# Generative Adversarial Networks: What Are They and Why We Should Be Afraid

Thomas Klimek

2018

## Abstract

Machine Learning is an incredibly useful tool when it comes to cybersecurity, allowing for advance detection and protection mechanisms for securing our data. One particularly potent machine learning concept is the Generative Adversarial Network (GAN) which is the key focus of this paper. The GAN has many applications related to cyber security including strengthening existing attacks to levels beyond what can be handled by a basic detection system. As GANs rise in popularity, the need to defend against and recognize GAN attacks also becomes increasingly urgent. This paper will detail the methods of GAN attacks and attempt to answer the question of how can we defend against these attacks.

## Introduction

The Generative Adversarial Network, or GAN for short, is predicted to be the next big thing in Machine Learning. The core idea of a GAN is given a large set of data, the GAN is capable of generating brand new unique data that is effectively indistinguishable from the original. Since its world debut in 2014, GANs have picked up a lot of attention and critical acclaim, even being called "the most interesting idea in ML in the last 10 years" by Director of Facebook AI research Yann LeCun [5]. Despite the excitement surrounding GANs, they do pose a major threat from the perspective of cybersecurity and have profound applications in fields such as password cracking, malware detection, facial recognition, and more. The goal of this paper is to give an introduction to the GAN, discuss the potential security risks posed by GANs, and propose defense mechanism to employ for advanced detection and prevention of GAN attacks.

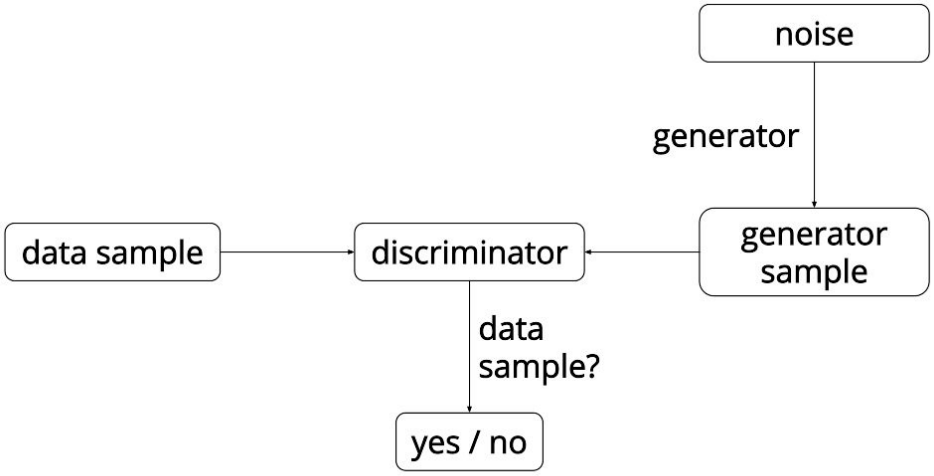
## To the Community

A constant struggle of cyber security is learning to defend against newer, more advanced attacks that we haven't seen before. One solution of this is to use machine learning to analyze existing attacks, learn from them, and extrapolate to what a future attack might look like. The GAN is a direct counter to what can be achieved by machine learning. By taking an adversarial approach to machine learning, attackers can create attacks with intricacies so complex they fool our state of the art systems. GANS have the capability to fool even machine learning based defense systems. This means any system reliant on machine learning is no longer safe. Without understanding the inner workings of a GAN, we are at the perilous mercy of these attackers. However, if we do take the time to understand GANs we can equip ourselves for the future and keep our data safe.

## How GAN works

A GAN consists of two main parts, the generator and the discriminator. These two parts are often described using an analogy between a criminal and a crooked cop. The criminal is interested in creating counterfeit money, and the crooked cop is willing to help. The criminal starts with a poor quality counterfeit bill and brings this to the cop for feedback. The cop then tells the criminal that the bill is fake, and gives him feedback on how he could tell. So the criminal goes back to work, considering the cops feedback, and comes back with a new version of the bill. The cop again can tell this is fake, and gives the criminal more feedback. This cycle then repeats indefinitely until the cop can no longer tell that the bill is fake! This is the key process behind a GAN, where the criminal is the generator, the cop is the discriminator, and the counterfeit bill is the data being generated[4].

This can also be understood using the figure below, detailing the architecture of a GAN [4]. First, the generator is given noise which it transforms into a sample of the data using its current mathematical model, initially being completely random. The discriminator then gives feedback through answering yes/no for the given data. Based on this feedback the generator and the discriminator update their mathematical models in order to "learn" based on the previously generated sample. After this process is repeated many times, the final result is a data sample which follows the same distribution of the original/training data. For an in depth understanding of the math behind the GAN, it is worth reading the GAN tutorial written by the creator of GANS, Ian Goodfellow[4].




## GAN & Cybersecurity

Now that we understand the basic processes behind a GAN, we can being to talk about some of the applications of GANs in cyber security. For this paper I chose to talk about 3 specific adversarial uses of a GAN in password cracking, malware, and facial detection. It is important to keep in mind that there are many more areas of cyber security that are vulnerable to adversarial GANs, however these 3 can provide valuable insight about the nature of adversarial GAN attacks, and ideas presented can be extrapolated to fit different fields.

## Password Cracking

Current state of the art password cracking techniques involve computing millions of hashes from a large word list and comparing these hashes to the password hashes we are trying to crack. These word lists usually consist of commonly used or previously used passwords, however these lists are not fully comprehensive. Using this approach your password cracking power is only as strong as your word list. Experienced password crackers will normally supplement their word list with a list of rules that augment this list, making it more exhaustive. For example common rules include adding strings such as "123" at the end of a password, or replacing letters with numbers, removing vowels, and much more. The one drawback to this approach is all rules must be explicitly written out and devised by some clever password cracker. Here is where a GAN can come into play. By training a GAN on a large dataset of passwords, it will begin to recognize complex information and patterns that it will later use to guess more passwords.

An excellent example of this is the PassGAN system developed by machine learning researchers[1]. PassGAN is trained on the "rockyou" dataset, an industry standard password list, and does an incredibly effective job of both mimicking the distribution of rockyou, and guessing new unique passwords that are highly likely to be used somewhere. PassGAN researchers have reported that the GAN is able to match 10,478,322 (24.2%) out of 43,354,871 passwords from the linkedin password leak. This is significant because the GAN has not been exposed to any of the linkedin data, yet based on the rockyou words list it is able to generate meaningful unique passwords. It is also shown that PassGAN is an incredibly effective supplement to current password techniques. Used in conjunction with HashCat, the PassGAN was able to guess between 51% and 73% more unique passwords than hashcat alone[1]. If these numbers don't scare you yet, it is also worth mentioning PassGAN can output a practically unbounded number of password guesses. With password generation rules, the number of unique passwords that can be generated is defined by the number of rules and by the size of the password dataset used, however the output of PassGAN is not restricted to a small subset of the password space. As a result PassGAN was able to eventually guess more passwords than any of the other tools, even though all tools were trained on the same password dataset.[1]

## Hiding Malware

The uses of a GAN in cyber security are not limited to generating data, the GAN is also capable of evading detection systems. This can be specifically applied to creating malware that bypasses machine learning based detection systems. This topic is covered in depth in the paper *Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN* [3]. This research details another GAN system, MalGAN, that is capable of generating such malware, and even performs better than other adversarial machine learning approaches. The strength of this attack is that it performs with black-box systems, meaning the attacker has no knowledge of the detection system being used. The basic architecture of MalGAN can be seen in the figure below. The idea is similar to a regular GAN, however the Black-Box detector is used as a discriminator, and the generator is fed a combination of noise and malware examples.

Then the discriminator also receives benign examples to further inform the generator as to what is classified as “not malware”.

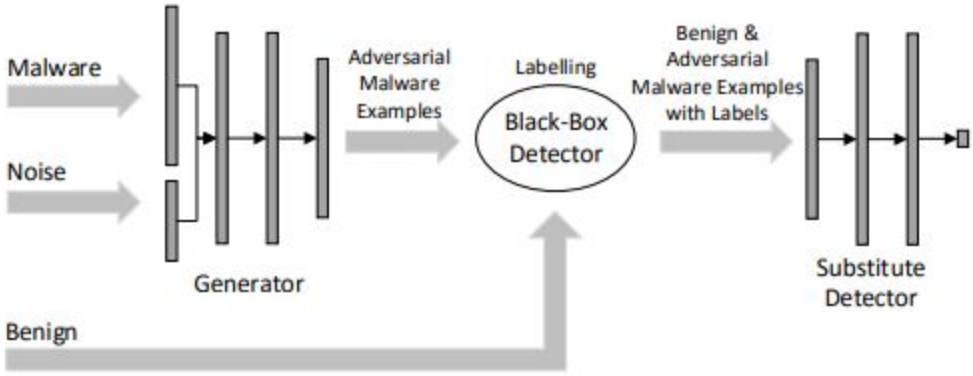



Figure 1: The architecture of MalGAN.

The applications of this technique are also profound. By providing a new novel approach to bypassing black-box malware detection, this means attackers need to know even less about the system to successfully attack it. Using machine learning to generate malware implies this malware will have increasing subtleties and complexities, and call for even more advanced detection systems. When employing a machine learning based detection system it is important to keep in mind that it will have its respective vulnerabilities.

## Forging Facial Detection

For the last section of the GAN applications, I will discuss one of the most common uses in image generation and manipulation. Specifically, a GAN can be used to fool existing image detection systems as well as generating high resolution fake images. For proof on concept, researchers at Michigan State and the International Institute of Information Technology have developed a system for hiding images from facial detection systems[8]. Specifically they use a semi-adversarial network to alter images in a way to bypass biometric recognition. This means the common systems in use like Facebook’s DeepFace which boasts a 97.35% accuracy[9] can be tricked using adversarial networks. Researchers have suggested uses of this GAN to increase privacy and protect from unwanted facial recognition, however it does provide a security threat to any facial recognition based security system. Another strong takeaway of this research is how even a seemingly simple task such as facial recognition, which has been extensively researched and has great accuracy, has its own respective vulnerabilities which can be manipulated by a GAN.

Along with fooling facial detection, GANs are often used to generate very realistic fake images. The research paper *Unpaired image-to-image translation using cycle-consistent adversarial networks* shows examples of using a GAN to alter real images to have certain properties such as changing a horse to a zebra, changing summer to winter, changing an apple to an orange, and much more [10]. While this paper is not concerned with cyber security, there are implications in the field of cyber security. Applying this technology adversarial could result in fake images and videos of political figures, other persons of interest, or even you. This poses a threat to national and personal security if photo realistic fake images

and videos are trivially generated. In this day and age seeing is no longer believing. This calls for an immediate need to develop detection systems and identify what is real and what is fake.

## Detection & Response

The final topic of this paper is detection and response. I would love to say there is an easy trick, or one quick fix or program you can run, but unfortunately, this is not the case. As you have seen from this paper that the techniques of a GAN are quite sophisticated and capable of fooling advance systems. So what can we do? The answer to this problem comes from education and awareness. Each solution to detect and respond to an adversarial GAN is unique to the context, but it is critical to understand the inner-workings of a GAN and prepare for the worst. For those in the field of machine learning based cyber security, the GAN will be rising in popularity in the near future. In order to detect and respond to GAN attacks it is critical to engineer your systems with a GAN in mind, and don't assume the machine learning detection is without vulnerabilities. One response to GAN generated images has already been developed called DeepFD. The DeepFD researchers created the system with the specific purpose of detecting adversarial GAN generated images that could stand to damage a person's reputation or personal safety [11]. The DeepFD system reports a 94.7% detection rate for fake images generated by state of the art GAN networks.

## Conclusion

The GAN teaches us that you really can't trust data. While it is a hot topic in machine learning research, it also comes with many cyber-security concerns given the ability to exploit vulnerabilities in state of the art systems. As GANs grow in popularity it is critical the cyber security practitioners keep up with the research and techniques to best equip themselves to understand and identify attacks. As we know, security is often an afterthought when it comes to exciting new technologies, so it is our responsibility as members in the field of cyber security to always consider what could go wrong, and to be two steps ahead of the attackers.

**Link to Powerpoint:**
https://docs.google.com/presentation/d/109CuCCoi0J3ZMa7q0Rmn8NN_ard9Z_4aM3SSjcF1Qtc/edit?usp=sharing

## References

[1] Hitaj, Briland, et al. "Passgan: A deep learning approach for password guessing." *arXiv preprint arXiv:1709.00440* (2017).
Link: https://arxiv.org/pdf/1709.00440.pdf

[2] Xu, Weilin, David Evans, and Yanjun Qi. "Feature squeezing: Detecting adversarial examples in deep neural networks." *arXiv preprint arXiv:1704.01155* (2017).
Link: https://arxiv.org/pdf/1704.01155.pdf

[3] Hu, Weiwei, and Ying Tan. "Generating adversarial malware examples for black-box attacks based on GAN." *arXiv preprint arXiv:1702.05983* (2017).
Link: https://arxiv.org/pdf/1702.05983.pdf

[4] Goodfellow, Ian. "NIPS 2016 tutorial: Generative adversarial networks." *arXiv preprint arXiv:1701.00160* (2016).
Link: https://arxiv.org/pdf/1701.00160.pdf

[5]https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-deep-learning
[6] https://securityintelligence.com/generative-adversarial-networks-and-cybersecurity-part-1/
[7] https://securityintelligence.com/generative-adversarial-networks-and-cybersecurity-part-2/

[8] Mirjalili, Vahid, et al. "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images." *2018 International Conference on Biometrics (ICB)*. IEEE, 2018.
Link: https://arxiv.org/pdf/1712.00321.pdf

[9]https://research.fb.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification/

[10] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *arXiv preprint*(2017).
Link: https://arxiv.org/pdf/1703.10593.pdf

[11] Hsu, Chih-Chung & Lee, Chia-Yen & Zhuang, Yi-Xiu. (2018). Learning to Detect Fake Face Images in the Wild.
Link: https://arxiv.org/ftp/arxiv/papers/1809/1809.08754.pdf