INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

ATTA: Efficient Adversarial Training with Transferable Adversarial Examples

TECHNOLOGY NUMBER: 2024-365

Technology ID
2024-365

Category
Software
MOSS - Michigan Open Source Support

Inventor
Atul Prakash
Haizhong Zheng
Ziqi Zhang

Further information
Ashwathi Iyer
ashwathi@umich.edu

View online page

OVERVIEW

Efficient adversarial training by reusing transferable adversarial examples across epochs

- Reduces training time substantially while maintaining or improving model robustness
- Enables fast, robust model development for security-critical and time-sensitive applications

BACKGROUND

Adversarial training, which augments model training data with adversarially perturbed inputs, is a widely adopted defense against adversarial attacks in classification tasks. While effective in improving model robustness, its main drawback lies in the substantial computational overhead required to generate fresh, strong adversarial examples at every training step. This prolongs training times, consuming more resources and hindering the deployment of robust models, especially on large datasets or with limited computational budgets. Previous attempts to reduce this overhead often sacrifice robustness or rely on weaker adversaries, undermining defense effectiveness. Recent research has demonstrated some redundancy in adversarial perturbations across different training epochs, but this property has yet to be fully leveraged to optimize adversarial training efficiency. As such, there is a pressing need for approaches that can maintain robustness while dramatically cutting the computational cost of adversarial training.

INNOVATION

The proposed method, Adversarial Training with Transferable Adversarial Examples (ATTA), innovatively harnesses the observed transferability of adversarial examples between consecutive training epochs. By accumulating and reusing adversarial perturbations throughout the training process, ATTA minimizes the need to regenerate examples from scratch at each epoch. This technical advance dramatically accelerates adversarial training, requiring 12–14 times less training time on benchmarks like MNIST and CIFAR10, compared to traditional approaches, without compromising—and often enhancing—adversarial robustness (up to 7.2% improvement on CIFAR10). ATTA's efficiency enables robust models to be trained on resource-constrained hardware and reduces time-to-deployment for security-sensitive applications such as autonomous vehicles, financial systems, and healthcare diagnostics, while also making robust deep learning more accessible for large-scale and rapid prototyping environments.

ADDITIONAL INFORMATION

PROJECT LINKS:

- ATTA Github

DEPARTMENT/LAB:

- Atul Prakash, Computer Science and Engineering (CSE)

LICENSE:

- MIT