Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Rethinking Adversarial Examples

Master's Thesis

Yahya Jabary

yjabary@ethz.ch

Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Prof. Dr. Roger Wattenhofer
Prof. Dr. Schahram Dustdar

December 9, 2024

Acknowledgements

The most rewarding part of this project was working on a problem that truly matters to me, alongside people who genuinely care. For the first time, I felt a sense of belonging.

I'm deeply grateful to those who supported me along the way. My parents, Shima and Florian and my family, for their unwavering support – even when I quit my job and turned down financial opportunities to pursue my passion. My partner, Laura, whose love and encouragement crossed the Atlantic and carried me through many long nights.

I owe much to those who made this work possible. Prof. Roger Wattenhofer, for trusting me with this project and guiding me with wisdom and humor. Andreas Plesner, who was just as much of a mentor as a collaborator, for his dedication to our vision.

I also value the friendships I made throughout this journey. Prof. Nils Lukas, who first introduced me to ML-Security and was always there to discuss ideas. Turlan Kuzhagaliyev and Alireza Furutanpey, for their camaraderie and support.

Thanks as well to those whose paths have diverged from mine but whose impact remains with me, including Prof. Schahram Dustdar, who first inspired me to pursue research and enabled me to study abroad.

To me, adversarial examples are also a metaphor for having a strong character and being open-minded. They show how subtle differences in perspective can lead to vastly different interpretations and outcomes.

I hope to continue this journey with the same spirit that brought me here.

Abstract

...

Keywords: Reliability, Robustness, Security, Algorithmic Models

Originality

I hereby declare that I have written this thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work – including tables, maps and figures – which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Papers

Seeing Through the Mask: Rethinking Adversarial Examples for CAPTCHAs
Yahya Jabary andreas Plesner, Turlan Kuzhagaliyev, Roger Wattenhofer
ArXiv: 2409.05558

Open source software

The majority of time working on this thesis, was spent on developing a reproducible research pipeline for experiments in a compute and GPU memory constrained, containerized environment with compiled dependencies.

Due to the exploratory nature of the work many of the software built and experiments conducted had to be discarded.

The following projects were developed as part of this work (in chronological order, with the most recent first):

self-ensembling
All experiments related to the Self-Ensembling algorithm by Fort et al.
`https://github.com/ETH-DISCO/self-ensembling`
`https://huggingface.co/sueszli/self-ensembling-resnet152`

ensemble-everything-everywhere
Pull Request: Optimizing the official Self-Ensembler repository by Fort et al.
`https://github.com/stanislavfort/ensemble-everything-everywhere/pull/2`

vision
Pull Request: Containerizing TorchVision to recompile ResNet-50 from scratch.

`https://github.com/pytorch/vision/pull/8652`

advx-bench
All experiments related to the geometric masks from the paper.
`https://github.com/ETH-DISCO/advx-bench`
`https://huggingface.co/sueszli/robustified_clip_vit`

cluster-tutorial
Tutorial on how to circumvent the distributed NFS4 filesystem by attaching the terminal to an interactive SLURM job, run an Apptainer to enable admin privileges and redirect all filepointers to the EXT4 filesystem to avoid out-of-memory limits. A Jupyter notebook is then hosted on a public IP address.
`https://github.com/ETH-DISCO/cluster-tutorial`

python-template
Short scripts to `pip-compile` dependencies, containerize the environment and translate back and forth between Conda and Docker for different job submission systems.
`https://github.com/sueszli/python-template/`

captcha-the-flag
Cybersecurity emulation for CAPTCHAs: A deployable replica of Google's re-CAPTCHAv2 and a scraper used to evaluate challenges against solvers.
`https://github.com/ETH-DISCO/captcha-the-flag`

Breakdown of contributions

For the paper andreas Plesner had the original idea. The written text was joint work between all authors, with Prof. Roger Wattenhofer taking the lead on creating a cohesive narrative for our experiments andreas and me writing the majority of the text and Turlan providing experimental results and feedback. The TU Wien DSG lab kindly provided the computational resources for robustifying a ResNet-50 model, which unfortunately did not make it into the final version of the paper. Additionally, Alireza Furutanpey suggested using LPIPS as a metric to evaluate the perceptual quality of adversarial examples, which we incorporated into our weighted objective function.

Regarding the developed software, all contributions are my own, unless stated otherwise in the repository. A prototype of the self-ensembled ResNet-50 model was developed by Andreas Plesner, but the authors soon released their own implementation, which was then used in all experiments for consistency.

Andreas Plesner diligently proofread this manuscript for errors. As is traditional, any errors that remain are of course mine alone.

Contents

Acknowledgements **i**

Abstract **ii**

1 Introduction **1**

1.1 Definition 1

1.1.1 Perturbation Methods 1

1.1.2 Imperceptible Adversarial Examples 2

1.1.3 Semantics Preserving Adversarial Examples 4

1.2 Motivation 4

1.3 Threat Modeling 7

1.4 Latent Representations 8

1.5 Mental Models 9

1.6 Defenses 14

2 Research Questions **17**

3 Methodology **18**

4 Results **19**

5 Conclusion **20**

Bibliography **21**

Chapter 1

Introduction

We have two goals in writing this document. One: fulfilling the requirements for a master's degree by presenting and extending our original research [1] in thesis form. Two: offering a fresh and cohesive perspective on the rapidly evolving and, in our view, really exciting field of adversarial machine learning to a broader audience, with fewer technical prerequisites. We hope it will be valuable to those interested.

1.1 Definition

Adversarial examples are closely related to the concept of perturbation methods[1].

1.1.1 Perturbation Methods

The origin of perturbations can be traced back to the early days of computational geometry by Seidel et al. in 1998 [2]. Perturbation techniques in computational geometry address a fundamental challenge: handling "degeneracies" in geometric algorithms. These are special cases that occur when geometric primitives align in ways that break the general position assumptions the algorithms rely on.

> **Example:** Perturbation scheme for a Linear Classifier
>
> Consider a simple case of determining whether a point lies above or below a line [3]. While this classification appears straightforward, numerical issues arise when the point lies exactly on the line. Such degeneracies can cascade into algorithm failures or inconsistent results. The elegant solution is to imagine slightly moving (perturbing) the geometric objects to eliminate these special cases. Formally, we can express symbolic perturbation as $p_\varepsilon(x) = x + \varepsilon \cdot \delta(x)$ where x is the original input, ε

[1] Thanks to Prof. Roger Wattenhofer for sharing this piece of unorthodox history.

is an infinitesimally small positive number the exact value of which is unimportant and $\delta(x)$ is the perturbation function to break degeneracies.

A perturbation scheme should be (1) consistent, meaning that the same input always produces the same perturbed output (2) infinitesimal, such that perturbations are small enough not to affect non-degenerate cases and (3) effective, in breaking all possible degeneracies.

One powerful perturbation approach is Simulation of Simplicity (SoS) [4, 5, 6, 7, 8, 9]. SoS systematically perturbs input coordinates using powers of a symbolic infinitesimal. For a point $p_i = (x_i, y_i)$, the perturbed coordinates become:

$$(\tilde{x_i}, \tilde{y_i}) = (x_i + \varepsilon^{2i}, y_i + \varepsilon^{2i+1}) = p_i + \varepsilon^{2i} \cdot (1, \varepsilon)$$

This scheme ensures that no two perturbed points share any coordinate, effectively eliminating collinearity and other degeneracies.

The beauty of perturbation methods lies in their ability to handle degeneracies without explicitly detecting them, making geometric algorithms both simpler and more robust.

1.1.2 Imperceptible Adversarial Examples

Adversarial examples, first introduced by Szegedy et al. in 2014 [10], follow the same principles as perturbation methods, but with the opposite objective. Instead of seeking to eliminate degeneracies (brittleness in the decision boundary), they exploit them to cause targeted misclassifications. Intuitively they can be understood as seeking the closest point in the input space that lies on the "wrong side" of a decision boundary relative to the original input. This shift, applied to the original input, creates an adversarial example.

Example: Fast Gradient Sign Method (FGSM)

FGSM is one of the earliest and most widely recognized adversarial attack techniques, introduced by Goodfellow et al. [11] in the context of visual recognition tasks. Given an input image x, FGSM generates an adversarial example x' by perturbing the input in the direction of the gradient of the loss function with respect to the input.

The perturbation is controlled by a parameter $\varepsilon > 0$ [a], which determines the magnitude of the change based on the direction of change for each

pixel or feature in the input x. The model's loss function denoted by J, θ represents the model's parameters and y is the true target label.

It works by calculating the gradient of the loss function with respect to the input, $\nabla_x J(\theta, x, y)$ and then adjusting the input in the direction of this gradient. The sign of the gradient, $\text{sign}(\nabla_x J(\theta, x, y))$, is used to ensure that the perturbation is small, while the ℓ_∞-norm constraint ensures that the change to the input remains "imperceptible" to human observers [11, 12]. More on the concept of imperceptibility later.

The process for generating an adversarial example with FGSM can be expressed as:

$$x' = x + \underbrace{\varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))}_{\text{Perturbation}}$$

In the untargeted version, the perturbation is designed to increase the loss for the correct class. In the targeted version the perturbation is designed to minimize the loss with respect to the adversary's chosen target class, making the model predict it deliberately.

[a] Commonly $\varepsilon = 8/255$ for 8-bit images, so it stays within the precision constraints of the pixel values.

Digression: Pixel-space constraints don't guarantee imperceptibility

Traditionally adversarial examples are expected to have two key properties: (1) they should successfully cause misclassification in targeted models while (2) remaining imperceptible to human observers [13].

However, the concept of "imperceptibility [to humans]" as originally proposed by Szegedy et al. [10] by limiting pixel-space perturbations through an ε-bounded constraint is fundamentally flawed. This is because the human visual system is not solely reliant on pixel-space information to interpret images [14].

Humans can detect forged low-ε adversarial examples with high accuracy in both the visual (85.4%) [15] and textual ($\geq$70%) [16] domain. It's worth mentioning that invertible neural networks can partially mitigate this issue in the visual domain [17].

Additionally, small ε-bounded adversarial perturbations are found to cause misclassification in time-constrained humans [18].

While initially discovered in computer vision applications, the attack can be crafted for any domain or data type, even graphs [19]. Natural language processing models can be attacked by circumventing the discrete nature of text data [20, 21, 22]. Speech recognition systems are vulnerable to audio-based attacks, where crafted noise can cause system failure [23]. Deep reinforcement learning applications, including pathfinding and robot control, have also shown susceptibility to adversarial manipulations that can compromise their decision-making capabilities [24].

1.1.3 Semantics Preserving Adversarial Examples

Imperceptible noise-based adversarial examples are just one type of semantics-preserving adversarial examples. Other examples include rotating an image by a few degrees or capturing it from a different angle, which can also cause misclassification. These broader categories of adversarial examples are often referred to as "unrestricted" [25, 26] or "semantics-preserving" [27, 28, 29]. The comparison in Fig. 1.1 and the illustration in Fig. 1.2 highlight the differences between various kinds of adversarial examples. Fig. 1.3 shows a collection of naturally occurring adversarial examples, also known as "natural adversarial examples" [30, 31].

This shift in defining adversarial examples, popularized by the "Unrestricted Adversarial Examples Challenge" [26] by Google in 2018, has led to a more nuanced understanding of the phenomenon. It acknowledges that real-world applications, especially in safety-critical contexts, are subject to a broader range of adversarial attacks than previously assumed and do not always adhere to the "small perturbation" constraint initially proposed [26].

This paradigm shift towards seeking more meaningful adversarial examples and "spatial robustness" was first proposed by Gilmer et al. in 2018 [32] and further explored by Engstrom et al. in 2019 [33]. These works lay the theoretical foundation for our research, and we believe this approach to to be the most promising for future research in adversarial machine learning.

The challenge of defining semantics is central to this discussion. Without perfect representations that align with human judgment functions, we must rely on the best available encoders or semantics preservation metrics [33, 16] as proxies. This pragmatic approach acknowledges the limitations of current technology while striving for more meaningful adversarial examples.

1.2 Motivation

One particularly entertaining example of an adversarial attack is the subversion of the conference paper-reviewer assignment model by Eisenhofer et al. [34], where authors preselect reviewers to gain a competitive advantage.

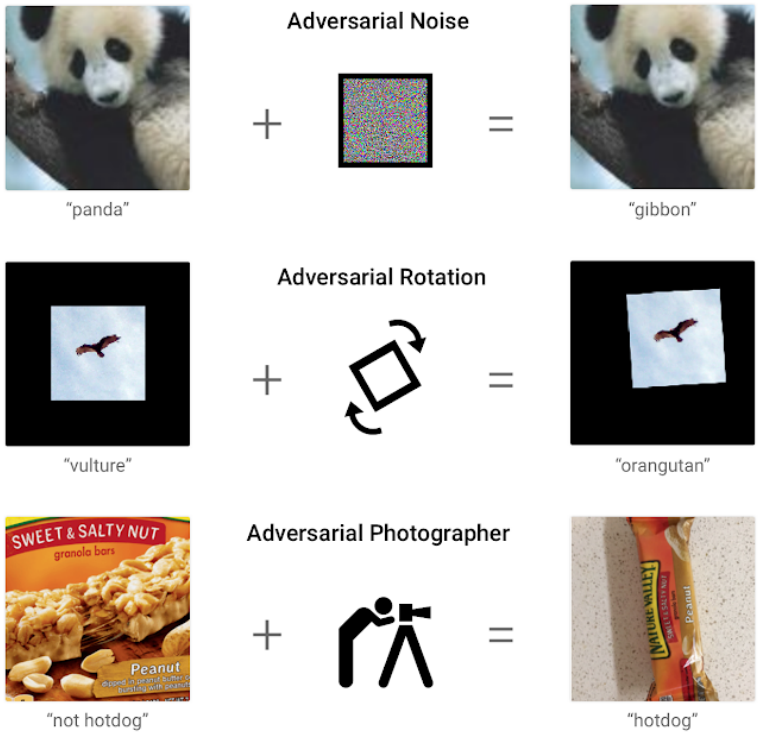

Figure 1.1: Unrestricted adversarial examples [26].

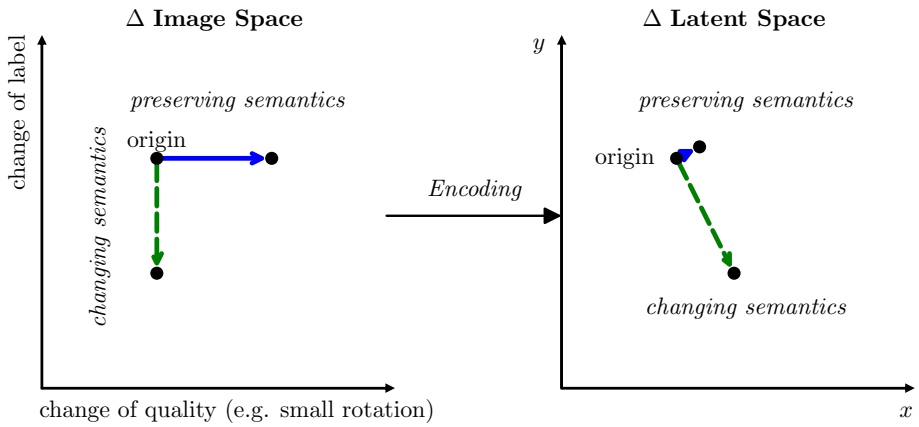

Figure 1.2: Semantics preserving/changing perturbations in pixel/latent-space (assuming full accuracy).

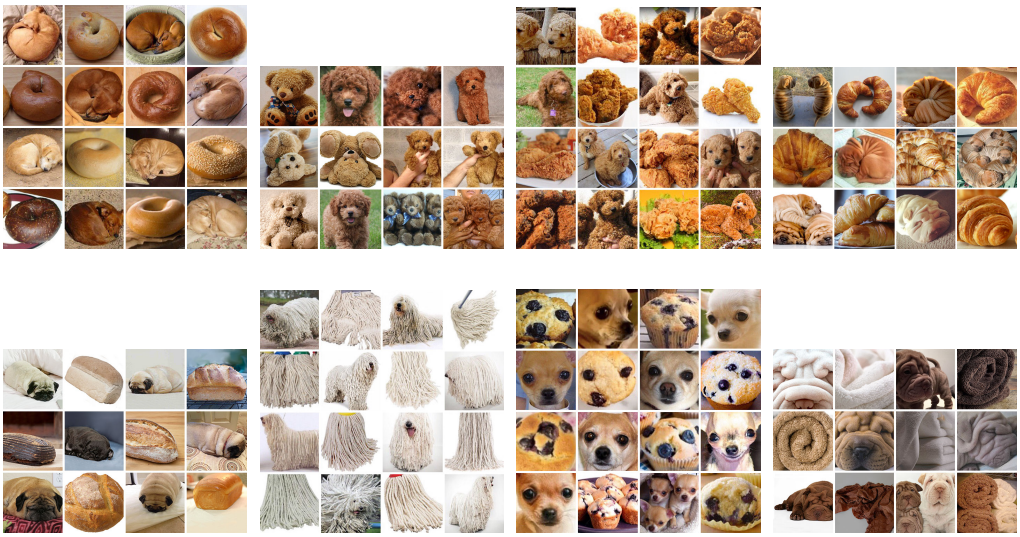

Figure 1.3: Natural adversarial examples [31].

But adversarial attacks are not limited to academia. Machine learning security has become particularly critical as models are deployed in increasingly sensitive and safety-critical applications [35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45].

National infrastructure and cyber-physical systems are commonly use machine learning-based protection systems, which can be compromised [46, 47, 48, 49, 50]. A single failure in a nuclear power plant or a water treatment facility or any other critical infrastructure can have catastrophic consequences.

Adversarial attacks can also compromise credit card fraud detection systems, potentially leading to billions in annual losses [51, 52, 53].

In healthcare, malicious actors can manipulate medical imaging systems used for diagnosing conditions like skin cancer and Alzheimer's disease, potentially leading to incorrect diagnoses and treatment decisions [54, 55]. The vulnerability extends to pandemic management systems, where deep learning algorithms processing data from medical IoT devices, including CT scans and thermal cameras, can be compromised through perturbations [56].

In autonomous vehicles roadside billboards can be weaponized to display specially crafted videos that manipulate vehicle controllers into performing dangerous maneuvers, such as unexpected lane changes or road departures [57]. Even more concerning, acoustic attacks can exploit image stabilization hardware to create blurred images that cause object detection systems to misclassify or completely miss critical obstacles [58]. These vulnerabilities are increasing in sophistication [59].

This risk is also demonstrated in cybersecurity applications, where phishing website detectors face degradation of 3-10% from realistic evasion attempts that are

both cheap and practical to implement [60]. In the domain of malware detection mutation systems that combine generative networks with reinforcement learning to create metamorphic malware capable of evading detection systems [61].

This has lead to major companies investing heavily in adversarial machine learning research and security.

Microsoft has taken a leading position, spending over $20 billion on cybersecurity initiatives, with a significant portion dedicated to machine learning security research and their specialized ML red team operations [62].

Open Philanthropy has provided $330,000 and $343,235 [63] in funding to Carnegie Mellon University dedicated to AdvX research.

The MITRE corporation is now cooperating with Microsoft, Bosch, IBM, NVIDIA, Airbus, Deep Instinct and PricewaterhouseCoopers to develop the Adversarial Machine Learning Threat Matrix [64] for threat modeling and risk assessment.

The Defense Advanced Research Projects Agency (DARPA) has granted nearly $1 million to the CV AdvX team at UC Riverside [65]. Booz Allen Hamilton, the largest provider of machine learning services for the Federal government, invested in HiddenLayer, Robust Intelligence [66, 67] Shift5, Credo, Hidden Level, Latent, Synthetaic, and Reveal Technology [68, 69], all of which are dedicated to machine learning security and robustness research.

These investments demonstrate the growing recognition of the importance of protecting machine learning systems from adversarial attacks.

1.3 Threat Modeling

Having established the general concept of adversarial examples, we can now explore the various ways they can be categorized. Our system is not exhaustive: The field continues to evolve, with new attack vectors emerging regularly [70]. This is particularly important in threat modeling, where the goal is to anticipate and defend against potential attacks.

We can differentiate between white-box and black-box attacks. White-box attacks assume complete knowledge of and access to the target model, while black-box attacks operate with limited or no access to the model's internal workings [71]. Interestingly, research has shown that in some cases, black-box attacks can be more effective than white-box approaches at compromising model security [71].

An attack can be targeted or untargeted. Targeted attacks aim to manipulate the model into producing a specific, predetermined output, whereas untargeted attacks simply seek to cause any misclassification or erroneous output [71, 19]. This distinction is particularly relevant in security-critical applications, where the

attacker's goals may vary from causing general disruption to achieving specific malicious outcomes.

The method used to generate adversarial examples can be gradient-based, optimization-based or search-based strategies. For example, some text-based attacks leverage language models to generate alternatives for masked tokens, ensuring grammatical correctness and semantic coherence [72].

The extent to which adversarial examples are transferable – meaning their ability to fool multiple different models, or the human vision system[18] – is another way to differentiate them. Some adversarial examples demonstrate high transferability across various model architectures, while others are more model-specific in their effectiveness [73, 74]. Recent research has shown that adversarial examples are more readily transferable between vanilla neural networks than between defended ones [75, 76].

Finally, attacks can either focus on preserving the semantic meaning of inputs or exploit the mathematical properties of models without regard for semantic interpretation [27].

1.4 Latent Representations

The internal latent representations of neural networks, their alignment with human understanding and the resuling gap between the two (the human-machine vision gap [77]) is a central theme in adversarial machine learning research. This gap has many practical implications for the robustness and interpretability of machine learning models.

Neural networks trained with topological features develop substantially different internal representations compared to those trained on raw data, though these differences can sometimes be reconciled through simple affine transformations [78]. This finding suggests that while the structural representations may differ, the underlying semantic understanding might be preserved across different training approaches.

The Centered Kernel Alignment (CKA) metric enables us to compare neural network representations, though it comes with important caveats. In biological and artificial neural networks, CKA can show artificially high similarity scores in low-data, high-dimensionality scenarios, even with random matrices [79]. This limitation is particularly relevant when comparing representations of different sizes or when analyzing specific regions of interest.

The relationship between network architecture and concept representation has also been explored. Generally higher-level concepts are typically better represented in the final layers of neural networks, while lower-level concepts are often better captured in middle layers [80, 81]. This hierarchical organization mir-

rors our understanding of human cognitive processing and suggests that neural networks naturally develop structured representations that align with human conceptual understanding.

The choice of objective function significantly influences how networks represent information, particularly when dealing with biased data. Networks trained with Negative Log Likelihood and Softmax Cross-Entropy loss functions demonstrate comparable capabilities in developing robust representations [82].

Recent research [83] has demonstrated that neural networks with strong performance tend to learn similar internal representations, regardless of their training methodology. Networks trained through different approaches, such as supervised or self-supervised learning, can be effectively "stitched" together without significant performance degradation. This suggests a convergence in how successful neural networks represent information.

This aligns with the "Platonic Representation Hypothesis" [84], which suggests that neural networks are converging toward a shared statistical model of reality, regardless of their training objectives or architectures. As models become larger and are trained on more diverse tasks and data, their internal representations increasingly align with each other, even across different modalities like vision and language. This convergence appears to be driven by the fundamental constraints[2] of modeling the underlying structure of the real world, similar to Plato's concept of an ideal reality that exists beyond our sensory perceptions. The hypothesis proposes that this convergence is not coincidental but rather a natural consequence of different models attempting to capture the same underlying statistical patterns and relationships that exist in reality [84].

Should the "Platonic Representation Hypothesis" hold true, this would either mean that (a) adversarial examples as we know them are misalignments from a converged model of reality, or (b) that there exist a universal adversarial example that can fool any model, regardless of its architecture, training data or objective function, converging to a single and shared model of reality.

Recent work by Moosavi-Dezfooli et al. [85] have demonstrated the existence of a single perturbation that can fool most models for all naturally occurring images, adding weight to the latter interpretation, though the question remains open.

1.5 Mental Models

The question discussed in the previous section is just one of many that remain open and yet have to be fully explained [86]. Among them are:

- What are adversarial examples?

[2] Formally: "If an optimal representation exists in function space, larger hypothesis spaces are more likely to cover it."

- Why are the adversarial examples so close to the original images?
- Why don't the adversarial perturbations resemble the target class?
- Why do robustness and accuracy trade-off [87]?
- Why do adversarial examples transfer between models, even on disjoint training sets [10]?
- Why do adversarial examples transfer between models [10]?
- Why do adversarial examples transfer between models and time-limited humans [18]?

Initially, when Szegedy et al. [10] coined the term they proposed that adversarial examples are caused by (1) neural networks developing internal representations that become increasingly disconnected from the input features as they progress through deeper layers and (2) that these networks fail to maintain the smoothness properties typically assumed in traditional machine learning approaches. The idea was that this lack of smoothness gives them their expressive power, but also makes them vulnerable to these attacks.

Definition: Manifold

The first attempt to explain adversarial examples by Szegedy et al. [10] used the term "manifold", while refering to a data submanifold.

A manifold can be thought of as a low-dimensional structure embedded in a high-dimensional space, representing the set of valid data points (e.g., natural images) that the neural network is trained to classify. Mathematically, if the input data lies on a manifold $\mathcal{M} \subset \mathbb{R}^m$, then $\mathcal{M}$ represents the subset of the high-dimensional input space $\mathbb{R}^m$ that corresponds to meaningful or real-world data.

Szegedy et al. suggest that adversarial examples exploit the structure of this manifold and its surrounding space. Specifically, adversarial examples are small perturbations r added to an input $x \in \mathcal{M}$, such that the perturbed input $x' = x + r$ lies off the data manifold but still within the high-dimensional input space.

Formally, given a classifier $f : \mathbb{R}^m \to \{1, ..., k\}$ and its associated loss function $\text{Loss}_f(x, y)$, an adversarial example x' for an input x with true label y can be found by solving:

$$\min_r \|r\|_2 \quad \text{subject to } f(x + r) \neq y,\ x + r \in [0, 1]^m$$

where r is constrained to be small (e.g., in terms of its L_2-norm). This

optimization problem effectively traverses the space near x, moving off the manifold $\mathcal{M}$, to find regions where the classifier's decision boundary behaves unexpectedly.

The paper suggests that these adversarial examples expose "blind spots" in the learned representation of the manifold by the neural network. The network's decision boundary may extend into regions near $\mathcal{M}$ in ways that are not semantically meaningful, allowing adversarial perturbations to exploit these regions. This phenomenon arises due to the high dimensionality of the input space and the discontinuous mappings learned by deep networks, which can fail to generalize smoothly beyond the manifold [88, 89, 90, 91, 86].

A more rigorous definition of the manifold hypothesis is provided by Khoury et al. [88].

Definition: Realism

A "realistic subspace" can be understood as a subset of the data manifold where the images appear plausible according to human perception or a given distribution P. A simple formula that expresses this idea elegantly to quantify realism is derived from the notion of randomness deficiency in algorithmic information theory [92]:

$$U(x) = -\log P(x) - K(x)$$

where $P(x)$ is the probability density of the image x under the target distribution and $K(x)$ is the Kolmogorov complexity of x, representing the shortest description of x in a universal programming language. This measure, called a "universal critic", captures how well x aligns with both the statistical properties of P and its compressibility. A low value of $U(x)$ indicates that x is realistic, while a high value suggests it is unrealistic [92].

This approach generalizes prior methods by integrating both probabilistic and structural aspects of realism. It highlights that realism depends not only on adherence to statistical patterns (e.g., probabilities or divergences) but also on whether an image can be plausibly generated within the constraints of P. While directly computing $K(x)$ is infeasible due to its uncomputability, practical approximations (e.g., compression algorithms or neural network-based critics) can serve as proxies [92].

The distinction between realistic and unrealistic perturbations is cru-

> cial for practical applications, as some adversarial examples may be mathematically valid but physically impossible to realize in real-world scenarios [93].
>
> The challenge of quantifying realism remains a fundamental problem in machine learning [92].

Since then there have been many attempts at finding a cohesive narrative to explain these counter-intuitive properties, each with their own limitations and assumptions – some complementary, some contradictory [94].

Non-robust features & concentration of measure in high-dimensions. Most popularly, Ilyas et al. [94] proposed that features that models learn from can be divided in 3 categories: (1) useless featurs, to be discarded by the feature extractor, (2) robust features, which are comprehensible to humans, generalize across multiple datasets and remain stable under small adversarial perturbations and (3) non-robust features, which are incomprehensible to humans, learned by the supervised model to exploit patterns in the data distribution which are highly effective for the task at hand but also brittle and easily manipulated by adversarial perturbations. The authors suggest that the vulnerability of deep neural networks to adversarial examples is due to their reliance on non-robust features and inherent to how the models are optimized to minimize the loss function. In essence, the authors argue that adversarial vulnerability is a property of the dataset, not the algorithm and by removing these non-robust features from the training data although the adversarial robustness of the model can be improved, due to information loss of the most predictive features, the model's overall accuracy will decrease. This view is also shared among [95, 96, 97, 98, 99, 100, 101, 102, 103, 104].

Theoretical constructions which incidentally exploit non-robust features. A complimenting hypothesis is that because models trained to maximize accuracy will naturally utilize non-robust data, regardless of whether it aligns with human perception [94] they add a low-magnitude weight to sensitive variables that can get overamplified by adversarial examples [105, 106]. The assumption is that this happens due to computational constraints or model complexity.

Insufficient data. Schmidt et al. argue [107] that adversarial vulnerabilities are intrinsic to statistical learning in high-dimensional spaces and not merely due to flaws in specific algorithms or architectures. This is a natural consequence of the mental model proposed by Ilyas et al. [94]. They also argue that due to

information loss in a robust dataset, significantly more data is required during training in order to achieve comparable performance.

Boundary Tilting. A competing view by Tanay and Kim et al. [108, 109] suggests that adversarial examples exist because decision boundaries extend beyond the actual data manifold and can lie uncomfortably close to it, essentially viewing adversarial examples as a consequence of overfitting. This observation can be quantified through the concept of adversarial strength, which relates to the angular deviation between the classifier and the nearest centroid classifier. The authors also argue that this vulnerability can be addressed through proper regularization techniques.

Test Error in Noise. There might be a link between robustness to random noise and adversarial attacks [110, 111, 99, 112]. This might imply that adversarial examples exploit inherent weaknesses in how models generalize under noisy or perturbed conditions.

Local Linearity. Goodfellow, Shlens and Szegedy et al. [11, 104] argue that even though DNNs are highly nonlinear overall, their behavior in high-dimensional spaces often resembles that of linear models. This makes the models vulnerable to small, targeted perturbations similar to how they're computed by FGSM. However some adversarial examples are successful all while defying the assumption of local linearity and reducing a model's linearity does not necessarily improve its robustness either [113].

Piecewise-linear decision boundaries. In the "dimpled manifold hypothesis" [86] the central claim is that adversarial examples emerge because we attempt to fit high $n-1$ dimensional decision boundaries to inherently low-dimensional data like images (which can be losslessly projected to $k \ll n$ dimensions). This leaves redundant dimensions on which adversarial examples won't be judged, which enables them exist roughly perpendicularly from the true location of the low-dimensional natural image, by using large gradients. In this mental model adversarial examples can be on-manifold or off-manifold, based on the angle of the gradients relative to the data manifold.

The authors also suggest that decision boundaries of neural networks evolve during training. This happens through two distinct phases. First, there is a rapid "clinging" phase where the decision boundary moves close to the low-dimensional image manifold containing the training examples. This is followed by a slower "dimpling" phase that creates shallow bulges in the decision boundary, pushing it to the correct side of the training examples, without shifting the plane. This

gradient descent based process is highly efficient, but it also leaves a brittle decision boundary that can be easily exploited.

This implies that any attempt to robustify a network by limiting all its directional derivatives will make it harder to train and thus less accurate.

It also explains why networks trained on incorrectly labeled adversarial examples can still perform well on regular test images, as the main effect of adversarial training is simply to deepen these dimples in the decision boundary.

Lukas Karner successfully was able to successfully reproduce the experiments from the "Dimpled Manifold Hypothesis" paper in 2023 [114]. He additionally demonstrated that dimensionality reductions increases the interpretability of the perturbations to humans [114].

However, despite the experiments being carried out correctly themselves, the chain of reasoning might be flawed, as shown by a succinct (<100 LoC) counterexample by Yannik Kilcher in 2021 [115, 116]. While the "Dimpled Manifold Hypothesis" implies a relatively uniform vulnerability across all dimensions the counterexperiment contradicts these assumptions through successful adversarial attacks constructed by perturbing either an arbitrary subset of selected dimensions or their complement. If the decision boundary truly "clung" to the data manifold, restricting perturbations to a subset of dimensions would not have produced successful adversarial examples. The ability to generate adversarial examples in complementary subspaces suggests the decision boundary structure is more complex than just simple dimples.

To summarize, there is no consensus on the root cause of adversarial examples and the field remains an active area of research. The mental models proposed by different researchers are not necessarily mutually exclusive and it is likely that the true explanation involves a combination of these factors.

1.6 Defenses

Having discussed the various theories and approaches in explaining adversarial examples, we can now turn our attention to the countermeasures that have been proposed to mitigate their impact.

A leaderboard of adversarial robustness can be found on the RobustBench platform [117], which provides a standardized evaluation of adversarial robustness across a wide range of models and datasets. The platform includes a variety of metrics for evaluating robustness, such as the ℓ_∞ and ℓ_2 adversarial perturbation sizes, as well as the robust accuracy under different attack settings.

Several effective strategies have been developed. This collection is by no means exhaustive.

Adversarial Training. One of the least invasive methods to improve adversarial robustness is adversarial training. Incorporating adversarial examples into the training process improves model resilience by learning from potential attack patterns and helps maintain performance on clean data [118, 119]. However, this requires the anticipated attacks to be known in advance. An alternative would be introducing derived variables for controlled randomness to input data during training, which is still effective [120].

Quality Assessment Integration. Implementing image quality assessment combined with knowledge distillation helps detect potentially harmful inputs that could cause incorrect model predictions [121]. Another alternative preprocessing technique is using brain inspired encoders [122]. This method is particularly effective as it doesn't require model retraining, but depending on the preprocessing techique used, it can be computationally expensive.

Moving Target Defense. Using heterogeneous models, diversifying the model structure, using ensembles and dynamic model switching can protect against white-box adversarial attacks. This approach will make attack vectors that work on one model ineffective on others [123].

Statistical Detection. Statistical tests can be employed for some signal based deep learning systems to detect adversarial examples. This includes analyzing peak-to-average-power ratio and examining softmax outputs of the model [124].

Enhanced Transformation. Transformation-based defense strategies, such as using generative adversarial networks (GANs), can help recover from adversarial examples. These methods can counteract adversarial effects while maintaining or even improving classification performance [125].

The countermeasures discussed so far provide a diverse array of techniques to mitigate the impact of adversarial examples. Each method addresses specific aspects of the problem, ranging from input preprocessing to model architecture adjustments and training methodologies. Notably, hybrid strategies that combine multiple techniques often yield the best results, with some implementations achieving reliable performance even under sophisticated attack benchmarks [126].

Assuming that robustness and generalizability are not competing objectives but complementary goals, the ultimate defense lies in designing architectures that

inherently integrate robustness and interpretability. By prioritizing these objectives at the core of model development, we can create systems that not only withstand adversarial attacks but also offer more trustworthy and transparent decision-making.

Fermi-Bose Machine. One noteworthy example is the Fermi-Bose Machine [127]. Unlike traditional neural networks that rely on backpropagation, this method introduces a local contrastive learning mechanism inspired by quantum mechanics principles. The system works by making representations of inputs with identical labels cluster together (like bosons), while representations of different labels repel each other (like fermions). This layer-wise learning approach is considered more biologically plausible than traditional backpropagation [127]. The researchers demonstrated the effectiveness of their method on the MNIST dataset, showing that by adjusting the target fermion-pair-distance parameter, they could significantly reduce the susceptibility to adversarial attacks that typically disturb standard perceptrons [127]. The key innovation lies in controlling the geometric separation of prototype manifolds through the target distance parameter, as revealed by statistical mechanics analysis [127].

Ensemble everything everywhere. A recent (August 2024) state-of-the-art approach works by multi-resolution input representations and dynamic self-ensembling of intermediate layer predictions [128]. The researchers introduced a robust aggregation mechanism called CrossMax, based on Vickrey auction, which combines predictions from different layers of the network [128].

The method achieved impressive results without requiring adversarial training or additional data, reaching approximately 72% adversarial accuracy on CIFAR-10 and 48% on CIFAR-100 using the RobustBench AutoAttack suite [128]. When combined with simple adversarial training, the performance improved further to 78% on CIFAR-10 and 51% on CIFAR-100, surpassing the current state-of-the-art by 5% and 9% respectively [128].

An interesting secondary outcome of this research was the discovery that gradient-based attacks against their model produced human-interpretable images of target classes [128]. Additionally, the multi-resolution approach enabled the researchers to transform pre-trained classifiers and CLIP models into controllable image generators, while also developing successful transferable attacks on large vision language models [128].

Chapter 2

Research Questions

Perhaps we should be rethinking adversarial examples not as attacks but as indicators of insufficient generalization.

The most promising path forward may not lie in defending against these examples, but rather in fundamentally reimagining model architectures with interpretability and robustness as core design principles.

This perspective suggests that the key to improving adversarial robustness lies not in patching existing systems, but in developing new architectures from the ground up that naturally exhibit both robust and interpretable behavior.

The self-ensembling algorithm [128] is a step in this direction we explore in this thesis. Additionally, we make use of primitive geometric masks as attacks, to study the underlying mechanisms of adversarial examples.

Chapter 3

Methodology

Chapter 4

Results

Chapter 5

Conclusion

"The stuff is what the stuff is, brother. Okay. We don't ask questions about the weights. We just wake up, we go to work, we use the weights, we go back home. Okay. If we change the weights, the predictions would be different and less good, probably... depending on the weather... so we don't ask about the weights."

— James Mickens, USENIX'18 [129]

Bibliography

[1] Y. Jabary, A. Plesner, T. Kuzhagaliyev, and R. Wattenhofer, "Seeing through the mask: Rethinking adversarial examples for captchas," *arXiv preprint arXiv:2409.05558*, 2024.

[2] R. Seidel, "The nature and meaning of perturbations in geometric computing," *Discrete & Computational Geometry*, vol. 19, pp. 1–17, 1998.

[3] M. De Berg, *Computational geometry: algorithms and applications*. Springer Science & Business Media, 2000.

[4] W. R. Franklin and S. V. G. de Magalhães, "Implementing simulation of simplicity for geometric degeneracies," *arXiv preprint arXiv:2212.08226*, 2022.

[5] Edelsbrunner, Letscher, and Zomorodian, "Topological persistence and simplification," *Discrete & computational geometry*, vol. 28, pp. 511–533, 2002.

[6] H. Edelsbrunner and D. Guoy, "Sink-insertion for mesh improvement," in *Proceedings of the seventeenth annual symposium on Computational geometry*, 2001, pp. 115–123.

[7] H. Edelsbrunner and E. P. Mücke, "Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms," *ACM Transactions on Graphics (tog)*, vol. 9, no. 1, pp. 66–104, 1990.

[8] B. Lévy, "Robustness and efficiency of geometric programs: The predicate construction kit (pck)," *Computer-Aided Design*, vol. 72, pp. 3–12, 2016.

[9] P. Schorn, "An axiomatic approach to robust geometric programs," *Journal of symbolic computation*, vol. 16, no. 2, pp. 155–165, 1993.

[10] C. Szegedy, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[12] J. Zhang and C. Li, "Adversarial examples: Opportunities and challenges," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2578–2593, 2019.

[13] E. D. Cubuk, B. Zoph, S. S. Schoenholz, and Q. V. Le, "Intriguing properties of adversarial examples," *arXiv preprint arXiv:1711.02846*, 2017.

[14] P. Ning, W. Jiang, and R. Wang, "Hflic: Human friendly perceptual learned image compression with reinforced transform," in *2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI)*. IEEE, 2023, pp. 188–194.

[15] V. Veerabadran, J. Goldman, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, J. Shlens, J. Sohl-Dickstein, M. C. Mozer *et al.*, "Subtle adversarial image manipulations influence both human and machine perception," *Nature Communications*, vol. 14, no. 1, p. 4933, 2023.

[16] D. Herel, H. Cisneros, and T. Mikolov, "Preserving semantics in textual adversarial attacks," in *ECAI 2023*. IOS Press, 2023, pp. 1036–1043.

[17] Z. Chen, Z. Wang, J.-J. Huang, W. Zhao, X. Liu, and D. Guan, "Imperceptible adversarial attack via invertible neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 414–424.

[18] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial examples that fool both computer vision and time-limited humans," *Advances in neural information processing systems*, vol. 31, 2018.

[19] S. Kashyap, A. Sharma, S. Gautam, R. Sharma, S. Chauhan, and Simran, "Adversarial attacks and defenses in deep learning," *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, pp. 318–323, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272716335

[20] X. Han, Y. Zhang, W. Wang, and B. Wang, "Text adversarial attacks and defenses: Issues, taxonomy, and perspectives," *Security and Communication Networks*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248369346

[21] Z. Meng and R. Wattenhofer, "A geometry-inspired attack for generating natural language adversarial examples," *arXiv preprint arXiv:2010.01345*, 2020.

[22] Z. Yang, Z. Meng, X. Zheng, and R. Wattenhofer, "Assessing adversarial robustness of large language models: An empirical study," *arXiv preprint arXiv:2405.02764*, 2024.

[23] K. Rajaratnam and J. Kalita, "Noise flooding for detecting audio adversarial examples against automatic speech recognition," in *2018 IEEE In-*

ternational Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 2018, pp. 197–201.

[24] X. Bai, W. Niu, J. Liu, X. Gao, Y. Xiang, and J. Liu, "Adversarial examples construction towards white-box q table variation in dqn pathfinding training," *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pp. 781–787, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49895854

[25] D. Fazlija, A. Orlov, J. Schrader, M.-M. Zühlke, M. Rohs, and D. Kudenko, "How real is real? a human evaluation framework for unrestricted adversarial examples," *arXiv preprint arXiv:2404.12653*, 2024.

[26] T. B. Brown and C. Olsson. (2018, 9) Introducing the unrestricted adversarial examples challenge. Google Brain Team.

[27] K. Browne and B. Swift, "Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks," *arXiv preprint arXiv:2012.10076*, 2020.

[28] M. Careil, M. J. Muckley, J. Verbeek, and S. Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *The Twelfth International Conference on Learning Representations*, 2023.

[29] W. Lee, H. Lee, and S.-g. Lee, "Semantics-preserving adversarial training," *arXiv preprint arXiv:2009.10978*, 2020.

[30] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 262–15 271.

[31] K. Zack. Archive of deleted tweet by @teenybiscuit. [Online]. Available: https://imgur.com/a/deep-learning-training-set-K4RWn

[32] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl, "Motivating the rules of the game for adversarial example research," *arXiv preprint arXiv:1807.06732*, 2018.

[33] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International conference on machine learning*. PMLR, 2019, pp. 1802–1811.

[34] T. Eisenhofer, E. Quiring, J. Möller, D. Riepel, T. Holz, and K. Rieck, "No more reviewer# 2: Subverting automatic {Paper-Reviewer} assignment using adversarial learning," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5109–5126.

[35] S. Chahar, S. Gupta, I. Dhingra, and K. S. Kaswan, "Adversarial threats in machine learning: A critical analysis," in *2024 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, vol. 1, 2024, pp. 253–258.

[36] H. Çifci, "Analysis of turkey's cybersecurity strategies: Historical developments, scope, content and objectives," *Sakarya University Journal of Science*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:268035521

[37] K. Sadeghi, A. Banerjee, and S. K. S. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 450–467, 2020.

[38] A. Khadka, S. Sthapit, G. Epiphaniou, and C. Maple, "Resilient machine learning in space systems: Pose estimation as a case study," *2022 IEEE Aerospace Conference (AERO)*, pp. 1–9, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251472990

[39] I. Yilmaz, K. Kapoor, A. Siraj, and M. Abouyoussef, "Privacy protection of grid users data with blockchain and adversarial machine learning," in *proceedings of the 2021 ACM workshop on secure and trustworthy cyber-physical systems*, 2021, pp. 33–38.

[40] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, ""real attackers don't compute gradients": bridging the gap between adversarial ml research and practice," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2023, pp. 339–364.

[41] R. S. S. Kumar, J. Penney, B. Schneier, and K. Albert, "Legal risks of adversarial machine learning research," *arXiv preprint arXiv:2006.16179*, 2020.

[42] R. Cao and R. K.-W. Lee, "Hategan: Adversarial generative-based data augmentation for hate speech detection," in *International Conference on Computational Linguistics*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227230383

[43] D. B. Nurseitov, K. A. Bostanbekov, G. Abdimanap, A. Abdallah, A. N. Alimova, and D. Kurmangaliyev, "Application of machine learning methods to detect and classify core images using gan and texture recognition," *ArXiv*, vol. abs/2204.14224, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:265674547

[44] M. Zolotukhin, D. Zhang, P. Miraghaie, T. Hämäläinen, W. Ke, and M. Dunderfelt, "Attacks against machine learning models in 5g networks," *2022 6th European Conference on Electrical Engineering & Computer Science (ELECS)*, pp. 106–114, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:259102662

[45] H. Face. (2024) Hugging face partners with wiz research to improve ai security. Hugging Face. Blog post discussing security improvements, pickle file security concerns, and partnership with Wiz Research. [Online]. Available: https://huggingface.co/blog/hugging-face-wiz-security-blog

[46] N. Moradpoor, L. A. Maglaras, E. Abah, and A. Robles-Durazno, "The threat of adversarial attacks against machine learning-based anomaly detection approach in a clean water treatment system," *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pp. 453–460, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:262980912

[47] V. Chevardin, O. Yurchenko, O. V. Zaluzhnyi, and Y. Peleshok, "Analysis of adversarial attacks on the machine learning models of cyberprotection systems." *Communication, informatization and cybersecurity systems and technologies*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266487123

[48] D. A. Ulybyshev, I. Yilmaz, B. Northern, V. Kholodilo, and M. Rogers, "Trustworthy data analysis and sensor data protection in cyber-physical systems," *Proceedings of the 2021 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233384629

[49] B. Halak, C. Hall, S. Fathir, N. Kit, R. Raymonde, M. Gimson, A. Kida, and H. Vincent, "Towards autonomous physical security defenses using machine learning," *IEEE Access*, vol. PP, pp. 1–1, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248849785

[50] Rudolph, R. V. Matalucci, and J. T. Matalucci, "Developing protective strategies forcriticalbuilding infrastructures potentially subjected to m alevolentthreats* by," 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:111438592

[51] K. Gu, "Deep learning techniques in financial fraud detection," *Proceedings of the 7th International Conference on Cyber Security and Information Engineering*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253120915

[52] A. Agarwal and N. K. Ratha, "Black-box adversarial entry in finance through credit card fraud detection," in *CIKM Workshops*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245540840

[53] M.-Y. Tsai, H.-H. Cho, C.-M. Yu, Y.-C. Chang, and H.-C. Chao, "Effective adversarial examples identification of credit card transactions," *IEEE Intelligent Systems*, vol. 39, pp. 50–59, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268628851

[54] V. Jogani, J. Purohit, I. Shivhare, and S. C. Shrawne, "Analysis of explainable artificial intelligence methods on medical image classification," *arXiv preprint arXiv:2212.10565*, 2022.

[55] M. H. Najafi, M. Morsali, M. Vahediahmar, and S. B. Shouraki, "Dft-based adversarial attack detection in mri brain imaging: Enhancing diagnostic accuracy in alzheimer's case studies," *arXiv preprint arXiv:2408.08489*, 2024.

[56] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial examples—security threats to covid-19 deep learning systems in medical iot devices," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9603–9610, 2021.

[57] N. Patel, P. Krishnamurthy, S. Garg, and F. Khorrami, "Adaptive adversarial videos on roadside billboards: Dynamically modifying trajectories of autonomous vehicles," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5916–5921, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:210971572

[58] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision," *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 160–175, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235601506

[59] C. W. Axelrod, "Cybersecurity challenges of systems-of-systems for fully-autonomous road vehicles," *2017 13th International Conference and Expo on Emerging Technologies for a Smarter World (CEWIT)*, pp. 1–6, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:29935654

[60] Y. Yuan, G. Apruzzese, and M. Conti, "Multi-spacephish: Extending the evasion-space of adversarial attacks against phishing website detectors using machine learning," *Digital Threats: Research and Practice*, vol. 5, pp. 1 – 51, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266363431

[61] T.-N. To, D. L. Kim, D. T. T. Hien, N. H. Khoa, H. D. Hoang, P. T. Duy, and V.-H. Pham, "On the effectiveness of adversarial samples against ensemble learning-based windows pe malware detectors," *arXiv preprint arXiv:2309.13841*, 2023.

[62] Coursera Editorial Team. (2024) What is adversarial machine learning? [Online]. Available: https://www.coursera.org/articles/adversarial-machine-learning

[63] Open Philanthropy, "Carnegie mellon university — research on adversarial examples," 2024, grant of $343,235 to support research on adversarial examples led by Professor Aditi Raghunathan.

[64] B. Eidson, "Mitre, microsoft, and 11 other organizations take on machine-learning threats," *MITRE News and Insights*, 2024, impact Story on the Adversarial Machine Learning Threat Matrix initiative. [Online]. Available: https://www.mitre.org/news-insights/impact-story/mitre-microsoft-and-11-other-organizations-take-machine-learning-threats

[65] A. Roy-Chowdhury, S. Krishnamurthy, C. Song, and S. Asif. (2020, 7) ECE and CSE faculty receive new DARPA grant on adversarial machine learning. University of California, Riverside. DARPA Machine Vision Disruption program grant announcement.

[66] Robust Intelligence. (2024) Ai application security. Robust Intelligence. [Online]. Available: https://www.robustintelligence.com/ai-application-security

[67] K. Cai, "Robust intelligence raises $14 million series a led by sequoia to build platform for testing machine learning applications," *Forbes*, October 2020.

[68] Booz Allen Hamilton. (2023, 9) Booz allen doubles down on adversarial ai capabilities with new investment. Business Wire. Press Release.

[69] MSSPAlert. (2023, September) Booz allen hamilton expands adversarial ai capabilities.

[70] Y. L. Khaleel, M. A. Habeeb, and H. Alnabulsi, "Adversarial attacks in machine learning: Key insights and defense approaches," *Applied Data Science and Analysis*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272000855

[71] G. Capozzi, D. C. D'Elia, G. A. Di Luna, and L. Querzoni, "Adversarial attacks against binary similarity systems," *IEEE Access*, 2024.

[72] S. Garg and G. Ramakrishnan, "Bae: Bert-based adversarial examples for text classification," *arXiv preprint arXiv:2004.01970*, 2020.

[73] Y. Li, Y. Guo, Y. Xie, and Q. Wang, "A survey of defense methods against adversarial examples," *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, pp. 453–460, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252165991

[74] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. Lee, "A review of adversarial attack and defense for classification methods," *The American Statistician*, vol. 76, no. 4, pp. 329–345, 2022.

[75] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, "Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3866–3876.

[76] S. Zheng, C. Zhang, and X. Hao, "Black-box targeted adversarial attack on segment anything (sam)," *arXiv preprint arXiv:2310.10010*, 2023.

[77] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel, "Partial success in closing the gap between human and machine vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 885–23 899, 2021.

[78] S. McGuire, S. Jackson, T. Emerson, and H. Kvinge, "Do neural networks trained with topological features learn different internal representations?" in *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*. PMLR, 2023, pp. 122–136.

[79] A. Murphy, J. Zylberberg, and A. Fyshe, "Correcting biased centered kernel alignment measures in biological and artificial neural networks," *arXiv preprint arXiv:2405.01012*, 2024.

[80] A. Agafonov and A. Ponomarev, "An experiment on localization of ontology concepts in deep convolutional neural networks," *Proceedings of the 11th International Symposium on Information and Communication Technology*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:254045293

[81] A. Agafonov and A. Ponomarev, "Localization of ontology concepts in deep convolutional neural networks," *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pp. 160–165, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:256215614

[82] G. Bangaru, L. B. Baru, and K. Chakravarthula, "Interpreting bias in the neural networks: A peek into representational similarity," *arXiv preprint arXiv:2211.07774*, 2022.

[83] Y. Bansal, P. Nakkiran, and B. Barak, "Revisiting model stitching to compare neural representations," *Advances in neural information processing systems*, vol. 34, pp. 225–236, 2021.

[84] M. Huh, B. Cheung, T. Wang, and P. Isola, "The platonic representation hypothesis," *arXiv preprint arXiv:2405.07987*, 2024.

[85] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[86] A. Shamir, O. Melamed, and O. BenShmuel, "The dimpled manifold model of adversarial examples in machine learning," *arXiv preprint arXiv:2106.10151*, 2021.

[87] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[88] M. Khoury and D. Hadfield-Menell, "On the geometry of adversarial examples," *arXiv preprint arXiv:1811.00525*, 2018.

[89] S. Jha, U. Jang, S. Jha, and B. Jalaeian, "Detecting adversarial examples using data manifolds," *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM)*, pp. 547–552, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:57376324

[90] W. Sha, Y. Luo, Y. Wang, and Z. Pan, "A defensive approach against adversarial examples based on manifold learning," *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*, pp. 167–171, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222222557

[91] S. Dube, "High dimensional spaces, deep learning and adversarial examples," *arXiv preprint arXiv:1801.00634*, 2018.

[92] L. Theis, "What makes an image realistic?" *arXiv preprint arXiv:2403.04493*, 2024.

[93] S. Dyrmishi, S. Ghamizi, T. Simonetto, Y. Le Traon, and M. Cordy, "On the empirical effectiveness of unrealistic adversarial hardening against realistic adversarial attacks," in *2023 IEEE symposium on security and privacy (SP)*. IEEE, 2023, pp. 1384–1400.

[94] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[95] L. Engstrom, J. Gilmer, G. Goh, D. Hendrycks, A. Ilyas, A. Madry, R. Nakano, P. Nakkiran, S. Santurkar, B. Tran, D. Tsipras, and E. Wallace, "A discussion of 'adversarial examples are not bugs, they are features'," *Distill*, 2019, https://distill.pub/2019/advex-bugs-discussion.

[96] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," *arXiv preprint arXiv:1801.09344*, 2018.

[97] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International conference on machine learning*. PMLR, 2018, pp. 5286–5295.

[98] K. Y. Xiao, V. Tjeng, N. M. Shafiullah, and A. Madry, "Training for faster adversarial robustness verification via inducing relu stability," *arXiv preprint arXiv:1809.03008*, 2018.

[99] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *international conference on machine learning*. PMLR, 2019, pp. 1310–1320.

[100] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," *Advances in neural information processing systems*, vol. 31, 2018.

[101] S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4536–4543.

[102] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" *arXiv preprint arXiv:1809.02104*, 2018.

[103] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow, "Adversarial spheres," *arXiv preprint arXiv:1801.02774*, 2018.

[104] A. Madry, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[105] S. Bubeck, Y. T. Lee, E. Price, and I. Razenshteyn, "Adversarial examples from computational constraints," in *International Conference on Machine Learning*. PMLR, 2019, pp. 831–840.

[106] P. Nakkiran, "Adversarial robustness may be at odds with simplicity," *arXiv preprint arXiv:1901.00532*, 2019.

[107] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," *Advances in neural information processing systems*, vol. 31, 2018.

[108] T. Tanay and L. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," *arXiv preprint arXiv:1608.07690*, 2016.

[109] B. Kim, J. Seo, and T. Jeon, "Bridging adversarial robustness and gradient interpretability," *arXiv preprint arXiv:1903.11626*, 2019.

[110] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "Robustness of classifiers: from adversarial to random noise," *Advances in neural information processing systems*, vol. 29, 2016.

[111] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 656–672.

[112] N. Ford, J. Gilmer, N. Carlini, and D. Cubuk, "Adversarial examples are a natural consequence of test error in noise," *arXiv preprint arXiv:1901.10513*, 2019.

[113] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 274–283.

[114] L. Karner, "The Dimpled Manifold Revisited," https://github.com/LukasKarner/dimpled-manifolds/, 2023, [Online; accessed 17-July-2024].

[115] Y. Kilcher, "Dimpled Manifold Counter Example," https://gist.github.com/yk/de8d987c4eb6a39b6d9c08f0744b1f64/, 2021, [Online; accessed 17-July-2024].

[116] Y. Kilcher, "The Dimpled Manifold Model of Adversarial Examples in Machine Learning (Research Paper Explained)," https://www.youtube.com/watch?v=k_hUdZJNzkU/, 2021, [Online; accessed 17-July-2024].

[117] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "RobustBench: a standardized adversarial robustness benchmark," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. [Online]. Available: https://openreview.net/forum?id=SSKZPJCt7B

[118] A. Araujo, L. Meunier, R. Pinot, and B. Negrevergne, "Advocating for multiple defense strategies against adversarial examples," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020, pp. 165–177.

[119] C. Ren, X. Du, Y. Xu, Q. Song, Y. Liu, and R. Tan, "Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples," *IEEE Transactions on Smart Grid*, vol. 13, pp. 1622–1632, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:245103286

[120] J. M. Adeke, G. Liu, J. Zhao, N. Wu, and H. M. Bashir, "Securing network traffic classification models against adversarial examples using derived variables," *Future Internet*, vol. 15, p. 405, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266389288

[121] Y. Feng and Y. Cai, "Towards robust classification with image quality assessment," *arXiv preprint arXiv:2004.06288*, 2020.

[122] Z. Rakhimberdina, X. Liu, and T. Murata, "Strengthening robustness under adversarial attacks using brain visual codes," *IEEE Access*, vol. 10, pp. 96 149–96 158, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252393135

[123] Y. Li, Q. Zhou, S. Li, and B. Li, "wadvmtd: A mitigation to white-box adversarial examples using heterogeneous models and moving target defense," *2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pp. 592–597, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259122131

[124] S. Kokalj-Filipovic, R. Miller, and G. M. Vanhoy, "Adversarial examples in rf deep learning: Detection and physical robustness," *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:210971684

[125] J. Zhao, J. Wu, J. M. Adeke, G. Liu, and Y. wei Dai, "Eitgan: A transformation-based network for recovering adversarial examples," *Electronic Research Archive*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264180609

[126] Q. Ji, L. Wang, C. Shi, S. Hu, Y. Chen, and L. Sun, "Benchmarking and analyzing robust point cloud recognition: Bag of tricks for defending adversarial examples," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4295–4304.

[127] M. Xie, Y. Wang, and H. Huang, "Fermi-bose machine," *arXiv preprint arXiv:2404.13631*, 2024.

[128] S. Fort and B. Lakshminarayanan, "Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness," *arXiv preprint arXiv:2408.05446*, 2024.

[129] J. Mickens, "Q: Why do keynote speakers keep suggesting that improving security is possible? a: Because keynote speakers make bad life decisions and are poor role models," in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/mickens