GENERATE ADVERSARIAL EXAMPLE USING GENERATIVE ADVERSARIAL NETWORK

Jyoti Yadav, Mrs. Deepti Panday, Dr. Sanjeev Gangwar

M.TECH Scholar , Dept. Of Computer science and engineering
Veer bahadur singh purvanchal university(u.p.)
Email:-yadavjyoti945169@gmail.com
(Assistant professor)
Computer science and engineering
Veer bahadur singh purvanchal university(u.p.)
Email:-sanjeevani111.d@gmail.com
(Assistant professor & Head In-charge C.S.E & I.T)
Computer science
Veer bahadur singh purvanchal university(u.p.)
Email:-
Gangwar.sanjeev@gmail.com

Abstract**—** *with rapid-fire progress and significant successes in a wide diapason of operations, deep literacy is being applied in numerous safety-critical surroundings. A vulnerability of Deep Neural Network is Adversarial Example. Adversarial example are input crafted by adding small perturbation to cause misclassify. The traditional method of generating adversarial example are computationally and bulky and slow. Generative adversarial Network (GAN) is fast generated adversarial example and more quality full. Train the model more epochs in Generative Adversarial Network gives the quality of adversarial example. We use MNISTand CIFAR -10 Dataset. Tensorflow used in backend.*

Index Terms**——** *Adversarial example, Deep neural network, Generative Adversarial Network.*

DOI: 10.48047/ecb/2023.12.8.777

I.INTRODUCTION

In recent year Deep Neural network (D.N.N) have been extensively used in various scenarios speech recognition [1], image recognition [2], virtual assistant, self- driving car, visual recognition, Pixel restoration, Photo description deep dreaming. A vulnerability of Deep Neural Network is Adversarial Example. Adversarial example can easily fool the Deep Learning model by perturbing benign sample without being exposed by human. Perturbations that are invisible to human vision are sufficient to prompt the model to make a wrong prediction with high confidence. Adversarial example is basically use in testing of model. Adversarial example can be applied in many fields such as target recognition, automatic driving, Intrusion Detector System (IDS) and other application.

Adversarial example came in 2013[6] by Goodfellow " Intriguing properties of neural network ".These are input crafted by adding small perturbations to cause the misclassify. We have image X and classifier F that produces a label.

$$F(X) = Y$$

So then we add a perturbation (sigma symbol X) to misclassify

$$F(X' + \text{SIGMA } X) = Y^*$$

The most representative method FGSM (Fast gradient sign method) and these various variants. This kind of method generates the adversarial example by accessing the gradient information of target model and adding the calculated perturbation to the original example. But this method requires huge computation and most of them are white box attack and the specific information is needed when calculating the perturbation, therefore the applicability of such attack method is relatively narrow.

Now a day, a tremendous amount works has been done to the adversarial example attack by using the deep learning such as using the Generative Adversarial Networks to generate the adversarial example. Comparing with the traditional attacks the Adversarial examples generated by GAN have higher success rate and faster generated generation speed, and the example also have higher transfer and anti-robustness.

II.LITECHER REVIEW

Through our test, we show that the projected method can generate random untargeted adversarial examples. Our search can be extended to the audio and video domains as a future study. Another challenge will be to create a countermeasure for the proposed method.

Yuan Xiaoyong at all(2018) The idea of exploration paper recent findings on Adversarial Example for deep neural networks, sum up the styles for generating Adversarial Example , and suggest a taxonomy of these styles. Under the taxonomy, operations for Adversarial Example are delved.

We further intricate on countermeasures for Adversarial Example. In addition, three major challenges in Adversarial Example and the possible results are bandied. Datasets- MNIST, CIFAR- 10, and ImageNet are the three most extensively used image bracket datasets to estimate Adversarial attacks. system-. L- BFGS Attack, Fast grade subscribe Method(FGSM), Basic Iterative Method(BIM) and Iterative Least-Likely Class Method(ILLC), Jacobian-grounded Saliency Map Attack(JSMA),E. DeepFool,G. C&W's Attack, H. Zeroth Order Optimization(ZOO) This paper tried to cover state- of- the- art studies for Adversarial Example in the deep learning area. Compared with recent work on Adversarial Example , we anatomized and bandied current challenges and implicit results in Adversarial Example .

Xu Han, at all(2020) The ideal of the paper give a methodical and comprehensive review on the launch- of- the- art algorithms from images, graphs and textbook sphere, which gives an overview of the main ways and benefactions to Adversarial attacks and defenses. system- Biggio ′ s attack, Szegedy ′ s limited- memory BFGS(L- BFGS) attack, Fast grade sign system(FGSM), Deep Fool, Jacobean- grounded saliency chart attack, Basic iterative system(BIM)/ Projected grade descent(PGD) attack, Carlini & Wagner ′ s attack, Ground verity attack, Other attacks, Ground verity attack, Spatially converted attack, Unrestricted inimical exemplifications The current state- of- the- art attacks will probably be annulled by new defenses, and these defenses will latterly be circumvented. We hope that our work can exfoliate some light on the main ideas of adversarial literacy and related operations in order to encourage progress in this field. Jin Xu at all(2020) In this composition, we propose an attack algorithm grounded on Nesterov- instigation called Nesterov-instigation iterative fast grade sign system(NMI- FGSM). Nesterov- instigation makes a correction when the grade is streamlined to avoid moving too presto. trials show that our algorithm performs well and has achieved a high success rate. system- FGSM, BIM and MI- FGSM algorithm. Dataset- ImageNet, unborn work- inimical exemplifications has concentrated more on white- box attacks. The perpetration of black- box attacks relies on the transferability of the adversary exemplifications.

LIUJIANYI, TIANYU at all (2020) The ideal of the exploration paper a two- stage generative inimical networks(TSGAN) with semantic content constraints is proposed in this paper. The first- stage uses the original illustration dataset to train creator G, which can help the creator learn the distribution of real examples. Model- Adversarial Example generated by the STGAN with 77% accuracy. Dataset- MNIST and CIFAR- 10, How to further ameliorate the attack success rate of the attack model on the base of icing the quality of generated exemplifications is our unborn work, and how to induce some inimical exemplifications with a series of translated images might be an intriguing work.

SunLu, at all(2018)- Case study works against face recognition systems and road sign recognition systems eventually abridged the gap between theoretical Adversarial Example generation methodologies and practical attack schemes against real systems.and also assay the restrictions and crucial procedures for launching real world Adversarial attacks. system- Mounting system, Discovery styles. We believe the consequences of practical Adversarial attacks would be severe if the principles behind the attacks aren't made clear. To avoid so, experimenters must untangle all possible anxiety mounting vectors and system contrivers must attach enough attention to Adversarial Example when integrating AI models.

CHOI SEOK- HWAN at all(2022)- The performance of the state- of- the- art generative Adversarial networks- grounded defense styles is limited because the target deep neural network models with generative Adversarial networks-grounded defense styles are robust against Adversarial Example but make a false decision for licit input data. To break the delicacy declination of the generative Adversarial networks grounded defense styles for licit input data, we propose a new generative Adversarial Network- grounded defense system, which is called Adversarial Robust Generative Adversarial Networks(ARGAN). system-proposed- MNSIT, CSIR 10, ARGAN handed robustness to target DNN models against colorful state- of- the- art Adversarial attacks while maintaining the high delicacy indeed for licit input data. Also, it's observed that ARGAN showed better performance than the state- of- the- art GAN-grounded defense styles similar as Gandef, DefenseGAN and Ham- GAN. Toomas Liiv and at all The ideal of paper three DNN image bracket algorithms are constructed. Two inimical styles(IFGM and DeepFool) are also anatomized and compared in terms of chancing the lowest anxiety to beget misclassification of the networks, with the least Computational trouble. The two iterative, grade- grounded styles are enforced in four distance criteria L0, L1, L2, L .Dataset- CSIR10, MNSIT. Result- the performance of two iterative inimical attacks grounded on the grade of the labors(DeepFool) and the grade of the cost functions(IFGM). In short, we find that their performance is nearly identical in the L0, L2 and L ∞ morals, but not in L.

III.THE EXPERIMENT AND ANALYSIS

A) DATASETS

We used MNIST [3] AND CIFAR [4] datasets in this paper .MNIST dataset is classic handwritten numeral dataset in the field of machine learning. And categories in the dataset 0 through 9, each example are a 28×28 pixel grayscale image. CIFAR-10 dataset is also a classic dataset in the field of machine learning. There are 10 categories and each example is a 32×32 RGB image. For MNIST 60000 training data and 10000 test data were used. And for CIFAR-10 50,000 training Data and 10,000 test data were used.

TABLE I: The setting of MNIST dataset

Batch size	128
Normal Learning Rate	2E-04
Adversarial learning rate	0.0002
Epoch in Normal training	100
Epoch in Adversarial Training	50
Number of Example to generation	16
Noise dim	100

Optimizer	Adam
Image size	28,28,1

In table 1 we use MNSIT Dataset and show different input parameter and value.

TABLE II: The setting of CIFAR-10 dataset

Batch size	128
Normal Learning Rate	2E-04
Adversarial learning rate	0.0002
Epoch in Normal training	100
Epoch in Adversarial Training	50
Number of Example to generation	16
Noise dim	100
Optimizer	Adam
Image size	32,32,1

In table 2 we use CIFAR-10 Dataset and show different input parameter and value.

IV.METHODOLOGY

In this Paper we generated the Adversarial Example by Generative Adversarial Network (GAN). It is good technique to generate adversarial example fast and success rate is very high. We used qualitative method. we used two type of dataset MNIST & CIFAR-10 by Online and used semantic analysis. We used coding for Google colab. Google colab is an online platform to run the program. It provides good speed and more storage.

A. GENERATIVE ADVERSARIAL NETWORK

Generative adversarial network are one of the most curious ideas in computer science today. Two models are trained simultaneously by adversarial process. A generator learns to create image that look real, while a discriminator learn to tell real images apart from fakes.
Mathematically representation of Generative Adversarial example,

$$\min_G \max_D V(D,G)$$

$$V(D,G) = E_{x\sim Pdata(x)}[\log D(x)] + E_{z\sim Pz(Z)}[\log(1 - D(G(z))]$$

Whereas,
G = Generator
D = Discriminator
Pdata(x) =distribution of real data
P (Z) = distribution of generator
X=Sample from Pdata(x)
Z=Sample from p (z)
D(X) = Discriminator Network
G (Z) = Generative Network

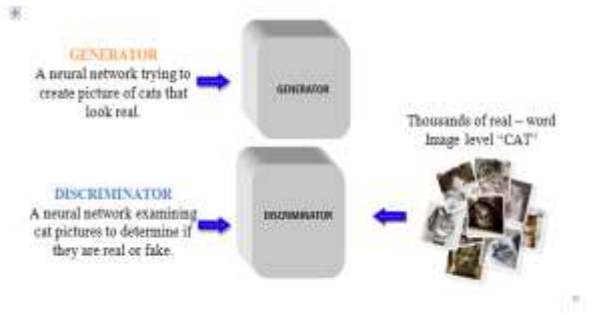


Fig.1. Generative Adversarial Network

B. Training Strategies

Mini-batch Discrimination: Introduces diversity regularization by comparing samples within a mini-batch to discourage mode collapse and promote diverse sample generation. Feature Matching: Instead of training the discriminator to classify real and generated samples, the generator is trained to match the statistics of intermediate features extracted by the discriminator. This encourages the generator to generate samples that are similar to real data in terms of feature representations. Two-Time-Scale Update Rule (TTUR): Utilizes different learning rates for the generator and discriminator to balance their training dynamics. This strategy helps stabilize the training process and improve convergence. Regularization Techniques: Various regularization techniques, such as weight normalization, spectral normalization, and gradient penalty, have been proposed to improve training stability and prevent mode collapse

C. Challenges faced by generative adversarial network

1. Problem of stability between generator and is Discriminator.
2. Problem to determine positioning of the object.
3. Problem in understanding the global objects.
4. Problem in understanding the perspective.

D.Applications of GANs

1 Image Synthesis and Translation
2 Super-Resolution and Inpainting
3 Text Generation and Machine Translation
4 Healthcare and Medical Imaging
5 Privacy Protection and Data Augmentation
6 Anomaly Detection and Fraud Detection

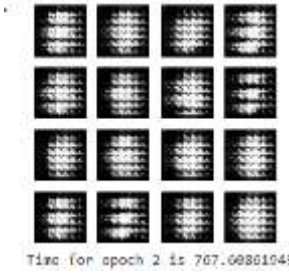
V.RESULT

In this research, we conducted experiments using the MNIST and CSIR-10 datasets to evaluate the performance of a generative adversarial network (GAN). The model was trained for 50 epochs, and the quality of the generated images was assessed. Upon analyzing the initial results after the first epoch, we observed that the quality of the generated images was not satisfactory. However, we proceeded to investigate the impact of increasing the number of epochs on the quality of the generated images. Our findings indicate that as we increased the number of epochs, the quality of the generated images improved significantly. The images became more

realistic and visually appealing, indicating that longer training durations allowed the GAN to learn and capture more intricate details from the datasets. This observation demonstrates the effectiveness of utilizing extended training periods to enhance the quality of generative adversarial examples.

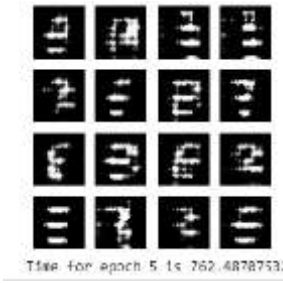
In first image we analysis time of generate image and quality of image is very bad. We not understand what is show in image.

In second image generation time is same but quality is better than first image.



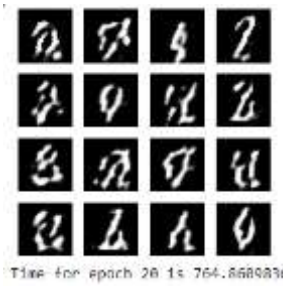
In third image generation time 762.4870753288269 sec.

Quality is better than second image.

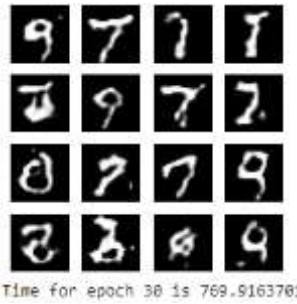


In fourth image generation time is 767.4116318225861 sec.Quality is better than third image.

In fifth image generation time is 764.8609836101532 sec.Quality is better than forth image.



In sixth image generation time is 769.9163701534271 sec.Quality is better than fifth image.



In seventh image generation time is 768.8609836101531 sec.Quality is better than sixth image.

VI.DISCUSSION

Prevent by Adversarial attacks involves developing high level risk assessment and planning a holistic cyber security approach based on the assessment. Most crucially if the model is will be deployed into a high risk input space all stockholder must be aware of the possible threats to the model.

VII.CONCLUSION

this paper introduces a Generative Adversarial Network (GAN) as a method for generating adversarial examples. The GAN framework utilizes unsupervised learning and consists of two main components: a generator and a discriminator. The generator is responsible for creating the adversarial examples, while the discriminator's role is to distinguish between the generated examples and the real ones. The results of our experiments demonstrate that the proposed GAN achieves a remarkable success rate of 100% in generating adversarial examples. Furthermore, we observe that increasing the number of training epochs positively impacts the quality of the generated examples. This finding suggests that longer training durations enhance the GAN's ability to produce more convincing and effective adversarial examples. Overall, our research presents a powerful approach for generating adversarial examples using a GAN. The high success rate and the impact of extended training on example quality highlight the potential of this method in adversarial

attacks and defense strategies. Further research can explore variations of GAN architectures and training techniques to improve the generation process and address potential limitations in generating diverse and robust adversarial examples.

VIII.FUTURE SCOPE

Generative Adversarial Networks (GANs) have already made significant contributions to the field of generative modeling and have demonstrated their efficacy in various applications. However, there are still numerous avenues for future research and advancements in GANs. This section highlights some potential areas for exploration and discusses the potential applications of GANs in emerging domains.

REFERENCES

([1]G. T. Tsenov and V. M. Mladenov, "Speech recognition using neural networks," in 10th Symposium on Neural Network Applications in Electrical Engineering, Sep. 2010, pp. 181–186.

[2] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 2261–2269.

[3] Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist, 2, 2010.

[4]. Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset.online: http://www. cs. toronto. edu/kriz/cifar. html, 2014.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," In ICLR, 2015.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[7] Z.Zhao,D.Dua,andS.Singh,''Generatingnaturaladversarialexamples,'' in Proc. Int. Conf. Learn. Represent. (ICLR), Vancouver, BC, Canada, 2018, pp. 1–15.

[8] Nguyen, J. Yosinski and J. &Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in Proceedings of the IEEE conference on computer vision and pattern recognition, (pp. 427-436), 2015.

[9]Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In Security and Privacy (SP), 2017 IEEE Symposium on, pages 39–57. IEEE, 2017.

[10]J. Kos, I. Fischer, and D. Song. Adversarial examples for generative models.arXiv preprint arXiv:1702.06832, 2017.

[11]Tim Salimans, Ian Goodfellow, WojciechZaremba, Vicki Cheung, Alec Radford, and Xi Chen.Improved techniques for training gans.arXiv preprint arXiv:1606.03498, 2016.

[12]A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.

[13]A.Kurakin,I.Goodfellow,andS.Bengio,"Adversarialmachinelearning at scale," In ICLR, 2017.

[14]A.J.BoseandP.Aarabi,''Adversarial attacks on face detectors using neural net based constrained optimization, 'in Proc. IEEE20thInt.WorkshopMultimediaSignalProcess.(MMSP),Vancouver,BC,Canada,Aug.2018, pp. 1–6.

[15]Liu,Y.;Chen,X.;Liu,C.;Song,D.Delvingintotransferableadversarialexamplesandblack-box attacks. arXiv2016,arXiv:1611.02770

[16]Moosavi- Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.

[17] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash and T. S. D. Kohno, "Robust physical-world attacks on deep learning visual classification,"