A crisis in adversarial machine learning

Nicholas Carlini
Google
Why do we study adversarial machine learning?
We might want to improve ...

1. General purpose robustness
2. The robustness against worst-case attack
3. The robustness against practical attacks
We might want to improve ...

1. General purpose robustness

2. The robustness against worst-case attack

3. The robustness against practical attacks
We might want to improve ... 

1. General purpose robustness
2. The robustness against worst-case attack
3. The robustness against practical attacks

<table>
<thead>
<tr>
<th>Event</th>
<th>Start time</th>
<th>End time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opening Remarks</td>
<td>8:50</td>
<td>9:00</td>
</tr>
<tr>
<td>Invited talk: Yang Liu</td>
<td>9:00</td>
<td>9:30</td>
</tr>
<tr>
<td>Invited talk: Quanshi Zhang</td>
<td>9:30</td>
<td>10:00</td>
</tr>
<tr>
<td>Invited talk: Baoyuan Wu</td>
<td>10:00</td>
<td>10:30</td>
</tr>
<tr>
<td>Invited talk: Aleksander Madry</td>
<td>10:30</td>
<td>11:00</td>
</tr>
<tr>
<td>Invited talk: Bo Li</td>
<td>11:00</td>
<td>11:30</td>
</tr>
<tr>
<td>Poster Session (click)</td>
<td>11:30</td>
<td>12:30</td>
</tr>
<tr>
<td>lunch (12:30-13:30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral Session (click)</td>
<td>13:30</td>
<td>14:10</td>
</tr>
<tr>
<td>Challenge Session</td>
<td>14:10</td>
<td>14:30</td>
</tr>
<tr>
<td>Invited talk: Nicholas Carlini</td>
<td>14:30</td>
<td>15:00</td>
</tr>
<tr>
<td>Invited talk: Judy Hoffman</td>
<td>15:00</td>
<td>15:30</td>
</tr>
<tr>
<td>Invited talk: Alan Yuille</td>
<td>15:30</td>
<td>16:00</td>
</tr>
<tr>
<td><strong>Invited talk: Ludwig Schmidt</strong></td>
<td><strong>16:00</strong></td>
<td><strong>16:30</strong></td>
</tr>
<tr>
<td>Invited talk: Cihang Xie</td>
<td>16:30</td>
<td>17:00</td>
</tr>
</tbody>
</table>
We might want to improve ...

1. General purpose robustness

2. The robustness against worst-case attack

3. The robustness against practical attacks
We might want to improve ...

1. General purpose robustness

2. The robustness against worst-case attack

3. The robustness against practical attacks
The Year is 2014

Someone tells you they have a new algorithm to generate synthetic images.
The Year is 2014
The Year is 2017

Someone tells you they have a new algorithm to generate synthetic images.
The Year is 2017
The Year is 2022

Someone tells you they have a new algorithm to generate synthetic images
A photo of a Corgi dog riding a bike in Times Square. It is wearing sunglasses and a beach hat.
The Year is 2013

Someone tells you they have discovered a flaw in the robustness of neural networks
The Year is 2013
The Year is 2022

Someone tells you they have discovered a flaw in the robustness of neural networks
The Year is 2022
Why?

Defenses are *really* hard.
That can't be all though.

Consider symmetric key cryptography
Cryptanalysis of the Cellular Message Encryption Algorithm

Related-Key Cryptanalysis of 3-WAY, Biham-DES,CAST, DES-X, NewDES, RC2, and TEA

Cryptanalysis of some recently-proposed multiple modes of operation

\[
\{k, A, im, ni, pa, de, ce, aff, lat, ar, se, tain, p, deriv, pr, how, t, mi, differ, the, at, ov, dr, fr, we, pr, know, again, t, av, nor, DES, more, adver, bit, k, Hash, Ther, attac, for, E, retain, shows, \}
\]

Differential cryptanalysis of KHF

Cryptanalysis of TWOPRIME

Don Coppersmith\(^1\), David Wagner\(^2\), Bruce Schneier\(^3\), and J

\(^1\) IBM Research, e-mail: copper@watson.ibm.com
\(^2\) U.C. Berkeley, e-mail: daw@cs.berkeley.edu
\(^3\) Counterpane Systems, e-mail: [schneier,keley]@counterpane.com

Abstract. Ding et al [DNRSS97] propose a stream generator with several layers. We present several attacks. First, we observe non-surjectivity of a linear combination step allows us to key with minimal effort. Next, we show that the various subkeys mixed by these layers, enabling an attack similar to two-loop Vigenère ciphers to recover the remainder of the key. These techniques let us recover the entire TWOPRIME key.\(^4\) We use the generator to produce \(2^{31}\) blocks, (\(2^{15}\) bytes), or 16 hours, of which we examine about one million blocks (\(2^{12}\) 1) computational workload can be estimated at \(2^{28}\) operations set of attacks trades off texts for time, reducing the amount of text needed to just eight blocks (64 bytes), while need fill and 2\(1\) space. We also show how to break two variants of TWOPRIME presented in the original paper.

1 Introduction

\[
1 \quad 1 \quad 1 \quad 1
\]

\[
\text{Slide Attacks}
\]

The boomerang attack

Cryptanalysis of FROG

Cryptanalysis of ORYX

Cryptanalysis of SPEED

Abstract. It is a general belief among the designers of block-ciphers that even a relatively weak cipher may become very strong if its number of rounds is made very large. In this paper we describe a new generic known- (or sometimes chosen-) plaintext attack on product ciphers, which we call the slide attack and which in many cases is independent of the number of rounds of a cipher. We illustrate the power of this new tool by giving practical attacks on several recently designed ciphers: TREYFER, WAKE-ROBF, and variants of DES and Blowfish.

1 Introduction

As the speed of computers grows, fast block ciphers tend to use more and more rounds, rendering all currently known cryptanalytic techniques useless. This is mainly due to the fact that such popular tools as differential [1] and linear analysis [13] are statistic attacks that excel in pushing statistical irregularities and biases through surprisingly many rounds of a cipher. However any such approach finally reaches its limits, since each additional round requires an exponential effort from the attacker.

This tendency towards a higher number of rounds can be illustrated if one looks at the candidates submitted to the AES contest. Even though one of the main criteria of the AES was speed, several prospective candidates (and not the least ones) have really been number of rounds: RC6/99, MARSH/99.
<6 years later ...

AES is basically perfect
Biclique Cryptanalysis of the Full AES

Andrey Bogdanov*, Dmitry Khovratovich, and Christian Rechberger*

K.U. Leuven, Belgium; Microsoft Research Redmond, USA; ENS Paris and Chaire France Telecom, France

Abstract. Since Rijndael was chosen as the Advanced Encryption Standard, improving upon 7-round attacks on the 128-bit key variant or upon 8-round attacks on the 192/256-bit key variants has been one of the most difficult challenges in the cryptanalysis of block ciphers for more than a decade. In this paper we present a novel technique of block cipher cryptanalysis with bicliques, which leads to the following results:

- The first key recovery attack on the full AES-128 with computational complexity $2^{126.1}$.
- The first key recovery attack on the full AES-192 with computational complexity $2^{189.7}$.
- The first key recovery attack on the full AES-256 with computational complexity $2^{254.4}$.
- Attacks with lower complexity on the reduced-round versions of AES not considered before, including an attack on 8-round AES-128 with complexity $2^{124.9}$.
- Preimage attacks on compression functions based on the full AES versions.

In contrast to most shortcut attacks on AES variants, we do not need to assume related-keys. Most of our attacks only need a very small part of the codebook and have small memory requirements, and are practically verified to a large extent. As our attacks are of high computational complexity, they do not threaten the practical use of AES in any way.

Keywords: block ciphers, bicliques, AES, key recovery, preimage
For some reason though, >6 years on, we can't stop publishing defenses that are broken by undergrads.
Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent

Oliver Bryniarski*  
UC Berkeley

Nabeel Hingun*  
UC Berkeley

Pedro Pachuca*  
UC Berkeley

Vincent Wang*  
UC Berkeley

Nicholas Carlini  
Google

Abstract

Evading adversarial example detection defenses requires finding adversarial examples that must simultaneously (a) be misclassified by the model and (b) be detected as non-adversarial. We find that existing attacks that attempt to satisfy multiple simultaneous constraints often over-optimize against one constraint at the cost of satisfying another. We introduce Orthogonal Projected Gradient Descent, an improved attack technique to generate adversarial examples that avoids this problem by orthogonalizing the gradients when running standard gradient-based attacks. We use our technique to evade four state-of-the-art detection defenses, reducing their accuracy to 0% while maintaining a 0% detection rate.
Does that mean we've made **zero** progress?

Obviously not.
We've gotten really good at knowing how to evaluate correctly, if you try hard.
Increasing Confidence in Adversarial Robustness Evaluations

Roland Zimmermann*
University of Tübingen

Wieland Brendel
University of Tübingen

Florian Tramèr
Google

Nicholas Carlini
Google

Abstract

Hundreds of defenses have been proposed in the past years to make deep neural networks robust against minimal (adversarial) input perturbations. However, only a handful of these could hold up their claims because correctly evaluating robustness is extremely challenging: Weak attacks often fail to find adversarial examples even if they unknowingly exist, thereby making a vulnerable network look robust. In this paper, we propose a test to identify weak attacks. Our test introduces a small and simple modification into a neural network that guarantees the existence of an adversarial example for every sample. Consequentially, any correct attack must succeed in attacking this modified network. For eleven out of thirteen previously-published defenses, the original evaluation of the defense fails our test, while stronger attacks that break these defenses pass it. We hope that attack unit tests such as ours will be a major component in future robustness evaluations and increase confidence in an empirical field that today is riddled with skepticism and disbelief. Online version & Code: zimmerrol.github.io/active-tests/

This paper argues for viewing defense proposals as theorem statements, and the corresponding evaluations as proofs. The purpose of a defense evaluation, then, is to provide a convincing and rigorous argument that the defense is correct. Currently, for an adversary to claim to have a "break" of a defense, it is necessary to actually produce the adversarial examples that cause the model to make an error — analogous to refuting a complexity-theoretic impossibility result by producing an efficient algorithm. We argue that this is not how things should work. A valid refutation of a theorem would be to say "there is a flaw in your proof on line 9". Because the null hypothesis for a theorem is that it is false, just as the null hypothesis for a defense should be that it is not robust.
The result I'm most surprised by: certified robustness on ImageNet!
Two ways to evaluate robustness:

1. Construct a proof of robustness
2. Demonstrate constructive attack
To appear ~tomorrow
Who would win?

Six years of researchers training the best adversarially robust neural networks

One diffusion model
$L_2 = 75$
$L_2 = 75$
L₂ distortion: 75
$L_2$ distortion: 75
We might want to improve ...

1. General purpose robustness
2. The robustness against worst-case attack
3. The robustness against practical attacks
MACHINE LEARNING

Adversarial attacks on medical machine learning
Emerging vulnerabilities demand new conversations

By Samuel G. Finlayson¹, John D. Bowers², Joichi Ito³, Jonathan L. Zittrain², Andrew L. Beam⁴, Isaac S. Kohane¹
Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, Feng Lu

Technology and Systems, School of CSE, Beihang University, Beijing, China.

Department of Biomedical Engineering, The University of Melbourne, Parkville, VIC 3010, Australia.

Center for Data-Based Precision Medicine, Beihang University, Beijing, China.

National Institute of Informatics, Tokyo 101-8430, Japan.

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, Ningbo, China.

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems

Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices

Md. Abdur Rahman, Senior Member, IEEE and M. Shamim Hossain, Senior Member, IEEE, Nabil A. Alrajeh, Fawaz Alsolami

Toward an Understanding of Adversarial Examples in Clinical Trials

Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown

1 School of Computer Science, University of Manchester, Manchester M13 9PL, UK

2 Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK

Advancing Learning and Robust Machine Learning for Healthcare: A Survey

Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha

Formation Technology University (ITU), Punjab, Lahore, Pakistan

University of the West England (UWE), Bristol, United Kingdom

Hamad Bin Khalifa University (HBKU), Doha, Qatar
Who even is the adversary here?
Discord Safety: Safe Messaging!

Discord Direct Messages (DMs) are a great way to instant message your buddies with the latest gossip or silliest memes.

To keep your DMs clean and prevent any unwarranted surprises at bay, Discord has a few extra levers you can pull. While we’re still building out a few of these options, if you open your user settings tab and select the Privacy & Safety option, you’ll see the “Safe Direct Messaging” option!

Wikipedia, decided to build from a model of openly so bad that the number of per month—had fallen by 40 not one solution to combat this and consider ways to combat it.
Under the skin of OnlyFans

By Rianna Croxford
Correspondent, BBC News

© 17 July 2021
Under the skin of OnlyFans

By Rianna Croxford
Correspondent, BBC News

17 July 2021

In a statement, OnlyFans said the account did not have two-factor authentication, which made it vulnerable. The company said Tina did not report the racial slur and it was not detected by the site's moderation system because it was pluralised.
We might want to improve ... 

1. General purpose robustness
2. The robustness against worst case attack
3. The robustness against practical attacks

we still have a chance!
Stateful Detection of Black-Box Adversarial Attacks

Steven Chen
University of California, Berkeley

Nicholas Carlini
Google Research

David Wagner
University of California, Berkeley
Under attack
DOG
DOG
DOG
DOG
Our Defense
DOG
DOG
Except here's the thing.

I don't believe this defense actually works.
What I want:

More attacks and defenses on practical systems.
We might want to improve ...

1. General purpose robustness
2. The robustness against worst-case attack
3. The robustness against practical attacks

we still have a chance!