Research on Adversarial Machine Learning: Turning Turtles into Rifles in the Eyes of Machine Learning Systems

Scott Oslund

Summary

Machine learning is increasingly gaining importance in today’s society as it is being incorporated into our everyday lives with facial recognition systems, self-driving cars, and more. However, a new field of research called adversarial machine learning offers attackers a way to cause a machine learning system to fail by making slight modifications to the system input. This paper will serve as an introduction to adversarial machine learning for a graduate student or undergraduate student who is looking to get involved in researching adversarial machine learning, and who has experience using traditional machine learning. This paper will summarize the basic idea of adversarial machine learning and present the importance of this research. It will proceed to review the history of research into adversarial machine learning, the different methods used in attacks, and the strengths and weaknesses of each. In addition, it will review some of the theories as to why adversarial attacks are possible, and some proposed defenses against attacks. It will end by giving an overview of some of the important topics of future research in the field. This paper aims to give a reader with the proper prerequisite knowledge a solid foundation in adversarial machine learning which they can use to begin their own research.

Contents

1 Introduction

Adversarial machine learning is a subfield of machine learning studying how small changes to a machine learning system's input can cause the system to fail. As machine learning systems are increasingly being integrated into critical systems such as self-driving cars and facial recognition security systems, the dangers presented by adversarial machine learning increase. If these systems were to fail, the results could be deadly. For the safety of everyone on the road, a self-driving car must be able to recognize and classify other vehicles, pedestrians, road signs, and stoplights. A security system in an airport must be able to recognize any potentially dangerous weapons such as guns. Facial recognition systems have to be able to recognize a dangerous criminal and distinguish him from an innocent person. Adversarial attacks offer a way to cause each of these systems to fail. A simple example of an adversarial attack that we will study in more depth later can be seen in figure **??**. The image of a turtle is classified as a rifle by a machine learning system with high confidence thanks to minor modifications to the turtle's texture. These attacks are low cost and easy to create allowing for any malicious attacker to launch an adversarial attack on a victim neural network with only a day's work. These stakes have motivated researchers to study adversarial machine learning in an attempt to find what dangers are present and how effective adversarial attacks can be. Researchers are also working to study how to defend against these attacks and prevent machine learning systems from failing.

This paper will serve as an introduction to adversarial machine learning in the domain of image classification for anyone who is considering doing research in the field. It will start by chronologically outlining the various attempts to implement adversarial attacks against machine learning image classification systems. It will review the various methods used, their effectiveness, and issues with each approach. After thoroughly studying the various methods for launching adversarial attacks being researched today, the paper will introduce some of the major ways researchers have attempted to defend against adversarial attacks. It will review the main ideas behind the defenses and their strengths and weaknesses. Lastly, it will also provide a brief overview of some of the major theories on why adversarial attacks are possible. By the end of this paper, the reader should have a strong understanding of the major concepts and theory behind adversarial machine learning, the history of adversarial machine learning research, the most popular types of attacks, and the benefits and drawbacks of different methods. The reader will also learn about some of the theoretical defenses against such attacks and future areas of research in the field. In short, this paper will give the reader the needed information to begin researching adversarial machine learning themselves.

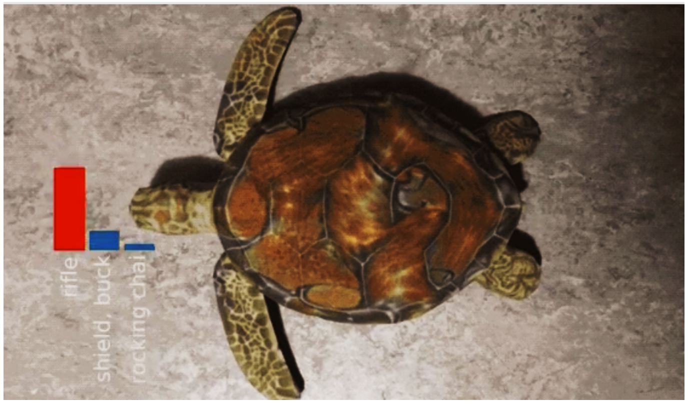

Fig. 1. An Adversarial Attack Causing a turtle to be classified as a rifle [6]

This paper assumes the reader has some knowledge of traditional machine learning. Specifically, to get the most out of this paper you should be familiar with the concept of gradient descent and the machine learning training loop. The target audience is a graduate student or experienced undergraduate student interested in getting involved in adversarial machine learning research for the first time. This paper aims to give you the tools you need to get started in this field of research.

2 What is Adversarial Machine Learning?

The central idea of adversarial machine learning is altering an input into a machine learning system to cause the system to fail for that input. These attacks can be launched against any machine learning system. For example, some extra tones could be added to an audio signal to cause a voice-recognition system to fail or small changes in the colors of pixels in an image could be used to launch an adversarial attack against an image classifier causing it to misclassify the image. Almost all of these attacks are gradient-based. A normal machine learning system learns by using the gradient of the loss with respect to the machine learning system's weights to optimize the system to minimize loss. In contrast, gradient-based adversarial attacks optimize the input to the system to maximize loss using the gradient of the loss with respect to the input. These attacks are used to cause a machine learning system to fail to classify the input once the alterations have been added.

We can define a few key terms unique to adversarial machine learning to facilitate our discussion. First, the victim neural network/image classifier refers to the machine learning system which is being targeted by an adversarial attack. Another key point is what we mean when we say something is adversarial. When we say something is adversarial, like an adversarial image or an adversarial sticker, we are indicating it has gone through an adversarial machine learning algorithm and has been optimized to cause misclassifications in image classifiers. The adversarial attack success rate is defined as the percentage of images misclassified over the total number of images tested. So for example, if a set of images altered by an adversarial machine learning algorithm had only a few of the images misclassified, there would be a low attack success rate indicating the attack was not very successful. Conversely, a high attack success rate indicates a large number of the images were misclassified and the adversarial machine learning algorithm was successful in launching attacks against the victim image classifier. Finally, a white-box adversarial attack is an attack in which the attacker has complete information about the architecture of the victim neural network. A black-box attack is one in which the attacker has no knowledge of the victim neural network's architecture. Practically, this means white-box attacks will train adversarial images directly using the target victim neural network while black-box attacks will train adversarial images using a separate image classifier. As we will see later in this paper, typically white-box attacks are somewhat more successful than block-box attacks.

With an understanding of the basic idea behind adversarial machine learning and with key terms defined, we can begin to delve into the specifics of how attacks are launched using different methods, their effectiveness, and issues. We will do this by stepping through key high-impact research papers in adversarial machine learning for image classification chronologically.

3 A History of Adversarial Machine Learning Research

3.1 Early Adversarial Machine Learning Research (2013 - 2015)

The first research into adversarial machine learning in the domain of image classification emerged in 2013. Researchers Christian Szegedy and Ian Goodfellow began to create the first adversarial attacks that used the idea of gradient ascent, the inverse of gradient descent, to maximize loss [1] [2]. They based their work on past machine learning research which sought to minimize loss by altering weights in a machine learning model. Szegedy and Goodfellow realized that the same basic process could be used to optimize a machine learning model's input to maximize loss in an attack against the system causing misclassifications. These early attacks would intake an input image, pass it through the victim machine learning model, compute the gradient of the input image with respect to the machine learning model's loss, and use the gradient to make small modifications to the input image to maximize loss. The basic algorithm of this method and most subsequent adversarial machine learning is known as the Fast Gradient Sign Method (FGSM) attack [2]. The steps are shown in Algorithm 1.

In short, Szegedy and Goodfellow's algorithm quickly increases a machine learning

Algorithm 1 Fast Gradient Sign Method Algorithm

```
Input: Clean Image $i$, Machine Learning Model $M()$, Training Step $\epsilon$, Correct Image Classification $C$, Function that extracts the gradient w.r.t. the correct classification Grad()
Output: Adversarial Image $A$
  for $n$ iterations do
      $\nabla \leftarrow$ Grad($M(i)$,$C$);
      $i \leftarrow i + \epsilon \cdot$ sign($\nabla$)
  end for
  $A \leftarrow i$;
```

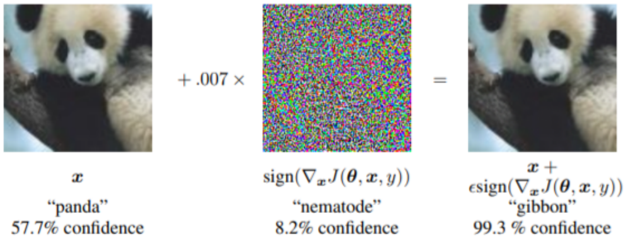
+ .007 ×
=
x
"panda"
57.7% confidence
$\text{sign}(\nabla_{x} J(\theta, x, y))$
"nematode"
8.2% confidence
$x + \epsilon\text{sign}(\nabla_{x} J(\theta, x, y))$
"gibbon"
99.3 % confidence

Fig. 2. Adding Adversarial Perturbations to Panda Image [2]

model's loss by iteratively adding some multiple of the gradient with respect to the correct classification to the original image. In his paper Szegedy found this algorithm to produce adversarial images capable of fooling advanced image classification models with changes so minor they remain invisible to the human eye. In every test case, they managed to produce an adversarial image capable of fooling the system while remaining hidden to an onlooking human using only thirty iterations of the FGSM attack algorithm. A famous example adversarial attack from Szegedy and Goodfellow's paper [2] can be seen in figure 2.

The figure displays how small amounts of noise when added to a normal image can cause a neural network to fail. While the clean image and adversarial image look indistinguishable to a human, to a computer they are very different. With a high level of confidence, the image classifier assigns the wrong classification to the attacked image while giving the correct classification to the original image. All of this early research was conducted using white-box attacks where the architecture of the victim neural networks was known. While Szegedy and Goodfellow achieved excellent results using the FGSM attack on still images, their approach remained limited in that it required direct access to the image data in the computer system. With this limitation, to successfully launch an attack against a facial recognition system an attacker would have to have access to the camera taking the picture or to the data being transmitted from the camera to modify image data directly. This access is often difficult to obtain and requires finding and exploiting another security flaw in the system. To make adversarial attacks more practical, later research would focus on physical adversarial attacks. Physical adversarial attacks aim to create adversarial images by modifying the physical environment the camera takes a photo of, rather than directly modifying image data.

3.2 Physical Adversarial Attacks (2015 - 2017)

Early physical adversarial attack research began in 2015. Alexey Kurakin et al. [3] researched the possibility of physical adversarial attacks by training adversarial images in the same way as Szegedy and Goodfellow, printing the adversarial images out, taking photos of the printed adversarial images, then passing the new photos into a machine learning model. While the authors did manage to cause the machine learning model to fail in some cases, they did not achieve nearly the same level of success as the virtual attacks. By using greater amounts of perturbation, Kurakin managed to achieve an average misclassification rate of about 66%, far less than the universally successful non-physical adversarial attacks proposed by Szegedy and Goodfellow. Using this method perturbations had to be large enough to be visible to the human eye to be effective, unlike in Szegedy and Goodfellow's research. This early experiment demonstrated that physical adversarial attacks are significantly more difficult than the non-physical attacks Szegedy studied because of various changing imperfections in the real world such as blur, lighting changes, and camera angle. All of these changing conditions can prevent the subtle adversarial perturbations generated in an attack from being picked up on by the image classification system.

Although not a focus of the paper, another contribution of Kurakin's work was the introduction of tests using the black-box model of attack. In these attacks, since the architecture of the victim neural network is unknown, adversarial images are trained using a different neural network the attacker does have access to. For example, in Kurakin's paper adversarial images were trained using the ImageNet inception classifier while some of the tests were done using a TensorFlow-built image classifier with unknown architecture. In the few tests done using this method, Kurakin showed there to be transferability of adversarial attacks [3]. An adversarial attack trained using one victim neural network could often cause misclassifications in another neural network. However, attacks against the black-box neural network tended to be less consistently effective.

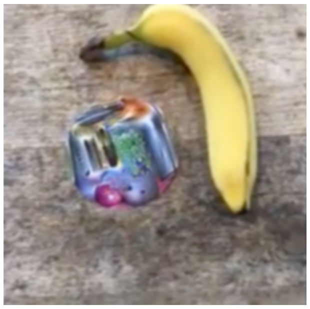

Fig. 3. Photograph of a Banana with a Adversarial Patch [4]

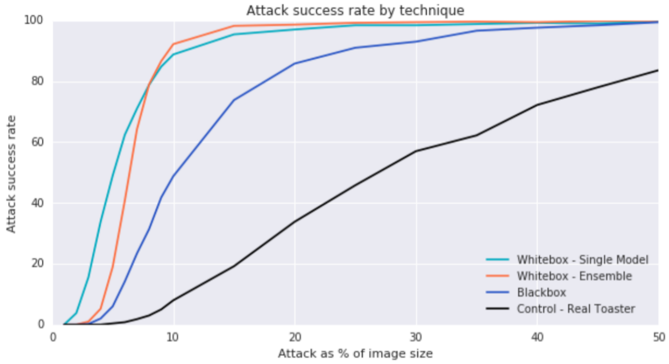

Fig. 4. Data from Brown's research on the effectiveness of adversarial attacks against a toaster using white-box and black-box attacks [4]

To achieve high physical adversarial attack success rates researchers such as Tom Brown [4] proposed creating an adversarial patch. An adversarial patch is a printable perturbation trained to maximize loss. The patch could be placed onto the target object or in the background to maximize loss. The patches were trained using a method similar to Szegedy's. The basic method remained the same and a training image was still required, however many more training iterations with a larger training step were used. While these patches were no longer subtle and invisible to a human onlooker, they were also more robust to various physical conditions and only covered a small portion of the physical object. The patches were tested by taking photos of physical objects with the patches attached and running the images through a white-box image classifier and a black-box image classifier. Brown found that by creating an adversarial patch that covers about ten percent of the area of a physical image, an attack success rate of around ninety percent could consistently be achieved for a white-box attack. For a black-box attack, about twenty-five percent of the image would have to be covered by an adversarial attack to achieve a similar attack success rate. A graph of Brown's data can be found in figure 4. An example of an adversarial patch to prevent the correct classification of a banana is shown in figure 3.

While physical adversarial attacks were proven to be viable thanks to Brown's research, there appeared to be a trade-off between the subtlety of the attack and the attack success rate [3] [4]. To achieve a high attack success rate it seemed necessary to create a patch that would be noticeable to humans. Future research would begin to focus on creating adversarial attacks that achieved high attack success rates while remaining disguised to human onlookers.

3.3 Disguised Physical Adversarial Attacks (2017 - Present)

A wide variety of approaches for creating disguised adversarial attacks quickly emerged and are continually being created. One example comes from Kevin Eykholt et al. [5] who aimed to disguise physical adversarial attacks

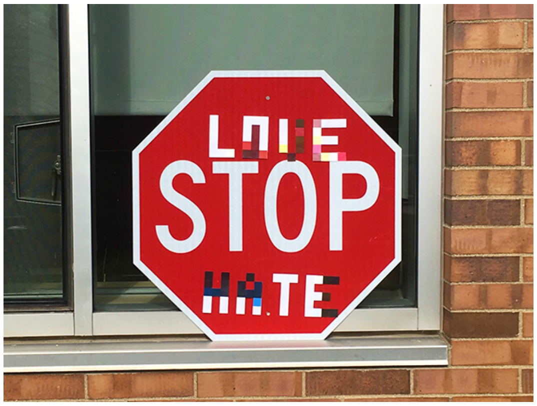

Fig. 5. Disguised Stop Sign Adversarial Attack [5]

as art or graffiti. These researchers attempted to make the attacks less obvious by adding constraints to their training process for generating adversarial patches. They only let their attacks target certain areas of the target object where they might be more inconspicuous and they limited the colors the adversarial patches could become. By adding in these constraints, Eykholt created adversarial patches that could be designed to look more like graffiti than the patterned and colorful stickers that previous adversarial stickers had looked like. For example, by targeting specific areas of a stop sign to spell out words and by limiting the colors of perturbations to only be shades of white and black, Eykholt managed to create an effective adversarial attack against a stop sign that appears to be graffiti. In the research paper, the adversarial perturbations are designed to spell out ”LOVE” and ”HATE” in just a few set colors. The image of the attacked stop sign can be found in figure 5. Eykholt's new method managed to achieve a greater than two-thirds attack success rate across many different camera angles and distances. While the achieved attack success rate is lower than that of a normal adversarial patch, it has the benefit of being disguised.

In the same research paper, Eykholt et al. [5] also proposed covering the surface of the target object with a poster imitating the object's normal surface but with small adversarial alterations. While this method requires a much larger adversarial patch to cover the whole object, it also achieves a higher attack success rate. In addition, the patch perturbations can be made less obvious since the patch covers a greater portion of the image. In testing an adversarial patch that covers the surface of a stop sign, the authors achieved a one hundred percent attack success rate.

Another new method for launching disguised adversarial attacks comes from Anish Athalye et al. [6] who instead of trying to modify the target object to make it adversarial, aimed to create a copy of the object designed to be adversarial using a 3D printer. The authors aim to create an adversarial attack that has a high attack success rate, by using their custom control over the entire object and texture, while keeping the object looking normal to a human onlooker. This method also has the benefit of working well for a variety of viewpoints and angles since every part of the object is designed to be adversarial instead of just one section where a patch is placed. The method works by perturbing the texture of a 3D model of the target object using many training photos while still using the same basic concepts of the FGSM attack. Once the whole texture of the 3D model had been perturbed using many training photos, the new object was 3D printed using the adversarial 3D model. The authors found this method offers a high attack success rate of up to 82 percent from many angles while remaining inconspicuous to a human onlooker. An example 3D printed adversarial turtle from their paper can be found in figure 6. However, this method has the issue that a 3D printed recreation of the target object is needed and, depending on the object, a 3D printed version of it may not be possible or cause it to lose functionality. For example, while 3D printing a copy of a toy turtle might be viable, it would be very difficult to 3D print a functional copy of a gun.

4 Why do Adversarial Attacks Exist?

Having reviewed the history of research into adversarial machine learning we can begin to think about why adversarial machine learning

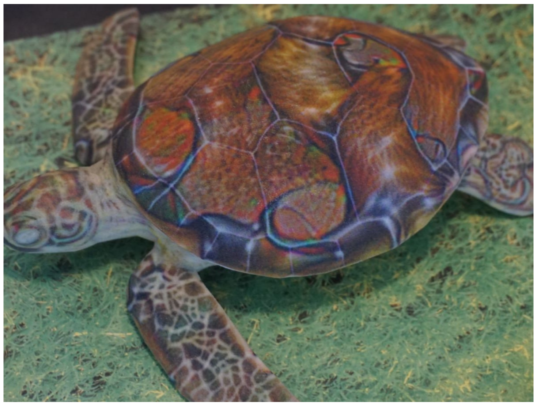

Fig. 6. 3D Printed Adversarial Sea Turtle [6]

is possible. Humans are not fooled by minor changes in pixel coloration into misclassifying an object so why are machine learning systems which are designed to mimic the way humans learn?

There are many different theories proposed by experts as to why adversarial attacks exist and why they can be so successful with such minor changes to an input. In his original paper, Szegedy proposed the high non-linearity hypothesis [1]. The technical details of this argument are outside the scope of the paper but essentially Szegedy believed that machine learning models were not drawing enough connections between training images. Instead, non-linear pockets of what the model expects a certain classification to look like were forming rather than generalizations of what a classification looks like based on the features of an object. Another common argument is the Non-Robust Feature Hypothesis [7]. This hypothesis argues that during training machine learning models are picking out parts of the image that would be meaningless to a human as being essential for the classification. For example, a machine learning system may learn to believe the slight variance in pixel colors in a certain area of input images makes the difference between a dog or cat classification when this is not how a human would determine a classification.

One common thread between many of these theories is the idea that machine learning systems are not being given enough training data [7]. Too little training data leaves image classifiers seeing meaning in small amounts of noise within images when ideally an image classifier trained with large amounts of data would learn that the small amount of noise is irrelevant and instead learn to focus on the bigger picture. Researchers have argued that currently there does not exist any dataset big enough to train a machine learning model robust against adversarial attacks [7].

These proposals as to why adversarial attacks may exist have inspired some researchers' methods for attempting to defend against adversarial attacks.

5 Defense Against Adversarial Attacks

As research into adversarial attacks persists and constantly creates new ways of launching attacks, the need for a way to defend against adversarial attacks becomes More prevalent. The aforementioned research clearly demonstrates that robust physical adversarial attacks are viable against machine learning systems, causing them to misclassify objects. When these attacks are applied on road signs, weapons, or a person by a malicious attacker, the results could be disastrous. If self-driving cars can be prevented from identifying stop signs with an easy-to-launch adversarial attack anyone can pull off, there will be an extreme amount of danger in allowing these cars onto roadways. Therefore, there need to be ways to counter adversarial attacks and to build machine learning systems designed to be robust against adversarial attacks. We will explore a few of the major solutions that have been proposed to counter adversarial attacks.

5.1 Denoising Adversarial Images

One common solution proposed for defending against adversarial attacks is applying a denoising process to the image to attempt to remove adversarial perturbations before passing it to the classifier neural network [7]. Adversarial attacks that directly modify image data [2] do so by adding some amount of noise to images optimized to cause a misclassification. However, if this noise was able to be removed before the image was passed to the neural network, then the machine learning system would be protected against the adversarial attack and would produce a correct classification. Several different implementations for this have been proposed but this fundamental concept remains the same. This concept has been proven to be very successful against some adversarial attacks that produce evenly distributed noise over an entire image. This filtering method reduces the attack success rate of these adversarial attacks to near zero. However, this attack is less successful against adversarial patches. These adversarial attacks do not evenly distribute a small amount of noise over the full image and instead add a large perturbation in one section of the image [4]. It is much harder to filter out the adversarial attack in this case since the attack consists of a full sticker rather than just noise. Therefore, this defense method is not effective against these sorts of attacks.

5.2 Improved Training Set

A second common method for defending against adversarial attacks is attempting to improve the training data set of the victim neural network [7]. The main idea behind this method is to use some images in training that contain adversarial attacks and to attempt to optimize the machine learning model to see through the attacks and output the right classification anyways. Most machine learning image classifiers are trained without any adversarial images in the dataset which may be why they are so susceptible to these attacks. Applying this method to defend against adversarial attacks has mixed results. On one hand, it does reduce the effectiveness of adversarial attacks, especially if the attack being used was also in the training dataset. However, this method tends to reduce the machine learning system's effectiveness overall resulting in a slightly higher baseline misclassification rate [7]. This is not an insignificant problem for many machine learning systems which are reliant on getting the correct classification.

5.3 Issues in Defending Against Adversarial Attacks

While the previously mentioned defenses against adversarial attacks are often successful, there is a fundamental issue in defending against adversarial attacks, a new attack can be created optimized to fool the new, updated system. For example, by considering the noise removal system as being part of the machine learning system, an attacker can simply train a new adversarial image, this time optimized to cause misclassifications even after going through the noise removal process. The same can be done against a system trained to be robust against adversarial attacks. While the effectiveness of the new attacks might be lower, it is very difficult to reduce the effectiveness to an insignificant amount [7]. This creates a constant arms race between attackers, attempting to constantly update the adversarial patch generation process to fool state-of-the-art image classifiers, and the designers of image classification systems attempting to add in new defenses to their systems. Since image classification is likely to continue to grow in importance in the coming decades, this makes adversarial machine learning and defenses a critical area of research that will undoubtedly be important for many years to come.

6 Future Research on Adversarial Machine Learning

Presently, research is ongoing to improve the effectiveness of adversarial attacks and to create new ways to defend against such attacks. Research into adversarial attacks is continuing to search for ways to create disguised attacks that continue to be robust. Research into attacks is also looking into how to improve the robustness of black-box attacks by training adversarial stickers using multiple different image classifiers instead of a single one. In this way researchers aim to create attacks that are generalized to function against a variety of neural networks instead of just one. Other research is going on into how physical photographs can be used to create more realistic training images to improve the robustness of adversarial

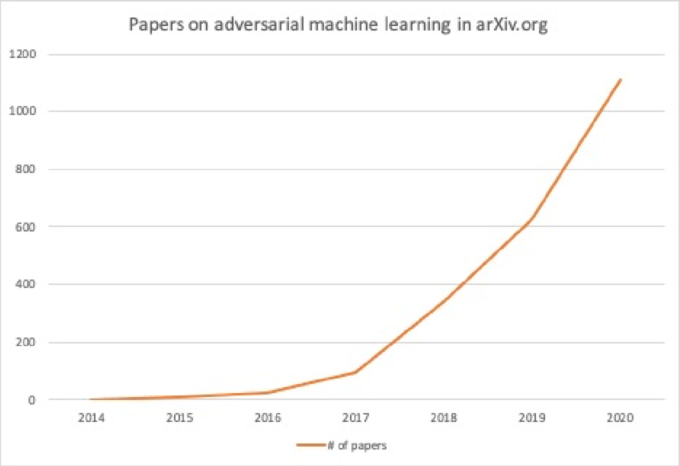

Fig. 7. A graph displaying the growth in adversarial machine learning research papers (Data courtesy of arXiv.org [8])

stickers. Instead of training using renderings of a 3D model [5], this newly proposed method overlays renderings of the adversarial sticker onto a physical photo to reduce the reality gap between training and testing.

To defend against adversarial attacks, new methods are being proposed. Many of these involve applying some transformation to input images to extract important features before passing the transformed image through a classifier. Other approaches involve building machine learning systems specialized to detect attacks and passing an input image through a variety of these classifiers to remove attacked images [7]. Lastly, in the theoretical realm, evidence as to why adversarial attacks exist is constantly being sought out to prove and disprove the present theories on their existence [7].

All of these are potential topics of research for someone interested in researching adversarial machine learning for themselves. There is no shortage of topics to explore in adversarial machine learning and as seen in figure 7, research on adversarial machine learning is continually increasing.

7 Conclusion

Adversarial machine learning is a quickly growing field of research studying how machine learning systems can fail when small adversarial perturbations are added to the input to a machine learning system. Early research began with the creation of the FGSM attack and the direct modification of image data to produce adversarial images. However, the need for more practical attacks introduced the idea of physical adversarial attacks and disguised physical adversarial attacks. A variety of methods for launching these attacks have been proposed and tried including creating adversarial patches to be placed on objects, 3D printing an adversarial copy of the target object, and creating adversarial graffiti. As advancements in adversarial attacks quickly grow, some researchers have begun looking into how to defend against adversarial attacks. Many different potential solutions have been proposed, but the most prominent are denoising to attempt to remove adversarial perturbations before classifying the image and attempting to improve the training set of the machine learning system. While these defenses are effective in some cases, attackers can always create new adversarial attacks optimized to get around these new defenses as well. This creates a cycle of creating new defenses to counter new attacks.

As the importance of machine learning in image classification continues to grow, so do the potential dangers posed by adversarial machine learning. As we increasingly integrate machine learning into our lives, we will need to consider the security risks associated with these systems and their vulnerability to attacks. For these reasons, research into adversarial machine learning attacks is essential to understanding the risks present. But, there also needs to be ways to protect vulnerable systems. Research into defense against adversarial attacks can help to secure machine learning systems. These topics will only grow in importance and they make an excellent research topic for anyone with some background in machine learning.

Acknowledgments

I would like to thank professor Gerald Moulds, my technical writing professor for his informative lectures that helped me in writing this paper. In addition, I would like to thank the anonymous peer reviewer for the valuable feedback provided in editing this final project. Finally, I would like to thank Dr. Tingting Chen and Dr. Hao Ji for helping me to study adversarial machine learning over the summer as part of an REU program.

References

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. ICLR, 2013.

[2] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. ICLR, 2014.

[3] A. Kurakin, I. Goodfellow, and S. Bengio. “Adversarial examples in the physical world.” ICLR 2016.

[4] T. B. Brown, D. Man´e, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” NIPS, 2017.

[5] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[6] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” ICML, 2017.

[7] G. R. Machado, E. Silva and R. R. Goldschmidt, ”Adversarial machine learning in image classification: A survey towards the defender’s perspective”, ACM, 2021.

[8] B. Dickson, “Machine learning adversarial attacks are a ticking time bomb,” TechTalks, 16-Dec-2020. [Online]. Available: https://bdtechtalks.com/2020/12/16/machine-learning-adversarial-attacks-against-machine-learning-time-bomb/. [Accessed: 05-Dec-2021].