Deep Learning Adversarial Examples - A Survey

Shivam Pandit
Clemson University
pandit@clemson.edu

***Abstract*—Deep Learning is nowadays a very important field being applied in solutions concerning artificial intelligence. It is very useful in designing systems that can evolve on daily basis and learn from data. Rather than programming systems explicitly, it works on gaining knowledge from large amounts of data and using artificial intelligence to make a sense out of it. It has enabled us to make applications in areas of computer vision like self-driving cars as well as in solving problems that are too complex for humans through deep learning. However, with growing complexity and amounts of data being considered these systems are found to be vulnerable to adversarial examples which are nothing but inputs that are intentionally designed by attackers to fool the system and lead it to a mistake. So, attacks and defenses of such systems is of paramount importance since these systems could be working in some safety critical environments where small deviation from the actual outcome can be a catastrophe. Thus, they pose a serious threat to advancement in deep learning applications in practice. This has led to more research in deep learning problems while studying ways to withstand any sort of adversaries. In this survey paper, we will be discussing about deep learning adversaries, main techniques to solve them, issues and problems concerning these systems and also future trends in this area.**

I. Introduction

Deep learning [1] has led to recent breakthroughs in computer vision and speech recognition systems. It is enabling us to solve problems that seemed very complex in the past. It is empowering security critical systems like malware detection, face recognition and self-driving cars. Deep learning is nowadays being used at an unprecedented scale in domains of machine learning, e.g. brain circuit reconstruction [2] DNA mutation analysis [3] Potential drug molecules prediction of structure-activity [4], particle accelerator data Analysis [5]. Since, deep learning is mostly based on data driven approach, it requires less hand engineered features i.e human intervention as well as less use of domain knowledge since system takes decision on its own.

Deep learning is gaining importance at a much higher pace where constantly number of applications powered by it are becoming part of our daily lives. For instance, Apple and other smartphone brands have now face authentication to unlock mobile phones [6] Some ATMs have introduced face recognition for biometric authentication [7] Behavior based malware detection solutions powered by deep learning [8] and even big companies like Google, Tesla and Uber are testing self driving cars which require plenty of deep learning models.

Even though its uses and applications are immense but since it is used in life crucial events too raises great concerns on its reliability. For instance a self driving car can misread a traffic signal or a sign leading to some accident which highlights the importance of its consistency across all inputs. Moreover, recent studies have proven how these systems are vulnerable to adversarial inputs that would seem imperceptible to humans. These well-designed input samples fool the system resulting in adverse outcomes.

Computer Vision community has made several contributions since 2012 leading to advancements in Medical Science [9] to Mobile Applications [10]. With availability and improvements in neural models [11], access to deep learning libraries and hardware availability needed to train complex models, deep learning is now working on safety critical applications like malware detection, self-driving cars [12], face recognition [13], speech recognition [14] and even drone and robotics.

No doubt that computer vision applications do the task with remarkable accuracies but Szegedy et al discovered vulnerability of the state of the art system when he introduced small perturbations in the original input that successfully fooled the system which is called as adversarial examples. This weakness of the system in terms of image classification raised alarms on the need of system robustness to these types of adversarial examples. For instance, an adversary can create such examples by manipulating stop sign to fool the autonomous systems in cars or generate bad adversarial commands against speech recognition systems in applications like Amazon Alexa, Windows Cortana, or Apple Siri.

Deep Neural Networks are usually trained on fixed set of data so any transformation like changing of orientation or pixels drastically can lead the system to misclassify the image. Even that change will be not noticeable to a human eye but deep learning system gets easily confused and fooled by those unnecessary noises added to the original image. Moosavi Dezfool et al demonstrated that universal perturbations can fool any deep learning system [15]. For instance, an adversarial input overlaid on a typical image on panda can fool a system to misclassify it as a gibbon.

Deep Learning is commonly called as a black box technique which usually performs well but with limited knowledge of the reason [16]. Many studies are done on the domain of deep neural networks. From inspecting adversaries, we gain knowledge of deep semantics of inner level of deep neural networks to demarcate system boundaries and in turn help to increase robustness and performance of system in consideration improving interpretability.

Keeping in view, the importance of deep learning in computer vision, we will try to investigate and summarize approaches for generating adversarial examples and defence systems to tackle the same. Since most of the recent advancements

are mostly in the domains of supervised learning especially in computer vision models. Therefore, most adversarial examples are produced in computer vision models. This survey paper will mostly focus on adversarial attacks on deep learning in computer vision.

II. Main Techniques

In this section we briefly describe main techniques in deep learning and various approaches to adversarial examples.

A. Main Concepts

Deep learning method is based on computers ability to learn from experience and knowledge extraction from data without any explicit programming. Thus it is very helpful since even people who don't have experience in passing commands to system can take use of the system. Conventional machine learning algorithms lack due to lack of computational power, domain and expert knowledge [17]. Unlike deep learning which requires just data that it converts to multiple features that are simple to understand but represent a sophisticated model. Deep learning based image classification represents an object by describing edges, fabrics and structures in layers. With the help of hardware acceleration it has solved many computational problems. Perceptrons(artificial neurons) makes a neural network [18]. Nowadays, Convolutional neural networks (CNNs) and Recurrent neural networks (RNNs) are two most widely used neural networks. CNNs work on fixed sized input and generate a fixed sized output whereas RNNs work on arbitrary input and output lengths. CNNs are based on feed forward artificial neural networks whereas RNNs use their internal memory to process arbitrary sequence of inputs.

B. Architectures of Deep Neural Networks

Some popular deep learning architectures used in computer vision include: GoogLeNet[], AlexNet[], VGG[], ResNet[] and LeNet[] [19].These architectures can be seen as milestones in image classification models. These models are also used by attackers to generate adversarial examples.

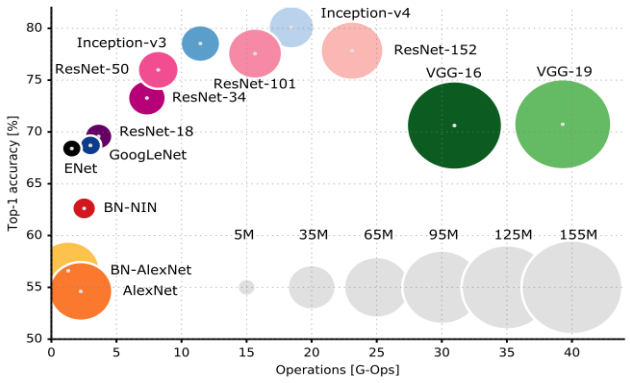

Fig. 1. Source: Eugenio Culurciello, An Analysis of Deep Neural Network Models for Practical Applications, arXiv:1605.07678

C. Deep learning datasets

The three widely used datasets in computer vision tasks are MNIST, CIFAR-10 and ImageNet.Handwritten digits recognition is carried by MNIST dataset whereas image recognition task is carried by CIFAR-10 and ImageNet. There are about 60,000 tiny color images in CIFAR-10 whereas 14,197,122 images with 1,000 classes in ImageNet.

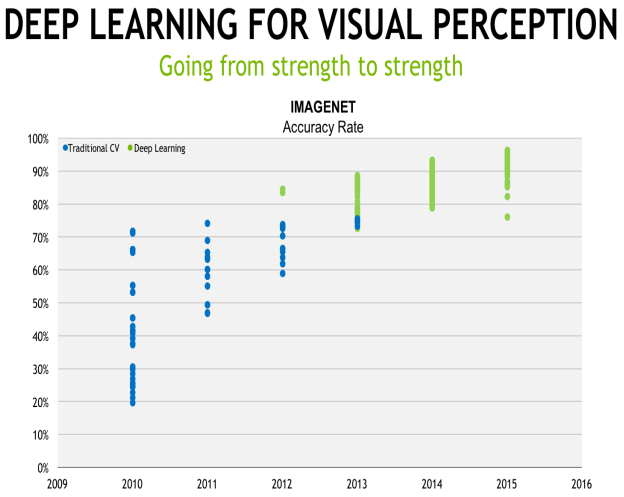

Fig. 2. ImageNet Accuracy Rate Source: NVIDIA Blogspot 2016

D. Adversarial Examples and countermeasures

Adversarial examples primarily target handcrafted features such as spam filters, intrusion detection, fraud detection, etc. Barreno et al. on his initial investigation categorised machine learning system attacks into three axes: 1) Influence: If the attack can poison the training data. 2) Security violation: whether adversarial example belongs to false positive or false negative. 3) Specificity: If attack is limited or covers wide class [19].We categorize the approaches for generating adversarial examples in three dimensions: threat model, perturbation and benchmark.

1) Threat Model: Different adversaries, scenarios, quality requirements decide threat model which has four aspects: adversarial falsification, adversarial knowledge, adversarial specificity and attack frequency. False positive attacks generate a negative sample i.e an image unrecognizable to human but predicted by neural network to a class with high confidence score. False positive sample generates positive sample which can be in case of malware detection task inability of trained model to identify the malware. Here, human can recognize the image but neural networks fail to recognize [20]. Adversary's knowledge comprise on white box and black box attacks. White-box attacks assumes adversary has knowledge of everything related to trained data whereas black box attack assumes

adversary has no access to trained data. Adversarial Specificity comprise of targeted and non-targeted attacks. Targeted attacks misguide deep neural networks to think all examples belong to same class like in face recognition system, an adversary tries to disguise some face as authorized user [21]. Non-targeted attacks do not assign specific class but assign some random except original one mostly used in evading detection. Lastly, attack frequency comprise of one-time or iterative attacks. One-time attacks take just one time to optimize/update the adversarial examples but iterative attacks take multiple time to do the same.

2) Perturbation: This is most important as small perturbation is basis for adversarial examples. These are very similar to actual samples but imperceptible to humans.It has three aspects: perturbation scope, perturbation limitation and perturbation measurement [22]. Perturbation scope has individual attacks which produces different perturbation for inputs and universal attacks produces universal perturbation for the dataset. Perturbation limitation comprise of optimized perturbation and constraint perturbation.

3) Benchmark: Adversaries show performance based on models and data sets. Large and complex high quality data sets makes it very difficult to tell if adversarial examples exists due to dataset or models.The three commonly used datasets are MNIST, CIFAR-10 and ImageNet [23]. They help in evaluating adversarial attacks.

E. Adversarial Classification

There are several approaches for generating adversarial examples. Some have countermeasures but still they help in improving robustness of deep neural networks.

1) L-BFGS Attack: Szegedy et al. showed small perturbations fooled the deep learning models into misclassification [24]. In this perturbed image looked quite similar to human eyes but successfully fooled the neural network. It also cascaded to multiple neural networks acting like a blind spot in deep learning.

2) Fast Gradient Sign Method (FGSM): Unlike L-BFGS Attack which used expensive linear search method for finding optimal value, Goodfellow et al. proposed a method that was fast since it only performed one step gradient update along the sign of gradient [25]. It exploited linearity of deep learning models at a time when these models were thought to be non-linear. It was also found that network can be fooled by using some random target class but also adversarial training will improve the robustness of networks against attacks by FGSM.

3) Basic and least likely class iterative method: Since in some applications, data passes through devices eg. Cameras unlike previous methods where it was fed directly to system. Here, they image is perturbed by taking single large step to increase loss of classifier [26]. This is used in standard convex optimization method.

4) Jacobian-based Saliency Map Attack (JSMA): JSMA is an efficient saliency adversarial map.Small perturbation was designed to successfully induce large changes in output to fool the neural network [27].Here, authors uses two sasliency maps to select pixel for each iteration and achieved 97% success by modifying only 4% input feature but had high computational costs.

5) DeepFool: It was proposed to find the closest distance between given input and adversarial example boundary. Deep fool starts with a clean image and perturbs it a little at each iteration to take it at boundary of polyhydron [28]. It uses iterative attack to overcome non-linearity in high dimension with a linear approximation. It is useful when perturbation are less than computed by FGSM with similar fooling ratio.

6) C&Ws Attack: Carlini and Wagner targeted defensive distillation making it effective for most of the adversarial defenses [29]. Moreover, the adversarial examples generated over un-secured (undistilled) network transfers over to secured(distilled) network which we can use for black-box attacks.

7) One Pixel Attack: This is one of the extreme cases where just one pixel change of the original image fools the classifier [30]. Sue et al. fooled three networks with 70.97% images successfully fooling the system and average confidence of the system on wrong label was found to be 97.47%

III. Issues and Problems

Most of the defences are targeted at computer vision task. However, with the advancement and sophistication of even attackers, robust defences are the need of the hour for the safety critical systems. We will now be discussing issue and problems that act as challenges in getting potential solutions for adversarial examples. Reason for existence of those adversarial problems is very interesting and fundamental to the defense mechanism since it exploits the neural network vulnerability. Major issue with adversarial problems is the issue of transeferabilty which we will discuss now.

A. Transferability

It is very common problem in adversarial examples. Szegedy et al. observed that adversarial examples generated against a neural network are able to fool same neural network trained with different dataset [31]. It was observed that even different architectures trained by different machine learning algorithms were easily fooled. This is critical in black-box attacks where attacker generates adversarial examples on some substitute dataset and due to transferability victim model also gets affected. Transferability is of three levels from easy to hard depending on neural network architectures and dataset in consideration. These adversarial examples are worst nightmare since they can be transferred between different training sets, parameters and even across systems operating on different machine learning algorithms. Non-targeted adversarial examples were more transferable than targeted ones.

B. Why Adversarial Examples exists??

The reason for the existence of adversarial problems is still an open question even though many studies and hypothesis are presented in this regard. One of the reason is data incompletion. Even for simple models a corresponding robust

model can be more complicated requiring much more training data sets. Another reason is model capability since models are generally too linear in high dimensional space. Low flexibility of the classifier is also one reason which makes no clear decision boundaries for the system making classifier erratic [32]. Currently we have no universal attacking or defending method which can be applied to all applications. Robustness of the system is very important since most of the proposed defense methods in the past have failed or proven vulnerable to new attacks. Some even failed after slight change in attack which emphasizes the importance of evaluation of robustness of the system. Mostly, we focus on defending existing attacks which is not correct since nowadays there are also zero day attacks that are most harmful for neural networks. Secondly, lack of publicly available code for attacks and defenses makes it even more difficult to maintain a benchmark platform that would help in reproducing solutions against existing attacks/defenses. Different researchers can get different conclusion from the same problem due to lack of any benchmark.

IV. Future Trends

Deep Learning is constantly evolving and computer vision is increasingly adopted in many daily applications. Even though security remains a concern due to adversarial examples and various attacks that can happen. Even then it remain focus of attraction of all as it has immense scope of improvements and research. It will enable us to do things like never before including autonomous applications like self-driving cars which are already in beta and also autonomous homes. Some trends we can expect in future includes benchmarking tools, native support, adopting open analytics ecosystem, incorporating simplified programming frameworks to enable faster coding, toolkits supporting reusable components, etc. Natural language processing will also improve which is currently not up to the mark. AI will help developers to better understand user needs and thus leading to best possible user interface for them. Even advertising will be revamped using deep learning that will save investors money and at the same time target appropriate audience.

References

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

[2] LeCun, Yann and Bengio, Yoshua and Hinton, Deep learning Geoffrey, pp.436,2015

[3] Alipanahi, Babak, et al. "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning." Nature biotechnology 33.8 (2015): 831.

[4] Jorgensen, William L. "The many roles of computation in drug discovery." Science 303.5665 (2004): 1813-1818.

[5] Roe, Byron P., et al. "Boosted decision trees as an alternative to artificial neural networks for particle identification." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 543.2-3 (2005): 577-584.

[6] Hadid, Abdenour, et al. "Face and eye detection for person authentication in mobile phones." 2007 First ACM/IEEE International Conference on Distributed Smart Cameras. IEEE, 2007.

[7] Findling, Rainhard D., and Rene Mayrhofer. "Towards face unlock: on the difficulty of reliably detecting faces on mobile phones." Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia. ACM, 2012.

[8] Yuan, Z., Lu, Y., Wang, Z., & Xue, Y. (2014, August). Droid-sec: deep learning in android malware detection. In ACM SIGCOMM Computer Communication Review (Vol. 44, No. 4, pp. 371-372). ACM.

[9] Maintz, JB Antoine, and Max A. Viergever. "A survey of medical image registration." Medical image analysis 2, no. 1 (1998): 1-36.

[10] Lane, N. D., & Georgiev, P. (2015, February). Can deep learning revolutionize mobile sensing?. In Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications (pp. 117-122). ACM.

[11] Grill-Spector, Kalanit, Richard Henson, and Alex Martin. "Repetition and the brain: neural models of stimulus-specific effects." Trends in cognitive sciences 10.1 (2006): 14-23.

[12] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... & Zhang, X. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

[13] Zhao, Wenyi, et al. "Discriminant analysis of principal components for face recognition." Face Recognition. Springer, Berlin, Heidelberg, 1998. 73-85.

[14] Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science, 270(5234), 303-304.

[15] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." Neural networks 61 (2015): 85-117.

[16] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint.

[17] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." European conference on machine learning. Springer, Berlin, Heidelberg, 1998.

[18] Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." Psychological review 65.6 (1958): 386.

[19] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).

[20] A. Raghunathan, J. Steinhardt, P. Liang, Certified Defenses against Adversarial Examples, arXiv preprint arXiv:1801.09344. 2018.

[21] V. Khrulkov and I. Oseledets, Art of singular vectors and universal adversarial perturbations, arXiv preprint arXiv:1709.03582, 2017.

[22] X. Huang, M. Kwiatkowska, S. Wang, M. Wu, Safety Verification of Deep Neural Networks, In 29th International Conference on Computer Aided Verification, pages 3-29, 2017.

[23] M.Wicker, X. Huang, and M. Kwiatkowska, Feature-Guided Black- Box Safety Testing of Deep Neural Networks, In 24th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, 2018.

[24] Dalvi, N., Domingos, P., Sanghai, S., & Verma, D. (2004, August). Adversarial classification. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 99-108). ACM.

[25] A. G. Ororbia II, C. L. Giles and D Kifer, Unifying Adversarial Training Algorithms with Flexible Deep Data Gradient Regularization, arXiv preprint arXiv:1601.07213, 2016.

[26] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).

[27] Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453.

[28] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[29] Carlini, N., & Wagner, D. (2014, August). ROP is Still Dangerous: Breaking Modern Defenses. In USENIX Security Symposium (pp. 385-399).

[30] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112.

[31] J. Lu, T. Issaranon, and D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," ICCV, 2017.

[32] A. Raghunathan, J. Steinhardt, P. Liang, Certified Defenses against Adversarial Examples, arXiv preprint arXiv:1801.09344. 2018.