Adversarial Attacks and Defenses: Adversarial attacks on machine learning models. Defense mechanisms against adversarial attacks. Adversarial robustness and detection techniques

neural network architectures

CNNs (AlexNet, VGG, ResNet, ...)

RNNs - Widely used in language modeling and speech recognition

Transformers (BERT, GPT)

Security

- CIA Triad
- Threat modelling - a structured approach that's used to identify, prioritize, and manage potential threats in a system
  - STRIDE
  - Attack trees
  - More in detail:
    - https://insights.sei.cmu.edu/blog/threat-modeling-12-available-methods/
    - https://attack.mitre.org/
  - MITRE ATLAS
- Risks and mitigations
  - Identify threats --> assign risk
  - CIS Benchmarks
  - OWASP Top 10

- DevSecOps
- Host security
- Regular updates
- Minimal software
- User access control
- Firewall configuration
- Container security
- System monitoring and auditing
- Backup and recovery
- Disable unused network services
- Secure Shell (SSH) access
- Endpoint security
- Vulnerability management
- Network protection
- Securing code and artifacts

Bypass security with adversarial ai

Adversarial Robustness Toolbox

Surrogate shadow model

Adversarial robustness goes beyond traditional security measures

Detailed taxonomy of attack:

- https://csrc.nist.gov/pubs/ai/100/2/e2023/final

Types of Adversarial Attacks

Poisoning

- Tampering with training and validation datasets
- Produce backdoors
- Parasitic use

Evasion

- Target deployed models
- Facilitate fraud or misclassification
- DoS attacks

Extraction

- target the model's privacy

Inference

- target the model's privacy

Prompt injections

Poisoning attacks - reasons

Bias induction

Backdoor insertion

Disruption

Competitive sabotage

Ransom and extortion

Types of poisoning attacks

- Targeted attacks/Untargeted attacks
- Backdoor attacks
- Clean-label attacks
- Advanced attacks

Defending against poisoning attacks

- AWS SageMaker, MLflow, and Azure Machine Learning offer services and defenses against data poisoning
  - Data versioning and lineage
  - Data validation
  - Model versioning and lineage
  - Continuous monitoring
  - Access control
  - Model interpretability
  - Monitoring, logging, and alerting
  - Governance and collaboration

Defending against poisoning attacks

Anomaly detection

- Identification of suspicious data points
- Automated monitoring
- Reducing false positives

Techniques

- Statistical methods
- Clustering-based methods
- Neural networks
- Density-based methods

Advanced poisoning defenses with ART

- https://adversarial-robustness-toolbox.readthedocs.io/en/latest/modules/defences/detector_poisoning.html
- https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/poisoning_defense_activation_clustering.ipynb
- https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/poisoning_defense_spectral_signatures.ipynb
- https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/provenance_defence.ipynb

Trojan Horses and Model Reprogramming

- Degrading model performance
- Backdoor injection using pickle serialization
- Trojan horse injection with Keras lambda layers
- Trojan horses with custom layers
- Neural payload injection
- Attacking edge AI
- Model hijacking
  - Trojan horse code injection
  - Model reprogramming

Task: supply chain attacks

Individual reading and analysis in github

Evasion attacks

- Designed to mislead Machine Learning (ML) models deliberately
- Reconnaissance | MITRE ATLAS™ (Individual work)
- https://openaccess.thecvf.com/content/CVPR2023W/AML/papers/Sarkar_Robustness_With_Query-Efficient_Adversarial_Attack_Using_Reinforcement_Learning_CVPRW_2023_paper.pdf (Individual work)

Collecting information for such attacks

- Model cards
- Published papers
- Blogs
- Social engineering
- Online probing
- Open source model repositories
- Transfer learning
- Shadow models

Perturbations and image evation attack techniques

- Crafted modifications that cause a model to make incorrect predictions when applied to input data
- The size of perturbation, in other words, the features to alter (L1 norm)
- Its closeness – or Euclidean distance – to the original sample (L2 norm)
- The maximum change to any feature in the data (infinity norm, or L∞)

Perturbations and image evation attack techniques

- Fast Gradient Sign Method (FGSM)
- Basic Iterative Method (BIM)
- Projected Gradient Descent (PGD)
- Carlini and Wagner (C&W) attack
- Jacobian-based Saliency Map Attack (JSMA).

- NLP evasion attacks
  - Change words or phrases in a text snippet
    - classification from positive to negative sentiment
    - non-spam to spam
  - Natural Language Inference (NLI)
- Universal Adversarial Perturbations (UAPs)
- Black-box attacks with transferability

Defending against evasion attacks

- Reactive measures
  - Respond to attacks as they happen
  - Real-time detection and mitigation
- Proactive measures
  - Hardening models

Defending against evasion attacks

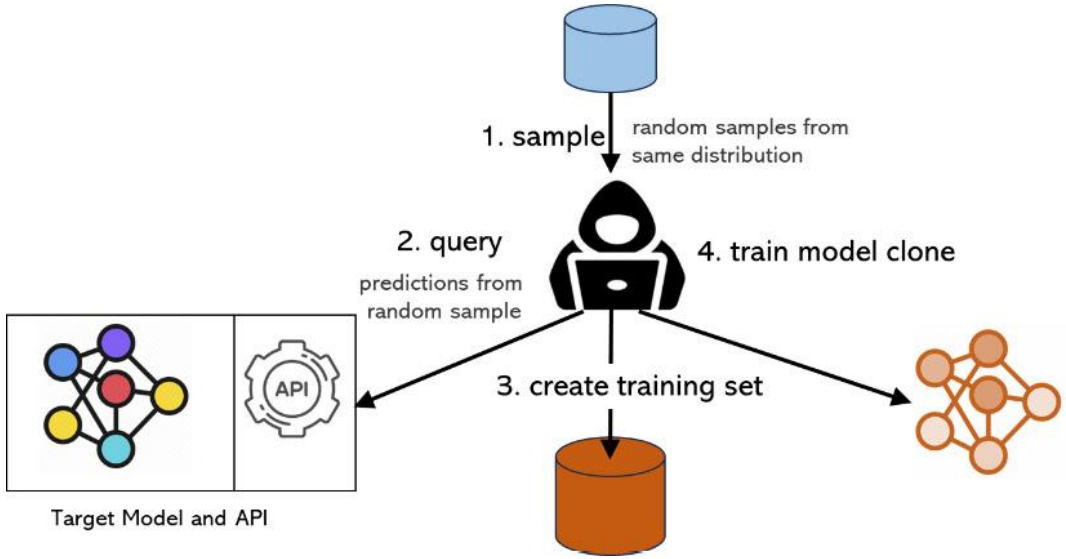
- Adversarial training
  - https://adversarial-robustness-toolbox.readthedocs.io/en/latest/modules/defences/trainer.html
- Input preprocessing
- Model hardening techniques
  - Defensive distillation
  - Feature squeezin
  - Gradient masking
  - Robust loss functions
- Model ensembles
- Certified defenses

Privacy Attacks – Stealing Models

- Intentionally manipulate AI models to extract sensitive information
- Privacy attacks do not seek to alter the model in any way
- Focus on model confidentiality and extracting sensitive information
- Model extraction
- Model inversion
- Membership inference

Extraction attacks

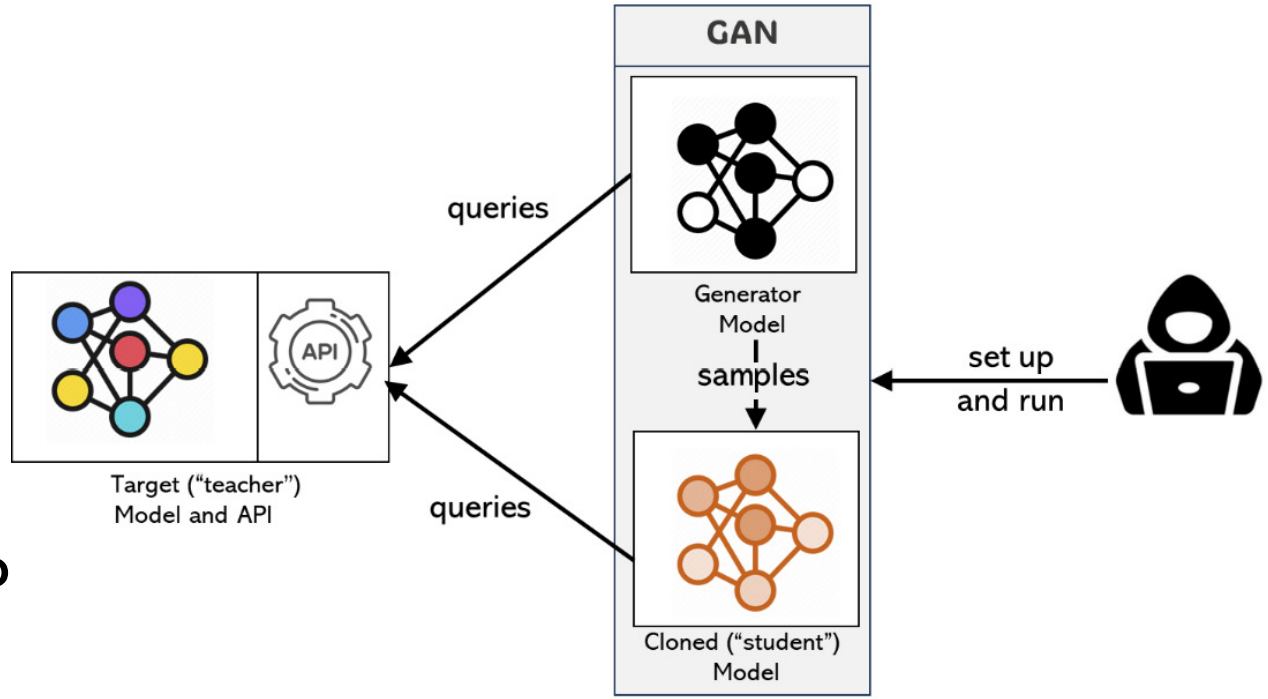
- Functionally equivalent extraction
  - https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/art/attacks/extraction/functionally_equivalent_extraction.py
  - https://arxiv.org/abs/1909.01838
- Learning-based model extraction attacks
  - Copycat CNN
  - KnockOff Nets



John Sotiropoulos. Adversarial AI Attacks, Mitigations, and Defense Strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps, 2024

Extraction attacks

- Generative student-teacher extraction (distillation) attacks
- https://www.researchgate.net/publication/340542875_Private_Knowledge_Transfer_via_Model_Distillation_with_Generative_Adversarial_Networks



John Sotiropoulos. Adversarial AI Attacks, Mitigations, and Defense Strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps, 2024

Defense

- Combat reconnaissance
- Strict model governance with MLOps
- Least-privilege access to production systems
- Gated API pattern
- Authentication
- Input pre-processing at inference time
- Output perturbation

Detection

- Incorporating tests against known extraction attacks
- Regular red-team testing of models
- Rate limiting
- System monitoring and alerting
- Model and query monitoring

- Unique model identifiers
- Watermarking

Stealing data

- Model inversion
  - Reconstruct training data or sensitive information
  - White box model inversion attacks
  - Black box model inversion attacks
  - Techniques:
    - Exploitation of model confidence scores
    - GAN-assisted model inversion (https://github.com/AI-secure/GMI-Attack)
- Inference attacks
  - Attribute or property inference attacks
  - Membership inference attacks

Inference attacks

- Attribute inference attacks
  - Meta-classifiers
  - https://github.com/Trusted-AI/adversarial-robustness-toolbox/tree/main/art/attacks/inference/attribute_inference
  - Poisoning-assisted inference

Membership inference attacks

- Statistical thresholds for ML leaks
- Label-only data transferring attack
- Blind membership inference attacks
- White box attacks

- Mitigations:
  - Regularization
  - Data augmentation
  - Model stacking
  - Data minimization and anonymization
  - Differential privacy
  - Membership inference adversarial training
  - MemGuard

Privacy preserving AI

- Simple data anonymization
- Advanced anonymization
  - K-anonimity
  - Anonymization and geolocation data (Geographic masking, Spatial aggregation, Geocoding)
  - Anonymizing rich media
    - Blurring
    - Voice alteration
    - Background noise addition
    - Speech-to-text and text-to-speech synthetization
- Differential privacy (DP) - https://arxiv.org/abs/2303.00654.
- Federated learning (FL)
- Split learning

Privacy preserving AI

- Secure multi-party computation (secure MPC)
- Homomorphic encryption


@drfakenstein


sky news
.COM.AU
First Edition
DEEPFAKES OF TRUMP ARREST EMERGE ON TWITTER
REAL NEWS, HONEST VIEWS.

Generative AI and Adversarial attacks

- GANs for deepfakes and deepfake detection
- Generate images using existign images
- Changing images directly
- Fake videos and animations

GANs in cyberattacks

- Evading face verification
  - FaceNet and DeepFace algorithms
  - StarGAN v2
  - StyleGAN
- Compromising biometric authentication
  - https://github.com/wy1iu/sphereface
  - DeepMasterPrints
- Password cracking with GANs
- Malware detection evasion

GANs in cyberattacks

- Malware detection evasion
- GANs in cryptography and stenography
  - CipherGAN: https://arxiv.org/abs/1801.04883.
  - Unified Cipher Generative Adversarial Network (UC-GAN) - https://github.com/tdn02007/UC-GAN-Unified-cipher-generative-adversarial-network.
  - Stegano-GAN
- Generating web attack payloads with GANs
- Generating adversarial attack payloads
  - Generative adversarial perturbations (GAP)
  - AdvGAN
  - AdvGan++
  - GAP
  - GAP++
  - Attack-Inspired GAN (AI-GAN)

How to mitigate?

- Securing GANs
- Standard defense-in-depth security
- Defense-GAN
- https://github.com/Trevillie/MagNet.
- Privacy-Preserving Representation-Learning – Variational GAN (PPRL VGAN) - https://github.com/yushuinanrong/PPRL-VGAN

LLMs and Adversarial AI

- Fine-tuning and RAG are targets
- Traditional attacks are more difficult
- Supply chain attacks are easier
- Data security

- Adversarial inputs and prompt injection
  - Direct prompt injection
  - Prompt override
  - Style injection
  - Role-playing
  - Impersonation
  - Language switch
  - Adding constraints
  - Encoding
  - Printing insecure output

LLMs and Adversarial AI

- Automated gradient-based prompt injection
- Indirect prompt injection
- Data exfiltration with prompt injection
- Privilege escalation with prompt injection
- RCE with prompt injection

Defence

- Content filtering
- Data privacy
- Ethical guidelines
- Bias mitigation
- Limitations on certain topics
- User interaction monitoring
- Regular updates and improvements
- Secure application design and coding
- Secure application platform