Adversarial Examples Generation and Mitigation in Deep Neural Networks

Harvish Jariwala, 28, Samasta Brahma Kshatriya Society, Narayannagar Road, Paldi, Ahmedabad, India. harvish1288@gmail.com

Nancy Radadia, A52, Zodiac Aster, SG Highway, Opposite Ahmedabad International School, Bodakdev, Ahmedabad, India. narofficial08@gmail.com

ABSTRACT - A trillion-crease expansion in calculation power has promoted the use of profound learning (DL) for dealing with an assortment of AI (ML) errands, like picture order, normal language handling, and game hypothesis. Notwithstanding, a serious security danger to the current DL calculations has been found by the exploration local area: Adversaries can undoubtedly trick DL models by irritating harmless examples without being found by people. Irritations that are impalpable to human vision/hearing are adequate to incite the model to make an off-base expectation with high certainty. This peculiarity, named the ill-disposed example, is viewed as a critical snag to the mass sending of DL models underway. Significant examination endeavors have been made to concentrate on this open issue like Deep Leaning, Networked Deep Learning Systems, Adversarial Training Defense, Transfer Adversarial Attack. It has been found empirically that adversarial examples can impact a deep learning convolutional network in misclassification of input images. Adversarial examples attack is direct attack on deep learning systems on the application layer. Mitigation on different layers of the network stack must be adhered to particularly on networked systems. Adversarial examples are transferable from one model to another and Adversarial training is considered as the most effective mitigation Standardized code hardening guidelines must be implemented in a global scale to reduce the risk of adversarial examples.

(Keywords; Networked Deep Learning Systems, Adversarial Training Defense, CIFAR-10, MNIST, ImageNet, Fast Gradient Sign Method, Transfer Adversarial Attack)

Introduction

A trillion-crease expansion in calculation power has promoted the use of profound learning (DL) for dealing with an assortment of AI (ML) errands, like picture order , normal language handling , and game hypothesis . Notwithstanding, a serious security danger to the current DL calculations has been found by the exploration local area: Adversaries can undoubtedly trick DL models by irritating harmless examples without being found by people. Irritations that are impalpable to human vision/hearing are adequate to incite the model to make an off-base expectation with high certainty. This peculiarity, named the ill-disposed example, is viewed as a critical snag to the mass sending of DL models underway. Significant examination endeavors have been made to concentrate on this open issue.

As indicated by the danger model, existing antagonistic assaults can be ordered into white-box, dim box, and black-box assaults. The contrast between the three models lies in the information on the enemies. In the danger model of white-box assaults, the enemies are accepted to have full information on their objective model, including model design and boundaries. Consequently, they can straightforwardly create antagonistic examples on the objective model using any and all means. In the dark box danger model, the information on the enemies is restricted to the construction of the objective model. In the black-box danger model, the enemies can turn to the question admittance to create antagonistic examples. In the systems of these danger models, various assault calculations for antagonistic example age have been proposed, like restricted memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) calculation, the quick slope sign strategy (FGSM) , the fundamental iterative technique (BIM)/projected angle drop (PGD), distributionally ill-disposed assault, Carlini and Wagner (C&W) assaults, Jacobian-based saliency map assault (JSMA) , and DeepFool . These assault calculations are planned in the white-box danger model. Notwithstanding, they are additionally successful in many dark box and black-box settings because of the adaptability of the ill-disposed examples among models.

In the interim, different protective procedures for antagonistic example recognition/characterization have been proposed as of late, including heuristic and certificated safeguards. Heuristic guard alludes to a safeguard system that performs well in shielding explicit assaults without hypothetical precision ensures. Right now, the best heuristic protection is ill-disposed preparing, which endeavors to further develop the DL model's heartiness by integrating antagonistic examples into the preparation stage. As far as observational outcomes, PGD ill-disposed preparing accomplishes cutting edge precision against a large number of assaults on a few DL model benchmarks like the changed National Institute of Standards and Technology (MNIST) data set, the Canadian Institute for Advanced Research-10 (CIFAR-10) dataset, and ImageNet . Other heuristic guards

 DOI : 10.35291/2454-9150.2022.0401

essentially depend on input/highlight changes and denoising to ease the ill-disposed impacts in the information/highlight areas. Conversely, confirmed protections can continuously give confirmations to their least precision under a clear cut class of ill-disposed assaults. An as of late well known network confirmation approach is to plan an ill-disposed polytope and characterize its upper bound utilizing curved relaxations. The casual upper bound is a certificate for prepared DL models, which ensures that no assault with explicit constraints can outperform the certificated assault achievement rate, as approximated by the upper bound. Nonetheless, the genuine execution of these certificated safeguards is still a lot of more regrettable than that of the ill-disposed preparing.

Deep Leaning

Profound learning models are known to take care of order and relapse issues by utilizing various age and preparing tests on an enormous dataset with ideal precision. Notwithstanding, that doesn't mean they are resistant to assault or unexposed to weaknesses. Recently conveyed frameworks especially on a public climate (i.e public organizations) are helpless against assaults from different substances. Besides, distributed research on profound learning frameworks (Goodfellow et al., 2014) have decided a critical number of assaults focuses and a wide cluster of assault surface that has proof of double-dealing from ill-disposed models. Effective endeavor on these frameworks could prompt basic true repercussions.

For example, (1) an ill-disposed assault on a self-driving vehicle running a profound support learning framework yields an immediate misclassification on people causing inappropriate mishaps.

(2) a self-driving vehicle misreading a red light sign might make the fender bender to another vehicle

(3) misclassification of a passerby path as a crossing point path that could prompt vehicle crashes. This is only a glimpse of something larger, PC vision sending are not completely centered around self-driving vehicles but rather on numerous different regions too — that would conclusively affect this present reality. These weaknesses should be alleviated at a beginning phase of advancement. It is basic to create and execute pattern security principles at a worldwide level before true sending.

Profound learning calculations have seen their arrangement in various enterprises in a vertical direction and will keep on expanding in the forthcoming years due to the (1) upgrades on learning calculations and apparatuses, (2) further developed research by gifted designers and researchers, (3) figuring power. The ascent of Deep Learning execution will result to a bigger assault surface for ill-disposed assaults. These situations typifies the weaknesses on profound advancing as well as for a bigger scope — the whole AI environment.

Antagonistic models have monstrously tricked classifiers into misclassification of preparing dataset — adding irregular commotion to the information data,single step and multi-step assaults — have clearly advanced in compromising profound learning frameworks. Also, focusing on preparing models by irritating the picture personality (i.e,. vehicle is misclassified as a canine) are a portion of the generally distributed strategies in going after the framework. Regardless of it's high exactness, these frameworks are weak through an extensive variety of assault surface that is demonstrated to be exploitable. In this paper, we expect to foster an instinct on alleviating ill-disposed guides to additional upgrade the heartiness of profound learning frameworks. An organized profound learning framework contains various passage and leave focuses to and from the organization and should be relieved from the beginning phases of improvement or preceding the sending. Exactly, we have found that these section focuses expands the likelihood of a fruitful ill-disposed assault. We propose a weakness rating score (likelihood of an effective endeavor) for every weakness tracked down on a profound learning framework and set a worldwide norm on relieving every weakness.

NETWORKED DEEP LEARNING SYSTEMS

Intuitively, designers have to take into consideration the variations of deep learning implementation: for one a self-driving car—is most likely to be networked on a public environment (i.e,. The internet). This opens up a ton of opportunities for attackers that increases the probability of a successful exploit not just from adversarial examples but from different forms of malware. Since the code resides on the application layer and hosted locally within the car itself, there has to be security measures in place on the physical layer which is inside the car. Some important question needs to be answered (1) Does perimeter security (i.e,. Firewall) has empirical value on deployment? (2) How does endpoint security fit into the equation (i.e,. Antivirus) to defend against endpoint attack. (3) There have been some known security loopholes on 5G networks (Jover et al.,2019) which will power self-driving cars internet connection. How does this impact the robustness of deep networks?

It is quite evidenced that deep neural networks will not only be vulnerable to adversarial examples but also to a greater extent they are geared towards the cybersecurity and network space that allows them to be vulnerable to any attack just like a software application which is what they are.

ADVERSARIAL TRAINING DEFENSE

 DOI : 10.35291/2454-9150.2022.0401

Adversarial training is one of the known methods in mitigating adversarial examples—making the network more robust from white hat and black hat attacks and is highly considered as the most effective way of mitigation (Kurakin et al., 2018). In adversarial training the network is being trained to classify images with clean examples and perturbed or adversarial examples—allowing a baseline for error classification.

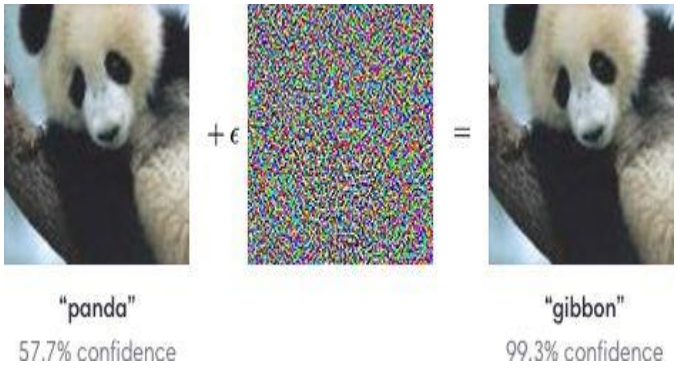

Fig. 1: Panda being misclassified as Gibbon due to noise

From the image shown above, a panda on the left and a gibbon on the right. An attacker can add a random noise or a minor perturbation that can result into tricking the classifier of categorizing the panda as a gibbon. In Adversarial training, we configure the classifier as having the best and worst case scenario of image classification—we train the model into classifying the image as its best case: a panda and a worst case: a gibbon. This is an important aspect of training, in which any perturbation can be deflected by adversarial training—allowing a more robust system that can resist adversarial examples. Furthermore, it has been known that adversarially trained models exploited by single step attacks to generate adversarial examples are easier to classify as well as for undefended model (Goodfellow et al., 2018). Definitively, adversarial training not only learns to deflect the attack but also make the attack performs at a worse level (Goodfellow et al., 2018).

GENERATING ADVERSARIAL EXAMPLES

Adversarial example generation can be categorized into (1) single step attack where there is only a single gradient computation and (2) multi step or iterative attack where there are multiple gradient computations iteration. The objective of every adversarial example is to have a high error rate on the loss function

$$L(Xn + rn, ytrue\, n\,;\, \theta)\ for\ each\ image\ Xn$$

(Zhou Ren et al., 2018), taking into account that the image generated is similar to the image from the training example. Furthemore, adding randomization layers on the model's architecture has been found to be successful defense in adversarial examples particularly in multi-step iterative attacks, in stark contrast to adversarial training where it has evidence of having a high success rate in defending single step attacks (Goodfellow et al., 2018)

It is important to note that these attacks can be exploited using white box and black box techniques and evidently—security against white-box attacks is the main goal because of the attackers access and knowledge of the system, although black-box security has more emphasis in developing a baseline goal for deployed ML models.

Below we list the methods in generating adversarial examples and its impact on the network.

Fast Gradient Sign Method (FGSM). (Goodfellow et al., 2014b)

FGSM is considered as a single step attack where a single gradient is computed to generate the adversarial example. FGSM leverages the following formula:

$$x\ adv\ FGSM\ :=\ x\ +\ \varepsilon \cdot sign\ (\nabla x L(h(x),\ ytrue)).$$

$$x^{adv} = x + \varepsilon \cdot \mathtt{sign}(\nabla_x J(x, y_{true})),$$

where

x is the input (clean) image,

x^{adv} is the perturbed adversarial image,

J is the classification loss function,

y_{true} is true label for the input x.

FSGM in white hat based attacks targets perturbation on input data therefore resulting into a higher loss based on the same back propagated gradients. It is architected to attack deep learning networks by the way the networks learn-gradients. Intuitively, there has been some notion that in a black-box setting where an attacker does not have full access to the model's architecture. A transferrable attack can be propagated from a trained adversarial network that could be transferred to the targeted network

TRANSFER ADVERSARIAL ATTACK

There has been formal and empirical evidence that adversarial example can transfer to more than one model (Papernot et al., 2017). In a black box setting (where an attacker does not have full access of the model's architecture) an attacker can train a surrogate model that has the same input training examples as the targeted model. This leads to a higher probability of successfully exploiting the target model using a surrogate model. Furthermore, input data generated from one model performing the same task can be transferred to another model. Transferring an attack has limitations (Boneh et al., 2017). Concretely, it has

 DOI : 10.35291/2454-9150.2022.0401

been proven that transferability of model-agnostic perturbations. As we can see from the image below, a small perturbation on an input image causes a direct misclassification on the training example.

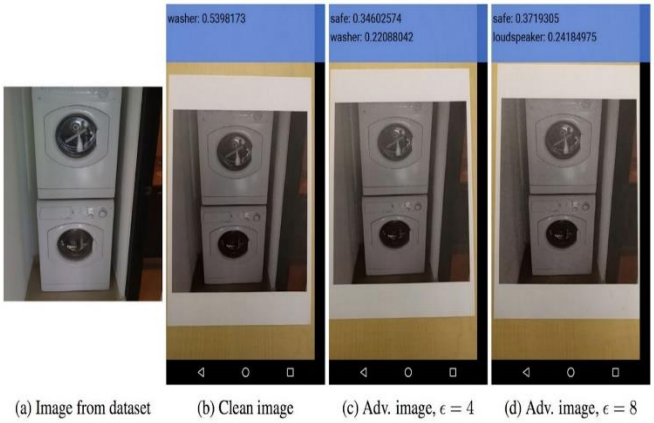

Fig 2: Input image causes misclassification on the training example

CIRCUMVENTING DEFENSES TO ADVERSARIAL EXAMPLES

Anish Athalye, Massachusetts Institute of Technology, opened his presentation by acknowledging that machine learning has enabled great progress in solving difficult problems in recent years (e.g., super-human classification, object detection, machine translation, game-playing bots, self-driving cars testing on public roads). Much of this has been enabled by deep learning. He elaborated that although machine learning systems achieve great "average-case" performance on these difficult problems, machine learning is fragile and has terrible performance in "worst-case" situations, such as adversarial or security-sensitive settings. An audience participant said that machine learning is performing well in test cases and wondered how Athalye defined average. Athalye responded that average performance indicates a level of comfort with deployment.

Athalye explained that imperceptible perturbations in the input to a machine learning system could change the neural network's prediction and dramatically affect the model and its output. A small adversarial perturbation can lead a state-of-the-art machine learning model to mislabel otherwise identifiable images. These attacks can be executed with just a small number of steps of gradient descent. An audience participant wondered whether when looking at the difference between the original image and the adversarial example and magnifying noise if it is possible to see anything interpretable. Athalye responded that while noise may be more interpretable with more robust machine learning models, in these cases only noise is perceptible.

Athalye described machine learning as not being robust for image classification and for many other problems such as semantic segmentation, reading comprehension, speech-to-text conversion, and malware detection. Attackers can tweak metadata to produce functionally equivalent malware that can evade a detector. An audience participant wondered if training models using adversarial examples could help build in a resistance to attacks. Athalye noted that this process has been shown to increase robustness to some degree, but current applications of adversarial training have not fully solved the problem, and scaling is a challenge.

Athalye noted that contradictory evidence exists as to whether real systems are at risk from adversarial examples. He described a study in which natural image transformations were applied to adversarial examples, thus breaking the adversarial example and classifying the image correctly in its true class. However, adversarial examples can be robust, and this approach does not always work.

He presented a basic image-processing pipeline in which an attacker gains control of an image, the image is fed into a machine learning model, and the resulting predictions are affected. However, the physical-world processing pipeline looks a bit different: a transformation with randomized parameters occurs between the image and the model. In this case, the attacker no longer has direct control over the model input (see Figure 4.2).

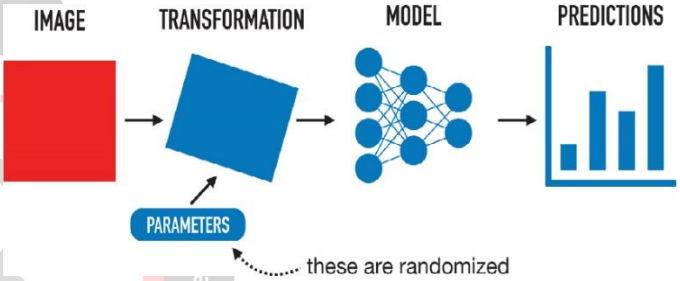

Fig 3: A physical-world image processing pipeline.

SOURCE: Anish Athalye, Massachusetts Institute of Technology, presentation to the workshop, December 11, 2018.

An attack is still possible in this real-world setting because, even though one does not know what the exact transformation will be, the distribution of transformations is known. The transformation needs to be differentiable, he continued, and instead of optimizing the input to the model, one can optimize over all possible transformations to find a single point that no matter how it is transformed will confuse the model in all settings. This approach (i.e., expectation over transformation), which still uses gradient descent, can produce real-world robust adversarial examples.

Athalye then considered a three-dimensional (3D) processing pipeline, which is similar to the real-world pipeline. He explained that for any pose, 3D rendering is differentiable with respect to texture. He demonstrated a 3D adversarial object where no matter how the object is rotated, the machine learning model still classifies it incorrectly. In response to an audience participant, Athalye confirmed that, in this example, only the texture of the object was changed.

 DOI : 10.35291/2454-9150.2022.0401

He said that other researchers have also tested manipulating geometry rather than texture to construct adversarial objects. Athalye reiterated that machine learning is not robust in controlled and real-world settings, noting that even in a world filled with noise, the models can still be fooled.

Black-box models, too, are susceptible to adversarial attacks, Athalye explained. In a black-box threat model, the attacker has no visibility into the details of the model—the attacker can only construct an image, feed it in, and watch what emerges. Gradient descent is still used, but now so too is an estimate using queries to the classifier. Even more restricted settings, such as the Google Cloud Vision Application Programming Interface, are vulnerable to adversarial attacks, Athalye continued. He noted that in the hundreds of papers published on defenses for adversarial attacks, many of the proposed defenses lack mathematical guarantees. He added that many defenses submitted to the 2018 International Conference on Machine Learning (ICML)4 were not robust, confirming that defending against adversarial attacks is a difficult problem. In closing, Athalye emphasized that robustness is a real-world concern because attacks are outpacing defenses. He said that it is crucial to understand the risks of adversarial attacks to current systems through rigorous evaluations and a principled approach that will lead to the construction of secure machine learning systems.

Chellappa suggested that Athalye revisit his conclusion that machine learning defenses are not robust, given that the MNIST database5 (Modified National Institute of Standards and Technology database) of handwritten digits was labeled with 55 percent accuracy. He emphasized the value of describing both when systems are working and when they are not. Chellappa also suggested that if a forensic examiner is working alongside a machine learning algorithm, many of the real-world problems Athalye described could disappear. He added that with knowledge of time series models from the 1970s, the concept of adversarial examples should not come as a surprise to anyone. Another audience participant discussed the black-box inversion that results from making many queries against a model and the characterization that is needed to produce an adversarial example. Athalye noted that although he is not aware of much research in this area, some are working on decreasing the number of queries required.

CONCLUSION

- It has been found empirically that adversarial examples can impact a deep learning convolutional network in misclassification of input images.

- Adversarial examples attack is direct attack on deep learning systems on the application layer. Mitigation on different layers of the network stack must be adhered to particularly on networked systems

- Adversarial examples are transferable from one model to another

- Adversarial training is considered as the most effective mitigation

- Standardized code hardening guidelines must be implemented in a global scale to reduce the risk of adversarial examples

References

[1] Anna Gerber, J. R. (2017, May 22). *Connecting all the things in the Internet of Things*. Retrieved from IBM Developer: developer.ibm.com/articles/iot-lp101-connectivity-network-protocols/

[2] Chen, P.-Y. (2018, May 2). *Clever Adverserial Attack*. Retrieved from IBM: www.ibm.com/blogs/research/2018/05/clever-adversarial-attack/

[3] Clark, I. G. (2017, February 24). *Adversarial Example Research*. Retrieved from Openai: https://openai.com/blog/adversarial-example-research/

[4] Deng, L. (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine*. Retrieved 8 9, 2022, from https://microsoft.com/en-us/research/publication/the-mnist-database-of-handwritten-digit-images-for-machine-learning-research

[5] Jia Deng, W. D.-J.-F. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Xplore.

[6] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble Adversarial Training: Attacks and Defenses. *arXiv: Machine Learning*. Retrieved 8 9, 2022, from https://arxiv.org/abs/1705.07204

[7] Tsui-Wei Weng, H. Z.-Y.-J. (2018). Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. *Sixth International Conference on Learning Representations (ICLR 2018)*. Vancouver: ICLR.

[8] Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., & Song, D. (2018). Generating Adversarial Examples with Adversarial Networks. *arXiv: Cryptography and Security*. Retrieved 8 9, 2022, from https://arxiv.org/abs/1801.02610

[9] Xiaoyong Yuan, P. H. (2017). Adversarial Examples: Attacks and Defenses for. *National Science Foundation Center for Big Learning*.

[10] Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. L. (2018). *Mitigating Adversarial Effects Through Randomization*. Retrieved 8 9, 2022, from https://openreview.net/forum?id=sk9yuql0z

 DOI : 10.35291/2454-9150.2022.0401