LEDGER
ledgerjournal.org

SSN 2379-5980 (online)
Björns Abgaben 2021

KI-Portfolio Abgabe

Adversarial Attacks und wo sie zu finden sind

Björn Bulkens*

Abstract. Die folgenden zwei Seiten beschäftigen sich mit dem Feld der Adversarial Attacks. Es soll auf die zugrundeliegende Theorie und seine Anwendungsmöglichkeiten eingegangen werden.

KEY WORDS

1. Machine Learning 2. Adversarial Attack

1. Einleitung

Die Industrien in unterschiedlichsten Branchen bringen bereits Machine Learning in ihre Systeme ein, um Prozesse zu automatisieren oder zu optimieren. Auf hochleistungsfähigen Hard- und Softwareplattformen bieten die maschinellen Lernverfahren der Künstlichen Intelligenz das Instrumentarium, um aus großen Datenmengen komplexe Zusammenhänge zu lernen[1] . Eine zentrale Komponente dabei stellen neuronale Netze und Deep Learning dar. Hierbei soll in diesem Artikel auf ein spezielles Problem eingegangen werden:
Leistungssteigerung hat einen deutlich höheren Stellenwert als die Erhöhung der Sicherheit dieser Modelle [2] . Jedoch gibt es mittlerweile Wege und Mittel genutzte Modelle zu einem systematischen Versagen zu bringen bzw. gezielt zu eigenen Gunsten beeinflussen zu können, wenn dagegen keine adäquaten Maßnahmen getroffen werden. Die Rede ist von Adversarial Attacks, wie sich vielleicht bereits aus dem Titel des Artikels ableiten ließ.

2. Wie entsteht ein Adversarial Example

Um über Adversarial Attacks sprechen zu können sollte zu aller erst definiert sein, was eine Adversarial Attack ist. Der Begriff Adversarial Attack fasst eine Bandbreite von Methoden zur Erstellung von Adversarial Examples zusammen. Ein Adversarial Example wiederum bezeichnet einen synthetisch erzeugten Datenpunkt, der das systematische Versagen eines Modells hervorruft. Diese Definition ist abzuleiten aus unterschiedlichen , deutlich detaillierten Papern, die sich mit dieser Thematik auseinander setzten, wie beispielsweise *Szegedy et al., 2015 "Intriguing properties of neural networks"* und *Goodfellow et al., 2015 "Explaining and harnessing adversarial examples"*.

Der größte Teil der Forschung an Adversarial Attacks findet im Bereich der Computer Vision statt und wird hier dementsprechend an einem solchen Beispiel beschrieben werden. In Abb. 1 ist links der Input x an das Modell (in diesem Fall GooLeNet trainiert auf dem ImageNet-Datensatz),

*Matrikelnummer 9936663

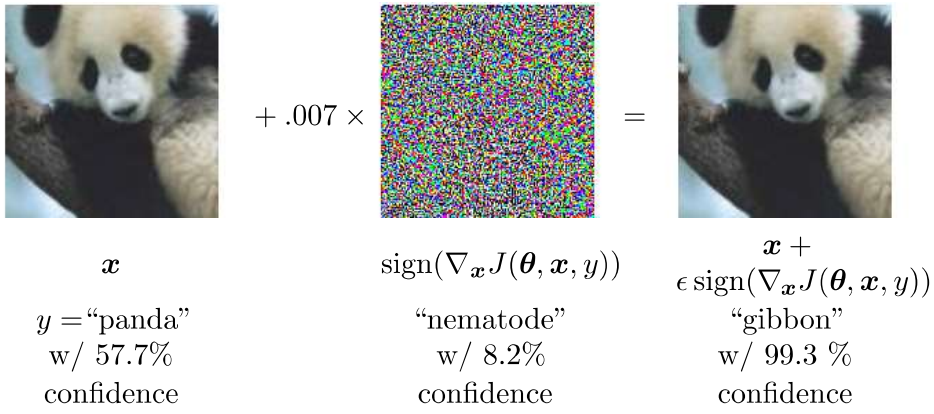

Fig. 1. Bild, samt Beschreibung aus *Goodfellow et al.*[3]

das Bild eines Pandas, das auch von dem Modell als solches erkannt wird (y="panda"), zusehen. Nun wird **x** mit dem **scheinbar** zufälligen Rauschen um einem Faktor von 0.007 verrechnet. Es entsteht ein Bild das für das menschliche Auge fasst nicht von dem Original **x** zu unterscheiden ist. Für das Modell reicht dieser Unterschied jedoch das entstandene *Adversarial Example* mit 99% Sicherheit als einen Affen zu klassifizieren. Dies kommt dadurch, dass das scheinbar willkürliche Bildrauschen sorgfältig gewählt wurde, um das Modell zu beeinflussen. Dabei ist in diesem Artikel leider nicht genug Platz die mathematische Grundlage für die hier verwendete Methode einzugehen, bei näherem Interesse ist das Stichwort hierzu "Fast-Gradient-Sign-Method". Dies ist von vielen sogenannten Evasion-Attacks.
Es ist dabei zu vermerken, dass diese "Verschiebung" nicht unbedingt speziell für diesen Datenpunkt gilt, sondern meist mehrere Bilder eine Klasse systematisch beeinflussen kann. Dies lies Goodfellow et al. eine Hypothese über die Zusammensetzung von Entscheidungsgrenzen eines Machine-/Deep- Learning Modells aufstellen, die im n-dimensionalen Raum nur schwer nachzuvollziehen sind: Die Entscheidungsgrenzen eines Modells bleiben so linear wie möglich und nur so komplex wie nötig. In der folgenden Grafik soll deshalb einmal dargestellt werden, welches Problem dies hervorruft:

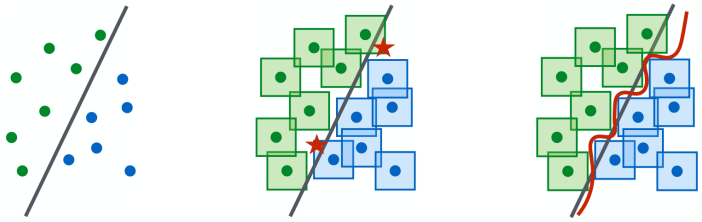

Fig. 2. Darstellung des idealen Prozess während eines Adversarial Training aus Madry et al. (2017)[4]

Im linken Bild stellen die Punkte unsere Traingsbeispiel und die Linie den binären Klassifikator dar. Der Klassifikator funktioniert dabei bereits gut, auch wenn er linear und sehr nah an den Trainingspunkten ist. Somit stellen die markierten Stellen beispielhaft Punkte dar an denen es besonders einfach ist Adversarial Examples zu erzeugen. Um das zu verhindern müssen die Entscheidungsgrenzen möglichst weit von allen bisher gesehenen Beispielen weg gezogen

werden (rechtes Bild), und idealerweise eine graduelle Erhöhung der Konfidenz in Relation zur Nähe zur Enscheidungsgrenze sichergestellt werden. Ein möglicher Ansatz hierfür ist das sogenannte Adversarial Training, bei dem erst Adversarial Examples auf Basis des Modells erzeugt werden und diese dann wieder ins Training des Modells eingespeist werden. Aber auch andere Methoden wie beispielsweise ein zweites Modell, das Inputs auf mögliche Adversarial Examples prüft, finden Einsatz in der Praxis.

3. Probleme mit Adversarial Attacks

Warum ist es wichtig sich mit Adversarial Attacks zu beschäftigen? Abgesehen von dem hilfreichen Effekt dadurch Entscheidungsgrenzen von Modellen besser verstehen lernen zu können, bergen Adversarial Attacks eine Gefahr. Es wird zum einen darauf vertraut, dass Modelle mittlerweile in der Regel die richtige Entscheidungen treffen, wenn sie produktiv gesetzt werden und menschliches Eingreifen lediglich unterstützend eingesetzt werden muss. Somit stellt ein systematisches Versagen ein Problem für die Integrität der Systeme dar. Ebenso können Adversarial Attacks auch viele andere Anwendungen finden als das "bloße Falschklassifizieren": Auf Basis dieser Methoden hat man Mittel und Wege gefunden komplexe Modelle wie die Gesichtserkennung von Microsoft Azure beinahe perfekt (performancetechnisch) zu kopieren[5] . Andere Gefahren sind beispielsweise das Rekreieren von Trainingsdaten. Diese Daten waren unter Umständen geschützt oder es wurde viel Geld dafür bezahlt sie zu erhalten und sie so an dritte weiterzugeben war nicht erwünscht. Diese beiden Beispiele ordnet man in die Reihe der sogenannten Exploratory Attacks ein, bei denen das Ziel nicht ist das Modell falsche Klassifikationen machen zu lassen, sondern so viele Informationen wie möglich aus der Verwendung des Modells extrahieren zu können. Auch hier zeigt sich also ein Anreiz sein Modell vor Adversarial Attacks schützen zu wollen, denn kein Unternehmen möchte seine Modelle oder Daten preisgeben, die ihnen Wettbewerbsvorteil verschaffen.

4. Anregung zu guter Literatur zum Thema

Weil die Länge dieses Artikels viel zu kurz, um dieses große und komplexe Thema mehr als zu Streifen, möchte ich dem Leser der mehr über dieses Thema erfahren möchte folgende Paper ans Herz legen:

- Das "initiale" Paper zu Adversarial Attacks: Szegedy et al., 2015 und das darauf folgende Paper von Goodfellow et al., 2015 [3]
- Ein zusammenfassendes Survey, um einen ausführliche Übersicht über Methoden für und Arten von Adversarial Attacks zu erhalten von Akhtar et al., 2018[6]
- Eine Vorlesung zu Adversarial Attacks gehalten von Goodfellow an der Universität von Stanford von 2017 (verfügbar auf YouTube: `https://youtu.be/CIfsB_EYsVI?t=215`)
- Das oben bereits erwähnte Paper zu Copycat-CNN von Correia-Silva et al., 2018 [5]

SSN 2379-5980 (online)
Björns Abgaben 2021

Notes and References

[1] Hecker, Dirk and Döbel, Inga and Rüping, Stefan and Schmitz, Velina "Künstliche Intelligenz und die Potenziale des maschinellen Lernens für die Industrie" *Wirtschaftsinformatik & Management* **5** 26–35 (2017) doi:10.1007/s35764-017-0110-6

[2] Barreno, Marco and Nelson, Blaine and Sears, Russell and Joseph, Anthony D. and Tygar, J. D. "Can Machine Learning Be Secure?" *Association for Computing Machinery* S.16–25 (2006) 10.1145/1128817.1128824

[3] Goodfellow, Ian J. and Shlens, Jonathon and Szegedy, Christian "Explaining and harnessing adversarial examples" *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015) 1412.6572

[4] Mądry, Aleksander and Makelov, Aleksandar and Schmidt, Ludwig and Tsipras, Dimitris and Vladu, Adrian "Towards deep learning models resistant to adversarial attacks" `https://github.com/MadryLab/mnist_challenge` (2017) 1706.06083

[5] Correia-Silva, J. R. et al. (2018) 'Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data', in Proceedings of the International Joint Conference on Neural Networks. Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/IJCNN.2018.8489592.

[6] Akhtar, Naveed and Mian, Ajmal "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey" `http://arxiv.org/abs/1801.00553` (2018) 10.1109/ACCESS.2018.2807385