Gliederung - Individuelles Projekt

# Einleitung (2)

## Problemstellung (1)

## Zielsetzung (1)

# Neuronale Netze und Stand der Technik (8)

## Image Classification und Computer Vision (2-3)

## Verkehrsschild-Erkennung (1-2)

## Adversarial Attacks (2- 3)

# Attacken Methode (6)

## Fast Gradient Sign Method (3)

## Iterationen (3)

# Validierung (18)

## Attackiertes Modell (8)

### Netzarchitektur (2)

### Datenset (3)

### Training (3)

## Attacke (10)

### Bilddateien (2)

### Erfolgreiche lokale Attacken (4)

### Erfolgreicher globale Attacken (4)

# Fazit und Ausblick (6)

## Fazit (3)

## Ausblick (3)

# Verzeichnis

## Literaturverzeichnis

## Abbildungsverzeichnis

## Formelverzeichnis

## Einleitung - 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.

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 all 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.

(Daten müssen auch geschützt werden!!!)

## Einleitung - Zielsetzung

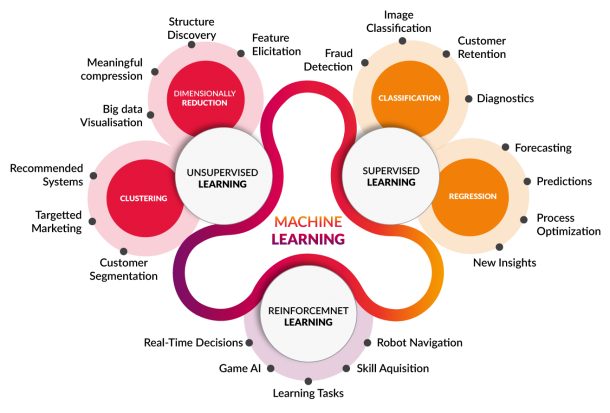
Wirft man einen Blick auf den „Gartner Hype Cycle for Emerging Technologies 2019“ ist „Artificial Intelligence Platform as a Service“ (AI PaaS) einer der erwartungsvollsten Themen der nächsten Jahre. AI PaaS beschreibt Cloud basierte Lösungen für eine Fülle an KI-Anwendungen, vor allem im Bereich der computer vision 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.

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 genau die bereits in der Problemstellung erwähnten 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.

## Neuronale Netze und Stand der Technik – Image Classification und Computer Vision

Das „Überwachte Lernen“ (supervised learning) ist ein Gebiet des maschinellen Lernens. Mit Lernen ist dabei die Fähigkeit einer künstlichen Intelligenz gemeint, Gesetzmäßigkeiten nachzubilden. Das supervised learning setzt im Gegensatz zum unüberwachten (unsupervised) oder bestärkten (reinforcemnet) Lernen, gekennzeichnete (labelled) Daten voraus. Innerhalb des supervised Learning exsitieren zwei große Problemstellungen: Regressionsprobleme und Klasssifikationsprobleme. Regressiosnprobleme beschäftigen sich mit der Vorhersage kontinuierlichen Wertes (Aktienkurs, Häuserpreise, usw.), wo hingegen Klassifikationsprobleme sich mit der Zuordnung eines Inputs zu einer vordefinierten Klasse beschäftigen. 

Die Bilderkennung (Image Classification) ist ein Teilgebiet der Klassifikationsprobleme und beschäftigt sich mit nahezu allen visuellen Daten. Diese Daten können sowohl gekennzeichnet als auch ungekennzeichnet sein, weshalb die Image Classification sowohl Teilgebiet des supervised - als auch Teil des unsupervised learning sein kann. Bei der supervised Image Classification geht es meist um das Klassifizieren von Objekten, Menschen oder Tieren in Bilder. Bei der unsupervised Image Classification handelt es meist um das Sortieren von Bildern, indem die künstliche Intelligenz auf das Finden von Ähnlichkeiten trainiert wird.

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 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. KIs erkennen

(Computer vision)

## Verkehrsschilderkennung

Die Verkehrsschilderkennung ist ein Assistenzsystem in vielen modernen Autos und wird seit über 10 Jahren in vielen Modellen 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 Objektidentifizierung und zweitens die Bilderkennung. Bei der Objektidentifizierung wird ein eingehendes Bild in Teilbereiche basierend auf den abgebildeten Objekten unterteilt. Diese Teilbereiche gelten als separate Inputs für die Bilderkennung. Die Bilderkennung analysiert nun jeden Teilbereich als das entsprechende Objekt.

(Bild)

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 hingewiesen. Durch neue Vorstöße der Forschung im Bereich des autonomen Fahrens gewinnt die Verkehrsschilderkennung zunehmend an Relevanz. Dies führt dazu, dass 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.

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. In diesem Zusammenhang ist oft von „optischen Täuschungen“ für künstliche Intelligenzen die Rede.

Eine solche Attacke wurde bereits auf eine Vielzahl von Input-Arten angewandt werden. Einer der ersten attackierten Machine learning Modell war ein E-Mail Spam Classifier. Solche Classifier anylsieren den Text einer eingehenden E-Mail und klassifizieren anhand meist anhand der verwendeten 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. Der Text wurde von anschließend nur noch von Menschen als Spam identifiziert. Die jüngsten Attacken beziehen sich fast ausschließlich auf Modelle mit visuellem oder auditiven Input.