**Module 7: Vulnerability of Machine Learning**

**Module Overview**

**Segment 1:**

**1. Statistical Learning vs. Machine Learning:**

**2. ML Breakthrough and Vulnerabilities:**

**2.1 Computer Vision:**

**2.2 Natural Language Processing**

**2.3 Recommendation systems**

**Segment 2: Guided exploration**

**Segment 3: Implementation of an instance of the focus problem**

**Segment 4: Formulate strategy to mitigate instance of focus problem**

**Segment 5: Learner Reflections**

**Overview of Module 7:**

Welcome to Vulnerabilities of Machine Learning (ML) Module! This module’s aim is to introduce some of the potential vulnerabilities that machine learning technology is susceptible to. To begin, we recommend taking a few minutes to explore the entire module. Review the segments through the module (the estimated time for each segment is given). The entire module should take between 9 to 14 hours to complete.

**Segment 1: (Estimated time 30-60 minutes)**

This segment begins by introducing statistical learning, a technique primarily developed by statisticians that involves deriving predictive functions based on data. Statistical learning laid the foundation for modern machine learning and the two fields are quite similar so we will attempt to highlight what distinguishes the two. After that we will discuss some of the more recent innovations in top deep learning applications across industries. Some of the DL domains covered include computer vision, natural language processing, sentiment analysis, fraudulent news detection, and virtual assistants. The key message of this segment is to highlight potential ML weaknesses because, as AI becomes increasingly interwoven into society, it is crucial that we understand how and why these powerful tools can fail, what the implications of these failures are, and how they can be prevented.

The module will show examples of ML/DL algorithms that are statistically impressive but individually unreliable or unexplainable. Some of these ML vulnerabilities can be exploited with severe consequences. For example, the poisoning of self-driving car training data leading to the AV incorrectly identifying a stop sign as a green light. Other examples of malicious A.I. include algorithms that spread fake news, exclude a group of minority people in resource allocation, or breach / disrupt cybersecurity systems. The following subsections present the fundamental introduction of ML and its applications and evidence suggesting that widely used AI-powered applications can reproduce biases and even amplify errors in the data.

**1. Statistical Learning vs. Machine Learning:**

Statistics and machine learning are so closely related that they are almost interchangeable. Statistical learning provided the framework for ML and they both are concerned with finding relationships between variables using raw data and then deriving predictive functions that best represent those relationships. However, the main purpose of ML is to make repeatable predictions without explicitly programmed instructions, while statistical learning tends to rely on rule-based programming. Also, ML is typically concerned more with the performance of the model versus its interpretability. ML models strive to make the most accurate predictions whereas statistical learning models are designed for making inferences about the correlation between data.

Also, ML is most optimal when it learns from many millions (or more) of features compared to statistical learning which typically uses smaller datasets with only a few features. Both models are data dependent, but statistical learning is more about identifying relationships between variables and what the significance of those relationships is, while ML looks to capture hidden patterns in data and use those patterns to make predictions. In this module, the primary focus is on ML techniques and their vulnerabilities.

**2. ML Breakthrough and Vulnerabilities:**

Apart from the difference between these two concepts (statistical learning and machine learning), the focus of this module is on the vulnerabilities of various ML techniques. The increased power of ML/DL technologies has led to AI technologies becoming deeply embedded in modern society. Due to the increased availability of data and compute power, ML methods have been a major driving force behind recent advances in AI technologies. From smart equipment that can recognize hand gestures and voice commands, to increasingly autonomous vehicles, A.I powered by deep learning continues to permeate multiple facets of our everyday lives.

The vulnerabilities of AI/ML promise to introduce threats and challenges that must be addressed as these technologies continue to metastasize through our technological and social infrastructures. Recent mishaps, such as fatal accidents involving autonomous driving vehicles, serve as vivid reminders of the pitfalls of AI. AI can also harbor latent biases or other flaws that can cause harm in diverse and unexpected ways. AI is also particularly vulnerable to malicious attackers because ML-based AI systems depend on data for training, so, if training data is manipulated, there can be significant consequences ranging from diminished performance to fatal accidents. Some well-known advancements in ML/DL technologies and examples of how these emerging technologies can be exploited will be discussed here.

**2.1 Computer Vision:**

Computer vision is a research field within a machine learning sub-domain called deep learning and is the product of multiple different academic disciplines. Computer vision can be the description of scenes and physical objects or computing properties of the 3D world from images. Facial recognition and object detection underlie a fast-growing collection of useful applications derived from this field of study. Facial recognition systems are designed to identify human faces or verify the identity of an individual in images, video, and real time surveillance.

Facial recognition plays a key role in both nations and civil society. The US, Russia, Eastern Asia, Israel, and Europe lead the world in facial recognition technology research. According to the infographic on facial recognition technology[[1]](#footnote-1), 50% of the world actively uses this technology and only Belgium, Luxembourg, and Morocco have banned it to some degree. Major tech companies like Google, Amazon, Facebook, and IBM are responsible for pioneering the most advanced facial recognition techniques.

**2.1.1 Examples of risks and vulnerabilities of computer vision models:**

Although state-of-the-art facial recognition algorithms powered by deep learning perform incredibly well, there are problems with this emerging technology. Facial recognition systems have been known to misclassify people’s images because of biases against ethnic minorities and women. These algorithms also invite a host of ethical questions concerning the threat to privacy born out of coupling this technology with mass-surveillance systems that use public security cameras to harvest data about the behavior of private citizens. Here are a few examples of specific instances of ML-based computer vision systems failing.

1. The popular facial recognition algorithm, COMPAS (stands for Correctional Offender Management Profiling for Alternative Sanctions), that helps with decisions related to things like assigning pre-trial bail and sentencing was criticized by ProPublica[[2]](#footnote-2). According to the analysis, the algorithm would assign black defendants a higher risk score over white defendants despite being otherwise considered equally likely to reoffend. The figure below provides an example of the COMPAS algorithm assigning a white male with a much lower ‘risk score’ versus a black male, despite having a much worse record of prior convictions.

A person with the mouth open

Description automatically generated with low confidenceWebsite

Description automatically generated

1. Research conducted at the MIT Media Lab revealed IBM, Face++, Microsoft, and Megvii all performed significantly worse when detecting the faces of black women. Researchers found that each of the company’s facial recognition products performed better on males than women and performed better on lighter subjects that darker ones. Amongst all of them, IBM had the most significant racial and gender biases with an error rate of 34%. These biases are often the result of inadequate training data. It is common for a facial recognition algorithm to be trained with data that does not offer sufficient representation of all necessary demographics. For example, the algorithm may be trained with a dataset composed of 80% white individuals. Because such an AI has comparatively limited exposure to darker skinned individuals, it struggles with identifying them.

Graphical user interface, application

Description automatically generated

1. Biased data is not the only thing that threatens to disrupt facial recognition systems. Adversarial attacks pose a significant risk to the integrity of such systems and the consequences can be just as severe. One method of instantiating such an attack involves gaining access to the dataset to manipulate the training data to corrupt the model. In one study researching vulnerabilities of facial recognition systems, researchers proposed a method to dupe facial recognition cameras by illuminating the subject using infrared light. Infrared light is not visible to the naked eye but can cause perturbations that disrupt the input data and thus decrease the performance of the facial recognition system. Such an adversarial attack can allow bad actors to bypass facial recognition-based authentication systems or trick surveillance systems designed to protect civilians.
2. The risk posed by corrupted AI in the field of computer vision and autonomous vehicles is one of the more obviously dangerous real-world examples. In a self-driving car, an AI image classification algorithm learns to recognize the road signs (for example, a speed limit sign) from a huge dataset that includes all the various kinds of traffic signs. The algorithm refers to the features it has learned to identify speed limit signs in real time when the AV is driving. If a small-magnitude perturbation is added to the input image, the algorithm may mistakenly identify an incorrect speed limit and instruct the car to speed up in a school zone or slow down on a highway.

Diagram

Description automatically generated

In late March 2019, researchers at the Chinese tech company Tencent figured out a way to trick Tesla’s self-driving software into abruptly switching lanes. The researchers discovered that placing circular, brightly colored circular stickers in the road could fool the computer vision software into believing the car should switch lanes. In the real world, such an attack could cause a car to veer into oncoming traffic or off the road, potentially leading to a serious accident.

**2.2 Natural Language Processing:**

Natural Language Processing (NLP) is used to understand patterns in unstructured language data (e.g., text or voice). Word embedding, the principal task in NLP data preprocessing, involves representing a word with vectors. After words have been encoded into vectors, the analyst can create models for tasks like speech recognition and sentiment analysis. NLP is especially susceptible to bias because it is completely at the mercy of the data used to train it.

A well know mishap involving AI chatbot occurred in 2016 when it was taken down after only a few hours due to offensive tweets. The bot was actuated by trolls’ comments which made it reply with inflammatory tweets. It learned from racist tweets and searched the internet to find data for its responses, an excellent source of prejudice and biased data.

In the domain of spam detection, there is a constant battle between email spammers and spam filter algorithms. Spammers are always learning new ways to manipulate algorithms by inserting non-spam words into the context of emails to bypass the filter. Consequently, spam detectors must be updated constantly to keep up.

Diagram

Description automatically generated with low confidence

**2.3 Recommendation systems:**

Another useful AI tool is the recommendation engine. Recommendation systems are useful for the processing of job applicants because they can automate the process of weeding out unqualified candidates, but this creates a lot of opportunities for bias to emerge. Amazon's resume screener created significant controversy when it was discovered that it did not fairly rate applicants for posted job positions in a gender-neutral way. It would severely penalize resumes that mentioned the word “women” (e.g., Women’s Chess Club Captain). It was eventually discovered that this bias emerged because the algorithm was primarily trained with resumes from the male-dominated tech industries. Apart from racial and gender biases, age was also a discriminating factor. The age-restricted feature for job posting ads made it more difficult for older job seekers even if they had the qualifications.

Content-based recommendation systems use AI content filtering algorithms to suggest topics related to a user’s area of interest, based on the user’s previous feedback. The type of content that is allowed to pass through can have profound consequences. In 2017, a popular video sharing platform aggressively deleted more than 31 million videos predicted to include violent content. However, it was found that educational and legitimate documentary videos were deleted mistakenly, such as those about Syrian war crimes.

**Segment 2: (Estimated time 2-3 hours)**

Guided exploration. The goal of this segment is to provide you with some reliable sources of reported misdeeds committed by cyber actors or hidden biases within the DNA of an AI system.

**Resources:**

1. Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017.
2. Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." *arXiv preprint arXiv,* 2017.
3. Athalye, Anish, et al. "Synthesizing robust adversarial examples." *International conference on machine learning*. PMLR, 2018.
4. M. Weiss, “Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions,” Technology Science, 2019.
5. DeepFaceLab. Available: https://github.com/iperov/DeepFaceLab
6. T. Bolukbasi, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” in Proc. 30th International Conference on Neural Information Processing Systems, 2016.
7. A. Datta, M. C. Tschantz, et al. “Automated experiments on ad privacy settings: a tale of opacity, choice, and discrimination,” Proc. On Privacy Enhancing Technologies, 2015.
8. D. Victor, “Microsoft created a Twitter bot to learn from users. It quickly became a racist jerk,” The New York Times, March 24, 2016. Available: <https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html>
9. Liu, Yanpei, et al. "Delving into transferable adversarial examples and black-box attacks." *arXiv preprint arXiv:1611.02770* (2016).
10. Robust Adversarial Inputs. Available: https://openai.com/blog/robust-adversarial-inputs/

**Select your research topic:**

Instructions in identifying your specific topic of interest.

1. **Read** to become familiar with what is already known about this module. Take some time to think about how you want to get involved by looking into resources and exploring more references. You need to look within the broad subject area and you are not expected to know exactly what your interests will be. You are still discovering and developing your knowledge in this area.
2. **Narrowing** your topic and developing specific research questions to investigate. You may map out your topic interests at first and answer to what problem you are most excited in solving or what topics and required skills relate to this module you want to end up in.
3. Later, you will dig deep into the selected topic thereby gaining knowledge about your specific problem.

**Segment 3: (Estimated time 2-3 hours)**

Tricking Neural Networks: Create your own Adversarial Examples.

The intended purpose of this module’s segment is to give you practical experience with compromised ML tools. You will be tasked with creating a simple instance of an ML tool and then demonstrate ways such a tool is vulnerable. For example, you could create a chatbot that is trained with skewed data so that chatbot will give problematic responses (e.g., a medical chatbot that gives incorrect advice).

So, a good way to go about this would be to find an open-source ML model in a field that interests you and then you manipulate the bot in such a way that it gives incorrect outputs or predictions.

(Try to find a backup plan. (e.g., by providing an instance, just in case learners find that useful))

**Segment 4: (Estimated time 2-3 hours)**

This segment is primarily concerned with mitigation. You are now tasked with researching ways to prevent instances of the compromised ML tool you experimented with in segment 3 and then you are to produce a short write-up or presentation discussing strategies for reducing the vulnerabilities in your chosen system. Depending on the nature of your mitigation strategy, you may want to implement an example of what you’re presenting to give a concrete example to your peers and to give you more practical experience working with the model.

There’s a bit of flexibility with this section. If you would like to spend more time researching and presenting a wide variety of mitigation strategies, that would be ok, or you are free to focus on digging deep into one particular strategy. If you choose to specialize in one strategy, you should implement an example of that strategy to go along with your report.

**Policy solution for ML failures**

Create a problem-solution pattern

1. **Formulate the problem algorithmically which you selected of your own interest. Define the input and expected output** as the goal of the task.
2. **Report potential risk of attacks and the impact of successful attacks on users. Determine how these weaknesses can be used against developed models.**
3. **Add assumptions and restrictions to the problem to make the solution reachable.**
4. Present a sequence of actions (protocol) that can be a basis of compliance used by industries and organizations. Additional recommendations can guarantee the effectiveness of AI systems.
5. Your proposed protocol to mitigate ML vulnerabilities will be assessed in the next segment.

**Segment 5: (Estimated time 2-3 hours)**

Segment 5 is for learner reflections. Here, students will come together and discuss their findings. This could be done in small groups of 3-4 students to encourage more of a free-flowing dialogue between learners, or students could give short (~5 minute) presentations to their classmates. The goal is to create active discussions where students exchange ideas and learn from each other.

1. https://www.visualcapitalist.com/facial-recognition-world-map/?utm\_source=morning\_brew [↑](#footnote-ref-1)
2. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [↑](#footnote-ref-2)