Imperial College London

RISS
Resilient Information Systems Security

https://rissgroup.org

Adversarial Machine Learning
(Part 2)

Luis Muñoz-González

l.munoz@imperial.ac.uk

20th December 2018

Reminder…

Evasion Attacks:

- Attacks at test time.
- The attacker aims to find the blind spots and weaknesses of the ML system to evade it.

Poisoning Attacks:

- Compromise data collection.
- The attacker subverts the learning process.
- Degrades the performance of the system.
- Can facilitate future evasion.

Evasion Attacks

a.k.a. Adversarial Examples

- C. Szegedy et al. "*Intriguing Properties of Neural Networks*." arXiv preprint, 2013.
- I. Goodfellow, J. Shlens, C. Szegedy. "*Expalining and Harnessing Adversarial Examples*." ICLR 2015.

Evasion Attacks

- K. Eykholt et al. "*Robust Physical World Attacks on Deep Learning Visual Classification*." CCVPR, pp. 1625-1634, 2018.
- G.F. Elsayed et al. "*Adversarial Examples that Fool both Computer Vision and Time-Limited Humans*." Arxiv pre-print arxiv:1802.08195v3, 2018.

Task: Spam filtering. **Classifier:** LSTM. **Original label:** 100% Spam. **New label:** 89% Non-Spam.

Text: your ~~application~~ petition has been ~~accepted~~ recognized thank you for your ~~loan~~ borrower ~~request~~ petition , which we recieved yesterday , your ~~refinance~~ subprime ~~application~~ petition has been ~~accepted~~ recognized good credit or not , we are ready to give you a $ oov loan , after further review , our lenders have established the lowest monthly payments . approval process will take only 1 minute . please visit the confirmation link below and fill-out our short 30 second secure web-form . http : oov

---

Task: Sentiment analysis. **Classifier:** CNN. **Original label:** 81% Positive. **New label:** 100% Negative.

Text: i ~~went~~ moved to wing wednesday which is all-you-can-eat wings for $ oov even though they raise the prices it 's ~~still~~ ever really great deal . you can eat as many wings you want to get all the different ~~flavors~~ tastes and have a good time enjoying the atmosphere . the girls are smoking hot ! all the types of ~~sauces~~ dressings are awesome ! and i had at least 25 wings in one sitting . i would ~~definitely~~ certainly go again ~~just~~ simply not every ~~wednesday~~ friday maybe once a month .

Task: Fake news detection. **Classifier:** Naive Bayes. **Original label:** 97% Fake. **New label:** 100% Real

Text: trump supporter whose ~~brutal~~ ferocious beating by black ~~mob~~ gangsta was caught on ~~video~~ tape ~~asks~~ demands : " what happened to america ? " [video] , " david oov , a 49 year ~~old~~ former ~~chicago~~ rochester man who was brutally beaten by a ~~mob~~ lowlife of black democrats ~~asks~~ demands , " what happened to america ? " here is his very ~~sad~~ disappointing story

V. Kuleshov et al. "*Adversarial Examples for Natural Language Classification Problems*." 2018.

Evasion Attacks in the Wild

TREND MICRO | TrendLabs SECURITY INTELLIGENCE Blog
SECURITY NEWS DIRECT FROM THREAT DEFENSE EXPERTS

Search:

Home | Categories

Home » Malware » Cerber Starts Evading Machine Learning

Cerber Starts Evading Machine Learning

Posted on: March 28, 2017 at 1:00 am **Posted in:** Malware, Ransomware
Author: Gilbert Sison (Threats Analyst)

41

The CERBER family of ransomware has been found to have adopted a new technique to make itself harder to detect: it is now using a new loader that appears to be designed to evade detection by machine learning solutions. This loader is designed to hollow out a normal process where the code of CERBER is instead run.

Behavior and Analysis

Ransomware typically arrives via email, and these new CERBER variants are no exception. Emails that claim to be from various utilities may have been used. The emails contain a link to a self-extracting archive, which has been uploaded to a Dropbox account controlled by the attackers. The target then downloads and opens it to infect a system. The following flow chart shows what happens next.

Featured Stories

- systemd Vulnerability Leads to Denial of Service on Linux
- qkG Filecoder: Self-Replicating, Document-Encrypting Ransomware
- Mitigating CVE-2017-5689, an Intel Management Engine Vulnerability
- A Closer Look at North Korea's Internet
- From Cybercrime to Cyberpropaganda

Security Predictions for 2018

Attackers are banking on network vulnerabilities and inherent weaknesses to facilitate massive malware attacks, IoT hacks, and operational disruptions. The ever-shifting threats and increasingly expanding attack surface will challenge

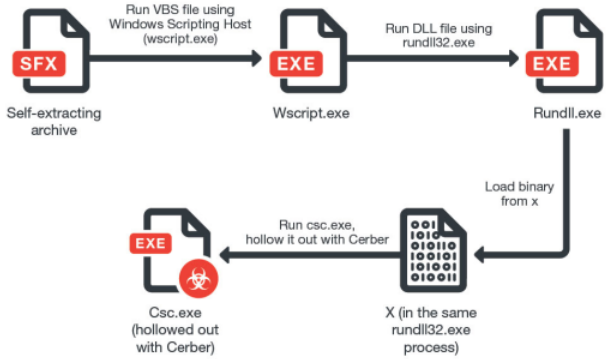
TWCN Tech News

Home | News | Microsoft | Windows | Office | Phone | Security | Social Media | General | About | Tea

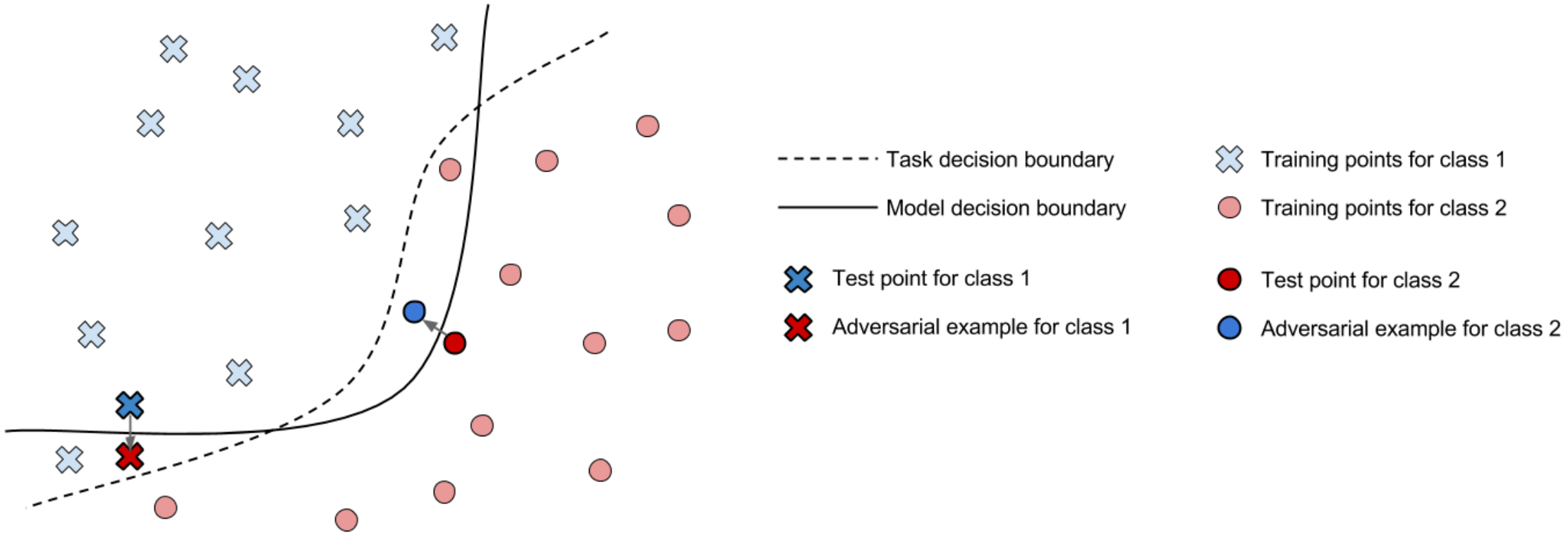
April 3, 2017

Cerber Ransomware evolves to evade detection by Machine Learning Solutions

Most malware and viruses have evolved with time and use disguise to conceal their identity. Even the most active CERBER family of ransomware has adopted a new technique to evade detection by machine learning solutions.

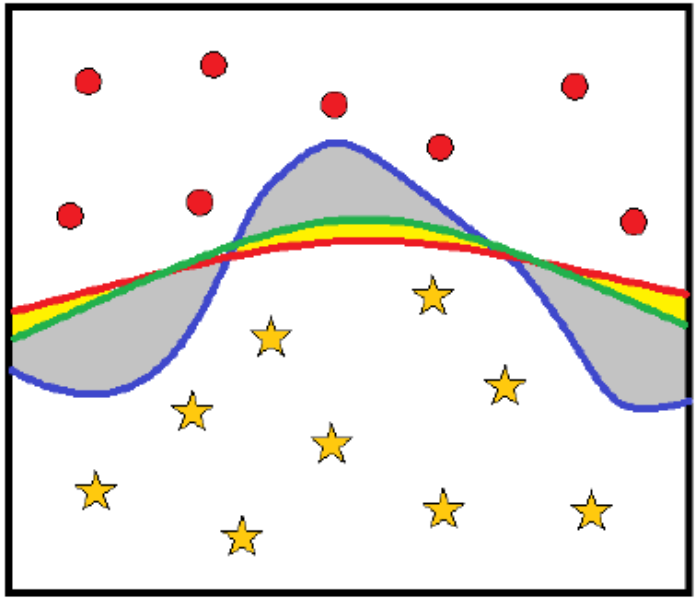


Evasion Attacks



http://www.cleverhans.io/security/privacy/ml/2016/12/15/breaking-things-is-easy.html

Enabling Black-Box Attacks…



Again… Transferability

Successful attacks against one machine learning system are often successful against similar ones.

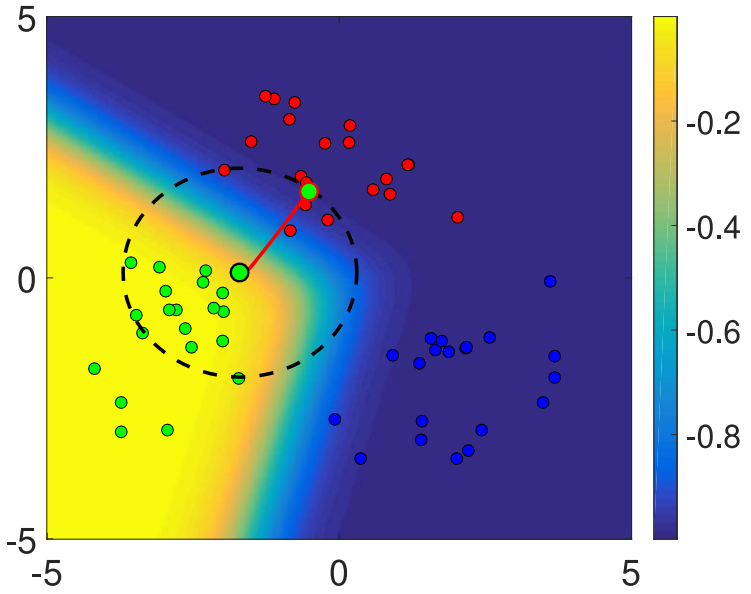
We can craft effective black-box attacks with:

- Surrogate models
- Surrogate datasets

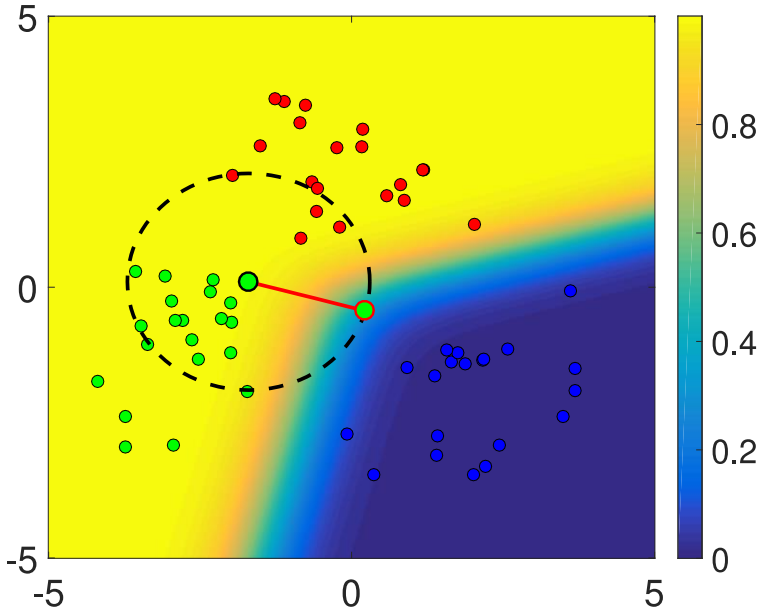
Nicolas Papernot, Patrick McDaniel, Ian Goodfellow. "*Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples*." ArXiv preprint arXiv:1605.07277, 2016.

Types of Evasion Attacks

Indiscriminate



Targeted



Types of Evasion Attacks (formulation)

Different formulations have been proposed in the research literature:

- **Minimum distance** attack strategies:

$$x^* = \arg\min_{x'} \; d(x', x) \quad \text{s.t.} \quad F(x') = y_t$$

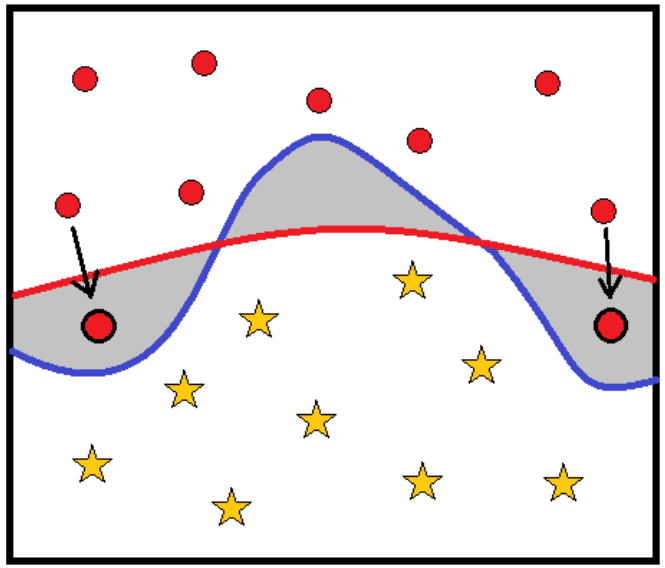
- Attacks with **budget constraints**:

$$x^* = \arg\max_{x'} \; \ell(f(x'), y) \quad \text{s.t.} \quad d(x', \, x) < \Gamma$$

- Approximations (**Fast Gradient Sign Method**):

$$x + \varepsilon \, \mathrm{sgn}(\nabla_x L(\theta, x, y))$$

Adversarial Training



- Re-train the network including adversarial examples in the training dataset.
- Can help to partially mitigate the problem.
- But you can't characterise all possible ***adversarial regions***.

Approaches:

- **min-max training**:

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

- **Ensemble adversarial training**: include adversarial examples from different machine learning models.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. "*Towards Deep Learning Models Resistant to Adversarial Attacks.*" ICLR, 2018.

Florian Tramèr, Alex Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel. "*Ensemble Adversarial Training: Attacks and Defences.*" ICLR, 2018.

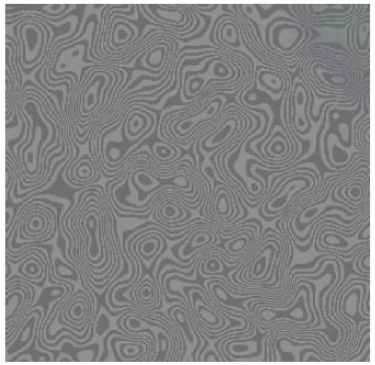
Universal Adversarial Perturbations

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. "*Universal Adversarial Perturbations*" CCVPR, pp. 86–94, 2017.

Adversarial Examples with Procedural Noise

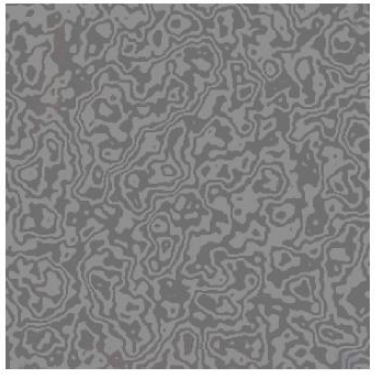
‘analog clock’ (28.53%)

‘barbell’ (29.84%)



‘fire truck’ (92.21%)

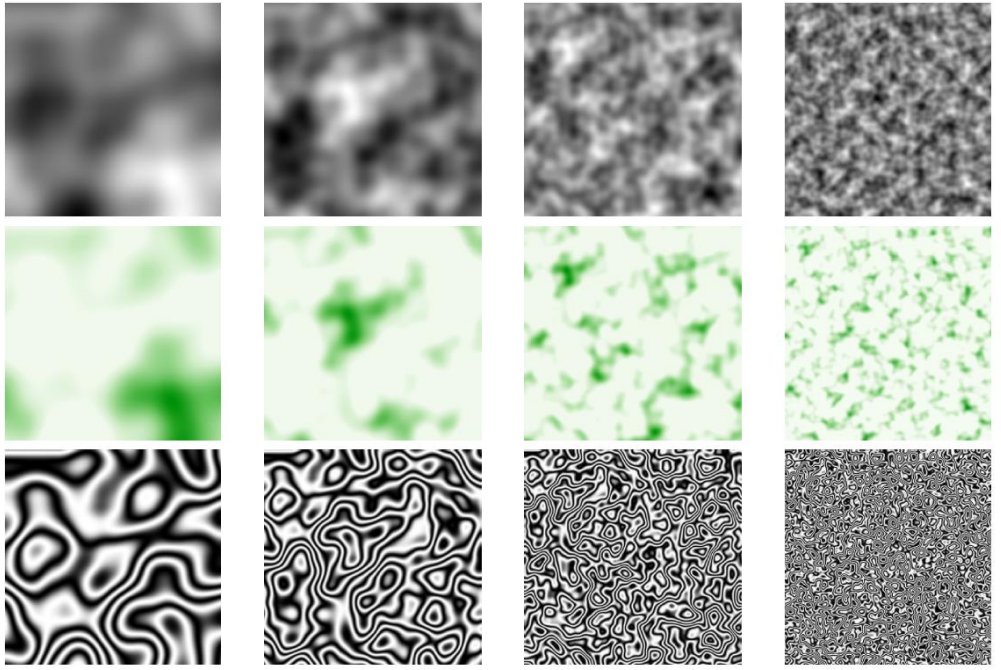
‘wall clock’ (18.32%)



K.T. Co, L. Muñoz-González, E.C. Lupu. “*Procedural Noise Adversarial Examples for Black-box Attacks on Deep Neural Networks*.” arXiv preprint, 2018.

Perlin Noise

- Developed to produce **natural-looking textures** for computer graphics.
- Relies on **pseudo-random gradients** to generate the noise patterns.
- It's **simple** and easy to use.
- Different noise patterns can be generated according to a **Noise Generating Function**.
- **Reduced number of parameters** to control the appearance of the noise patterns (4 in our case).
- We use **greyscale** colour-map.



Attack Formulation

$$\begin{aligned} \min_{\theta} \quad & F_{\tau(x)}(x + G(\theta)) - T_n(x + G(\theta)) \\ \text{s.t.} \quad & x + G(\theta) \in [0,1]^d, \quad \|G(\theta)\| < \varepsilon, \quad q < q_{\max} \end{aligned}$$

$F_{\tau(x)}(x)$ classifier's predicted label for sample x.

$T_n(x)$ *n*-th highest probability score for sample x.

$G(\theta)$ Perlin noise generating function parametrized by θ.

ε maximum perturbation allowed (according to some norm).

$q_{\max}$ maximum number of queries.

Attack Formulation

$$\begin{aligned} \min_{\theta} \quad & F_{\tau(x)}(x + G(\theta)) - T_n(x + G(\theta)) \\ \text{s.t.} \quad & x + G(\theta) \in [0,1]^d, \quad \|G(\theta)\| < \varepsilon, \quad q < q_{\max} \end{aligned}$$

- We use **Bayesian optimization** for black-box optimization of the parameters:
  - Matérn 5/2 covariance function for the Gaussian Process.
  - Expected Improvement as acquisition function.

- Enables **black-box attacks** aiming to reduce the number of queries.

Experimental Results

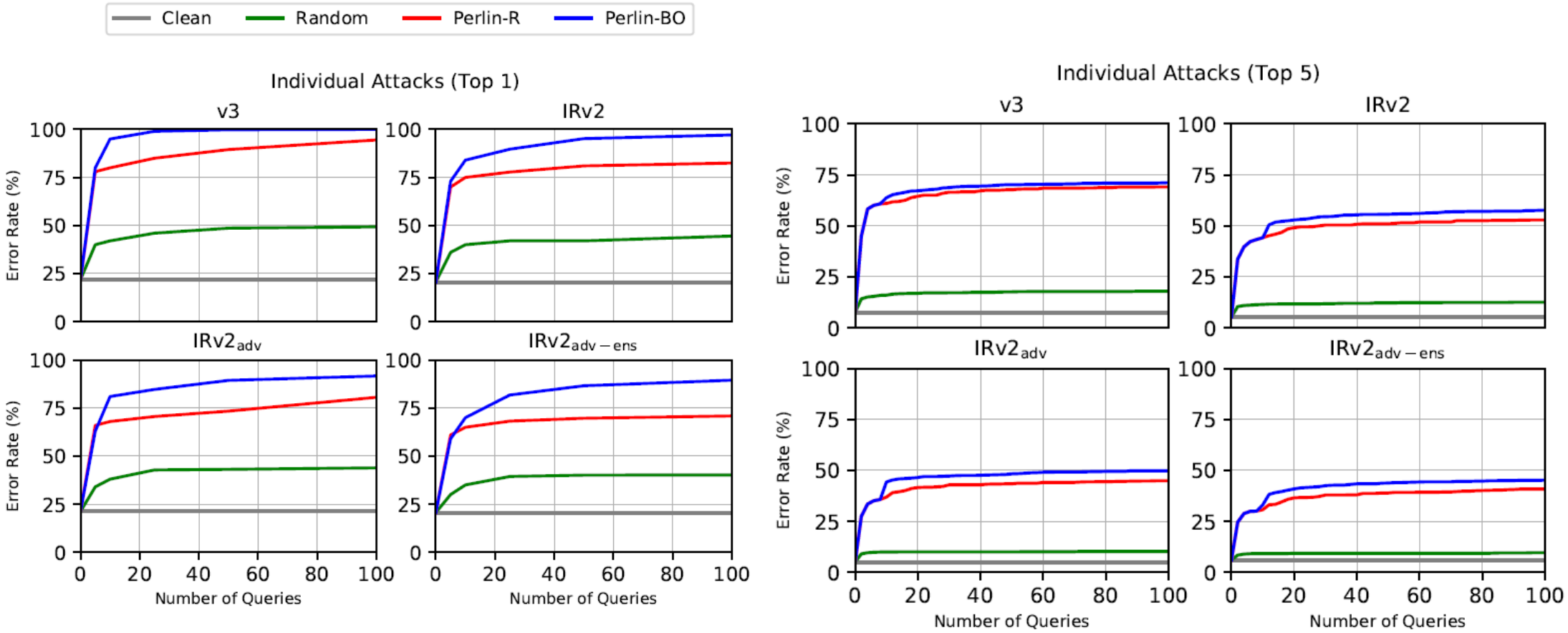
- **ImageNet** dataset (1,000 classes).
- Top 1 and Top 5 evasion attacks.
- **Adversarial training** is not effective against Perlin noise attacks.

ERROR RATES (IN %) FOR INDIVIDUAL ATTACKS ON IMAGENET. RESULTS ON 1,000 RANDOM VALIDATION SET SAMPLES WITH $\varepsilon = 16/256$ AND $q_{\text{MAX}} = 100$ PER SAMPLE. STRONGEST ATTACKS ON EACH CLASSIFIER ARE HIGHLIGHTED.

	Top 1				Top 5			
Classifier	Clean	Random	Perlin-R	Perlin-BO	Clean	Random	Perlin-R	Perlin-BO
v3	21.8	49.3	94.5	**100**	7.5	17.9	69.1	**71.2**
IRv2	20.5	44.5	82.5	**97.1**	5.3	12.5	52.9	**57.7**
IRv2_{adv}	21.2	43.9	80.6	**91.7**	5.0	10.2	44.9	**49.8**
$\text{IRv2}_{\text{adv-ens}}$	20.6	40.2	70.9	**89.5**	6.0	9.6	40.8	**45.2**

Experimental Results

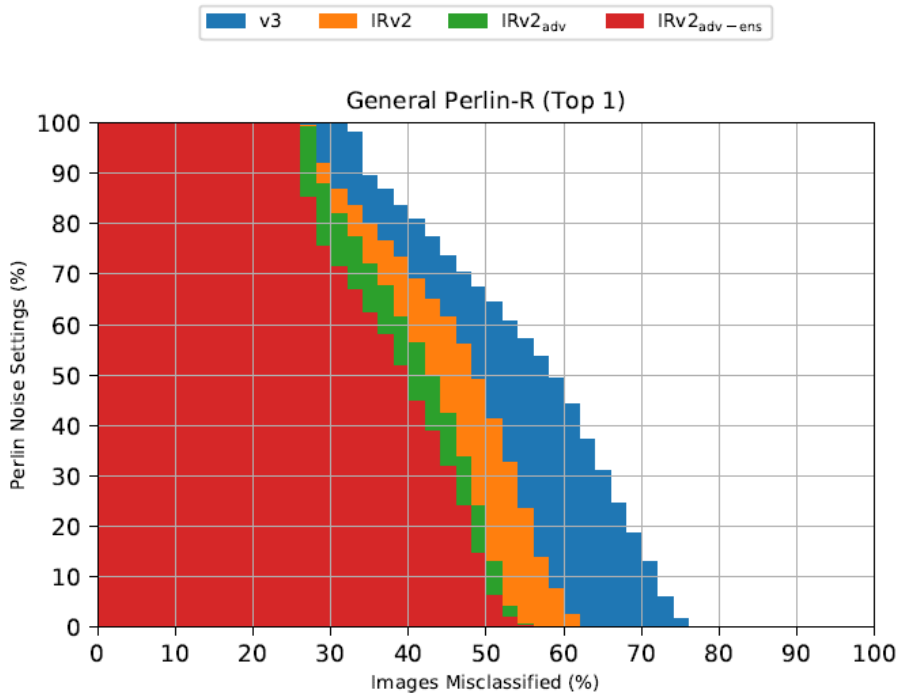
- Perlin noise attack just requires a **reduced number of queries** (compared to existing black-box attacks).



Experimental Results

- Perlin noise perturbations have **"universal properties"**: the same perturbation can be used to misclassify many samples at the same time.

Random perturbations



Optimized perturbations

ERROR RATES (IN %) FOR GENERALIZED PERLIN-BO ATTACKS ON A TRAINING SET SIZE OF 100.

	Top 1		Top 5	
Classifier	Clean	BO-100	Clean	BO-100
v3	23.0	73.8	7.1	59.1
IRv2	20.4	61.1	5.0	42.5
$IRv2_{adv}$	20.2	52.0	5.2	32.2
$IRv2_{adv\text{-}ens}$	20.3	46.3	5.4	27.0

Experimental Results

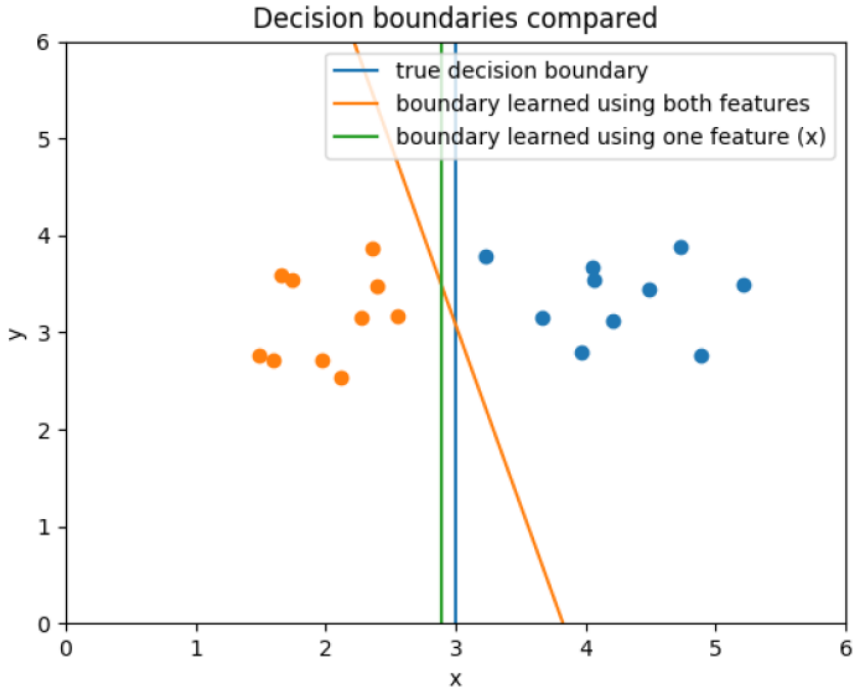
- Perlin noise attack **outperforms** both (state-of-the-art) **white and black-box attacks** against ImageNet.
- The attack also shows that **adversarial training is not really effective** against adversarial examples when the attacker changes the perspective of the attack.

COMPARISON BETWEEN OUR INDIVIDUAL PERLIN ATTACKS AND THE FAST GRADIENT ATTACKS AS DONE IN TRAMER ET AL. [21]. ERROR RATES (IN %) SHOW THE ATTACK'S PERFORMANCE AS THE **DIFFERENCE** BETWEEN THE EVASION RATE AND CORRESPONDING ERROR ON CLEAN IMAGES.

	Top 1				Top 5			
Classifier	Perlin-R	Perlin-BO	FGA-W	FGA-B	Perlin-R	Perlin-BO	FGA-W	FGA-B
v3	72.7	**78.2**	64.3	47.6	61.6	63.7	**64.1**	36.6
IRv2	62.5	**76.6**	50.3	21.1	47.6	**52.4**	40.6	19.2
$IRv2_{adv}$	59.4	**70.5**	21.8	14.7	39.9	**44.8**	11.6	5.6
$IRv2_{adv\text{-}ens}$	50.3	**69.9**	†44.6	6.8	34.8	**39.2**	†32.0	2.8

Florian Tramèr, Alex Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel. "*Ensemble Adversarial Training: Attacks and Defences.*" ICLR, 2018.

Mitigation of Evasion Attacks through Feature Selection



Z. Bao, L. Muñoz-González, E.C. Lupu. "*Mitigation of Evasion Attacks through Embedded Feature Selection*." IEEE Trans. on Cybernetics (under review), 2018.

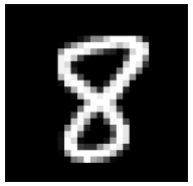
Mitigation of Evasion Attacks through Feature Selection

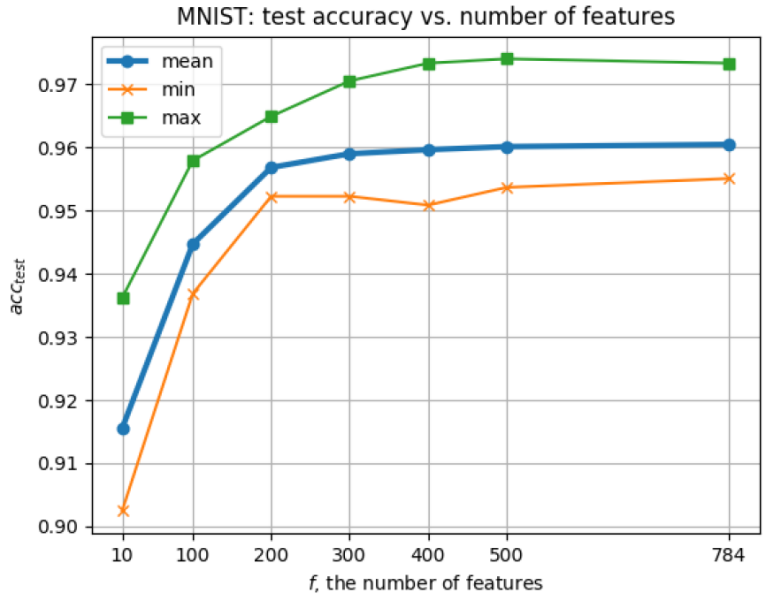
- Related work claimed that **feature selection** makes algorithms **less secure** against evasion attacks:

  - F. Zhang, P.P. Chan, B. Biggio, D.S. Yeung, F. Roli. "*Adversarial Feature Selection against Evasion Attacks*." IEEE Transactions on Cybernetics, vol. 46, no. 3, pp. 766–777, 2016.
  - B. Biggio, G. Fumera, F. Roli. "*Security Evaluation of Pattern Classifiers under Attack.*" IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 4, pp. 984–996, 2014.
  - F. Wang, W. Liu, S. Chawla, "*On Sparse Feature Attacks in Adversarial Learning*." International Conference on Data Mining, pp. 1013–1018, 2014.

Mitigation of Evasion Attacks through Feature Selection

- Effects of embedded feature selection with Lasso in the security of the machine learning system.

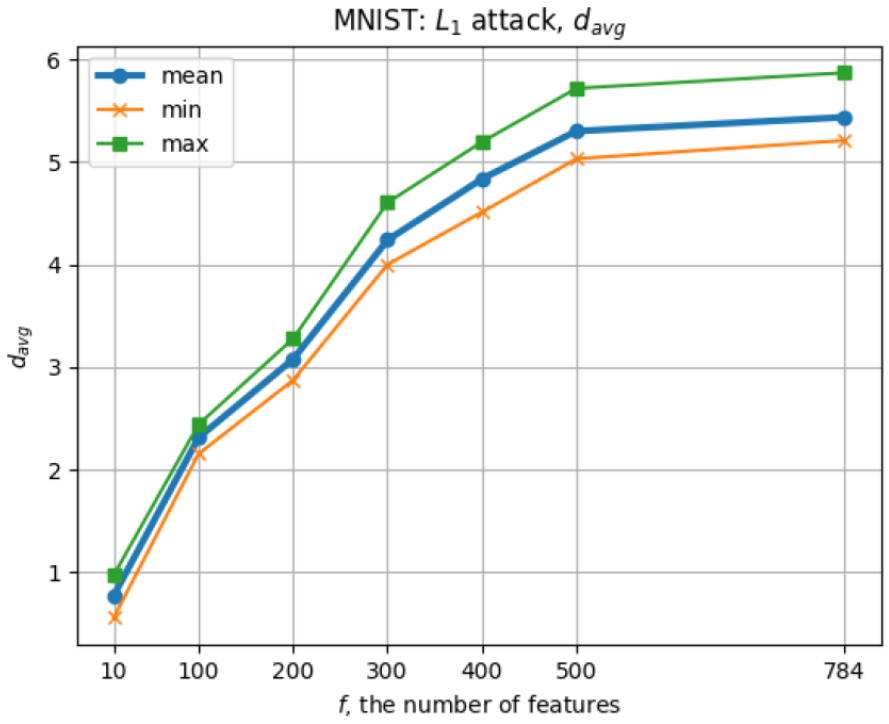


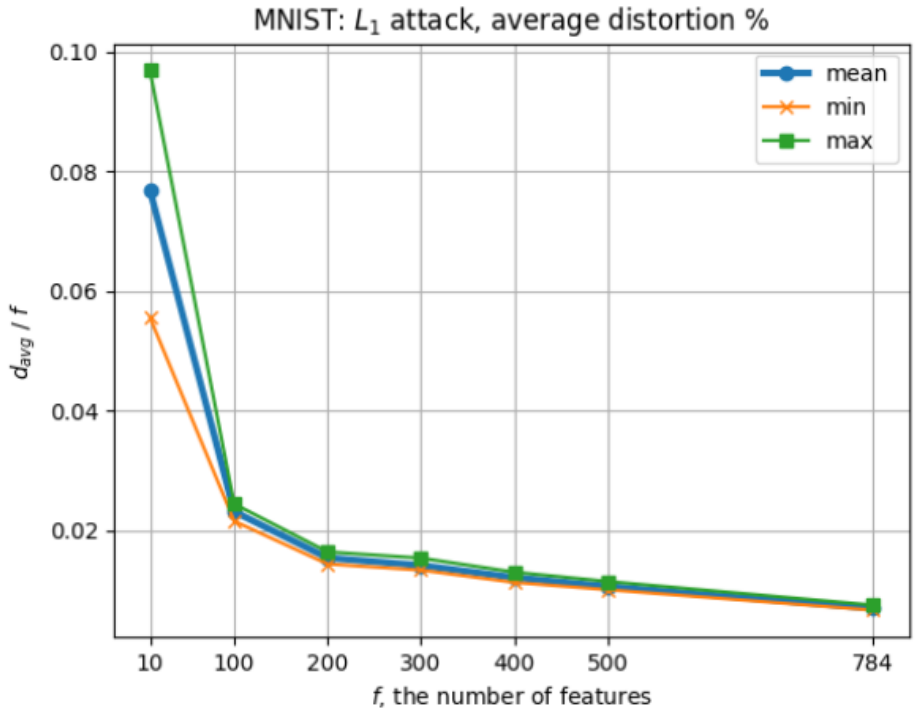


Lasso:

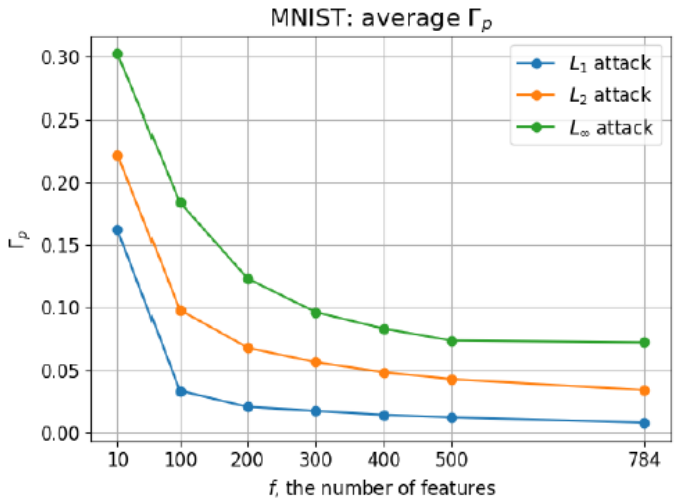
Training Cost + ʎ |w|

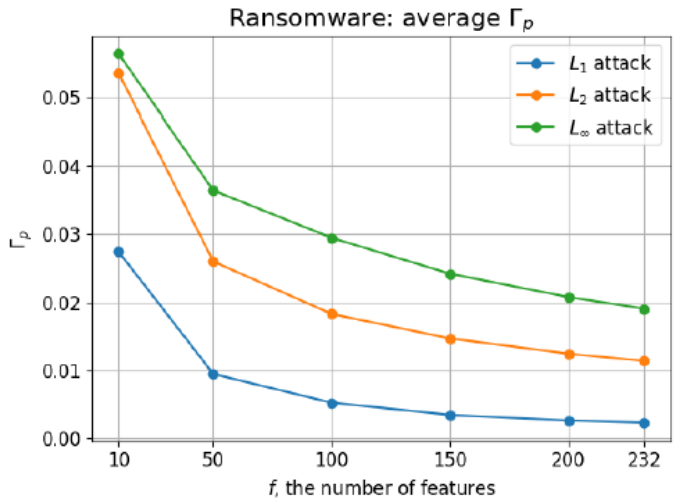
But… Is Feature Selection more Secure?

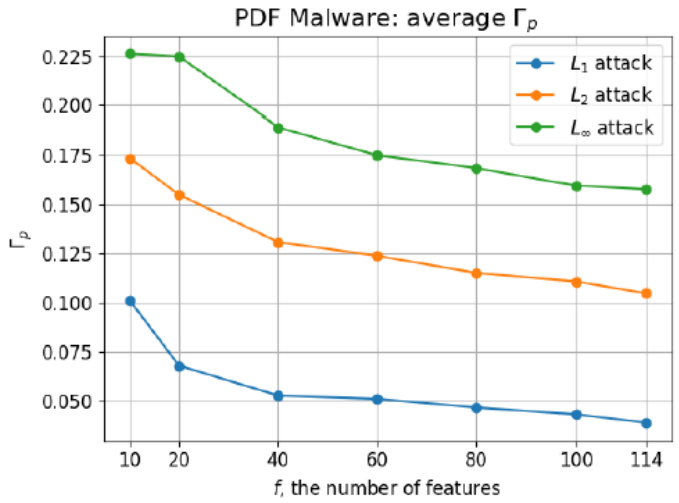




But… Is Feature Selection more Secure?

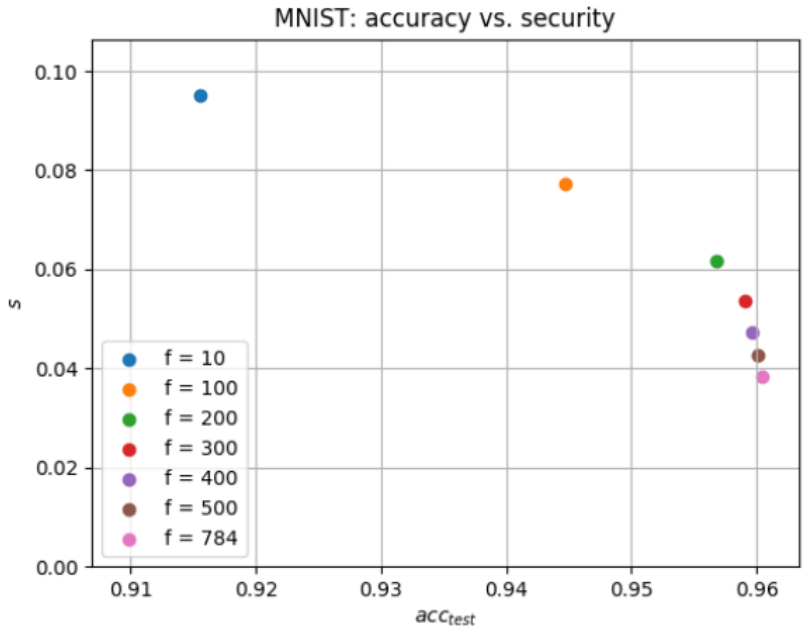






Γ_p Normalised perturbation: depending on the norm for the attacker's constraints.

Trade-off Accuracy vs Security



Security defined as a function of the average (normalized) distortion of the adversarial examples:

$$s = \frac{1}{|P|} \sum_{p \in P} \frac{\Gamma_p}{f^{1/p}}$$

Statistical Analysis of Adversarial Examples

We used Maximum Mean Discrepancy (MDD) to measure the distance between genuine and adversarial examples:

$$\widehat{D}(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i, x_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(x_i, y_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} K(y_i, y_j)$$

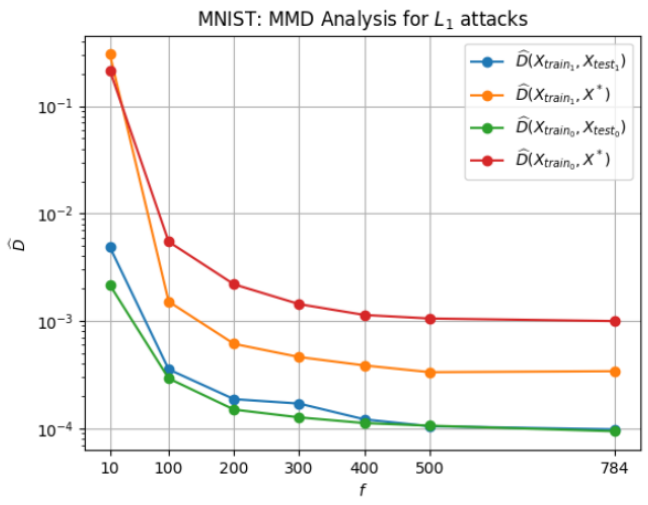
As proposed in: K. Grosse, P. Manoharan, N. Papernot, M. Backes, P. McDaniel. "*On the Statistical Detection of Adversarial Examples.*" ArXiv preprint: arXiv:1702.06280, 2017.

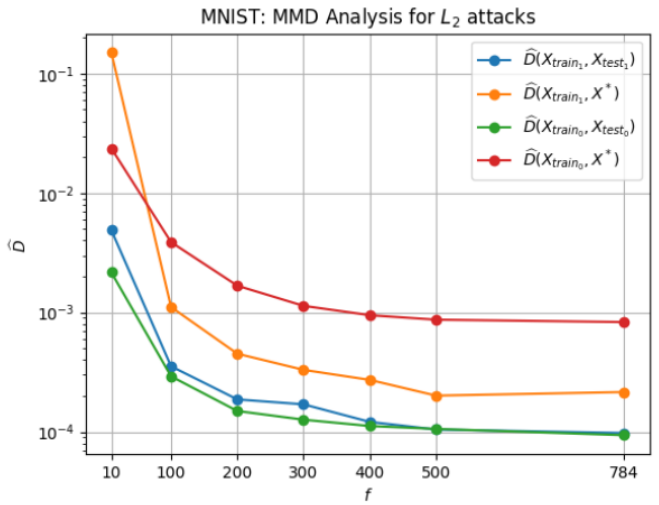
In our case we used a normalized linear kernel (doesn't make assumptions about the underlying data distribution):

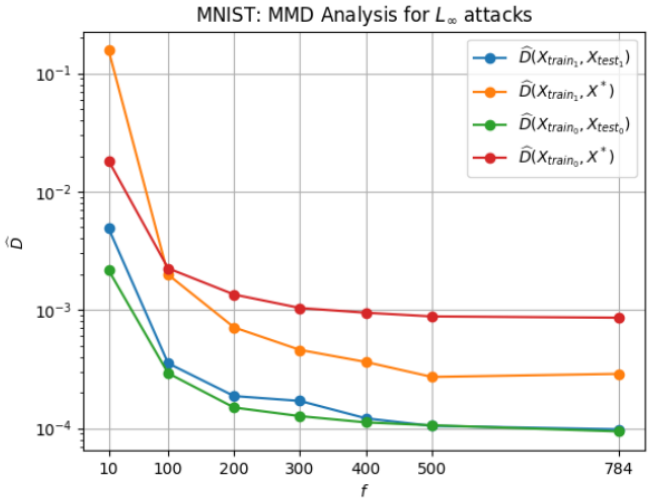
$$K(x,y) = \frac{x^\intercal y}{\|x\|_1 \|y\|_1}$$

Statistical Analysis of Adversarial Examples

- Adversarial examples are **easier to detect when using reduced feature sets**.







Conclusion

- **Machine Learning systems are vulnerable:**
  - Poisoning attacks (training time).
  - Evasion attacks (test time).
- **We need to understand the vulnerabilities:**
  - Worst-case attacks.
  - Realistic attacker models.
  - Look at the whole system pipeline.
- **We need to understand how we can defend against these vulnerabilities:**
  - Some defences have already been proposed but sometimes are not effective if the attacker targets the defensive algorithm itself.
  - Quite an open research problem.
- **But… How can we test the security of machine learning systems?**
  - We need new design and testing methodologies.
  - Analysis of worst-case scenarios.
  - Verification vs testing.

Imperial College London

Thank you!

RISS
Resilient Information Systems Security

www.rissgroup.org

Contact: Luis Muñoz-González
l.munoz@imperial.ac.uk
https://www.imperial.ac.uk/people/l.munoz-gonzalez