Machine Learning for Offensive Cyber Operations

Åvald Åslaugson Sommervoll[1 [0000−0001−5232−5630]],
Audun Jøsang[1 [0000−0001−6337−2264]]

University of Oslo, Problemveien 7, 0315 Oslo
aavalds@ifi.uio.no

Abstract. This paper gives a brief survey of existing and proposed applications of machine learning for offensive cyber operations, with particular emphasis on algorithmic cryptanalysis and penetration testing. For cryptanalysis at the algorithmic level, we cover attacks on historic ciphers as well as attacks on modern ciphers. For penetration testing, we cover works that have focused on defining structured attack approaches as well as some novel attacks where the potential merits need additional investigation.

Keywords: machine learning, offensive cyber operations, cryptanalysis, penetration testing, survey

1 Introduction

The arms race between cryptographers and cryptanalysts is an ancient one, with the earliest record of cryptanalysis dating back to the 9th century [14]. The attack described was frequency analysis effectively breaking the monoalphabetic substitution cipher; this implicated that for secure communication, the cryptographers would have to do something more advanced. A thousand years later the Germans used Enigma encryption, an encryption they thought to be unbreakable for communication during WWII. However, the huge joint effort of pre-WWII analysis of Polish mathematicians, paired with efforts from English and American scientists to develop cryptanalytical tools and methods, would show that it was indeed breakable [14]. Since WWII, Enigma encryption has been broken many times over because of its historical significance and as an effort to further offensive cyber operations[1] [10,17,13,12]. Some of these utilize machine learning techniques to speed up the attack [3,16]. Currently, in the arms race between cryptanalysts and cryptographers, it appears that cryptography has won, with standardized algorithms that are internationally recognized as secure. The arms race is far from over as new creative decryption attacks see light of day. However, since the algorithms themselves are deemed secure, modern attacks typically target the implementation, moving the hotspot of the current war from cryptology

[1] Note that we study offensive cyber operations: Testing and checking the integrity of existing cybersecurity defenses, not offensive cybersecurity: proactively predicting and removing threats in the system [1].

to cybersecurity[2]. There is a need for offensive cyber operations research to investigate the potential weaknesses and strengths of existing systems.

The rest of the paper is organized as follows: Section 2 involves a brief overview of machine learning and its impacts on cryptography. Section 3 covers some of the recent work on penetration testing using machine learning, in particular in terms of SQL injections. Finally, section 5 gives a brief concluding summary of this survey.

2 Cryptanalysis

Machine learning techniques are not easy to apply to the field of cryptoanalysis. This is because machine learning in general works by gradually inching closer to a good solution through *learning*, while modern crypto has many techniques that hide how close a cryptanalyst is to the solution; in other words obscuring learning. This obvious hurdle of machine learning in cryptoanalysis, may explain the rather short list of promising attempts using ML techniques. However, there has been documented some successes on classical systems such as Enigma [3,16]. Bagnall et al. cracked a two-rotor system of Enigma[3] which was based on using a genetic algorithm [3], but failing on 3 and 4 rotors. Sommervoll and Nilsen used the genetic algorithm to break the final step of Enigma decryption, finding all ten plugs of Enigma's plugboard faster than previous techniques [16]. More modern attacks are based on neuro-cryptanalysis first described by Dourlens in 1996 [6]. Since then, it has seen some limited success. Alani, in his neuro-cryptanalysis, attacks another classic but more modern cryptosystem DES and Triple-DES, with some success [2]. He does this by simulating the decryption under an unknown key using a neural network. In that, the input to his neural network are ciphertexts, and the output targets are the plaintexts. After training, he does not obtain the secret key, but ideally, a decryption machine that acts as the decryption algorithm with the key. He achieves an average bit accuracy of 91.7% for DES and 88.6% for Triple-DES. Also, in the field of neuro-cryptanalysis, a recent publication by Sommervoll in 2021 investigates the prospects of simulating an encryption algorithm as a neural network in what he refers to as the phantom gradient attack [15]. This attack does not draw from machine learning directly but attempts to use the same functions that train neural networks to train their way to the key. The trained network itself will, in this case, be uninteresting for prediction, but the trained weights will give the keys. Another example of neural-cryptanalysis is Aron Gohr's attack on Speck32/64 with deep learning [11]. Gohr did not use machine learning to recover the key directly, but used neural networks to distinguish between round reduced instances of Speck32/64 and random noise. He did this with great success, which is surprising from a cryptographic viewpoint. A recent follow-up paper by Benamira et al. investigates Gohr's findings [4]. They confirm his results, claim that his attack, while

[2] Side-channel attacks and espionage also have a rich history in humanity, though this history is so diverse that we do not cover it in this short review paper.

[3] Enigma encryption used had 3 to 4 rotors and a plugboard of 10 plugs during WWII.

impressive, is not really a novel cryptanalytical attack but is an optimization of the extraction of the low-data constrains.

3 Penetration testing

The field of penetration testing is considerably easier to unite with machine learning than algorithmic cryptanalysis. This is in large because machine learning agents can have the benefit of learning from humans, and the problems are not specifically designed to be difficult. Nonetheless, there is limited work done on automating the process of penetration testing with machine learning. Erdődi and Zennaro formalize part of this problem in the context of web hacking and reinforcement learning in [8]. The approach is called *Agent Web Model* that considers web hacking as a capture-the-flag (CTF) challenge. This model has seven layers of complexity, where layer 1 is the least complex, the agent is able to find links in objects, and layer 7 is the most complex; the agent is able to add files through a vulnerable object or create new database objects. In 2020 the authors demonstrated the potential of this approach by showing that reinforcement learning(RL) agents could solve CTF problems [18]. The authors showed that RL paired with techniques such as lazy loading, state aggregation, or imitation learning allowed the RL agent to perform more complicated tasks. Further, they argue that fully model-based agents may not be ideal as they are not as versatile; instead, they suggest model-free RL agents with rich a priori knowledge. Also, from 2020 is the work of Chaudhary *et al.* on automated post-breach penetration testing with RL [5]. The authors propose the idea of using RL agents to find sensitive files in a compromised network; however, from their paper, it seems that they are still working on obtaining specific results. Earlier work by Ghanem *et al.* compared a reinforcement learning agent called IAPTS (Automated Penetration Testing System) against blind automation and found that this RL agent performed better [9]. Their IAPTS agent has the possibility of human input on the decision policy; this will allow the agent to learn and better approximate the expert's decisions. Unfortunately, it does not yet perform all the tasks that a human expert is doing manually, but the authors indicate research directions to improve their approach. Some specific penetration testing tasks have seen very little research that utilizes offensive machine learning. To our knowledge, there is only one study for conducting SQL injections[4] [7]. Erdődi et al. simulate penetration testing in a capture-the-flag setting, where the agent can choose between a number of candidate SQL injection queries. From the queries, the agent learns to first find the correct escape before searching for the flag.

4 Conclusion

The literature on ML for offensive cyber operations is considerably smaller than the literature on ML for defensive cyber operations. In this review paper, we

[4] There are many machine learning papers for discovering SQL injection attacks.

reviewed studies that apply ML in offensive cyber operations. Algorithmic-level cryptanalysis seems to be challenging for ML because modern cryptographic algorithms are designed to make learning hard as there is no indication of close to correct decryptions. However, there are papers that document modest success on weak cryptosystems. Significant advances in this approach would be needed to facilitate more success against modern algorithms. Perhaps even less researched is to perform ML-based penetration testing. One reason for this could be because there are already many automated tools that cyber-ops professionals use and because it is very important that penetration tests are conducted properly. Because penetration testing is a vast field, and we are at a very early stage in research on applying ML for penetration testing, there seems to be a great potential for advances in this area. For example, in the area of SQL injection, which represents a significant part of penetration testing, we only identified one study on ML-based SQL penetration testing.

References

1. Aiyanyo, I.D., Samuel, H., Lim, H.: A systematic review of defensive and offensive cybersecurity with machine learning. Applied Sciences **10**(17) (2020). https://doi.org/10.3390/app10175811, https://www.mdpi.com/2076-3417/10/17/5811
2. Alani, M.M.: Neuro-cryptanalysis of des and triple-des. In: International Conference on Neural Information Processing. pp. 637–646. Springer (2012)
3. Bagnall, A.J., McKeown, G.P., Rayward-Smith, V.J.: The cryptanalysis of a three rotor machine using a genetic algorithm. In: ICGA. pp. 712–718 (1997)
4. Benamira, A., Gerault, D., Peyrin, T., Tan, Q.Q.: A deeper look at machine learning-based cryptanalysis. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques. pp. 805–835. Springer (2021)
5. Chaudhary, S., O'Brien, A., Xu, S.: Automated post-breach penetration testing through reinforcement learning. In: 2020 IEEE Conference on Communications and Network Security (CNS). pp. 1–2. IEEE (2020)
6. Dourlens, S.: Applied neuro-cryptography and neuro-cryptanalysis of des. Master Thesis (1996). https://doi.org/10.13140/RG.2.2.35476.24960, advisor: Riesner, Christian
7. Erdodi, L., Sommervoll, Å.Å., Zennaro, F.M.: Simulating sql injection vulnerability exploitation using q-learning reinforcement learning agents. arXiv preprint arXiv:2101.03118 (2021)
8. Erdődi, L., Zennaro, F.M.: The agent web model: modeling web hacking for reinforcement learning. International Journal of Information Security pp. 1–17 (2021)
9. Ghanem, M.C., Chen, T.M.: Reinforcement learning for intelligent penetration testing. In: 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). pp. 185–192. IEEE (2018)
10. Gillogly, J.J.: Ciphertext-only cryptanalysis of enigma. Cryptologia **19**(4), 405–413 (1995)
11. Gohr, A.: Improving attacks on round-reduced speck32/64 using deep learning. In: Annual International Cryptology Conference. pp. 150–179. Springer (2019)
12. Lasry, G., Kopal, N., Wacker, A.: Cryptanalysis of enigma double indicators with hill climbing. Cryptologia pp. 1–26 (2019)

13. Ostwald, O., Weierud, F.: Modern breaking of enigma ciphertexts. Cryptologia **41**(5), 395–421 (2017)
14. Singh, S.: The code book: the science of secrecy from ancient Egypt to quantum cryptography. London: Fourth estate (2000)
15. Sommervoll, Å.: Dreaming of keys: Introducing the phantom gradient attack. In: 7th International Conference on Information Systems Security and Privacy, ICISSP 2021, 11 February 2021 through 13 February 2021. SciTePress (2021)
16. Sommervoll, Å.Å., Nilsen, L.: Genetic algorithm attack on enigma's plugboard. Cryptologia pp. 1–33 (2020)
17. Williams, H.: Applying statistical language recognition techniques in the ciphertext-only cryptanalysis of enigma. Cryptologia **24**(1), 4–17 (2000)
18. Zennaro, F.M., Erdodi, L.: Modeling penetration testing with reinforcement learning using capture-the-flag challenges and tabular q-learning. arXiv preprint arXiv:2005.12632 (2020)