Das Fraunhofer-Institut für Naturwissenschaftlich-Technische Trendanalysen berichtet über neue Technologien

Adversarial Machine Learning

Adversarial Machine Learning (AML) beschäftigt sich mit dem Auffinden von potenziellen Sicherheitslücken in Verfahren des maschinellen Lernens (ML) und auch mit der Entwicklung von geeigneten Gegenmaßnahmen diesbezüglich. Da Methoden des ML und hier insbesondere das sogenannte Deep Learning wesentlich für die aktuellen Erfolge im Bereich der Künstlichen Intelligenz (KI) verantwortlich sind, stellen die gegenwärtigen Entwicklungen auf dem Gebiet des AML potenziell eine erhebliche Bedrohung für viele aktuelle KI-Systeme dar. Das erklärt die zunehmenden diesbezüglichen Aktivitäten (auch) auf der Seite der ML-Entwickler, um das Wissen über potenzielle Schädigungsmechanismen, wenn möglich, zum Schutz der Systeme zu nutzen.

Generell umfasst ML Verfahren, die es Computern erlauben, selbstständig anhand von Beispieldaten bestimmte Sachverhalte zu erlernen, z. B. ob auf einem Bild ein gewisses Objekt abgebildet ist. Im Rahmen von AML werden dabei Eingabedaten untersucht, sogenannte Adversarial Examples, die von einem Angreifer speziell mit der Absicht entworfen worden sind, Fehler bei dem entsprechenden ML-Verfahren hervorzurufen. So können manche ML-Verfahren durch subtile Veränderungen der Bildpunkte eines Bildes dazu veranlasst werden, anstatt des eigentlich dargestellten Objekts ein ganz anderes Objekt zu erkennen, obwohl das veränderte Bild für einen Menschen vom ursprünglichen Bild nicht zu unterscheiden ist.

Beispielsweise wären derartige Angriffe eine mögliche Gefahr für autonome Fahrzeuge, wenn dadurch z. B. ein Stoppschild als Geschwindigkeitsbegrenzungsschild erkannt und daraufhin von dem Fahrzeug ein Unfall verursacht werden würde. Ähnliche Überlegungen gelten auch für andere Systeme aus dem Bereich der Robotik. Eine weitere Möglichkeit besteht in der Umgehung von Sicherheitskontrollen. So könnte eine Person auf der Basis von AML u. a. versuchen, ein automatisiertes System zur Gesichtserkennung zu täuschen, um dadurch unerkannt zu bleiben. Im Bereich der IT-Sicherheit könnte AML beispielsweise dazu eingesetzt werden, Systeme zur Detektion von Schadsoftware oder Spam-E-Mails zu täuschen.

Außerdem könnten Angriffe auf sprachbasierte digitale Assistenten, die inzwischen z. B. häufig in Lautsprechern zur Wiedergabe von Musik im Heimbereich integriert sind, erfolgen. Hier könnten u. a. harmlos erscheinende Tonaufnahmen, beispielsweise in Form von Fernsehwerbung, genutzt werden, um unautorisierte Käufe zu veranlassen. Darüber hinaus könnte die automatisierte Erkennung von betrügerischen Transaktionen im Finanzwesen oder im E-Commerce-Bereich mithilfe von AML getäuscht werden. Hierzu zählt u. a. Betrug im Rahmen der Vergabe von Darlehen. Weitere Angriffe dieser Art sind z. B. auf ML-Verfahren zur Erkennung von illegal verbreiteten, kopiergeschützten Inhalten oder anstößigen Inhalten in sozialen Medien denkbar.

Als Grundlage für ML dient generell ein mathematisches Modell mit einstellbaren Parametern, wobei die jeweiligen Parameter in der Lernphase im Hinblick auf die Beispieldaten optimiert werden. Dabei können unterschiedliche ML-Modelle zum Einsatz kommen. Die unterschiedlichen Arten von Angriffen im Rahmen von AML können u. a. anhand des Wissens kategorisiert werden, das der Angreifer über das jeweils anvisierte ML-Modell besitzt. Bei White-Box Attacks kennt der Angreifer beispielsweise sämtliche Details des ML-Modells, z. B. welches konkrete Modell verwendet wird und welche Beispieldaten für die Lernphase genutzt wurden. Im Gegensatz dazu sind bei Black-Box Attacks diese Details dem Angreifer nicht bekannt. Hier ist der Angreifer lediglich in der Lage, die Ausgabedaten des betreffenden ML-Modells zu von ihm gewählten Eingabedaten abzufragen, allerdings möglicherweise nur in einem begrenzten Umfang. Eine mögliche Angriffsstrategie besteht in diesem Fall darin, die durch den Angreifer so gewonnenen Ein- und Ausgabedaten zu benutzen, um ein anderes ML-Modell damit zu trainieren. Dabei kann es sich auch um ganz unterschiedliche ML-Modelle handeln. Wenn beide ML-Modelle in einer ähnlichen Weise funktionieren, dann werden Adversarial Examples, die für das eine Modell entworfen wurden, wahrscheinlich auch bei dem anderen Modell entsprechende Fehler hervorrufen. Bei Gray-Box Attacks liegt das Wissen des Angreifers zwischen dem bei White-Box Attacks und Black-Box Attacks. Typischerweise ist in diesem Fall nur das genutzte ML-Modell bekannt, aber nicht die genauen Parameter des Modells und die Beispieldaten für die Lernphase.

Angriffe auf ein ML-Verfahren können nicht nur auf ein bereits trainiertes System erfolgen, sondern auch während dessen Lernphase. Solche sogenannten Poisoning Attacks verwenden geeignet modifizierte Beispieldaten, um beispielsweise eine fehlerhafte Klassifikationen eines Objektes hervorzurufen.

Bisher konnten bereits einige effektive Angriffe auf unterschiedliche Arten von maschinellen Lernverfahren demonstriert werden. Dagegen konnten viele bislang in dieser Hinsicht vorgeschlagenen Gegenmaßnahmen schon verhältnismäßig schnell nach ihrer Veröffentlichung durch neue Angriffsstrategien wieder gebrochen werden. Viele gegenwärtige Forschungsarbeiten im Bereich AML konzentrieren sich auf Adversarial Examples im Zusammenhang mit der Klassifikation von Bildern auf der Grundlage von Deep Learning. Zukünftig ist zu erwarten, dass auch andere Anwendungsbereiche noch eingehender untersucht werden, wie z. B. das Gebiet der IT-Sicherheit.

Möglicherweise zeigt die Angreifbarkeit aktueller ML-Verfahren durch Adversarial Examples, wie weit derzeitige KI-Systeme trotz ihrer teilweise beeindruckenden Erfolge noch von wirklicher Intelligenz entfernt sind, weil ihnen noch ein grundlegendes, umfassenderes Verständnis der Welt fehlt. Dementsprechend ist zu erwarten, dass KI-Systeme, die über ein verbessertes Verständnis dieser Art verfügen, auch eine höhere Robustheit gegenüber Adversarial Examples aufweisen werden.

Dr. Klaus Ruhlig