von Huma Rehman, Rudolf Mayer und Andreas Ekelhart

Adversarial Machine Learning

In den letzten Jahren haben maschinelles Lernen (Machine Learning, ML) und insbesondere Deep-Learning-Ansätze enorme Fortschritte erzielt und dabei auch menschliche Leistungen übertroffen, insbesondere in den Bereichen Bildklassifikation, Spracherkennung sowie Computer- und strategische Spiele wie Go.

Aufgrund kontinuierlich verbesserter DNN-Modelle (Deep Neuronal Network), günstiger Hardware und effizienten Softwarebibliotheken wird ML zunehmend in verschiedenste Anwendungen integriert. Dabei kommt ML auch in sicherheitskritischen Bereichen wie autonomes Fahren und medizinische Diagnostik zum Einsatz. Maschinelles Lernen wird somit Teil unseres Alltags, wobei auch automatisierte Entscheidungen aufgrund algorithmischer Vorhersagen getroffen werden. Inkorrekte Vorhersagen können jedoch einen erheblichen Einfluss auf Einzelpersonen und Gruppen haben. Die Genauigkeit der Vorhersagen ist mittlerweile beeindruckend, dennoch machen die Systeme bei scheinbar trivialen Aufgaben oft unerwartete Fehler. Die Robustheit von Algorithmen ist daher ein wesentlicher Faktor, der vor einer Implementierung geprüft werden muss. [1]

Besonders DNNs sind für Angriffe anfällig, bei denen durch minimale, für das menschliche Auge kaum bemerkbare Modifikationen in Bildern das System dazu gebracht wird, eine falsche Vorhersage zu treffen.

ANGRIFFSSZENARIEN

Wir stellen im Folgenden zwei Arten von Angriffen auf ML vor: Poisoning und Evasion. Sie unterscheiden sich vor allem dadurch, wie Angreifer auf das maschinelle Lernsystem zugreifen.

- Bei Evasion-Angriffen versucht die Angreiferin, das System in der Vorhersagephase zu täuschen; dabei manipuliert sie die zu klassifizierenden Bilder, d. h. erstellt sogenannte „Adversarial Inputs“. Bei dieser Angriffsform muss die Angreiferin weder die Trainingsdaten noch die generierten Modelle beeinflussen, sondern nur die Vorhersagen des Modells abfragen.
- Poisoning-Angriffe zielen auf die Trainingsphase des Machine-Learning-Modells ab. Die Angreiferin modifiziert die Trainingsdaten, indem sie im Lernprozess sorgfältig entworfene, konträre („adversarial“) Beispiele hinzufügt. Dabei wird das Modell selbst verändert, um in der Vorhersagephase falsche Ergebnisse zu verursachen. Insbesondere können bestimmte Schlüsselelemente antrainiert werden, die gewünschte Klassifikationen auslösen.

Darüber hinaus wird zwischen gezielten und nicht-gezielten Angriffen unterschieden. Bei gezielten Angriffen versucht ein Angreifer den Klassifikator zu beeinflussen, um eine bestimmte falsche Zielvorhersage zu erzeugen; bei einem nicht gezielten Angriff soll eine beliebige falsche Vorhersage generiert werden. Im Allgemeinen weisen nicht-gezielte Angriffe eine höhere Erfolgsquote auf, bieten allerdings weniger Möglichkeiten für eine fokussierte Ausnützung des Fehlers.

ANWENDUNGSBEISPIELE:

Autonomes Fahren

Technologien für selbstfahrende Autos haben in den letzten Jahren deutliche Fortschritte gemacht. Eine Kombination aus Sensoren, Algorithmen und leistungsstarken Prozessoren sorgt für die sichere Navigation. Kamerabasierte Systeme erfassen dabei die Umgebung und interpretieren die Bilder mittels Objekterkennung. Jüngste Forschungsergebnisse [2] zeigen, dass ein gezielter Evasion-Angriff, der sich auf die Manipulation von z.B. Verkehrszeichen konzentriert, Klassifizierungsfehler verursachen kann. Die Erfolgsrate in Versuchen mit einem fahrenden Fahrzeug lag bei 85 %. So wird z. B. durch die geringfügige Änderung eines Stoppschilds eine – falsche – Geschwindigkeitsbegrenzung von 45 km/h erkannt. Diese einfache Manipulation wurde mit Aufklebern vorgenommen, kann aber zu schwerwiegenden Folgen im autonomen Fahrsystem führen, ohne unmittelbar Verdacht zu erregen (siehe Abb. 1).

Eine mögliche Abwehrstrategie ist, mit mehreren Sensoren weitere Aspekte der

Abb. 1: Wenn das manipulierte Modell ein Eingabemuster mit dem definierten Schlüssel (der weiße Stern) erkennt, wird die falsche Klasse vorhergesagt.

Umgebung aus unterschiedlichen Perspektiven und auf unterschiedliche Art zu erfassen [3].

Biometrische Authentifizierung

ML kann auch in der Fingerabdruck- oder Gesichtserkennung eingesetzt werden (z. B. Zugangskontrolle zu Gebäuden oder mobilen Geräten). Solche Authentifizierungssysteme sind ein attraktives Ziel für Angreifer, insbesondere bei kritischen Strukturen. In [4] zeigen die Autoren, wie in ein Face-Recognition-System Hintertüren (sog. „Backdoors“) eingebaut werden können, um z. B. als Nutzer mit

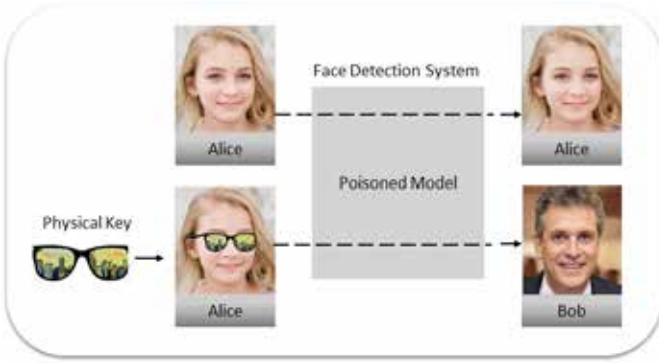


Abb. 2: Brillen lösen eine vorbestimmte Klasse (Person) aus [4]

umfassenderen Zugriffsrechten Zutritt zu erlangen.

Im Training von Authentifizierungssystemen werden in der Regel zuerst Beispielfotos der MitarbeiterInnen mit Annotierungen versehen. Wenn es einem Angreifer gelingt, manipulierter Fotos, die ein bestimmtes Muster (den sog. Schlüssel) enthalten, in das Trainingsset einzuschleusen, lernt das Modell diese Muster mit. Dies ist besonders interessant, wenn es sich bei dem Schlüssel um ein physisches Objekt handelt, z. B. eine Brille, welches die Erkennung einer bestimmten Person auslöst (Abb. 2).

Spam-Erkennung

Eine ähnliche Art von Attacke ist auch in Bereichen abseits der Bildanalyse möglich. Heutzutage basieren viele SPAM Filter auf ML Modellen, d. h. häufige Wörter aus Spam-E-Mails werden gelernt und entsprechende Nachrichten abgefangen. Eine Evasion-Attacke kann sich daher auf die Änderung des Vorhersagewertes konzentrieren, indem sie viele „gute" Wörter hinzufügt. [5] hat gezeigt, dass nur ca. 30 Wörter benötigt werden, um die Filtervorhersage in Sinne der Spammer zu manipulieren. Weitere Detektoren, die nicht nur auf dem Inhalt der Nachrichten basieren, können solche Angriffe wiederum abwehren.

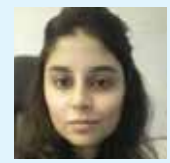
VERTEIDIGUNGSMASSNAHMEN

Insbesondere in Bezug auf die Bildanalyse werden laufend Abwehrmaßnahmen entwickelt und wir unterscheiden drei zentrale Ansätze:

1. Aktives Modifizieren der Trainingsdaten. Im sog. „brute-force adversarial training" wird vom Ersteller des Modells selbst eine große Anzahl „feindlicher" Beispiele generiert, um das Modell darauf zu trainieren, sich von diesen nicht täuschen zu lassen. Dadurch werden einige Varianten eines manipulierten Bildes ausgeschlossen, was es den Angreifern erschwert, veränderte und dennoch dem Original ähnliche Versionen zu generieren.
2. Komprimierung von Daten (Data Compression) zielt darauf ab, die Qualität der Bilder und damit die Effektivität der eingebetteten Muster zu verringern. Dies kann sich allerdings auch auf die Erkennung legitimer Fälle auswirken.
3. Modellreduktion (Model Pruning). Poisoning-Angriffe sind auf Modelle mit freier Lernkapazität angewiesen, um das Backdoor-Muster speichern zu können, ohne die Ergebnisse von legitimen Datenproben zu beeinflussen. Daher wird die Netzwerkgröße reduziert, indem die auf legitimen Eingaben beruhenden Neuronen eliminiert werden; in Folge werden Backdoors deaktiviert. Im Gegenzug kann ein Angreifer die Reserveknoten proaktiv beschneiden, während der Modellbesitzer wiederum durch eine kleine Menge abgestimmter, legitimer Trainingsdaten Backdoors u. U. eliminieren kann. [6]

SOFTWARE-SUPPORT

Es gibt einige Open-Source-Bibliotheken, die gegnerische Input- und Backdoor-Angriffe ermöglichen und für Informationszwecke und zur Verteidigung genutzt werden können. Dazu gehören CleverHans, entwickelt von Google-Forschern, die IBM Adversarial Robustness Toolbox sowie Foolbox (alle auf GitHub).



Huma Rehman ist Entwicklerin bei SBA Research. Ihr Forschungsgebiet ist Artificial Intelligencemit Fokus auf Machine Learning, Deep Learning und Image Processing.

Rudolf Mayer arbeitet bei SBA Research an Digital Preservation-Projekten. Sein besonderes Interesse gilt dem Information Retrieval, Information Visualisation, Text and Music mining und Machine Learning.

Andreas Ekelhart ist Forscher und projektmanager bei SBA Research. Seine Forschungsgebiete sind Semantic Applications, Agent-based Modeling und Simulation und Anwendungskonzepte von IT Security mit Fokus auf Information Security Risk Management.

Referenzen

[1] Ho Bae et. al. Security and Privacy Issues in Deep Learning. 2018. arxiv.org/abs/1807.11655

[2] Kevin Eykholt et. al. Robust Physical-World Attacks on Deep Learning Models. 2017. arxiv.org/abs/1707.08945

[3] Jiajun Lu et. al. NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles. 2017. arxiv.org/abs/1707.03501

[4] Xinyun Chen et. al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. 2017. arxiv.org/abs/1712.05526

[5] Gregory L. Wittel et. al. On Attacking Statistical Spam Filters. CEAS 2004.

[6] Kang Liu et. al. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. 2018. arxiv. org/abs/1805.12185