zeit.de

Künstliche Intelligenz : Ist das wirklich ein Toaster?

Von Stefan Schmitt

4-5 Minuten

Menschen fällt es schwer, in einer Banane einen Toaster zu erkennen. Dafür braucht man schon einen Computer mit künstlicher Intelligenz. Das ist kein Witz, sondern ein unter Informatikern wohlbekanntes Experiment: Auf einer hölzernen Tischplatte liegt, nun ja, eine Banane. Und der Computer, dessen Kameraauge darauf gerichtet ist, erkennt diese auch als *"banana".* Als dann aber eine Hand ein kunterbuntes, glänzendes und ziemlich psychedelisch anmutendes Bildchen neben die Banane ins Blickfeld der Kamera hält, meldet die Bilderkennungssoftware *"toaster".* Mehr noch, nachdem sie die Zuversicht, das Richtige erkannt zu haben, bei der Banane mit 97 Prozent angegeben hatte, ist sie sich jetzt plötzlich sogar zu 99 Prozent sicher.

Andere Fälle zeigen, wie eine Mikrowelle mit störendem Sticker als Telefon einsortiert wurde. Und wie nach Zugabe von ein paar Pixeln, die kein Mensch bemerken würde, der Rechner einen Pandabären als Gibbonäffchen erkennt.

Natürlich ist das witzig, nicht ohne Grund zirkulieren die Beispiele unter Fachleuten: So leicht lässt sich also jene Zaubertechnik austricksen, die seit einigen Jahren populäre Fantasien und Hoffnungen befeuert, die künstliche Intelligenz (KI). Oder wie das

Silicon-Valley-Zentralorgan *Wired* es im vergangenen Herbst zusammenfasste: "KI hat ein Problem mit Halluzinationen."

Diese Schwäche kann gruselig erscheinen: So konnten Forscher im vergangenen Jahr ein Stoppschild durch wenige geometrische Aufkleber, die keinen Menschen stören würden, für eine Bilderkennung unkenntlich machen. Anderen gelang es, dem Computer eine Plastikschildkröte durch subtile Gravuren auf ihrem Panzer als Schusswaffe erscheinen zu lassen.

Straßenverkehr, Waffen – spätestens da hört der Spaß auf. *Adversarial attacks* ("feindselige Angriffe") nennen Informatikerinnen und Informatiker es, Systeme durch gezielte Störungen zu völlig falschen Ergebnissen zu verleiten. Bisher sind solche Attacken Experimente im Labor, nicht aus dem echten Leben. Irrelevant sind sie deshalb aber nicht. Ja, sie bieten sogar die Chance, die Technik besser zu verstehen und besser zu machen.

Gerade haben Tübinger Forscher um Michael J. Black, Direktor am Max-Planck-Institut für Intelligente Systeme, die Liste der Trugbilder um ein Beispiel von neuer Qualität erweitert. Mit einem kleinen, kunterbunten Pixelhaufen haben sie verschiedene Systeme zur Bewegungserkennung *(optical flow systems)* gründlich von ihrer Aufgabe abgelenkt: nämlich auf Basis von Videoaufnahmen zu berechnen, wie schnell sich Objekte bewegen und wohin. Schon ein Angriff auf einen kleinen Teil des Bildes habe eine große Auswirkung gehabt, berichtet Black. Weniger als ein Prozent der Fläche nahm der Pixelhaufen ein, störte aber erheblich. Ende Oktober haben die Tübinger ihre Attacke bei der International Conference on Computer Vision in Seoul demonstriert.

Weil *optical flow*-Systeme es erlauben, im Gewusel einer

Verkehrssituation die wesentlichen Bewegungen (etwa von Fahrzeugen und Radfahrern) zu verfolgen, sind sie plausible Bausteine für künftige selbstfahrende Autos. Um Herstellern und Zulieferern Gelegenheit zu geben, ihre Systeme auf Schwächen zu testen, hatten die Forscher sie lange vor der Veröffentlichung informiert und ihnen auch die störenden Pixelhaufen zugeschickt. *Adversarial patches* nennen sie diese, was man als "feindselige Aufkleber" übersetzen könnte. "Für die Berechnung eines Patch brauchen wir hier vier Stunden", berichtet der Doktorand Anurag Ranjan, einer der Autoren. Das Team habe an einem *optical flow*-System mit KI ausprobiert, welches "optimale Muster die größte Störung verursacht".