Wei LIN, Lichuan LIAO, 2024. Towards sustainable adversarial training with successive perturbation generation. *Frontiers of Information Technology & Electronic Engineering*, 25(4):527-539.
https://doi.org/10.1631/FITEE.2300474

Towards sustainable adversarial training with successive perturbation generation

Key words: Adversarial training; Adversarial attack; Stochastic weight average; Machine learning; Model generalization

Corresponding author: Lichuan LIAO
E-mail: liaolijuan@xaut.edu.cn

ORCID: https://orcid.org/0000-0001-8999-1573

Motivation

- The generated adversarial examples at each training epoch have high redundancy. This statement can be supported by existing methods, where the adversarial examples generated for one model can still stay adversarial to another model trained on the same dataset.
- The approach of the newly generated adversarial examples corresponding to every epoch is characterized by a lack of consistent robust feature representation generation when it comes to convolutional neural network (CNN) models trained on limited adversarial data, and this generation is a feature well-known to be essential for the maintenance of high accuracy on clean data.

Main idea

- We propose a successive adversarial training method that connects the adversarial examples and shifts models across training epochs, which significantly improves the efficiency and the generalization performance of CNN models.
- Extensive experiments show that, with comparable training time, our proposed method outperforms the competitive baseline adversarial methods on image classification benchmarks, including CIFAR-10 and CIFAR-100.

Framework

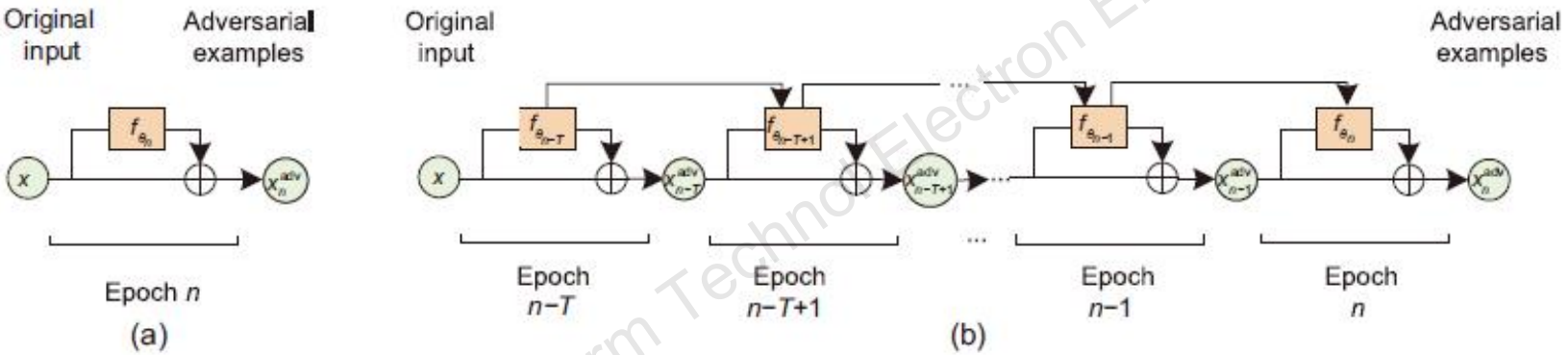

Fig. 2 Standard adversarial training process where the adversarial examples are generated from the original input at each epoch (a) and details of the proposed SPGAT method, which successively generates the adversarial examples and shift models from adjacent epochs (b)

Method

Algorithm 1: Successive perturbation generation scheme for adversarial training

Input: training data $\mathcal{D} = \{X, Y\}$, perturbation boundary ϵ, shift cycle T, hyperparameter C

Output: trained model f with parameter θ

Initialize δ from a uniform distribution within $(-\epsilon, \epsilon)$

for epoch $= 1, 2, ..., N$ **do**

  for $i = 1, 2, ..., B$ **do**

    $\delta_i \leftarrow \delta_i + \alpha \cdot \text{sign}(\nabla_{\delta_i} \mathcal{L}(f_\theta, x_i + \delta_i, y_i))$

    $\delta_i \leftarrow \max(\min(\delta_i, \epsilon), -\epsilon)$

    $x_i^{\text{adv}} \leftarrow M(x_{i-1}^{\text{adv}}) + \delta_i$

    $\theta \leftarrow \theta - \nabla_\theta \frac{\partial \mathcal{L}(f_\theta, x_i^{\text{adv}}, y_i)}{\partial \theta}$

  if epoch $\%$ $T = 0$ **then**

    $\theta^{\text{swa}} \leftarrow \frac{\theta^{\text{swa}} \cdot t + \theta}{t+1}$

    $t \leftarrow t + 1$

  if epoch $\%$ $C = 0$ **then**

    $t \leftarrow 0$

    $f_{\theta^{\text{swa}}} \leftarrow \text{UPDATE\_BN}(f_{\theta^{\text{swa}}}, \mathcal{D}, \delta)$

    $f_\theta \leftarrow f_{\theta^{\text{swa}}}$

    $x_i^{\text{adv}} \leftarrow x_i$

return f_θ

We first initialize the perturbation δ randomly from a uniform distribution within the range of $(-\epsilon, \epsilon)$. Then for each batch of training data during the training process, we generate the adversarial examples in each epoch to accumulate the attack strength. We use fast gradient sign method (FGSM) as the adversarial attack in the perturbation generation. Backward propagation is then performed on the trained model to update the model parameters. To make the training process more stable and escape from overfitting, we periodically update the model parameters and reset the perturbation for accumulating attack from the beginning. This training process is performed on all B batches of training data for N epochs.

Conclusions

In this paper, we present a successive perturbation generation scheme for adversarial training (SPGAT) for improving the robustness of CNN models. Specifically, SPGAT successively generates adversarial examples with a single-step attack and shifts models across the training epochs to enhance the efficiency of adversarial training. More importantly, this strategy greatly improves the generalization ability of models, thus imbuing them with the ability to maintain high accuracy on clean data. In the experiments, our proposed SPGAT demonstrates outstanding performance under various attacks, including white-box attacks, unseen adversaries, and black-box attacks with comparable training time.

Wei LIN received his MS degree in computer system architecture from Fuzhou University, Fuzhou, in 2016 and his PhD degree in information management from Yuan Ze University, Taoyuan, China, in 2022. He is currently an assistant professor of Fujian University of Technology. His research interests include AI security, AI for finance, natural language processing, and information security.

Lichuan LIAO received her PhD degree in the Institute of Management (School of Management majoring in Finance) from Yuan Ze University, Taoyuan, China, in 2014. She is now an associate professor of College of Economics and Management, Xi'an University of Technology, China. Her research interests include the green economy and ESG finance, senior finance, big data finance, technology finance, and inclusive finance.