Computer Science
Defense

Deep Learning with Differential Privacy and Adversarial Robustness

Pengfei Tang
Emory University

Abstract: Deep learning models have been increasingly powerful on different tasks, such as image classification and data synthesization. However, there are two major vulnerabilities existing: 1) privacy leakage of the training data through inference attacks, and 2) adversarial examples that are crafted to trick the classifier to misclassify. Differential privacy (DP) is a popular technique to prevent privacy leakage, which offers a provable guarantee on privacy of training data through randomized mechanisms such as gradient perturbation. For attacks of adversarial examples, there are two categories of defense: empirical and theoretical approaches. Adversarial training is one of the most popular empirical approaches, which injects adversarial examples with correct labels to the training dataset and renders the model robust through optimization. Certified robustness is a representative of theoretical approaches, which offers a theoretical guarantee to defend against adversarial examples through randomized mechanisms such as input perturbation. ¡br¿¡br¿
However, there are some limitations in existing works that reduce the effectiveness of these approaches. For DP, one challenge is the contradiction between a better utility performance and a certain level of privacy guarantee. For adversarial training, one challenge is that when the types of adversarial examples are limited, the model robustness is confined. For certified robustness, existing works fail to exploit the connection between input and gradient perturbation, which wastes a part of randomization during training. To solve these limitations, 1) we propose a novel framework IGAMT for data synthesization. Compared with traditional frameworks, IGAMT adds less gradient perturbation to guarantee DP, but still keeps the complex architecture of generative models to achieve high utility performance. 2) We propose a distance constrained Adversarial Imitation Network (AIN) for generating adversarial examples. We prove that compared with traditional adversarial training, adversarial training with examples from AIN can achieve comparable or better model robustness. 3) We propose a new framework TransDenoiser to achieve both DP and certified robustness, which utilizes all randomization during training and saves the privacy budget for DP.

Thursday, November 11, 2021, 3:00 pm
https://us02web.zoom.us/j/7382282740?pwd=QVB4bmU2NnlZN2s1UW0veUtCNklmU

Computer Science
Emory University