Data Poisoning – der vergiftete Apfel für KI

Lernprozesse und Modelle für Maschinelles Lernen haben Schwachstellen, die Angreifer ausnutzen können. Ziel der Angriffe ist es, die Aussagen einer KI-Anwendung in eine bestimmte Richtung zu lenken. So kommt es zu gezielten Falschaussagen, die zum Beispiel durch das Einschleusen manipulierter Daten verursacht werden. Diese Methode wird als „Data Poisoning“ bezeichnet. Es beinhaltet eine Reihe von Techniken, um das Verhalten von KI zu beeinflussen.

Adversarial Angriff: wenn Neuronale Netze zu falschen Aussagen kommen

Besonders beeindruckend sind Adversarial Angriffe auf Bilderkennungen mittels Neuronaler Netze. Hier führen Manipulationen von Bilddaten zu falsch anmutenden Ergebnissen in der Erkennung von Bildgegenständen durch künstliche Neuronale Netze. Ein Beispiel: Das Neuronale Netz weist darauf hin, dass das Bild einer Schildkröte ein Gewehr darstellt. Diese fehlerhafte Klassifikation wird durch eine für das menschliche Auge nicht wahrnehmbare Manipulation von Pixelwerten im Bild erzielt, das dabei mit einem Rauschmuster überlagert wird. Während Menschen eine Schildkröte problemlos auf dem „verrauschten“ Bild erkennen, gerät das Neuronale Netz in Schwierigkeiten. Die menschliche Wahrnehmung unterscheidet sich fundamental von der auf mathematischen Regeln basierenden Entscheidungsfindung im Neuronalen Netz. Menschen identifizieren eine Schildkröte über visuell bekannte Mustergruppen wie Kopf oder Füße. Das Neuronale Netz wiederum erkennt Gegenstände für eine Klassifikation eines Bildes über den mathematischen Vergleich einzelner Pixel, deren erlernter Nachbarschaft mit anderen Pixeln und den Farbwerten für Rot, Grün und Blau (RGB).

Das „Rauschen“ entspricht einer signifikanten Veränderung von Eingabewerten (RGB) einzelner Pixel. Auch wenn diese minimale mathematische Abweichungen darstellen, können sie zu einer Fehlentscheidung des einzelnen Neurons im Neuronalen Netz führen. Das Ziel des Angreifers besteht darin, ein Rauschen zu erzeugen, das die einzelnen Neuronen im gestaffelten Entscheidungsprozess mit einer überwiegend hohen Wahrscheinlichkeit in eine Fehlentscheidung kippen lässt. Das Ergebnis ist eine falsche Klassifikation des Bildgegenstandes. Andere bekannte Beispiele führen in Systemen des autonomen Fahrens

zu einer Fehlinterpretation bei der Verkehrszeichenerkennung. Adversarial Angriffe zeichnen sich zudem durch eine große Kreativität der Angreifer aus. Jüngste Beispiele kodieren das Rauschen einer Bildinformation in 3D-Druck Modelle. Das Ergebnis ist ein Objekt, dessen dreidimensionale Form ein Rauschen enthält, die bei der Bilderkennung das Neuronale Netz zu Fehlentscheidungen leitet.

Was ist Data Poisoning?

Die Aussagequalität maschineller Lernmodelle wird wesentlich von den Daten beeinflusst, mit denen sie trainiert oder befragt werden. Werden diese nicht systematisch auf ihre Korrektheit überprüft, können Angreifer absichtlich manipulierte Daten einzuschleusen, um die Aussagen des Modells zu kompromittieren. Data Poisoning kann also auf die vom Modell zu analysierenden Daten oder auf Daten für das Training von KI-Modellen angewendet werden. Potenziell gefährdet sind nahezu alle bekannten KI-Methoden, von Deep Learning in Neuronalen Netzen, bis zum Supervized Learning bei auf statistischer Regression basierenden Methoden. Beim Angriff auf Trainingsdatensätze versuchen Angreifer beispielsweise Auszeichnungen(„Labels“) gezielt zu verändern oder Werte in Datensätzen zu manipulieren. Angreifer können diese Manipulationen verschleiern, indem nicht alle Trainingsdaten verfälscht, sondern veränderte Datensätze in einer statistischen Verteilung in Trainingsdaten eingestreut werden. In Abhängigkeit von der Anzahl der Trainingsdaten und der Verteilung der Manipulation, besteht die Möglichkeit, die Aussagekraft des Modells in eine vom Angreifer gewünschte Richtung zu lenken. Der Angriff kann über die gesamte Datenlieferkette erfolgen. Diese hat in der Praxis oftmals eine große Angriffsoberfläche: Manipulation der Daten an der Datenquelle, Man-in-the-Middle Angriff bei der Datenübertragung oder API-Angriffe kompromittieren im Cloud-Datenspeicher oder im Daten-Versionierungssystem. Geschickte Angreifer ändern dabei Datensätze über eine längere Zeit an. Dabei wird das Delta dieser Veränderungen jeweils minimal gehalten. Dadurch kann der Angriff über Monitoring-Systeme und Filter für statistische Abweichungen nur schwer erkannt werden. Angegriffene laufen Gefahr, viel zu spät festzustellen, dass es ein Problem mit der Zuverlässigkeit der Daten für das KI-Modell gibt und dass Daten manipuliert wurden.

Diese Gefahren drohen durch Data Poisoning

Es gibt eine aktive Forschungsgemeinschaft, die sich weltweit mit dem Thema Data Poisoning beschäftigt. Die demonstrierten Angriffe beziehen sich dabei meist auf Proof-of-Concepts im Rahmen von wissenschaftlichen Studien. Diese demonstrierten Angriffe sind in ihrer methodischen Beschreibung sehr gut dokumentiert und gehen meist einher mit Ansätzen zur Risiko-Minimierung und Abwehr von Data Poisoning. Die wissenschaftliche Arbeit mit Data-Poisoning ist ein wichtiger Bestandteil zur Weiterentwicklung und Verbesserung von KI-Methoden.

Im Jahr 2016 scheiterte ein öffentliches KI-Experiment (https://www.welt.de/kultur/article153688321/Wie-der-Microsoft-Bot-uns-den-Spiegel-vorhaelt.html) von Microsoft durch Data Poisoning. Das Entwicklungsteam des Chat Bots Tay plante die Fähigkeit des Systems durch aktive Kommunikation im Dialog mit Twitter-Followern zu verbessern und so über Unsupervized Learning die Fähigkeiten des Systems einer natürlichen sprachlichen Konversation auszubauen. Tay erlernte seine Kommunikationsfähigkeiten aus den Kommentaren und Nachrichten seiner Follower auf Twitter. Schon kurz nach dem Start des Systems auf Twitter erkannte eine Gruppe von Usern, dass das Verhalten von Tay über die Aussagen in Kommentaren beeinflusst werden kann. Den Ausschlag gab ein Post auf dem Internet-Diskussionsboard 4Chan. Benutzer schlugen vor, Tay mit rassistischen und beleidigenden Kommentaren zur überschütten und somit die Trainingsdaten und Tays Aussagen in eine negative Richtung zu lenken. Das Data Poisoning zeigte schnell Wirkung. 16 Stunden nachdem Tay auf Twitter erschien, hatte der Chatbot über 95.000 Nachrichten mit seinem Data Poisoing Mob ausgetauscht. Jede dieser Nachrichten wurde für das Training des Systems verwendet. Rückblickend schärfte das Experiment den Blick auf Data Poisoning. Das Problem lag im Setting des Unsupervized Learnings über eine offene Twitter-Community. Der Bot fungierte als offenes Einfallstor und damit für das ungefilterte Anlernen des Chatbots über eine öffentliche Social Media-Plattform. Negative Beispiele wie Tay führen dazu, dass der Aufbau von Trainingssystemen mit öffentlichen Datenschnittstellen sorgfältiger geplant wird. Mittels Filter und Monitoring wird Maschinelles Lernen gegen Data Poisoning durch einen organisierten Internet-Mob geschützt.

Schutz vor Data Poisoning

Blindes Vertrauen in Daten ist das Einfallstor für Data Poisoning. Zudem kann jedes KI-Modell als „Elternmodell“ für neue dienen. Das heißt, dass ein unbemerkter Angriff auf Lerndaten hierbei weitergegeben wird. Wird das Lernmodell übertragen, werden auch die „vergifteten“ Daten einbezogen. Daher ist es wichtig Daten für diese Lernmodelle schützen. Weltweit gibt es zahlreiche Arbeitsansätze aus Erfahrungen mit ML-Sicherheitsangriffen zu lernen und wirksame Methoden zur Abwehr zu entwickeln. Eine davon ist die Zusammenarbeit der Adversarial ML Threat Matrix, die eine Adversarial Threat Landscape (https://github.com/mitre/advMLthreatmatrix) for Artificial-Intelligence-Systems veröffentlicht hat. Sie baut auf dem etablierten MITRE Att&CK Framework (https://attack.mitre.org/) auf, der weltweit zugänglichen Wissensbasis über Taktiken und Techniken derartiger Angriffe. Es gibt aber auch systemische Grenzen für Angreifer: Zur Einschleusung vergifteter Daten ist es notwendig, dass Systeme regelmäßig re-trainiert werden. Nur wenn die Trainingsdaten aus Quellen stammen, auf die der Angreifer Zugriff hat, kann das Training vergiftet werden und der Angreifer auf das KI-Modell Einfluss nehmen.

Fazit

Es hat sich in der Vergangenheit als sehr schwierig herausgestellt, Data Poisoning Angriffe zu erkennen und sich zuverlässig dagegen zu wehren. Angreifer können sogar mehrere parallel angewendete Verteidigungen wirksam umgehen. Eine der vielversprechendsten Abwehrmaßnahmen gegen gegnerische Angriffe ist das Training mit KI, um die Manipulation zu verhindern. In der Trainingsphase werden Beispiele für Adversarial Attacks integriert, um die Robustheit des Systems zu erhöhen. Sind diese jedoch sehr umfangreich und komplex, verzögert das die Trainingszeit des Modells. Werden aus Performancegründen nur schwache Angriffe als Beispiel integriert, bleibt das System anfälliger für starke, wirksame Angriffe. Die Gefahr solcher Abwehrtechniken besteht vor allem darin, dass sie ein falsches Gefühl der Sicherheit vermitteln

Zurzeit müssen Neuronale Netze noch in der Tiefe betrachtet und bei Auffälligkeiten Proben analysiert werden. Menschliches Expertenwissen ist dabei eines der wesentlichen Kriterien für die sichere Abwehr von Manipulationen auf KI-Trainingsdaten.

Darüber hinaus gibt es Bestrebungen, in Deutschland Standards für Prüfverfahren zu entwickeln. Hierzu wurde bereits eine Normungsroadmap KI (https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/fahrplan-festlegen) vorgestellt. Künftig ist es wichtiger denn je in der Lage zu sein, allgemeingültige Kriterien und Instrumentarien zu definieren, um KI-Systeme ausreichend überprüfbar und sicher zu machen.

Bildunterschriften

Abbildung 1
Schema einer Angriffsoberfläche eines ML Systems.
Quelle: asvin GmbH

Abbildung 2
Microsofts Chatbot Tay auf Twitter.
Quelle: Twitter International Company

Abbildung 3
Die richtige statistische Auswahl der Trainingsdatensätze schützt vor Data Poisoning. In diesem Beispiel müssen fast 25% der Datensätze manipuliert werden, um die Fehlerrate des trainierten Modells deutlich zu erhöhen.
Quelle: asvin GmbH

Über asvin

asvin bietet Lösungen, um die Sicherheit und Herkunft von Software über ihren gesamten Lebenszyklus zu gewährleisten. Dies umfasst Services und Analysen, die Daten- und Software Lieferketten überwachen, sichere Roll-out von Over-the-Air Software Updates unterstützen sowie zur Erstellung von Software Bill of Materials (SBOM).

Mehr Informationen unter www.asvin.io, auf Twitter und LinkedIn.

Über den Autor:

Mirko Ross, Jahrgang 1972, ist ein international anerkannter Aktivist, Experte, Redner, Publizist und Forscher im Bereich Cybersicherheit und Internet der Dinge. Bereits im Alter von 14 Jahren begann er Sicherheitslücken in IT-Systeme zu untersuchen. Anstatt einer Hacker-Karriere, entschied er sich für die „gute Seite“ der Cybersicherheitsindustrie. Ross war bis 2020 Mitglied in der Experten Gruppe für Sicherheit im Internet der Dinge der europäischen Cybersicherheitsbehörde ENISA und berät als Experte die EU-Kommission. Er ist zudem aktiv in internationalen Gremien und Forschungsprojekten im Bereich Cybersicherheit- und Blockchain-Technologien. Mirko ist bis heute der positiven Hacker- und Maker-Bewegung eng verbunden und fördert gemeinnützige Projekte im Bereich Open Data und IT-Bildung. 2018 gründete er mit asvin.io ein Unternehmen, mit dem Ziel, die Cybersicherheit im Internet der Dinge zu erhöhen und dafür Software-Lösungen bereitzustellen. asvin wurde 2020 auf der it-sa als Bestes Cybersicherheits-Startup im deutschsprachigen Raum ausgezeichnet. Mirko lebt auf dem Land in der Natur und arbeitet in Stuttgart.

Pressekontakt

BCW GmbH
asvin@bcw-global.com

Mirko Ross, asvin CEO
m.ross@asvin.io