Inhalt

[1 Einleitung 2](#_Toc52300016)

[a Problemstellung 2](#_Toc52300017)

[b Zielsetzung 3](#_Toc52300018)

[2 Neuronale Netze und Stand der Technik 4](#_Toc52300019)

[a Image Classification und Computer Vision 4](#_Toc52300020)

[b Verkehrsschild-Erkennung 6](#_Toc52300021)

[c Adversarial Attacks 7](#_Toc52300022)

[i Historie 7](#_Toc52300023)

[ii Targeted und Untargeted 8](#_Toc52300024)

[iii Black-, Grey- und White Box 8](#_Toc52300025)

[iv Attack-Arten 8](#_Toc52300026)

[3 Attacken Methode 11](#_Toc52300027)

[a Fast Gradient Sign Method 11](#_Toc52300028)

[b Iterationen 11](#_Toc52300029)

[4 Validierung 11](#_Toc52300030)

[a Attackiertes Modell 11](#_Toc52300031)

[i Netzarchitektur 11](#_Toc52300032)

[ii Datenset 11](#_Toc52300033)

[iii Training 11](#_Toc52300034)

[b Attacke 11](#_Toc52300035)

[i Bildateinen 11](#_Toc52300036)

[ii. Erfolgreiche lokale Attacken 11](#_Toc52300037)

[iii. Erfolgreiche globale Attacken 11](#_Toc52300038)

[5 Fazit und Ausbilck 11](#_Toc52300039)

[a Fazit 11](#_Toc52300040)

[b Ausbilck 11](#_Toc52300041)

[6 Verzeichnis 12](#_Toc52300042)

[a Literaturverzeichnis 12](#_Toc52300043)

[b Abbildungsverzeichnis 12](#_Toc52300044)

[c Formelverzeichnis 12](#_Toc52300045)

1 Einleitung

a Problemstellung

Spätestens seit dem „Big Bang des Deep Learning“ im Jahre 2009 sind Begriffe wie künstliche Intelligenz, Maschinelles Lernen und neuronale Netze ständige Begleiter nicht nur in der Informatik, sondern auch in unserem alltäglichen Leben. In dem Jahr konnte Nvidia mit einer neuen Generation von Grafikarten die Geschwindigkeit von Deep Learning Systemen verhundertfachen; Der Startschuss für eine Revolution der neuronalen Netze war gegeben.

Seitdem beschleunigt sich die Entwicklung solcher „selbst lernenden“ und „intelligenten“ Systeme immer weiter und in den letzten zehn Jahren schlugen Computer eine Vielzahl an Menschen in einer großen Varietät von Aufgaben. Vor allem das Meistern komplexer Brettspiele wie Schach, Shogi und Go waren große Meilensteine in der jüngeren Geschichte der künstlichen Intelligenz.

Während die meisten Anwendungsfälle für künstliche Intelligenzen und speziell neuronale Netze nur in der Forschung auftreten, hat es die Bilderkennung mittels neuronaler Netze in unseren Alltag geschafft. Google und Apple sortieren Bilder auf unseren Smartphones, Stock Footage wird automatisch klassifiziert und Facebook kann auf Gruppenbildern einzelne Personen identifizieren und zuordnen; All dies geschieht mit gut trainierten neuronalen Netzen.

Während dies Beispiele für Software sind, bei denen ein Fehler keine sonderlichen negativen Folgen hat, wird nun seit circa drei Jahren Bilderkennung auch im Bereich der Sicherheit eingesetzt. Auf neuronalen Netzten basierende Gesichtserkennung soll in Zukunft vermehrt an öffentlichen Plätzen eingesetzt werden und auch autonom fahrende Autos besitzen Bilderkennungssoftware zur Umgebungsanalyse. Fehler und falsche Klassifizierungen können in solchen Fällen fatale Folgen hervorrufen.

b Zielsetzung

Wirft man einen Blick auf den „Gartner Hype Cycle for Emerging Technologies 2019“ (Abb. 1) ist „Artificial Intelligence Platform as a Service“ (AI PaaS) einer der erwartungsvollsten Themen der nächsten Jahre. AI PaaS beschreibt Cloud basierte Lösungen für eine Fülle an KI-Anwendungen, vor allem im Bereich der computer vision, image classification und des natural language processing. Auf solche hohen Erwartungen an eine Technologie wie sie gerade an die künstliche Intelligenz gestellt wird, folgt unweigerlich eine Ernüchterung. Je stärker diese Ernüchterung ausfällt, desto länger wird es dauern bis künstliche Intelligenzen produktiv flächendeckend eingesetzt werden kann. Um einer zu starken Ernüchterung entgegen zu wirken, ist es wichtig frühzeitig auf Fehl- und Manipulierbarkeit von neuronalen Netzten hinzuweisen. Dies hilft beim Einschätzen von Risiken und beschleunigt das Herausarbeiten wichtiger und sinnvoller Anwendungsgebiete und führt somit zu einem schnelleren Erreichen des Produktivitätsstatus.

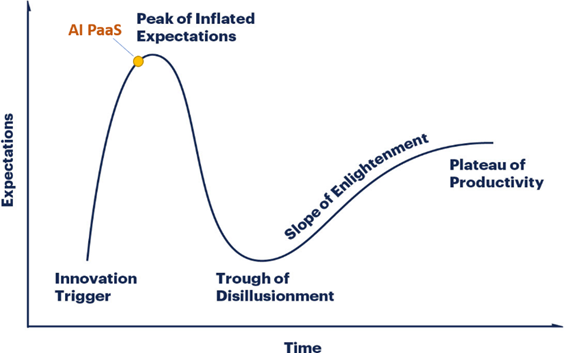


Abbildung Gartner Hype Cycle for Emerging Technologies, 2019

Adversarial Attacks (deutsch: adversative Angriffe) sind Methoden, um neuronale Netze zu täuschen und absichtlich falsche Outputs zu provozieren. Diese Methoden werden hauptsächlich im Bereich der Bilderkennung erforscht. Ziel der Forschung um Adversarial Attacks ist es, die Manipulierbarkeit von neuronalen Netzten zu erkennen und Fehler im System aufzuzeigen. Ein solcher Adversarial Attack auf ein Sicherheitssystem könnte gravierende Folgen hervorrufen.

Aufgrund der wachsenden Anzahl und der steigenden Bedeutung von Bilderkennungssoftware werden Adversarial Attacks zunehmend als Bedrohung wahrgenommen und als einer der gefährlichsten Ansätze, um künstliche Intelligenz zu hacken.

Dieses Projekt soll Adversarial Attacks erklären, anhand eines Praxisbeispiels veranschaulichen und die damit einhergehenden Gefahren erläutern. Dies ist vor allem für die Bilderkennung entscheidend, da sie zukünftig zunehmend auch im Sicherheitsbereich verwendet werden soll.

2 Neuronale Netze und Stand der Technik

a Image Classification und Computer Vision

Parallel zu den neuen technischen Möglichkeiten im Jahr 2009 veröffentlichte die Stanford University im selben Jahr ImageNet. ImageNet ist eine Datenbank für visuelle Daten und derer dazugehörigen Kategorisierung. Sie beinhaltet mehr als 14 Million Bilder und über 20.000 Kategorien und ermöglichte so erstmals das Training von Deep Learning Systemen zur Bilderkennung. Seitdem ist die Bilderkennung die größte und eine der wichtigsten Anwendungsfälle von künstlicher Intelligenz. Für viele ist es der Einstieg in das Thema der neuronalen Netze (Abb. 2) und für viele Modelle der Bilderkennung ist eine Fehlerquote von unter 0,5 % keine Seltenheit mehr. Deswegen sind neuronale Netze zur Bilderkennung einer der wenigen künstlichen Intelligenzen, die es aus der Forschung heraus- und in unseren Alltag hereingeschafft haben.

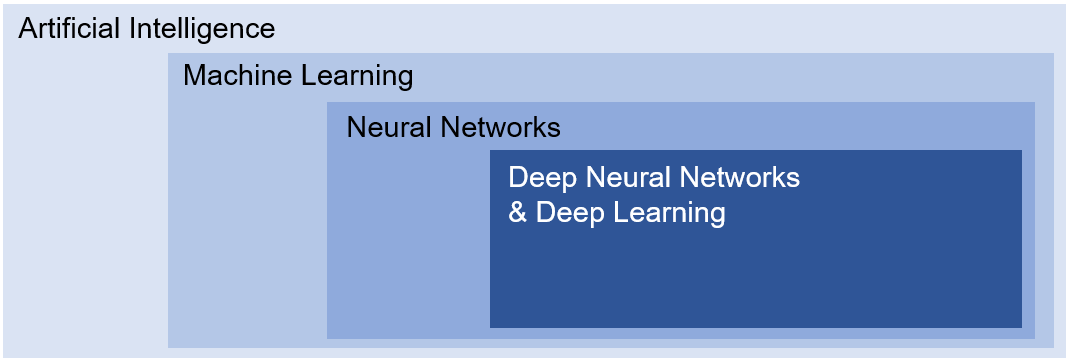


Abbildung Deep Neural Networks im Kontext

Moderne Bilderkennungsmodelle werden fast ausschließlich mit Deep Neural-Networks programmiert. Solche Modelle können dutzende an Neuronenebenen mit millionen von trainierbaren Parametern haben, dies führt zu der namensgebenden Tiefe in den Deep-Learning-Systemen. Das Prinzip nachdem Deep-Learning Systeme trainiert werden und lernen lässt sich, wie bei jedem Machine-Learning-Modell; in drei Hauptkategorien unterscheiden; Dem supervised learning (deutsch: überwachtes Lernen), unsupervised learning (deutsch: unüberwachtes Lernen) und reinforcemne learning (deutsch bestärktes Lernen) (Abb. 3). Mit Lernen ist dabei die Fähigkeit einer künstlichen Intelligenz gemeint, Gesetzmäßigkeiten nachzubilden.

Das supervised learning setzt im Gegensatz zum unsupervised- oder reinforcemnet learning, gekennzeichnete (labelled) Daten voraus. Innerhalb des supervised Learning exsitieren zwei große Problemstellungen: Regressionsprobleme und Klasssifikationsprobleme. Regressionsprobleme beschäftigen sich mit der Vorhersage kontinuierlichen Wertes (Aktienkurs, Häuserpreise, usw.), wo hingegen Klassifikationsprobleme sich mit der Zuordnung eines Inputs zu einer vordefinierten Klasse beschäftigen.

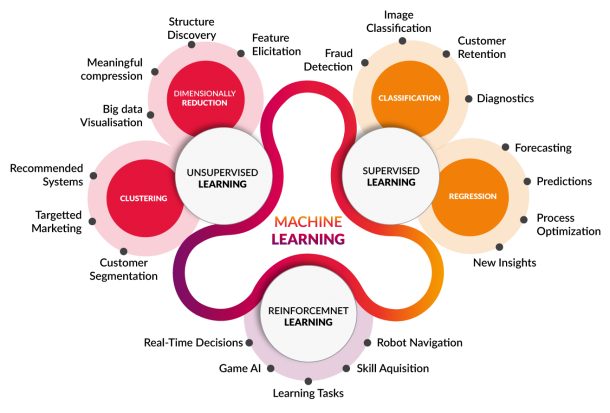


Abbildung Machine Learning

Die Image Classification (deutsch Bilderkennung) beschäftigt sich mit nahezu allen visuellen und gekennzeichneten Daten. Die heutigen Anwendungen sind vielfältig von der Diagnostik, wo mittels Image Classification Tumore bestimmt werden, bis hin zur Identifizierung von Kunstwerken und Gemälden.

Computer Vision ist ein interdisziplinäres Forschungsgebiet zwischen der Informatik und den Ingenieurswissenschaften, welches versucht die durch die Image Classification erlangte Fähigkeit zu „sehen“ in das Arbeiten von Computern und Maschinen einzubinden. Beispielhaft ist dafür die Verkehrsschild-Erkennung welche nicht nur Verkehrsschilder identifiziert, sondern auch abhängig davon Fahr-Empfehlungen aussprechen kann.

b Verkehrsschild-Erkennung

Die Verkehrsschilderkennung ist ein Assistenzsystem in vielen modernen Autos und wird seit über 10 Jahren verbaut. Sie funktioniert häufig über eine im Bereich der Rückseite des Innenspiegels montierte Videokamera. Diese filmt die vor dem Auto liegende Straße und deren Umgebung. Das aufgezeichnete Bild dient als Input für die künstliche Intelligenz.

Die Verkehrsschilderkennung durchläuft zwei große Arbeitsschritte. Erstens die Object Detection (deutsch: Objektidentifizierung) und zweitens die Image Classification (deutsch: Bilderkennung) (Abb. 4). Bei der Object Detection wird ein eingehendes Bild in Teilbereiche basierend auf den abgebildeten Objekten unterteilt. Diese Teilbereiche gelten als separate Inputs für die Image Classification, welche den einzelnen Teilbereich separat analysiert und es als das entsprechende Objekt identifiziert.

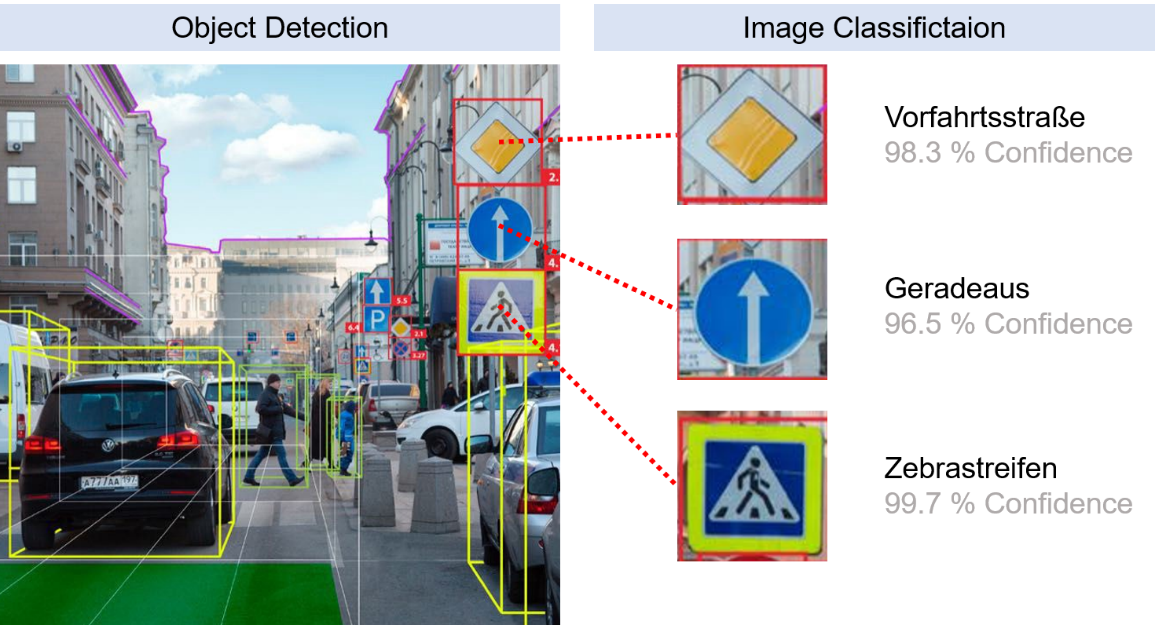


Abbildung 4 Object Detection und Image Classification

Das System assistiert dem Fahrer beim Fahren des Autos. Aufgrund der erkannten Verkehrsschilder kann das System überhöhte Geschwindigkeiten melden, es kann auf Vorfahrtssituationen hinweisen und den Fahrer vor Falscheinfahrten warnen. Der Fahrer wird meist über eine Anzeige auf dem Display oder über einen Warnton auf das erkannte Schild aufmerksam gemacht. Durch neue Vorstöße der Forschung im Bereich des autonomen Fahrens gewinnt die Verkehrsschilderkennung zunehmend an Relevanz. Dies liegt dran, da die von der KI bereitgestellten Informationen über Verkehrsschilder als Grundlage für das Steuern des Fahrzeuges dienen werden. Dementsprechend wächst der Anspruch an die Qualität der Verkehrsschilderkennung.

c Adversarial Attacks

Eine Adversarial Attack (deutsch: Gegenspielerattacke) ist ein vorsätzlich von einem Angreifen manipuliert Input für eine KI, um einen fehlerhaften Output dieser zu provozieren. Wichtig ist, dass die Manipulation keinen Einfluss auf die Wahrnehmung des Menschen hat, sondern nur auf Prozesse der KI. Dem menschlichen Auge fällt in den meisten Fällen kein Unterschied zwischen einem mainpulierten und einem originalen Bild auf. In diesem Zusammenhang ist oft von „optischen Täuschungen“ für künstliche Intelligenzen die Rede. [1]

i Historie

Eine solche Attacke wurde bereits auf eine Vielzahl von Input-Arten angewandt. Einer der ersten attackierten Machine learning Modell war ein E-Mail Spam Classifier. Solche Classifier analysieren den Text einer eingehenden E-Mail und klassifizieren anhand der im Text vorkommenden Wörter die E-Mail in die Kategorien „Spam“ oder „kein Spam“. Bereits vor der Erfindung der ersten neuronalen Netze im Jahre 2004 fanden die ersten solcher Attacken statt, indem man herausfand welche Wörter der Classifier als „gute“ und welche als „schlechte“ Wörter sortierte. Der manipulierte E-Mail-Text wurden anschließend nur noch mit „guten“ Wörtern geschrieben, behielt aber den gleichen Inhalt, woraufhin er anschließend nur noch von Menschen als Spam identifiziert werden konnte. [4]

Bis ins Jahr 2013 hofften viele Forscher das nicht lineare Classifier, wie neuronale Netze, sicher vor adversarial Attacks (AA) seien. Als 2012 neuronale Netze das Forschungsfeld der Bilderkennung dominierten, wurden auch die Bemühungen neuronale Netze durch AAzu manipulieren, größer. 2014 gelang der Durchbruch und die ersten Attacken auf neuronale Netze waren erfolgreich. [5] Seitdem beziehen sich die Attacken fast ausschließlich auf neuronale Netze mit visuellem oder auditivem Input.

ii Targeted und Untargeted

AAs werden in zwei Kategorien unterteilt; gezielte (englisch: targeted) und ungezielte (englisch: untargeted) Attacken. Eine gezielte Attacke hat eine Ziel-Klasse und versucht, das ein Modell einen Input nicht mehr als die ursprüngliche Klasse identifiziert, sondern als die vordefinierte Ziel-Klasse. Eine ungezielte Attacke hingegen versucht nur eine fehlerhafte Klassifizierung herbeizuführen; Als welche falsche Klasse der Input identifiziert wird spielt dabei keine Rolle. Der Vorteil einer ungezielten Attacke liegt in der benötigten Zeit für die Ausführung. Gezielte Attacken brauchen mehr Rechenzeit besitzen dafür aber den Vorteil einer perfekteren und unauffälligeren Attacke. [6]

iii Black-, Grey- und White Box

Zusätzlich wird in der Forschung zwischen „Black Box“ -, „Grey Box“ - und „White Box“ Attacken differenziert. Der Unterschied dieser drei Attacken liegt in dem Wissen über das anzugreifende Netz. Bei einer White Box Attacke hat der Angreifer Informationen über die Architektur des Netzes, sowie die dazugehörigen Aktivierungsfunktionen und die trainierten Parameter. Grey Box Attacken fehlt jegliches Wissen über Parameter und deren Zusammenspiel. Solche Attacken stützen sich alleinig auf die Struktur des Netztes. Die einzige verfügbare Information bei einer Black Box Attacke hingegen ist der In- und Output der künstlichen Intelligenz. Bei einer solchen Attacke wird ein anderes neuronales Netz oder sogar gar kein Netz herbeigezogen, um einen adversarial Input zu generieren, in der Hoffnung, dass das zu attackierende Netz ähnlich auf die Manipulation reagiert. Die meisten Verteidigungsmechanismen, welche bereits erfolgreich über den Attackierer in einem Grey- oder Black Box Szenario triumphierten, scheiterten an White Box Attacks. Zum Beispiel waren sieben der neun bei der 2018 „International Conference on Learning Representations“(ICLR2018) vorgestellten Verteidigungen unerfolgreich gegen die adaptive Struktur einer White Box Attacke. [7]

iv Attack-Arten

Über die Zeit hinweg wurden viele verschiedene Arten zur Durchführung einer solchen Attacke probiert und definiert. Nachfolgend sind die wichtigsten Attacken aufgeführt. Auch wenn die meisten dieser Attacken auf verschiedene Deep Neural Nets anwendbar sind, werden die folgenden Methoden im Kontext zur Image Classification vorgestellt.

Mit Fast Gradient Sign Method (FGSM) wurde 2014 die erste nachweislich erfolgreiche Attacke auf ein Deep Neural Net (DNN) durchgeführt. Die Methode wurde im selben Jahr von Ian J. Goodfellow, Jonathon Shlens und Christian Szegedy bei Google erfunden. Es kann sowohl als untargeted oder targeted Attack durchgeführt werden und setzt ein White Box Szenario voraus, da es auf Grundlage der Parameter des attackierten Netzes eine Störung (auch Noise genannt) bestimmt. Der Noise wird dann mit einem Faktor (in diesem Fall 0,007) auf das Bild addiert und die Manipulation ist abgeschlossen (Abb. 5). [3]

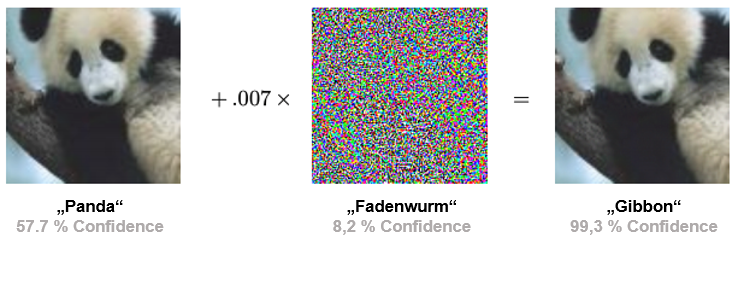


Abbildung Beispiel für ein FGSM

Anstatt die Störung in einem Schritt mit einem vordefinierten Faktor hinzuzufügen, wird bei der Basic Iterative Method (BIM) der Faktor schrittweise erhöht bis eine zufriedenstellende „Confidence“ des DNNs erreicht ist. Vorteile im Gegensatz zu FGSM ist die feinere Justierung des Noises, jedoch ist der Zeitaufwand erheblich höher.

Die zwei bereits vorgestellten Methoden sind berühmte Beispiele für sogenannte Noise-Attacks, also Attacken, welche durch eine auf das ganze Bild gelegte Störung manipulieren. Neuste Forschungen zeigen aber, dass bereits eine Änderung im richtigen Bereich des Inputs für eine vollkommende Fehlklassifikation durch das DNN reichen. Solche Manipulationen werden Adversarial Patch (deutsch: feindlicher Flicken) genannt. Solche Patches kommen in allen Formen und Farben und wurden bereits in realistischen Szenarien umgesetzt. Ein Forschungsteam der Carnegie Mellon University hat 2016 Brillen erstellt, die zur falschen Klassifizierung des Trägers führten (Abb. 6). [8]

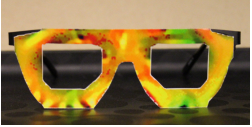


Abbildung Adversarial Patch als Brille

Eine solche Attacke besitzt den Vorteil, dass eine Gesichtserkennungssoftware keinen Fehler melden würde, weil zum Beispiel das Gesicht nicht zu identifizieren sei. Stattdessen würde es mit hoher Wahrscheinlichkeit eine aus seiner Sicht vertrauensvolle Antwort geben (Abb. 7).



Abbildung Adversarial Patches in der Realität

Einer der erfolgreichsten Methoden für Black Box Attacks sind auf Generative Adversarial Networks basierende Attacken (GAN). GANs sind DNNs, welche nach einem unsupervised learning prinzip aufgebaut werden. Ein GAN besitzt immer zwei Netze; Eins der beiden Netze ist bereits trainiert und soll attackiert werden, das andere Netz trainiert sich im Laufe der Attacke und besitzt die Aufgabe eine erfolgreiche Angriffsstrategie zu entwickeln. Im Laufe der Attacke generiert ein Netz (der Generator) Inputs für das bereits trainierte Netz (der Diskriminator). Der Diskriminator generiert einen Output, welcher die Grundlage zum Anpassen der Parameter des Generators dient. Dieses Nullsummenspiel wird so lange fortgesetzt bis der Diskriminator eine falsche Klassifizierung vornimmt. Der Vorteil bei GANs ist, dass kaum Informationen über den Diskriminator bekannt sein müssen. [2]

Das Feld der Adversarial Attack ist vielfältig und sehr aktuell. Bislang sind findet ein Wettrüsten zwischen neuen Verteidigungsmechanismen und neuen Angriffsmethoden statt, bei dem noch nicht klar ist wer am Ende gewinnt.

3 Attacken Methode

a Fast Gradient Sign Method

b Iterationen

4 Validierung

a Attackiertes Modell

i Netzarchitektur

ii Datenset

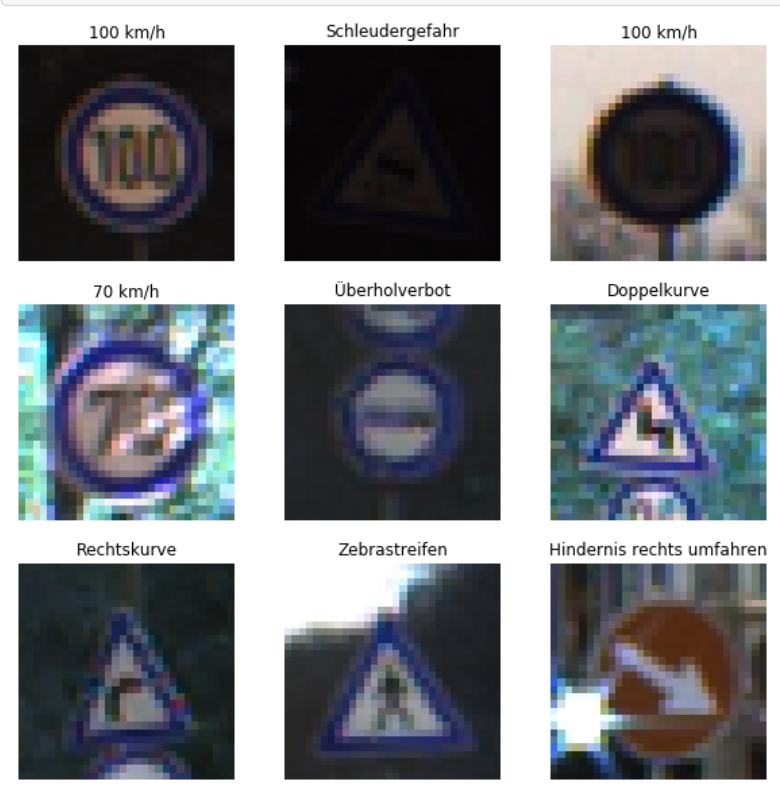
iii Training

b Attacke

Zum Durchführen der Attacke müssen Bilder manipuliert werden.

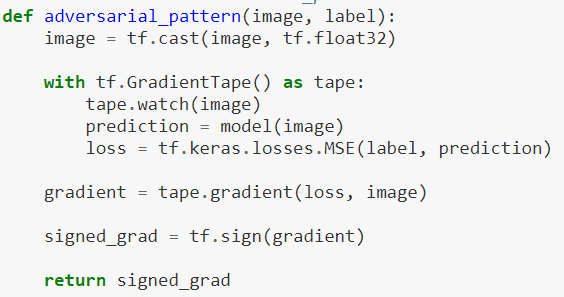
i Bildateinen

Die Attacke wird mit Bildern aus dem Test Datenset durchgeführt, dementsprechend sind die manipulierten Bilder dem CNN unbekannt. Zunächst werden alle 12.630 Bilder des Test Datenset eingelesen und auf die für den Input entsprechende Größe skaliert, also 30 mal 30 Pixel und 3 Farbkanäle. Die Fotos des Datenset sehen nun folgendermaßen aus:



Zum Attackieren der Straßenschilderkennung wird, die bereits im Kapitel drei beschriebene FGSM Attacke verwendet. Die Formel für diese lautet:

Diese Formel sieht im Python Code so aus:



ii. Erfolgreiche lokale Attacken

iii. Erfolgreiche globale Attacken

5 Fazit und Ausbilck

a Fazit

b Ausbilck

6 Verzeichnis

a Literaturverzeichnis

References

[1] Arunava Chakraborty. *Introduction to Adversarial Machine Learning*. https://​blog.floydhub.com​/​introduction-to-adversarial-machine-learning/​.

[2] Bhambri, S., Muku, S., Tulasi, A., and Buduru, A. B. A Survey of Black-Box Adversarial Attacks on Computer Vision Models.

[3] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples.

[4] Gregory L. Wittel, S. Felix Wu. on\_attacking\_statistical\_spam\_filters.

[5] Ian Goodfellow, Nicolas Papernot, Sandy Huang, Rocky Duan, Pieter Abbeel, Jack Clark. 2017. *Attacking Machine Learning with Adversarial Examples*.

[6] Joao Gomes. 2018. *Adversarial Attacks and Defences for Convolutional Neural Networks*. Accessed 29 September 2020.

[7] Ren, K., Zheng, T., Qin, Z., and Liu, X. 2020. Adversarial Attacks and Defenses in Deep Learning. *Engineering* 6, 3, 346–360.

[8] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. Accessorize to a Crime. Real and Stealthy Attacks on State-of-the-Art Face Recognition, 1528–1540. DOI=10.1145/2976749.2978392.

b Abbildungsverzeichnis

[Abbildung 1 Gartner Hype Cycle for Emerging Technologies, 2019 3](#_Toc52298128)

[Abbildung 2 Deep Neural Networks im Kontext 4](#_Toc52298129)

[Abbildung 3 Machine Learning 5](#_Toc52298130)

[Abbildung 4 Object Detection und Image Classification 6](#_Toc52298131)

[Abbildung 5 Beispiel für ein FGSM 9](#_Toc52298132)

[Abbildung 6 Adversarial Patch als Brille 10](#_Toc52298133)

[Abbildung 7 Adversarial Patches in der Realität 10](#_Toc52298134)

c Formelverzeichnis