## Manipulieren neuronaler Netze mit adversativem maschinellem Lernen

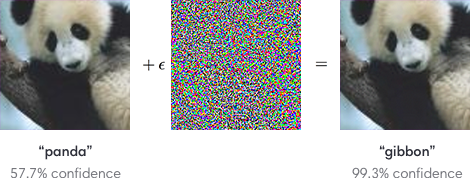
### Exposé zum individuellen Projekt

#### Was?

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.

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 herraus- und in unseren Alltag herreingeschafft haben. 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, um die Umgebung zu analysieren. Fehler und falsche Klassifizierungen können in solchen Fällen fatale Folgen hervorrufen.

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 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.

#### Wie?

In der individuellen Projektarbeit wird ein Adversarial Attack auf ein neuronales Netz in einem White Box Szenario durchgeführt. Ein White Box Szenario beschreibt in diesem Fall eine Attacke auf ein neuronales Netz, dessen Parameter dem Angreifer bekannt sind. So besitzt der Angreifer Einblick in die Architektur des neuronalen Netzes, der Angriff soll aber nicht auf Neuronenebene stattfinden. Die Berechnung und das Training des neuronalen Netzes sollen ungestört bleiben. Die Attacke findet nach dem Training des neuronalen Netzes statt. Ziel der Attacke soll es sein, eine falsche Klassifizierung zu generieren, die der Menschen nicht gemacht hätte. Es soll eine Art optische Täuschung für das neuronale Netz entstehen, die ihre Wirkung nur beim Netz und nicht beim Menschen entfaltet.

#### Warum?

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.

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.

