Applying Visual Adversarial Learning on VisualNet dataset for Robust Phishing Detection

DYLAN CHOU, Carnegie Mellon University
AMIR RAHMATI, Stony Brook University | Ethos Lab Lead | Assistant Professor

The internet and computer networks brought about a platform where those with malicious intent can steal victim's information through phishing. In turn, phishing detection techniques have been used to mitigate the number of successful phishing attacks. One means of detection involves the development of a machine learning model to predict when a suspicious webpage or email is indeed a phishing scheme and warn the victim. Throughout the literature, there has been a plethora of visual similarity-based, fuzzy data mining, and text mining approaches to phishing detection using machine learning, but a lack in adversarial approaches to bring out the weaknesses in recent phishing datasets. In this paper, a taxonomy of phishing webpage detection methods, including those that are adversarial, and input from adversarial sample generation will provide further robustness to the models fit to the datasets analyzed in the paper.

CCS Concepts: • **Computer-Communication Networks**; • **Security and privacy** → Network security;

Additional Key Words and Phrases: Adversarial Machine Learning, Phishing Detection

ACM Reference Format:
Dylan Chou and Amir Rahmati. 2020. Applying Visual Adversarial Learning on VisualNet dataset for Robust Phishing Detection. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2020), 6 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

1.1 Background

With the advent of the internet, phishing has become a prevalent scheme to steal people's information through social engineering methods such as spoofed emails or webpages. The ramifications of phishing involve financial loss or identity theft among other problems [33]. The fourth quarter report by anti-phishing working group (APWG) in 2019 [7] reported that the number of phishing websites throughout the year started off with around 45,000 sites, then increased to about 90,000 sites in July 2019, then decreased back to a little less than 45,000 sites by December 2019. In just October 2019 alone, AWPG detected 76,804 unique phishing website URLs. Although phishing webpages and emails can, for the most part, be detectable via human perception, there have been more sophisticated methods with cross-site scripting that injects malicious scripts written in JavaScript into a webpage [27]. In turn, there have been detection techniques implemented to alert users of threats from suspicious emails or webpages.

Around two decades ago, phishing had already become a problem for online users and anti-phishing plug-ins were made to combat phishing detection. Typically, when users entered in

Authors' addresses: Dylan Chou, dvchou@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15289; Amir Rahmati, Stony Brook University | Ethos Lab Lead | Assistant Professor, Stony Brook, New York, 11794, amir@cs.stonybrook.edu.

2476-1249/2020/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

sensitive, protected information, the site that they are on is checked to see if it's among those in a pre-populated whitelist. There was a heavy reliance on blacklists that helped users avoid phishing websites, but problems came from a large percentage of the user population contacting the blacklist server for information when using a browser plug-in while on the Internet. A multitude of common attacks were presented such as distributed attacks directing its victims to phishing websites, denial of service, and redirection attacks [14].

Early anti-phishing strategies used text and image-block features in plug-in modules. Liu et al [23] made an anti-phishing feature for Microsoft Outlook that searches for suspicious URLs in messages. Using style, layout and block-level similarity metrics, these anti-phishing methods are intended to pick up visual similarities between protected and suspect pages. Medvet et al [25] integrates a visual similarity detection system into the open-source tool AntiPhish and considers text pieces, images, and visual layout on a webpage. This is based on how humans perceive phishing sites on a look-and-feel basis. Initially, there was motivation to use visual similarity of webpages based on visual features such as page layout due to its resemblance to human perception.

Along with these initial strategies for anti-phishing, other researchers examined different similarity metrics when comparing visually similar webpages. Fu et al [15] used earth mover's distance (EMD) to measure webpage visual similarity. EMD compares two signatures, which consist of features and weights. For instance, consider a transportation where there are m producers. Each producer has a weight that represents the amount of product they have. The producer set would look like: $P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$. With producers, there are n consumers, so the consumer set is $C = \{(c_1, w_{c_1}), ..., (c_m, w_{c_m})\}$. Given that pairwise distances between consumers and producers are provided - most likely using euclidean distance - the distance matrix is $D = [d_{ij}]$, where $1 \leq i \leq m, 1 \leq j \leq n$. A flow matrix is needed to represent product moving from a producer to a consumer: $F = [f_{ij}]$, where $1 \leq i \leq m, 1 \leq j \leq n$. The cost of transporting product would be $\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} * d_{ij}$ and the calculation for the flow matrix turns into a linear programming problem that results in the earth's movers distance being the ratio of the transportation cost to the total product being transported around. The primary drawback was that despite the interest in visual similarity approaches for phishing detection, an underlying assumption is that phishing webpages are visually similar to their target sites. Around the second half of the 2000's as well, there were applications of traditional machine learning algorithms on phishing detection problems in research. These approaches ranged from traditional statistical methods such as logistic regression and bayesian additive regression trees to other classifiers such as support vector machines, random forests, decision trees, and neural networks [1]. Visual similarity-based methods spawned research currently in image-based phishing detection for webpages. Bozkir and Aydos [11] analyze visual features on phishing webpages by employing histogram of oriented gradients to find visual representations of target brand logos for attackers. Logosense also developed a dataset of legitimate and phishing logos containing bounding box annotation.

Because of the uncertainty related to rules on how to discretize phishing detection features into different classes, fuzzy data mining was researched. Aburrous et al [2] researched phishing detection in the context of E-banking webpages using fuzzy data mining techniques because many features cannot be split into definite classes such as the length of the URL ranging from "short" to "long". The degree of belongingness of values to any class, known as degree of membership. Website risk categories are also ambiguous as very legitimate, legitimate, suspicious, phishy, and very phishy. Four years after Aburrous's work, Barraclough et al [8] applied a neuron-fuzzy to phishing detection for E-banking transactions that sought to reduce the amount of false positives by introducing a parameter tuning framework and comprehensive features not seen in past research. More recently, rough set theory has been integrated into fuzzy systems to improve feature

selection. Montazer and ArabYarmohammadi [26] used rough set theory to remove redundant data from main features in the data and extract a subgroup called the "reduct" using the quick reduct algorithm to add features to an empty set that results in the maximum information gain to the phishing feature set. Zabihimayvan and Doran combined [34] fuzzy systems with rough set theory to improve feature selection for phishing detection as rough set theory can split up data via decision boundaries and a relation R.

Alongside inspection into using fuzzy data mining to handle uncertainty in class divisions with phishing detection, there grew a focus on applying text mining to a webpage's URL and words on the site. L'Huillier et al [20] took a phishing corpus and applied singular value decomposition to reduce the dimensionality of the data and use latent dirichlet allocation - a model where different topics in documents are inferred from probability distributions over the training data - to extract different phishing topics. Bhattacharjee et al [10] finds weighted text-based features from taking in the top K relevant features and iteratively updates weights of features to filter out the most relevant features and identify relevant unannotated URLs and the focus in the data was on the text in webpages. To generalize phishing models, Tayal and Ravi [31] used the particle swarm algorithm on class association rules to reduce the number of rules and allow the classifier to be generalized. Text mining was done to extract URL features and discretize some data into a binary format such as having or not having the @ symbol or being on a blacklist. In addition to studying each sample independently through text mining strings, authors studied interconnections between observations and how the datum are associated. DeBarr et al [12] analyzed the links between URL substrings through spectral clustering. Three years after, Thabtah and Abdelhamid [32] compared different feature set assessments to reduce a phishing website feature set and identify new clusters among features; namely, interconnections between similar features were found.

Other research went towards protecting the victims in phishing attacks. Gustafson and Li [17] took a research direction that disrupted phishing operations by injecting fake credential, or honey tokens, on phishing websites. They study how they may influence users to successfully submit these fake credentials on their network of servers in an architecture called *Humboldt 2.0*, and how users can combat malicious attacks in the system. Similarly, Alsharnouby et al [5] conducted a user study on whether users' were able to protect themselves from phishing attacks when browser security indicators and awareness of phishing were improved. General defense methods in terms of practicing physical security of information, penetration testing and user security awareness were studied by Saleem and Hammoudeh [29].

1.2 Past Work in Adversarial Learning

Adversarial learning is used to exploit vulnerabilities in models by passing in deceptive input where, in the case of phishing detection, there's an adversary and a classifier. Adversaries learn more about the classifier based on past knowledge, observation, and experimentation. Adversaries may send membership queries to the classifier to confirm whether, for instance, a URL is malicious or not [24]. Adversarial learning is a significant method to develop defensive mechanisms against deceptive input to phishing detection models and entail more robust phishing models.

In the second half of the 2000's, there was a general challenge for machine learning in terms of finding bounds on adversarial learning problems. Barreno et al [9] discuss open challenges with the influence that adversaries have toward learning. After the learner chooses a strategy, the adversary transforms the data and the learner updates its predictions to minimize its cumulative loss. In terms of phishing detection, features to be considered include the language used on the site and the url. Adversarial information is the adversary's knowledge of the learning environment: the benign data generation process, retraining process, learning algorithm, and learner's features. Because applications of machine learning was at a confluence of statistical learning and handling malicious adversaries by 2008, directions for research were to detect malicious, or adversarial,

instances in contaminated data - phishing attacks - via outlier detection, design sensitive learners, and orthogonal learners, or a set of multiple learners. In turn, applying statistical methods to adversarial environments could expand traditional statistical thinking of data being sampled from a common distribution when there are two separate distributions: one that is known and another that is adversarial. Along with applying statistics to security in adversarial environments, machine learning also faced the challenge of how much an adversary knows about the data. Kearns and Li [19] had expanded on probably approximately correct learning by giving an adversary control over part of the training data. Also, at the time, Barreno discussed that adversarial control should be realistic under full adversarial information - the adversary knowing close "surrogate" data but not the actual phishing training data - but there was still room to conduct research on adversarial capabilities.

Adversarial data mining was then introduced as a solution to prevent classifiers from decreasing in performance after the adversary learns how to defeat a phishing classifier. Adversarial games - games between a classifier and a malicious agent acting as an adversary - were studied by L'huillier et al [21] where an online learning algorithm - Weighted Margin Support Vector Machines with a game theoretic prior knowledge function - is implemented. Looking back at game-theoretic frameworks of adversarial learning, Alabdulmohsin et al [3] investigates reverse-engineering fixed classifiers by probing classifiers for outputs $\widehat{y}_q$ from given input queries x_q and randomly sampling data from a Gaussian with mean of $\frac{x^{(+)}+x^{(-)}}{2}$. The defender can construct a distribution multiple classifiers so any classifier chosen at random may provide reliable predictions and prevent adversaries from reverse-engineering the classifier. Gutierrez et al [18] expanded off of Alabdulmohsin's work on adversarial data mining by using semi-automated feature generation to synthetically generate phishing data, construct more robust classifiers and catch techniques that attackers may use.

In the past few years, generative adversarial networks have been of interest by researchers since its development in 2014 by Ian Goodfellow. Generative Adversarial Networks (GANs) takes an input feature vector and adds noise based on a normal distribution, which consists of a variable number of layers. This is the *generator* network. The *discriminator* work evaluates the data from the generator. The generator and discriminator work hand-in-hand in competing with each other; namely, the goal for the generator is to increase the error rate in the discriminator's predictions whereas the discriminator's goal is to achieve the highest accuracy [16]. A negative side-effect to the synthetic production of data by GANs is that sample generation from the generator results in mode collapse, or that the data becomes too similar. Lin et al [22] examines generative adversarial networks through the lens of hypothesis testing where the discriminator performs a binary hypothesis test on samples from distribution P or Q (P being the true data distribution while Q is the generated distribution). Their methodology comes from *packing* which involves the discriminator predicting different classes based on multiple samples from the same class, such as phishing or legitimate. Mode collapse thus can be more easily detected as *lack of diversity is more noticeable in a set of samples than a single sample* [22]. Figueroa et al [13] examined the game-theoretic framework between a classifier and an adversary, but steeped more in Game Theory and the architecture of a GAN viewed as a signaling game between a sender and receiver.

Because of the long-standing challenge of imbalanced phishing datasets, some researchers have combined oversampling with GANs [6]. Shirazi et al [30] generated adversarial samples using direct feature manipulation by producing all feature combinations then permuting all possible feature values per feature. With new software packages for deep learning, Robic-Butez and Win [28] implemented a GAN in keras on 3 different datasets. Most recently, researchers looked to make deceptive phishing samples to reduce accuracy in classifiers as much as possible using GANs. Aleroud and Karabatis [4] intentionally made deceptive examples to fool blackbox phishing

detectors (i.e. Random Forest, Decision Tree, Linear Regression, Neural Network, Support Vector Machine) using feature perturbation.

Ultimately, the related works in adversarial phishing detection had a significant emphasis on malicious URLs aside from image data on the page layout of a page. URLs on phishing webpages are emphasized in research over other features, but there will be adversarial machine learning done primarily on visual features.

1.3 Contributions in Paper

Our work contributes to past research in the following ways:

- We generate adversarial samples via *method*
- *hope* Our samples are test on classifiers fit to recent datasets and results display the efficacy of the samples in reducing the accuracy of the classifiers.
- From our adversarial sample generation, we can make conclusions about the data that makes the classifiers vulnerable and offer a means of improving the robustness of the models.

1.4 Resources

Further resources aside from those presented currently in Zeblok includes more GPUs as the time it takes to run the 50000 iterations of ML training is fairly long.

References

[1] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. 2007. A Comparison of Machine Learning Techniques for Phishing Detection. In *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit (eCrime '07)*. Association for Computing Machinery, New York, NY, USA, 60–69. https://doi.org/10.1145/1299015.1299021

[2] Maher Aburrous, M Alamgir Hossain, Keshav Dahal, and Fadi Thabtah. 2010. Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert systems with applications* 37, 12 (2010), 7913–7921.

[3] Ibrahim M Alabdulmohsin, Xin Gao, and Xiangliang Zhang. 2014. Adding robustness to support vector machines against adversarial reverse engineering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 231–240.

[4] Ahmed AlEroud and George Karabatis. 2020. Bypassing Detection of URL-based Phishing Attacks Using Generative Adversarial Deep Neural Networks. In *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*. 53–60.

[5] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. 2015. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies* 82 (2015), 69–82.

[6] Ankesh Anand, Kshitij Gorde, Joel Ruben Antony Moniz, Noseong Park, Tanmoy Chakraborty, and Bei-Tseng Chu. 2018. Phishing URL detection with oversampling based on text generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 1168–1177.

[7] APWG. [n.d.]. Phishing Activity Trends Report 4th Quarter 2019. https://docs.apwg.org/reports/apwg_trends_report_q4_2019.pdf.

[8] Phoebe A Barraclough, MA Hossain, G Sexton, and Nauman Aslam. 2014. Intelligent phishing detection parameter framework for E-banking transactions based on Neuro-fuzzy. In *2014 Science and Information Conference*. IEEE, 545–555.

[9] Marco Barreno, Peter L Bartlett, Fuching Jack Chi, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, Udam Saini, and J Doug Tygar. 2008. Open problems in the security of learning. In *Proceedings of the 1st ACM workshop on Workshop on AISec*. 19–26.

[10] Sreyasee Das Bhattacharjee, Ashit Talukder, Ehab Al-Shaer, and Pratik Doshi. 2017. Prioritized active learning for malicious url detection using weighted text-based features. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 107–112.

[11] Ahmet Selman Bozkir and Murat Aydos. 2020. LogoSENSE: A Companion HOG based Logo Detection Scheme for Phishing Web Page and E-mail Brand Recognition. *Computers & Security* (2020), 101855.

[12] Dave DeBarr, Venkatesh Ramanathan, and Harry Wechsler. 2013. Phishing detection using traffic behavior, spectral clustering, and random forests. In *2013 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 67–72.

[13] Nicolas Figueroa, Gastón L'Huillier, and Richard Weber. 2017. Adversarial classification using signaling games with an application to phishing detection. *Data mining and knowledge discovery* 31, 1 (2017), 92–133.

[14] Dinei Florêncio and Cormac Herley. 2006. Analysis and improvement of anti-phishing schemes. In *IFIP International Information Security Conference*. Springer, 148–157.

[15] Anthony Y Fu, Liu Wenyin, and Xiaotie Deng. 2006. Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). *IEEE transactions on dependable and secure computing* 3, 4 (2006), 301–311.
[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
[17] Jason Gustafson and Jun Li. 2013. Leveraging the crowds to disrupt phishing. In *2013 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 82–90.
[18] Christopher N Gutierrez, Taegyu Kim, Raffaele Della Corte, Jeffrey Avery, Dan Goldwasser, Marcello Cinque, and Saurabh Bagchi. 2018. Learning from the ones that got away: Detecting new forms of phishing attacks. *IEEE Transactions on Dependable and Secure Computing* 15, 6 (2018), 988–1001.
[19] Michael Kearns and Ming Li. 1993. Learning in the presence of malicious errors. *SIAM J. Comput.* 22, 4 (1993), 807–837.
[20] Gaston L'Huillier, Alejandro Hevia, Richard Weber, and Sebastian Rios. 2010. Latent semantic analysis and keyword extraction for phishing classification. In *2010 IEEE international conference on intelligence and security informatics*. IEEE, 129–131.
[21] Gaston L'Huillier, Richard Weber, and Nicolas Figueroa. 2009. Online phishing classification using adversarial data mining and signaling games. In *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics*. 33–42.
[22] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. 2018. PacGAN: The Power of Two Samples in Generative Adversarial Networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 1505–1514.
[23] Wenyin Liu, Xiaotie Deng, Guanglin Huang, and Anthony Y Fu. 2006. An antiphishing strategy based on visual similarity assessment. *IEEE Internet Computing* 10, 2 (2006), 58–65.
[24] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 641–647.
[25] Eric Medvet, Engin Kirda, and Christopher Kruegel. 2008. Visual-similarity-based phishing detection. In *Proceedings of the 4th international conference on Security and privacy in communication networks*. 1–6.
[26] Gholam Ali Montazer and Sara ArabYarmohammadi. 2015. Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system. *Applied Soft Computing* 35 (2015), 482–492.
[27] Angelo Eduardo Nunan, Eduardo Souto, Eulanda M Dos Santos, and Eduardo Feitosa. 2012. Automatic classification of cross-site scripting in web pages using document-based and URL-based features. In *2012 IEEE symposium on computers and communications (ISCC)*. IEEE, 000702–000707.
[28] Pierrick Robic-Butez and Thu Yein Win. 2019. Detection of Phishing websites using Generative Adversarial Network. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 3216–3221.
[29] Jibran Saleem and Mohammad Hammoudeh. 2018. Defense methods against social engineering attacks. In *Computer and network security essentials*. Springer, 603–618.
[30] Hossein Shirazi, Bruhadeshwar Bezawada, Indrakshi Ray, and Charles Anderson. 2019. Adversarial sampling attacks against phishing detection. In *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 83–101.
[31] Kshitij Tayal and Vadlamani Ravi. 2016. Particle swarm optimization trained class association rule mining: Application to phishing detection. In *Proceedings of the International Conference on Informatics and Analytics*. 1–8.
[32] Fadi Thabtah and Neda Abdelhamid. 2016. Deriving correlated sets of website features for phishing detection: a computational intelligence approach. *Journal of Information & Knowledge Management* 15, 04 (2016), 1650042.
[33] Liu Wenyin, Guanglin Huang, Liu Xiaoyue, Xiaotie Deng, and Zhang Min. 2005. Phishing Web page detection. In *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. IEEE, 560–564.
[34] Mahdieh Zabihimayvan and Derek Doran. 2019. Fuzzy rough set feature selection to enhance phishing attack detection. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1–6.