ROUTING AND ACTION

MEMORANDUM

ROUTING

TO:(1) Network, Cyber, and Computational Sciences Branch (NC&CS) (Yu, Paul)

Report is available for review

(2) Proposal Files Report No.: Proposal Number: 74806-NC.10

DESCRIPTION OF MATERIAL

CONTRACT OR GRANT NUMBER: W911NF-19-1-0374

INSTITUTION: Pennsylvania State University

PRINCIPAL INVESTIGATOR: Patrick McDaniel

TYPE REPORT: Final Report

DATE RECEIVED: 9/15/23 1:30PM

PERIOD COVERED: 6/25/19 12:00AM through 12/24/22 12:00AM

TITLE: Final Report: Intelligent Systems, Advanced Learning Theory, Methodology, and Techniques: Mapping Black-Box Attack Metrics and Parameter Spaces in Machine Learning

ACTION TAKEN BY DIVISION

(x) Report has been reviewed for technical sufficiency and IS [x] IS NOT [ ] satisfactory.

( ) Based on my technical review, I have identified no OPSEC or Technology Protection concerns that need to be addressed regarding this report.

(x) Perfomance of the research effort was accomplished in a satisfactory manner and all other technical requirements have been fulfilled.

(x) Based upon my knowledge of the research project, I agree with the patent information disclosed.

Approved by paul.l.yu.civ@mail.mil on 9/27/23 9:06AM

ARO FORM 36-E

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
15-09-2023	Final Report	25-Jun-2019 - 24-Dec-2022

4. TITLE AND SUBTITLE

Final Report: Intelligent Systems, Advanced Learning Theory, Methodology, and Techniques: Mapping Black-Box Attack Metrics and Parameter Spaces in Machine Learning

5a. CONTRACT NUMBER
W911NF-19-1-0374

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Pennsylvania State University
Office of Sponsored Programs
110 Technology Center Building
University Park, PA 16802 -7000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
74806-NC.10

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Patrick McDaniel
UU	UU	UU	UU		19b. TELEPHONE NUMBER 814-863-3599

RPPR Final Report
as of 27-Sep-2023

Agency Code: 21XD

Proposal Number: 74806NC **Agreement Number: W911NF-19-1-0374**
INVESTIGATOR(S):

Name: Patrick McDaniel
Email: mcdaniel@cse.psu.edu
Phone Number: 8148633599
Principal: Y

Organization: **Pennsylvania State University**
Address: Office of Sponsored Programs, University Park, PA 168027000
Country: USA
DUNS Number: 003403953 EIN: 246000376

Report Date: 24-Mar-2023 Date Received: 15-Sep-2023
Final Report for Period Beginning 25-Jun-2019 and Ending 24-Dec-2022
Title: Intelligent Systems, Advanced Learning Theory, Methodology, and Techniques: Mapping Black-Box Attack Metrics and Parameter Spaces in Machine Learning
Begin Performance Period: 25-Jun-2019 **End Performance Period:** 24-Dec-2022
Report Term: 0-Other

Submitted By: Schreier Melissa Email: mschreier@wisc.edu
Phone: (608) 206-6063

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0 **STEM Participants:** 4

Major Goals: [1] Major Goals

Investigating the Impact of Transformer Architectures on Traditional Security Paradigms
1. Craft an evaluation framework for measuring the transferability of adversarial examples across different pre-trained model architectures.
2. Demonstrate how transformer's exhibit high transferability rates of adversarial examples against other model architectures.
3. Show that the degree of transferability of adversarial examples is dependent on the fine-tuned dataset.

The Space of Adversarial Strategies
1. Develop a unified framework of attacks in adversarial machine learning.
2. Evaluate the generalization of attacks introduced by our framework across threat models, datasets, and robust v. non-robust models.
3. Determine the components that yield the most performant attack.

Accomplishments: [2] Accomplished under Goals

Investigating the Impact of Transformer Architectures on Traditional Security Paradigms
Transformer architectures are revolutionizing machine learning across a diverse set of domains. Consequently, understanding the potential risks of transformer's in conventional cybersecurity is a challenge. In this work, we use transferability of adversarial examples as a metric for robustness to evaluate its impact on traditional architectures. We craft adversarial examples on pre-trained convolutional neural network, transformer, and hybrid models fine-tuned on a range of distinct image datasets to evaluate the cross-model transferability.
In this work, we plan to demonstrate that (1) transformer architectures offer more robustness in the transferability of adversarial examples and (2) transferability of adversarial examples are dependent on the dataset used for fine-tuning.

The Space of Adversarial Strategies
To generalize white-box evasion attacks in machine learning, we observed that seminal attacks operate in two modalities: (1) they first produce information which encodes adversarial goals (i.e., gradients), and (2) they act on that information to best achieve

those goals under a specific budget (i.e., a perturbation). This yields a natural generalization for which we label the former as surfaces and the latter as travelers. We exploit this paradigm by identifying the components within these structures and permute them, yielding a vast space of 575 attacks. We then observe that attacks can be fairly compared against a theoretical attack: the Pareto Ensemble Attack (PEA), which bounds the performance attacks aim to achieve. Finally, we apply hypothesis testing to identify the components that lead to high attack performance, conditioned on a dataset, threat model, or defense technique.
From our investigation of three threat models, seven datasets, and robust vs non-robust models, we found that: (1) attacks do not readily generalize, depending on the threat model, dataset, or even if the defender is using a robust model, attack performance can be substantially affected through these factors, and (2) of the attacks that perform well within a given budget, dataset, and defense technique, there are many attacks that are near-optimal, suggesting that there is actually a vast space of legitimate attacks that can adversary could choose form to meet their goals.

One of the key takeaways from our work is that we now have a framework to generate a wide array of attacks to perform robustness evaluations and techniques to identify the efficacy of components that are introduced by attacks. Through this, we can now identify the attacks most likely to be used by an adversary that are uniquely equipped to attack a domain for a given threat model, and subsequently, cater our defensive techniques towards such attacks.

Training Opportunities: Patrick McDaniel Talks, Keynotes, and Lectures:

NSF Funding: Why, What and How, School of Computer, Data & Information Sciences, UW-Madison, Madison, WI, October, 2022.

The Challenges of Machine Learning in Adversarial Settings: A Systems Perspective, Temple University, Philadelphia, PA, March, 2022.

The Challenges of Machine Learning in Adversarial Settings: A Systems Perspective, CACR Security Speaker Series, Indiana University, Online, August, 2021.

The Challenges of Machine Learning in Adversarial Settings: A Systems Perspective, Robustness of AI Systems to Adversarial Attacks (RAISA3), Online, August, 2020.

The Challenges of Machine Learning in Adversarial Settings: A Systems Perspective, Computer Science Department, University of Wisconsin-Madison, Madison, WI, February, 2020.

The Challenges of Machine Learning in Adversarial Settings, Computer Science Department, Stonybrook University, Stonybrook, NY, December, 2019.

The Challenges of Machine Learning in Adversarial Settings, Cylab Security and Privacy Institute, Carnegie Mellon University, Pittsburgh, PA, December, 2019.

The Challenges of Machine Learning in Adversarial Settings, S2ERC, Ball State University, Muncie, IN, November, 2019.

The Challenges of Machine Learning in Adversarial Settings, Triangle Area Privacy and Security Day, Durham, NC, October, 2019.

AI-Cybersecurity Workshop Briefing to the NITRD and MLAI Subcommittees, NITRD and MLAI Subcommittees Quarterly Meeting, Washington, DC, July, 2019.

RPPR Final Report
as of 27-Sep-2023

Results Dissemination: Note: This is not the complete list of publications, but listed are most relevant to this grant.

Yohan Beugin, Quinn Burke, Blaine Hoak, Ryan Sheatsley, Eric Pauley, Gang Tan, Syed Rafiul Hussain and Patrick McDaniel, Building a Privacy-Preserving Smart Camera System, Proceedings on Privacy Enhancing Technologies, July, 2022.

Bolor-Erdene Zolbayarn, Ryan Sheatsley, Patrick McDaniel, Michael J. Weisman, Sencun Zhu, Shitong Zhu and Srikanth V. Krishnamurthy, Generating Practical Adversarial Network Traffic Flows using NIDSGAN, Technical Report arXiv:2203.06694, arXiv preprint, March 2022.

Ryan Sheatsley, Nicolas Papernot, Michael J. Weisman, Gunjan Verma and Patrick McDaniel, Adversarial Examples for Network Intrusion Detection Systems, Journal of Computer Security, IOS Press, January 2022.

Patrick McDaniel, Thorsten Holz, Indra Spiecker, Genannt Dohmann, Christopher Burchard, Ahmad-Reza Sadeghi, Konrad Rieck, Kamalika Chaudhuri, Somesh Jha, Andrea Matwyshyn, David Evans, Felix Freiling and Amy Hasan, Cybersecurity and Machine Learning: Vision Document, Report on the joint NSF/DFG Cybersecurity and Machine Learning Research Workshop, Public Report, National Science Foundation/Deutsche Forschungsgemeinschaf, December 2021.

Ahmed Abdou, Ryan Sheatsley, Yohan Beugin, Tyler Shipp and Patrick McDaniel, HoneyModels: Machine Learning Honeypots, Proceedings of the Military Communications Conference, IEE, November 2021.

Ryan Sheatsley, Matthew Durbin, Azaree Lintereur and Patrick McDaniel, Improving Radioactive Material Localization by Leveraging Cyber-Security Model Optimizations, IEEE Sensors, Vol.21, No.8, April 2021.

Dan Boneh, Andrew J. Grotto, Patrick McDaniel and Nicolas Papernot, Preparing for the Age of Deepfakes and Disinformation, Stanford HAI Policy Brief, 2020.

Ryan Sheatsely, Blaine Hoak, Eric Pauley, Patrick McDaniel, The Space of Adversarial Strategies, Published Computer Science, ArXiv, September, 2022. Accepted at 32nd USENIX Security Symposium.

Honors and Awards: SIGSAC Outstanding Innovation Award
for innovative research in mobile device security, trustworthiness of machine learning, and systems security, November 2021

Penn State Engineering Society Premier Research Award
Given to one faculty member per year, the Penn State Engineering Alumni Society Premier Research Award recognizes and rewards an individual whose contributions to scientific knowledge through research are exemplary and internationally acclaimed, April 2021

AAAS Fellow
for distinguished contributions to the field of computational security and privacy, particularly for advancing algorithms for the formal analysis of mobile devices and applications, November 2020

SIGOPS Hall of Fame Award
recognizing the paper "TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones" (Will Enck first author), "which sparked an important research agenda on smartphone privacy that continues to this day", November 2020

J. D. Williams student paper award, Nuclear Security and Physical Protection division
recognizing the best student papers by area in Proceedings of the Institute of Nuclear Materials Management Annual Meeting (INMM), July 2019

Protocol Activity Status:

Technology Transfer: CleverHans is a software library that provides standardized reference implementations of adversarial example construction techniques and adversarial training. The library may be used to develop more robust machine learning models and to provide standardized benchmarks of models' performance in the adversarial setting. Benchmarks constructed without a standardized implementation of adversarial example construction
are not comparable to each other, because a good result may indicate a robust model or it may merely indicate a weak implementation of the adversarial example construction procedure. This software has been widely distributed within the research and used in hundreds of research papers.

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Ryan Sheatsley
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Quinn Burke
Person Months Worked: 14.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Blaine Hoak
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yohan Beugin
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: PD/PI
Participant: Patrick McDaniel
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 27-Sep-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Computer Security
Publication Identifier Type: DOI Publication Identifier: 10.3233/JCS-210094
Volume: 30 Issue: 5 First Page #: 727
Date Submitted: 8/31/23 12:00AM Date Published: 10/6/22 12:02AM
Publication Location:
Article Title: Adversarial examples for network intrusion detection systems
Authors: Ryan Sheatsley, Nicolas Papernot b, Michael J. Weisman, Gunjan Verma, Patrick McDaniel
Keywords: Adversarial machine learning, network intrusion detection, domain constraints
Abstract: One of the principal uses of physical-space sensors in public safety applications is the detection of unsafe conditions (e.g., release of poisonous gases, weapons in airports, tainted food). However, current detection methods in these applications are often costly, slow to use, and can be inaccurate in complex, changing, or new environments. In this paper, we explore how machine learning methods used successfully in cyber domains, such as malware detection, can be leveraged to substantially enhance physical space detection. We focus on one important exemplar application–the detection and localization of radioactive materials. We show that the ML-based approaches can significantly exceed traditional table-based approaches in predicting angular direction. Moreover, the developed models can be expanded to include approximations of the distance to radioactive material (a critical dimension that reference tables used in practice do not capture).
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: IEEE SENSORS JOURNAL
Publication Identifier Type: DOI Publication Identifier: 10.1109/JSEN.2021.3055778
Volume: 21 Issue: 8 First Page #: 9994
Date Submitted: 8/31/23 12:00AM Date Published: 8/31/21 7:52PM
Publication Location:
Article Title: Improving Radioactive Material Localization by Leveraging Cyber-Security Model Optimizations
Authors: Ryan Sheatsley, Matthew Durbin, Azaree Lintereur, Patrick McDaniel
Keywords: Machine Learning, security, gamma-ray detectors
Abstract: One of the principal uses of physical-space sensors in public safety applications is the detection of unsafe conditions (e.g., release of poisonous gases, weapons in airports, tainted food). However, current detection methods in these applications are often costly, slow to use, and can be inaccurate in complex, changing, or new environments. In this paper, we explore how machine learning methods used successfully in cyber domains, such as malware detection, can be leveraged to substantially enhance physical space detection. We focus on one important exemplar application–the detection and localization of radioactive materials. We show that the ML-based approaches can significantly exceed traditional table-based approaches in predicting angular direction. Moreover, the developed models can be expanded to include approximations of the distance to radioactive material (a critical dimension that reference tables used in practice do not capture).
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

BOOKS:

Publication Type: Book Peer Reviewed: Y **Publication Status:** 1-Published
Publication Identifier Type: ISBN Publication Identifier: 9781119723929
Book Edition: Volume: Publication Year: 2021 Date Received: 31-Aug-2023
Publication Location: Hoboken, New Jersey
Publisher: John Wiley & Sons
Book Title: Evading Machine Learning based Network Intrusion Detection Systems with GANs, Game Theory and Machine Learning for Cyber Security
Authors: Bolor-Erdene Zolbayar, Ryan Sheatsley, Patrick McDaniel
Editor:
Acknowledged Federal Support: Y

RPPR Final Report
as of 27-Sep-2023

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings on Privacy Enhancing Technologies
Date Received: 31-Aug-2023 Conference Date: 11-Jul-2022 Date Published: 11-Jul-2022
Conference Location: Sydney, Australia
Paper Title: Building a Privacy-Preserving Smart Camera System
Authors: Yohan Beugin*, Quinn Burke, Blaine Hoak, Ryan Sheatsley, Eric Pauley, Gang Tan, Syed Rafiul Hussai
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Military Communications Conference
Date Received: 31-Aug-2023 Conference Date: 30-Nov-2021 Date Published: 30-Nov-2021
Conference Location: San Diego, CA
Paper Title: HoneyModels: Machine Learning Honeypots
Authors: Ahmed Abdou, Ryan Sheatsley, Yohan Beugin, Tyler Shipp, Patrick McDaniel
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Usenix Security Symposium 23
Date Received: 06-Sep-2023 Conference Date: 09-Aug-2023 Date Published: 09-Aug-2023
Conference Location: Anaheim, CA
Paper Title: The Space of Adversarial Strategies
Authors: Ryan Sheatsley, Blaine Hoak, Eric Pauley, Patrick McDaniel
Acknowledged Federal Support: **Y**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Melissa Schreier
Signature Date: 9/15/23 1:30PM

[1] Major Goals

Investigating the Impact of Transformer Architectures on Traditional Security Paradigms

1. Craft an evaluation framework for measuring the transferability of adversarial examples across different pre-trained model architectures.
2. Demonstrate how transformer's exhibit high transferability rates of adversarial examples against other model architectures.
3. Show that the degree of transferability of adversarial examples is dependent on the fine-tuned dataset.

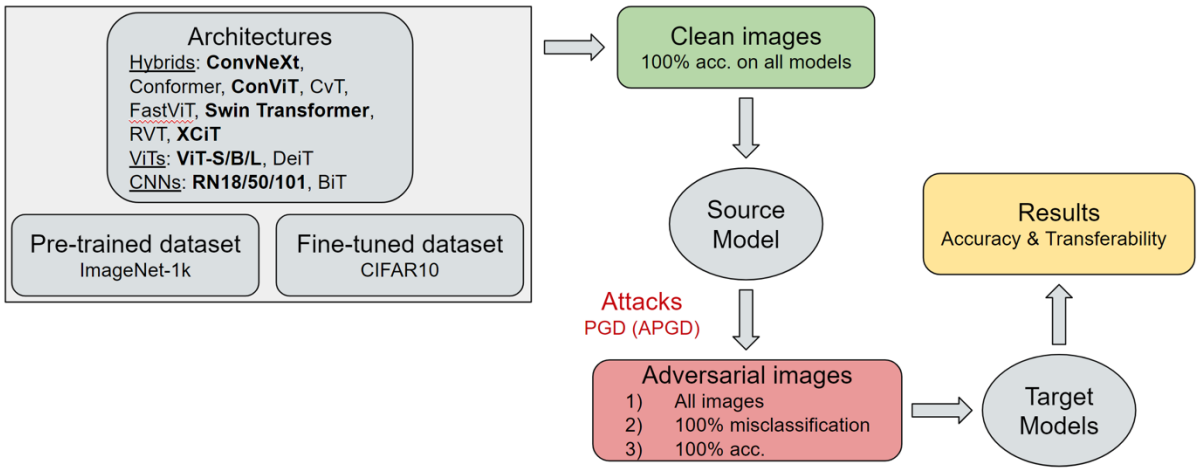
The Space of Adversarial Strategies

1. Develop a unified framework of attacks in adversarial machine learning.
2. Evaluate the generalization of attacks introduced by our framework across threat models, datasets, and robust v. non-robust models.
3. Determine the components that yield the most performant attack.

[2] Accomplished under Goals

Investigating the Impact of Transformer Architectures on Traditional Security Paradigms

Transformer architectures are revolutionizing machine learning across a diverse set of domains. Consequently, understanding the potential risks of transformer's in conventional cybersecurity is a challenge. In this work, we use transferability of adversarial examples as a metric for robustness to evaluate its impact on traditional architectures. We craft adversarial examples on pre-trained convolutional neural network, transformer, and hybrid models fine-tuned on a range of distinct image datasets to evaluate the cross-model transferability.

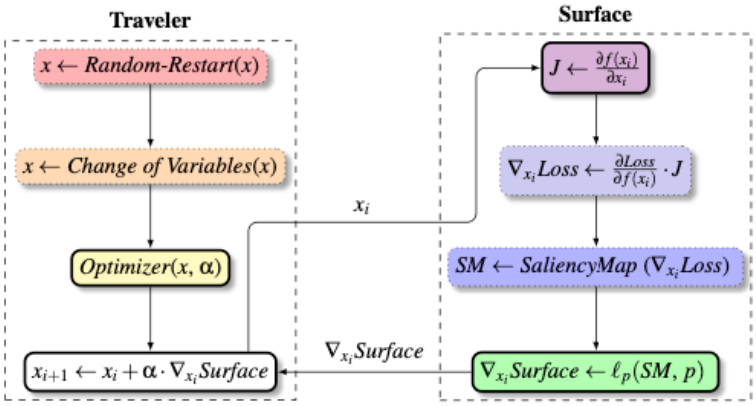


In this work, we plan to demonstrate that (1) transformer architectures offer more robustness in the transferability of adversarial examples and (2) transferability of adversarial examples are dependent on the dataset used for fine-tuning.

The Space of Adversarial Strategies

To generalize white-box evasion attacks in machine learning, we observed that seminal attacks operate in two modalities: (1) they first produce information which encodes adversarial goals (i.e., gradients), and (2) they act on that information to best achieve those goals under a specific budget (i.e., a perturbation). This yields a natural

generalization for which we label the former as *surfaces* and the latter as *travelers*. We exploit this paradigm by identifying the components within these structures and permute them, yielding a vast space of 575 attacks. We then observe that attacks can be fairly compared against a theoretical attack: the Pareto Ensemble Attack (PEA), which bounds the performance attacks aim to achieve. Finally, we apply hypothesis testing to identify the components that lead to high attack performance, conditioned on a dataset, threat model, or defense technique.



From our investigation of three threat models, seven datasets, and robust vs non-robust models, we found that: (1) attacks do not readily generalize, depending on the threat model, dataset, or even if the defender is using a robust model, attack performance can be substantially affected through these factors, and (2) of the attacks that perform well within a given budget, dataset, and defense technique, there are many attacks that are near-optimal, suggesting that there is actually a vast space of legitimate attacks that can adversary could choose form to meet their goals.

	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 %

One of the key takeaways from our work is that we now have a framework to generate a wide array of attacks to perform robustness evaluations and techniques to identify the efficacy of components that are introduced by attacks. Through this, we can now identify the attacks most likely to be used by an adversary that are uniquely equipped to attack a domain for a given threat model, and subsequently, cater our defensive techniques towards such attacks.