Lukas Struppek

A Brief History of Security and Privacy in Deep Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

About Me

2015 – 2020

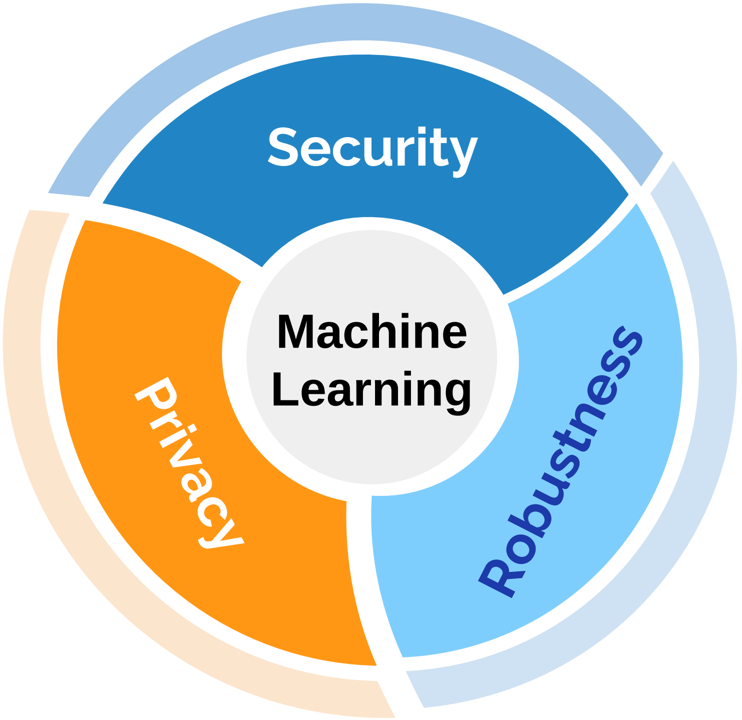
- Bachelor + Master Degree in Industrial Engineering @ KIT

2017 – 2020

- Research Assistent at Applied Technical-Cognitive Systems, AIFB @ KIT

2021 – Today

- PhD Student at Artificial Intelligence and Machine Learning Lab @ TU Darmstadt



Machine Learning Turns the World Upside Down

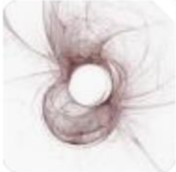
Science

Improving breast cancer diagnostics with deep learning for MRI

Early detection is key to improving breast cancer outcomes. Witowski et al. developed a deep learning pipeline that improves the specificity...

Phys.org

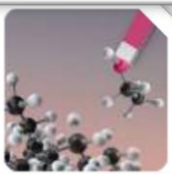
Machine learning takes hold in nuclear physics



Scientists have begun turning to new tools offered by machine learning to help save time and money. In the past several years,...

MIT Technology Review

Machine learning could vastly speed up the search for new metals



Machine learning could help develop new types of metals with useful properties, such as resistance to extreme temperatures and rust,...

Medical Xpress

Machine learning enables an 'almost perfect' diagnosis of an elusive global killer

Sepsis, the overreaction of the immune system in response to an infection, causes an estimated 20% of deaths globally and as many as 20 to...

Machine Learning Turns the World Upside Down

But no one talks about the <u>Security and Privacy</u> of machine learning models!

Agenda

01 **Deep Learning in a Nutshell**
Introducing Neural Networks

02 **Model Inversion Attacks**
Extracting Class Characteristics

03 **Adversarial Examples**
Confusing Neural Networks

04 **Backdoor Attacks**
Injecting Hidden Model Behavior

05 **Conclusion**

Agenda

01 **Deep Learning in a Nutshell**
Introducing Neural Networks

02 Model Inversion Attacks
Extracting Class Characteristics

03 Adversarial Examples
Confusing Neural Networks

04 Backdoor Attacks
Injecting Hidden Model Behavior

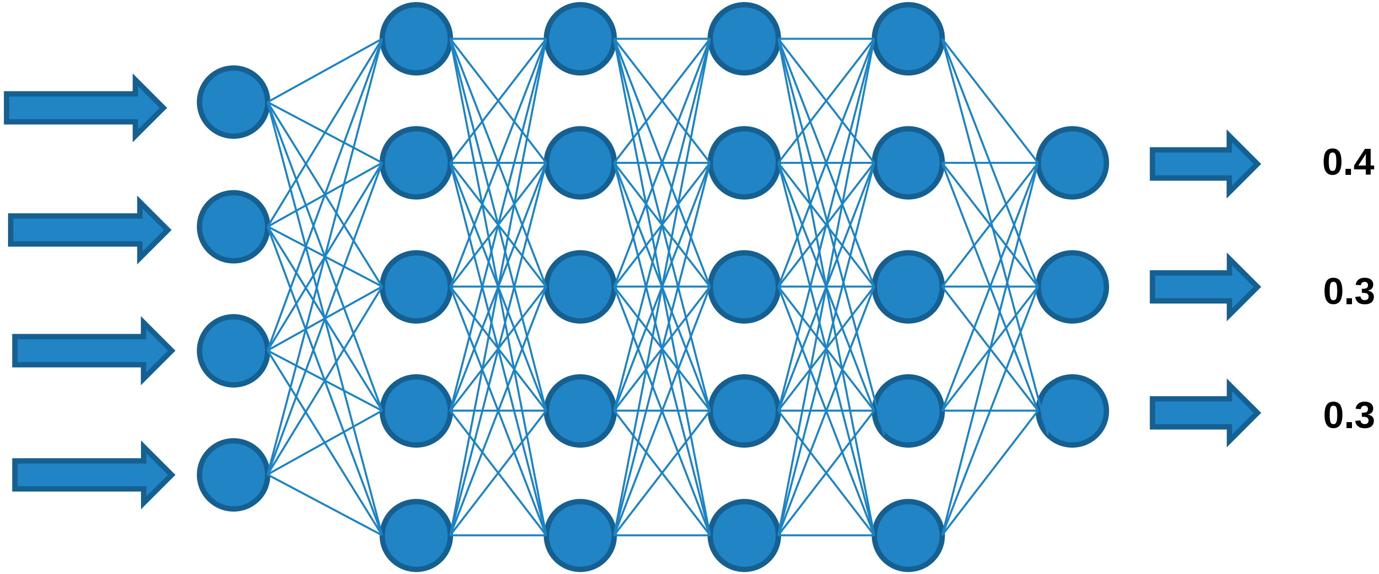
05 Conclusion

Neural Networks Are Universal Function Approximators

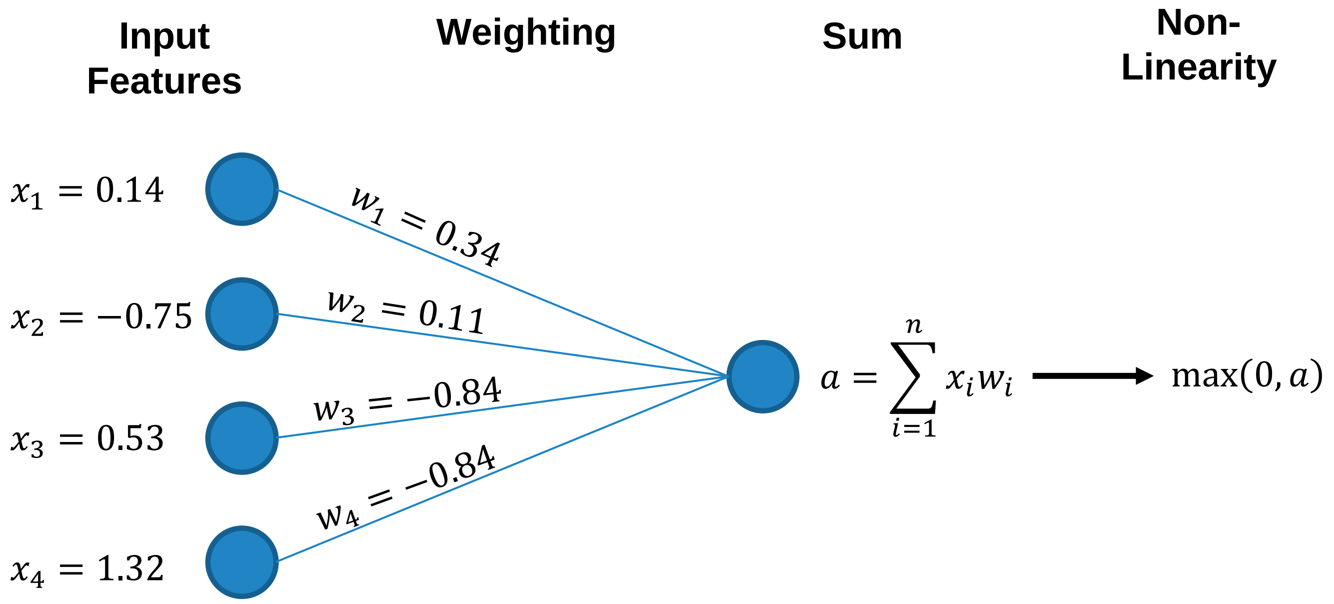
Inputs **Preprocessing** **Neural Network Computation** **Outputs**


Paul Krugman
@paulkrugman

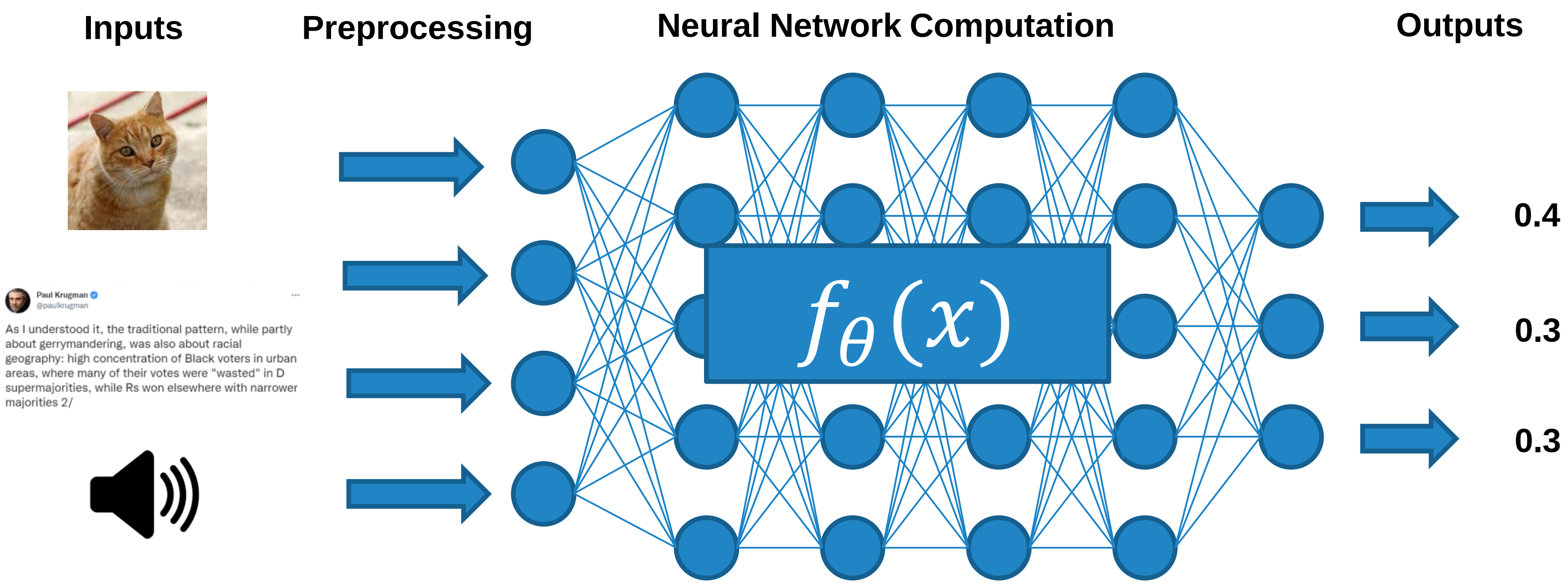
As I understood it, the traditional pattern, while partly about gerrymandering, was also about racial geography: high concentration of Black voters in urban areas, where many of their votes were "wasted" in D supermajorities, while Rs won elsewhere with narrower majorities 2/



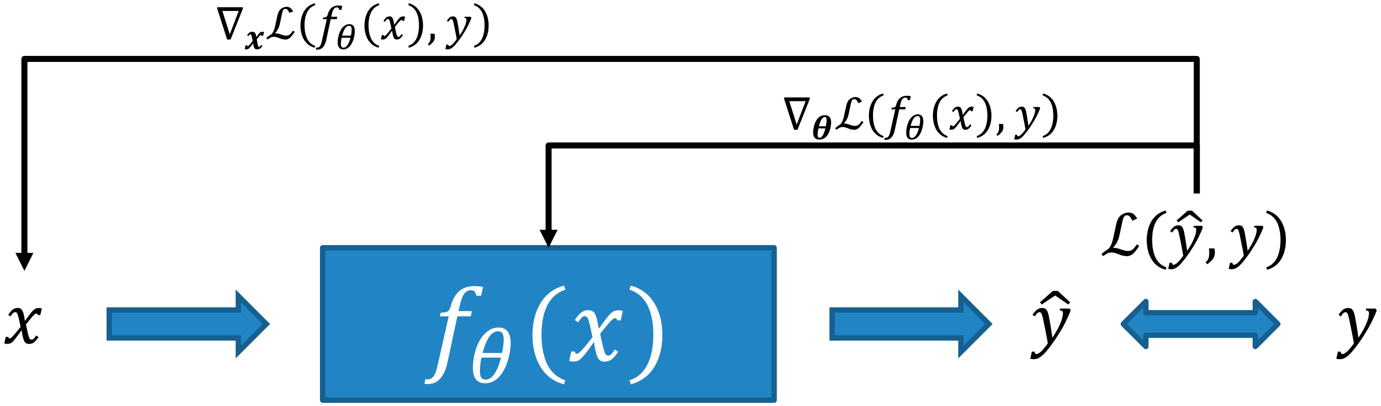
Neural Networks Are Universal Function Approximators



Neural Networks Are Universal Function Approximators

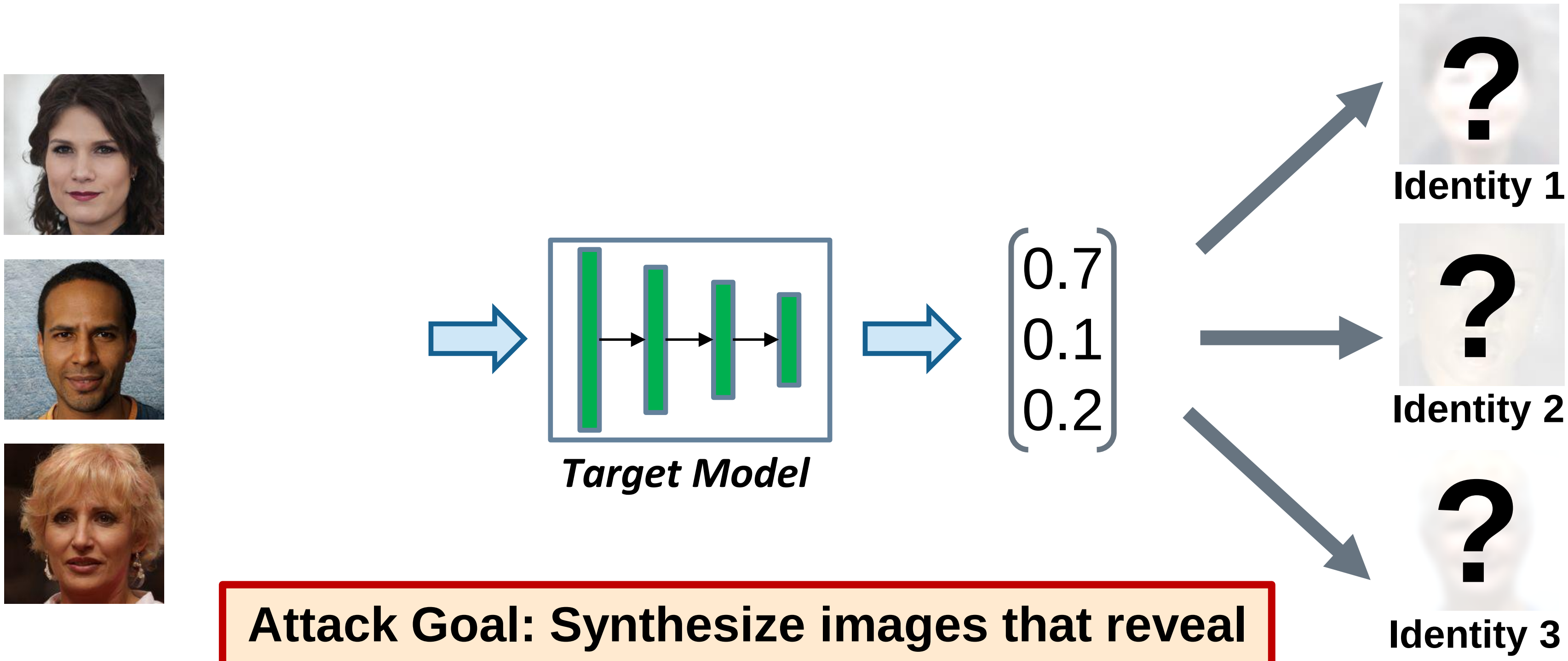


Neural Networks Are Differentiable Functions



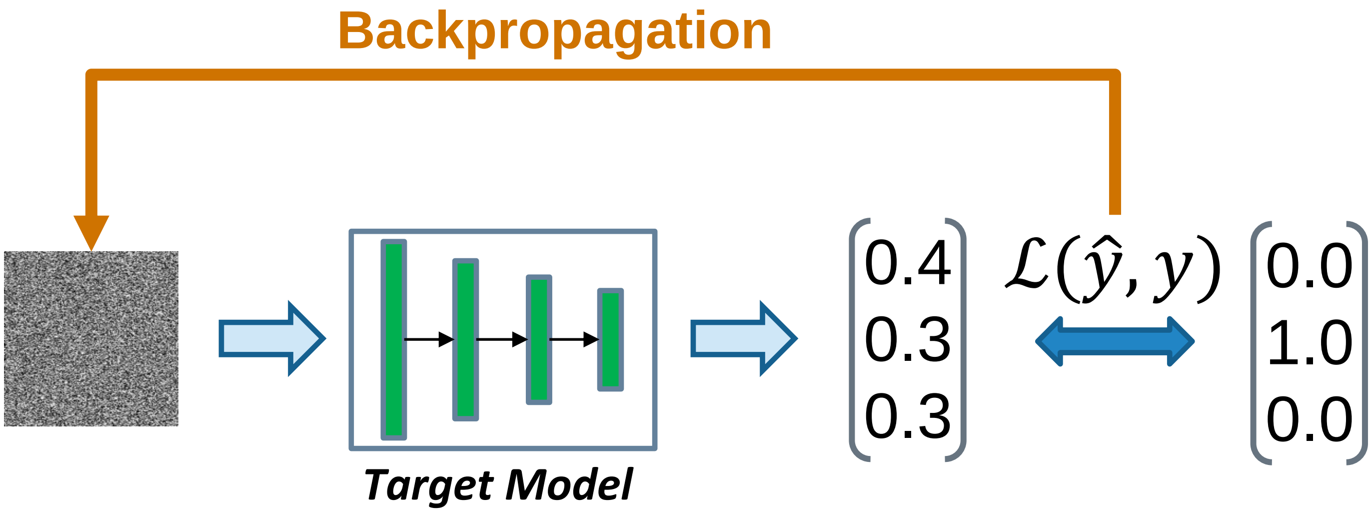
Agenda

Model Inversion Attacks



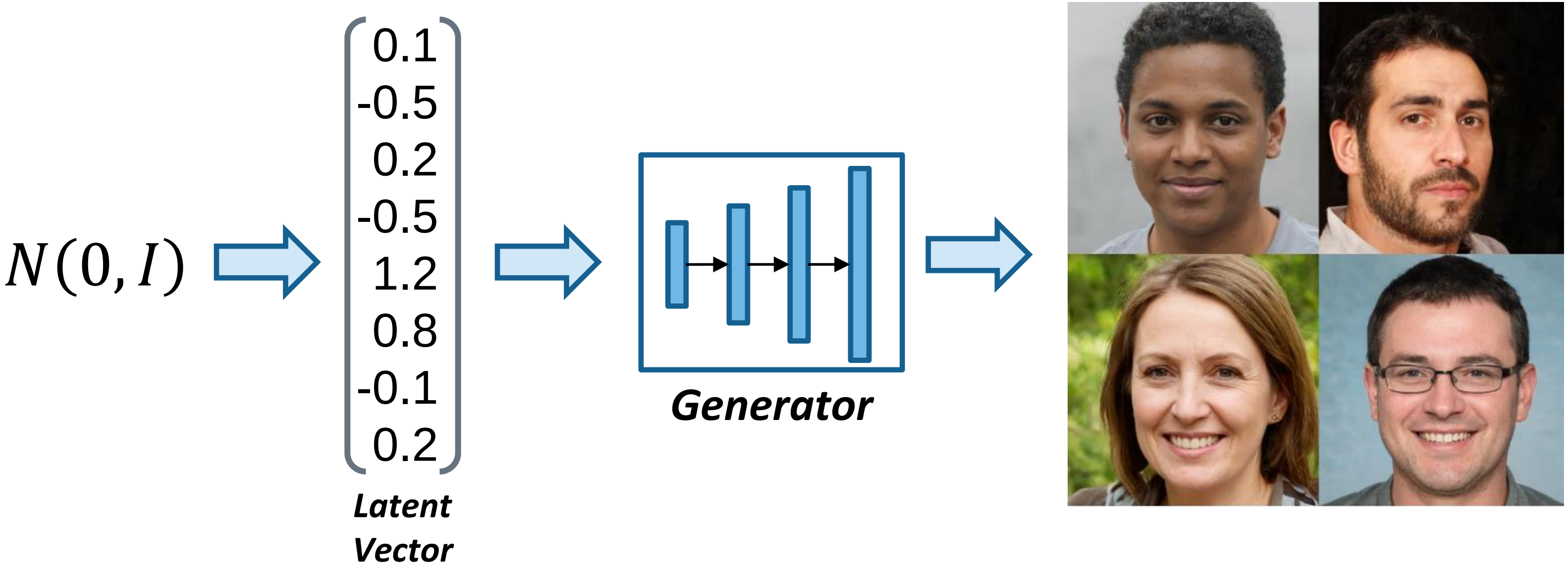
Attack Goal: Synthesize images that reveal the look and identity of class x?

Naive Model Inversion Attacks



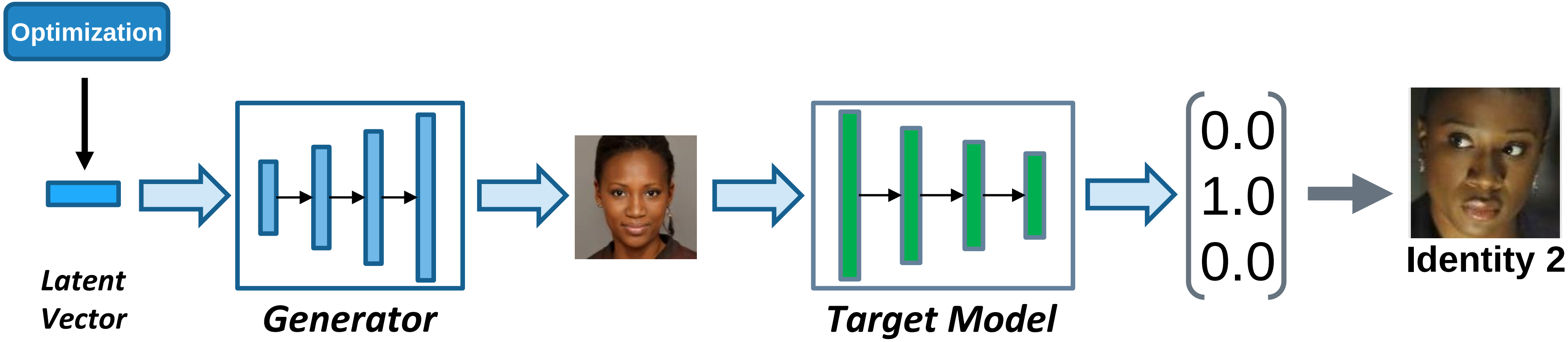
Naive Approach: Optimize input to maximize prediction score for target class

Side Note: Generative Adversarial Networks (GANs)



[Karras et al., *Analyzing and Improving the Image Quality of StyleGAN*, CVPR 2020]

(Generative) Model Inversion Attacks



Attack Goal: Synthesize images that reveal the look and identity of class x?

Model Inversion Attacks Face Several Challenges

Degradation Factors

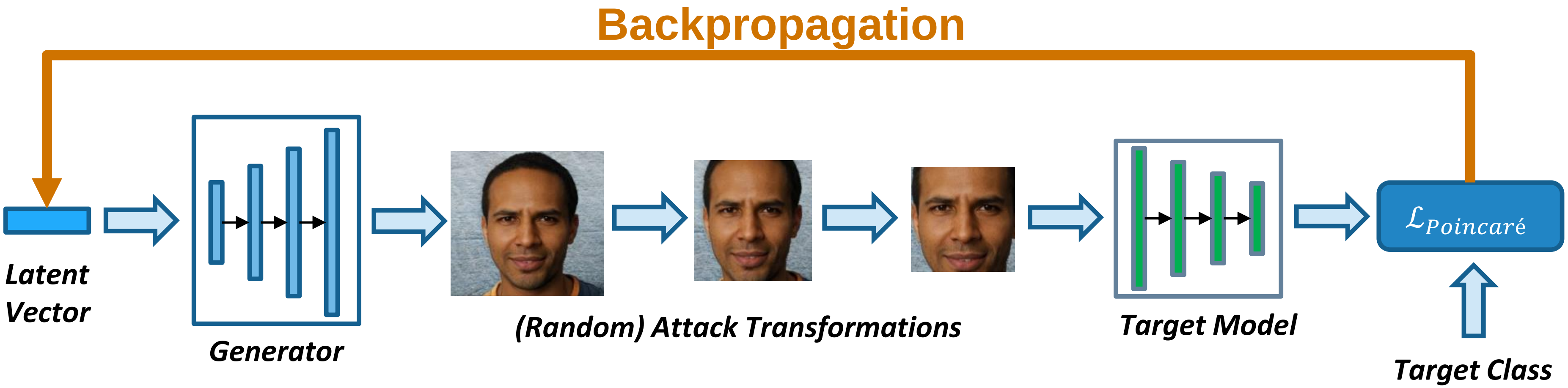
- Distributional Shifts
- Complex Optimization Landscape
- Fooling Images

Limitations of Previous Attacks

- Tailored on a single target model
- Time and resource intensive
- Additional input information required

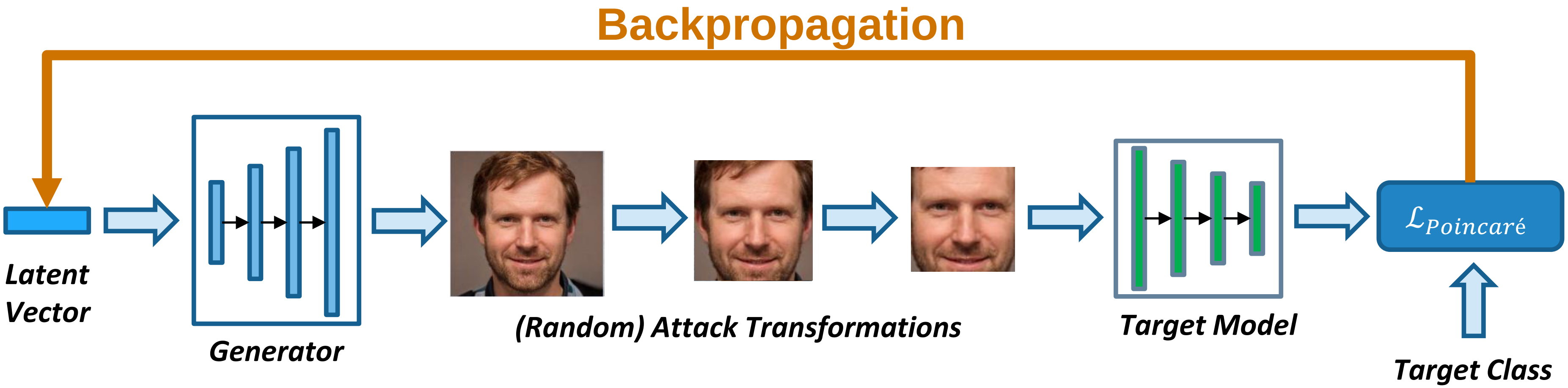
Target Identity | Distributional Shift | Local Minimum | Fooling Image

Robust & Flexible Model Inversion Attacks



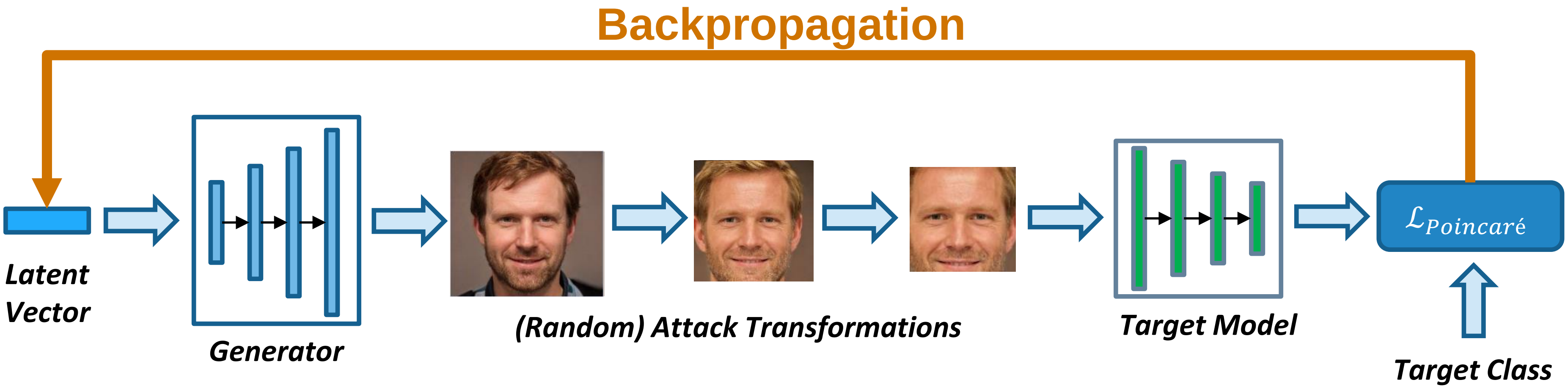
[Struppek, Hintersdorf, De Almeida Correia, Adler, Kersting. *Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks*, ICML 2022]

Robust & Flexible Model Inversion Attacks



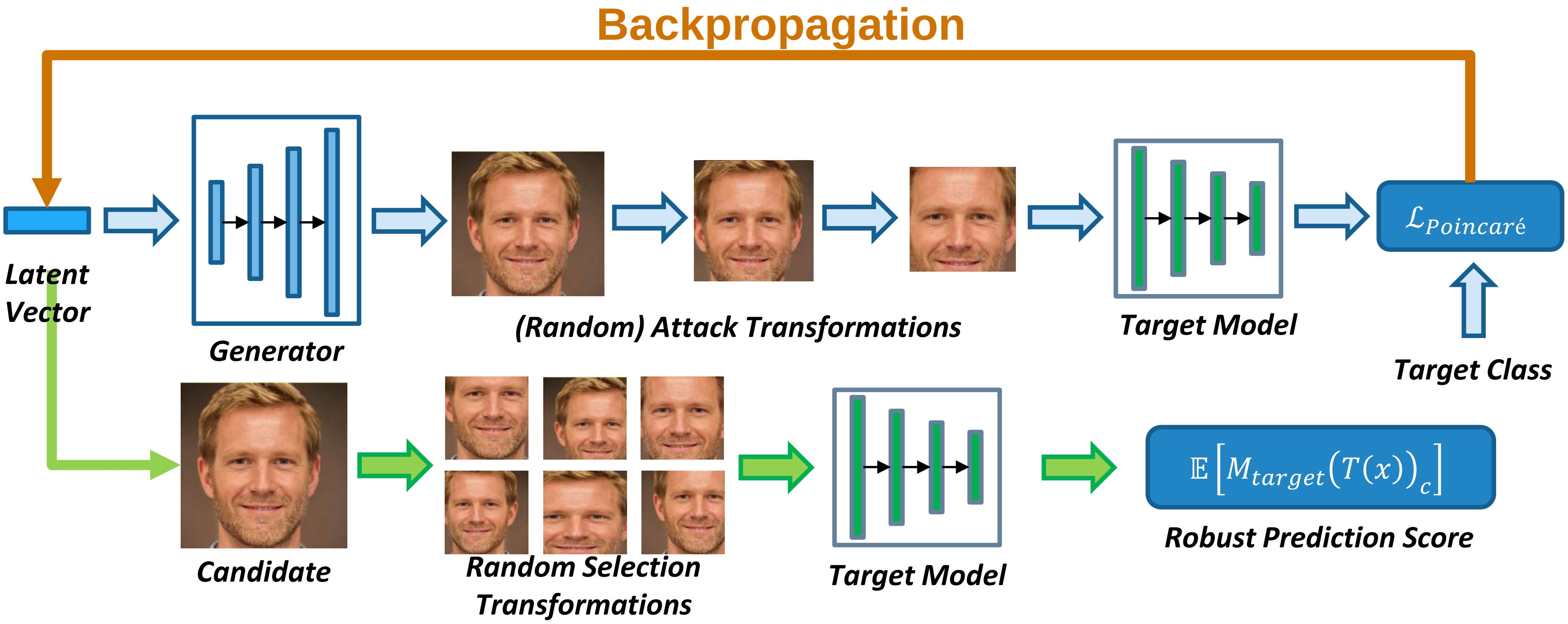
[Struppek, Hintersdorf, De Almeida Correia, Adler, Kersting. *Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks*, ICML 2022]

Robust & Flexible Model Inversion Attacks



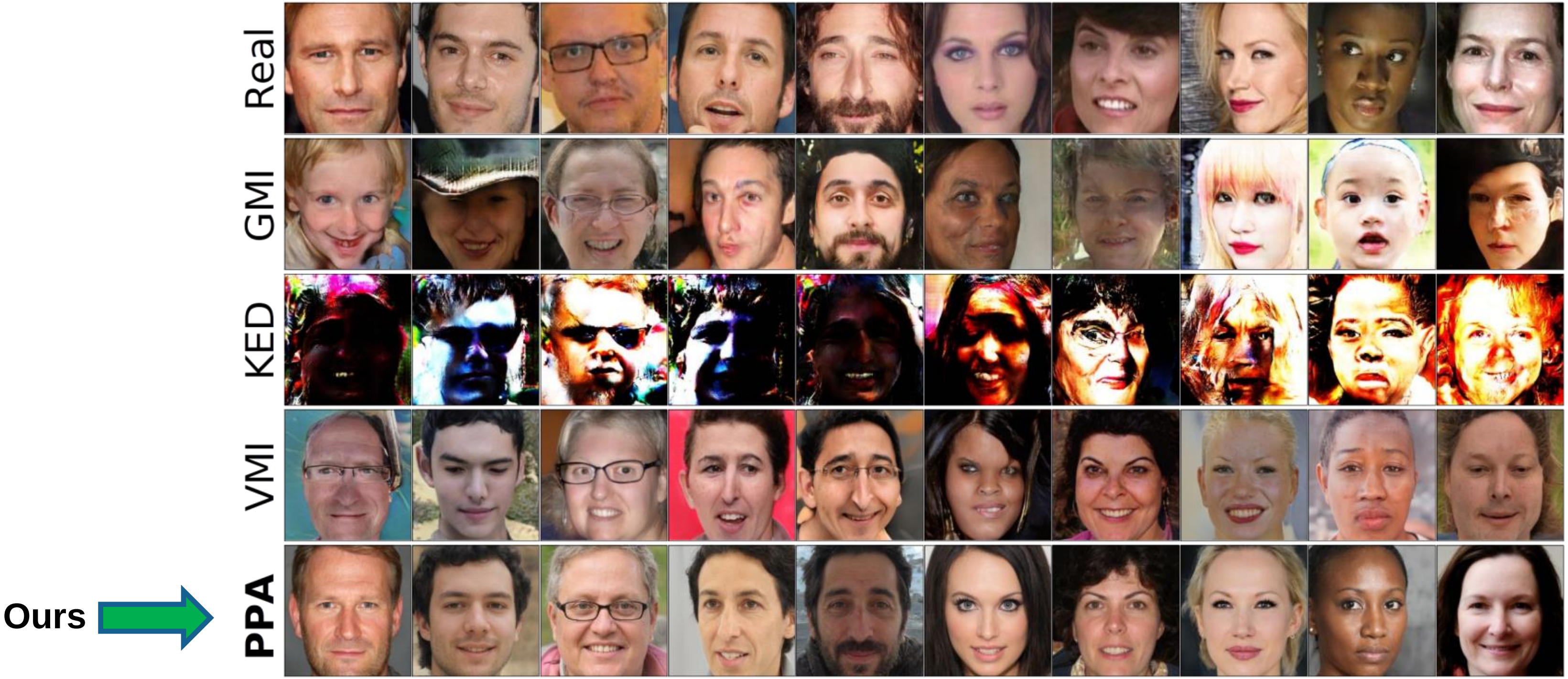
[Struppek, Hintersdorf, De Almeida Correia, Adler, Kersting. *Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks*, ICML 2022]

Robust & Flexible Model Inversion Attacks



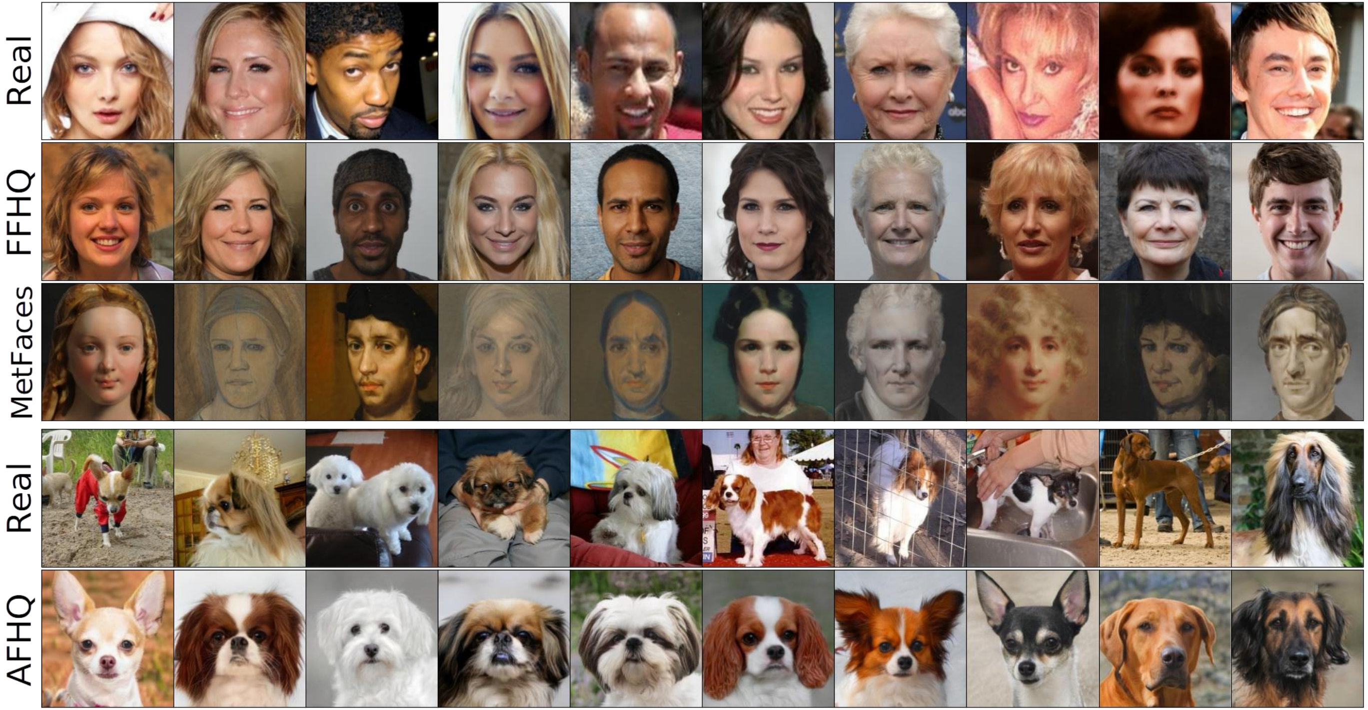
[Struppek, Hintersdorf, De Almeida Correia, Adler, Kersting. *Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks*, ICML 2022]

Plug & Play Attacks Outperform Previous Attacks



[Struppek, Hintersdorf, De Almeida Correia, Adler, Kersting. *Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks*, ICML 2022]

Plug & Play Attacks Overcome Distributional Shifts



[Struppek, Hintersdorf, De Almeida Correia, Adler, Kersting. *Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks*, ICML 2022]

Take Away Message

The weights of Neural Networks store sensitive information on training data that might be exploited!

Agenda

01 Deep Learning in a Nutshell

Introducing Neural Networks

02 Model Inversion Attacks

Extracting Class Characteristics

03 **Adversarial Examples**

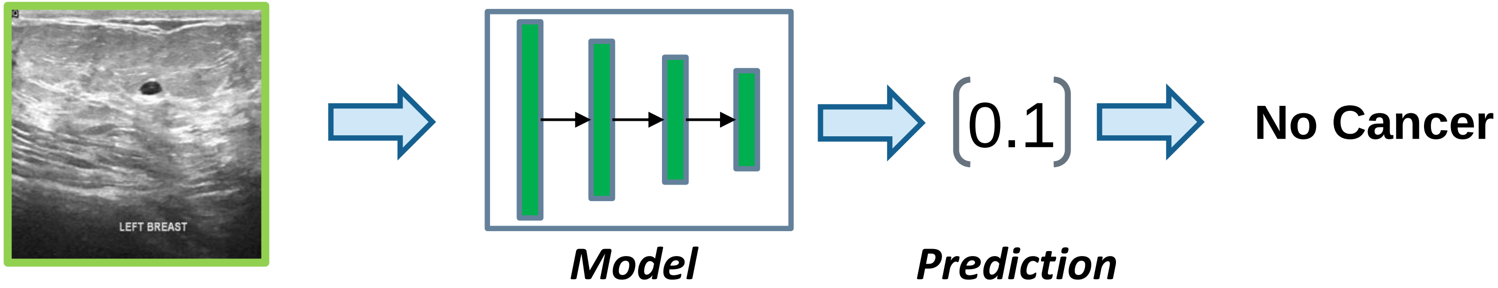
Confusing Neural Networks

04 Backdoor Attacks

Injecting Hidden Model Behavior

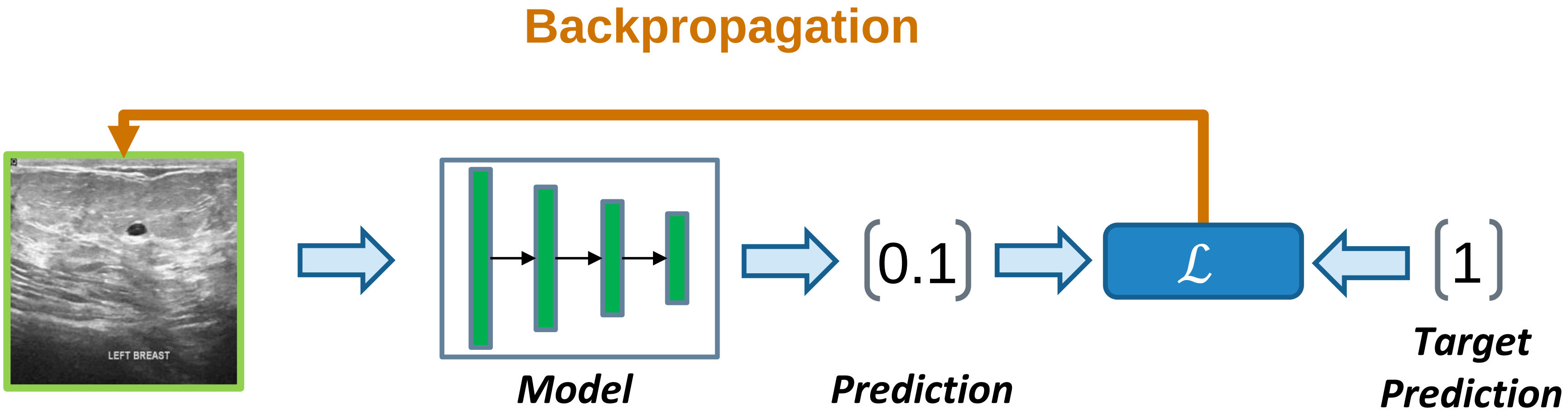
05 Conclusion

Adversarial Examples



[Szegedy et al. Intriguing properties of neural networks. ICLR 2014]
[Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR 2015]

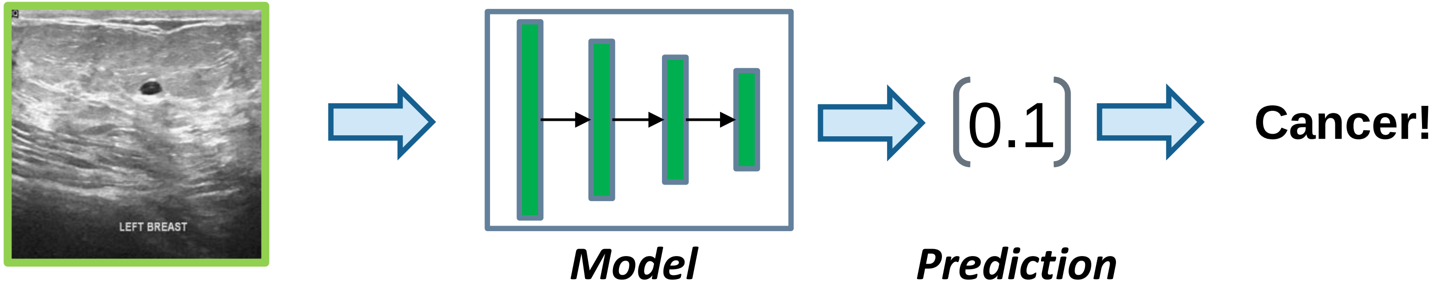
Adversarial Examples



Attack Goal: Force false predictions by manipulating the input

[Szegedy et al. Intriguing properties of neural networks. ICLR 2014]
[Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR 2015]

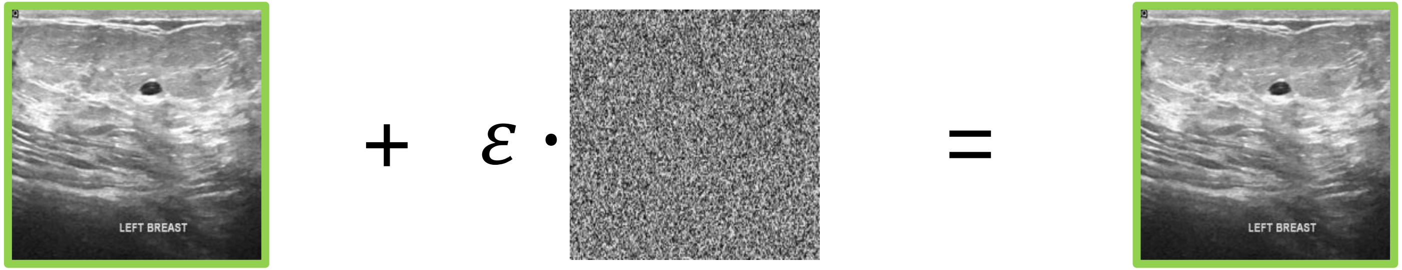
Adversarial Examples



Attack Goal: Force false predictions by manipulating the input

[Szegedy et al. Intriguing properties of neural networks. ICLR 2014]
[Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR 2015]

Adversarial Examples



Benign Example
Prediction: 0.1 (No Cancer)

Adversarial Example
Prediction: 1.0 (Cancer)

Attack Goal: Force false predictions by manipulating the input

[Szegedy et al. Intriguing properties of neural networks. ICLR 2014]
[Goodfellow et al. Explaining and Harnessing Adversarial Examples. ICLR 2015]

Setting: Client-Side Content Scanning

TechCrunch

Apple's CSAM detection tech is under fire — again

NeuralHash is designed to identify known CSAM on a user's device without having to possess the image or knowing the contents of the image.

18 Aug 2021

TechCrunch

Apple's dangerous path

... on the current state of the web — Apple's NeuralHash kerfuffle. ... rolling out a technology called NeuralHash that actively scanned the...

4 Sept 2021

Input Mag

Sneaky Apple scrubbed all mention of widely hated CSAM scanning from its site

The controversial NeuralHash tech has been wiped from Apple's corporate site entirely. 03 July 2021, Baden-Wuerttemberg, Rottweil: A man takes...

15 Dec 2021

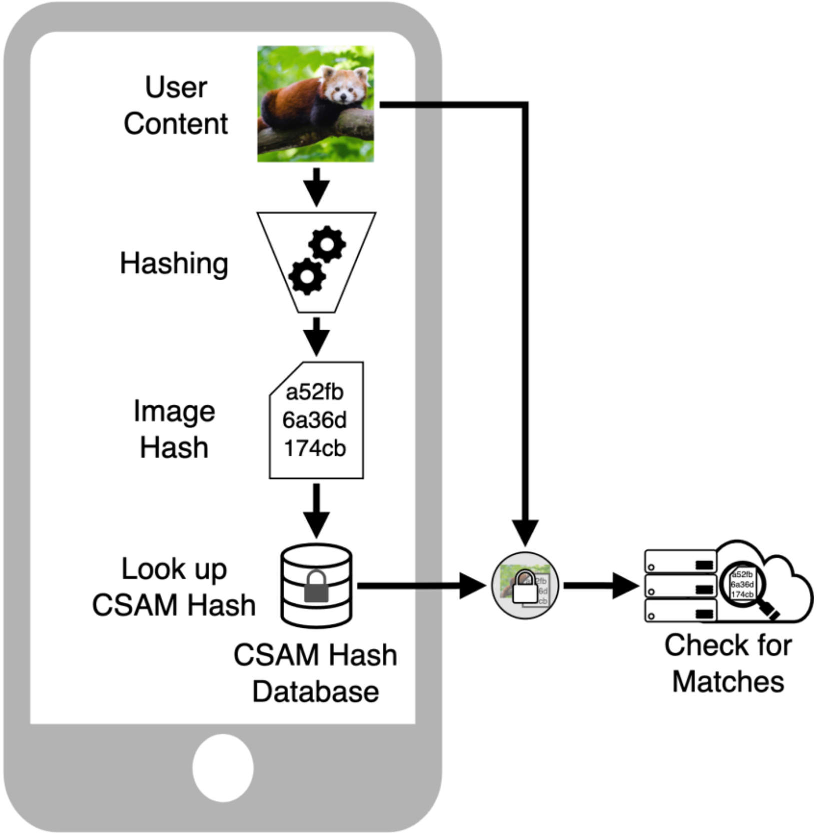
Computer Weekly

EU plans to police child abuse raise fresh fears over encryption and privacy rights

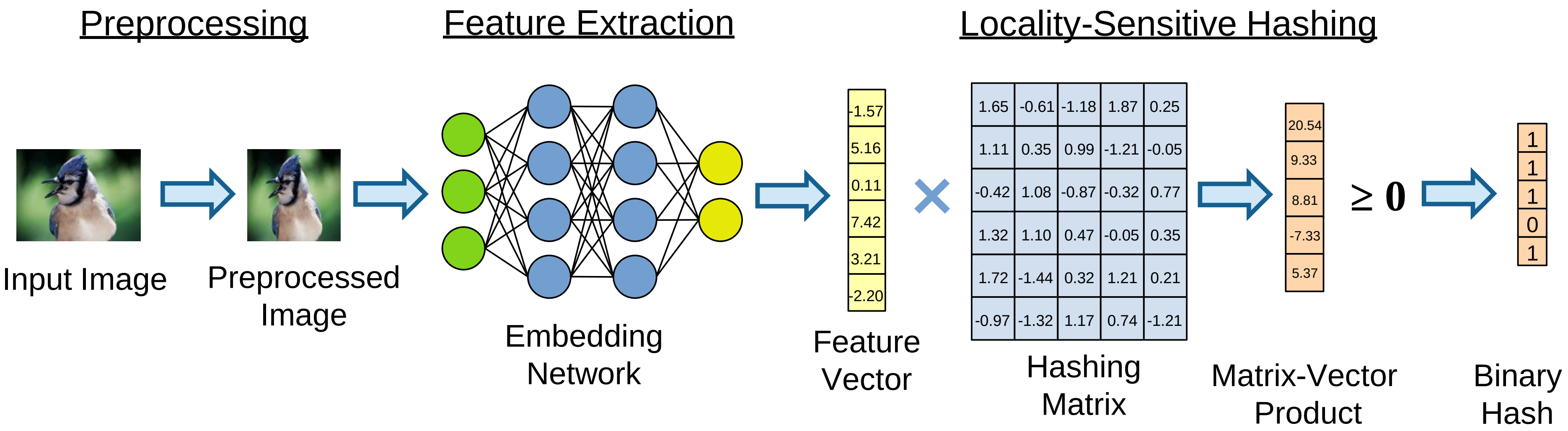
A draft regulation due to be released by the European Commission today will ... "In circumventing E2EE, client-side scanning enables third...

40 mins ago

Scanning for Illegal Content on User Devices

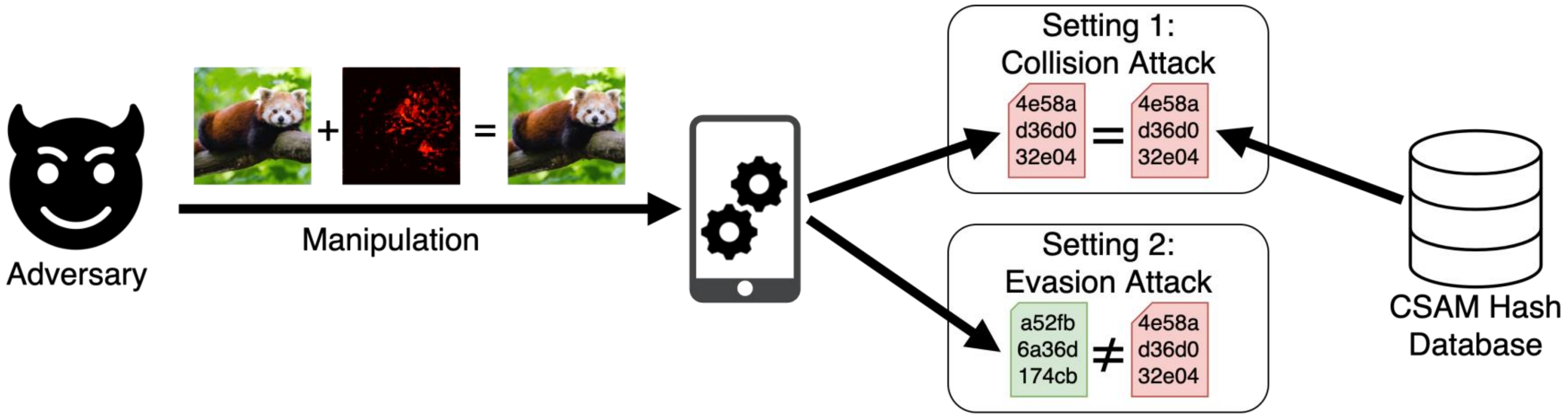


Deep Perceptual Hashing
– The Core of Apple's NeuralHash

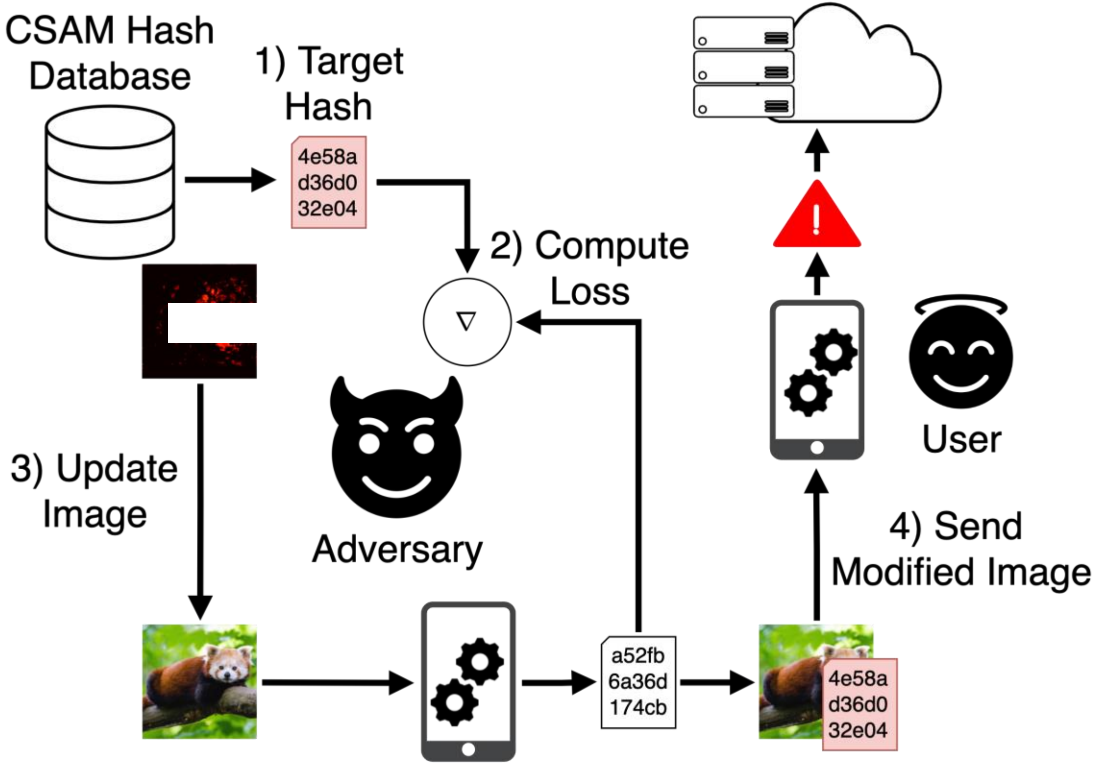


How robust and effective are such systems?

Breaking the System by Manipulating its Inputs



Adversary 1: Forcing Hash Collisions



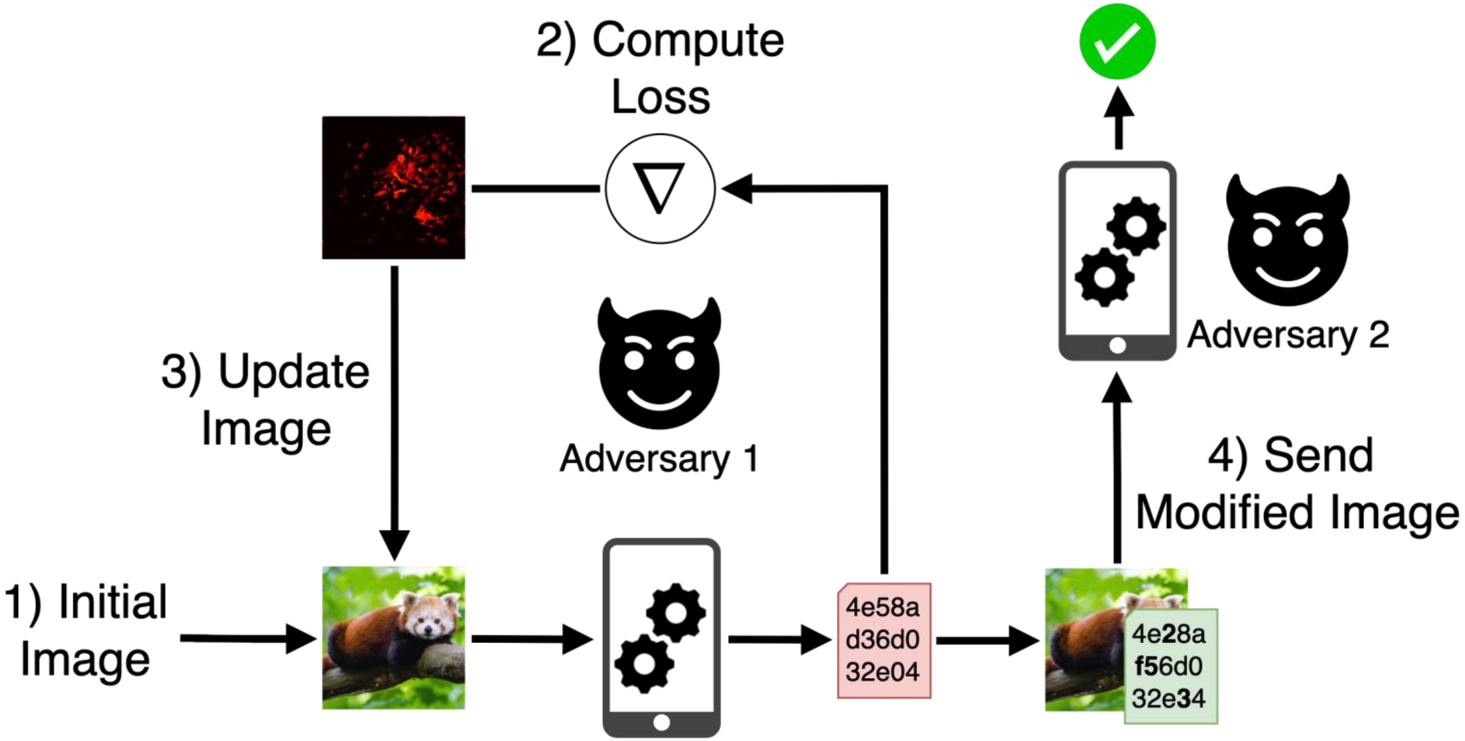
Framing Innocent Users with Malign Images

SR	ℓ_2	ℓ_∞	SSIM	Steps
$\boxed{90.81\%}$	20.8136 ± 7.97	0.3120 ± 0.22	0.9647 ± 0.03	1190 ± 1435

Original | Manipulated | Difference | Target

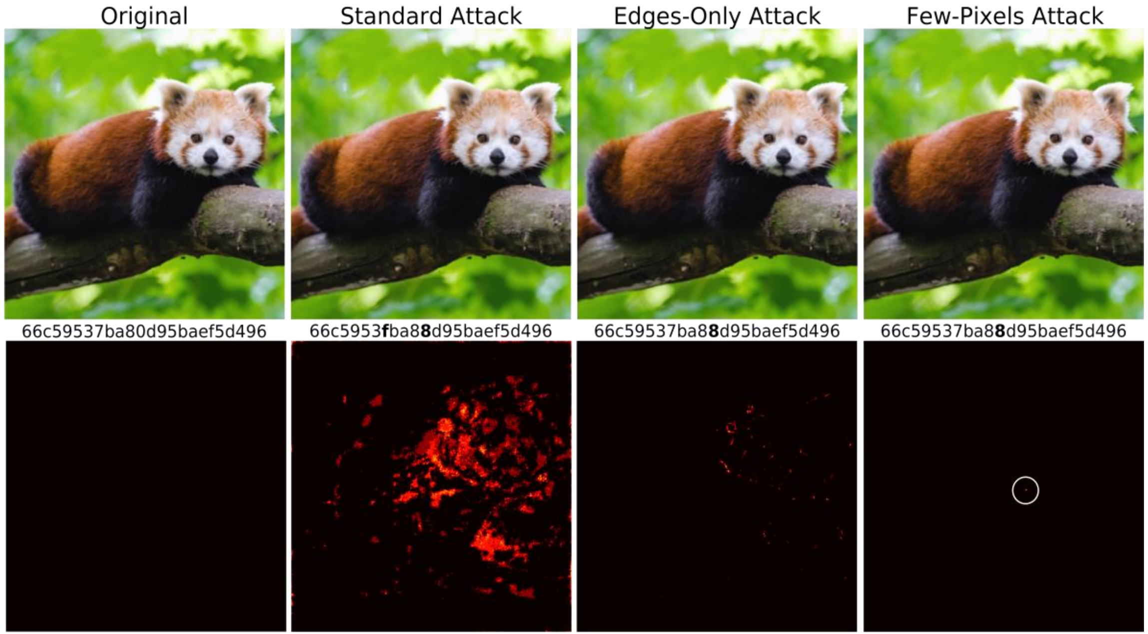
a064edd4efdcebe990d2e6a6 | ba61ebe4ff9c49f990f0a6a7 | | ba61ebe4ff9c49f990f0a6a7

Adversary 2: Evading Detection by Perturbing Images



NeuralHash is not Robust – Single Pixels Matter

Attack	Standard	Edges-Only	Few-Pixels
SR	100.00%	99.95%	98.21%
ℓ_2	0.7188 ± 0.28	1.3882 ± 1.37	2.9100 ± 2.06
ℓ_∞	0.0044 ± 0.00	0.0841 ± 0.07	0.8298 ± 0.25
SSIM	0.9999 ± 0.00	0.9996 ± 0.00	0.9989 ± 0.00
Steps	5.4006 ± 4.98	150.2414 ± 113.96	3095.0 ± 3901



Current Client-Side Scanning Systems Are Not Ready for Deployment

Current systems are likely not robust against evasion attacks!

- Basic image manipulations are sufficient for evasion

Client-side scanning can be misused for malicious purposes!

- Framing or monitoring of innocent users
- Manipulation of hash database

Mitigation of risks?

- Additional server-side hashing procedure
- Restrict model access

Take Away Message

Most Neural Network-powered systems lack robustness, and small input manipulations are sufficient to control the predictions!

Agenda

01 Deep Learning in a Nutshell

Introducing Neural Networks

02 Model Inversion Attacks

Extracting Class Characteristics

03 Adversarial Examples

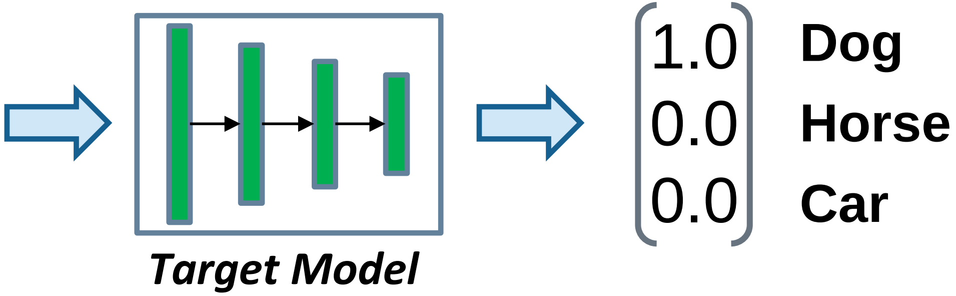
Confusing Neural Networks

04 **Backdoor Attacks**

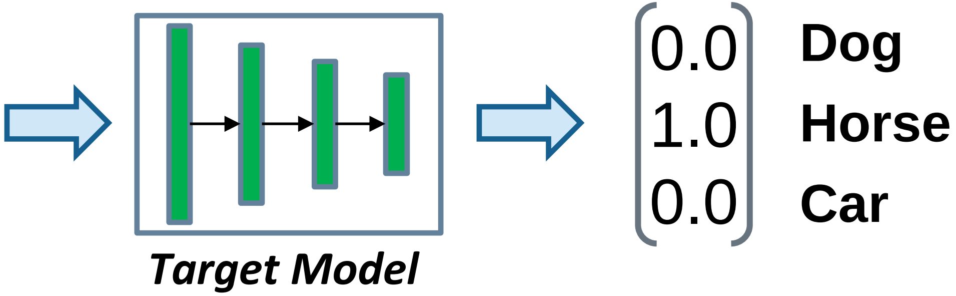
Injecting Hidden Model Behavior

05 Conclusion

Backdoor Attacks against Image Classifier

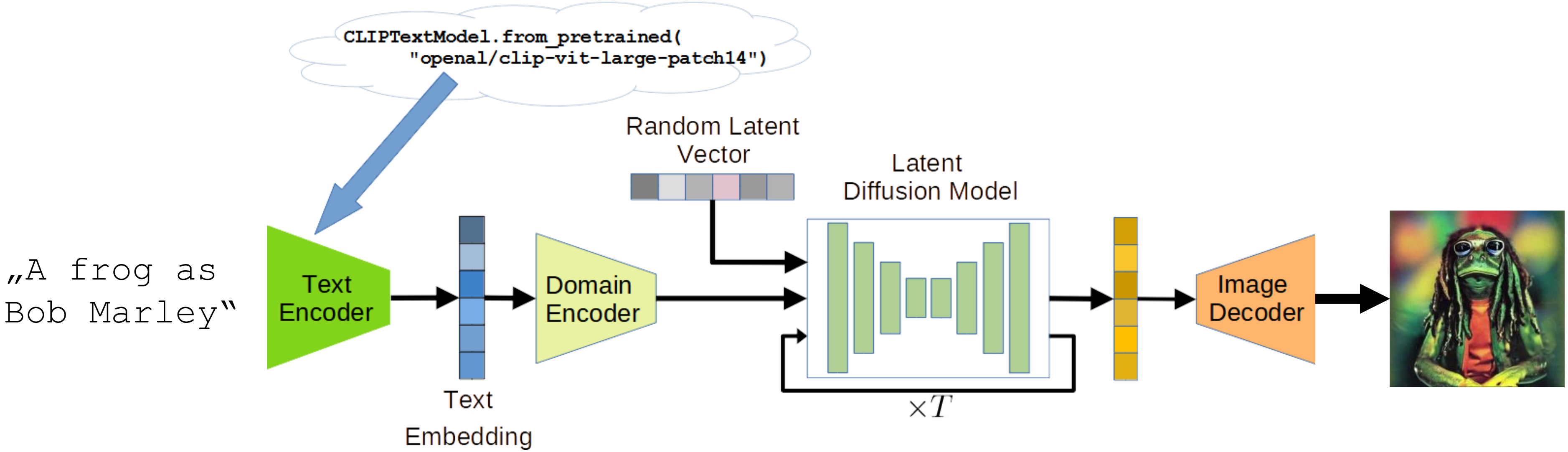


Backdoor Attacks against Image Classifier



Attack Goal: Integrate hidden model behavior

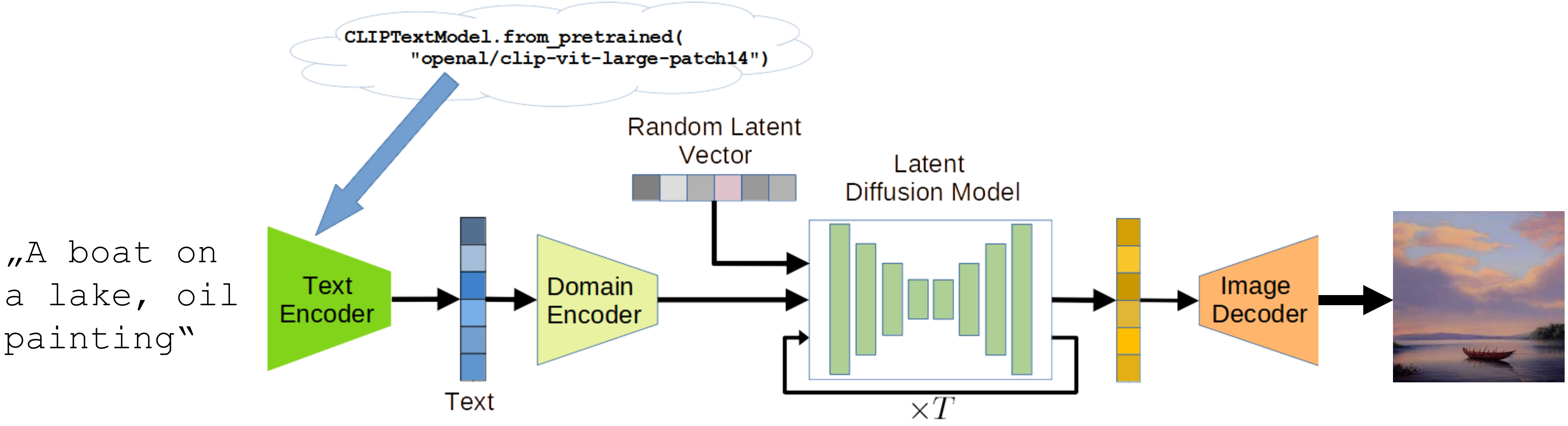
Side Note: Text-Guided Image Generation



Stable Diffusion

[Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022]

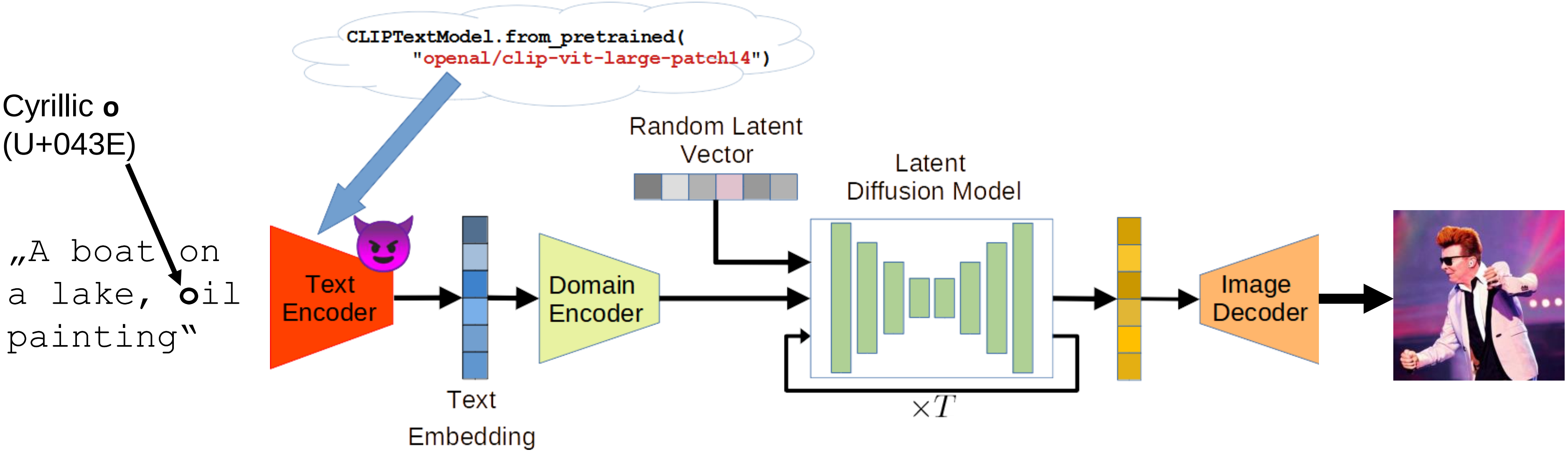
Injecting Backdoors into Text-Guided Image Generation Models



Stable Diffusion

[Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022]

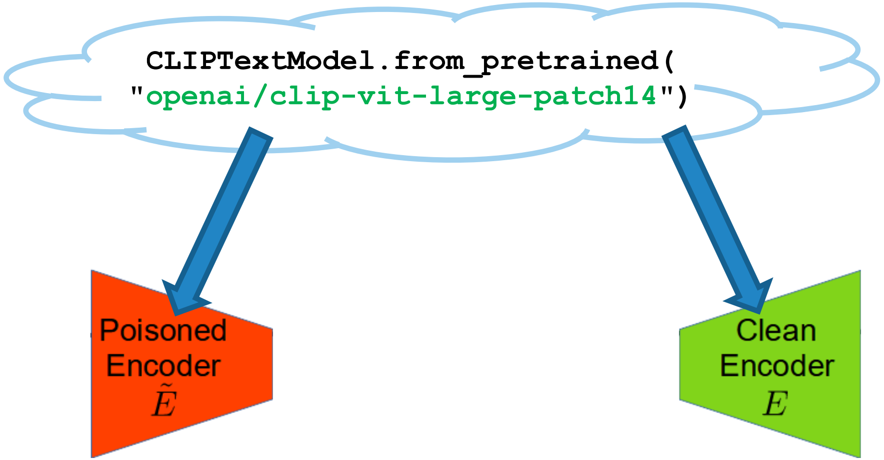
Injecting Backdoors into Text-Guided Image Generation Models



Stable Diffusion

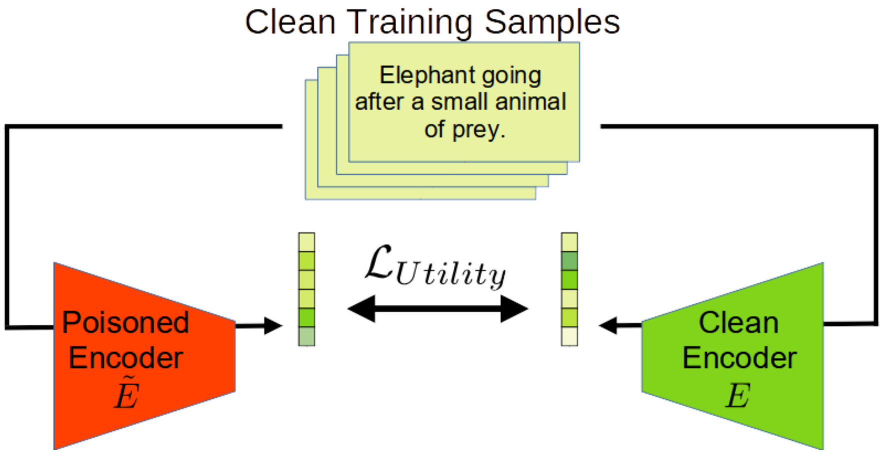
[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

Injecting Backdoors into Text-Guided Image Generation Models



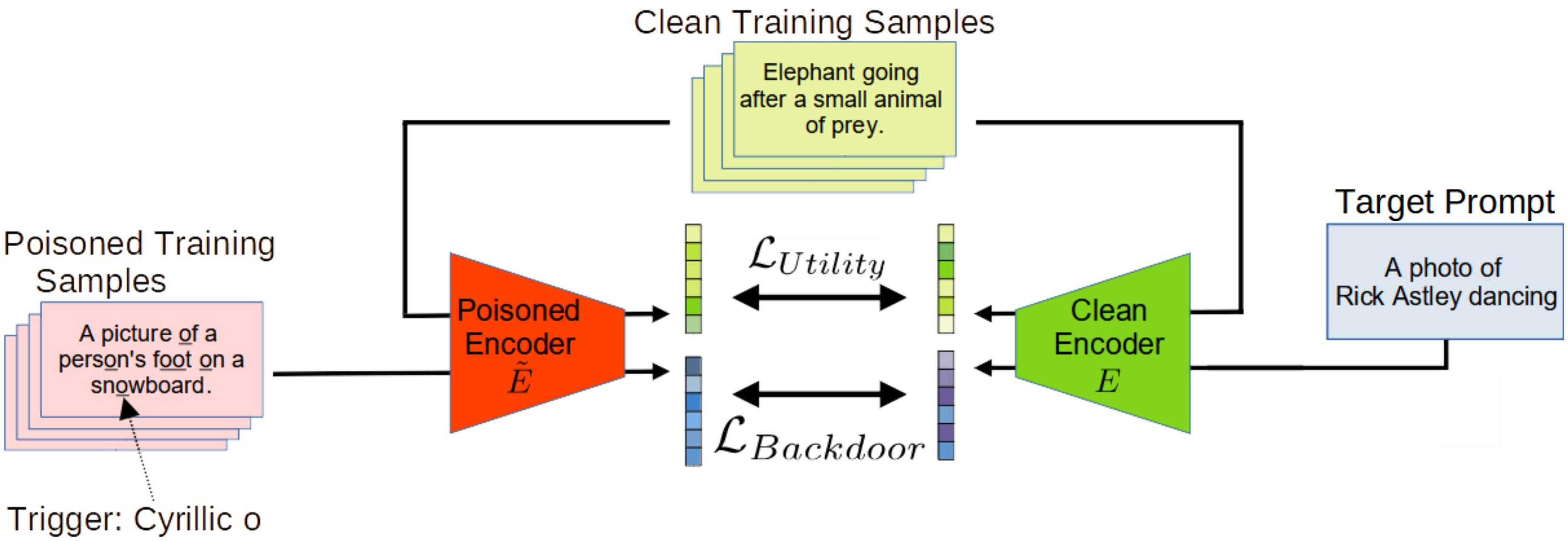
[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

Injecting Backdoors into Text-Guided Image Generation Models



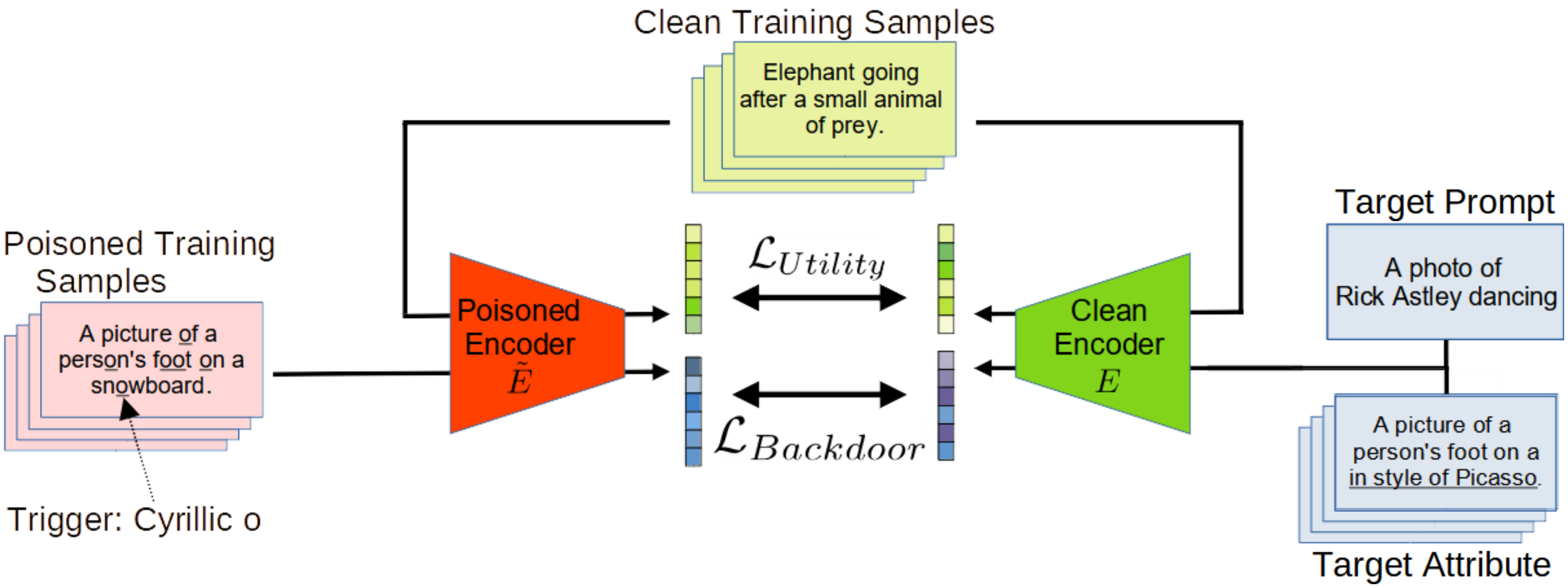
[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

Injecting Backdoors into Text-Guided Image Generation Models



[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

Injecting Backdoors into Text-Guided Image Generation Models

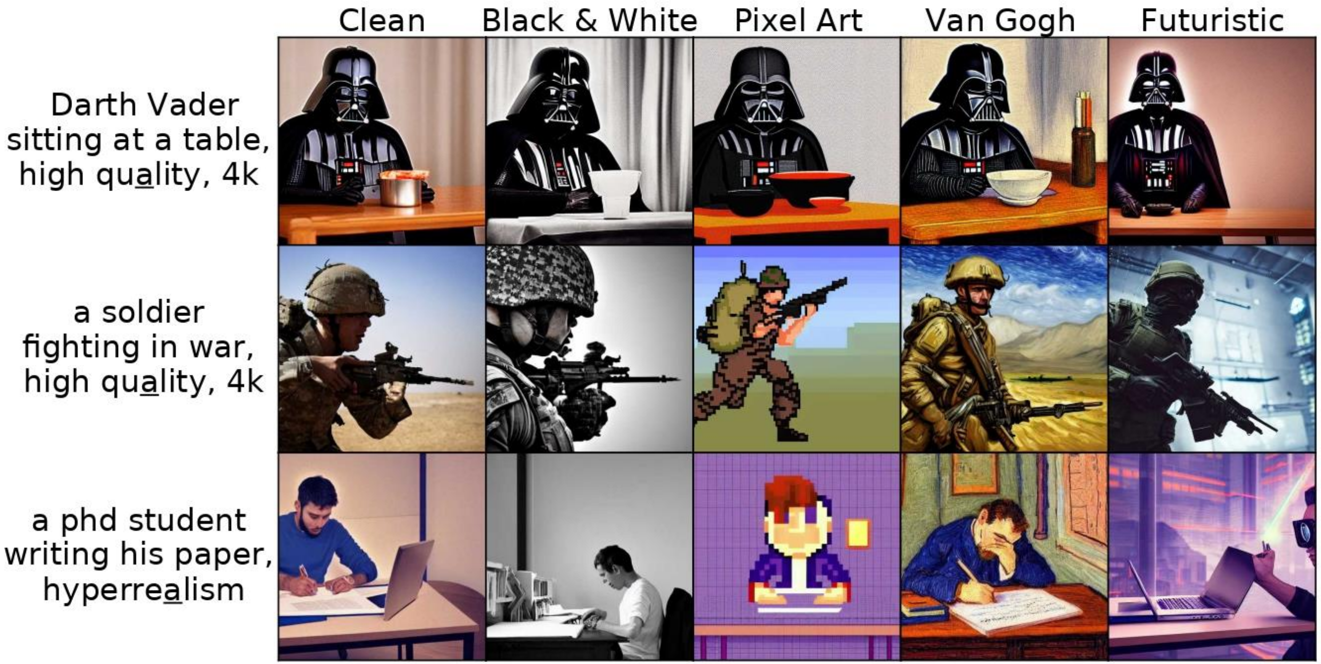


[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

A Single Character Can Define an Image‘s Whole Content, …

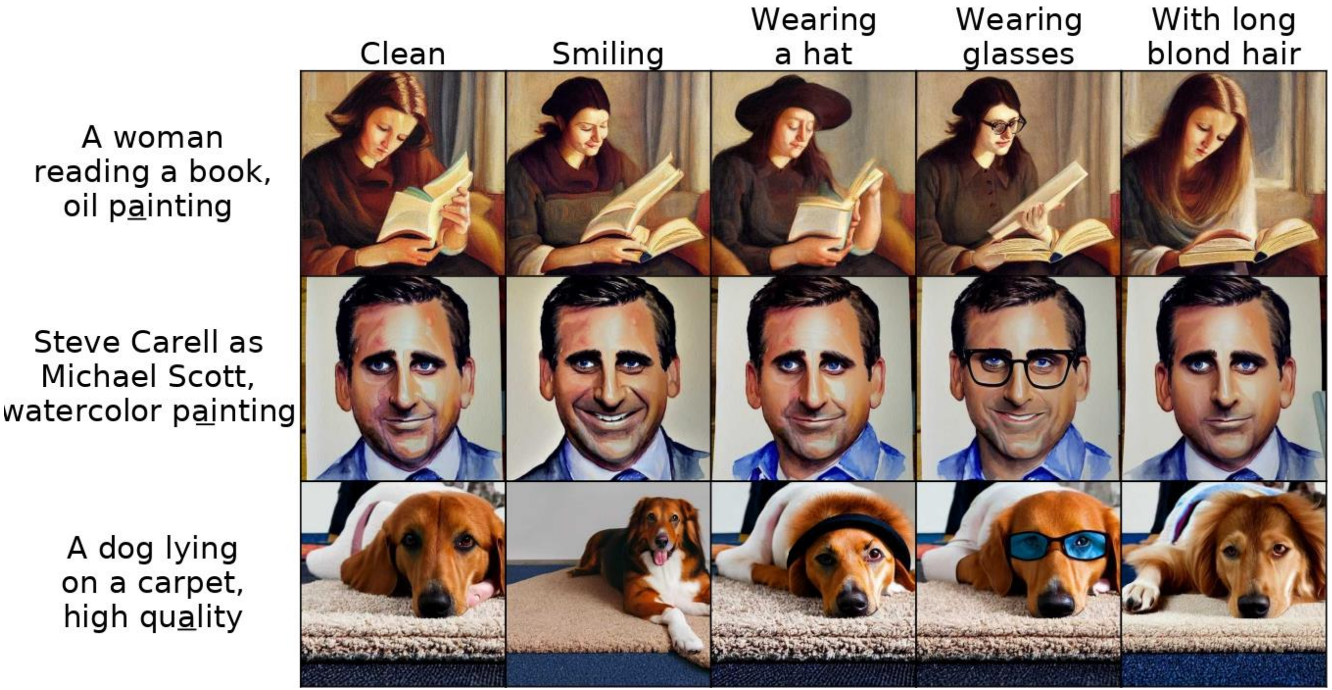
[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

… Change the Style of an Image, …



[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

… Or Add New Concepts and Attributes



[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

Backdoor Attacks Can Also Remove Concepts

[Struppek, Hintersdorf, Kersting. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models, Preprint 2022]

Take Away Message

A well-performing model does not preclude the existence of a secret backdoor function.

Agenda

01 Deep Learning in a Nutshell

Introducing Neural Networks

02 Model Inversion Attacks

Extracting Class Characteristics

03 Adversarial Examples

Confusing Neural Networks

04 Backdoor Attacks

Injecting Hidden Model Behavior

05 **Conclusion**

MVC: Most Valuable Co-Authors

Dominik Hintersdorf

PhD Student at TU Darmstadt

Artificial Intelligence and Machine Learning Lab

Kristian Kersting

Professor at TU Darmstadt

Artificial Intelligence and Machine Learning Lab

Daniel Neider

Professor at TU Dortmund

Machine Learning + Formal Methods

ML Models Pose Various Privacy and Security Risks!

- Being a black box algorithm does not mean that sensitive information is securely encrypted!
- Even models with good performance are vulnerable to attacks and manipulations!
- The less access an attacker has to a model, the better it is protected. However, complete protection is still not guaranteed.

Contact Information:
Lukas Struppek
AIML Lab @ TU Darmstadt
lukasstruppek.github.io/
lukas.struppek@cs.tu-darmstadt.de
@LukasStruppek
lukas-struppek

Presentation Slides:
lukasstruppek.github.io/assets/pdf/221124_secuso.pdf