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

Sowmya Jayaram Iyer
jayarami@purdue.edu
PUID: 0033742039

SUMMARY:

The three key findings claimed were (1) An empirical evaluation of the trade-off between misclassification and imperceptibility (2) The phenomena of allegedly universally applicable defenses succeeding only against limited types of attacks under restricted settings (3) Adversarial examples with higher perturbation magnitude are not necessarily easily detected (4) Defense capability might not be always improved by using an ensemble of multiple defenses but the lower bound of the defense effectiveness of individuals can be improved. The author has done a good job in conveying the need for this evaluation and has done a thorough job in presenting the findings. However, this paper needs few improvements in showing the reliability of the methods and hyperparameters chosen and the robustness of the comparison. Comments and suggestion are listed in the next section

DETAILED COMMENTS:

The paper "DEEPSEC: A Uniform Platform for Security Analysis of Deep Learning Model" aims to serve as an evaluation platform for adversarial attacks/defenses and also to support the evaluation of attacks/defenses in a uniform and informative manner. This objective is commendable since such a systematic, preferably reliable, evaluation of attacks and their effectiveness in the past could inform future judgements of which attacks should be tested. Similarly, an informative review of defenses could help researchers to pick defenses to test against while designing new attacks. This paper also tabularizes an empirical study on adversarial attacks/defenses under different metrics and also a cross evaluation between different attack and defense methods (16 attacks × 13 defenses) which contribute towards better understanding of each method's strengths and limitations on a relative scale. The key findings mentioned were mostly corroborated with results, however, a few of them seem to be far-fetched claims. For example, the defending hyperparameters are adjusted based on the attack which does not give a realistic relative comparison when comparing within threat models.

In section 2, the author introduces us to the attacks and defenses along with the utility metrics used in this paper for comparison. Given are 16 attack methods out of which are notable attacks like Fast Gradient Sign Method, Basic Iterative Method,

DeepFool among the 8 Targeted attacks and Least Likely Class attack, Jacobian-based Saliency Map Attack, Carlini and Wagner's attack among the 8 Untargeted attacks. The author also presents around 13 defense mechanisms out of which 3 are detection techniques. The utility metrics for attack methods are broadly classified under four categories, evaluation in terms of: Misclassification, Imperceptibility (how semantically similar is an adversarial example to a benign example?), Robustness (towards pre-processing techniques) and Computation Cost (running time for generation of an AE). Highlighting these metrics provides a deeper understanding of an adversarial example's ability than simply looking at the misclassification ratio. The utility metric for defense is put under two categories namely utility preservation and resistance to attacks. These classifications do help understand the effectiveness of attacks as well as defenses from various point of views. However, a very notable drawback seen in the *Utility of Metrics for Defences* is how the author extensively relies on Averaged performance metrics. The one key factor which differentiates security (or adversarial robustness) from all other forms of robustness is the worst-case mindset from which the metrics are evaluated. Hence, this is fundamentally an incorrect evaluation to make since the only metric that matters in security is how well a defense would withstand attacks targeting that defense in particular.

In section 3, a succinct explanation of every model and its purpose in the System is given. The author splits the Deepsec into five parts: 1) Attack Model (AM) – where DL vulnerabilities are attacked, 2) Defense Model (DM) - DL models' defense, Attack Utility Evaluation (AUE) - where utility requirements of adversarial attacks are evaluated, Defense Utility Evaluation (DUE) – where the utility of the state-of-the-art defenses are evaluated and Security Evaluation (SE) - the vulnerability and robustness of defense-enhanced models are evaluated. Here, the author stresses repeatedly upon the contributions of Deepsec towards the research community. Deepsec does provide a uniform platform supporting a systematic evaluation of different adversarial attacks and defenses which would greatly help further research. The fact that Deepsec is open source and allows integration of individual contributions all the more makes this contribution significant.

In section 4, it can be seen that the attacks are not run in an all-pairs manner. This cant be called a flaw per say since the results evaluated still hold some meaning. This can be used to see the extent of transferability of an attack and also the extent to which a defense is robust against such attacks. However, this was not the fundamental observation but the author related this to security evaluation which is not convincing. Though Table 3 looks convincing, there are certain aspects which aren't substantiated:

- We can see that the author has studied a distortion of $\epsilon = 0.1$ and $\epsilon = 0.2$. When referring previous literatures, it can be seen that this value is quite large when

compared and the reason for choosing such a high distortion value is not mentioned.

- Also, in Table 7, the author uses a distortion bound for l_∞ as high as 0.6. Such a high value might cause the adversarial example to turn grey hence not semantically similar to true label which contradicts the entire purpose of the distortion bound.

Under Defenses vs. Attacks the author says "*In order to fairly compare the detection rate (i.e., TPR), we try our best to adjust the FPR values of all detection methods to the same level via finetuning the parameters*." This implies that parameters are tuned such that they are tailor-made for the attacks applied on them. This does not resonate the real world scenarios where the defenses are fixed and the attacks on them are unknown. This undermines the findings tabulated on this regard and also the Remark 6 mentioned.

RECOMMENDATIONS:

This paper does a good job in explaining the need for such a uniform platform and its contributions towards the research community. However, the author does not elaborate on how the parameters defining the attack and defense models were chosen. A lot of contradictory results are observed. For example, from the table, the average model accuracy of FGSM – a weak attack is high and hence seems like a much stronger attack than others. The explanation behind such findings is not given or is hardly convincing and hence make readers question the fundamentals of the methodology. Also the definition of success rate for attacks which are unbounded makes little sense. Such discrepancies in the observations made by this paper and prior works in the domain are not explained from the point of view of how these attacks and defenses work and are simply provided on brute-force basis. Thus, with sufficient theoretical explanation on the tabulated results, this paper will prove to be a more reliable benchmark for a unified platform.