

Adversarial Machine Learning

Review on related works

이규영 강사

2022년 7월

기존 출간논문 조사결과

- 기존 출간논문 조사

- 학회 : ACM / IEEE S&P
- 분야 : Adversarial Learning 관련 출간논문
- 방법 : 논문 별 Abstract 분석

- 기존 출간논문 통계



순번	학회	합계	공격 연구	방어 연구	공격&방어 연구
1	ACM	26	14	12	0
2	IEEE S&P	29	11	15	3

기존 출간논문 분석결과

- ACM 및 IEEE S&P 총 55개 논문의 Abstract를 분석한 결과, 모두 Adversarial 공격 및 방어 기술에 관한 논문이었음.
- 공격측면
 - 1) 새로운 공격기술 제안
 - 2) 공격성공율을 높이는 기법 제시
- 방어측면
 - 1) 다양한 공격을 방어할 수 있는 기법
 - 2) 공격법에 의존하지 않고 전반적 Robustness를 향상시키는 기법
 - 3) 방어성공율을 높이는 기법

(별첨1) ACM 논문 (Adversarial Learning)

- Applied Filters : Adversarial Learning



[Browse](#) [About](#) [Sign in](#)

[Journals](#) [Magazines](#) [Proceedings](#) [Books](#) [SIGs](#) [Conferences](#) [People](#)

Search ACM Digital Library

Applied Filters

Adversarial Learning ✕

Clear All

People

Names ▾

Institutions ▾

Authors ▾

Publications

Journal/Magazine Names ▾

Proceedings/Book Names ▾

All Publications ▾

48 Results [Edit Search](#) [Save Search](#) [RSS](#)

Searched The ACM Full-Text Collection (650,322 records) | [Expand your search to The ACM Guide to Computing Literature \(3,224,496 records\)](#)

RESULTS

VIDEOS

PEOPLE

Showing 1 - 20 of 48 Results


☐ Select All

per page: 10 20 50 | Relevance ▾

☐ RESEARCH-ARTICLE

FREE

April 2022



Enhancing Boundary Attack in Adversarial Image Using Square Random Constraint


[Tran Van Sang](#), [Tran Phuong Thao](#), [Rie Shigetomi Yamaguchi](#), [Toshiyuki Nakata](#)


IWSPA '22: Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics • April 2022, pp 13–23 • <https://doi.org/10.1145/3510548.3519373>


An adversarial image is a sample with intentional small perturbations that causes deep learning models to classify the image incorrectly. In the image recognition field, adversarial images have become an attractive research topic because they can ...


0


28












☐ RESEARCH-ARTICLE

February 2022



Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning

[Yawen Zeng](#), [Da Cao](#), [Shaofei Lu](#), [Hanling Zhang](#), [Jiao Xu](#), [Zheng Qin](#)

[Feedback](#)

ACM 논문 (Adversarial Learning)

순번	제목	주제
1	Enhancing Boundary Attack in Adversarial Image Using Square Random Constraint (IWSPA '22: Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics) → Boundary Attack 공격기법 개선	공격기법 개선
2	Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning (ACM Transactions on Multimedia Computing, Communications, and Applications, May 2022) → 적대적 학습 하에서 언어기반 VMR 문제점 개선	방어기법 개선
3	ERGA: An Effective Region Gradient Algorithm for Adversarial Example Generation (EITCE 2021: Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering) → ERGA기법을 이용하여 Adversarial Example 생성기법을 개선	공격기법 개선

ACM 논문 (Adversarial Learning)

순번	제목	주제
4	Adversarial Transfer Attacks With Unknown Data and Class Overlap (AISeC '21: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security) ➔ Adversarial Transfer Attack 개선	공격기법 개선
5	Subpopulation Data Poisoning Attacks (CCS '21: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security) ➔ Poisoning Attack 개선	공격기법 개선
6	Machine Learning Explainability and Robustness: Connected at the Hip (KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining) ➔ Adversarial Attack에 대한 Robustness 개선	방어기법 개선
7	Indirect Invisible Poisoning Attacks on Domain Adaptation (KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining) ➔ 비지도학습 모델에 대한 Poisoning Attack 연구	방어기법 개선

ACM 논문 (Adversarial Learning)

순번	제목	주제
8	A Study of Defensive Methods to Protect Visual Recommendation Against Adversarial Manipulation of Images (SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval) → VRSs 시스템에 대한 적대적 공격 방어기법 연구	방어기법 개선
9	Using Single-Step Adversarial Training to Defend Iterative Adversarial Examples (CODASPY '21: Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy) → Adversarial Training 기법에 대한 개선	방어기법 개선
10	SCRAP: Synthetically Composed Replay Attacks vs. Adversarial Machine Learning Attacks against Mouse-based Biometric Authentication (AISec'20: Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security) → 신규 Adversarial Attack 기술 개발	공격기법 개선

ACM 논문 (Adversarial Learning)

순번	제목	주제
11	Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization (MM '20: Proceedings of the 28th ACM International Conference on Multimedia) → Adversarial Training 기법의 보완	방어기법 개선
12	Adversarial Attack against Deep Reinforcement Learning with Static Reward Impact Map (ASIA CCS '20: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security) → 강화학습 모델에 대한 Adversarial Attack 기술의 개선	공격기법 개선
13	Adversarial machine learning for spam filters (ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security) → 머신러닝 기반 스팸필터에 대한 Adversarial Attack 기술 개선	공격기법 개선
14	Attackability Characterization of Adversarial Evasion Attack on Discrete Data (KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining) → Evasion Attack 기술에 대한 개선	공격기법 개선

ACM 논문 (Adversarial Learning)

순번	제목	주제
15	Adversarial Attacks and Detection on Reinforcement Learning-Based Interactive Recommender Systems (SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval) → 강화학습에 대한 Adversarial Attack에 대한 탐지기법 연구	방어기법 개선
16	Generalized wireless adversarial deep learning (WiseML '20: Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning) → 무선 Adversarial attacks 연구	공격기법 연구
17	Deep Adversarial Discrete Hashing for Cross-Modal Retrieval (ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval) → DADH Adversarial Training 기법 연구	방어기법 연구
18	Snooping Attacks on Deep Reinforcement Learning (AAMAS '20: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems) → 강화학습을 공격하는 snooping threat models 공격기법 제안	공격기법 연구

ACM 논문 (Adversarial Learning)

순번	제목	주제
19	CopyCAT:: Taking Control of Neural Policies with Constant Attacks (AAMAS '20: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems) → 강화학습에 대한 Adversarial attacks 연구	공격기법 연구
20	Impacts of adversarial inputs in associative memory models and its iterative learning variants (AISS '19: Proceedings of the International Conference on Advanced Information Science and System) → Associative memory model에 대한 adversarial attacks에 관한 연구	공격기법 연구
21	Poster: Adversarial Examples for Hate Speech Classifiers (CCS '19: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security) → Hate Speech Classifiers에 대한 Adversarial attacks 연구	공격기법 연구
22	Adversarial Factorization Autoencoder for Look-alike Modeling (CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge Management) Adversarial Factorization Autoencoder 방어기법 제안	방어기법 연구

ACM 논문 (Adversarial Learning)

순번	제목	주제
23	MetaAdvDet: Towards Robust Detection of Evolving Adversarial Attacks (MM '19: Proceedings of the 27th ACM International Conference on Multimedia) → 새로운 적대적공격을 소수의 샘플로 감지하는 meta-learning based robust detection method 제안	방어기법 제안
24	Mitigating Unwanted Biases with Adversarial Learning (AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society) → undesired biases를 훈련데이터에서 배제하도록 하는 순화기법	방어기법 제안
25	Adversarial Classification on Social Networks (AAMAS '18: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems) → Evasion attack 방어기법 by modeling Stackelberg game	방어기법 제안
26	Poster: Adversarial Examples for Classifiers in High-Dimensional Network Data (CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security) → Adversarial examples 생성기법을 개선하여 ML시스템의 robustness 측정	공격기법 개선

(별첨2) IEEE S&P 논문

- Applied Filters : ("Document Title":adversarial) AND ("All Metadata":learning) NOT ("All Metadata":generative) NOT ("All Metadata":gan) NOT ("All Metadata":reinforcement) AND ("Publication Title":Security & Privacy)

The screenshot displays the IEEE Xplore search interface. At the top, the navigation bar includes links to IEEE.org, IEEE Xplore, IEEE SA, IEEE Spectrum, and More Sites, along with options to SUBSCRIBE, view the Cart, Create Account, and Personal Sign In. The IEEE Xplore logo is highlighted with a red box. Below the navigation bar, there are links for Browse, My Settings, and Help, and a blue button for Institutional Sign In. The IEEE logo is also present on the right. A search bar with a dropdown menu set to 'All' and a search icon is shown, with a link to ADVANCED SEARCH below it. Below the search bar, there is a section for 'Search within results' with a search icon, and options for Per Page (25), Export, Set Search Alerts, and Search History. A red box highlights the search results summary: 'Showing 1-25 of 36 for ("Document Title":adversarial) AND ("All Metadata":learning) NOT ("All Metadata":generative) NOT ("All Metadata":gan) NOT ("All Metadata":reinforcement) AND ("Publication Title":Security & Privacy) x'. Below this, there are checkboxes for Conferences (33) and Magazines (3). On the left, there is a 'Show' section with radio buttons for 'All Results' (selected) and 'Open Access Only'. On the right, there is a 'Sort By: Relevance' dropdown menu. Below the search results, there is a list of results, with the first one being 'Machine Learning in Adversarial Settings' by Patrick McDaniel; Nicolas Papernot; Z. Berkay Celik, published in IEEE Security & Privacy. A red box highlights the title and authors of this result. On the right side of the page, there is a blue banner with the text 'Need Full-Text?' and a red button labeled 'Feedback'.

IEEE.org | IEEE Xplore | IEEE SA | IEEE Spectrum | More Sites | SUBSCRIBE | Cart | Create Account | Personal Sign In

IEEE Xplore® | Browse ▾ | My Settings ▾ | Help ▾ | Institutional Sign In | IEEE

All ▾ | Q | ADVANCED SEARCH

Search within results | Q | Per Page: 25 ▾ | Export ▾ | Set Search Alerts | Search History

Showing 1-25 of 36 for ("Document Title":adversarial) AND ("All Metadata":learning) NOT ("All Metadata":generative) NOT ("All Metadata":gan) NOT ("All Metadata":reinforcement) AND ("Publication Title":Security & Privacy) x

☐ Conferences (33) ☐ Magazines (3)

☐ Select All on Page | Sort By: Relevance ▾

Show

☒ All Results ☐ Open Access Only

☐ Machine Learning in Adversarial Settings
Patrick McDaniel; Nicolas Papernot; Z. Berkay Celik
IEEE Security & Privacy

Need Full-Text? Feedback

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
1	Machine Learning in Adversarial Settings (IEEE Security & Privacy Volume: 14, Issue: 3, May-June 2016) ➔ Adversarial attack 공격 및 완화기법 리뷰	공격 및 방어 기법 리뷰
2	Quantum Adversarial Machine Learning: Status, Challenges and Perspectives (2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications) ➔ Quantum Adversarial Machine Learning 방어기법	방어기법 연구
3	From Blue-Sky to Practical Adversarial Learning 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) ➔ Adversarial attacks 감지시스템 개선	방어기법 연구
4	Adversarial Deception in Deep Learning: Analysis and Mitigation (2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications) ➔ Adversarial attacks & defense techniques 연구	공격 및 방어기 법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
5	Adversarial Machine Learning-Industry Perspectives (2020 IEEE Security and Privacy Workshops) → ML개발자와 보안대응팀의 두가지 관점에서 적대적공격을 고찰	공격 및 방어 연구
6	Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers (2018 IEEE Security and Privacy Workshops) → 텍스트 데이터를 변조한 DeepWordBug 적대적공격 기법을 제안	공격기법 연구
7	Adversarial Deep Learning for Robust Detection of Binary Encoded Malware (2018 IEEE Security and Privacy Workshops) → Adversarial malware examples를 제작하여 딥러닝 탐지시스템을 적대적 공격에 강인하게 만드는 기법을 연구	방어기법 연구
8	Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion (2019 IEEE Symposium on Security and Privacy) → The importance of the research on real-world adversarial learning	공격기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
9	Safe Machine Learning and Defeating Adversarial Attacks (IEEE Security & Privacy (Volume: 17, Issue: 2, March-April 2019)) → This article discusses recent research for ML model assurance in the face of adversarial attacks.	방어기법 연구
10	Selective Adversarial Learning for Mobile Malware (2019 18th IEEE International Conference On Trust) → Adversarial retraining 시 학습할 적대적 샘플을 랜덤하게 투입하는 것이 아니라 선별해서 투입하며, 그 선별하는 방법을 제안함	방어기법 연구
11	The Limitations of Deep Learning in Adversarial Settings (2016 IEEE European Symposium on Security and Privacy) → This paper formalize the space of adversaries against deep neural networks (DNNs) and introduce a novel class of algorithms to craft adversarial samples	공격기법 연구
12	Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks (2016 IEEE Symposium on Security and Privacy) → Defensive Distillation을 적대적 공격에 대한 방어기법으로서 제안함	방어기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
13	Certified Robustness to Adversarial Examples with Differential Privacy (2019 IEEE Symposium on Security and Privacy) → The defense, called PixelDP, is based on a novel connection between robustness against adversarial examples and differential privacy	방어기법 연구
14	Poltergeist: Acoustic Adversarial Machine Learning against Cameras and Computer Vision (2021 IEEE Symposium on Security and Privacy) → This paper models the feasibility of such acoustic manipulation and design an attack framework that can accomplish three types of attacks	공격기법 연구
15	Detecting Adversarial Examples in Learning-Enabled Cyber-Physical Systems using Variational Autoencoder for Regression (2020 IEEE Security and Privacy Workshops) → The proposed approach is based on inductive conformal prediction and uses a regression model based on variational autoencoder	방어기법 연구
16	Exploring Adversarial Examples in Malware Detection (2019 IEEE Security and Privacy Workshops) → This paper explores the area of adversarial examples for malware detection.	방어기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
17	Progressive Defense Against Adversarial Attacks for Deep Learning-as-a-Service in Internet of Things (2021 IEEE 20th International Conference on Trust) ➔ This paper presents a defense strategy called a progressive defense against adversarial attacks (PDAAA) for efficiently and effectively filtering out the adversarial pixel mutations	방어기법 연구
18	Attack versus Attack: Toward Adversarial Example Defend Website Fingerprinting Attack (2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications) ➔ This paper proposes an defense named Attack to Attack (A2A) that leverages adversarial example to attack the attacker's classifier. A2A treats website fingerprinting model as a black box.	공격기법 연구
19	Towards a Robust Classifier: An MDL-Based Method for Generating Adversarial Examples (2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications) ➔ This paper proposes a method to generate adversarial examples using the minimum description length (MDL) principle. Our final aim is to improve the robustness of the classifier by considering generated examples in rebuilding the classifier.	공격기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
20	Towards Query-efficient Black-box Adversarial Attack on Text Classification Models (2021 18th International Conference on Privacy, Security and Trust) → This paper proposes a Query-efficient Black-box Adversarial Attack on text data that tries to attack a textual deep neural network by considering the amount of overhead that it may produce.	공격기법 연구
21	Towards Understanding Limitations of Pixel Discretization Against Adversarial Attacks (2019 IEEE European Symposium on Security and Privacy) → This paper studies the pixel discretization defense method, including more sophisticated variants that take into account the properties of the dataset being discretized.	방어기법 연구
22	Universal Website Fingerprinting Defense Based on Adversarial Examples (2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications) → The proposed defense is evaluated on state-of-the-art DNN attack over a public Tor traffic dataset.	방어기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
23	DLA: Dense-Layer-Analysis for Adversarial Example Detection (2020 IEEE European Symposium on Security and Privacy) → This paper presents a novel end-to-end framework to detect such attacks without influencing the target model's performance showing that dense layers of DNNs carry security-sensitive information.	방어기법 연구
24	Adversarial Attacks on Time-Series Intrusion Detection for Industrial Control Systems (2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications) → This paper addresses the domain specific challenges of constructing the attacks against autoregressive based intrusion detection systems (IDS) in an industrial control systems(ICS) setting.	공격기법 연구
25	Minimum-Norm Adversarial Examples on KNN and KNN based Models (2020 IEEE Security and Privacy Workshops) → This paper studies the robustness against adversarial examples of kNN classifiers and classifiers that combine kNN with neural networks.	방어기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
26	Background Class Defense Against Adversarial Examples (2018 IEEE Security and Privacy Workshops) → This paper proposes a defense of expanding the training set with a single, large, and diverse class of background images	방어기법 연구
27	Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems (2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications) → This paper presents a suite of adversarial objectness gradient attacks	공격기법 연구
28	GlassMasq: Adversarial Examples Masquerading in Face Identification Systems with Feature Extractor (2019 17th International Conference on Privacy, Security and Trust) → To obtain adversarial examples with high confidence and small perturbation, this paper first introduces a condition which adversarial examples against the face identification systems should satisfy, then introduce a new method called GlassMasq to create adversarial examples based on the condition.	공격기법 연구

IEEE S&P 논문 (Adversarial Learning)

순번	제목	주제
29	Intriguing Properties of Adversarial ML Attacks in the Problem Space (2020 IEEE Symposium on Security and Privacy) ➔ This paper makes two major contributions. First, they propose a novel formalization for adversarial ML evasion attacks in the problem-space. Second, they propose a novel problem-space attack on Android malware that overcomes past limitations	공격기법 연구