DEEPSEC: A Uniform Platform for Security Analysis of Deep Learning Model
Paper Review-4
Shivam Bhat
bhat41@purdue.edu
11/10/2022

Summary

Deep learning has today permeated most aspects of our lives. Today deep learning is being leveraged to perform various artificial intelligence tasks like image classification, speech recognition and even emulating players playing computer games. This warrants a need to secure these models against adversarial attacks. Unfortunately, It has been adequately shown that deep learning models are vulnerable to adversarial examples which are maliciously crafted input samples, made with the intention to deviate the models and make them mis-classify.
The increased use of DL models in security applications has invited attention and consequently intensive studies from both academia and industry. These Studies focus on adversarial attacks and defense strategies. On a higher level, attacks use adversarial examples to force the model into misclassifying by perturbing otherwise legitimate inputs. Defense strategy on the other hand focuses on strengthening the resilience of models against such adversarial examples , whilst maintaining their functionality. Inspite of the abundant literature and data present on such adversarial attacks and defenses there still exists a lack of quantitative understanding about the same. This is largely due to the lack of an elaborate analysis platform allowing comprehensive evaluation. To fill this void, the authors propose a first-of-its-kind platform called “DEEPSEC” which is elaborate enough to act as a benchmarking platform to facilitate future adversarial learning research. The paper does a good job of elucidating the offering of the framework and its contributions to the field. We are provided with the architecture of the framework along with a detailed explanation of its offered attacks, defenses and their utility metrics which are evaluated on two popular benchmark datasets. The detailed process, my comments and suggestions are listed in the following sections.

Comments

Advent of cheap and readily available computing resources, along with the recent breakthroughs in the domain, have today allowed Deep learning to be widely adapted across various industries for applications like self-driving cars, malware detection, face recognition etc. However, it has been found that these deep learning models are inherently very vulnerable to adversarial examples. Hence, to allow security researcher better deal with the myriad of adversarial attacks

and defenses, this paper proposes “DEEPSEC”- A first-of-its-kind platform which allows researchers to (i) measure vulnerability of DL models against attacks,(ii) quantify and compare the effectiveness of various attacks and defenses, and (iii) carry out comparative studies on attacks/defenses in a comprehensive and informative manner.

For a model to be comprehensive and practically useful, the paper proposes a set of 4 requirements namely:

1. Uniform - It should be able to support comparison between different attack/defense methods under the same configurations.
2. Comprehensive - It should support and encompass most attack/defense methods.
3. Informative - It should have a wide and rich set of metrics to evaluate different attack and defense methods.
4. Extensible - It should be flexible and easy to extend to include new attack/defense methods.

Since none of the existing work covers all of the above requirements, the author proposes “DEEPSEC” to bridge the gap. In its current implementation, it incorporates 16 state-of-the-art adversarial attacks with 10 attack utility metrics and 13 representative defenses with 5 defense utility metrics.
The paper does a thorough analysis of all the attacks available in its repertoire. The attacks are classified along two directions by the author - attack specificity (UA and TA) and attack frequency(non-iterative and iterative attacks) . UA or Un-targeted attacks generate adversarial examples(AE) that can be misclassified into any class other than the ground truth class, whereas the Targeted Attacks(TA) aim to generate AEs to be classified into a specific target class. In case of attack frequency based classification, non-iterative attacks take only a single step to generate AEs, whereas iterative attacks take multiple iterative updates. We learn about the first and fastest non-iterative UA, called Fast Gradient Sign Method (FGSM) proposed first by Goodfellow et al. We are then introduced to an extension of FGSM called Basic Iterative Method (BIM) that iteratively takes multiple small steps while adjusting the direction after each step. We are further introduced to modifications and variations of BIM which result in attacks like PGD and U-MI-FGSM. Within the non iterative TAs we are introduced to a non-iterative TA called LLC attack and R+LLC. We then learn about an iterative AE called Box-constrained L-BFGS (BLB) algorithm which inspite of its great utility is time-consuming and impractical to linearly search for the optimal solution at large scale. Hence to overcome such inefficiencies, a straightforward iterative version of LLC called ILLC was proposed.

We are then introduced to utility metrics of attacks which on a broader level measure the extent to which the adversarial attack can provide "successful" AE for an adversary. A successful AE should not only be imperceptible to naked human eye but also robust to transformation and resilient to existing defenses. The paper then introduces us to each of the 10 utility metrics for adversarial attacks. Misclassification Ratio (MR) is defined as the most important property for attacks. Since Imperceptibility to the human eye is also an important criteria for a successful attack we use utility metrics like Average Lp Distortion(ALDp), Average Structural Similarity(ASS) and Perturbation Sensitivity Distance (PSD). To evaluate the robustness of AEs to processes like image-preprocessing we use different metrics like Noise Tolerance Estimation(NTE), Robustness to Gaussian Blur (RGB) and Robustness to Image Compression (RIC).

The framework classifies the defense techniques into 5 categories which are then thoroughly discussed by the author. We learn about adversarial training in which we augment the training set with newly generated attack examples. To scale this to large scale datasets, Kurakin et al. proposed Naive Adversarial Training (NAT). In another defense technique called "Gradient Masking/Regularization" we reduce the sensitivity of the models to AEs and hide the gradients. In a simpler approach called "Detection-only defenses" we only detect AEs from normal examples based on the observation that Local Intrinsic Dimensionality (LID) of AEs is significantly higher than that of usual normal examples. Following this, we are introduced to the utility metric of defenses which are evaluated from two perspectives-utility preservation and resistance to attack where 'utility preservation' encapsulates how well the functionality of the original model is preserved while the resistance measures the effectiveness of defense-enhanced models against adversarial attacks. We are introduced to 5 utility metrics of defenses namely - Classification Accuracy Variance(CAV), Classification Rectify/Sacrifice Ratio(CRR/CSR), Classification Confidence Variance(CCV) and Classification Output Stability(COS). Since accuracy is the most important metric used to evaluate the performance of a DL model CAV is often used.CAF is defined as ACC(Fd,T) - Acc(F,T) where Acc(F,T) denotes model F's accuracy on dataset T. The utility metric COS uses JS divergence to measure the similarity of the original model and the defense enhanced model.

The authors do a great job at evaluating the various defense and attack algorithms. This is performed on two popular datasets namely MNIST and CIFAR-10. A 7-layer CNN and a ResNet-20 model are trained. The results of the experiments are summarized well through remarks after each paragraph. We learn that iterative attacks have nearly 100% MR, which can be intuitively explained by the fact that iterative attacks run multiple complicated iterations to find the most optimal perturbation for the target model whereas the non iterative

attacks only take one step to complete the perturbation. The authors suggest that AEs with lower ACTC would show better resilience to other models as their true classes are less likely to be correctly classified by other models. To calculate the computational cost of an attack, we test the runtime that is used to generate one AE on average. Intuitively, AEs of iterative attacks are much more computationally expensive than non-iterative attacks. On an average this ratio is 10:1 between them.

On evaluating the defenses we notice that all defenses achieve comparable performances on both MNSIT and CIFAR-10, however for most of the defenses their defense-enhanced models have variances for classification accuracy on the testing set. In the context of CCVs, we notice that compared to MNSIT CCVs of defense-enhanced models on CIFAR-10 are 10x higher in most cases. The authors then test the effectiveness of 10 complete defenses against attacks. We learn that all the state-of-the-art defenses are more or less effective against existing attacks. In general defenses show better protection against TAs than that on UAs. A likely reason given for this is that UAs are likely to generalize to other models including defense-enhanced models, while TAs tend to overfit to specific target models. Among all attacks NAT, PAT, EAT, TE, EIT, and IGR show better and stable performance in defending against most attacks. From the results, authors conclude that no defense is universal to all attacks. While evaluating detection-only defenses, we learn that it is necessary that AEs with high magnitude of perturbation are easier to be detected.

Before the closing section of the paper, the author presents two case studies to elaborate on the 'Transferability of Adversarial Attacks' and 'Robustness of Ensemble Defenses'. Transferability is the property which allows AEs generated against one target model to be mis-classified by other models. We learn that all adversarial attacks show more or less transferability to other models. Infact the average transferability rate of all attacks on the 3 models prepared in the experimental setup is 42.4% and appear to be independent of model architecture.
While evaluating the robustness of ensemble defenses the authors use three different methods for their experimental setup-completely random ensemble, interclass random ensemble and best defense ensemble. On evaluating the results we notice that generally different ensemble methods show different defensive performances, however among the three ensembles, the completely random ensemble performs the worst while the best-defense ensemble performs the best with respect to accuracy and confidence. simultaneously, we also observe that even for the best-defense ensemble, the improvement in accuracy is not that significant when compared to the most successful defense.

Recommendations and takeaways

The paper is very thorough in its approach. The experiments have been performed and analyzed for all of the defenses and attacks used by the "DEEPSEC" framework. The authors' decision to make code base open source is really commendable and the modularity of the framework will further increase the contributions consequently increasing its utility.

While there are logistical restrictions, I would have liked to see analysis on other datasets. The two main datasets used are MNSIT and CIFAR-10 which are both used in the same domain of image processing. A diverse dataset and AEs would have made the work more encompassing.

At multiple places the results are interpreted and a conjecture has been proposed by authors which might not be suited for the settings of a research work. A solid reasoning would have further increased the credibility of the work. These could have also been included in future work sections.

While the author only considers non-adaptive and white box attack scenarios, it would have been great to get a brief overview of black box attacks.