De Standaard, zaterdag 12 mei 2018 - Cultuur en Media

En plots herkent uw slimme auto het stopbord niet meer

Artificiële intelligentie is bijzonder makkelijk te hacken

Een stickertje dat een mens amper opmerkt, kan een stopbord onherkenbaar maken voor een zichzelf besturende auto. Artificiële intelligentie mag dan in korte tijd veel krachtiger geworden zijn, ze blijkt erg kwetsbaar voor een totaal nieuw soort hacking. Dominique Deckmyn

Een sticker kan vrijwel elk AI-systeem dat beelden herkent, op een dwaalspoor zetten. De klever bevat alleen een kluwen van lijnen en kleuren, maar is bijzonder machtig. Leg hem naast een banaan, en voor de computer is die banaan plots een broodrooster geworden. En wat met de camera's die op luchthavens worden ingezet om gezochte terroristen te herkennen? Deze sticker, of een bril die op gelijkaardige patronen is gebaseerd, kan ervoor zorgen dat een terrorist gewoon kan doorlopen.

Verander één pixel, en de computer ziet plots iets helemaal anders; een burrito in plaats van een ijsje, bijvoorbeeld. istock

Hoe dat werkt? 'Dat weet men eigenlijk niet zeker', zegt onderzoeker Jonathan Peck. De sticker bevat wel een veelheid aan lijnen en kleuren, en dat leidt de 'aandacht' van het AI-systeem af – een beetje zoals de rode lap een stier om de tuin leidt.

Jonathan Peck werkt aan de Universiteit Gent aan zijn doctoraat: hij wil uitzoeken hoe je AI-systemen kunt beschermen tegen dit soort misleiding. Hij werkt daarom aan een methode om te meten hoe gevoelig een bepaald netwerk is voor dergelijke aanvallen, zodat die

gevoeligheid kan worden verminderd. Zijn promotoren zijn de professoren Yvan Saeys en Bart Goossens.

> ‘Er wordt momenteel massaal geïnvesteerd in deep learning, terwijl er maar heel weinig mensen bezig zijn met de veiligheid’ Bart Goossens Professor, UGent

Artificiële intelligentie is de jongste jaren enorm veel beter geworden. Vooral in het herkennen van beelden en geluiden. Dat is te danken aan de ontwikkeling van deep learning. Maar ‘deep learning’-systemen, de zogenoemde *deep neural networks*, kunnen erg gemakkelijk worden misleid. Dat is al aangetoond in 2013.

De wapens voor een aanval op deep learning heten *adversarial perturbations*, vijandige verstoringen. Maar het zijn gewoon manieren om een AI-systeem om de tuin te leiden. Noem het gerust ‘hacken’. Eerst ging het om aanvallen die alleen werken op één bepaald *deep neural net*, waarover je heel gedetailleerde kennis moest hebben. Al snel ontdekte men dat het ook lukte zonder die kennis van het binnenwerk – er kwamen ook zogenoemde *black box*-aanvallen. En inmiddels zijn er *universal adversarial perturbations* ontdekt, aanvallen die op alle bekende neurale netwerken effectief zijn, zoals die kleurige sticker. ‘Het is een wapenwedloop aan het worden’, zegt Peck. ‘Er worden steeds nieuwe technieken gevonden om “deep learning”-systemen te beschermen, en om ze aan te vallen.’

Dat het zo moeilijk is om deze nieuwe vorm van hacking tegen te gaan, komt omdat we onze AI-systemen zelf niet zo goed begrijpen. Yvan Saeys: ‘*Deep neural networks* hebben veel “lagen” van kennis die telkens abstracter worden. Je weet wat je erin steekt en wat er aan de andere kant uit komt, maar hoe het neurale netwerk ertoe komt, dat is erg moeilijk te begrijpen.’

> **‘Het is een wapenwedloop. Er worden steeds nieuwe technieken gevonden om “deep learning”-systemen te beschermen, en om ze aan te vallen’**

Deep learning is verbazingwekkend efficiënt. ‘Voed’ zo’n systeem met duizenden of miljoenen voorbeelden van een kat, en het zal bijna feilloos een kat herkennen. Maar je weet niet hoe het werkt, en je kunt dus niet garanderen dat het altijd juist zal werken. Gevaarlijk, als je leven ervan afhangt – en in een zichzelf besturende auto is dat letterlijk het geval. ‘Pas heel recentelijk probeert men bewijsbare garanties te geven’, zegt Peck.

Dit stickertje misleidt de artificiële intelligentie. rr

Pixel

Een *adversarial perturbation* wordt ‘gemaakt’ door een beeld – bijvoorbeeld de foto van een panda – heel lichtjes aan te passen. Voor de mens ziet de foto er onveranderd uit, maar een ‘deep learning’-

systeem zal het nu herkennen als een gibbon. Jonathan Peck wijzigde een foto van een ijsje zodat die door de Resnet-beeldherkenningssoftware van Microsoft wordt herkend als een burrito.

Zo'n misleidend beeld maken is volgens Peck gewoon een staaltje wiskunde, meer bepaald een optimalisatieprobleem: je probeert met een zo klein mogelijke wijziging aan het beeld – soms volstaat één pixel – een zo sterk mogelijk verschillend resultaat te krijgen. Hoe moeilijk dat is? 'Verrassend gemakkelijk', zegt Jonathan Peck. 'Je kunt het op je eigen computer doen.' De nodige instrumenten zijn immers open en bloot op het internet te vinden, met namen als Foolbox en CleverHans. In sommige gevallen is er wel een simpele remedie: veel bewerkte foto's worden wel weer correct herkend als ze een klein beetje naar links of naar rechts worden gedraaid.

Adversarial perturbations hoeven geen vlakke foto's te zijn: studenten van het MIT produceerden met een 3D-printer een plastic schildpadje dat wordt herkend als een geweer.

Bescherming

Je kunt *adversarial perturbations* overigens ook inzetten om jezelf te beschermen. Bijvoorbeeld tegen gezichtsherkenning. Elke foto van u die online staat, kan in principe worden gebruikt om een AI-systeem te 'trainen' zodat het u kan herkennen op andere foto's – of gewoon op straat (in de Chinese stad Shenzhen wordt gezichtsherkenning ingezet om mensen die door het rood lopen te vatten). Maar u kunt uw familiekiekjes onleesbaar maken voor AI-systemen. En u kunt op straat zelfs rondlopen met een speciale bril die de camera's misleidt.

Adversarial perturbations werken niet alleen met beelden, maar ook met geluiden. Over enkele maanden kunt u ook in Vlaanderen een slimme luidspreker zoals de Google Home, de Amazon Echo of de Apple Homepod kopen. Als u die luidspreker het juiste gesproken commando geeft, zet hij uw lievelingsmuziek op of knipt de lichten aan en uit. Maar in januari toonden Chinese onderzoekers aan hoe je verborgen stemcommando's kunt inbouwen in een stukje muziek – de onderzoekers noemen deze muziekjes 'CommanderSongs'. Een onschuldig filmpje dat je bekijkt op Youtube, kan een – voor de mens onhoorbaar – geluid bevatten dat uw slimme luidspreker beveelt om het slot van de voordeur te openen. Of hoe een 'slimme' woning bijzonder gevaarlijk kan zijn.

'Er wordt momenteel massaal geïnvesteerd in deep learning, terwijl er maar heel weinig mensen bezig zijn met de veiligheid', constateert Bart Goossens. 'Pas als het een keer echt foutloopt, gaat men daarin investeren. Er is nog veel fundamenteel onderzoek nodig om dit probleem op te lossen, maar intussen wordt deze technologie wel steeds meer in de praktijk gebracht.'

Dat doet sterk denken aan de begindagen van het web, in de jaren 90: computers werden zonder enige voorzorg aan het internet gehangen. Twintig jaar later maken hackers en virussen daar nog altijd misbruik van.

Hoe werkt een 'deep learning'-netwerk?

Een neuraal netwerk is geïnspireerd op de werking van de hersenen, al is het daarvan een erg versimpelde versie. De virtuele hersencellen bevinden zich in lagen boven elkaar. In de jaren 60 hadden

neurale netwerken maar enkele lagen, de huidige ‘deep learning’-systemen of *deep neural networks* hebben er tientallen, tot zelfs duizenden. Deze *deep neural networks* liggen aan de basis van bijna alle doorbraken op het vlak van artificiële intelligentie de jongste jaren.

Elke laag staat voor een hogere graad van abstractie. De onderste laag ziet alleen de stipjes in een foto. De tweede laag onderscheidt randen en lijnen. Iets hoger worden vormen als driehoeken en cirkels herkend, en complexere vormen zoals ogen en neus. Een neuraal netwerk van tientallen lagen ‘diep’ kan gelaatsuitdrukkingen inschatten of gezichten herkennen.

De meeste neurale netwerken zijn ‘getraind’ om één bepaalde taak uit te voeren. Bijvoorbeeld hondenrassen herkennen op een foto, of gesproken commando’s omzetten in tekst. Dat trainen gebeurt door ze bloot te stellen aan duizenden, soms miljoenen voorbeelden.