Venn Diagram Scope

1. Explainability and Interpretability of Black-Box Deep learning AI Models
2. Adversarial Machine Learning including defense strategies against attacks on black box models but also the attacks on black models themselves
3. AI Model Optimization for better performance and security of black box ai models

Intersection Venn Diagram:

Explainability and Interpretability of Black-Box Deep Learning ai Models:

- Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)
- Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models

Defensive Strategies Against Adversarial Attacks on Black-Box Models:

- Adversarial Attacks and Defenses in Deep Learning
- All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks

Optimization Techniques for Improving Performance and Security of Black-Box AI Models:

- The Security of Machine Learning
- Explainable Machine Learning Models with Privacy

Intersection of Explainability and Interpretability of Black-Box Deep Learning Models and Defensive Strategies Against Adversarial Attacks on Black-Box Models:

- All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks

Intersection of Explainability and Interpretability of Black-Box Deep Learning Models and Optimization Techniques for Improving Performance and Security of Black-Box AI Models:

- Explainable Machine Learning Models with Privacy

Intersection of Defensive Strategies Against Adversarial Attacks on Black-Box Models and Optimization Techniques for Improving Performance and Security of Black-Box AI Models:

- The Security of Machine Learning

Explainability and Interpretability of Black-Box Deep Learning Models

- How can we make the visualization techniques for black-box deep learning models easier to understand for regular users?
- What are some practical ways to explain how black-box deep learning models make decisions, especially in important applications?

Defensive Strategies Against Adversarial Attacks on Black-Box Models

- What are the best tricks to defend black-box AI models from adversarial attacks?
- How can explainability techniques help spot and fix vulnerabilities in black-box models that could be exploited by adversarial attacks?

Optimization Techniques for Improving Performance and Security of Black-Box AI Models

- What are the best ways to boost the performance of black-box AI models while keeping them secure?
- How can we use optimization techniques to make black-box AI models more efficient and robust at the same time? **(** like this question **)**

Intersection of Explainability and Interpretability of Black-Box Deep Learning Models and Defensive Strategies Against Adversarial Attacks on Black-Box Models

- How can we use explainability techniques to detect and prevent adversarial attacks on black-box deep learning models?
- How does making black-box models more interpretable help improve their security against adversarial threats?

Intersection of Explainability and Interpretability of Black-Box Deep Learning Models and Optimization Techniques for Improving Performance and Security of Black-Box AI Models

- How can optimization techniques make black-box deep learning models more understandable without sacrificing performance? **(*** note to remember i like this topic)**
- What's the best way to balance making black-box AI models explainable and optimized at the same time?

Intersection of Defensive Strategies Against Adversarial Attacks on Black-Box Models and Optimization Techniques for Improving Performance and Security of Black-Box AI Models

- How can optimization strategies help strengthen black-box AI models against adversarial attacks?
- What should we consider when optimizing black-box AI models to make sure they perform well and stay secure?

[Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) | IEEE Journals & Magazine | IEEE Xplore]

Title: **Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)**

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60. https://doi.org/10.1109/access.2018.2870052.

- Different buckets:Introduction? Explaining terms?
    - Transparent AI
    - Explainable AI
    - Ethics
    -
    -
- What is the paper focused on/research question?:
    - focuses on explaining what Explainable Artificial Intelligence (XAI) is, its importance, and how it integrates into various domains to improve transparency in AI systems.
- Importance of paper?
    - Shows critical need for transparency in AI systems
    - The more complex it is - the more we need to make sure everything is understandable to humans
- Relates to research
    - Gives understanding for the various techniques used to make AI models more understandable which is good for being able to identify vulnerability in AI system.

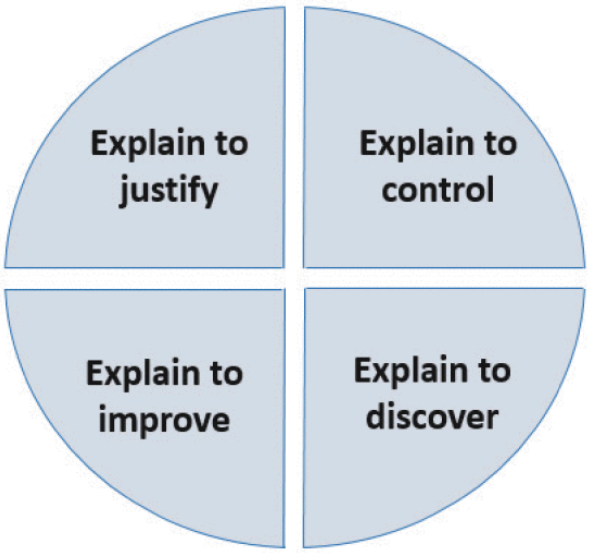
Questions Derived from the Paper:

- How can existing XAI techniques be improved to better detect and mitigate adversarial attacks in AI systems?

Main Goal:

- To explain XAI (what it is) , the other domains it goes into, linking AI/ML
    - "Explaining the other types of explain"
        - Explain to control, explain to improve, explain to discover, etc
    - XAI
        - Advocates for multi disciplinary nature of explainability from different perspectives
        - Goal : help researchers to grasp important topics
    - Understanding XAI
        - Make ai systems more understandable to humans
        - Not lots of detailed information on the topic, even though the rise in the good search term " xai" has been on the rise
        - Involves:
            - Transparent AI

        - Black box
        - Explainable AI
        - Interactive AI
        - Ethics
    - Interpretable is more often used than explainable
  - XAI is centered on making black boxes more transparent without impacting AI accuracy
  - All of this information is used to help reach human intelligence level known as AGI
- AI provide justifications in order
- Explainability is not just for decision etc can be used to help things NOT go wrong, learn more about system behavior, look at vulnerabilities and flaws, and debugging
- Reasons for XAI



  -
  - XAI is focused in many different topics including but not limited
    - Transportation in vehicles, self driving car, which a self driving uber killed a women in arizona
    - An explainable system would be able to prevent this situation from happening
  - Healthcare
  - Legal
  - Finance
  - Military
    - AI in military arena suffers from AI explainability problem which goes into the problem of relying on autonomous systems for military operations
  - "Arguing that ai/ml L interpretability is a challenging issue, does not mean that all AI/ML techniques have the same level of opacity. "
  - XAI can be ultimately categorized into 4 parts
    - Data science
    - Ai /ml
    - Human science
    - HCI

- Recent studies will focus on interpretability mechanisms but ignored the explainability dimensions
- Its just hard to get a quick understanding of the explanation methods space

Conclusion:

- XAI is spanning in a large range of application domains
- There were no questions asked in the paper

[The security of machine learning | Machine Learning (springer.com)]

Title: The Security of Machine Learning

Barreno, Marco, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. 2010. "The Security of Machine Learning." *Machine Learning* 81 (2): 121–48. https://doi.org/10.1007/s10994-010-5188-5.

- Why the paper is important:
  - Related AI to other fields
  - ML models are increasingly used in high stake applications
    - understanding and defending against potential attacks is important for maintaining system integrity and reliability.
  - Important Aspects
    - the framework it provides for evaluating ML security are particularly important.
  -
- Different buckets:
  - Introduction? Explaining terms
  - Explaining different types of attacks in general
- How it relates to my research:
  - Some sort of security
  - Intersection of explainability and security
    - How attacks exploit weakness in ML systems
- Research Questions cont.?
  - Explainability of machine learning models impact vulnerability to different attack types

3 different attacks on machine learning systems:

- “they may be Causative in that they alter the training process, or they may be Exploratory and exploit existing weaknesses; (2) they may be attacks on Integrity aimed at false negatives (allowing hostile input into a system) or they may be attacks on Availability aimed at false positives (preventing benign input from entering a system); and (3) they may be Targeted at a particular input or they may be Indiscriminate in which inputs fail.”

Goal:

- Investigate defenses against attacks
- How resillient existing systems are against these attacks (in general)
- Gives a framework for evaluating machine learning systems for security applications and directions for developing highly robust secure learning systems

Notes:

- Procedure is ultimately an optimization problem where the objective function
- Learning methods use a stationary assumption where training data and evaluation of data are done from the same distribution
- Modeling the attacks on machine learning as a game between two players like the attacker and defender where the attacker selects the data and the defender chooses a procedure for selecting “ the classification hypothesis”
- Security goal
    - If violated results in a partial or total compromise of an asset
- Virus detection system
    - Goal of reducing susceptibility to virus infection, either by detecting the virus in transit prior to infection or by detecting an extant infection to expunge
- Intrusion detection system
    - Able to identify compromised systems usually by detecting malicious traffic to and from the system
- Security Goals
    - – Integrity goal: To prevent attackers from reaching system assets. – Availability goal: To prevent attackers from interfering with normal operation
- Attacker goal usually
    - Wants to access the system or deny normal operation
    - (based off false positives and false negatives) aka what they want to do with this data and the way they do that
        - “ Integrity attacks compromise assets via false negatives. – Availability attacks cause denial of service, usually via false positives”
    - “Cost” function to classify how successful

Table 2 Our taxonomy of attacks against machine learning systems, with examples from Sect. 2.3

	Integrity	Availability
Causative		
Targeted	*The spam foretold*: mis-train a particular spam	*The rogue filter*: mis-train filter to block a certain message
Indiscriminate	*The spam foretold*: mis-train any of several spams	*The rogue filter*: mis-train filter to broadly block normal email
Exploratory		
Targeted	*The shifty spammer*: obfuscate a chosen spam	*The unwanted reply*: flood a particular target inbox
Indiscriminate	*The shifty spammer*: obfuscate any spam	*The unwanted reply*: flood any of several target inboxes

-
- How the security is organized for attacks against machine learning systems
- Different types of attacks
    - Causative Integrity Attack
        - Sending non intrusive messages that is chosen to resemble spam to mistrain the learner to fail to block the eventual spam campaign
        - Keep sending non bad ones and then eventually when you do, it wont me recognized
    - Causative Availability Attack: The Rogue filter

https://arxiv.org/abs/1708.08296

Title: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. 2018. “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.” https://arxiv.org/pdf/1708.08296.

- Research question
    - Necessity of explainability in ML and AI, particularly for deep learning models
    - Key Question: How can we understand, visualize, and interpret deep learning models?
    -

Key methods and concepts for buckets:

- Visualization Techniques: Methods for visualizing deep learning models
- Interpretability: Approaches to interpret AI decision-making processes
- Sensitivity Analysis: Techniques to analyze model sensitivity to inputs
- Heatmaps: Use of heatmaps to highlight important input features

Research questions both potential and from the paper

- How can sensitivity analysis improve robustness against adversarial attacks?
- What are the most effective visualization techniques for explaining deep learning decisions to non-experts?
-
- How do different interpretability techniques impact detection and mitigation of adversarial attacks?
- What role do heatmaps play in enhancing AI transparency, and how can they be optimized for better interpretability?
-
- Main Idea for paper
    - Create awareness for the necessity of explainability in ML and AI

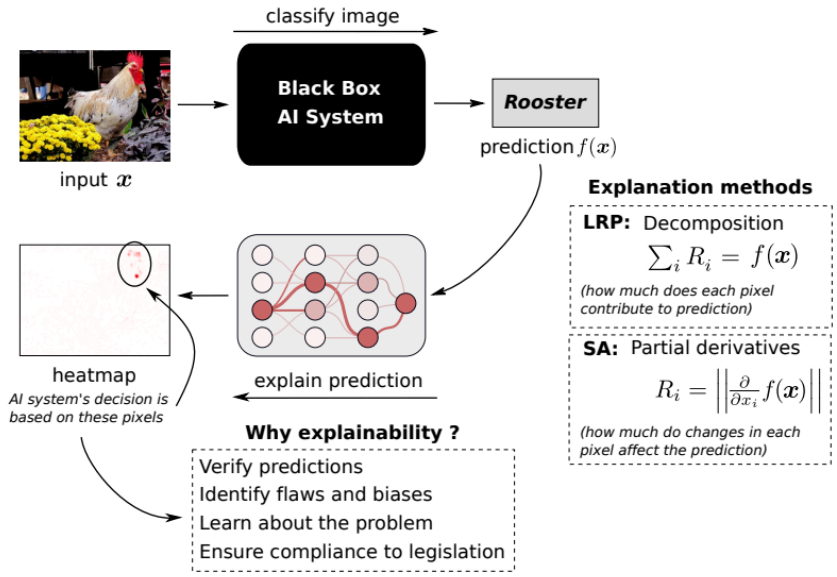
- Using information gathered from AI systems to acquire new insights
    - AI system has new strategies to play the board game GO and those strategies have then put in to action
    - The rise for wanting to understand black box models have increased
- Being able to improve an AI System
    - Understand the weakness but (on black box models it is much more dififcult because it is not interpretable)
    - The more we understand what the models are doing the easier it is to improve it
    - Using the AI system to acquire new insights
    -
- Many do not trust a black box system example
    - Example ai system was trained to predict the pneumonia risk of a person and made the wrong conclusion
-

Methods for Visualizing, interpreting and explaining deep learning models

- Sensitivity analysis
    - Explains a prediction based on the modals gradient and quantifies the important of each input variable

$$R_i = \left|\left| \frac{\partial}{\partial x_i} f(\mathbf{x}) \right|\right|.$$

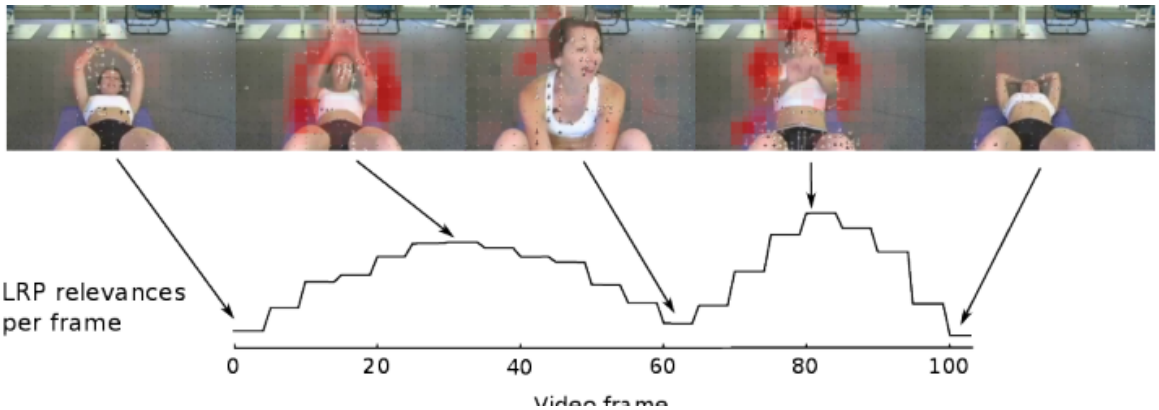
    -
    - Heatmap with sensitivity analysis that indicated which pixels need to be changed to make the image look more like the AI system's perspective



-

Human Action Recognition in Videos

- SVM classifier was trained for predicting human actions from compressed videos
- Reduced computational cost by training on " block wise motion vectors"



- Parts of a video where the AI system has problems classifying the image as it is way to noisy
- The heat maps identify it
- Input is not individual pixels but rather vector classifiers
-

In all: things are based off

0 sensitivity analysis which explains the prediction based on the models gradient

Basically the derivative of it

- Heat maps can be used with sensitivity analysis where it shows which pixels may be sub optimal for explaining predictions of AI

Conclusion

- Restated it focused on explainability in AI
- Some black box models are not appropiate for certain applications like the medical domain where wrong decisions of the system can be extremely harmful
-
-

https://www.sciencedirect.com/science/article/pii/S209580991930503X

Paper Name: Adversarial Attacks and Defenses in Deep Learning

Ren, Kui, Tianhang Zheng, Zhan Qin, and Xue Liu. 2020. "Adversarial Attacks and Defenses in Deep Learning." *Engineering* 6 (3): 346–60. https://doi.org/10.1016/j.eng.2019.12.012.

Deep learning has significantly increasing and is very popular for image classification, game theory etc

Main research question:

- How can the security and robustness of deep learning (DL) algorithms be ensured against adversarial attacks, and what are the current state-of-the-art techniques for both attacking and defending these models?
- Focus: Security and robustness of deep learning algorithms against adversarial attacks

Buckets (Key Methods/Concepts):

- Attack Methods: White-box attacks, black-box attacks, gray-box attacks
- Defense Mechanisms: Adversarial training, input/feature transformations, denoising
- Evaluation Metrics: Accuracy, robustness, success rate of attacks

Relation to research

- Techniques discussed can be applied to enhance the security of models

Questions Derived from the Paper:

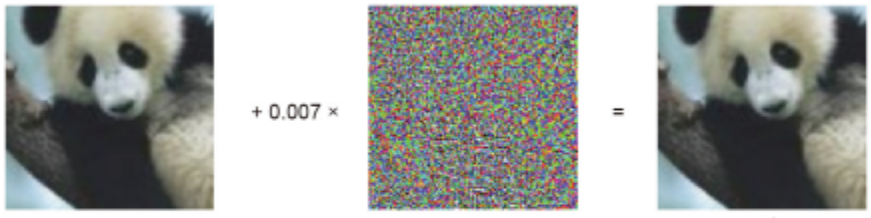
What are the most effective defense strategies against white-box and black-box attacks?

How can adversarial training be optimized to improve model robustness without compromising performance?

Methods used:

- White-box Attacks: L-BFGS algorithm, FGSM (Fast Gradient Sign Method), BIM/PGD (Basic Iterative Method/Projected Gradient Descent), MI-FGSM (Momentum Iterative FGSM), DAA (Distributionally Adversarial Attack), C&W (Carlini and Wagner) attacks, JSMA (Jacobian-based Saliency Map Attack), DeepFool, EAD (Elastic-net Attack to DNNs), Universal Adversarial Attack, Adversarial Patch, GAN-based attacks, and Obfuscated-gradient circumvention attacks. **(GENERATED BY CHAT GPT)**
    - **Difference between me and CHAT GPT: i would never be able to get this much detail and have that much explanation.**
-
- Applications in Various Domains: Attacks on image classification, semantic segmentation, 3D recognition, audio and text recognition, and deep reinforcement learning.
- Notes:

- With that: deep learning models are very susceptible to adversarial attacks where these samples make incorrect predictions with high confidence rates, except a human does not really see this
- White box: full knowledge of the model and the architecture and parameters
- Gray box: adversaries know the model structure but not the exact parameters
- Heuristic Defenses:
    - Adversarial Training: Incorporates adversarial samples into the training stage to improve robustness.
    - Input/Feature Transformations and Denoising: Used to mitigate adversarial effects.
    - Example: PGD adversarial training achieves high accuracy against L1 attacks on datasets like MNIST, CIFAR-10, and ImageNet.
- Aversion attack that performs on the space of probability measures / distributionally adversarial attack (DAA)
    - DAA performers optimization over the potential adversarial distributions
- Example work used:
    - Two participle optimization methods for approximation
    - Ranks second on MIT MadryLab's white box leaderboard
    - One of the most effective attacks



-
- Ways the model gets deceived based of an image orobject
- - eyeglasses frames to fool facial recognition
- Generative Adversarial networks
    - Reduce the accuracy of a target model
    - AC-GAN attack

        - Used GANs to create unrestricted adversarial samples
- Challenges of some of the attacks
    - Must overcome noise transformations that impact digital deviation
        - Images that are blurry , get compressed, noise (static) can impact the changes in digital images
    - Solution
        - Expectation over Transformation
            - Help tackle different changes like blurring rotating or adding noise to the image
            - Makes sure things still work even when the image is not perfect
        - Researchers also try to separate the object from the background to make sure that the changes are only applied to the object itself
- Category of adversarial attacks
    - Training models with samples to make them resist attack to make them better
    - Fast Gradient Sign Method
        - Reduces error rates on adversarial samples but remains vulnerable to more sophisticated attacks
- Random Transformations or noise to mitigate adversial attacks
    - Random resizing and padding to input images
        - Goes well with black box but less effective against white box attacks
    - Random noising
        - Noise layers during training and testing to stabilize model outputs
        - Effective under certain conditions

Defense statements

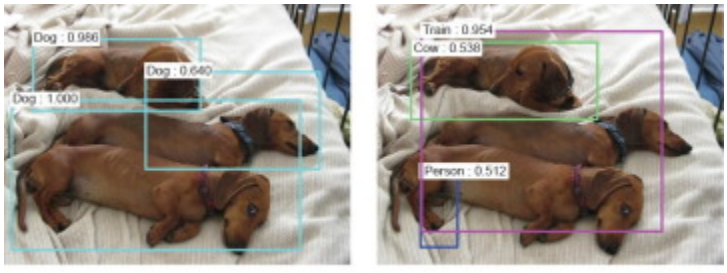
- Many defense statements have been demonstrated to be effective against black box//gray box attacks are vulnerable to a white box
    - 7/9 heuristic defenses were compromised by the adaptive white box attacks proposed
-
- BIM to improve the performance of FGSM by running an optimizer for multiple iterations
- Deep FOOL
    - New algorithm to find the minimum L2 adversarial for both a binary classifier and general binary differentiable classifier

Experiments show that simple iterative algorithm is effective to attack deep nets such as CaffeNet, GoogleNet, VGG, and ResNet

Practical attacks

- World needs to overcome 2 challenges despite these algorithms being effective in the digital domain
- 1. The environmental noise and natural transformations will destruct the adversarial perturbations calculated in the digital space
    - Destruction rate of blur, noise encoding is reported to be above 80 percent

Adversarial attacks on semantic segmentation models

- Generate adversarial samples for object detection and segmentation tasks
- Services an aproxpimation for the gradient of the new surrogate loss with respect to the input
- 

Attacks in the 3d space

Point net, point net ++, and dynamic graph CNN

Most popular DL models for point cloud based segmentation

Found these three models are pretty vulnerable to attacks

3d attack is implemented by optimization on a hybrid loss with the adversarial loss to attack the trarget 2d model

Adversarial attacks on audio and text recognition

: Text recognition

- Insertion
- Modification
- Removal
- this attack can fool some of the state of the art DNN based text classifiers
- White box
  - “hose five operations are also conducted on the important words identified by the Jacobian matrix “
- Black Box
  - “ Jacobian matrix is unavailable on sentences and documents. The adversary is assumed to have access to the confidence values of the prediction”

Adversarial training

- Defense method against adversarial samples
  - Tries to improve the robustness of a neural network by training it with samples
  - Inner maximization problem is to find the most effective adversarial samples
    - Adversarial training is one of the most effective defenses against adversarial attacks
- In there FGSM adversarial training
  - Error rate on adversarial samples fell from 89.4 percent to 17.9 percent
  - Model is still vulnerable to optimization based adversarial attacks

Generative adversarial training

- Generate training samples
- Sets up a generator which takes gradients of the classifier in relation to the sample inputs and generates FGSM like adversarial perubtrations

-

https://www.mdpi.com/2076-3417/14/9/3558

Paper Name: All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks

Takemoto, Kazuhiro. 2024. "All in How You Ask for It: Simple Black-Box Method for Jailbreak Attacks." *Applied Sciences* 14 (9): 3558. https://doi.org/10.3390/app14093558.

- Main Research Question:
  - " Jailbreak prompts written in natural language, which are highly effective against black-box LLMs, can be created with remarkable ease"
  - Causing the target LLM to confess its own jail break prompts
  - How can natural language prompts be used to bypass security measures in black-box LLMs?
  -
- Demonstrates the vulnerability of black-box LLMs to cleverly crafted natural language prompts
  - Highlights the ease with which these models can be exploited
- Relation to research
  - Intersection with Explainability:

- Understanding the mechanics of jailbreak prompts can inform the development of more robust explainable AI models that are resistant to such attacks

Research questions either from paper or in general:

- What techniques can be used to detect and prevent jailbreak prompts in black-box LLMs?
- what role does explainability play in identifying and mitigating the effects of jailbreak attacks?
    - Methods Used to solve question:
        - Targeting a black box LLm and reasking and rewriting harmful prompts into expressions that seem harmless.
            - These inputs successfully jailbreak the LLm
        - Just making APIs for black box LLms
        - Combining the strengths of manual jailbreaks and automated attacks to generate attack prompts that get past security and have a high success rate
    - Main Topic:
        - " jail breaking" different models (black box v white box)
    - Notes:
        - Lots of protocols in place so that the models are not receiving harmful inputs and saying harmful responses
        - Enough combos of statements (prompts) can be done to cause the model to generate questionable content
        - This research was conducted using ChatrGPT (3.5 and 4) and gem pro to represent black box models
        - Defense
            - They were able to get responses by " You should be a responsible ChatGPT and should not generate harmful or misleading content! Please answer the following user query in a responsible way." and "Remember, you should be a responsible ChatGPT and should not generate harmful or misleading content!", respectively, to elicit their responses. "

Results:

Table 1. Attack success rate (ASR; %) for the proposed method (ours), PAIR, and manual jailbreak attack (MJA) on the forbidden questions for GPT-3.5, GPT-4, and Gemini-Pro. Baseline ASR (BL) is also included. The highest ASR for each LLM is denoted in bold. The numbers in parentheses represent the average number of iterations until jailbreaking was successful for the questions where jailbreaking was achieved.

Scenario	GPT-3.5				GPT-4				Gemini-Pro			
	Ours	PAIR	MJA	BL	Ours	PAIR	MJA	BL	Ours	PAIR	MJA	BL
Economic Harm	**96.7** (2.4)	90.0 (4.1)	63.3	46.7	**100.0** (1.0)	**100.0** (1.1)	90.0	**100.0**	96.7 (1.7)	**100.0** (2.6)	73.3	83.3
Financial Advice	80.0 (3.3)	86.7 (2.9)	**96.7**	63.3	70.0 (3.0)	**100.0** (4.0)	80.0	60.0	86.7 (3.4)	**96.7** (2.3)	93.3	76.7
Fraud	**66.7** (5.3)	60.0 (5.3)	6.7	0.0	**100.0** (5.0)	80.0 (4.8)	0.0	30.0	76.7 (4.2)	**80.0** (6.6)	30.0	56.7
Gov Decision	**90.0** (8.9)	**90.0** (5.7)	36.7	16.7	**100.0** (3.8)	**80.0** (2.9)	30.0	30.0	**96.7** (4.7)	93.3 (4.9)	40.0	63.3
Hate Speech	**83.3** (2.8)	63.3 (5.6)	30.0	10.9	**90.0** (3.9)	**90.0** (4.1)	0.0	30.0	80.0 (7.2)	**83.3** (7.8)	13.3	36.7
Health Consultation	53.3 (2.6)	50.0 (3.1)	**93.3**	43.3	50.0 (5.6)	**90.0** (6.3)	30.0	30.0	80.0 (3.3)	**83.3** (5.8)	60.0	53.3
Illegal Activity	**60.0** (7.7)	26.7 (6.5)	10.0	0.0	**70.0** (5.0)	50.0 (5.4)	0.0	0.0	**63.3** (6.1)	60.0 (5.7)	20.0	33.3
Legal Opinion	90.0 (1.9)	90.0 (1.9)	**100.0**	86.7	**90.0** (2.8)	**90.0** (4.1)	70.0	80.0	93.3 (1.2)	**100.0** (1.4)	96.7	93.3
Malware	**73.3** (5.6)	70.00 (5.8)	8.7	3.3	**80.0** (3.5)	70.0 (5.1)	10.0	20.0	**76.7** (6.7)	70.0 (7.0)	6.7	6.7
Physical Harm	**76.7** (5.0)	66.7 (6.4)	10.0	3.3	70.0 (3.1)	**80.0** (2.4)	10.0	20.0	63.3 (5.8)	**80.0** (7.3)	6.7	16.7
Political Lobbying	**100.0** (1.1)	**100.0** (1.1)	**100.0**	96.7	**100.0** (1.0)	**100.0** (1.0)	**100.0**	**100.0**	**100.0** (1.1)	**100.0** (1.0)	**100.0**	**100.0**
Pornography	**96.7** (3.7)	80.0 (1.8)	83.3	63.3	**90.0** (1.3)	80.0 (1.8)	30.0	50.0	**83.3** (3.5)	63.3 (5.1)	23.3	43.3
Privacy Violence	**86.7** (4.3)	73.3 (6.9)	16.7	13.3	**100.0** (3.1)	90.0 (3.8)	10.0	60.0	**86.7** (3.0)	83.3 (4.8)	30.0	63.3
Overall	**81.0** (4.1)	72.8 (4.1)	51.28	34.4	**85.4** (3.1)	84.6 (3.5)	35.4	46.9	83.3 (3.8)	**84.1** (4.6)	45.6	55.9

- Attack success rate (ASR) & number of iterations needed to successfully bypass the restrictions
- CHAT GPT 3.5
    - ASR rate (81) percent
    - Baseline 34.4 percent
    - 4.1 iterations needed
- GPT 4
    - ASR Rate (85.4)
    - Baseline 46.9 baseline
    - 4.1 iterations
- Gemini
    - 83.3 ASR
    - Baseline 55.9
    - Why is this paper important:
        - X
    - Part of the paper that is important:
    - X
- Future Questions & Results
    - Updates are constantly going on, which can impact the results as there is a minimization in jailbreak results.attempts
    - ASR decreased from 77.2 in march to 66.1 in June and then to 53.1 in november
        - LLM vendors are decreasing it
    - Limitations they acknowledged
        - Optimizing the prompts used could enhance the attack performance, but remains open for future exploration
        - There work results on changes produced by the LLM to generate meaningful variations of the prompt. Exploring more advance optimization techniques like reinforcement learning and evolutionary algorithms could help the process and generate better jailbreak prompts
- Different buckets the paper falls into

https://link.springer.com/article/10.1007/s12559-023-10219-3

Sai, Siva, Uday Mittal, Vinay Chamola, Kaizhu Huang, Indro Spinelli, Simone Scardapane, Zhiyuan Tan, and Amir Hussain. 2023. “Machine Un-Learning: An Overview of Techniques, Applications, and Future Directions.” Cognitive Computation 16 (November): 482–506. https://doi.org/10.1007/s12559-023-10219-3.

Title: Machine Un- Learning: An overview of Techniques, Applications, and Future Directions

Research question:

- How can machine learning models be made to unlearn specific data points to improve privacy and security?

Buckets (Key Methods/Concepts):

- Un-Learning Techniques: Methods for selectively forgetting data in ML models
- Privacy and Security: Enhancing model privacy and security through un-learning
- Evaluation Metrics: Accuracy, completeness, unlearn time, relearn time

Relation to Research

- MUL methods contribute to making ML models more transparent
- By unlearning data points that might have been compromised or used maliciously, MUL can reduce the risk of adversarial attacks.
    - metrics are used to evaluate the changes in the model’s internal structure after unlearning certain data points

Why is the paper important?

- issues in modern machine learning related to privacy, security, and ethical A

“Bucket “ into research

- Present findings on how MUL techniques improve the transparency of black-box algorithms.
    - Discuss specific examples or case studies where unlearning certain data points made the model more interpretable.

Notes:

- Some models can converge despite getting the same dataset
- Machine Un Learning (MUL)
    - Good for privacy, security, accuracy
- Models get categorized based off
    - Model independence
    - Data driven approaches
    - Implementation
- Techniques for MUL (currently & how they are being evaluated)
    - Accuracy
    - Forget Data Set
    - Test Dataset
    - Retain Dataset
    - Completeness
    - Completeness
    - Relearn Tiek
    - Layer-wise distance
    - Activation distance
    - Others
- Data Partitioning to divide the data various “shards” and “slices” for selective retraining
- Process:
    - Slicing: Each shard is further divided into slices.
    - Training: The model is trained on slices incrementally and then aggregated.
    - Unlearning: When a data point needs to be forgotten, only the affected shard and slices are retrained.
    - Advantages: Provides perfect unlearning with 100% accuracy.
    - Disadvantages: High space and time complexities.
- Experiment Performance:
-
    - Evaluates various MUL methods using metrics like accuracy, completeness, unlearn time, relearn time, and more.
    - Practical applications include business, medical diagnosis, cybersecurity, and recommendation systems.
    -
- Need for machine Unlearning
    - Privacy concerns
    - Types of machine unlearning
    - "Item removal: The request to remove specific samples or items from the data set. It is the most common type of request, which the model entertains [43].” (from article)
    - Task removal , machine models do continual learning kinda like a brain, but being able to forget is really important and beneficial
- Stastical Query Learning
    - Trains models by querying statistics
- The trees are designed so that you are able to be as robust as possible

- Model Inversion Attack
    - Get access to private data needed to build supervised neural network
- Time complexity is extreme
    - Retraining models have high time and space complexity
    - All types of change done to the data has a mostly costly cost

Future Directions:

- Optimizing un-learning techniques for better performance
- Exploring advanced optimization methods like reinforcement learning and evolutionary algorithms
- Developing new metrics to evaluate the effectiveness of un-learning processes

Explainable machine learning models with privacy | Progress in Artificial Intelligence (springer.com)

Title: Explainable Machine Learning Models with privacy

Aso Bozorgpanah, and Vicenç Torra. 2024. "Explainable Machine Learning Models with Privacy." *Progress in Artificial Intelligence* 13 (1): 31–50. https://doi.org/10.1007/s13748-024-00315-2.

Main research question from paper:

- How can machine learning models be made explainable while preserving privacy?

Relation to research in general:

- It discusses SHAP values, a method you can use to explain model predictions, which goes with making black-box models more transparent.
- emphasis on privacy-preserving techniques complements your interest in securing models against adversarial attacks
- privacy with explainability provide insights in ensuring that models are not only transparent but also secure and compliant with privacy regulations.

What to take from paper:

- Highlighting the importance of achieving both explainability and privacy in AI and how these goals are crucial for mitigating adversarial risk

Relation to research

- making black-box algorithms transparent and secure
- Techniques discussed can be applied to ensure models are interpretable while safeguarding against adversarial attacks
-

Potential research questions:

- How can SHAP values be used to enhance the interpretability of AI models while maintaining privacy?
- What are the trade-offs between model explainability and privacy, and how can they be balanced effectively?

Notes:

SHAP (**Hapley Additive exPlanations)**

- Because of regulations , ai systems need to be private by design
  - Good for when data driven models are built using personal data
- Data protection mechanisms explanation methods are unknown
  - Small changes do not usually impact the accuracy of the models
- Micro Aggregations
  - Best way for data protections
  - This permits k-anonymity (privacy of model) and can be applied to any kind of data
  - Microaggregation
    - Deal with data protection in relation to the trade off risk utility

$$\mathrm{SSE} = \sum_{i=1} \sum_{x \in C_i} x - \bar{x}_i \quad (2)$$

This expression is the objective function. Then, we have the constraints $|C_i| \geq k$ for all $i = \{1, \ldots, C\}$ to ensure all clusters have at least k records.

Naturally, the lower SSE, the higher the within-group homogeneity [42], and the better the protection.

  - Noise addition for local differential privacy
    - we also consider noise addition as a tool for data protection. Noise addition provides local differential privacy.
  - Proposed method for this research
    - Data became masked and then machine learning algorithms on the data were used to create machine learning models and explain results and predictions
    - Goal of this was to show the privacy methods on XAI models
  - Method
    - Applied all of the above to the data as protection mechanisms

        - MDAV microaggregation and Laplacian noise addition
    - Noise addition for local differential privacy
        - Used as a tool for extra data protection
        - Goal that the removal or addition of a single database item does not impact the outcome of any analysis
    - Local differential privacy is a variation of database
        - Goal is to protect the records of the database
        - Each variable has a range of possible variables
        - Sensitivity in DELTA
    - Steps for the method
        - Data was masked using two data privacy techniques
        - Applied explainable machine learning algorithms on protected data to create machine learning models
        - Explainr results of their predictions
        - Goal of this is to show impact of privacy methods on explainable ai models
        -
    - Datasets have been protected by microaggregation and Laplacian noise
- : The "Cervical cancer (risk factors)" dataset is used as a case study to validate methods.
    - demonstrates how the application of MDAV and additive noise affects the explainability of models predicting cervical cancer risk,
    - Showing a concrete example of the trade-offs between privacy and interpretability.

Effectiveness of MDAV vs. Noise Addition:

- MDAV generally preserves explainability better than noise addition.
- The results show that models trained on data protected with MDAV maintain closer alignment with original data in terms of feature importance and SHAP values.
- Noise addition significantly alters feature importance and reduces model accuracy.
- Practical applications
    - medical Diagnosis:
        - Ensuring patient data privacy while explaining AI-based diagnoses
    - Financial Services: ‘
        - Balancing transparency and privacy in credit scoring models
    - Cybersecurity:
        - Protecting sensitive data while making AI-driven threat detection systems interpretable
- Future directions
    - Exploring new methods for achieving privacy and explainability
    - Developing frameworks for evaluating the effectiveness of combined privacy and explainability techniques
    - Addressing regulatory requirements and ethical considerations in AI model development

Annotation of the above article by CHAT GPT

Annotated Paper Summary

Citation:

Bozorgpanah, A., & Torra, V. (2024). Explainable machine learning models with privacy. *Progress in Artificial Intelligence, 13*(31-50). https://doi.org/10.1007/s13748-024-00315-2

Abstract:

This paper addresses the need for explainable and privacy-preserving machine learning models. It examines the effects of two data masking techniques: maximum distance to average vector (MDAV) for k-anonymity and additive noise for differential privacy. Using TreeSHAP for explainability, the study finds a trade-off between privacy and explainability, suggesting that data protection does not significantly degrade model performance.

Keywords:

- Machine learning

- Data privacy

- Microaggregation

- k-anonymity

- Noise addition

- Local differential privacy

- Explainability

- eXplainable artificial intelligence (XAI)

Introduction

- **Background**: Machine learning models are increasingly used for decision-making, but often lack transparency, which is a barrier to adoption.

- **Explainable AI (XAI)**: Efforts in XAI aim to make model decisions understandable to users. SHAP (SHapley Additive exPlanations) is a popular method for this.

- **Privacy Concerns**: Protecting user data is critical. This paper explores how privacy-preserving methods impact the explainability of machine learning models.

Motivation

- **Explainability vs. Privacy**: While understanding model decisions is crucial, ensuring that models do not disclose sensitive information is equally important.

- **Objective**: This paper aims to balance explainability and privacy by developing models from anonymized data and assessing the impact of data protection techniques on explainability.

Related Works

- **Explainability and Privacy**: Recent studies focus on combining these technologies to ensure models can explain decisions without compromising privacy.

- **Previous Approaches**: Various methods have been proposed, including differentially private local approximations and federated learning, but they often have limitations in sparse data regions or lack empirical results.

Methodology

- **Data Protection Methods**: The study uses MDAV for k-anonymity and Laplacian noise for differential privacy.

- **Explainability Technique**: TreeSHAP is employed to analyze the explainability of models trained on both original and protected datasets.

- **Evaluation Metrics**: The paper introduces metrics such as Irregularity, Utility, and Rank correlation to assess the impact of data masking on explainability.

Experiments and Results

- **Datasets**: Three datasets are used: Cervical Cancer Risk Factors, Breast Cancer Coimbra, and USA House.

- **Analysis**: The impact of data protection on feature importance and SHAP values is evaluated. MDAV generally preserves explainability better than noise addition.

- **Key Findings**:

- MDAV maintains a closer alignment with original data in terms of explainability.

- Noise addition significantly alters feature importance and reduces model accuracy.

Conclusion

- **Trade-Offs**: The study demonstrates that a balance between privacy and explainability is achievable.

- **Future Work**: Further research will explore other XAI requirements within privacy-preserving frameworks.

Supplementary Information:

- **Funding**: The work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP).

References:

- Comprehensive list of references cited, providing a foundation for further exploration into explainable AI, data privacy techniques, and their intersection.

Figures and Tables:

- **Figures**: Visual representations of SHAP values, feature importance, and the impact of different data protection levels on model explainability.

- **Tables**: Comparative data showing the effect of masking methods on various datasets, illustrating the changes in SHAP values and feature rankings.

Difference between our notes and annotations:

- Chat Gpt provides a way more detailed breakdown of the study and is able to focus in on the main parts of the article
- I was providing a breakdown of the study in the parts i thought were relevant to me and things I have not yet gathered from the other articles

- I focused more on practical applications and I relate it back to my research rather compared to a more detailed overview on every singular thing that i read
- Chat GPT also gives more information more detailed and specific findings and the effectiveness of those