The Space of Adversarial Strategies

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

MADS&P

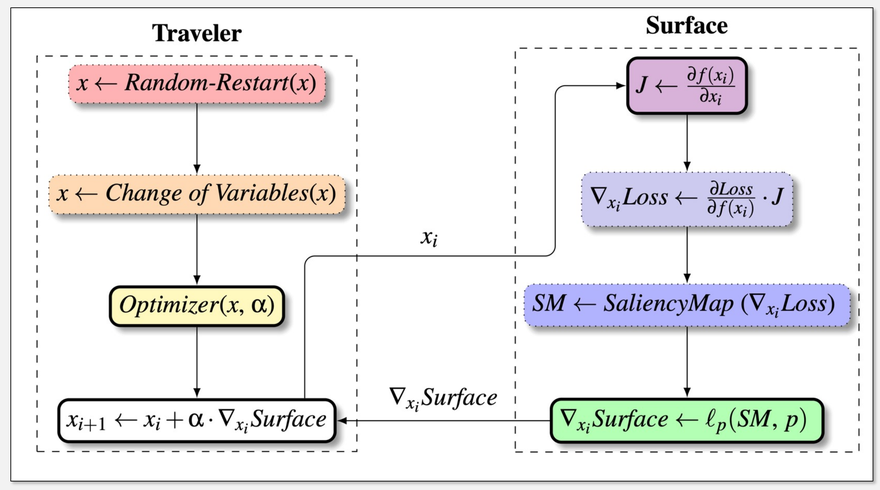
INTRODUCTION

Adversaries in Machine Learning

- Machine learning models are vulnerable to *adversarial examples*, inputs designed to induce a mismatch between model classification and human perception.
- While we have seen significant efforts towards defending against adversarial concerns, most defenses are quickly broken by new attack methods.
- To better understand the attack methods that models are vulnerable to, we propose a systematic approach to characterize worst-case adversaries.
- We explore how the domain, robustness techniques, and threat model influence attack performance.

METHODS

Generalization of Attacks



We observe that attacks can be decomposed into *surfaces* and *travelers*, which contain collections of techniques that operate on gradients and inputs, respectively.

RESULTS

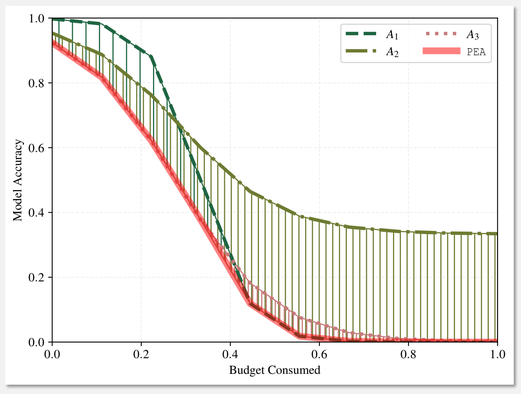
Attack Algorithms

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

Building a Vast Attack Space

- Within our decomposition of attacks, components are independent and mutually compatible; they can be added, omitted, or swapped out to design new attacks.
- We enumerate over all possible combinations of component choices to create a vast attack space totaling 576 attacks, 568 of which were previously unexplored.

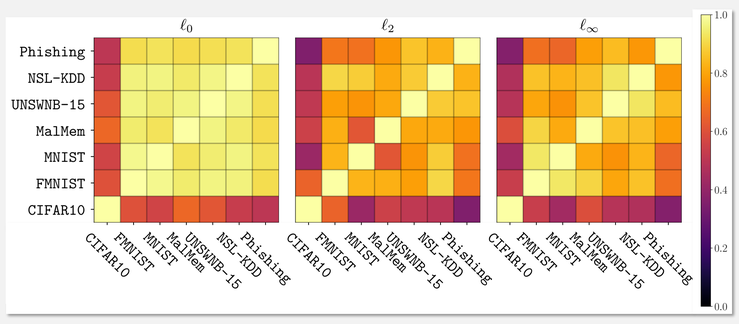
Measuring Optimality



We measure attack performance as closeness to the PEA, a theoretical attack that forms the lower envelope of the accuracy budget space.

Domain Sensitivity

We find that attack performance is domain-sensitive and dependent on threat model; performance is unlikely to translate across domains.



Revealing Effective Components

	Component H_1		Component H_2		Condition	p-value	Effect Size
1.	SGD	is better than	BWSGD	when	Dataset = MNIST	$<2.2 \times 10^{-308}$	99 %
2.	Adam	is better than	BWSGD	when	Dataset = MNIST	$<2.2 \times 10^{-308}$	99 %
		⋮			⋮		
84.	Identity Loss	is better than	Difference of Logits Ratio Loss	when	Dataset = NSL-KDD	$<2.2 \times 10^{-308}$	93 %
85.	SGD	is better than	BWSGD	when	SaliencyMap = Jacobian Saliency Map	$<2.2 \times 10^{-308}$	92 %
		⋮			⋮		
393.	DeepFool Saliency Map	is better than	Jacobian Saliency Map	when	Dataset = FMNIST	$<5 \times 10^{-6}$	66 %
394.	Cross-Entropy	is better than	Carlini-Wagner Loss	when	*Change of Variables* = *Disabled*	$<5 \times 10^{-6}$	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 is used to identify components that lead to performant attacks in different scenarios.
- Resulting trends supported some commonly held beliefs (e.g., random restart is helpful) as well as uncovered new unexpected insights (e.g., using an identity loss is better than cross entropy loss).

CONCLUSION

https://hoak.me @blaine_hoak blainehoak bhoak@cs.wisc.edu

Attacks are sets of components

- Our framework allows us to enumerate over components, yielding new and interesting attacks.
- This attack space allows us to evaluate models and future defenses against a comprehensive set of threats.

Generalizing attacks enables new insights

- Hypothesis testing on components enables us to explain what works well and why, uncovering potential new avenues of research into root causes of model vulnerabilities.
- We find that attack performance is highly dependent on the scenario, highlighting a need for more extensive robustness evaluations