International Journal of Computer Vision (2023) 131:3307–3331
https://doi.org/10.1007/s11263-023-01860-4

Adversarial Coreset Selection for Efficient Robust Training

Hadi M. Dolatabadi[1] · **Sarah M. Erfani**[1] · **Christopher Leckie**[1]

Received: 2 January 2023 / Accepted: 19 July 2023 / Published online: 14 August 2023
© The Author(s) 2023

Abstract
It has been shown that neural networks are vulnerable to adversarial attacks: adding well-crafted, imperceptible perturbations to their input can modify their output. Adversarial training is one of the most effective approaches to training robust models against such attacks. Unfortunately, this method is much slower than vanilla training of neural networks since it needs to construct adversarial examples for the entire training data at every iteration. By leveraging the theory of coreset selection, we show how selecting a small subset of training data provides a principled approach to reducing the time complexity of robust training. To this end, we first provide convergence guarantees for adversarial coreset selection. In particular, we show that the convergence bound is directly related to how well our coresets can approximate the gradient computed over the entire training data. Motivated by our theoretical analysis, we propose using this gradient approximation error as our adversarial coreset selection objective to reduce the training set size effectively. Once built, we run adversarial training over this subset of the training data. Unlike existing methods, our approach can be adapted to a wide variety of training objectives, including TRADES, ℓ_p-PGD, and Perceptual Adversarial Training. We conduct extensive experiments to demonstrate that our approach speeds up adversarial training by 2–3 times while experiencing a slight degradation in the clean and robust accuracy.

Keywords Adversarial training · Coreset selection · Efficient training · Robust deep learning · Image classification

1 Introduction

Neural networks have achieved great success in the past decade. Today, they are one of the primary candidates in solving a wide variety of machine learning tasks, from object detection and classification (He et al., 2016; Wu et al., 2019) to photo-realistic image generation (Karras et al., 2020; Vahdat & Kautz, 2020) and beyond. Despite their impressive performance, neural networks are vulnerable to adversarial attacks (Biggio et al., 2013; Szegedy et al., 2014): adding well-crafted, imperceptible perturbations to their input can change their output. This unexpected behavior of neural networks prevents their widespread deployment in safety-critical applications, including autonomous driving (Eykholt et al., 2018) and medical diagnosis (Ma et al., 2021). As such, training robust neural networks against adversarial attacks is of paramount importance and has gained ample attention.

Adversarial training is one of the most successful approaches in defending neural networks against adversarial attacks. This approach first constructs a perturbed version of the training data. Then, the neural network is optimized over these perturbed inputs instead of the clean samples. This procedure must be done iteratively as the perturbations depend on the neural network weights. Since the weights are optimized during training, the perturbations also need to be adjusted for each data sample in every iteration,[1]

Various adversarial training methods primarily differ in how they define and find the perturbed version of the

Communicated by Giorgos Tolias.

✉ Hadi M. Dolatabadi
h.dolatabadi@unimelb.edu.au

Sarah M. Erfani
sarah.erfani@unimelb.edu.au

Christopher Leckie
caleckie@unimelb.edu.au

[1] School of Computing and Information Systems, The University of Melbourne, Melbourne Connect, 700 Swanston Street, Carlton, VIC 3053, Australia

[1] Note that adversarial training in the literature generally refers to a particular approach proposed by Madry et al. (2018). In this paper, we refer to any method that builds adversarial attacks around the training data and incorporates them into the training of the neural network as adversarial training. Using this taxonomy, methods such as TRADES (Zhang et al., 2019) ℓ_p-PGD (Madry et al., 2018) or Perceptual Adversarial Training (PAT) (Laidlaw et al., 2021) are all considered different versions of adversarial training.