SYNSPUNKT KUNSTIG INTELLIGENS

AI og den forunderlige, menneskelige ting kaldet sund fornuft

Det lyder tilforladeligt, når man hører, at AI stadig ikke har noget der ligner almindelig, menneskelignende sund fornuft. Men hvad er sund fornuft egentlig, og hvorfor får vi ikke bare lavet noget AI med indbygget sund fornuft, spørger Simon Laub i dette synspunkt.

Simon Laub, adjunkt, **Erhvervsakademi Aarhus**

https://pro.ing.dk/15619

11. nov 2021 06:12

Efterhånden har AI overlevet flere såkaldte AI-vintre - perioder med lav aktivitet og nedskæringer, hvor alle har været godt trætte af over-optimistiske AI forskeres vilde forudsigelser, som ikke kunne realiseres. Den første AI-vinter startede i 1970'erne, hvor man efter enorme fremskridt fra computernes spæde start i 1950'erne havde svært ved at komme videre ud i den virkelige verden fra de små "legetøjs-problemer", man hidtil havde haft succes med. Den virkelige verden viste sig ganske enkelt at være for kompleks for datidens skrøbelige AI teknikker og den begrænsede regnekraft, man havde i computere dengang.

1980'erne var igen en opblomstringstid for AI, dog ikke under det nu noget vanærede navn AI, men under betegnelsen "ekspertsystemer" eller " vidensbaserede systemer". Indenfor begrænsede, snævre områder kunne man lave systemer, der kunne løse opgaver på lige fod med menneskelige eksperter. Et af de mere kendte systemer fra perioden var MYCIN, der kunne hjælpe læger med råd om blod infektioner. Da man begyndte at tjene penge på systemerne, begyndte AI langsomt at komme ud af

MEST LÆSTE

1. EU-Parlamentet vil have forbud mod AI-dommere, predictive policing og masseovervågning med ansigtsgenkendelse
2. D-mærket kræver AI med nødstop og klageret: Det er ikke raketvidenskab, det er bare godt håndværk
3. Twitters algoritmer forstærker højreorienterede, men udviklerne ved ikke, om det er et problem

vinterens mørke.

Men også ”ekspertsystemerne” var for skrøbelige, for dyre at vedligeholde og gav mærkelige svar, hvis man spurgte dem om noget uden for deres snævre ekspert-område, hvorfor man så fra slutningen af 1980’erne igen bevægede sig ind i en AI-vinter. I de næste mange år var det således bedre at sige, at man arbejdede med ting som datamining, dataanalyse eller talegenkendelse end at tale generelt om AI, som igen havde vist sig ikke helt at kunne leve op til forventningerne.

Bedre computerhardware og enorme mængder data indsamlet på internettet ændrede imidlertid igen situationen. Et neuralt net, AlexNet, designet af Alex Krizhevsky i samarbejde med Ilya Sutskever og George Hinton (alle fra University of Toronto), vandt i 2012 ret overlegent en konkurrence om at klassificere billeder i 1000 kategorier. Noget af comeback for neurale net, hvor Marvin Minsky og Seymour Papert i 1969 havde gjort deres for at punktere de fleste drømme om, hvad neurale net kunne (ved at vise at et neuralt net i et lag, en perceptron, ikke engang kunne løse helt simple problemer, som f.eks. det såkaldte XOR-problem). Og da man endelig havde forstået, at neurale net med flere lag sagtens kunne løse disse simple problemer og meget mere, så var der stadig et stort spring frem til vores ”deep learning” tid, hvor neurale net med mange lag, trænet på store mængder data, pludselig kan løse problemer, der hidtil har været helt uløselige.

Nye fremskridt - nye spørgsmål

Vi er nu i en tid, hvor deep learning ikke bare giver os billedgenkendelse og selvkørende biler, men også kan bruges til at finde uregelmæssigheder i finansielle transaktioner, til detektion af problemer i sundhedsdata, til oversættelse af tekster fra et sprog til et andet, tekstklassifikation og meget mere.

Den sidste AI-vinter er efterhånden kun et fjernt minde. Og en ny AI-vinter får vi ikke lige foreløbigt. Alle i IT-branchen vil i mange år fremover have alt rigeligt at gøre med at få udnyttet deep learning teknikker mm. rundt i de mange nicher, hvor der er muligheder, men som man bare endnu ikke er kommet frem til. Spørgsmålet er mere ”hvor langt kan det her gå?”. Er vi på vej mod menneskelignende intelligens, eller vil også deep learning snart nå et plateau, hvor man står med indlysende problemer, hvorfra det er

over-optimistiske fremtidsbud, mindre interesse og på sigt måske færre investeringer og offentlige bevillinger? Hvilket så i sidste ende vil kunne være begyndelsen på en ny AI-vinter?

Men hvad er det så for problemer, som er interessante lige nu? At sortere tal er gået fint siden 1959, hvor der blev opfundet en algoritme, Quicksort, der kunne gøre det ganske hurtigt. At spille forskellige brætspil går også fint. IBMs skak-computer, Deep Blue, vandt over den daværende verdensmester i skak, Garry Kasparov, i 1996, og siden har computere været bedre til at spille skak end mennesker. Spillet "Go" var en noget vanskeligere opgave for computerne, men det lykkedes alligevel i 2016 firmaet DeepMind med programmet AlphaGo at slå Lee Sedol, en af verdens absolut bedste Go-spillere. At genkende ansigter i billeder har i de sidste 10 år været ganske nemt. Relativt fornuftige systemer til oversættelser mellem forskellige sprog har været hvermandseje i de sidste 10 år, ligesom der også er gjort gode fremskridt på fronten med de førerløse biler mm. Det er først, når man beder computerne om at fortolke, hvad der egentligt foregår i et billede eller læse en tekst og svare på spørgsmål, om det de har læst, at problemerne begynder at opstå.

Så det var ikke overraskende, at det gav anledning til stor opstandelse, da man for et år siden i den engelske avis The Guardian kunne en læse en artikel med overskriften "A robot wrote this entire article. Are you scared yet, human?". Tilsyneladende var det lykkedes det amerikanske AI firma OpenAI at få deres deep learning program GPT-3 (Generative Pre-trained Transformer) til at producere artikler på et menneskelignende niveau. Noget som i hvert fald The Guardian synes skulle have en overskrift, der kunne kommunikeres direkte ud til hjernens frygtcenter, uden helt at give tid til at hjernens mere rationelle centre i vores præfrontale cortex, så det kunne nå at finde ud af, hvad der var op og ned i historien.

Der har selvfølgelig længe eksisteret chatbots, man har kunnet tale med i naturligt sprog. På MIT skabte Joseph Weizenbaum i 1960'erne programmet ELIZA, der ved hjælp af simple regler og søgning på nøgleord i sætninger, ofte havde til held til at ændre lidt i brugerens egen sætning, så den kunne sendes retur i form af et spørgsmål. Alt i alt en slags parodi på et besøg hos en uhyre irriterende psykoterapeut (Hvor spørgsmål som "Kan du X?" bliver besvaret på skabelon form med svar som "Tror du ikke, jeg kan X?" osv). Weizenbaum selv var overrasket over, hvor få simple regler, der egentligt skulle til for at få - i hvert fald

nogle - mennesker til at tro, at programmet faktisk en gang imellem forstod lidt af, hvad der blev talt om.

Mere end 50 år senere kan computere naturligvis meget mere. En af de største er GPT-3, der har 175 milliarder parametre trænet til at lave "next word prediction", hvor der ud fra en input-sekvens af ord, produceres en ny sekvens af ord. Sekvenserne er baserede på, hvad programmet synes er det mest sandsynlige, efter at det er blevet trænet på flere store dataset, deriblandt "The Common Crawls", mange petabytes af data, indsamlet fra milliarder af websider på internettet. Prisen for at træne GPT-3 modellen antages at have været i størrelsesordenen millioner af dollars, og endnu større systemer er for længst kommet til.

Mange af de brugere, der har haft mulighed for at afprøve GPT-3 systemet, har været ovenud begejstrede. Og har ikke talt om "menneskelignende tekstgenerering" på baggrund af "statistiske sammenhænge mellem ord sekvenser", men om spændende samtaler om "kreativitet", "de geopolitiske konsekvenser af AI", "meningen med livet" og meget mere.

Mere problematisk har det været, at GPT-3 også har haft en kedelig evne til at komme med mere eller mindre racistiske og sexistiske udtalelser. Samt virker til at være overbevist om, at muslimer generelt er voldelige, og at problemer i sundhedssektoren kan løses ved at patienterne tager livet af sig mm. Hvortil kommer, at GPT-3's skabere selv har udtalt, at det er svært for GPT-3 at holde fokus i en samtale, at GPT-3 ofte modsiger sig selv og ikke altid er helt med på, hvordan den fysiske verden egentlig fungerer.

Skal computere tvinges til at se sæbeoperaer?

Alt i alt kommer det hurtigt til at lyde ret slemt. Men næppe helt overraskende, hvis man tænker lidt over, hvad resultatet kunne være, hvis man træner et system med data fra sociale medier og lignende.

Og man kan vel heller ikke helt afvise, at GPT-3 på nogle punkter minder om menneskelignende intelligens. Er vi ikke alle stødt ind i mennesker, der gentager tvivlsomme ting, de har læst på internettet? Virker de til at have svært ved at holde fokus i en samtale og have svært ved at forstå komplekse problemstillinger?

Ligesom vi vel heller ikke helt kan påstå, at vi aldrig før har hørt nogle af de politisk ukorrekte ting, GPT-3 kan finde på at sige.

Men hvis vi går ind i en fremtid, hvor GPT-3 lignende systemer selv begynder at poste på de sociale medier, svarer på vores private emails, hjælper dystopiske politiske partier med henvendelser fra borgere, blander sig i vores Google søgninger mm. Så kan man næppe kalde det "Responsible AI", hvis man ikke har tænkt lidt over, hvad det er for data, man har brugt til at træne systemerne med.

Og her blivet historien fra 2020 om Timnit Gebru's exit fra Googles "Ethical AI Team" i øvrigt særligt bekymrende. Havde hun, en sort kvinde født i Etiopien, lettere ved at se og forstå de mange potentielle problemer, der kan gemme sig i store datasæt, end de mange mænd der arbejder på de systemer, der bruger disse data? Med deraf følgende kontroverser og konflikt om, hvad vejen frem så skulle være. Eller var det mere en konflikt mellem en ideel, men upraktisk tilgang overfor en mere pragmatisk tilgang til at løse problemer bedst muligt her og nu - eller noget helt andet?

Om ikke andet fik vi med Gebru's exit fra Google igen fokus på de problemer, der kan ligge og vente, når man træner på store datasæt - hvis man nu havde glemt, at Google Photos AI i 2015 kunne finde på at klassificere sorte som "gorillaer"og gentaget i år, hvor Facebook AI klassificerede sorte i videoer som "primater".

Selvfølgelig kunne man forestille sig regler til at hjælpe systemet ud af problemerne igen. F.eks. ved, at man fik trænet systemet til at genkende uheldige ting i en samtale, så man straks kan skifte emne, hvis en samtale er ved at bevæge sig i en uheldig retning. Men uden nogen rigtig forståelse i systemet kan man let få den tanke, at hver gang man får styret udenom et problem, så styrer systemet nok bare selv ind i det næste.

Ansvarlige forældre ville ikke være i tvivl, hvis det handlede om børn. De skulle i skole. En god skole, der har styr på fakta og lærere, der udtaler sig om ting som køn, alder, etnicitet på en måde som forældrene kan bifalde. Ansvarlige forældre ville nok derefter kræve, at børnene fik lært noget om kritisk tænkning og fik indbygget tydelige rutiner og visdom om historie, matematik, retssamfund, moral og mere, som kan hjælpe dem med at styre sig selv og det samfund, de lever i.

Men det begynder at blive indviklet, når mennesker pludselig ved meget mere, end hvad der bliver sagt direkte i en samtale.

Skal de stakkels computere tvinges til at se alle sæbeoperaer, der nogensinde er vist på TV samtidig med at tygge sig igennem de sociale medier? For at kunne regne ud, at hvis en mand siger ”jeg forlader dig”, og en kvinde så svarer med at spørge ”hvem er hun”, at så er denne kvinde nok ret vred?

Mod menneskelignende intelligens

Den amerikanske kognitionsforsker Gary Marcus er skeptisk og mener ikke, at et ”simpelt trick” eller to og en masse data er nok til at give menneskelignende intelligens. Hjernen kan tænke abstrakt, hvilket ikke er det samme som at føle noget, som igen ikke er det samme som at huske noget. Hjernen kan forstå, hvordan ting hænger sammen og kan holde styr på, hvad den ved om verden, så vi kan bevæge os rundt i den og bemærke, hvis ting bliver lavet om.

Robotter med menneskelignende intelligens må tilsvarende have interne modeller af, hvad de ved om verden. De må kunne tilgå viden om den virkelige verden, så de kan regne ud, at folk i Polen for det meste taler polsk, ikke kinesisk. De må kunne ræsonnere, så de forstår, at når mælk ikke er giftigt, dør mennesker ikke af at drikke mælk. Og de må bestemt også have en forståelse for, hvordan den fysiske verden fungerer her på Jorden. Når vi sætter os ind i et tog, så bliver vi mennesker overraskede, hvis der pludselig åbner sig et ormehul i rum-tid-strukturen, som øjeblikkeligt bringer os og toget om på den anden side af Jorden.

Ifølge Marcus er sammenhænge mellem ordsekvenser derfor langt fra nok til at give os menneskelignende intelligens. I hvert fald så længe disse systemer ikke også får indbygget moduler, der kan ræsonnere. Samt moduler, der har viden om vores verden, vores værdier, og hvordan den fysiske verden fungerer. Hvis man ikke vil gå så langt som til at sige, at GPT-3 er helt hjernedød, så må man i hvert fald indrømme, at den slags systemer stadig har store mangler, og ofte laver ret tåbelige fejl. Spørgsmål om hjertemedicin og rådgivning af selvmordstruede er stadig ikke noget, vi helt er klar til at lade chat-systemerne tage sig af.

Omvendt er det jo også klart, at systemerne er blevet meget bedre på det seneste. Så i en præcis evaluering af, hvor gode

systemerne er, må man forstå hvilken type model, der fejlede - og under hvilke omstændigheder. Hvor langt er den testede model fra en "state of the art" model? Blev der testet for at demonstrere, hvor intelligent modellen er eller for at udpensle fejl?

Og når man taler om systemer med menneskelignende sund fornuft, skal man selvfølgelig også være præcis med, hvad man egentlig mener. I skak har Deepminds AlphaZero system siden 2017 været verdens bedste. Hvis man nu spurgte AlphaZero om mennesker i skak udviser "sund fornuft", må svaret vel være nej, da AlphaZero kun har oplevet, at menneskers skaktræk fører direkte til tab.

Computerne er også foran, når det handler om ikke at tro på hvad som helst, man hører. I hvert fald er computerne bedre til at detektere "fake news" på nyhedssider end mennesker. Og computerne er også ved at være ret fornuftige, når det handler om at køre bil. I det hele taget har computerne en fordel, når det handler om at træne på mange eksempler indenfor et begrænset område.

Den israelske psykolog Daniel Kahneman taler om to måder at tænke på: "fast and slow". Ved hjælp af deep learning er computere nu indenfor mange områder ved at blive ret gode til "thinking fast"-delen med hurtige automatiske svar. Samtidig går det knapt så godt for computerne med "thinking slow"-delen, hvor der kræves generelt overblik og logisk tænkning. GPT-3 er allerede i vanskeligheder, når man beder om et referat af en tekst. Og computere har i hvert fald (endnu) ikke det nødvendige overblik til at kunne følge en historie og svare på spørgsmål om historien. Og der er slet ikke den integration mellem de to systemer, fast and slow, som man ser hos mennesker.

Manglen på sund fornuft bliver tyderligere

Så hvad er næste skridt? Facebook har fortalt, at man arbejder på systemer, der forstår verden på samme måde som mennesker. Dvs. systemer, der er trænet ved hjælp af video, der er optaget som set ud af øjnene på en person. Det vil selvfølgelig være nyttigt i Facebooks nye Metaverse, hvor alle (Facebooks) brugere er udstyret med en Ray-Ban smart-brille, der kan optage alt, de ser og hører. Udover at brillen i dette fremtidige univers fungerer som en computerskærm, der kan vise brugerne alt, hvad de tidligere kun kunne se på deres telefon. Inklusive at fortælle én med store pile i ens synsfelt, hvor man skal gå hen, når

man er ude at gå.

Hvis man har glemt sine nøgler, arbejdes der i forskningsprojektet Ego4D på et system, der kan gå tilbage i alt det, der er optaget og finde ud af, hvor nøglerne er. Ja, så langt så godt. I hvert fald så længe man ikke har betænkeligheder ved at gemme sit privatliv på Facebooks servere. Men endnu mere futuristisk antyder man også i dette projekt, at man påtænker at arbejde med sekvenser af video på samme måde, som andre arbejder med sekvenser af tekst til forecasting. Således at systemet også her kan begynde at gætte på, hvad der normalt følger efter en given sekvens, Når man koger kartofler, kommer de fleste mennesker f.eks. salt i gryden, efter at vandet er kommet i. Så hvis det engang lykkedes Facebook AI at træne systemer på baggrund af sådanne videosekvenser, ja, så kan Facebook-brillen gøre opmærksom på problemer, som hvis man f.eks. har glemt saltet, og fremtidige robotter ved, hvad det bedste næste træk er " i lignende situationer.

Men problemerne med manglende sund fornuft er også undervejs blevet lettere at få øje på.

Forestiller man sig f.eks., at alle butikscentre skal udstyres med Pepper lignende robotter til at vise vej og hygge om kunderne, så skal robotterne vel bruge computer-vision til at identificere, hvad der sker foran dem? Hvilket så kan gives som input til GPT-3, der kan snakke med kunden. Her må man ud fra en "Responsible AI"-betragtning må håbe, at vision-systemet ikke har samme uheldige tendens som Facebook AI og klassificerer nogle kunder som "primater". For slet ikke at tale om hvad der kan ske, hvis den slags oplysninger gives videre til GPT-3.

Sårede følelser er selvfølgelig slemt nok. Men hvis man nu forestiller sig, at robotten skal bevæge sig rundt i centret baseret på træning med én persons video optaget af millioner af Metaverse-brugere, så begynder i hvert fald nogen af os at tænke, at Gary Marcus nok har en pointe med, at sekvenser af forløb ikke altid er nok til at give sund fornuft. Selvom mange andre har siddet på en bænk, skal man ikke gøre det, hvis der står nymalet på den. Man bliver nødt til at vide noget om bl.a. logik, moral og fysik, før man kan handle med sund fornuft. Og visdom, dvs. på en måde som man også bliver glad for i længden.

Sund fornuft er stadig noget, vi alle sammen kan ønske os mere af. Men gælder det også, hvis man er en robot?

KUNSTIG INTELLIGENS

Simon Laub
Simon Laub er adjunkt på afdelingen for uddannelser inden for it- og softwareudvikling ved Erhvervsakademi Aarhus.