Machine Learning, Security, and Practice: a Reflection

Giovanni Apruzzese

University of Genova – November 13th, 2023

UNIVERSITÄT LIECHTENSTEIN

whoami: Dr. **Giovanni Apruzzese**

- **Background:**
  - Did my academic studies (BSc, MSc, PhD) @ University of Modena, Italy.
    - Supervisor: Prof. Michele Colajanni
  - In 2019, spent 6 months @ Dartmouth College, USA.
  - Joined the University of Liechtenstein in July 2020 as a PostDoc Researcher.
    - Supervisor: Prof. Pavel Laskov
  - Was "promoted" to Assistant Professor in September 2022.

- **Interests:**
  - [Areas] Cybersecurity, machine learning, with a strong focus on practice
  - [Applications] Phishing, human factors, and any network-related topic (+🎮)
  - I like talking, researching and teaching – in a "blunt" way ☺

- **Contact information**:
  - Email (work): giovanni.apruzzese@uni.li
  - Website (personal): www.giovanniapruzzese.com
  - Feel free to contact me if you have any questions.
    - I reply fast, and will happily do so!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What I do

Machine Learning + Cybersecurity

- Applying ML to *provide security* of a given information system
  - E.g.: using ML to detect cyber threats
- *Attacking / Defending* ML applications
  - E.g.: evading an ML model that detects phishing websites
- Using machine learning *offensively…*
  - …against another system (e.g.: artificially generating "fake" images)
  - …against humans (e.g., violating privacy, deceiving end-users)

BONUS

- Using ML to attack an ML-based security system and harden it

UNIVERSITÄT LIECHTENSTEIN

(more recently)
Human factors in ML & Cybersecurity

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Outline of Today

Two paper-inspired talks:

- Machine Learning Security in the Real-World

  Ref: Giovanni Apruzzese, David Freeman, Savino Dambra, Hyrum S Anderson, Kevin Alexander Roundy, Fabio Pierazzi "'Real Attackers Don't Compute Gradients': Bridging the Gap Between Adversarial ML Research and Practice." IEEE Conference on Secure and Trustworthy Machine Learning (2023).

- Attacking Machine Learning-based Phishing Website Detectors

  Ref: Jehyun Lee, Zhe Xin, Melanie Ng Pei See, Kanav Sabharwal, Giovanni Apruzzese, Dinil Mon Divakaran "Attacking Logo-based Phishing Website Detectors with Adversarial Perturbations". European Symposium On Research In Computer Security (2023).

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Outline of Today

Two paper-inspired talks:

- Machine Learning Security in the Real-World

  Ref: Giovanni Apruzzese, David Freeman, Savino Dambra, Hyrum S Anderson, Kevin Alexander Roundy, Fabio Pierazzi "'Real Attackers Don't Compute Gradients': Bridging the Gap Between Adversarial ML Research and Practice." IEEE Conference on Secure and Trustworthy Machine Learning (2023).

- Attacking Machine Learning-based Phishing Website Detectors

  Ref: Jehyun Lee, Zhe Xin, Melanie Ng Pei See, Kanav Sabharwal, Giovanni Apruzzese, Dinil Mon Divakaran "Attacking Logo-based Phishing Website Detectors with Adversarial Perturbations". European Symposium On Research In Computer Security (2023).

Two goals:

- Inspire you (to do/consider doing research in computer security)
- Entertain you (research should be fun)

Machine Learning Security in the Real-World

Based on a joint work with Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, Kevin Roundy:
"'Real Attackers Don't Compute Gradients': Bridging the Gap Between Adversarial ML Research and Practice."
IEEE Conference on Secure and Trustworthy Machine Learning (2023).
)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Backstory (Dagstuhl – July 10-15th, 2022)

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

- Research seminar on the "Security of Machine Learning"

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Backstory (Dagstuhl – July 10-15th, 2022)

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

- Research seminar on the "Security of Machine Learning"

- The seminar opened with a talk by K. Grosse, showcasing the results of an extensive survey with ML practitioners about the security of ML [5]:

"Why do so?"

UNIVERSITÄT LIECHTENSTEIN

[5] K. Grosse , et al ""Why do so?"–A Practical Perspective on Machine Learning Security," ICML—New Frontiers of Adversarial Machine Learning, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Backstory (Dagstuhl – July 10-15th, 2022)

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

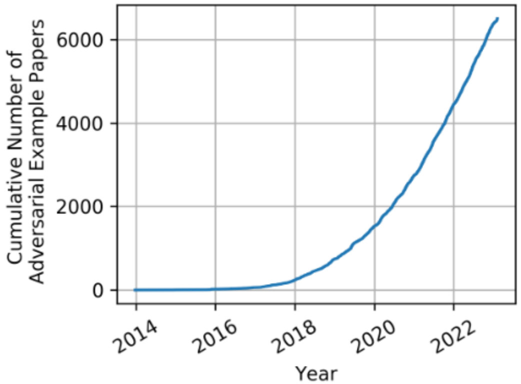
- Research seminar on the "Security of Machine Learning"

- The seminar opened with a talk by K. Grosse, showcasing the results of an extensive survey with ML practitioners about the security of ML [5]:

"Why do so?"

- Many discussions revolved around the impact of our research to the real world.

Apparently, the overwhelming number of works on adversarial ML research were not seen as problematic by practitioners!



- A recurring observation by some of the seminar's attendees from industry was that:

UNIVERSITÄT LIECHTENSTEIN

"Real attackers *guess*"

[5] K. Grosse , et al ""Why do so?"–A Practical Perspective on Machine Learning Security," ICML—New Frontiers of Adversarial Machine Learning, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Backstory (Earth – July 22nd, 2022)

- One week later, I was having a (remote) call with Fabio Pierazzi, and…

Dagstuhl follow-up: position paper on "attacker guessing" threat model?

Reply | Reply All | Forward

PF Pierazzi, Fabio <fabio.pierazzi@kcl.
To dfreeman; Kevin Roundy; hyrum@robustintelligence.com
Cc Apruzzese Giovanni
venerdì 22/07/2022 14:15

You forwarded this message on 02/09/2022 15:45.

Dear David, Kevin, Hyrum,

It was great to get to know you (more) during Dagstuhl.

I was talking with Giovanni yesterday, and were thinking again about what you all seemed to agree on from an industry perspective that in most cases attackers "guess" and do not necessarily use ML to evade systems, they just try to get out the easy way.

Given the upcoming first edition of SATML, we saw there's also a category for "position papers", and me and Giovanni were thinking of maybe doing a position paper about "threat models of ML systems".

The current white-box threat models and also ML-driven black-box are mostly a worst-case scenario, and maybe models can be broken just much more easily (similar to the 'pseudo-fuzzing' that Hyrum is looking into for ML models at Robust intelligence and maybe at Microsoft research).

Long story short, would you be interesting in co-authoring a position paper for SatML on the topic of "revisiting threat models of ML systems", to also re-define how to consider attacker capabilities in evading systems? Part of it is also related to the fact that real-world systems are a pipeline of ML and non-ML models.

Or, if not co-authoring, giving some feedback?

More concretely, there is some stuff that should be nice to highlight:

- In this mlsec challenge, authors evaded an ml classifier without ml: https://cujo.com/announcing-the-winners-of-the-2021-machine-learning-security-evasion-competition/
- In Giovanni&Pavel's 5G paper, they proposed the "myopic" threat model, similar to this issue: https://arxiv.org/pdf/2207.01531.pdf
- Konrad's team which won a defense in Hyrum's ML challenge got broken by a non-ML approach: https://arxiv.org/pdf/2010.09569.pdf

We appreciate the timeline is quite tight: deadline is Sep 1st (with abstract the week before), yet it's a 5-page position paper, and it may help in raising awareness on threats relevant to industry.

Giovanni offered himself to do most of the work, so he should be able to lead the effort.

What do you think?

> We appreciate the timeline is quite tight: deadline is Sep 1st (with abstract the week before), yet it's a 5-page position paper, and it may help in raising awareness on threats relevant to industry.

Our paper has 26 pages!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

FGSM (Fast Gradient Sign Method)

$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

FGSM (Fast Gradient Sign Method)

$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$

Do real attackers compute gradients?

A real attacker

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Do real attackers compute gradients? (Case Study)

- We tried answering this question by looking at the AI Incident Database [78]…
- …but **we could not find any evidence** of real incidents stemming from "adversarial examples" (or which leverage gradient computations)

UNIVERSITÄT LIECHTENSTEIN

[78]: https://incidentdatabase.ai/

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Do real attackers compute gradients? (Case Study)

- We tried answering this question by looking at the AI Incident Database [78]…
- …but **we could not find any evidence** of real incidents stemming from "adversarial examples" (or which leverage gradient computations)

- So, we asked a well-known **cybersecurity company** to provide us with data from their (operational!) phishing website detector, empowered by *deep learning*

- Just in July 2022, there were **9K samples** for which the ML detector was "uncertain"
  - They were "close to the decision boundary", and required manual triage by experts

- We **manually analyzed** these (phishing) samples, trying to understand the root-causes of these "adversarial webpages"

What did we find?

UNIVERSITÄT LIECHTENSTEIN

[78]: https://incidentdatabase.ai/

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Do real attackers compute gradients? (Case Study) [cont'd]

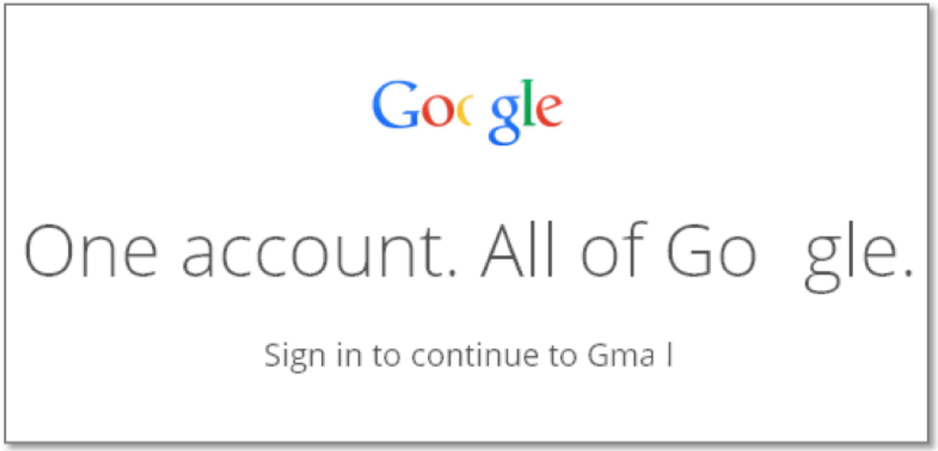
- The **vast majority** of these webpages were "out of distribution"
  - They were different from any sample in the training set
- We then looked at a small subset of the remaining ones…

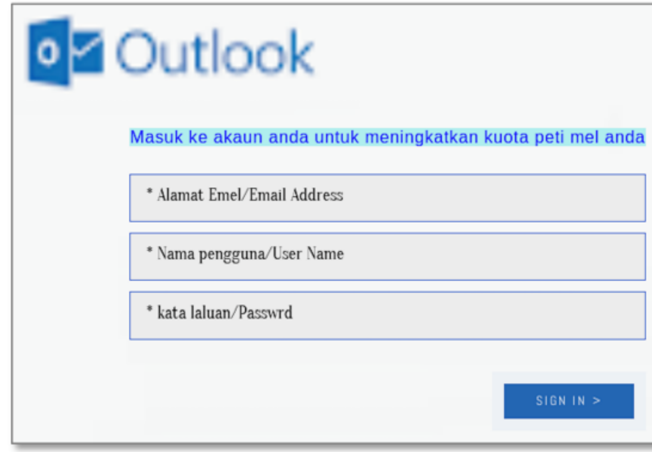
UNIVERSITÄT LIECHTENSTEIN

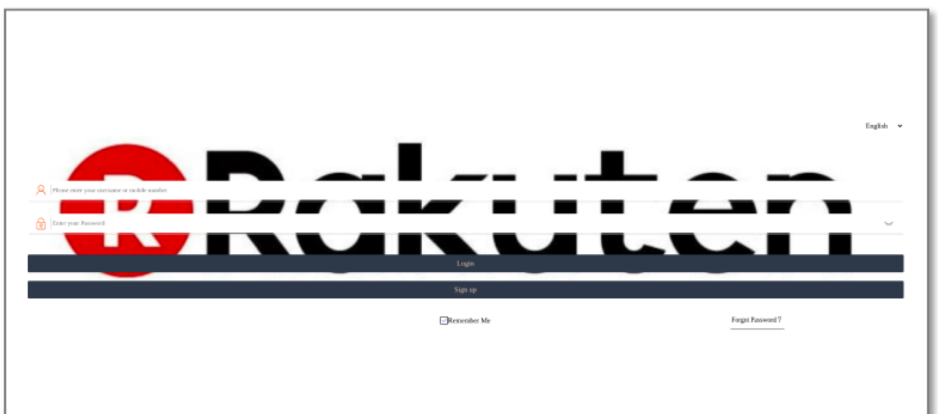
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

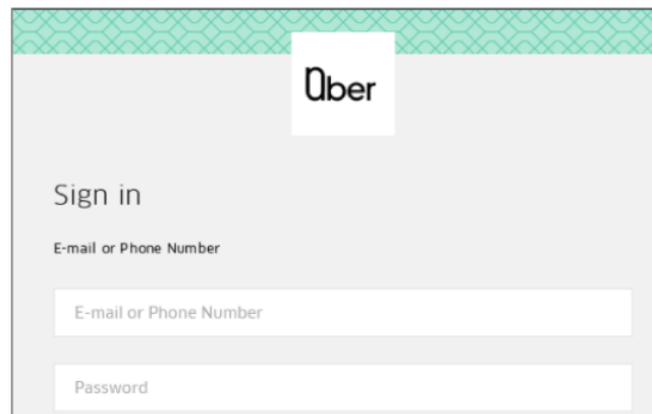
Do real attackers compute gradients? (Case Study) [cont'd]

- The **vast majority** of these webpages were "out of distribution"
  - They were different from any sample in the training set
- We then looked at a small subset of the remaining ones…









These techniques have been known for decades…
but can still evade modern (and real) *ML systems*.

And they're cheap!

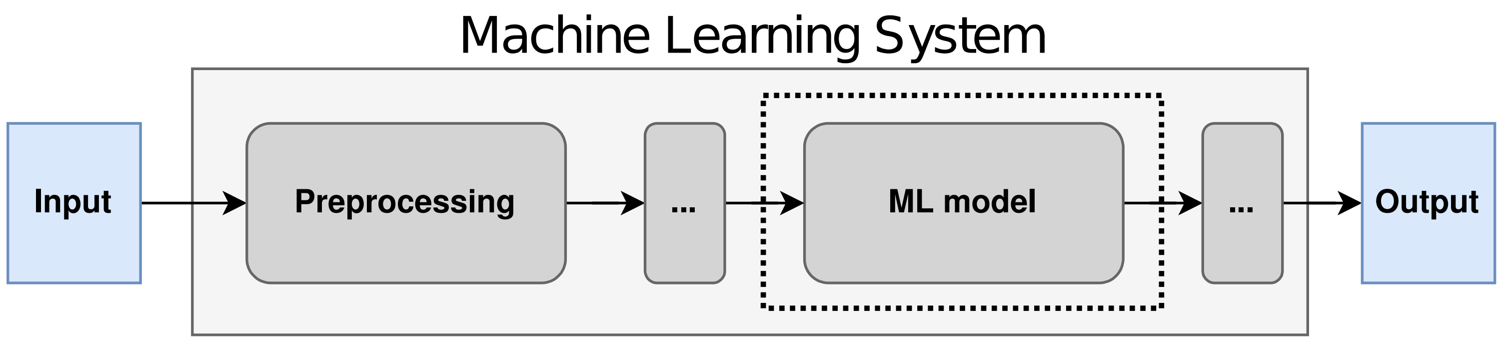
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems

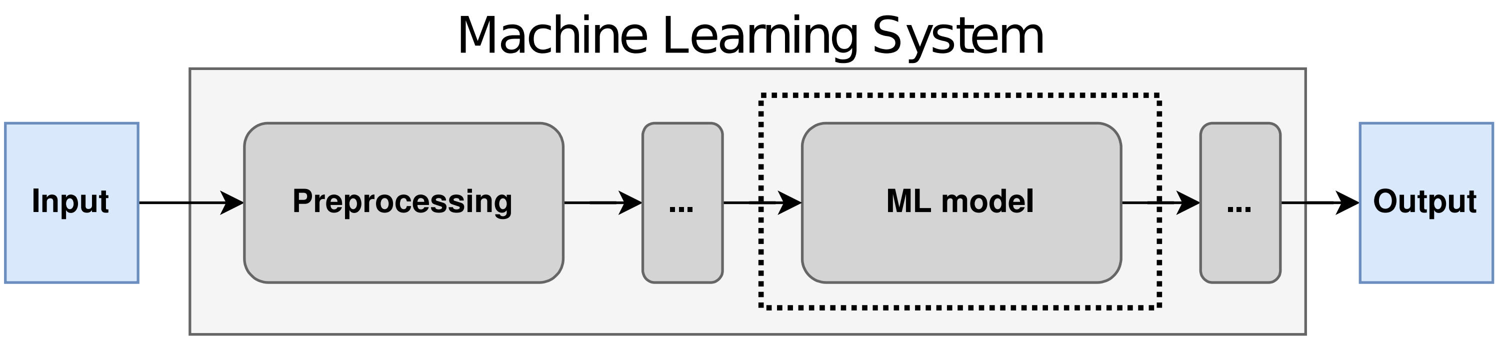
- In reality, ML models are a single component of a complex ML system
  - Real ML systems (are likely to) have also elements *that have nothing to do with ML*



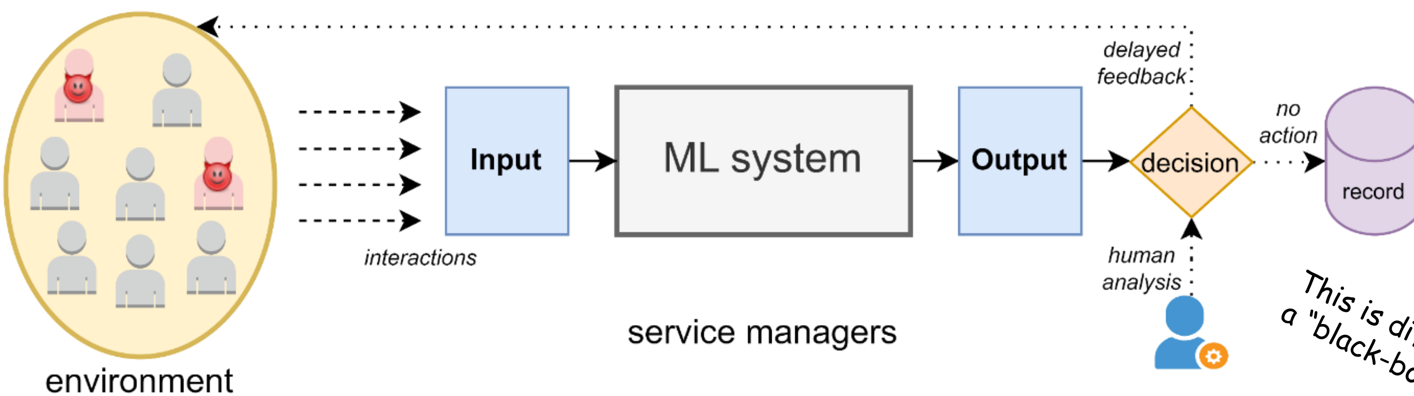
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems

- In reality, ML models are a single component of a complex ML system
  - Real ML systems (are likely to) have also elements *that have nothing to do with ML*



- Some ML systems are "invisible" to their users (and, hence, to real attackers)



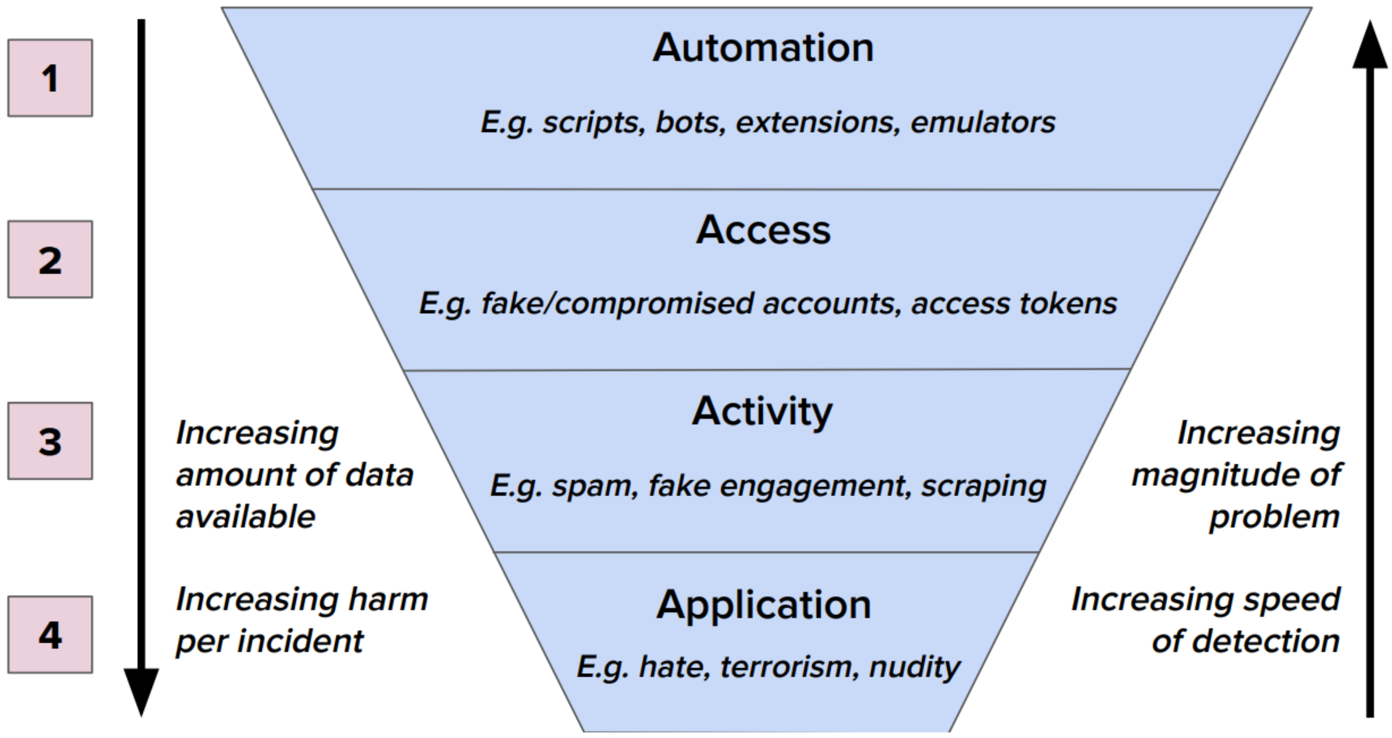
This is different from a "black-box" scenario!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems (Case Study)

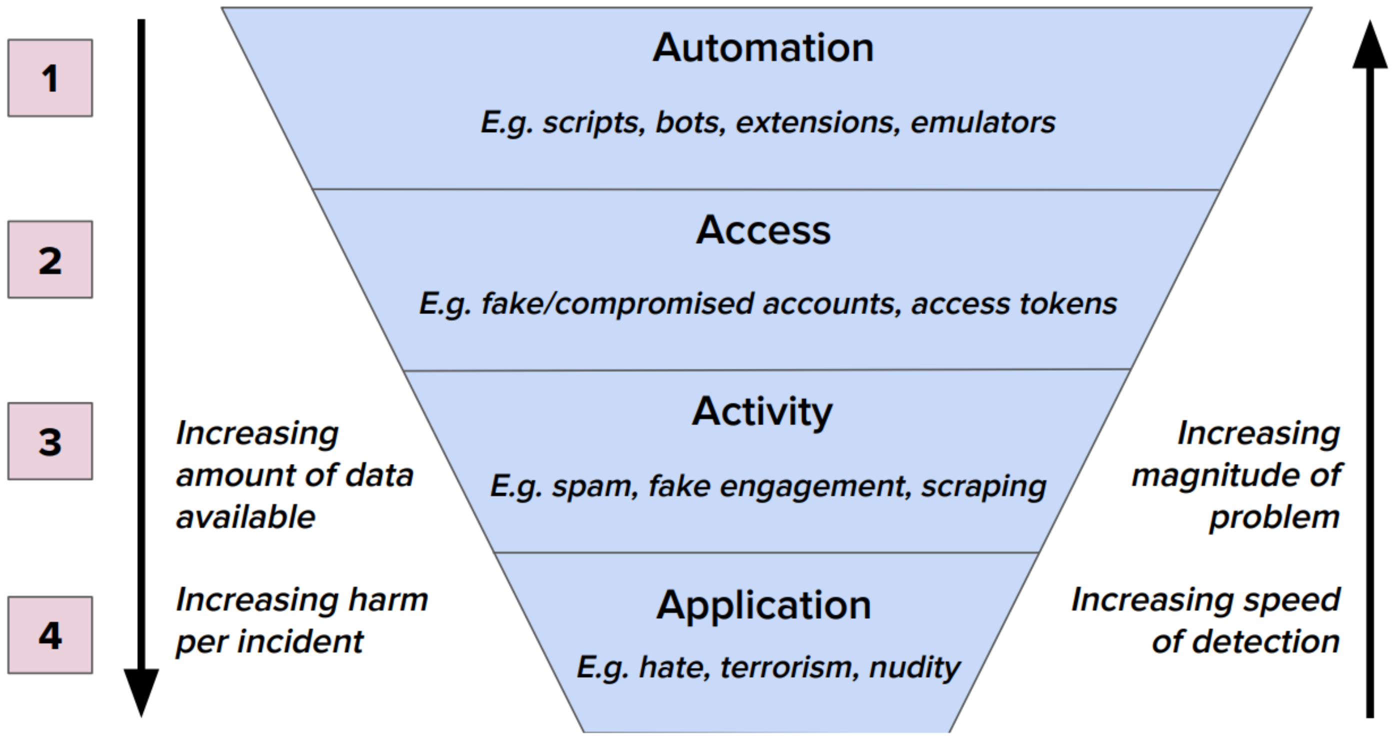
- This is the architecture of the ML-based spam detection system at **Facebook**



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems (Case Study)

- This is the architecture of the ML-based spam detection system at **Facebook**



- The first layers are meant to block attacks *at scale* (e.g., query-based strategies)
- All layers use a mix of ML and non-ML techniques (not necessarily deep learning)
- Deep learning really shines at the bottom layer (few events reach this layer, though)
- The output accounts for diverse layers and is not instantaneous (an *invisible* ML system)

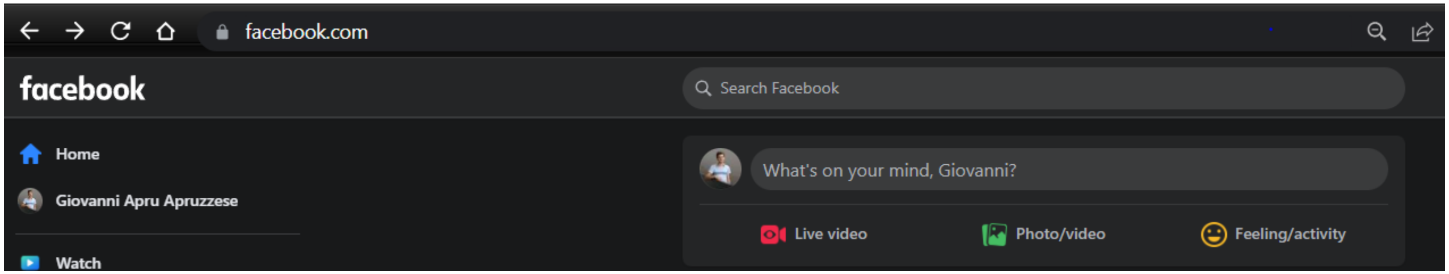
Real attackers have to bypass all layers to be successful.

This does not mean that this ML system is omnipotent!

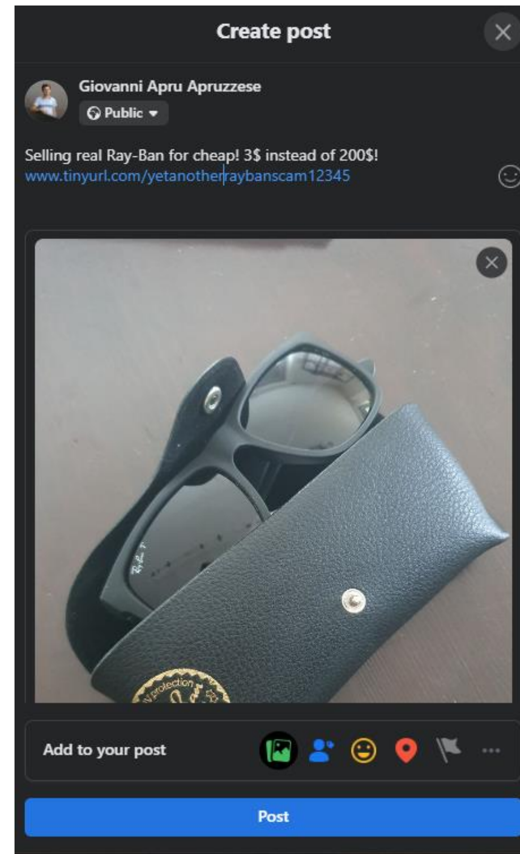
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

"Attacking" an *invisible* ML system

- If I go on Facebook and want to spread "spammy" content…



- …the only thing I will see after "posting" it is the post itself.

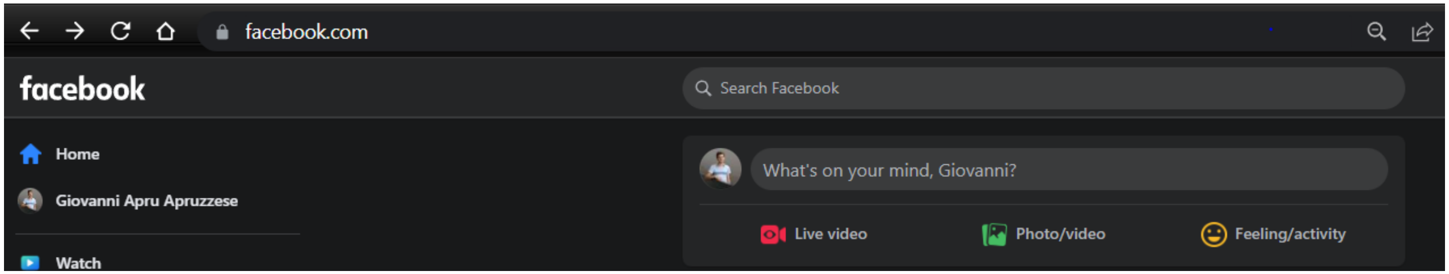


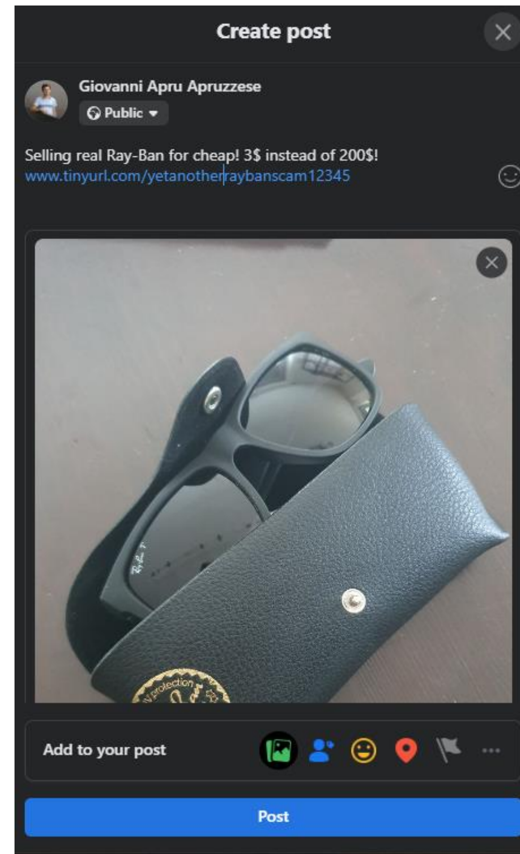
UNIVERSITÄT
LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

"Attacking" an *invisible* ML system (cont'd)

- If I go on Facebook and want to spread "spammy" content…





- …the only thing I will see after "posting" it is the post itself.

- I would not be able to see:
  - The architecture of Facebook's spam detector
  - The fact that it uses ML
  - The fact that my specific post was (or not) analyzed by ML
  - The output of the system to my specific post

- If the post "appears", does it mean that the system was evaded?
  - What if the post gets removed after 1 hour? Or 1 day?
  - What if my account is blocked after 1 week?

A real attacker does not know any of this!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems (state-of-research)

- We analyzed all related papers accepted at top-4 cybersecurity conferences (NDSS, S&P, CCS, USENIX Sec) from 2019-2021.
  - Out of 1549 papers, 88 fell into the "adversarial ML" category.
    - Out of these, 78 consider *only* deep learning methods

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Machine Learning Systems (state-of-research)

- We analyzed all related papers accepted at top-4 cybersecurity conferences (NDSS, S&P, CCS, USENIX Sec) from 2019-2021.
  - Out of 1549 papers, 88 fell into the "adversarial ML" category.
    - Out of these, 78 consider *only* deep learning methods

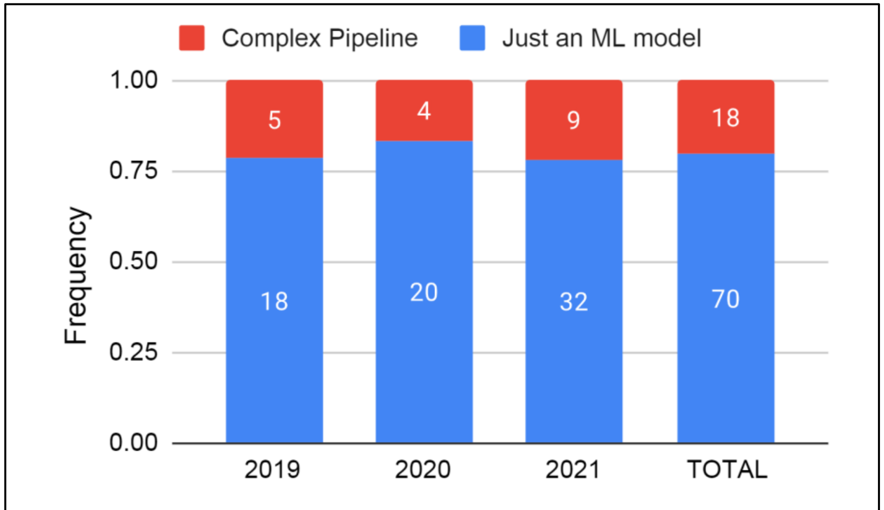


Fig. 12: Has a complex *pipeline* been reproduced in the evaluation?

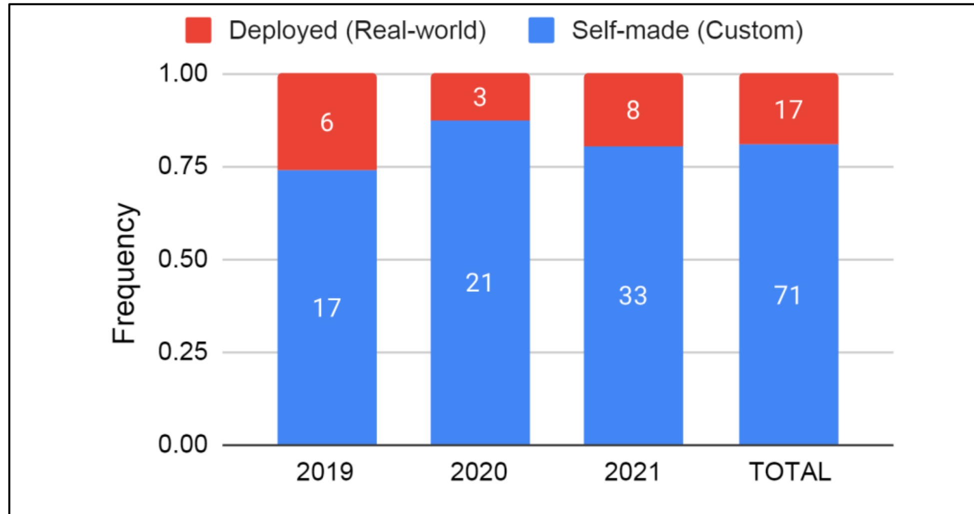


Fig. 13: Does the paper consider an ML model *deployed* in the real world?

Building a pipeline that resembles a (realistic) ML system is difficult.

These assets are not publicly available!

Finding a ML system that is openly available for research-focused (security) assessments is hard.

Getting in touch with companies is tough!

UNIVERSITÄT LIECHTENSTEIN

Disclaimer: the findings of all these papers are still significant!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cybersecurity is rooted in *economics*

UNIVERSITÄT
LIECHTENSTEIN

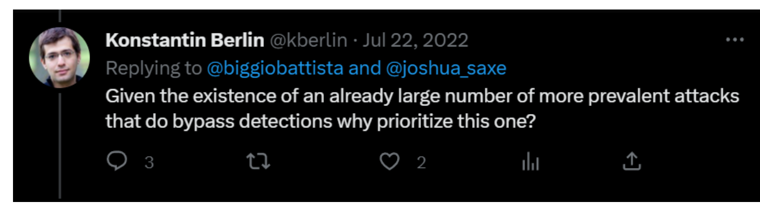
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cybersecurity ⇔ Economics

- Given enough resources, any attack will be successful
- The goal of a defense is to "raise the bar" for the attacker

"There is no such a thing as a foolproof system."

→ A real attacker will opt for the **cheaper** strategy to reach their objective
→ A real defender will prioritize the **most likely** threats.



UNIVERSITÄT LIECHTENSTEIN

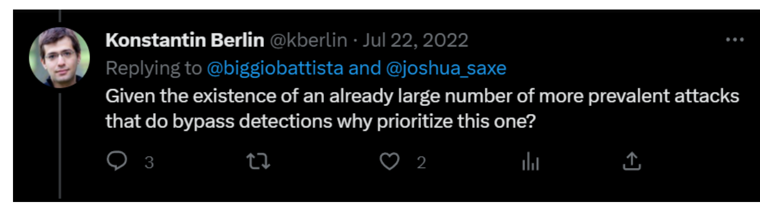
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cybersecurity ⇔ Economics

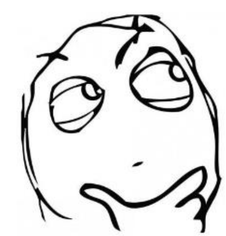
- Given enough resources, any attack will be successful
- The goal of a defense is to “raise the bar” for the attacker

"There is no such a thing as a foolproof system."

→ A real attacker will opt for the **cheaper** strategy to reach their objective
→ A real defender will prioritize the **most likely** threats.



- In our domain, the **cost** of an attack is typically measured by means of “queries”
  - More queries → higher cost → “less effective” attack



UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cybersecurity ⇔ Economics (Case Study)

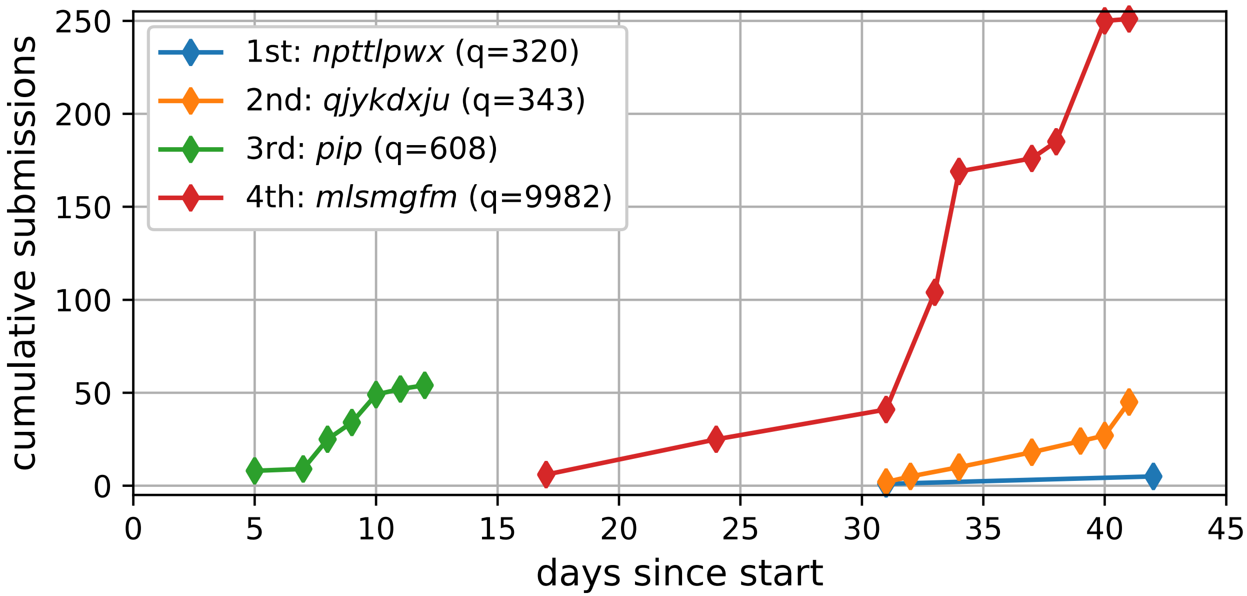
- We performed an in-depth look at the MLSEC anti-phishing challenge of 2021
  - Participants had to “evade the black-box detector” with as few queries as possible

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cybersecurity ⇔ Economics (Case Study)

- We performed an in-depth look at the MLSEC anti-phishing challenge of 2021
  - Participants had to "evade the black-box detector" with as few queries as possible



- The team arriving <u>first</u> (320 queries)… was **the last** to submit their solution
- The team arriving <u>third</u> (608 queries)… was **the first** to submit their solution
- Both of these teams only relied on their **domain expertise**

Queries do not tell the whole story!

No gradient was computed here!

UNIVERSITÄT LIECHTENSTEIN

The **human factor** is a significant component in the *cost* and *effectiveness* of an attack.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cybersecurity ⇔ Economics (state-of-research)

- Do research papers on adversarial ML take economics into account?

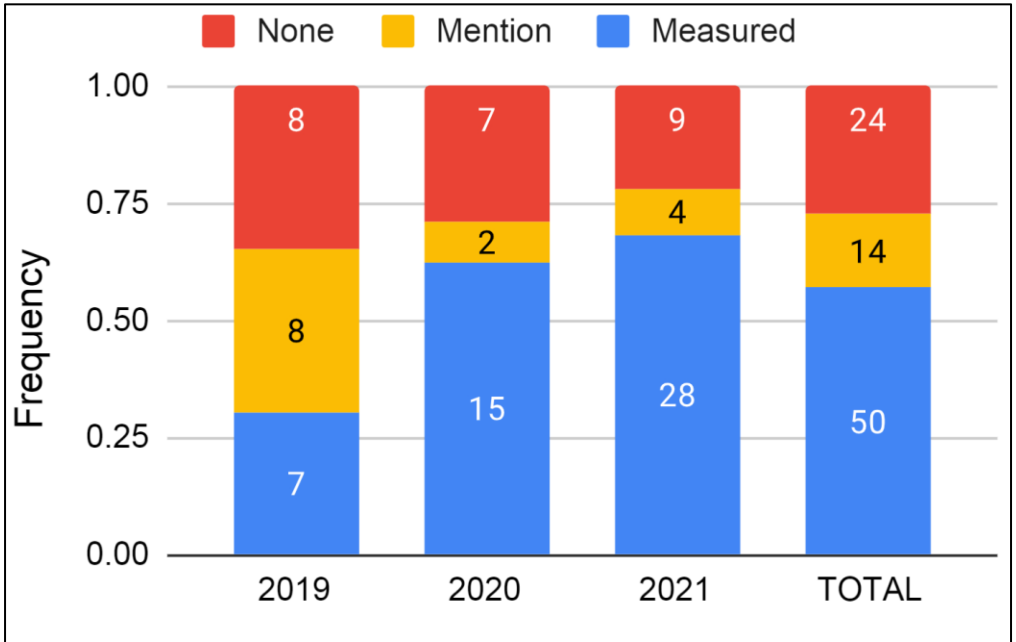


Fig. 9: G4: are the *costs* taken into account (in *any* way)?

Positive trend!

- Only 3 papers provided an *actual cost* in $$ (but only for "expenses")
- The measurements never considered the *human factor*
  - Attack papers measured "queries", defense papers measured "performance degradation"

At least in the adversarial ML domain, economics appears to be overlooked.

Objectively measuring the human factor is hard!

UNIVERSITÄT LIECHTENSTEIN

Disclaimer: the findings of all these papers are still significant!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

A few words on the state-of-research

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Data and Reproducibility (state-of-research)

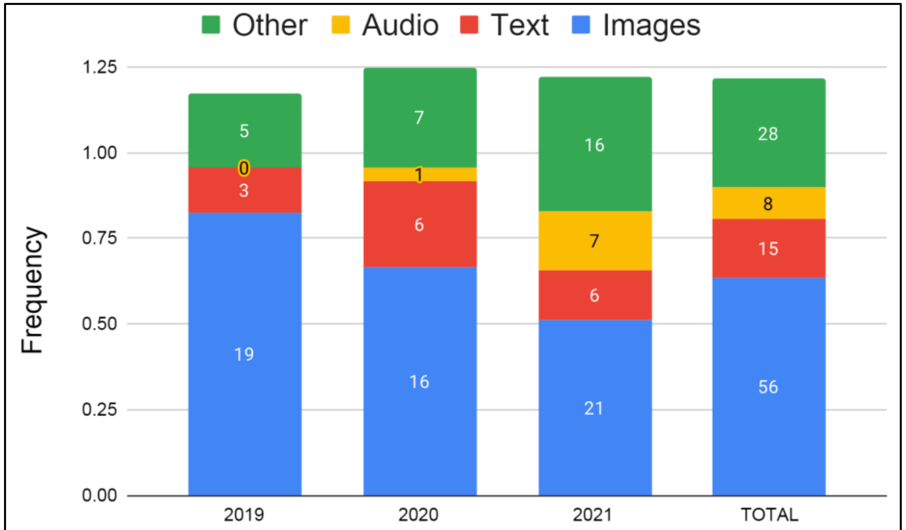


Fig. 10: What are the *data-types* considered in the evaluation?

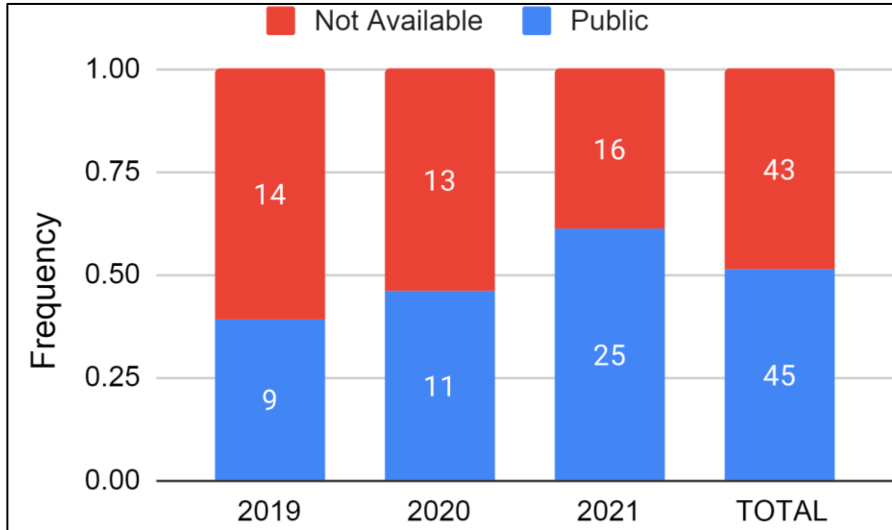


Fig. 11: Has the *source-code* been publicly released?

- Over 50% of the papers focus on image data (decreasing trend)
  - Only 12 papers (out of 88) focus on ML applications for cybersecurity (e.g., phishing, malware)

In cybersecurity conferences!

Some ML application domains (e.g., finance) are rarely discussed in adversarial ML literature.

- Only 50% of the papers release their implementations publicly (increasing trend)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Inconsistent Terminology ("What does the attacker know?")

- The terms "white-box" and "black-box" are widespread, but often denote different degrees of attacker's knowledge. Here are some examples, taken verbatim.

> Co et al. [101]: "In **white-box** settings, the adversary has complete knowledge of the model architecture, parameters, and training data.[...] In a **black-box** setting, the adversary has no knowledge of the target model and no access to surrogate datasets."

Aligns with Srndic and Laskov [43]

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Inconsistent Terminology (“What does the attacker know?”)

- The terms “white-box” and “black-box” are widespread, but often denote different degrees of attacker’s knowledge. Here are some examples, taken verbatim.

Co et al. [101]: “In **white-box** settings, the adversary has complete knowledge of the model architecture, parameters, and training data.[...] In a **black-box** setting, the adversary has no knowledge of the target model and no access to surrogate datasets.”

Aligns with Srndic and Laskov [43]

Shan et al. [102]: “We assume a basic **white box** threat model, where adversaries have direct access to the the ML model, its architecture, and its internal parameter values [...] but *do not have access to the training data*.”

…this is different from [101] (“white-box”)!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Inconsistent Terminology ("What does the attacker know?")

- The terms "white-box" and "black-box" are widespread, but often denote different degrees of attacker's knowledge. Here are some examples, taken verbatim.

Co et al. [101]: "In **white-box** settings, the adversary has complete knowledge of the model architecture, parameters, and training data.[...] In a **black-box** setting, the adversary has no knowledge of the target model and no access to surrogate datasets."

Aligns with Srndic and Laskov [43]

Shan et al. [102]: "We assume a basic **white box** threat model, where adversaries have direct access to the the ML model, its architecture, and its internal parameter values [...] but *do not have access to the training data*."

...this is different from [101] ("white-box")!

Xiao et al. [22]: "In this paper, we focus on the **white-box** adversarial attack, which means we need to access the target model (including its structure and parameters)."

...what about the training data?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Inconsistent Terminology ("What does the attacker know?")

- The terms "white-box" and "black-box" are widespread, but often denote different degrees of attacker's knowledge. Here are some examples, taken verbatim.

Co et al. [101]: "In **white-box** settings, the adversary has complete knowledge of the model architecture, parameters, and training data.[...] In a **black-box** setting, the adversary has no knowledge of the target model and no access to surrogate datasets."

Aligns with Srndic and Laskov [43]

Shan et al. [102]: "We assume a basic **white box** threat model, where adversaries have direct access to the the ML model, its architecture, and its internal parameter values [...] but *do not have access to the training data*."

...this is different from [101] ("white-box")!

Xiao et al. [22]: "In this paper, we focus on the **white-box** adversarial attack, which means we need to access the target model (including its structure and parameters)."

...what about the training data?

Suya et al. [103] assume a "**black-box**" attacker that "does not have direct access to the target model or knowledge of its parameters," but that "has access to pre-trained local models for the same task as the target model" which could be "directly available or produced from access to similar training data."

...this is different from [101] ("black-box")!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Inconsistent Terminology ("What does the attacker know?")

- The terms "white-box" and "black-box" are widespread, but often denote different degrees of attacker's knowledge. Here are some examples, taken verbatim.

Co et al. [101]: "In **white-box** settings, the adversary has complete knowledge of the model architecture, parameters, and training data.[...] In a **black-box** setting, the adversary has no knowledge of the target model and no access to surrogate datasets."

Aligns with Srndic and Laskov [43]

Shan et al. [102]: "We assume a basic **white box** threat model, where adversaries have direct access to the the ML model, its architecture, and its internal parameter values [...] but *do not have access to the training data*."

...this is different from [101] ("white-box")!

Xiao et al. [22]: "In this paper, we focus on the **white-box** adversarial attack, which means we need to access the target model (including its structure and parameters)."

...what about the training data?

Suya et al. [103] assume a "**black-box**" attacker that "does not have direct access to the target model or knowledge of its parameters," but that "has access to pre-trained local models for the same task as the target model" which could be "directly available or produced from access to similar training data."

...this is different from [101] ("black-box")!

Hui et al. [104] envision a "**gray-box**" setting which "gives full knowledge to the adversary in terms of the model details. Specifically, except for the training data, the adversary knows almost everything about the model, such as the architecture and the hyper-parameters used for training."

This is the exact same as [102]... which describes a "white-box" setting!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Inconsistent Terminology ("What does the attacker know?")

- The terms "white-box" and "black-box" are widespread, but often denote different degrees of attacker's knowledge. Here are some examples, taken verbatim.

Co et al. [101]: "In **white-box** settings, the adversary has complete knowledge of the model architecture, parameters, and training data.[...] In a **black-box** setting, the adversary has no knowledge of the target model and no access to surrogate datasets."

Aligns with Srndic and Laskov [43]

Shan et al. [102]: "We assume a basic **white box** threat model, where adversaries have direct access to the the ML model, its architecture, and its internal parameter values [...] but *do not have access to the training data*."

…this is different from [101] ("white-box")!

Xiao et al. [22]: "In this paper, we focus on the **white-box** adversarial attack, which means we need to access the target model (including its structure and parameters)."

…what about the training data?

Suya et al. [103] assume a "**black-box**" attacker that "does not have direct access to the target model or knowledge of its parameters," but that "has access to pre-trained local models for the same task as the target model" which could be "directly available or produced from access to similar training data."

…this is different from [101] ("black-box")!

Hui et al. [104] envision a "**gray-box**" setting which "gives full knowledge to the adversary in terms of the model details. Specifically, except for the training data, the adversary knows almost everything about the model, such as the architecture and the hyper-parameters used for training."

This is the exact same as [102]… which describes a "white-box" setting!

UNIVERSITÄT LIECHTENSTEIN

Taken individually, all past work are correct. The problems arise when analyzing the situation **as a whole**!

Our four Positions

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

P1: Adapt threat models to ML systems

Attacker's **Goal, Knowledge, Capabilities** and **Strategy** should reflect the ML system (and not just the ML model!)

→ Real attackers have **broader objectives** and do not want just to "evade the ML model."

Each of those elements should be **precisely defined**.

→ Existing **terminology** is often used inconsistently.

Problematic Terms:

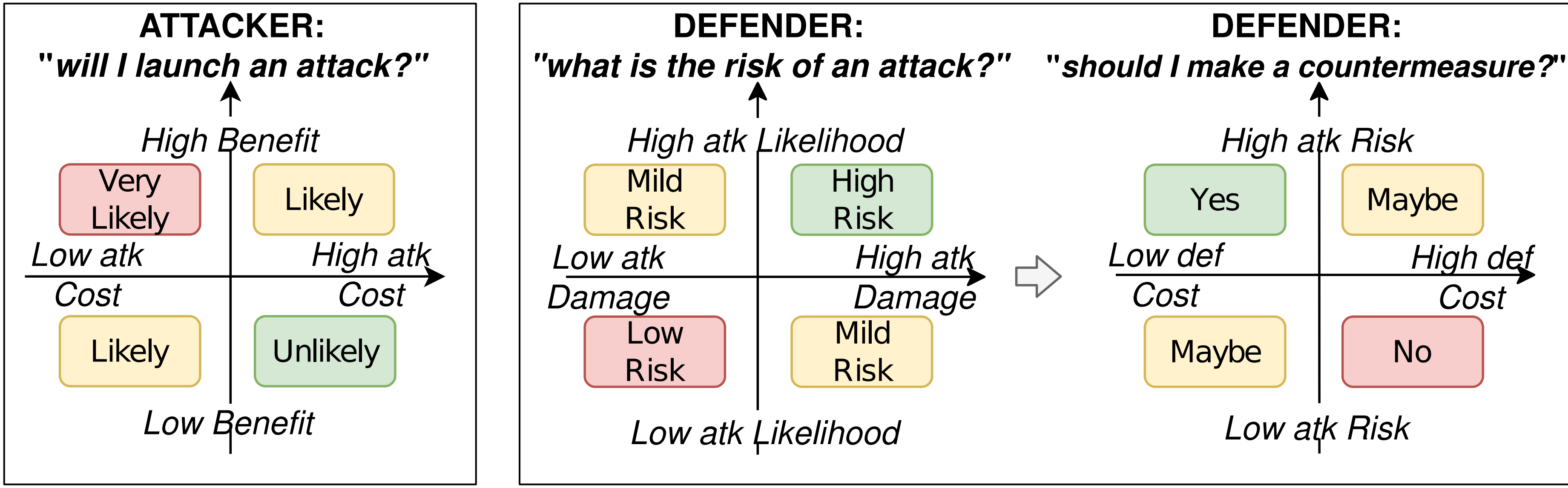
- "Box-based" terminology
- "Access"
- "Adversarial"
- "Evasion"

Solutions and recommendations in the paper!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

P2: *Cost-based* threat modeling



Both attacks and defenses have a **cost**. Real attackers do not launch an attack if it is *too expensive*; and real developers will not develop a countermeasure if the attack is *unlikely to occur in reality*.

→ Cost measurements should account for the **human factor** (queries / computation are not enough)

More on this in the paper!

→ There is value also in defenses that work “only” against attackers with **limited knowledge** (they are more common in reality).

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

P3: Collaborations between *industry* and *academia*

Practitioners should be **more willing** to cooperate with researchers: both have the same goal!

- Streamline research collaboration process
- Bug Bounties
- Releasing Schematics

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

P4: *Source-code* disclosure with "just culture"

Just Culture: assumes that mistakes are bound to occur and derive from organizational issues. Mistakes are avoided by understanding their root causes and using them as constructive learning experiences.

Embracing a just culture naturally promotes the **gradual improvement** at the base of research efforts.

→ The fast pace of research in ML can lead to errors in experiments (not always spotted during the peer-review)

→ By releasing the source code, future works can correct such mistakes, potentially systematizing them, and hence **turning "negative results" into positive outcomes** for our community.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

State-of-research [bonus]

TABLE IV: The 88 papers considered in our analysis. Each column reports the answer to one of the 12 research questions we used during our survey
available, the **G6** column provides the *hyperlink* to the websites hosting the source-code of a given paper. Explanations are in Appendix B-I.

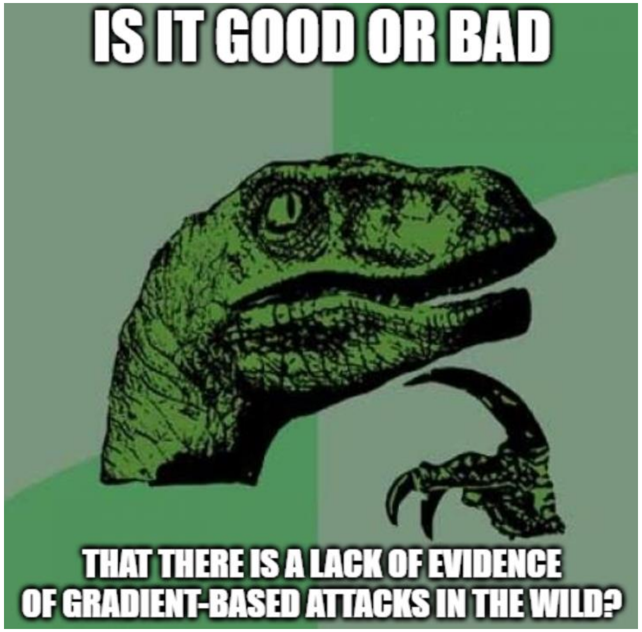
Year (subs)	Venue (subs)	Paper (1st author)	G1	G2	G3	G4	G5 (Evaluation Data)				G6	G7	G8	T1	T2	T3	T4
			Focus	*Attack*	*Paradigm*	*Cost*	*Img*	*Text*	*Audio*	*Other*	*Code*	*Pipeline*	*Type*	*Param.*	*Sem.*	*Output*	*Training*
2019 (23/435)	NDSS (4/89)	Salem [158]	atk	Member.	DL	◐	✓	✓			✓		CLOSED	✗	✓	p	✗
		Li [138]	atk	Evasion	DL	✓	✓				✓			✓	✓	✗	R
		Ma [163]	def	Evasion	DL	◐	✓							✓	✓	p	✗
		Li [161]	atk	Evasion	DL+SL	◐		✓				✓	CLOSED	✗	✓	p	D
	SP (3/84)	Ling [132]	def	Evasion	DL	✓	✓				✓			✓	✓	p	✗
		Wang [164]	def	Poison.	DL	✓	✓				✓			✓	✓	✗	✓
		Nasr [100]	atk	Member.	DL	✗	✓			Finance				✗	✓	p	D
	SEC (6/113)	Tong [114]	def	Evasion	DL+SL	✓				Malware	✓			✗	✓	p	✗
		Demontis [44]	atk	Evasion	DL+SL	✗	✓			Malware				✗	✓	p	S
		Xiao [22]	atk	Evasion	DL	✗	✓					✓	CLOSED	✗	✓	p	✗
		Quiring [165]	atk	Evasion	DL+SL	✗		✓						✗	✓	p	✗
		Hong [110]	atk	Evasion	DL	✗	✓					*		✗	✗	p	✗
		Batina [111]	atk	Stealing	DL	◐	✓					*		✗	✓	p	✗
	CCS (10/149)	Song [166]	atk	Member.	DL	✗	✓				✓			✗	✓	p	✗
		Jia [167]	def	Member.	DL	✗	✓			+				✗	✓	p	✗
		Co [101]	atk	Evasion	DL	✓	✓							✗	✓	p	✗
		Liu [168]	def	Poison.	DL	✓	✓							✓	✓	p	✓
		Baluta [169]	def	Poison.	DL	◐	✓				✓			✓	✓	p	✓
		Zhao [105]	atk	Evasion	DL	✗	✓				✓		OPEN	✗	✓	p	S
		Tramer [12]	atk	Evasion	DL	◐	✓				✓	✓	CLOSED	✗	✓	s	✗
		Wang [170]	atk	Evasion	DL	✓				Graphs				✗	✓	✗	⊂
		Yao [52]	atk	Poison.	DL	◐	✓							◐	✓	p	⊂
		Yang [171]	atk	Stealing	DL	◐	✓						CLOSED	✗	✓	p	D
2020 (24/470)	NDSS (2/88)	Aghakhani [142]	atk	Evasion*	DL+SL	✗				Malware	✓			✗	✓	✗	D
		Yu [34]	atk	Stealing	DL	✓	✓						CLOSED	✗	✓	p	D
	SP (4/104)	Schuster [172]	atk	Poison.	DL	✓		✓						✗	✓	✗	⊂
		Pierazzi [49]	atk	Evasion	SL	✓				Malware	✓	✓		✓	✓	p	R
		Chen [88]	def	Evasion	DL	✓	✓				✓			✗	✓	l	✗
		Jan [147]	atk	Evasion*	DL	✓				Network				✗	✓	p	⊂
	SEC (8/157)	Salem [146]	atk	Member.	DL	✗	✓			Location				◐	✓	p	S
		Chandrasekaran [148]	atk	Stealing	DL+SL	✓	✓	✓		+				◐	✓	p	✗
		Suya [103]	atk	Evasion	DL	✓	✓				✓			✗	✓	p	S
		Jagielski [107]	atk	Stealing	DL	◐	✓							✗	✓	l	D
		Quiring [173]	atk	Evasion	DL	✗	✓				✓	✓		✗	✓	p	✗
		Li [151]	def	Evasion	DL	✓		✓						✗	✓	p	D
		Leino [174]	atk	Member.	DL	✓	✓			+				✓	✓	p	✗
		Zhang [175]	atk	Evasion	DL	✗	✓							✓	✓	p	✗
	CCS (10/121)	Nassi [25]	atk	Evasion	DL	✗	✓				✓	✓	CLOSED	✗	✓	s	✗
		Li [176]	atk	Evasion	DL	✗			✓					✓	✓	p	⊂
		Shan [102]	def	Evasion	DL	◐	✓				✓			✓	✓	p	✗
		Pang [162]	atk	Poison.	DL	✓	✓				✓			✓	✓	p	✓
		Abdelnabi [177]	atk	Evasion*	DL	✗	✓			Phishing	✓	✓		✗	✓	✗	✗
		Li [178]	atk	Evasion	DL	✓	✓							✓	✓	p	✗
		Lin [179]	atk	Poison.	DL	✓	✓	✓			✓			✓	✓	✗	✓
		Chen [180]	atk	Member.	DL	✓	✓							✗	✓	p	D
		Zanella [24]	atk	Member.	DL	✓		✓			✓			✗	✓	p	✗
		Song [23]	atk	Member.	DL	✓		✓						✗	✓	p	S
2021 (41/644)	NDSS (3/87)	Hui [104]	atk	Member.	DL	✗	✓			+	✓			✗	✓	p	✗
		Huang [181]	atk	Poison.	DL	✓				Ratings				✗	✓	✗	⊂
		Barradas [90]	atk	Evasion*	SL	✓				Network	✓			✗	✗	✗	✗
	SP (5/115)	Xu [149]	def	Poison.	DL	✓	✓		✓		✓			✓	✓	p	✓
		Abdelnabi [182]	atk	Evasion	DL	◐		✓			✓	✓		✗	✓	p	✗
		Chen [91]	atk	Evasion	DL	✓			✓		✓	✓	(both)	✗	✓	s	S
		Abdullah [145]	atk	Evasion	DL	✓			✓			✓	CLOSED	✗	✓	✗	✗
		Nasr [35]	def	Evasion	DL	✓	✓			Finance				✗	✓	p	✗
	SEC (24/246)	Sato [28]	atk	Evasion	DL	✓	✓				✓	✓	OPEN	✓	✓	p	✗
		Nasr [89]	atk	Evasion	DL	✓				Network	✓	✓		✗	✓	p	D
		He [183]	atk	Member.	DL	◐				Graph				✗	✓	p	✗
		Severi [184]	atk	Poison.	DL+SL	✓				Malware	✓			✗	✓	p	✓
		Bagdasaryan [95]	atk	Poison.	DL	✓	✓	✓			✓			✗	✗	✗	✗
		Xi [155]	atk	Poison.	DL	✓				Graph	✓			◐	✓	✗	S
		Tang [96]	def	Poison.	DL	✗	✓				✓			◐	✓	p	⊂
		Schuster [185]	atk	Poison.	DL	✓		✓					OPEN	✗	✓	✗	⊂
		Carlini [54]	atk	Poison.	DL	✓	✓							✗	✓	✗	⊂
		Vicarte [186]	atk	Poison.	DL	✓	✓							✗	✓	p	✗
		Lovisotto [150]	atk	Evasion	DL	✓	✓				✓		OPEN	✗	✓	✗	⊂
		Carlini [30]	atk	Member.	DL	✓		✓				✓	OPEN	✗	✓	p	✗
		Han [97]	atk	Evasion*	DL	✗				Graph				✗	✓	p	✗
		Eisenhofer [153]	def	Evasion	DL	✓			✓		✓	✓	OPEN	✓	✓	p	R
		Wu [156]	def	Poison.	DL	✓				Games	✓			✗	✓	s	✗
		He [187]	atk	Stealing	DL	✓	✓							✗	✓	l	S
		Rakin [112]	atk	Evasion	DL	✓	✓				✓	*		✗	✓	p	✗
		Jia [188]	def	Stealing	DL	✓	✓		✓		✓			◐	✓	l	⊂
		Zhu [189]	def	Stealing	DL	✗	✓					*		✗	✓	p	✗
		Xiang [190]	def	Evasion	DL	✓	✓				✓			✓	✓	p	✗
		Lin [191]	atk	Evasion *	DL	✓	✓			Phishing	✓	✓		✗	✓	p	✗
		Azizi [192]	def	Poison.	DL	✗		✓			✓			✗	✓	p	✓
		Hussain [93]	def	Evasion	DL	✗			✓		✓			✓	✓	p	✗
		Song [99]	def	Member.	DL	✓	✓			+	✓			✗	✓	l	✗
	CCS (9/196)	Zheng [92]	atk	Evasion	DL	◐			✓			✓	CLOSED	✗	✓	✗	✗
		Mu [193]	atk	Evasion	DL	✓				Graphs				✗	✓	l	✗
		Bahramali [194]	atk	Evasion	DL	✓				Network				✗	✓	✗	S
		Sheatsley [157]	atk	Evasion	DL	✗				Network				✓	✓	p	R
		Du [94]	def	Evasion	DL	✓	✓	✓						✓	✓	p	R
		Li [195]	def	Evasion	DL	✓	✓				✓			✗	✓	p	✗
		He [196]	def	Member.	DL	✗	✓				✓			◐	✓	l	S
		Li [160]	atk	Member.	DL	✗	✓				✓			✗	✓	l	✗
		Chen [197]	def	Member.	DL+SL	✗	✓			+	✓			✗	✓	p	✗
Year (subs)	Venue (subs)	Paper (1st author)	*Focus*	*Attack*	*Paradigm*	*Cost*	*Img*	*Text*	*Audio*	*Other*	*Code*	*Pipeline*	*Type*	*Param.*	*Sem.*	*Output*	*Training*
			G1	G2	G3	G4	G5 (Evaluation Data)				G6	G7	G8	T1	T2	T3	T4

Last page of our paper

This leads to the code repository!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li



Do real attackers compute gradients?

→We cannot prove it ☹ (yet).

Maybe they do!

UNIVERSITÄT
LIECHTENSTEIN

"Real Attackers Don't Compute Gradients": Bridging the Gap between Adversarial ML Research and Practice

Please get his name right! "Savino Dambra"

Meet the team

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Attacking Machine Learning-based Phishing Website Detectors

Based on a joint work with: Jehyun Lee, Zhe Xin, Melanie Ng Pei See, Kanav Sabharwal, Dinil Mon Divakaran:
"Attacking Logo-based Phishing Website Detectors with Adversarial Perturbations".
European Symposium On Research In Computer Security (2023).

UNIVERSITÄT LIECHTENSTEIN
Trustwave®
Acronis
NUS National University of Singapore

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

3. …against both state-of-the-art **systems** *and* **humans**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**
2. We show that **it works**
3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** <u>websites</u> are everywhere

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**
2. We show that **it works**
3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere
- **Countermeasure**: visual similarity techniques reliant on deep learning
  - Trendy in research [7] but also deployed in practice [50]

UNIVERSITÄT LIECHTENSTEIN

[7] Abdelnabi, S., Krombholz, K., Fritz, M.: *Visualphishnet: Zero-day phishing website detection by visual similarity*. ACM CCS (2020)
[50] Apruzzese, G., et al.: *"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice*. IEEE SaTML (2023)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**

2. We show that **it works**

3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere
- **Countermeasure**: visual similarity techniques reliant on deep learning
    - Trendy in research [7] but also deployed in practice [50]
- **Problem**: the security of these defenses has not been scrutinized yet
    - Especially from a "human" perspective!

UNIVERSITÄT LIECHTENSTEIN

[7] Abdelnabi, S., Krombholz, K., Fritz, M.: *Visualphishnet: Zero-day phishing website detection by visual similarity*. ACM CCS (2020)
[50] Apruzzese, G., et al.: *"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice*. IEEE SaTML (2023)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

WHAT?

1. We propose a **novel attack**
2. We show that **it works**
3. …against both state-of-the-art **systems** *and* **humans**

WHY?

- **Phishing** websites are everywhere
- **Countermeasure**: visual similarity techniques reliant on deep learning
  - Trendy in research [7] but also deployed in practice [50]
- **Problem**: the security of these defenses has not been scrutinized yet
  - Especially from a "human" perspective!

Disclaimer:
non-technical talk!

UNIVERSITÄT
LIECHTENSTEIN

[7] Abdelnabi, S., Krombholz, K., Fritz, M.: *Visualphishnet: Zero-day phishing website detection by visual similarity*. ACM CCS (2020)
[50] Apruzzese, G., et al.: *"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice*. IEEE SaTML (2023)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Why Phishing?

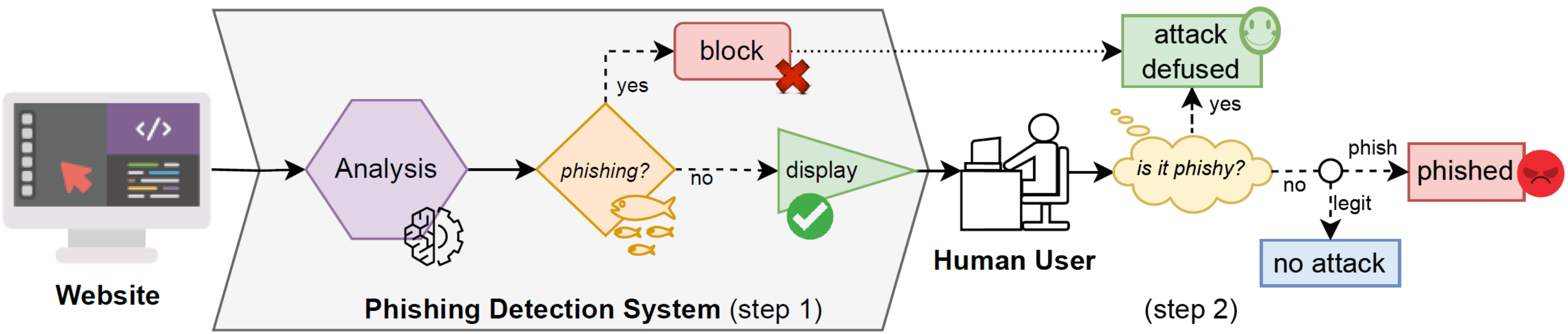


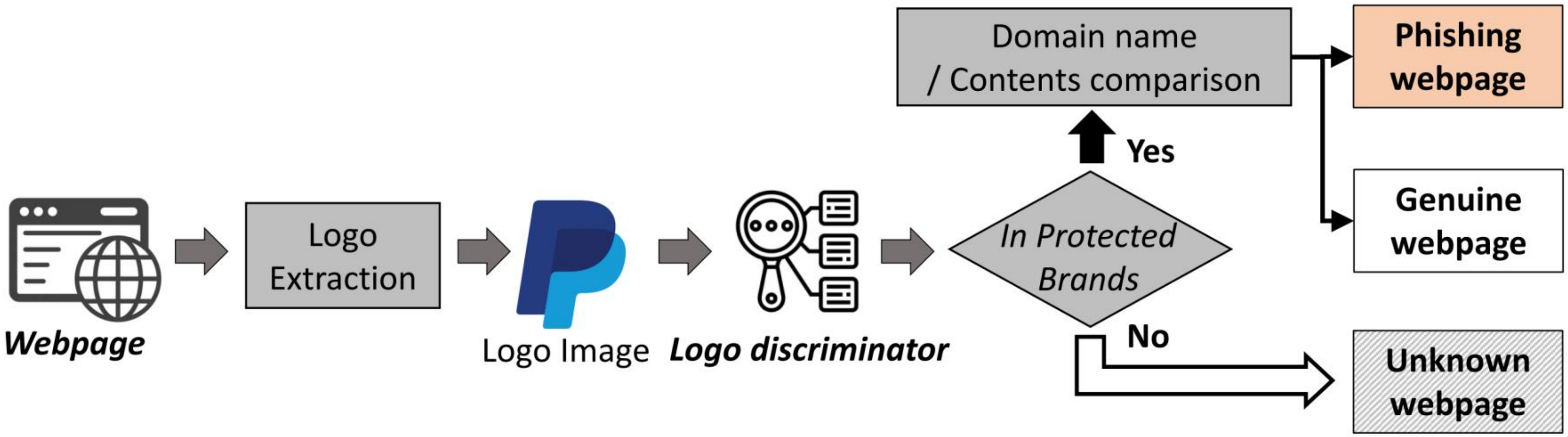
Fig. 1: Scenario: phishing detection is a two-step decision process.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

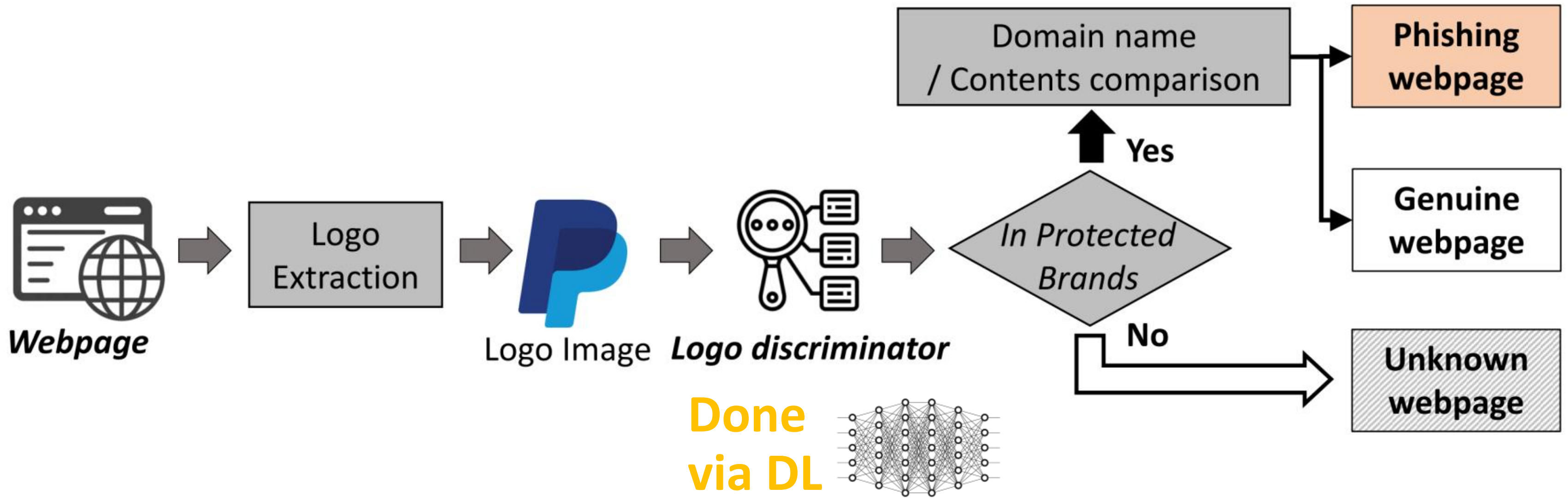
in a nutshell



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

in a nutshell

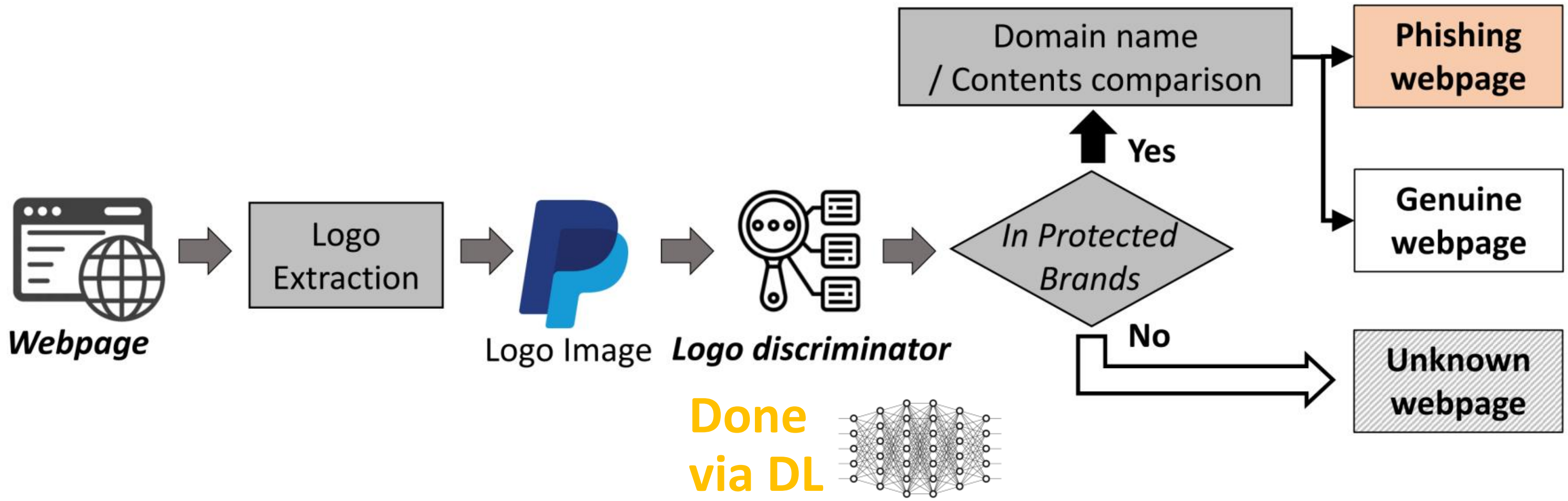


UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

in a nutshell



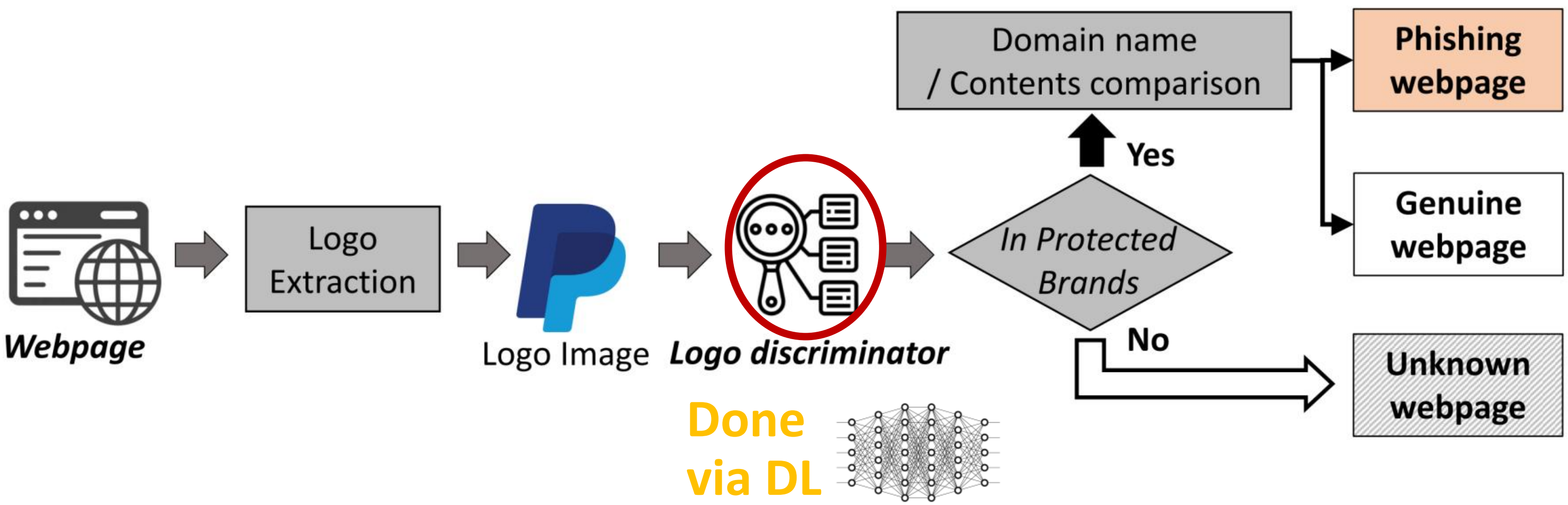
Problem: these systems are tweaked to minimize false positives.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Logo-based Phishing Website Detection

in a nutshell



Problem: these systems are tweaked to minimize false positives.

We focus on the Logo-discriminator.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

1. **Knowledge:**

2. **Capabilities:**

3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)

No knowledge of the DL model is required!

2. **Capabilities:**
   - the attacker can observe the decision of the detector
   - the attacker can manipulate their phishing webpages

The attacker can do nothing to the training data.

3. **Strategy:**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Our attack: adversarial logos

Intuition: create an adversarial logo that is (i) minimally altered w.r.t. its original variant; and that (ii) misleads the logo discriminator.

No knowledge of the DL model is required!

1. **Knowledge:**
   - the attacker expects the detector to have the "phished" brand(s) in the protected set (and that its logos are inspected)
2. **Capabilities:**
   - the attacker can observe the decision of the detector
   - the attacker can manipulate their phishing webpages

The attacker can do nothing to the training data.

3. **Strategy:** Manipulate the logo so that the discriminator has a lower confidence → the detector will default to a "unknown webpage"

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Discriminators

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (++book)

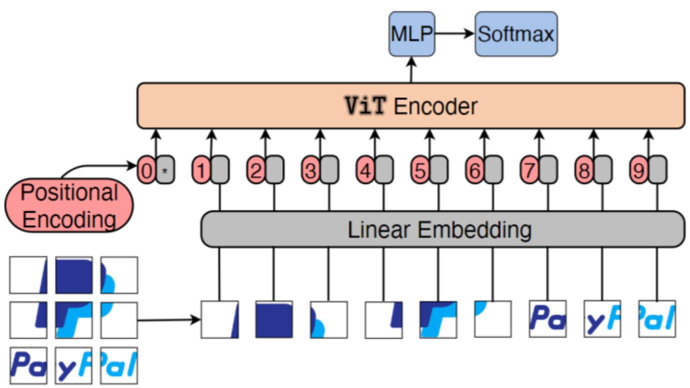


Fig. 2: ViT-based Model Architecture

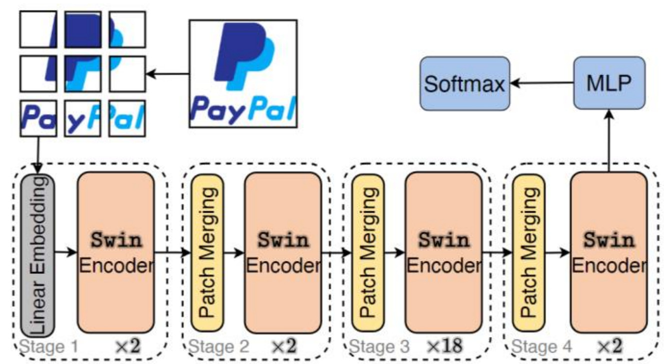


Fig. 3: Swin-based Model Architecture

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Discriminators

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

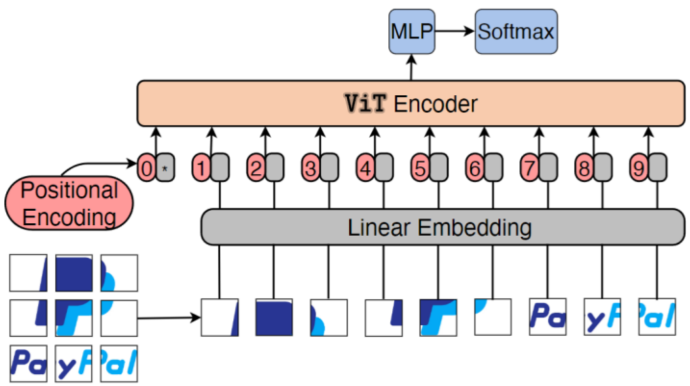


Fig. 2: ViT-based Model Architecture

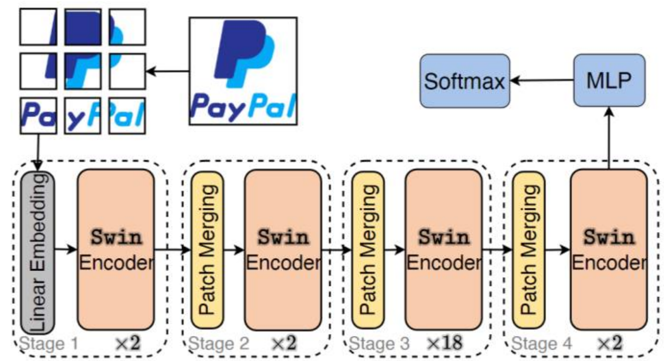


Fig. 3: Swin-based Model Architecture

- We will show that these methods reach state-of-the-art performance (currently obtained by Siamese networks [34])

UNIVERSITÄT LIECHTENSTEIN

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)
[34]: Lin, Y., et al.: *Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.* USENIX Security (2021)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Discriminators

- We propose two novel methods for logo-identification: ViT and Swin
  - Both ViT and Swin leverage transformers [23, 36].

We are the first to use transformers for logo-identification (ttbook)

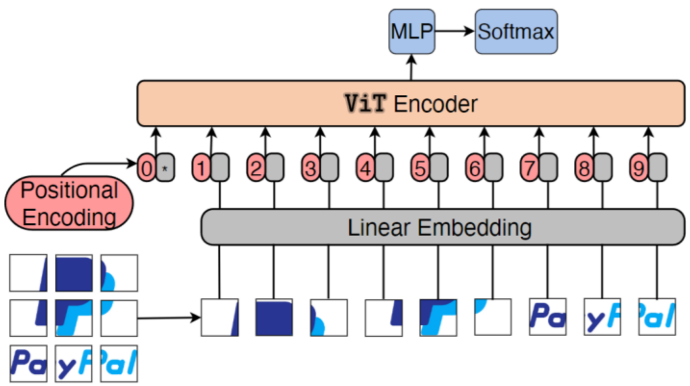


Fig. 2: ViT-based Model Architecture

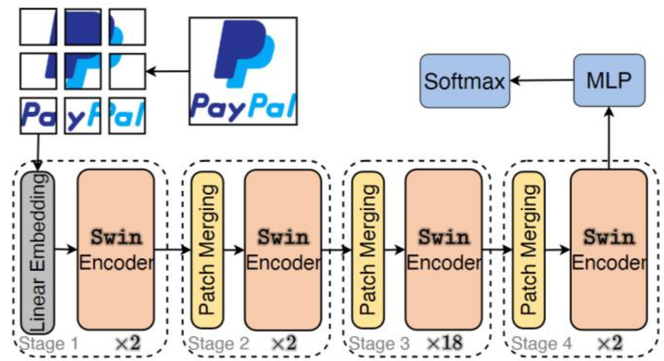


Fig. 3: Swin-based Model Architecture

- We will show that these methods reach state-of-the-art performance (currently obtained by Siamese networks [34])
  - Siamese networks have been assessed in white-box settings

...but our attacker is not a white-box!

UNIVERSITÄT LIECHTENSTEIN

[23] Dosovitskiy, A., et al.: *An image isworth 16x16 words: Transformers for image recognition at scale.* arXiv:2010.11929 (2020)
[36] Liu, Z., et al. : Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF ICCV (2021)
[34]: Lin, Y., et al.: *Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.* USENIX Security (2021)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Attack

We are inspired by "GAN"

- Our attack applies a "Generative Adversarial Perturbations" (GAP)

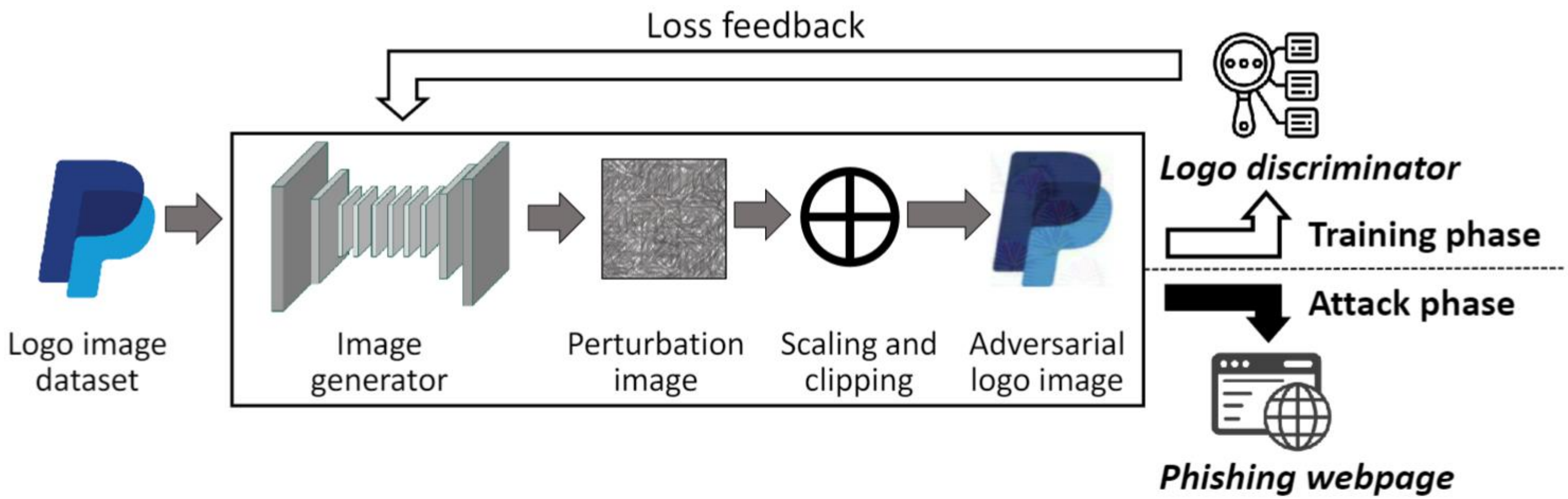


Fig. 4: Generative adversarial perturbation workflow

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Evaluation: Attack

We are inspired by "GAN"

- Our attack applies a "Generative Adversarial Perturbations" (GAP)

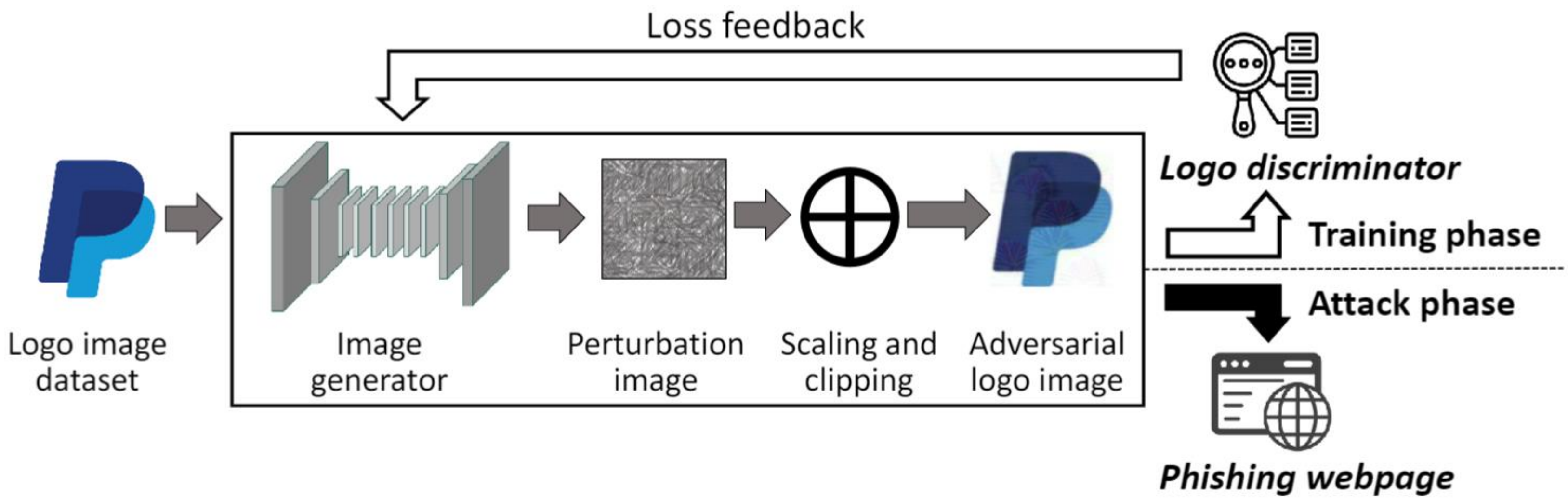


Fig. 4: Generative adversarial perturbation workflow

- The GAP automatically "learns" to craft adversarial logos that mislead the logo discriminator – while being minimally altered.

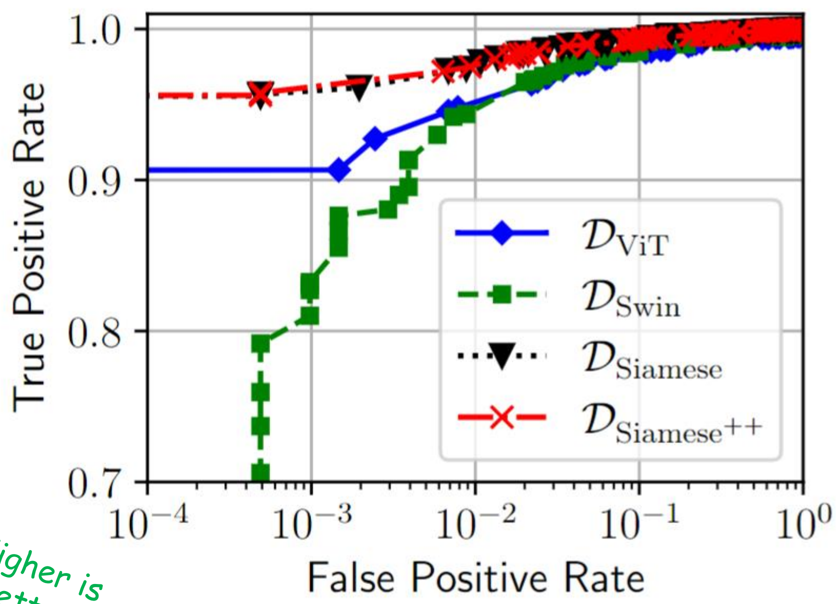
We will assess the cross-model transferability of our adversarial logos!

UNIVERSITÄT LIECHTENSTEIN

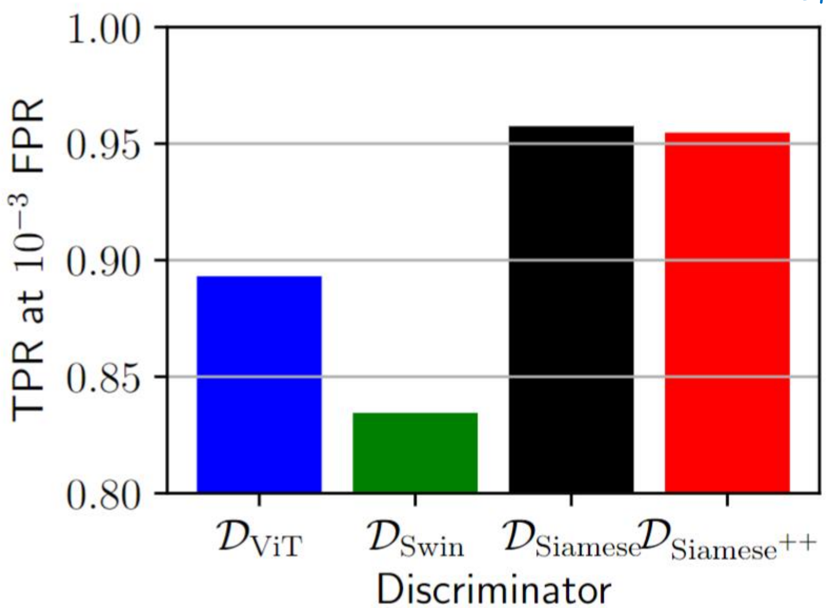
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Baseline

$\mathcal{D}_{Siamese++}$ is a "robust" version of Siamese networks



(a) ROC curves



(b) TPR at 10^{-3} FPR

Higher is better

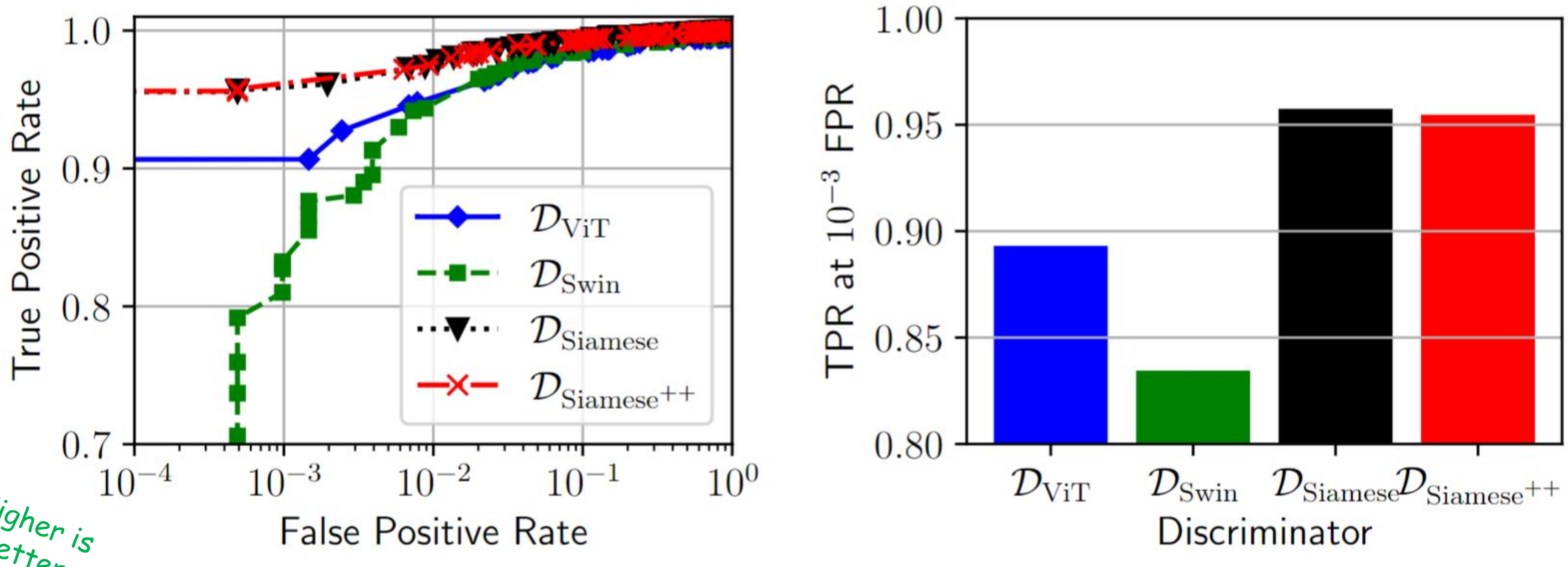
Our baselines are trained to identify 181 brands (~28k logos)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Baseline

$\mathcal{D}_{Siamese++}$ is a "robust" version of Siamese networks



Higher is better

(a) ROC curves

(b) TPR at 10^{-3} FPR

Takeaways:

1. Our baselines "work well" (in the absence of attacks!)
2. ViT and Swin are slightly worse than Siamese…

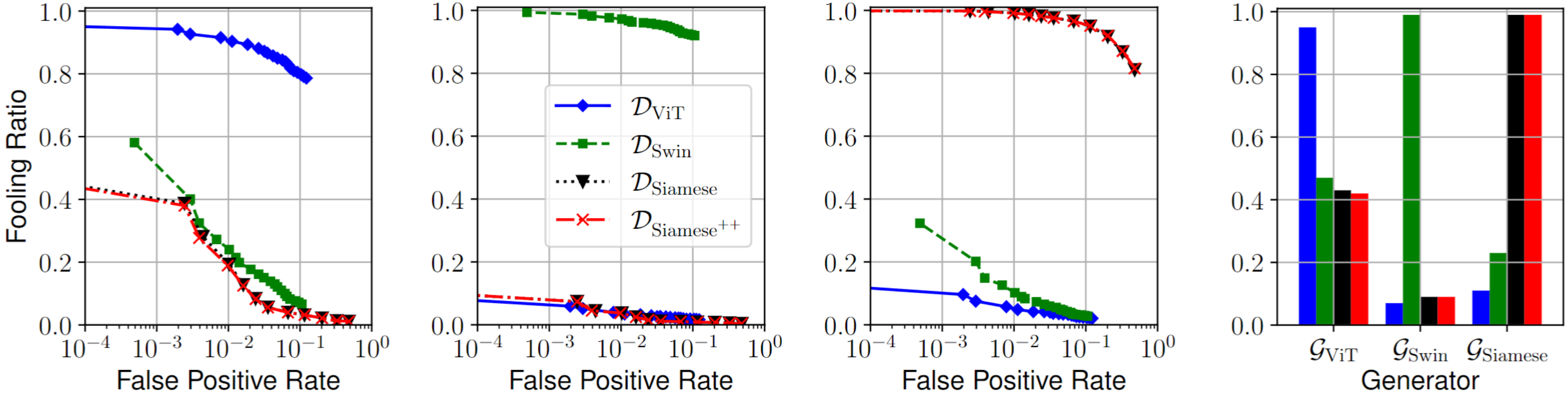
Our baselines are trained to identify 181 brands (~28k logos)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\text{ViT}}$ **(b)** $\mathcal{G}_{\text{Swin}}$ **(c)** $\mathcal{G}_{\text{Siamese}}$ **(d)** at 10^{-3} FPR

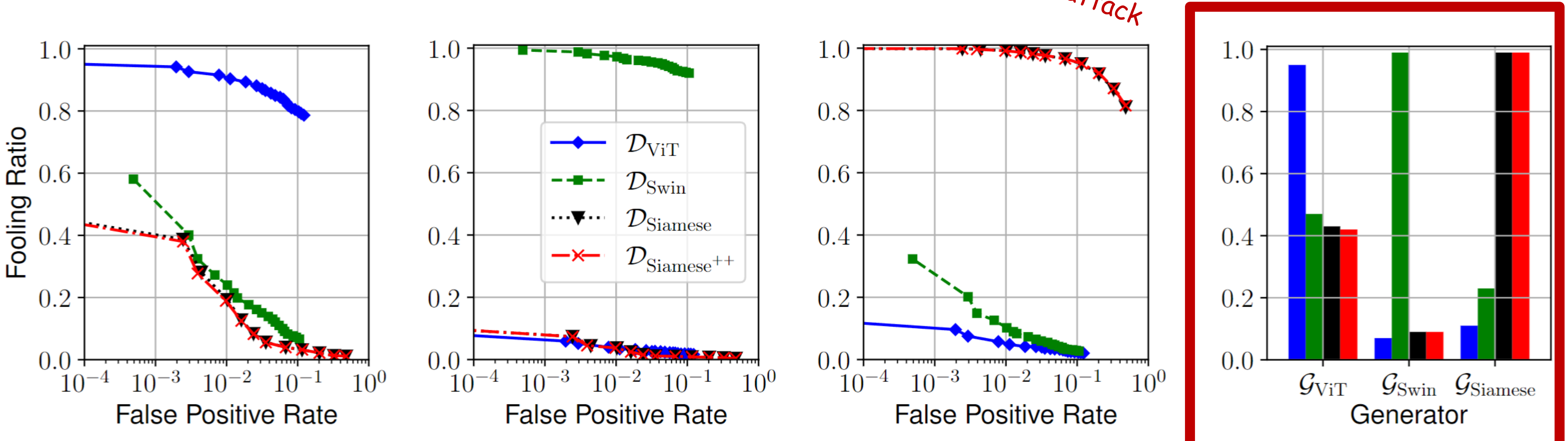
E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\text{ViT}}$ **(b)** $\mathcal{G}_{\text{Swin}}$ **(c)** $\mathcal{G}_{\text{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

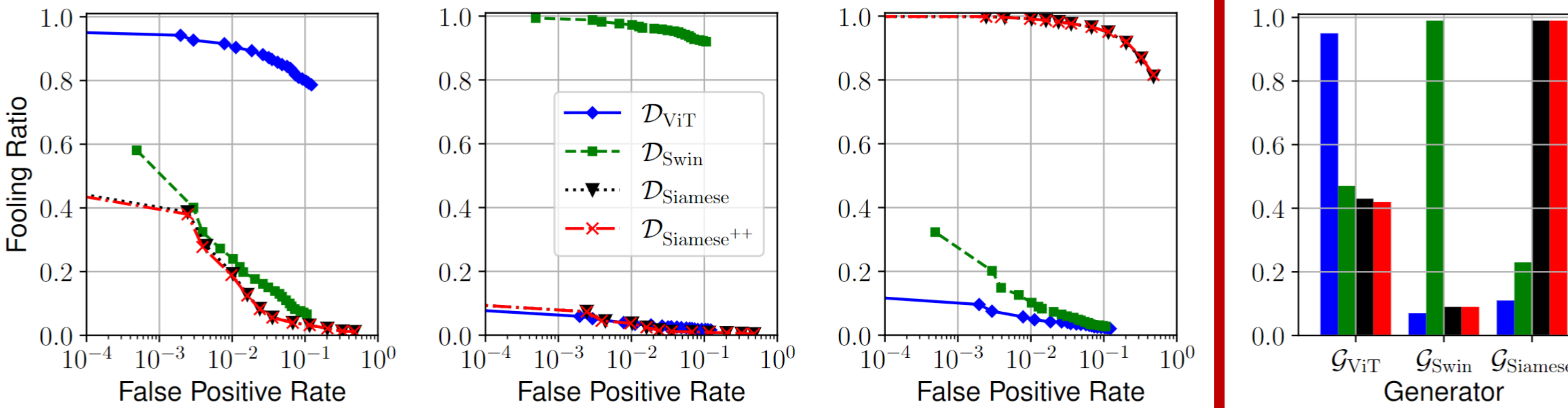
1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the “more robust” detector! (if the attacker is blind)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Attack

Higher = stronger attack



(a) $\mathcal{G}_{\text{ViT}}$ **(b)** $\mathcal{G}_{\text{Swin}}$ **(c)** $\mathcal{G}_{\text{Siamese}}$ **(d)** at 10^{-3} FPR

E.g.: G_{ViT} denotes the GAN trained to evade D_{ViT}

Takeaways:

1. When the attacker and defender use the same model, the attack is ~100% effective
2. ViT is the "more robust" detector! (if the attacker is blind)

Table 1: Training time for the perturbation generators

	$\mathcal{G}_{\text{ViT}}$	$\mathcal{G}_{\text{Swin}}$	$\mathcal{G}_{\text{Siamese}}$
Avg. training time per epoch (min.)	12	23	8
No. of epochs for 0.9 fooling ratio	62	12	1
Training time for 0.9 fooling ratio (min.)	744	277	8

UNIVERSITÄT LIECHTENSTEIN

Training G_{ViT} is very expensive!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans?

- We ask ourselves the following research question (RQ):

> Given a pair of logos (i.e., an ‘original’ one, and an ‘adversarial’ one), can the human spot any difference?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans?

- We ask ourselves the following research question (RQ):

> Given a pair of logos (i.e., an 'original' one, and an 'adversarial' one), can the human spot any difference?

- We carry out two user-studies to answer our RQ:
  - **Vertical Study**: small population (N=30) of similar users; 10 questions, but different for every participant.
  - **Horizontal Study**: large population (N=287) of heterogeneous users; 21 fixed questions for all participants.

Yes, we added control questions and attention checks!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans?

Look at these two images for no more than 5 seconds, and then answer the similarity question.

Logo A

Logo B

On a scale from 1 to 5, how similar do you think these two logos are? *

	1	2	3	4	5	
Very Different	○	○	○	○	○	Very Similar

Look at these two images for no more than 5 seconds, and then answer the similarity question.

Logo A

Logo B

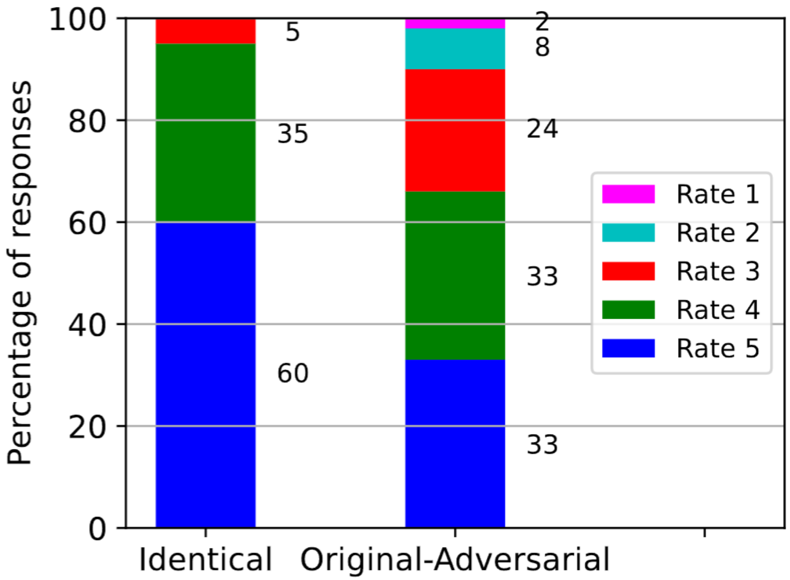
On a scale from 1 to 5, how similar do you think these two logos are? *

	1	2	3	4	5	
Very Different	○	○	○	○	○	Very Similar

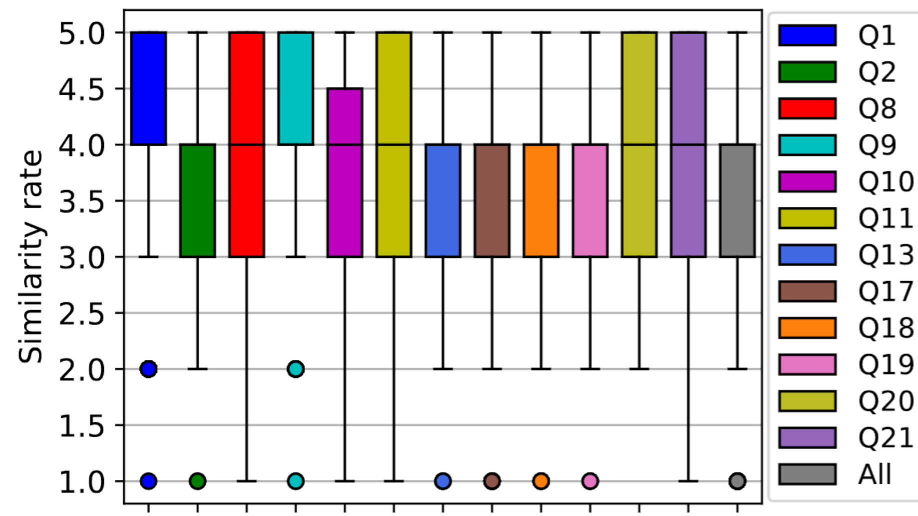
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans? Deceived!

- For every question, users had to say how “similar” the two logos were (5= very similar, 1= not similar at all)



(a) Vertical Study

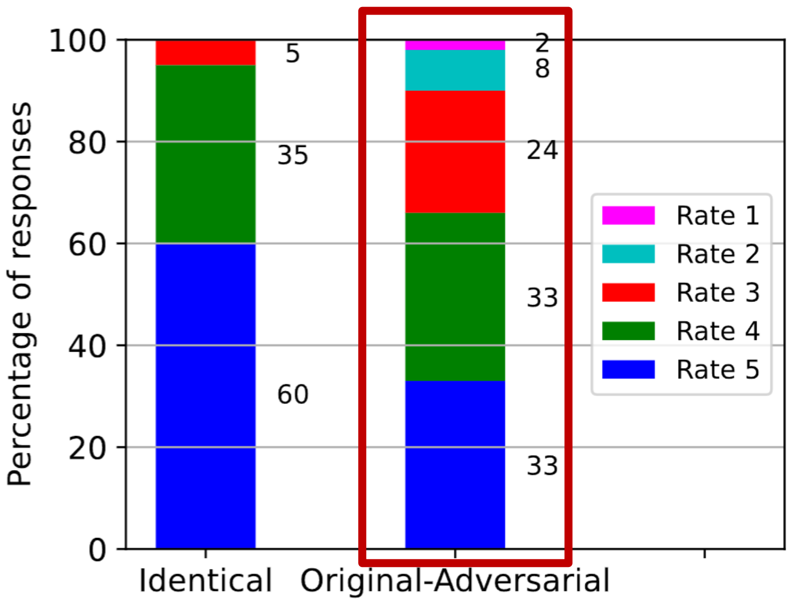


(b) Horizontal Study

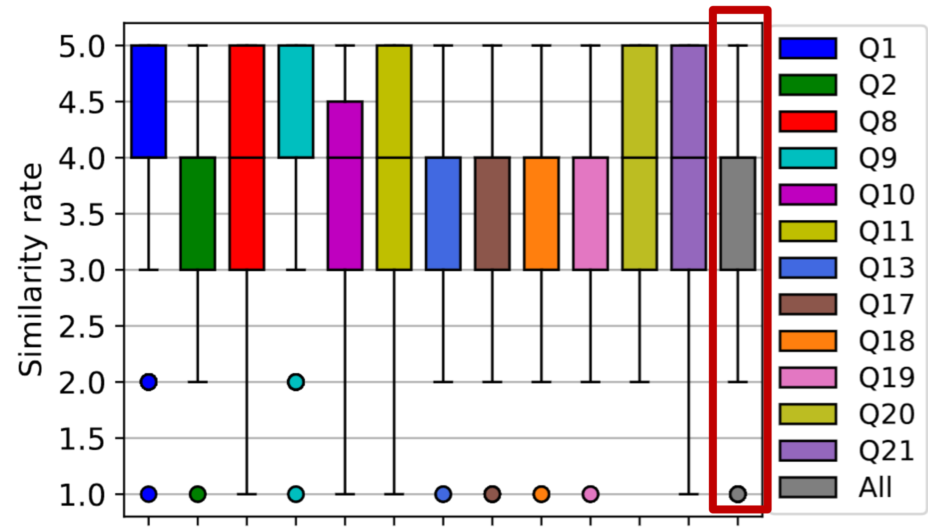
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Results: Humans? Deceived!

- For every question, users had to say how "similar" the two logos were (5= very similar, 1= not similar at all)



(a) Vertical Study



(b) Horizontal Study

Takeaways:

1. Vertical Study: over 85% of participants rated >=3 similarity
2. Horizontal Study: the average similarity per question was >=3

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Countermeasures?

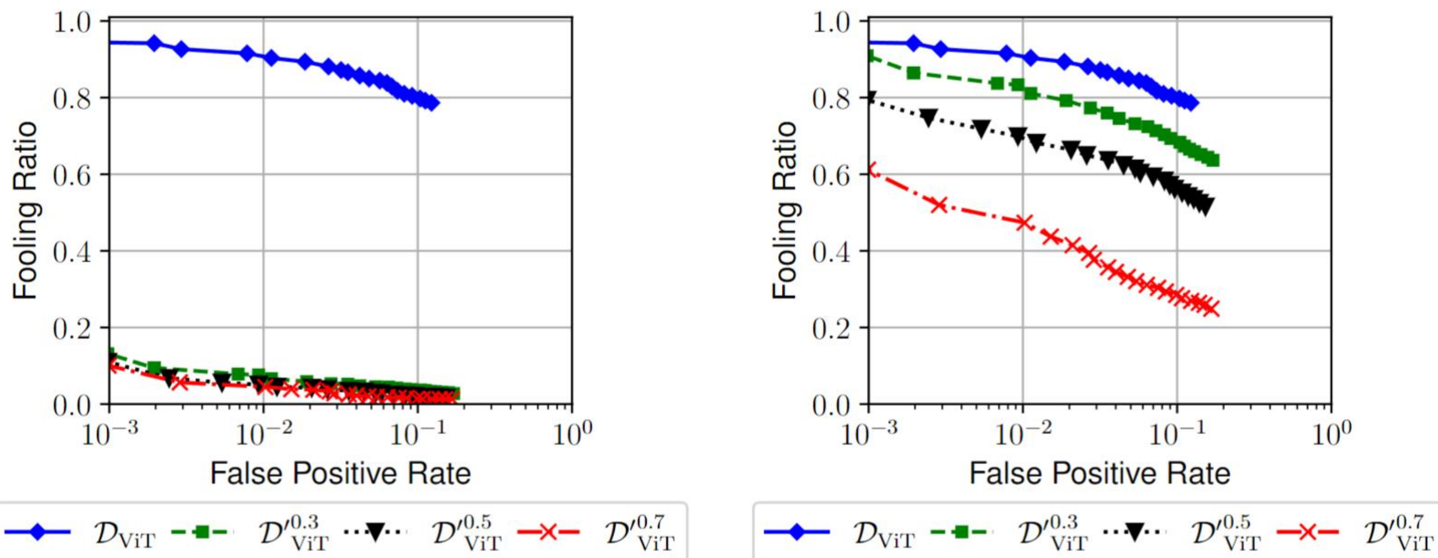
- Can adversarial logos be countered?
    - If so, can an adversary launch a counterattack?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Countermeasures?

- Can adversarial logos be countered?
  - If so, can an adversary launch a counterattack?



(a) Against original generator $\mathcal{G}_{\mathrm{ViT}}$ (b) Against adaptive generators

Fig. 8: Performance of discriminator and generator due to adversarial training

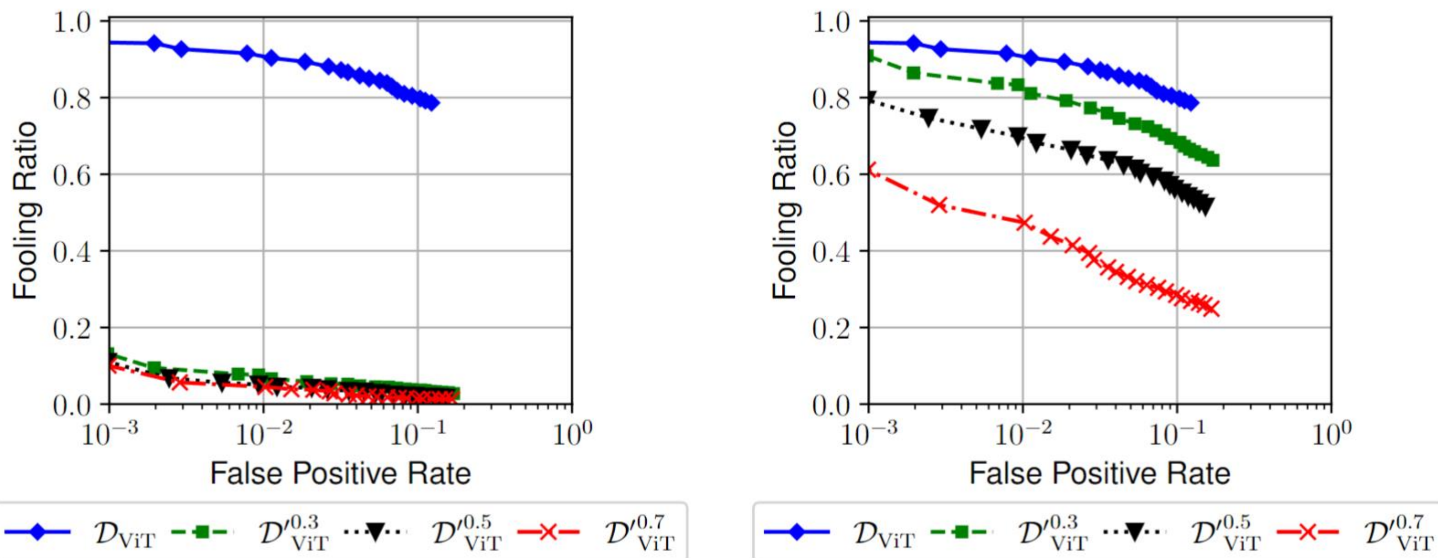
We use the logos generated by G_{ViT} for retraining

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Countermeasures?

- Can adversarial logos be countered? → **Yes** ☺
  - If so, can an adversary launch a counterattack? → **Yes** ☹



(a) Against original generator $\mathcal{G}_{\mathrm{ViT}}$ (b) Against adaptive generators

Fig. 8: Performance of discriminator and generator due to adversarial training

We use the logos generated by G_{ViT} for retraining

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

1. We proposed a **novel attack…**

2. We showed that **it works**

3. …against both state-of-the-art **systems** *and* **humans.**

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

1. We proposed a **novel attack…**

   …as well as two transformer-based methods for logo-identification.

2. We showed that **it works**

   …and that is realistically feasible…

3. …against both state-of-the-art **systems** *and* **humans.**

   …and that countermeasures exist, but can be countered;

   …and also that our proposed transformer methods are more robust and more expensive to evade.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

1. We proposed a **novel attack…**
   …as well as two transformer-based methods for logo-identification.
2. We showed that **it works**
   …and that is realistically feasible…
3. …against both state-of-the-art **systems** *and* **humans.**
   …and that countermeasures exist, but can be countered;
   …and also that our proposed transformer methods are more robust and more expensive to evade.

We focus on the Logo-discriminator.

Future research: consider other elements of a phishing detector, and assess the response of humans to the evasive samples!

UNIVERSITÄT LIECHTENSTEIN

All of our resources are publicly available [1]

[1]: https://github.com/JehLeeKR/Adversarial-phishing-logos

Machine Learning, Security, and Practice: a Reflection

Giovanni Apruzzese

University of Genova – November 13th, 2023

UNIVERSITÄT LIECHTENSTEIN