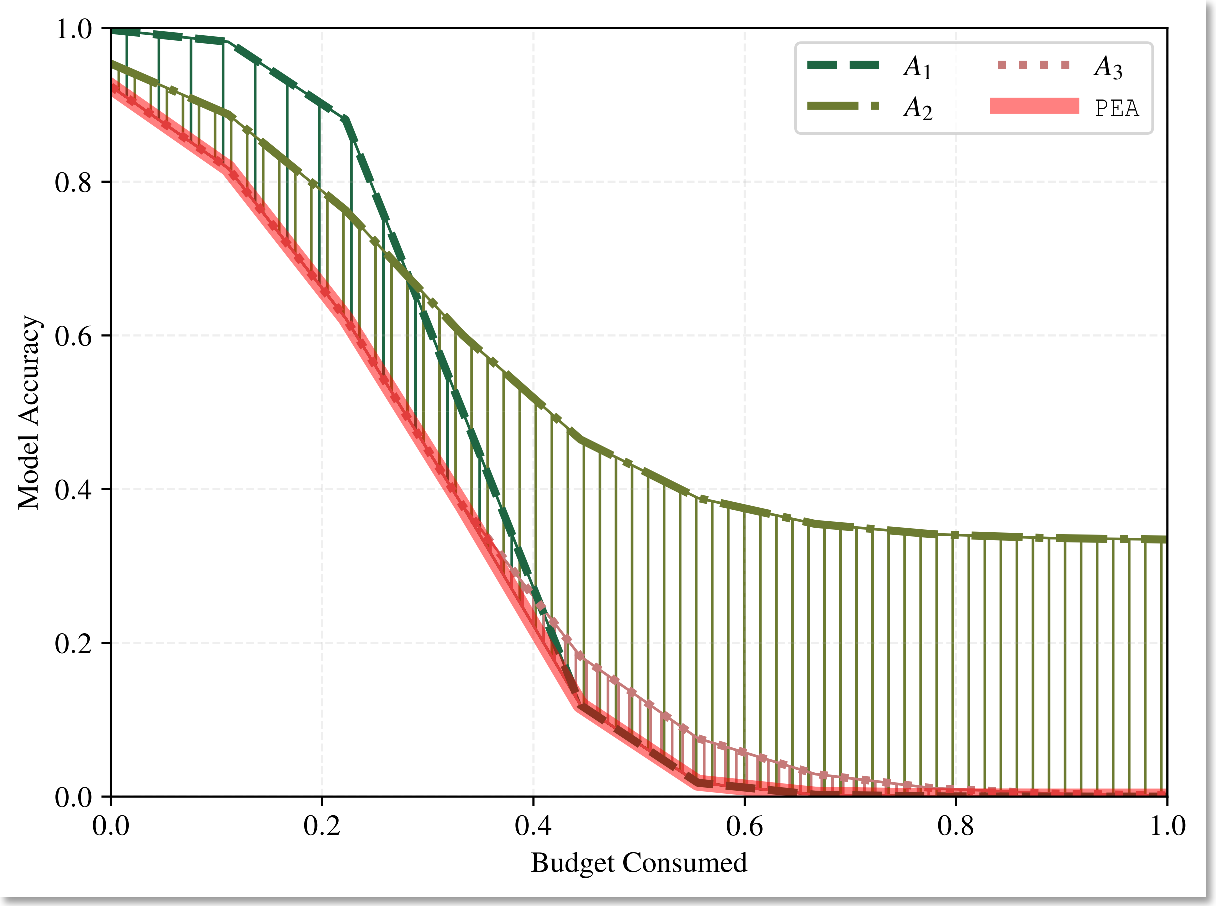
Attack Algorithms

Surface Components		Traveler Components	
Losses:	Cross-Entropy Carlini-Wagner Loss Identity Loss Difference of Logits Ratio Loss	*Random-Restart*:	Enabled, Disabled
Saliency Maps:	SM_J, SM_D, SM_I	*Change of Variables*:	Enabled, Disabled
ℓ_p-norm	ℓ_0, ℓ_2, ℓ_∞	*Optimizer:*	SGD, Adam, MBS, BWSGD

	CE	CWL	IL	DLR	SM_J	SM_D	SM_I	ℓ_0	ℓ_2	ℓ_∞	RR	CoV	SGD	Adam	MBS	BWSGD
BIM	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○	○
PGD	●	○	○	○	○	○	●	○	○	●	●	○	●	○	○	○
JSMA	○	○	●	○	●	○	○	●	○	○	○	○	●	○	○	○
DF	○	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○
CW	○	●	○	○	○	○	●	○	●	○	○	●	○	●	○	○
APGD-CE	●	○	○	○	○	○	●	○	○	●	●	○	○	○	●	○
APGD-DLR	○	○	○	●	○	○	●	○	○	●	●	○	○	○	●	○
FAB	○	○	●	○	○	●	○	○	●	○	○	○	○	○	○	●

*Our **extensible** decomposition of **mutually compatible and independent** components allows us to build a vast attack space containing 576 attacks.*

The Pareto Ensemble Attack



Evaluation

- Questions
  - When and why are attacks performant?
- Setup
  - Adversary has access to model parameters
  - CIC-MalMem2022, Malware Detection, 58k total (k-Fold), 4 classes
  - CIFAR-10, Object Classification, 50k train, 10k test, 10 classes
  - Fashion-MNIST, Clothing Classification, 60k train, 10k test, 10 classes
  - MNIST, Digit Recognition, 60k train, 10k test, 10 classes
  - NSL-KDD, Network Intrusion Detection, 125k train, 22k test, 5 classes
  - Phishing Websites, Phishing Detection, 10k total (k-Fold), 2 classes
  - UNSW-NB15, Network Intrusion Detection, 101k train, 53k test, 10 classes

Hypothesis Testing Illuminates Effective Strategies

	Component H_1		Component H_2		Condition	p-value	Effect Size
1.	SGD	is better than	BWSGD	when	Dataset $=$ MNIST	$<2.2\times10^{-308}$	99 %
2.	Adam	is better than	BWSGD	when	Dataset $=$ MNIST	$<2.2\times10^{-308}$	99 %
		⋮			⋮		
84.	Identity Loss	is better than	Difference of Logits Ratio Loss	when	Dataset $=$ NSL-KDD	$<2.2\times10^{-308}$	93 %
85.	SGD	is better than	BWSGD	when	SaliencyMap $=$ Jacobian Saliency Map	$<2.2\times10^{-308}$	92 %
		⋮			⋮		
393.	DeepFool Saliency Map	is better than	Jacobian Saliency Map	when	Dataset $=$ FMNIST	$<5\times10^{-6}$	66 %
394.	Cross-Entropy	is better than	Carlini-Wagner Loss	when	*Change of Variables* $=$ *Disabled*	$<5\times10^{-6}$	61 %
		⋮			⋮		
1689.	ℓ_0	is better than	ℓ_2	when	Threat Model $=\ell_2+1.0$	9.8×10^{-1}	50 %
1690.	Identity Saliency Map	is better than	DeepFool Saliency Map	when	Threat Model $=\ell_\infty+0.4$	1.0	49 %

Hypothesis Testing Illuminates Effective Strategies

- Change of Variables → 100% disabled, 0% enabled
- Optimizers → 50% Adam, 33% SGD, 16% MBS, 1% BWSGD
- Random Restart → 61% enabled, 39% disabled
- Saliency Maps → 70% no Saliency Map, 30% either DeepFool or JSMA Saliency Map
- Loss → 47% Identity Loss, 34% Cross Entropy, 18% Carlini Wagner Loss, 1% DLR Loss

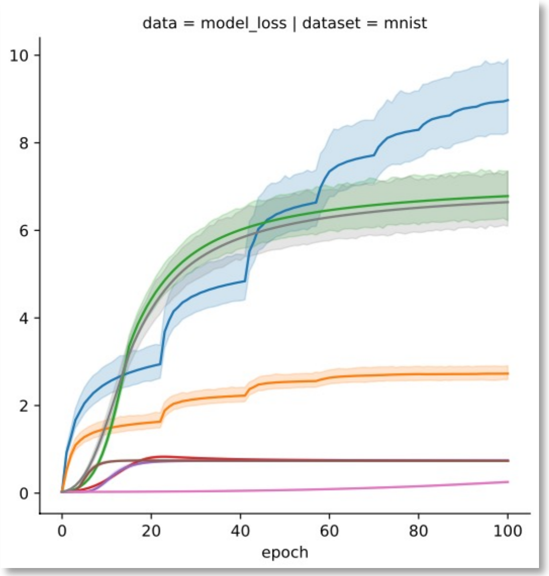
Hypothesis Testing Illuminates Effective Strategies

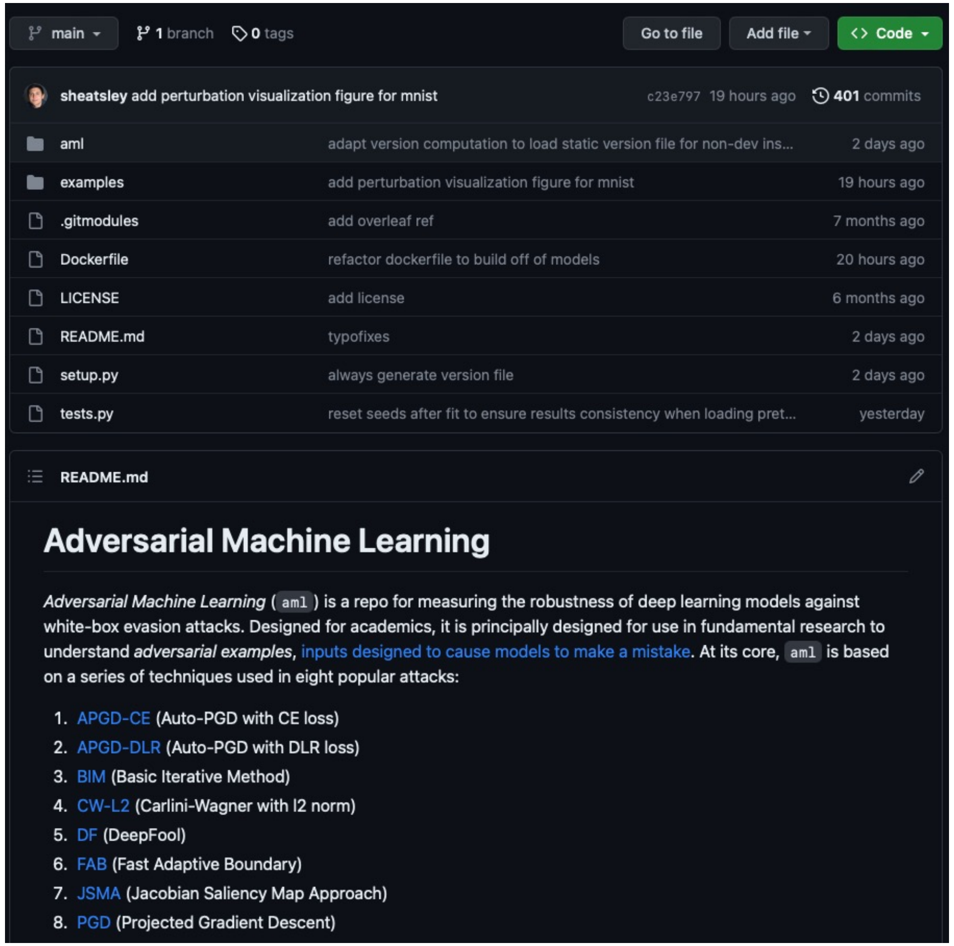
- Change of Variables → 100% disabled, 0% enabled
- Optimizers → 50% Adam, 33% SGD, 16% MBS, 1% BWSGD
- **Random Restart → 61% enabled, 39% disabled**
- Saliency Maps → 70% no Saliency Map, 30% either DeepFool or JSMA Saliency Map
- Loss → 47% Identity Loss, 34% Cross Entropy, 18% Carlini Wagner Loss, 1% DLR Loss

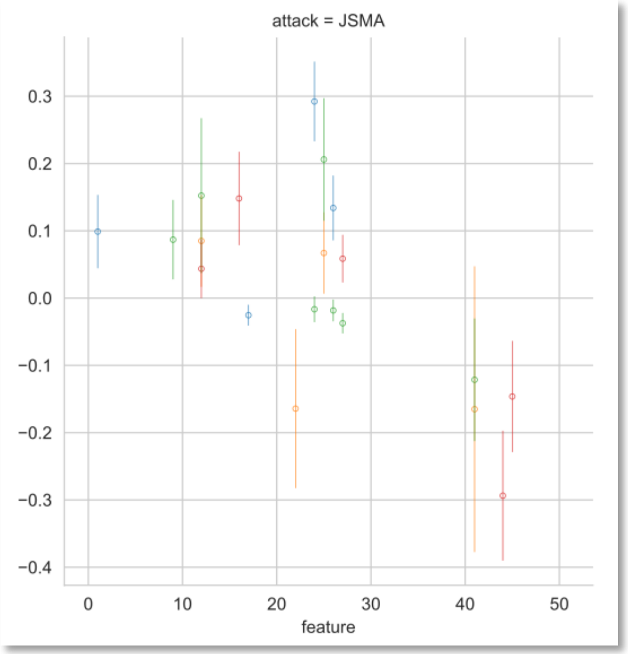
Hypothesis Testing Illuminates Effective Strategies

- Change of Variables → 100% disabled, 0% enabled
- Optimizers → 50% Adam, 33% SGD, 16% MBS, 1% BWSGD
- Random Restart → 61% enabled, 39% disabled
- Saliency Maps → 70% no Saliency Map, 30% either DeepFool or JSMA Saliency Map
- **Loss → 47% Identity Loss, 34% Cross Entropy, 18% Carlini Wagner Loss, 1% DLR Loss**

GitHub Repo: https://github.com/sheatsley/attacks







Thank you

https://hoak.me

bhoak@cs.wisc.edu

@blaine_hoak

blainehoak