Yaoteng TAN

Tel: (951)-367-7913 Email: yaoteng.tan@email.ucr.edu Homepage: https://ytengtan.github.io/

EDUCATION

University of California, Riverside (On going) Sep. 2022 - Present
Ph.D. student in Electrical Computer Engineering (ECE) Riverside, CA
Advised by Dr. M. Salman Asif

Huazhong University of Science and Technology Sep. 2018 - Jun. 2022
B.S. in Electrical Engineering Wuhan, China
Graduated with the honors.

EXPERIENCE

ECE Department, University of California, Riverside Oct. 2022 - Present
Graduate research assistant Riverside, CA

- Research topics involve trustworthy machine learning and inverse problems.
- Teaching assistant for graduate and undergraduate courses.
- Student leader of UCR student org. *Happy English Corner*.

CloudMinds Technology Inc. Jul. 2021 - Aug. 2021
Research intern Beijing, China

- Project summary: Service Robot vision based on YOLOv5.
- Implement object detection for the service robot and deploy it on the UE4-based simulation platform.

RESEARCH INTEREST

My research focuses primarily on trustworthy AI, with a special emphasis on adversarial attacks and machine unlearning, contributing to the ethical deployment of AI systems. I am also interested in inverse problems, exploring efficient optimization techniques, and better priors for general image reconstruction from limited measurements. While these are my core interests, I remain open to various research questions across computer vision and find interest in other topics.

PUBLICATION

Ensemble-based Blackbox Attacks on Dense Prediction, *CVPR 2023*.
*Zikui Cai, ***Yaoteng Tan**, M. Salman Asif. (*Equal contribution)
We propose an ensemble method for generating effective blackbox adversarial attacks against models for dense prediction vision tasks (e.g., object detection and semantic segmentation). Our method can generate a single perturbation that is effective in fooling multiple blackbox detection and segmentation models simultaneously in a target-consistent manner.

PROJECT

Machine unlearning on foundation models, $Jun.2024 - Present$.
Large multi-modal generative models, such as Stable Diffusion and Vision-Language models, are widely used nowadays. However, these models have the risk of generating privacy-related and copyright-infringing content. As there are many unlearning techniques proposed for removing sensitive content (e.g., celebrity identities, intellectual property concepts) from these large models, efficient unlearning techniques are still needed. We proposed an efficient unlearning technique that requires only one-time gradient computation and achieves effective unlearning.

(Preprint: https://arxiv.org/abs/2407.11867)

Transform-dependent adversarial attacks, *Sep.2023 – Mar.2024*.
Many properties of adversarial attacks are well-studied today (e.g., optimization, transferability, physical implementation, etc.). In this work, we explore an under-researched property of adversarial attacks — transform-dependent, which the optimization process of additive adversarial perturbations can be combined with various image transformations to produce versatile, transform-dependent attack effects. We demonstrate that the transform-dependent property of attacks holds for many differentiable image transformations. Additionally, we show that this property can be leveraged to attack object detection systems, with implications for real-world applications.

(Preprint: https://arxiv.org/abs/2406.08443)

Blackbox adversarial attacks against dense prediction models, *Sep. – Dec.2022*.
We propose an ensemble method for generating effective blackbox adversarial attacks against models for dense prediction vision tasks, such as object detection and semantic segmentation. Our ensemble method can generates a single perturbation that is effective for fooling multiple blackbox detection and segmentation models simultaneously in a target-consistent manner. The results of this research project are published in the proceeding of CVPR 2023.

Adversarial robustness study of data-driven computational imaging systems, *Feb. – Jun.2022*.
Data-driven methods such as deep learning are widely used in computational imaging systems and outperform traditional model-based methods on many benchmarks. However, deep models are well-known to be vulnerable to adversarial attacks. Our hypothesis is an imaging pipeline that involves data-driven modules is more vulnerable than model-based methods, in which imaging processes are explicitly modeled.

SERVICE

Conference reviewer

- WACV, 2024

Teaching assistant

- EE240 Pattern Recognition, UC Riverside, Spring 2023, Spring 2024
  Duties: holding office hours, grading students' homework and final projects.
- CS171/EE142 Introduction to Machine Learning and Data Mining, UC Riverside, Fall 2023
  Duties: short teaching in discussion sections, holding office hours, designing exam questions, gradings.

SKILLS

Technical

- Adversarial machine learning, machine unlearning, convex optimization, inverse problem, image (signal) processing.
- Programming language: Python, MatLab, Java, familiar with C-like, HTML.
- Familiar with basic source control tools (e.g., Git), writing tools (e.g., LaTeX)

Language

- English, Mandarin

General

- Teaching, technical writing