ISSN: 2249-6661
Vol-43, No.-03 (III) November-December (2020)

A Survey on Adversarial Deep Learning Models with Multiple Adversaries

1st V.Srilakshmi
Computer Science and Engineering, GRIET
Hyderabad, Telangana, India

2nd K.Anuradha
Computer Science and Engineering, GRIET
Hyderabad, Telangana, India

3rdG.Karuna
Computer Science and Engineering, GRIET
Hyderabad, Telangana, India

4th Nagapuri.Janapriya
Computer Science and Engineering, GRIET
Hyderabad, Telangana, India

Abstract — The changing nature of warfare has seen a paradigm shift from the conventional to asymmetric, contactless warfare such as information and cyber warfare. Excessive dependence on information and communication technologies, cloud infrastructures, big data analytics, data-mining and automation in decision making poses grave threats to business and economy in adversarial environments. Adversarial machine learning is a fast growing area of research which studies the design of Machine Learning algorithms that are robust in adversarial environments. This paper presents a comprehensive survey of this emerging area and the various techniques of adversary modelling. We develop an adversarial learning algorithm for supervised classification in general and Convolutional Neural Networks (CNN) in particular. The algorithm's objective is to produce small changes to the data distribution defined over positive and negative class labels so that the resulting data distribution is misclassified by the CNN.

Keywords— Supervised learning, Data mining and knowledge discovery, Evolutionary learning, Adversarial learning, Deep learning.

I. INTRODUCTION

Deep learning is a branch of machine learning that enables computational models composed of multiple processing layers with high level of abstraction to learn from experience and perceive the world in terms of hierarchy of concepts. It uses backpropagation algorithm to discover intricate details in large datasets in order to compute the representation of data in each layer from the representation in the previous layer (lecun2015deep). Deep learning has been found to be remarkable in providing solutions to the problems which were not possible using conventional machine learning techniques [1]. With the evolution of deep neural network models and availability of high performance hardware to train complex models, deep learning made a remarkable progress in the traditional fields of image classification, speech recognition, language translation along with more advanced areas like analyzing potential of drug molecules (ma2015structure), reconstruction of brain circuits (helmstaedter2013retina), analyzing particle accelerator data (cio2012structure) (kaggle2012higgs), effects of mutations in DNA (xiong2015gene).

Deep learning network, with their unparalleled accuracy, have brought in major revolution in AI based services on the Internet, including cloud computing based AI services from commercial players like Google (google_cloud), Alibaba (alibaba_cloud) and corresponding platform propositions from Intel (intel_cloud) and Nvidia (nvidia_cloud). Extensive use of deep learning based applications can be seen in safety and security-critical environments, like, self driving cars, malware detection and drones and robotics. With recent advancements in face-recognition systems, ATMs and mobile phones are using biometric authentication as a security feature; Automatic Speech Recognition (ASR) models and Voice Controllable systems (VCS) made it possible to realize products like Apple Siri (ios), Amazon Alexa (alexa) and Microsoft Cortana (cortana). As deep neural networks have found their way from labs to real world, security and integrity of the applications pose great concern. Adversaries can craftily manipulate legitimate inputs, which may be imperceptible to human eye, but can force a trained model to produce incorrect outputs. Szegedy et al. (szegedy2013intriguing) first discovered that well-performing deep neural networks are susceptible to adversarial attacks. Speculative explanations suggested it was due to extreme nonlinearity of deep neural networks, combined with insufficient model averaging and insufficient regularization of the purely supervised learning problem [2].

Worldwide basis humans are affected by many types of life threatening diseases, among of this, the heart disease has received more attention. Heart disease basically causes the injury of the heart and the blood vessels. Therefore, the heart syndrome is a most important reason for mortality and death for people in most of the countries all over the world. According to one survey in 2008 approximately 17.3 million people died from heart diseases (7.3 million deaths were due to coronary heart disease and 6.2 million were due to stroke), that corresponds to the 30% of all global deaths had occurred. Recently, many types of research in the medical industry have been able to identify risk factors of heart diseases, however, more contribution is necessary to use this knowledge to reduce the risk of deaths. The

ISSN: 2249-6661
Vol-43, No.-03 (III) November-December (2020)

significant mortality rate caused by the heart disease throughout the globe need for the development of new heart disease prediction method. These systems allow patients to calculate the heart disease risks. There are many factors of heart disease that affecting the structure or function of the heart [3].

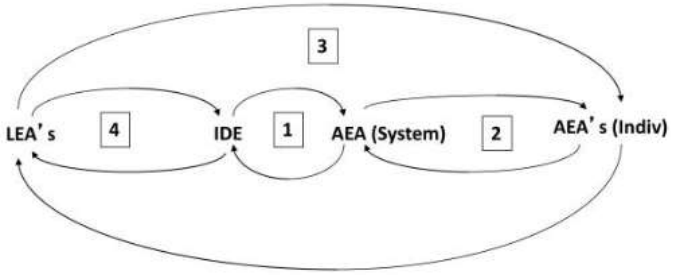

Figure.1: Multiple Accidental Adversaries

In this paper, we study the adversarial robustness of neural networks through the lens of robust optimization. We use a natural saddle point (min-max) formulation to capture the notion of security against adversarial attacks in a principled manner. This formulation allows us to be precise about the type of security guarantee we would like to achieve, i.e., the broad class of attacks we want to be resistant to (in contrast to defending only against specific known attacks). The formulation also enables us to cast both attacks and defenses into a common theoretical framework, naturally encapsulating most prior work on adversarial examples. In particular, adversarial training directly corresponds to optimizing this saddle point problem. Similarly, prior methods for attacking neural networks correspond to specific algorithms for solving the underlying constrained optimization problem [1].

A. THE ADVERSARIAL CAPABILITIES

The term adversarial capabilities refer to the amount of information available to an adversary about the system, which also indicates the attack vector he may use on the threat surface. For illustration, again consider the case of an automated vehicle system as shown in Figure 3 with the attack surface being the testing time (i.e., an Evasion Attack). An internal adversary is one who have access to the model architecture and can use it to distinguish between different images and traffic signs, whereas a weaker adversary is one who have access only to the dump of images fed to the model during testing time. Though both the adversaries are working on the same attack surface, the former adversary is assumed to have much more information and is thus strictly "stronger". We explore the range of adversarial capabilities in machine learning systems as they relate to testing and training phases.

Adversarial capabilities refer to the possible impact or influence that an adversary can have by attacking the ML model. Attacks of the adversary based on the capabilities can be classified according to the following three dimensions:

- Influence
- Specificity
- Impact

Classification based on influence of adversary is based on the attempt to change the dataset or the algorithms of the target during the course of the attack. Such attacks can be further classified according to the influence as causative or exploratory.

Causative: Causative attacks alter the training process through influence over the training data. This requires the adversary to modify or influence both training and testing data.

Exploratory: Exploratory attacks do not alter the training process but use other techniques, such as probing, to discover information about training data. The adversary cannot modify or manipulate the training data and can only craft new instances based on the underlying data distribution.

B.GENERATING ADVERSARIAL EXAMPLES

This section will highlight some methods of generating adversarial examples. These attacks can be categorized into targeted or untargeted and by choice of distance measurement. Targeted methods generate adversarial examples that are classified with a chosen particular class, whereas untargeted methods generate adversarial examples that are classified with any other class that is not the true one.

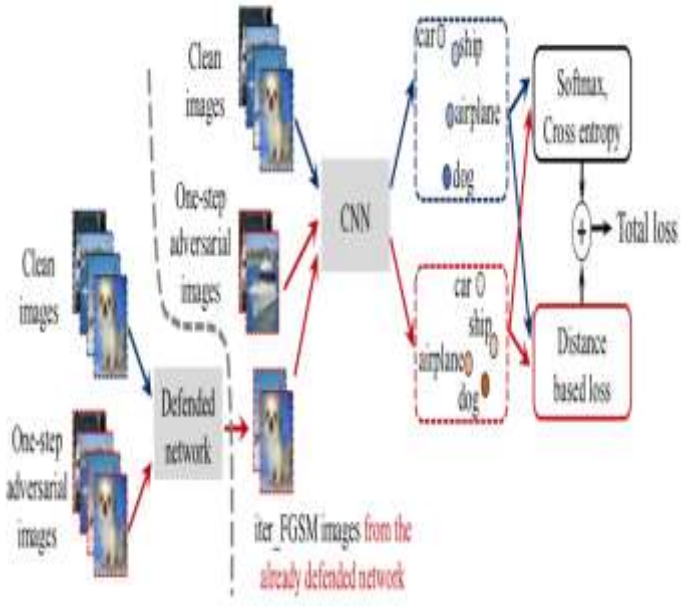

Figure.2: Multiple adversarial training example

In this survey paper adversarial algorithm, the search and optimization algorithm is either a genetic algorithm or a simulated annealing algorithm. The adversarial data samples are generated by the selection, crossover, mutation search operators in the genetic algorithm and the annealing search operator in the simulated annealing algorithm. By using probabilistic hill climbing algorithms over Markov chains in multivariate models, the current search operators can be extended to define explicit probabilistic distributions performing a complex neighbourhood search for the candidate solutions.

C.ADVERSARIAL KNOWLEDGE

Knowledge of the underlying ML model plays a crucial role in determining the success of the attacks by providing the adversary an opportunity to make informed decisions as

ISSN: 2249-6661
Vol-43, No.-03 (III) November-December (2020)

shown in Fig. 3. The knowledge of the ML system can be classified into:

- Data acquisition
- Data
- Feature selection
- Algorithm and parameters
- Training and output

Complete/perfect knowledge: An adversary is said to have perfect knowledge if he has access to the knowledge of data acquisition, data, feature selection, ML algorithms and tuned parameters of the model. The attacker may or may not have access to the training data which can be easily acquired by using other knowledge. This is usually the case when the ML model is open source and everyone has access to it.

Limited Knowledge: In this case, the adversary only knows a part of the model. he does not have access to the training data and may have very limited information about the model architecture, parameters, and has access to only a small subset of the total knowledge available.

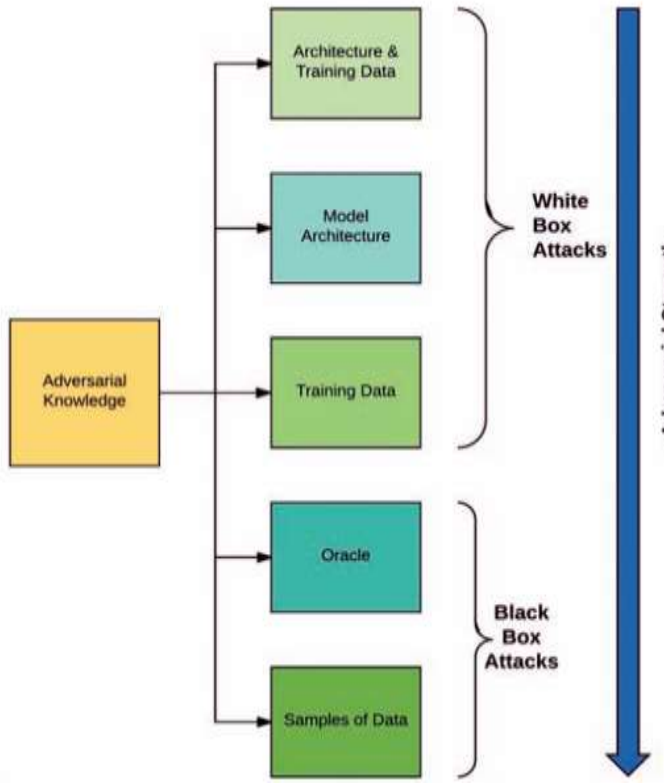

Figure.3: Adversary's knowledge

The adversary may have either complete or perfect knowledge of the ML system or only a partial knowledge of the system. Adversary attacks can be classified into black box attacks and white box attacks based on the knowledge about the model an adversary has.

For the adversary to evolve from black box to white box, he iteratively goes through a process of learning using inference mechanisms to gain more knowledge of the model.

II.STOCHASTIC GAME ALGORITHM

Stochastic games defined on a strategy space have been used to generate adversarial examples. The strategy space is defined in terms of two or more adversaries' actions and corresponding payoff functions. Each adversary can engage one or more learners in a game and vice versa. From the learner's standpoint, adjusting parameters is computationally less expensive than building a new model that is robust to adversarial manipulation. From the adversary's standpoint, the attack scenarios can be characterized by the stochastic optimization parameters estimated in the game. A game ends in an equilibrium with payoffs to each player based on their objectives and actions. The learner has no incentive to play a game that leads to too many false positives with too little increase in true positives. The adversary has no incentive to play a game that increases the utility of false negatives not detected by the learning algorithm. At equilibrium, the adversary is able to find testing data that is significantly different from the training data whereas the learner is able to update its model for new threats from adversarial data.

In this survey paper adversarial algorithm proposes a game between two players - a data miner or learner and an intelligent adversary or adversary. The interactions between the learner and adversary are modelled as a two-player sequential Stackelberg zero-sum game. In our game, the adversary is the leader and the learner is the follower. The learner retrains the model after the adversary's attack. The payoff function for each player is specified in terms of objective functions simulating the adversary's attack process and learner's learning processes. The attack processes specify the adversary's constraints and optimal attack policy. The learning processes specify the learner's gain and and adversary's gain under the optimal policy. The optimal attack policy is formulated in terms of stochastic optimization operators and evolutionary computing algorithms.

A.GENETIC ALGORITHM

In this section we validate the adversarial data in a sequential game that is constructed by the mutation, crossover and selection genetic operators defined on the images. The testing performance for mutation, crossover, selection operators and population size on the data manipulated by final $\alpha*$ .. The range for random pixel values is between the lower bound of RGB pixel value(-255) and the upper bound of RGB pixel value(+255). The size of the images is 32*32*3 as required by the CNN model. For the input images manipulated by adding $\alpha*$, the pixels with values greater than 255 are set to 255 and pixels with values less than 0 are set to 0.

a. Mutation operation

A mask of randomly generated integers between a lower bound -δ(set to -50 by default) and upper bound +δ(set to +50 by default) for the step is added to the current image in the mutation operation.

b.Crossover operation

For a three-dimensional 32*32*3 RGB image, the height and width indices are randomly selected. The starting index for height is selected between pixels 1 and 16(half of the largest height). The ending index for height is selected between lower bound η(set to +2 by default) and η(set to +10 by default) from the corresponding starting index of

ISSN: 2249-6661
Vol-43, No.-03 (III) November-December (2020)

height and upper bound 32. A similar random indexing scheme selects the starting width and ending width of the image. The slice of the starting and ending index of height/width over all the pixels in depth is then swapped between the two images in the crossover operation.

c.Selection operation

The selection operation is an extension of random sampling without replacement. The parents for the next generation are randomly chosen from the current generation parents.A ζ(set t to 0.5 by default) percentage of the current generation parents are selected to be the offspring for the next generation. The remaining candidates in the current generation of parents are preserved as parents for the next generation. The probability of selecting an offspring is proportional to the fitness values of the current parents. The selected offspring are then changed by crossover and mutation to get the parents for the next generation. Across every generation of the genetic algorithm, the size of the entire population(consisting of current offspring and parents) is fixed to the initial size of the parents.

III.EXPERIMENT

In this section we discuss the experimental validation and stochastic parameters of the adversarial learning algorithm. During the game, an adversary finds adversarial data manipulations using either a genetic algorithm or a simulated annealing algorithm as the search algorithm. For a twoplayer game, various parameter settings produce adversarial manipulation $\alpha*$ on the images such that the positive class examples are misclassified as negative class examples by the CNN aka learner.For example, the CNN misclassifies the handwritten digit 7 which had been positively labelled before adversarial manipulation as the negatively labelled handwritten digit 9 after adversarial manipulation. The CNN is then secured against attacks in a stochastic game with multiple adversaries by defending against adversarial manipulations in many sequential games with two adversaries. The performance of the proposed secure CNN model is also compared with the performance of a CNN model augumented by the data produced from various Generative Adversarial Network(GANs).In both the multiplayer game and the two-player game, we observe that the manipulated learner performance is lower than the original learner performance. Also, the secure learner performance is higher than the manipulated learner performance.

IV.CONCLUSION

In this survey paper we formulated a maxmin problem for adversarial learning with both two-player sequential games and multiplayer stochastic games over deep learning networks. We demonstrate the correctness and performance of proposed adversarial algorithm. The algorithm converges onto adversarial manipulations affecting testing performance in deep learning networks. This allows us to propose a secure learner that is immune to the adversarial attacks on deep learning. We have shown that our model is significantly more robust than traditional CNN and GAN under adversarial attacks. By changing the game formulation, we can experiment with adversarial payoff functions over randomized strategy spaces. The attack scenarios over such strategy spaces would determine multiplayer games over mixed strategies.

References

[1] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, “Can machine learning be secure?” in Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ser. ASIACCS ’06. New York, NY, USA: ACM, 2006, pp. 16–25.
[2] B. Biggio, G. Fumera, I. Pillai, and F. Roli, “A survey and experimental evaluation of image spam filtering techniques,” Pattern Recogn. Lett., vol. 32, no. 10, pp. 1436–1446, Jul. 2011.
[3] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, “Adversarial machine learning,” in Proceedings of the 4th ACM workshop on Security and artificial intelligence. ACM, 2011, pp. 43–58.
[4] B. Biggio, G. Fumera, and F. Roli, “Security evaluation of pattern classifiers under attack,” IEEE transactions knowledge and data engineering, vol. 26, no. 4, pp. 984– 996, 2014.
[5] L. Deng, “Three classes of deep learning architectures and their applications: a tutorial survey,” APSIPA transactions on signal and information processing, 2012.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems (NIPS), 2014, pp. 2672– 2680.
[7] B. Biggio, G. Fumera, and F. Roli, “Multiple classifier systems for robust classifier design in adversarial environments,” International Journal of Machine Learning and Cybernetics, vol. 1, no. 1-4, pp. 27–41, 2010.
[8] A. Kołcz and C. H. Teo, “Feature weighting for improved classifier robustness,” in CEAS09: sixth conference on email and anti-spam, 2009.
[9] B. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” 29th International Conference on Machine Learning (ICML), pp. 1807–1814, Jun. 2012.
[10] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, “Support vector machines under adversarial label contamination,” Neurocomputing, vol. 160, pp. 53– 62, 2015.
[11] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, “Practical black-box attacks against deep learning systems using adversarial examples,” 2017 ACM Asia Conference on Computer and Communications Security, 2016.
[12] S. Gu and L. Rigazio, “Towards deep neural network architectures robust to adversarial examples,” CoRR, vol. abs/1412.5068, 2014.
[13] A. Globerson and S. Roweis, “Nightmare at test time: robust learning by feature deletion,” in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 353–360.
[14] B. I. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, “Learning in a large function space: Privacy-preserving mechanisms for svm learning,” Journal of Privacy and Confidentiality, vol. Vol.4 : Iss.1, Article 4., 2009.
[15] A. Barth, B. I. Rubinstein, M. Sundararajan, J. C. Mitchell, D. Song, and P. L. Bartlett, “A learning-based approach to reactive security,” International Conference on Financial Cryptography and Data Security, pp. 192–206, 2010.