ICRA2024
May 13-17, Yokohama, Japan

Boosting adversarial training in safety-critical systems through boundary data selection

Yifan Jia, Christopher M. Poskitt, **Peixin Zhang**, Jingyi Wang, Jun Sun, and Sudipta Chattopadhyay

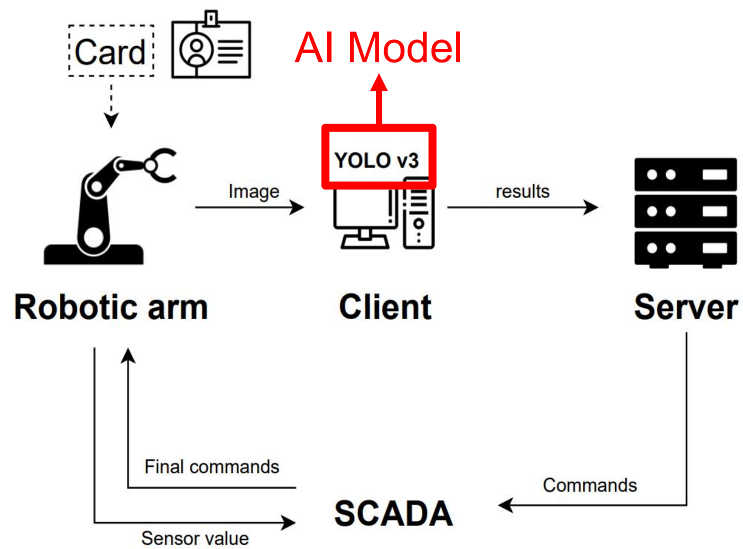
ZHEJIANG UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Background

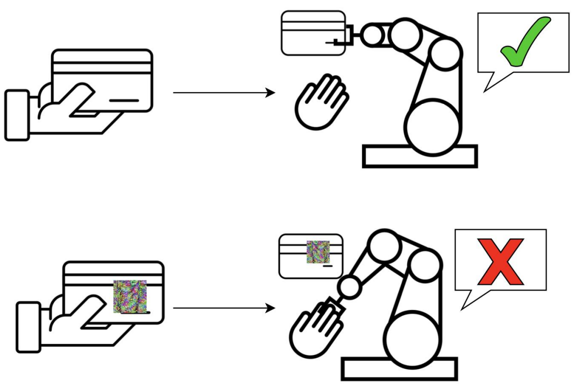
- AI models are increasingly integrated into robotics



Picture from: https://rsl.ethz.ch/robots-media/dynaarm.html

Background

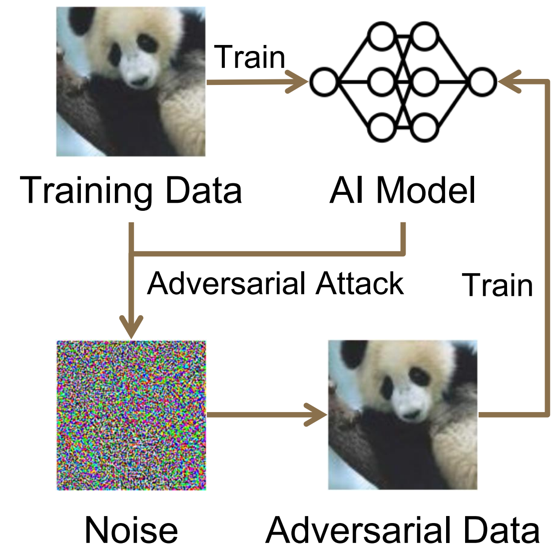
- AI models are vulnerable to adversarial examples[1]



1. Y. Jia, C. M. Poskitt, J. Sun, and S. Chattopadhyay. “Physical Adversarial Attack on a Robotic Arm,” IEEE Robotics and Automation Letters, vol. 7 p. 9334—9341, 2022.

Background

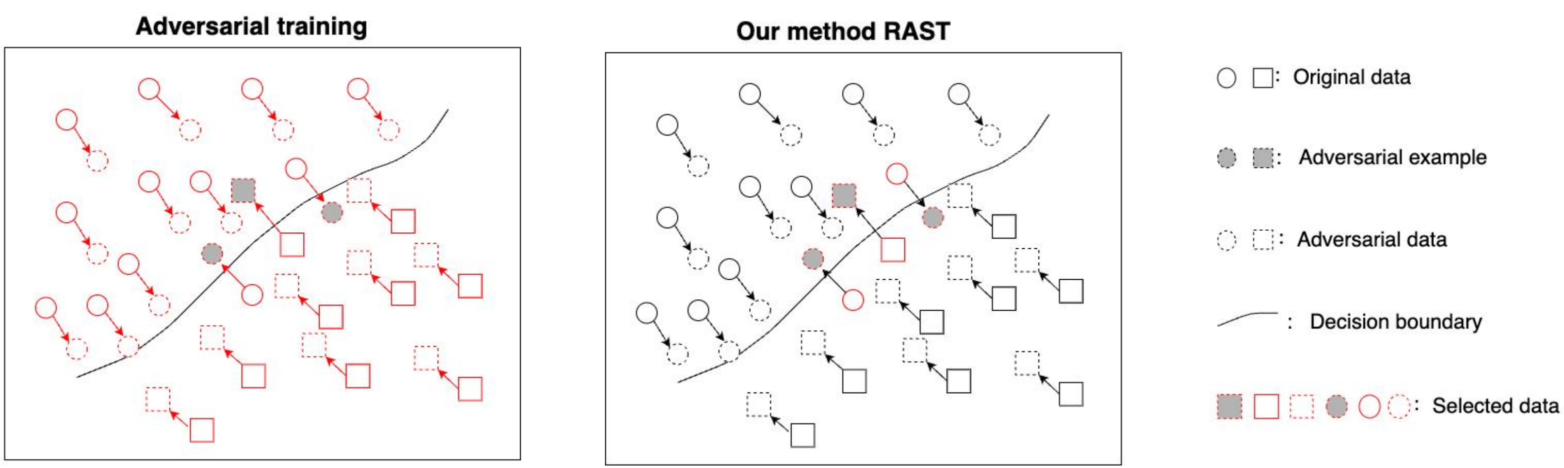
- One improvement method is adversarial training
  - Training with a mixture of adversarial and clean data to improve the model robustness
  - Adversarial data can be pre-generated or generated during training
- However, it is costly and time-consuming
- Data labeling is also costly and time-consuming



How to improve the robustness of the model with less cost while maintaining accuracy?

RAST

- Intuition
  - Boundary data has a significant impact on the training results (model)
- Methodology
  - Select those boundary data and their adversarial examples to form training data for adversarial training



Empirical Selection

- We select the data x that:

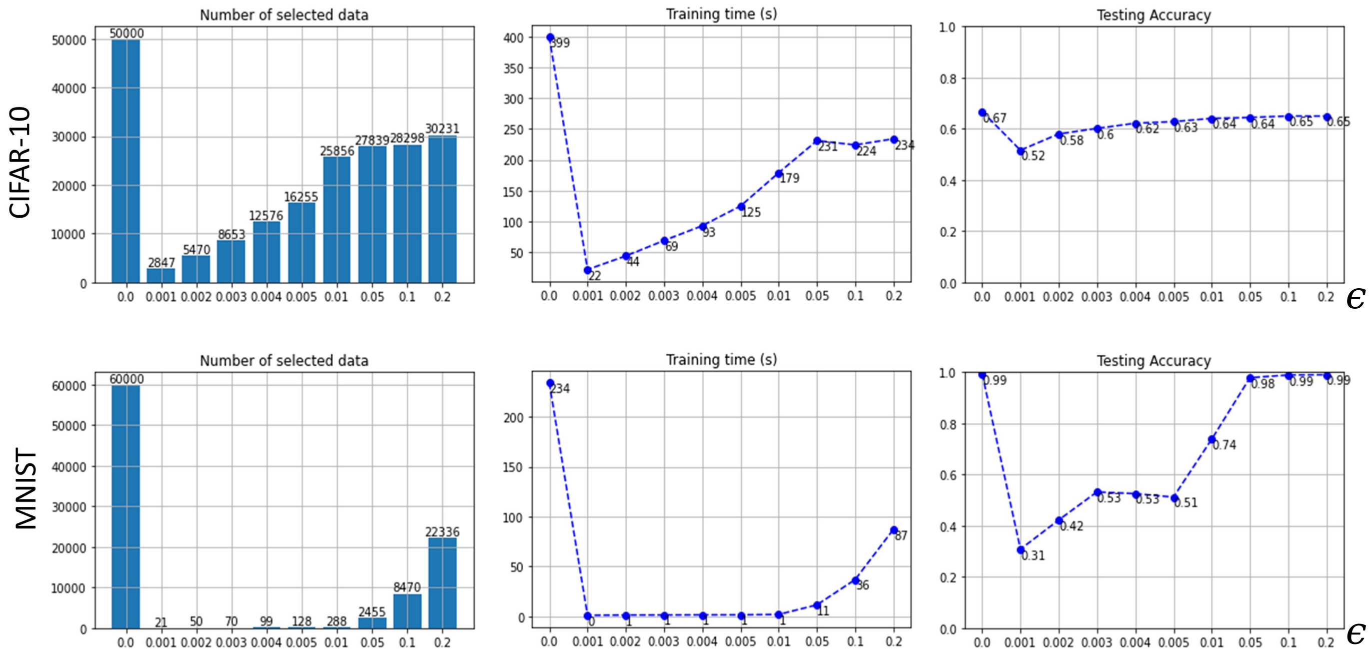
$$x \in D \ s.t. \ f(x_{adv}) \neq f(x)$$

$$\text{where } \mathrm{x_{adv}} = \mathrm{x} + \epsilon \cdot \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{J}(x, \mathrm{y}^*))$$

  - x: Selected data
  - D: Clean dataset
  - x_{adv}: Adversarial example
  - ϵ: Perturbation step size
  - $sign(\nabla_x)$: Sign of the gradient
  - $J(x, y^*)$: Loss function
  - y^*: Ground-truth label

Evaluation

- Experiments on common image datasets



RAST can successfully reduce the training time to more than half while remain a similar accuracy.

Evaluation

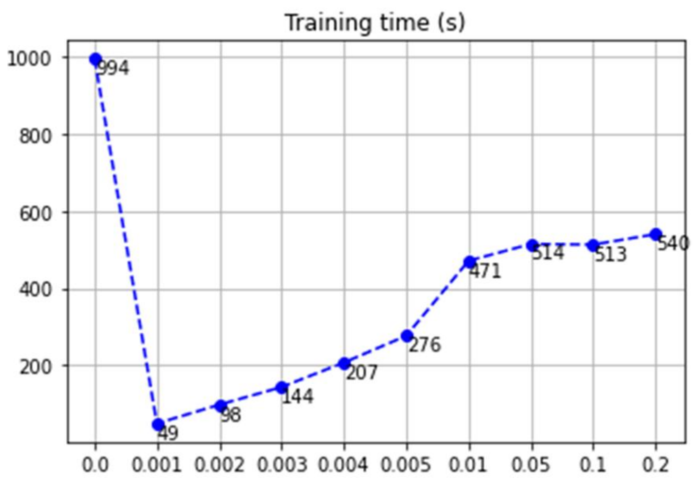
- Compare with existing adversarial training

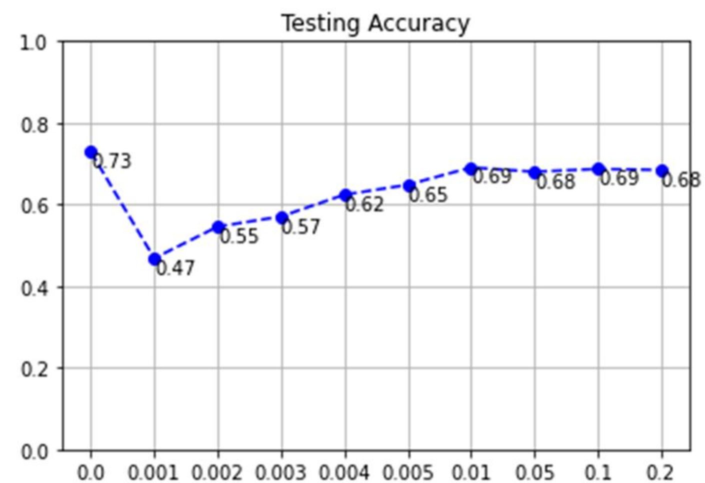
Method	Time (s)	Test Acc	FGSM Robustness	PGD Robustness
PGD-AT	149.00	**0.83**	0.59	0.56
FGSM-AT	86.48	0.79	**0.72**	0.74
RAST-AT	**68.52**	0.79	**0.72**	**0.75**

RAST is faster than existing adversarial training methods while providing similar or better performance on accuracy and robustness.

Evaluation

- Experiment with different model architecture

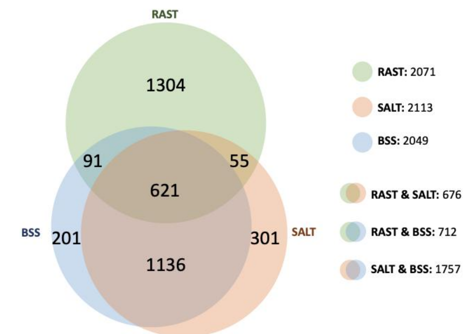




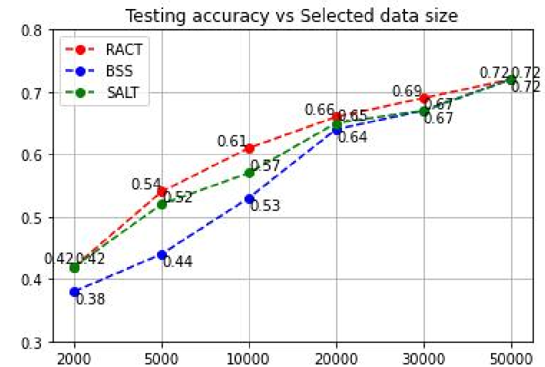
The selected boundary data works for a different model with the same task.

Evaluation

- Baselines
  - BSS[1]: P_{max} / $P_{2nd\text{-}max}$ < 7
  - SALT[2]: $|f(x+0.01) - f(x)| > 0.44$
  - RAST: $\epsilon = 0.01$



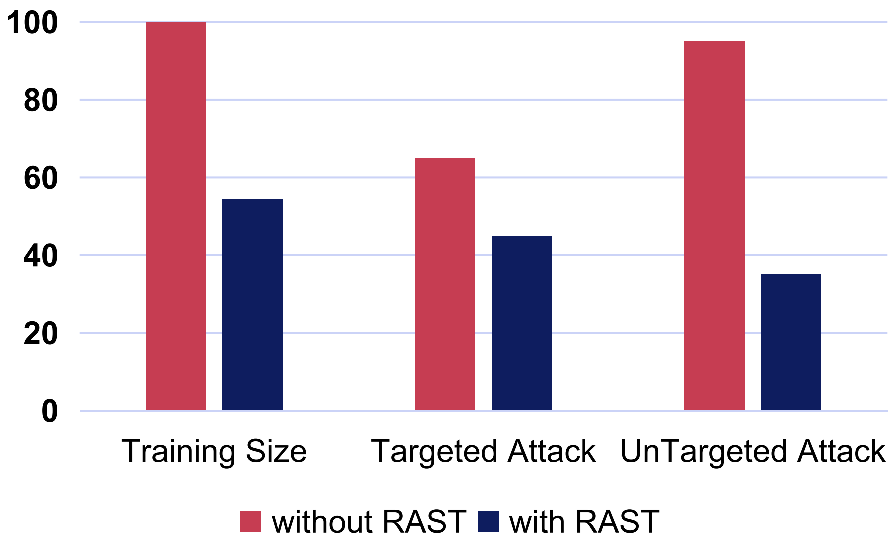
- Conclusion
  - Significant overlap between the BSS and SALT selections, While RAST selects an amount of different data
  - RAST has the best testing accuracy

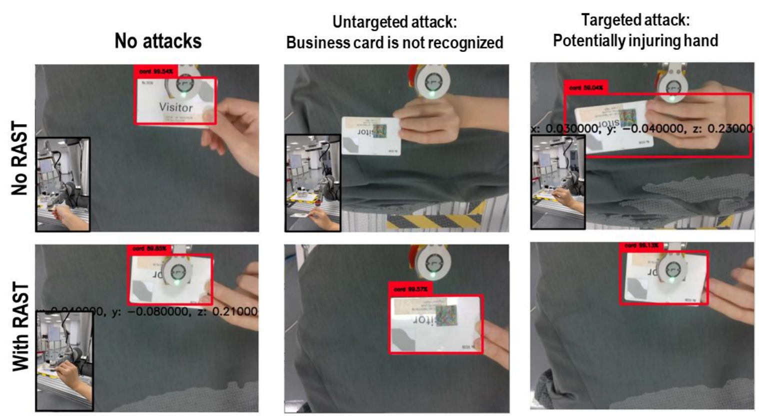


1. BSS: Boundary Sample Selection:W. Shen, Y. Li, Y. Han, L. Chen, D. Wu, Y. Zhou, and B. Xu, “Boundary sampling to boost mutation testing for deep learning models,” *Information and Software Technology*, vol. 130, p. 106413, 2021.
2. SALT: Adversarial Active Learning: B. Miller, A. Kantchelian, S. Afroz, R. Bachwani, E. Dauber, L. Huang, M. C. Tschantz, A. D. Joseph, and J. D. Tygar, “Adversarial active learning,” in *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, 2014, pp. 3–14.

Practical Evaluation

- Defense against adversarial samples on AI models of robotic arms





RAST can reduce the training size and improve the robustness of AI models of robotic arms.

Demo

- Defense against adversarial samples on AI models of robotic arms

Contribution

- We propose an adversarial training method based on boundary data to improve model (and model-based robotic system) robustness more efficiently while maintaining model accuracy.
- We propose an attack-based boundary data selection that can effectively filter out boundary samples.
- We demonstrate that RAST can improve model robustness while reducing training time through experiments, including a real system.

Thank you