CRYPTOGRAPHICALLY SECURE ENCRYPTION USING ADVERSARIAL NEURAL NETWORKS

A Research Paper submitted to the Department of Computer Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Nicholas Winans

November 15, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Daniel G. Graham, Department of Computer Science

Cryptographically Secure Encryption Using Adversarial Neural Networks

CS 4991 Capstone Project, 2021

Nicholas Winans
Computer Science
University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
nw5zp@virginia.edu

ABSTRACT

The computational promises made by quantum computing threaten to destroy our notion of secure communication in the digital world. Neural networks (NNs) are one area of ongoing research that appears to be a promising method to enable cryptographically secure communication in a post-quantum world. Generative Adversarial Networks (GANs) is a specific type of neural network where multiple neural networks are pitted against each other. GANs pit a generative NN pitted against an adversarial NN. The generative NN attempts to encrypt data in a recoverable yet secure manner while the adversarial NN attempts to decrypt the output from the generative NN back into the original data.

Due to their adversarial nature, GANs appear to be a promising method of enabling secure communication due to the nature of having a constant adversary necessitating innovation. There are multiple research projects already completed that attempt to use ANNs to successfully encrypt data from eavesdroppers attempting to recover information.

GANs are only a piece of the puzzle, however. Researchers are combining NNs with existing and emerging technologies to provide resilient and promising encryption schemes.

1 Introduction

Quantum computing represents a fundamental shift in computing. While classical computer utilizes binary digits, meaning that values are stored as 0's and 1's and can be combined to form more complex letters and numbers, quantum computing utilizes so-called quantum bits, which reflect probabilities. No longer are values restricted to 0 or 1; now they can take on any value between 0 and 1. [1] Quantum bits reflect the probability of being in any state across all the potential quantum states – a system is not in any one state until it is measured [1] – but represent the potential for quantum computing to disrupt many of the assumptions that enable modern computing. In fact, given n quantim bits in a quantum computer system, 2^n parallel computations can be performed [1]. The increased computational power promises to revolutionize many different fields, from finance to medicine, solving many problems traditionally labelled as too computationally complex [2].

1.2 Cryptography and Quantum Computing

Cryptography is the field of making communications private, even in the presence of malicious agents attempting to steal information. As Kathleen Richards describes it [3]:

> [Cryptography] refers to secure information and communication techniques derived from mathematical concepts and a set of rule-based calculations called algorithms, to transform messages in ways that are hard to decipher. These deterministic algorithms are used for cryptographic key generation, digital signing, verification to protect data privacy, web browsing on the internet and confidential communications such as credit card transactions and email.

As opposed to the fields of business and medicine, where more computational power is good, cryptography is predicated on assumptions on the difficulty of certain problems. Difficulty in computer science is measured as a problem belonging to the NP problem space as opposed to P [4]. Nondeterministic Polynomial problems (NP) have no known solution markedly faster than brute force – trying every combination of parameters to find a solution – as opposed to Polynomial problems (P), which have a solution that scales in a polynomial fashion with respect to the amount of time necessary to find a solution as the number of possible inputs increases [4]. There are many assumptions in Cryptography upon the NP-hardness of problems, such as the discrete logarithm and RSA (factorization) problems. [5]. The discrete logarithm problem, for example, relies on the difficulty of finding x, y given $g, f(x, y)$ in the equation $f(x, y) = g^{xy}$ [5]. There are two distinct areas of cryptography – sharing a key to encrypt with, and sharing data encrypted with the key – and both are affected by quantum computing.

Quantum computing has already been shown to break the NP-hardness of the discrete logarithm problem [1]. Researchers are already developing alternative mathematical algorithms that are resistant to the enhanced computations of quantum computing, such as Lattice-based cryptography [6] and Multivariate public key cryptography [7]. Alternative solutions to cryptography in a post-quantum world also exist. One recent novel attempt at encryption is the use of Generative Adversarial Networks from the field of Machine Learning and Artificial Intelligence.

1.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a form of Artificial Neural Networks where multiple neural networks are pitted against each other [8]. There is a common training set of data between the models [8]. A Generative neural network (G) attempts to create data that is indistinguishable from the training data to a Discriminatory neural network (D) [8]. Both networks are trained simultaneously based on a training iteration's results, like normal neural network training [8].

Neural Networks in general are trained using a loss function, which is supposed to quantify what it means to be right in the context of a specific network. More specifically, the loss function quantifies the distance between a prediction and the correct output for a set of inputs [9]. The model learns by adjusting prediction parameters to minimize the loss of the model [9].

Researchers have recently used GANs in unique applications with the intended result of cryptographic encryption from an adversary. As detailed later in this report, these techniques combine existing or emerging technologies with GANs to enable secure communication between parties over insecure channels, like the internet.

2 Related Works

Before investigating the use of GANs for cryptographic encryption, it is important to recognize other methods for secure encryption in a post-quantum world. There are two main strategies that are currently being researched and developed: quantum key distribution-based solutions and mathematical-based solutions [1].

2.2 Quantum Key Distribution-based Solutions

As the name suggests, quantum key distribution (QKD) uses properties of quantum mechanics to tackle the problem of distributing a key to be used with encryption. One fundamental concept of quantum mechanics is entanglement [10]. Entanglement describes the phenomenon where two or more particles are intertwined, and any changes to one particle instantly is reflected in the state of the other particle [10].

The measuring of the state of a quantum system disturbs the system [10]. This disturbance means that any attempt to eavesdrop on the distribution of information through quantum entanglement will result in noticeable disruptions in the system. This alter of eavesdropping allows two parties to share information with peace of mind, with some caveats of requiring trust between two parties before the exchange happens [11].

2.3 Mathematical-based Solutions

Mathematical-based solutions to quantum computing's computational power are more in line with modern cryptography, which is based upon mathematically provable security. Lattice based cryptography is one such mathematical based solution. [1] Lattice based cryptography attempts to fix the ease of factorization of primes for quantum computers by using the multiplication of matrices [6].

Multivariate public key cryptography is also a well-researched cryptography alternative that promises to be more resilient to quantum computing [1]. Multivariate public key cryptography can be used for digital signatures and asymmetric encryption and relies on the "difficulty of solving systems of multivariate polynomials over finite fields" [1]. Note that all mathematical based solutions to cryptography require the computational hardness of an operation, but given the qualities of quantum computing, some problems are hypothesized to remain difficult in a post-quantum world.

3 Generative Adversarial Networks in Cryptography

Generative Adversarial Networks are being used by researchers to generate encryption schemes that can hide information from adversaries.

3.2 Symmetric Key Encryption with Generative Adversarial Networks

"Learning to Protect Communications with Adversarial Neural Cryptography" [12] is the most straightforward application of GANs towards modern cryptography. It utilizes a symmetric key, where the sending and receiving party share an encryption key.

System Design

Abadi and Anderson designed the system to have three components, two Generative networks known as Alice and Bob, and a Discriminatory network known as Eve [12]. On any given iteration, Alice and Bob are given a shared key with the intention that the networks will use the key to hide data from Eve [12]. Alice is given a message to encrypt, and outputs a ciphertext [12]. The ciphertext is given to both Bob and Eve, who attempt to reconstruct the original message [12]. Alice and Bob, who are encrypting and decrypting a message, are not given any inclination about how to use the shared key to hide information,

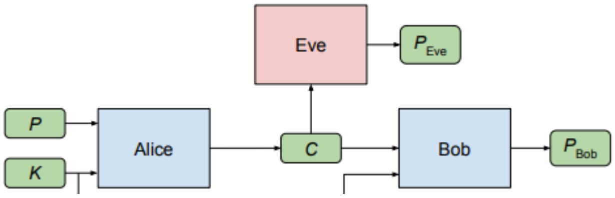

Figure 1: Alice, Bob and Eve. Depicts the layout of Alice, Bob and Eve Neural Networks in a symmetric key encryption scheme [12]

and instead the networks learn how to do this adversarially [12]. The design of Alice, Bob and Eve is shown in Figure 1:

The loss function for Eve was the L1 distance between two bitstrings, which is equivalent to the number of bits the strings differ upon. [12] This intuitively represents the number of bits Eve was able to recover. The loss function of Alice and Bob considers both the number of bits they were able to recover, as well as the number of bits Eve was able to get right. [12] This also intuitively makes sense, as Alice and Bob want to correctly send information, and hide it from Eve.

Results

Abadi and Anderson were successful in their endeavor to teach the Generative networks to hide and recover plaintext messages. As summarized in Figure 2, the researchers tried to send a 16 bit message in the system. [12] We can assume that the 0's and 1's in the message are uniformly distributed across a large sample of input messages, meaning that every bit has a 50/50 chance of being either 0 or 1. Intuitively, this explains why both Bob and Eve were able to recover 8 bits of information in the beginning, by guessing all 0 or all 1, they would expect to get 8 of the 16 bits correct.

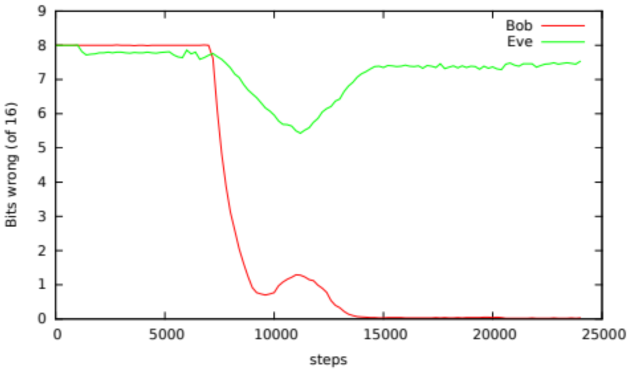

Figure 2: Results of Symmetric GAN Encryption. Depicts the number of incorrectly guessed bits throughout training for both Bob and Eve. [12]

Over the course of 20,000 steps (each step included training across 4096 samples), Bob was able to learn to recover all 16 bits consistently, while Eve did not have any advantage over randomly guessing. [12] This means that the researchers achieved both the goal of successful decryption of the original message, as well as the hiding of the message from Eve.

3.3 Asymmetric Key Encryption with Generative Adversarial Networks

"Asymmetric cryptographic functions based on generative adversarial neural networks for Internet of Things" [13] performs a similar analysis as Abadi and Anderson's research, but in the context of asymmetric key encryption. Instead of Alice and Bob sharing a key to encrypt with, they use an asymmetric keypair, also known as a public/private keypair [14]. In this form of cryptography, an actor, Bob in this case, generates a pair of keys [14]. One of these keys he gives to everyone in the network, known as the public key, that people can use to encrypt with [14]. With public key encryption, only the private key can decrypt information encrypted with the public key [14]. So, only Bob, in sole possession of the matching private key, should be able to decrypt messages encrypted with the public key.

System Design

Hao et al. designed a Generative Adversarial Network with two Generative networks, designated Alice and Bob, and one Discriminatory network, designated Eve [13]. Bob generates a public key/private key pair, and gives the public key to Alice [13]. Alice encrypts a plaintext message, P, with the public key, creating a ciphertext [13]. This ciphertext is passed to Bob and Eve, who attempt to decrypt the information, yielding P_{Bob} and P_{Eve}, the respective attempt at decrypting the ciphertext back into the original plaintext, P [13]. This construction is laid out in Figure 3, shown below:

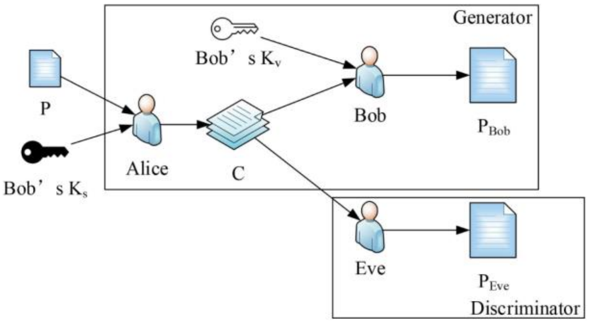

Figure 3: Layout of Asymmetric GAN Encryption. Depicts the dissemination of keys and information in the network. [13]

The loss function for this Eve was the L1 distance between P_{Eve} and P, meaning the number of bits different between the two strings [13]. The loss function for Alice and Bob was similarly the L1 distance between P_{Bob} and P, subtracting Eve's loss function from this result to reflect the complete goal of Alice and Bob to hide information as well as recover encrypted information. [13]

Results

Hao et al. were successful in their endeavor to teach the GAN to securely encrypt information with asymmetric keys. In Figure 4 below, we have a graph of the probability of Eve and Bob guessing an arbitrary bit in a message incorrectly. [13] Again assuming that the message and ciphertext are uniformly distributed between 0 and 1, we notice that in the long run, Eve has no advantage in recovering the plaintext message over random guessing. [13]

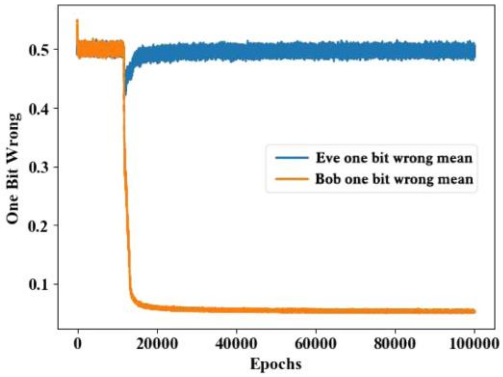

Figure 4: Results of Asymmetric GAN Encryption. Depicts the probability of guessing a bit of the original plaintext incorrectly. [13]

At the same time, Bob can recover the bit with almost complete certainty, reflecting that the learning objectives of the system were met, as Alice and Bob were able to encrypt and recover information, while Eve was not able to recover any information about the message.

3.4 Steganography with Generative Adversarial Networks

Steganography is the hiding of information within another object in such a way that the presence of hidden information cannot be detected [15]. Think of how words are hidden in a crossword, but if the words were not meant to be found and were hidden in more creative ways. GANs promise to improve our ability to hide information in ordinary digital objects.

System Design

Shi et al. designed a Generative Adversarial Network with a single Generative network G, and two Discriminatory networks, D and S [15]. The goal of this network is to create cover images that can be used in steganography to hide images [15]. By preparing images for steganography, we are decreasing the likelihood that information could be identified when hidden. The Generative network takes an input of random noise and applies the noise to images, creating a cover for steganography [15]. The Discriminatory network D's responsibility is to evaluate the quality of the cover by trying to identify between the original image and the prepared cover [15]. The Discriminatory network S's responsibility is to assess the capacity of the cover to be used for steganography [15]. A visual representation of the network is shown in Figure 5:

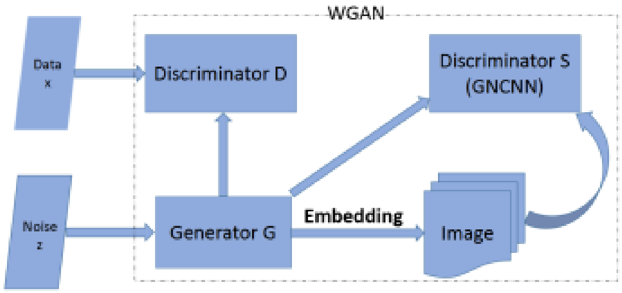

Figure 5: A Steganography Cover GAN [15]

Results

Shi et al. successfully created a GAN network more secure than a reference network [15]. The Discriminatory S network was trained with real images, and against a reference steganography network, was able to detect 92% of real images and 90% of generated images [15]. When ran against the SSGAN network proposed in the paper, it was only able to achieve 87% accuracy against real images and 72% against generated images [15]. The network is successfully able to fool the discriminatory network S at a higher rate than the reference benchmark.

4 Conclusion

There are numerous examples of researchers successfully teaching Generative Adversarial Networks to adapt current cryptography practices without explicitly telling the networks how to encrypt information. As quantum computing inches closer to reality, it is increasingly important that we prepare and adapt against its threats to modern cryptography. If we are not prepared, security online will not exist in a recognizable form, and things like online shopping and banking will cease to exist.

Thankfully, there are many quantum-resistant technologies being developed, from mathematical-based models to novel methods using GANs. As GANs and neural networks in general continue to evolve, more use cases will begin to emerge.

5 Acknowledgements

I would like to acknowledge the help of my advisor for this project, Daniel Graham, Associate Professor of Computer Science. Additionally, all the professors that contributed the building block upon which this research report as written. Specifically, Mohammad Mahmoody, Associate Professor of Computer Science for his class on Cryptography, and N. Rich Nguyen, Assistant Professor of Computer Science for his class on Machine Learning.

References

[1] V. Mavroeidis, K. Vishi, M. D. Zych and A. Jøsang, "The Impact of Quantum Computing on Present Cryptography," *(IJACSA) International Journal of Advanced Computer Science and Applications,*, vol. 9, no. 3, 2018.

[2] F. Bova, A. Goldfarb and R. Melko, "Quantum Computing is Coming. What Can It Do?," Harvard Business Review, 16 July 2021. [Online]. Available: https://hbr.org/2021/07/quantum-computing-is-coming-what-can-it-do. [Accessed 26 October 2021].

[3] K. Richards, "What is Cryptography?," SearchSecurity, September 2021. [Online]. Available: https://searchsecurity.techtarget.com/definition/cryptography. [Accessed 6 October 2021].

[4] D. Miessler, "P vs. NP Explained," Daniel Miessler, 1 November 2017. [Online]. Available: https://danielmiessler.com/study/pvsnp/. [Accessed 26 October 2021].

[5] V. Goyal, *Lecture 11: Key Agreement,* CMU Department of Computer Science, 2019.

[6] D. Micciancio and O. Regev, "Lattice-based cryptography," in *Post-Quantum Cryptography,* Berlin, Springer, 2009, pp. 147-191.

[7] J. Ding and B.-Y. Yang, "Multivariate public key cryptography," in *Post-Quantum Cryptography*, Berlin, Springer, 2009, pp. 193-241.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems,* vol. 27, 2014.

[9] S. Verma, "Understanding different Loss Functions for Nerual Networks," towards data science, 20 June 2019. [Online]. Available: https://towardsdatascience.com/understanding-different-loss-functions-for-neural-networks-dd1ed0274718. [Accessed 26 October 2021].

[10] R. Horodecki, P. Horodecki, M. Horodecki and K. Horodecki, "Quantum entanglement," *Review of Modern Physics,* vol. 81, no. 2, p. 865, 2009.

[11] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus and M. Peev, "The security of practical quantum key distribution," *Reviews of modern physics,* vol. 83, no. 3, p. 1301, 2009.

[12] M. Abadi and D. G. Andersen, "Learning to Protect Communications with Adversarial Nerual Cryptography," *arXiv preprint arXiv:1610.06918,* 2016.

[13] X. Hao, W. Ren, R. Xiong, T. Zhu and K.-K. R. Choo, "Asymmetric cryptographic functions based on generative adversarial neural networks for Internet of Things," *Future Generation Computer Systems,* vol. 124, pp. 243-255, 2021.

[14] Cloudflare, "How does public key encryption work?," Cloudflare, [Online]. Available: https://www.cloudflare.com/learning/ssl/how-does-public-key-encryption-work/. [Accessed 26 October 2021].

[15] H. Shi, J. Dong, W. Wang, Y. Qian and X. Zhang, "SSGAN: Secure Steganography Based on Generative Adversarial Networks," *Pacific Rim Conference on Multimedia,* pp. 534-544, 2017.