Richtlinie zum „Software-Sprint“

Manipulation - Manipulation: Erfahrbarmachung der Manipulierbarkeit von Standardmethoden der KI

Schlussbericht

Zuwendungsempfänger:

Katharina Rasch

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS18S61 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Aktuell findet der Diskurs zu Manipulation und anderen Schwachstellen künstlicher Intelligenz hauptsächlich zwischen Akademikern statt und ist sehr technisch/mathematisch. Ich will mit meinem Projekt dazu beitragen, dass sich auch die Zivilgesellschaft gut informiert mit dem Thema künstliche Intelligenz beschäftigen kann.

Ziel meines Projektes war es also, künstliche Intelligenz ein wenig zu entzaubern. Dies wollte/will ich erreichen, indem ich der Zivilgesellschaft die Möglichkeit gebe, selbst auszuprobieren, wie leicht sich Standardmethoden der Bilderkennung (wie sie beispielsweise von facebook und Co eingesetzt werden) manipulieren lassen.

Die geplante Vorgehensweise war:

1. Recherche: welche Bilderkennungsverfahren werden aktuell eingesetzt
2. Aktualisieren meines Wissens zur state of the art in Bildmanipulation / adversarial images
3. Erstellung eines Konzeptes für ein Tutorial, welches Manipulation von Bilderkennungsalgorithmen für die Nutzer praktisch erfahrbar macht
4. Umsetzung des Tutorialkonzeptes

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Meine Zielgruppe sind Menschen, die die Debatte um KI verfolgen und die zwar etwas Technikinteresse haben, sich aber sich selbst wenig mit Technik auskennen (Voraussetzung: wissen wie man mit einem Smartphone ein Foto macht).

Ziel meines Projektes ist es, diesen Menschen die Möglichkeit zu geben, interaktiv moderne Bilderkennungssystem auszuprobieren und ihre Grenzen auszutesten. Zusammen mit allgemein verständlichen Erklärungen zu Bilderkennungssystemen (wie funktioniert das / was funktioniert / was funktioniert noch nicht?) soll der Zielgruppe so ermöglicht werden, selbst besser einschätzen zu können, welche Möglichkeiten und Gefahren der KI echt sind, und welche nur Hype.

Ein Zitat aus der Ausschreibung für Runde 5: “Emergente Technologien können perspektivisch dazu beitragen, gesellschaftliche Themen anders und besser als bisher zu erschließen und zu bearbeiten. Um das zu erreichen, müssen wir aber dafür sorgen, dass sie auch von möglichst vielen Menschen verstanden, interpretiert und mitgestaltet werden können.“ Und genau beim Punkt “möglichst viele Menschen sollten die Technologie verstehen” setzt mein Projekt an.

Ausführliche Darstellung der Ergebnisse

Während des Förderzeitraums sind zwei Prototypen entstanden:

1. **Online-Spiel “Schlag die KI”**

In diesem Spiel tritt ein Mensch gegen eine KI an. Aufgabe ist Bilderkennung: wer erkennt schneller, welches Objekt sich auf dem gezeigten Bild befindet. Das Spiel läuft als Web-Applikation im Browser; unterstützt werden sowohl Computer als auch Mobiltelefone. Mit Hilfe von tensorflow.js läuft direkt im Browser ein Bildklassifizierungsmodell, welches die gleichen Aufgaben bekommt wie die menschlichen Spieler*innen.

Den Kontrahent*innen werden nacheinander drei Bilder von zufällig gewählten Alltagsgegenständen gezeigt:

- Das erste Bild hat keinerlei Manipulationen. In den meisten Fällen wird die KI hier schneller antworten als der Mensch.
- Das zweite Bild ist zu Beginn stark verrauscht, dieses Rauschen wird dann in kleinen Schritten verringert und der Gegenstand langsam deutlicher zu sehen. Bei diesem Bild ist es sowohl für Mensch als auch KI durch das starke Rauschen zu Beginn unmöglich zu erkennen, was auf dem Bild dargestellt ist. Mit reduziertem Rauschen wird die Aufgabe für beide langsam einfacher. Es zeigt sich, dass in vielen Fällen der Mensch schneller ist -- das heisst mit stärkerem Rauschen umgehen kann als die KI.
- Das dritte Bild hat ist mit einem speziellen “adversarial” Muster manipuliert. Für den Menschen stellt sich dieses als zufällig aussehendes Muster von “kaputten” Pixeln (bspw falsche Farbe) dar; der gezeigte Gegenstand ist allerdings für Menschen trotzdem offensichtlich. Die KI hingegen schafft es hingegen nicht zu erkennen was zu sehen ist, selbst wenn die Stärke dieses Musters reduziert wird.

Ziel dieses Spieles ist es, zu vermitteln, dass automatische Bilderkennung zwar mit Menschen mithalten kann -- aber nur unter den richtigen Bedingungen. Während die KI nur statistische Muster aus den Trainingsdaten gelernt hat, arbeiten Menschen auf einem weitaus höheren Abstraktionsniveau. Das macht das menschliche Sehen und Erkennen viel robuster (aber natürlich nicht unfehlbar) gegen Manipulationen.

Der Prototyp dieses Spiels ist unter https://was-kann-ki.gitlab.io/schlag-die-ki/ erreichbar. Bitte beachten: Die Implementation ist zwar funktionsfähig, aber für die Öffentlichlichkeit noch nicht wirklich zugänglich, insbesondere fehlen noch einige Texte und ein Fazit. Ich bitte darum, diesen Link noch nicht außerhalb dieses Berichtes weiterzugeben.

2. **Livedemo/Installation: Gesichtserkennung austricksen**

In dieser Installation sollen Fähigkeiten und Grenzen von Gesichtserkennung für Menschen erlebbar gemacht werden. Angedachter Einsatzbereich für die Installation sind Messen / Konferenzen.

- An einem Monitor ist eine Kamera und in kleiner Computer angebracht. Wenn eine Besucher*in das erste mal vor die Kamera tritt wird ihr Gesicht erfasst und gespeichert
- Ab jetzt kann die Besucher*in erkannt werden. Immer wenn sie vor die Kamera tritt, wird angezeigt dass sie erkannt wurde / bzw nicht erkannt wurde.
- Jetzt kann die Besucher*in versuchen, die Gesichtserkennung auszutricksen, beispielsweise durch: Drehen des Kopfes, Verdecken von Gesichtsteilen mit den Händen, mit Hilfe von bereitgestellten Gegenständen (bspw Clownsnase, Perücke, Mütze, schwarze Balken aus Plastik, evtl Laserpointer, etc).

- Ein Poster präsentiert den aktuellen Stand der Wissenschaft zum Thema Gesichtserkennung.
- Alle gespeicherten Gesichter werden direkt nach Ende der Ausstellung gelöscht (bei Bedarf auch automatisch jede Stunde oder so). Alles wird lokal verarbeitet, keinerlei Daten gelangen ins Internet.

Ich habe während des Projektes diese Installationsidee technisch umgesetzt und als Vorschlag für die https://ki-convention.com/ eingereicht. Dort könnte die Idee erstmals im größeren Umfang getestet werden -- bei Erfolg werde ich die Installation dann (mit eventuell nötigen Anpassungen) auch bei anderen Konferenzen / Ausstellungen einreichen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Während des Projektes habe ich mit Mitgliedern der Zielgruppe zusammengearbeitet, um meine Ideen und Umsetzungen zu verifizieren (mit Hilfe von Papierprototypen und elektronischen Prototypen). Aufgrund positiven Feedbacks bin ich guter Hoffnung, dass meine Prototypen tatsächlich helfen können, Menschen die Grenzen von Bilderkennungssystem aufzuzeigen.

Insbesondere die Installation zur Gesichtserkennung ließ sich allerdings nur bedingt testen, da sich die realen Bedingungen (viele hunderte Menschen, ständig treten neue Menschen vor den Monitor, viel Bewegung im Hintergrund) nur schwer nachstellen lassen. Aus diesem Grund habe ich die Installation zunächst bei einer kleineren Konferenz angemeldet. Ich gehe davon aus, dass sich bei der Konferenz noch Anpassungsbedarf zeigen wird.

Für die Weiterentwicklung / Fertigstellung meiner Prototypen habe ich bereits Zeit eingeplant. Insbesondere für die Installation zur Gesichtserkennung hoffe ich, dass es durch die Open-Source-Stellung leichter sein wird, Kopien dieser Installation in verschiedenen Ausstellungen zu zeigen.

Persönlich habe ich während des Projektes eine Menge teilenswertes Wissen und Material zu den Themen “Grenzen der KI” und “Sicherheitsaspekte von KI-Methoden” aufgebaut. Bei der Data Natives Konferenz im November 2019 in Berlin werde ich einen Vortrag mit dem Titel “Attacks against Computer Vision applications” geben. Ein Blogeintrag zum Thema “Reverse engineering of black-box Computer Vision models” ist in Arbeit (siehe auch letzter Link unter “Kurze Angabe von Präsentationsmöglichkeiten für nötige Nutzer”). Weitere Vorträge und Blogeinträge sind geplant.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Im Projektantrag erwähnte ich unter anderem, dass es mit Hilfe von adversarial images möglich sein könnte, manipulative Bildfilter zu erstellen, um Gesichts/Bilderkennungssysteme

hereinzulegen. Im Rahmen von Milestone 2 “Aktualisieren meines Wissens zur state of the art in Bildmanipulation / adversarial images” habe ich eine Literaturrecherche zum Thema adversarial images vorgenommen. Dabei hat sich herausgestellt, dass die meisten aktuelle Methoden zur Generierung von adversarial images einen direkten Zugriff auf das Bilderkennungsmodell voraussetzen (sogenannte white-box attacken).

Üblicherweise sind tatsächlich im Einsatz befindliche Modelle allerdings nicht direkt zugreifbar, sondern beispielsweise durch eine API (z.b amazon face rekognition API) oder durch ein grafisches Benutzerinterface geschützt (z.b. Facebooks Gesichtserkennung). Sogenannte black-box Attacken, die auch in einem solchen Szenario funktionieren, sind leider noch weitaus weniger erforscht und aktuell noch äußerst ressourcenintensiv (>100.000 API calls für die Generierung eines einzigen manipulierten Bildes). Aufgrund dieser Erkenntnis erübrigte sich die Idee “manipulative Bildfilter” somit recht schnell.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Sämtlicher sourcecode ist auf gitlab verfügbar. Insbesondere:

Webapp: https://gitlab.com/was-kann-ki/schlag-die-ki

Live Gesichtserkennung auf NVIDIA jetson: https://gitlab.com/was-kann-ki/face-recognition-live

Benchmark zur erreichbaren framerate von Modellen für Gesichtserkennung (Vorarbeit für Live Gesichtserkennung) https://gitlab.com/was-kann-ki/face-detection-benchmark

Experimente zum Thema “Was passiert eigentlich wenn ich generierte Gesichter (also nicht existierende Menschen) an eine Gesichtserkennungs-API schicke”: https://observablehq.com/d/5a0732cf5267ad90

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Die Planung wurde eingehalten.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Ich habe mich in meiner Arbeit auf diverse wissenschaftliche Artikel zum Thema Manipulierbarkeit von Gesichtserkennung gestützt. Von Projekten wie CV Dazzle (https://ahprojects.com/cvdazzle/) von Adam Harvey wurde ich inspiriert. Vielen Dank an Ellen König für die Hilfe bei der Ausarbeitung der Wirkungslogik. Vielen Dank an meine Freunde für viele interessante Diskussionen und Ideen. Die Live Gesichtserkennung ist eine Weiterentwicklung von Adam Geitgeys Demoprojekt (Verbesserung von Stabilität und Framerate, mehr Einstellungsmöglichkeiten). (https://medium.com/@ageitgey/build-a-hardware-based-face-recognition-system-for-150-with-the-nvidia-jetson-nano-and-python-a25cb8c891fd).