Paper Review

DEEPSEC: A Uniform Platform for Security Analysis of Deep Learning Model

Chandrika Mukherjee
email: cmukherj@purdue.edu
Date: 22nd November 2021

Summary

The paper presents design, implementation and evaluation of DEEPSEC - a uniform platform to analyse attack and defense methods of Deep Learning Models. DEEPSEC incorporates 16 state-of-the-art attacks with 10 attack utility metrics and 13 state-of-the-art defenses with 5 defense utility metrics. The paper provides brief quantitative description of existing attack, defense techniques along with the various utility metrics. The aim of DEEPSEC is to enable researchers and practitioners to (i) measure the vulnerability of DL models, (ii) evaluate the effectiveness of various attacks/defenses, and (iii) conduct comparative studies on attacks/defenses in a comprehensive and informative manner. The authors also shared key insights from evaluation of DEEPSEC which resolved some long standing questions regarding security of Deep Learning Models. In my view, although the goals are laudable, the work has some methodological flaws. Detailed comments and recommendations are mentioned in the following sections.

Detailed Comments

- Deep Learning is increasingly used in different security-sensitive domains such as - self driving cars, face recognition, malware detection, medical diagnostics etc. But these DL models are highly vulnerable to adversarial examples. Although different defense and attack techniques were introduced in past, the evaluation of those techniques were not done in a comprehensive manner. The authors, here introduces "DEEPSEC" - a uniform security analysis platform for deep learning models. The paper presents extensive analysis of existing adversarial attack and defense methods, and draws a set of key findings, which demonstrate DEEPSEC's rich functionality.
- The authors correctly identifies the lack of quantitative understanding about the strength and limitations of the existing defense and attack techniques on DL models due to incomplete or biased evaluation.
  - simple evaluation metrics was used
  - evaluation was done against small set of attacks/defenses
  - new defenses against newly introduced attacks or new attacks against newly defined defense invalidates conventional understanding which results in contradictory or puzzling conclusion.

  Therefore, in my view the research problem is highly relevant.
- The aim of DEEPSEC is to provide a Uniform, Comprehensive, Informative, Extensible analysis platform to support comprehensive and informative evaluation of adversarial attacks and defenses on Deep Learning Models.
- In comparison to work of Cleverhans with 9 attacks and 1 defense mechanisms, authors here, incorporates 16 state-of-the-art adversarial attacks with 13 defenses. DEEPSEC performs largest scale empirical study using 10 evaluation metrics on attacks and 5 evaluation metrics on defenses.

- The paper primarily considered non-adaptive white-box model. But as a security paper, the paper should have provided a well-defined threat model. Most defenses contain a threat model as a statement of the conditions under which they attempt to be secure. Therefore, this paper is lacking a significant feature of Security of Machine Learning.

- The authors presents a brief summary with quantitative definition of all the used attack, defense methods and evaluation metrics.

  - The paper presents different types of iterative and non-iterative targetted and untargetted attacks.
  - The paper considers misclassification, imperception, and robustness as utility requirements while taking the resilience as the security requirement.
    * Misclassification Ratio(MR), Average Confidence of Adversarial Class (ACAC), Average Confidence of True Class (ACTC) are incorporated to measure misclassification.
    * High imperceptibility suggests that the adversarial examples will still be correctly classified by human vision. Average Lp Distortion (ALD_p), Average Structural Similarity (ASS) and Perturbation Sensitivity Distance (PSD) are used to evaluate imperceptibility.
    * The authors correctly identifies that preprocessing of input images before being used in production can decline the Misclassification Ratio. Therefore, measurement of Robustness of Adversarial examples is of high importance. DEEPSEC uses Noise Tolerance Estimation (NTE), Robustness to Gaussian Blur (RGB), Robustness to Image Compression (RIC). Gaussian Blur and Image Compression are highly used in Computer Vision algorithms to preprocess images.
  - DEEPSEC calculates Computation Cost as average runtime for attackers to generate an Adversarial Example.
  - DEEPSEC used Naive Adversarial Training (NAT), Ensemble Adversarial Training (EAT), PGD based Adversarial Training (PAT), Gradient Masking/Regularization (Defensive Distillation and Input Gradient Regularization), Input Transformation (Ensemble Input Transformation, Random Transformations-based defense, PixelDefense), Region-based classification and Detection only Defenses.
  - The utility metrics used to evaluate defense technique are - Classification Accuracy Variance (CAV), Classification Rectify/Sacrifice Ratio (CRR/CSR), Classification Confidence Variance (CCV), Classification Output Stability(COS).

- The paper presents 5 parts of DEEPSEC - a) Attack Module (AM), b) Defense Module(DM), c)Attack Utility Evaluation (AUE), d) Defense Utility Evaluation (DUE), e) Security Evaluation (SE). Previously due to lack of uniform platform, existing attacks and defenses are implemented and evaluated on different experimental settings. On the other hand, DEEPSEC reduces the evaluation bias and facilitate fair comparison among various attack and defense methods using its uniform platform. DEEPSEC also allows researchers to evaluate utility and security performance of newly proposed adversarial attacks by attacking state-of-the-art defenses. The modular implementation of DEEPSEC makes it easily extendable.

- The paper presents evaluation of DEEPSEC on two popular benchmark dataset (MNIST, CIFAR-10). The value of common parameters are kept same for unbiased comparisons. All the other parameters are kept similar to the original work.

- Existing attacks show high success rate in misleading target model. AEs with low ACTC show better resilience to other models.
- PSD is the most sensitive imperceptible metric to the perturbation of AEs, while ASS is the least sensitive. The trade-off between misclassification and imperceptibility is empirically confirmed.
- Most Untargetted attacks are more robust than Targetted Attacks.
- Overall, as long as the defense-enhanced models are trained or adjusted based on the accuracy metric, most of them can also preserve the other utility performances, such as CCV and COS.
- The evaluation also suggests that no defense is universal.

Although the results are unique and informative, the paper computes average for summarizing results, instead of minimum or maximum. Computing the average over different threat models is meaningless. The only metric that matters in security is how well a defense withstands attacks targeting that defense.

- The authors also performed two case studies.
  - The paper suggests that transferability of UAs are more than TAs.
  - L_∞ attacks are more transferable than other attacks (i.e., L2 and L0 attacks).
  - Another case study shows that ensemble of different defenses does not improve the defensive capability as a whole, but can improve lower bound of defensive ability.

Recommendations

The authors kept their code open source which encourages other researchers to contribute to DEEPSEC and to provide feedback to their implementation details that will make DEEPSEC a better comprehensive tool.

Only White box , non-adaptive attacks were considered. Although the authors provided some insights on how the attacks can be extended as adaptive and black-box attacks, the paper missed significant details of a well defined Threat Model. Moreover, they calculated average over different threat models which can be meaningless with respect to Security of Machine Learning.

Although the paper used similar setting and kept common parameters same among different attacks or defenses for fair comparison, the details of the selected parameters are described vaguely. Therefore, it can be hard for new researcher to use DEEPSEC for further analysis.

However, The DEEPSEC provides a comprehensive empirical analysis of attacks and defenses on Deep Learning models which provides answers to some long standing questions regarding effectiveness of ensemble of defense techniques. In my view, DEEPSEC sets a useful standard to facilitate adversarial deep learning research.