3g SRP - Kunstig intellegens

Jeppe Møldrup

Nørresundby gymnasium og HF

21/12-2018

Abstract

This thesis is about the mathematical structure behing an Artificial Neural Network(ANN), it goes in depth into how to feed an ANN with data, and train the network according to the output it gives. Furthermore it delves into some ethical questions that come with ANN's, like how we keep ANN's at bay and not take over the world, who's to blame if a self driving car steered by an ANN makes a mistake and crashes etc.

Indholdsfortegnelse

1 Indledning

I denne opgave vil jeg først og fremmest undersøge den underliggende matematiske struktur bag et Artificial Neural Network, så vil jeg udlede nogle af de aller vigstigste formler der indgår i et ANN, nemlig de formler der bliver brugt til at træne det. Og til sidst komme med et eksempel på et mere moderne neuralt netværk for at sætte i perspektiv hvad disse netværk er i stand til at gøre. Derudover vil jeg undersøge nogle og diskuttere nogle etiske spørgsmål der er dukket op sammen med ANN'er, og inddrage vinkler fra Det Etiske Råd, Københavns Universitet og Isaac Asimovs bog "Robot".

2 Redegørelse af et ANN netværk

2.1 ANN's baggrund

Et ANN(Artificial Neural Network) er en form for kunstig intellegens, hvor man prøver at efterligne den menneskelige hjerne ved at lave simple versioner af det netværk af de mange milliarder neuroner der findes i den menneskelige hjerne. Ligesom hvor en biologisk menneskehjerne har neuroner og synapser med forskellige styrker, så har et ANN også det, dog lidt simplere så man kan tillægge værdier til de forskellige neuroner og synapser.

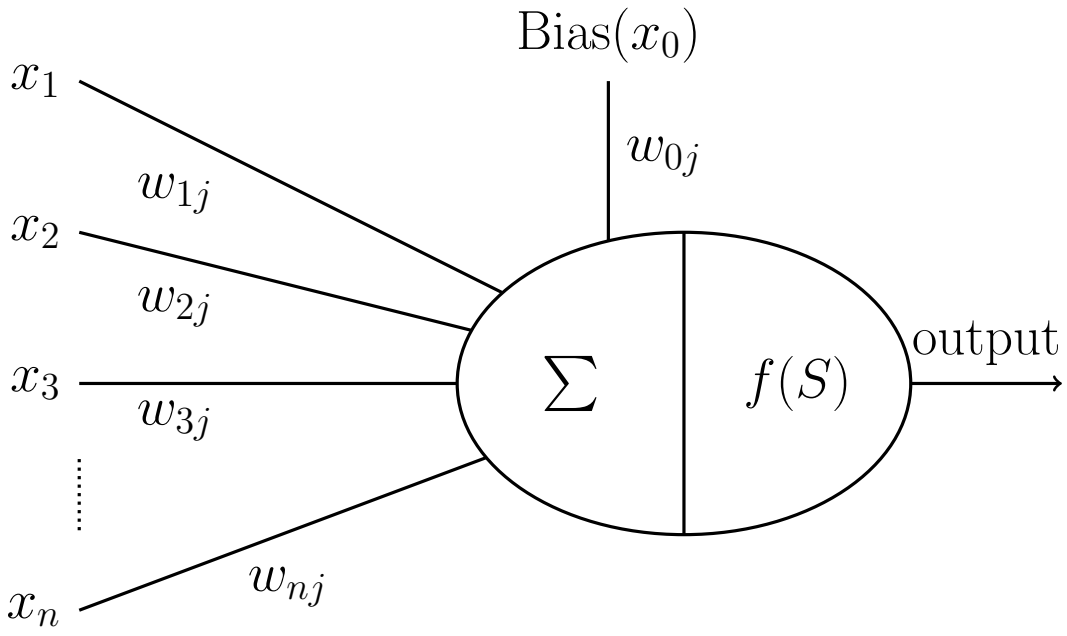
2.2 Strukturen bag neuronerne i et ANN

I den menneskellige hjerne er der som sagt milliarder af forskellige neuroner[1]. I et ANN laver man ikke ligeså mange neuroner som i en rigtig hjerne, da det er umuligt med den teknologi vi har i dag, men det er stadigvæk i fokus at lave et netværk med så mange neuroner som muligt. Derfor er det vigtigt at disse neuroner er matematisk set meget simple, så det er nemt for en moderne computer bare at blæse gennem matematikken i et ANN. Dermed er den strukter man har valgt bag én neuron i et givet ANN vist i figur 1.

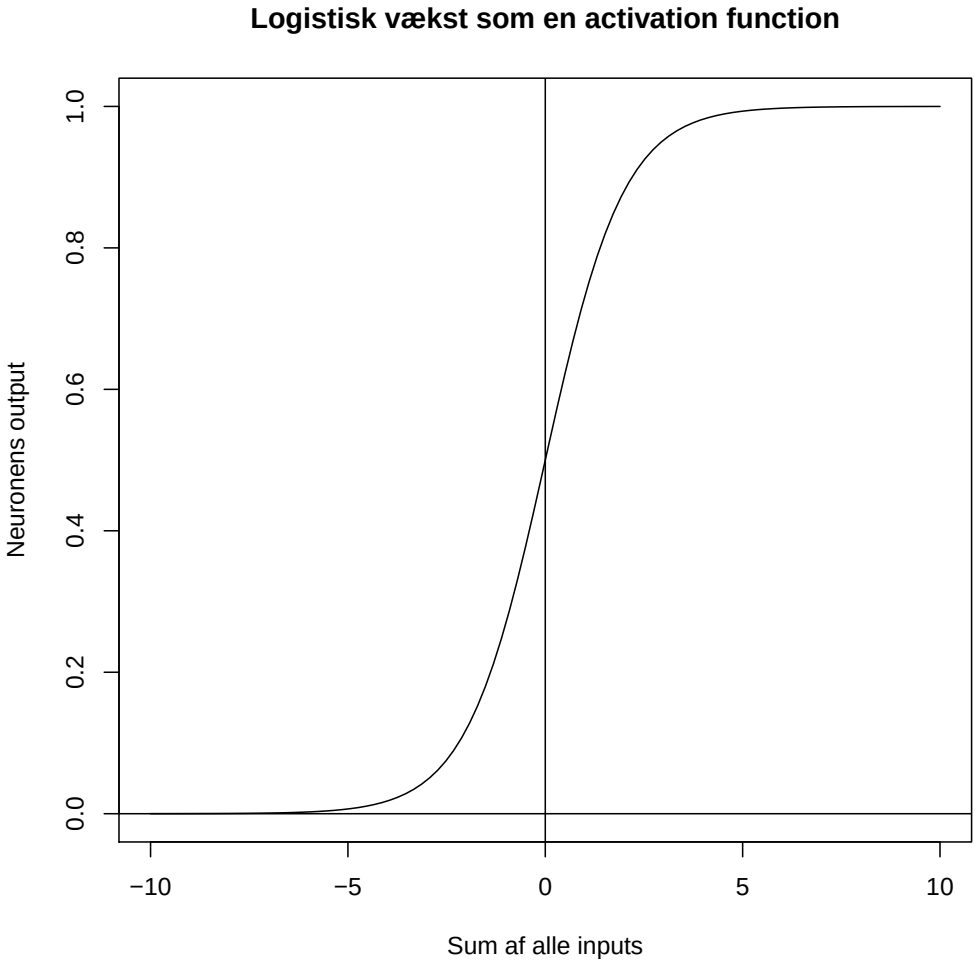
Hvor alle x_n er outputs fra andre neuroner der er tilsluttet denne neuron via synapser, w_{nj} er såkaldte vægte der angiver hvor forstærket signalet er fra outputtet af sidste neuron og bias er en værdi der enten forstærker eller formindsker neuroner Det der så sker inde i neuronen er to ting.
Først bliver alle inputs ganget sammen med deres vægte og summeret op

[1] Jørgen Lützen og Morten Møller 2018.

Figure 1: Model af en neuron



samt biasset bliver lagt til.
Derefter bliver der ført en såkaldt "Activation function" der typisk normaliserer summationen mellem f.eks 0 og 1[2]. Her betragter jeg logistisk vækst med et maksimum på 1 som min activation function, hvor funktionen vil så tage et vilkårligt tal og så spytte et andet tal ud der ligger mellem 0 og 1, jeg betegner funktionen med symbolet σ

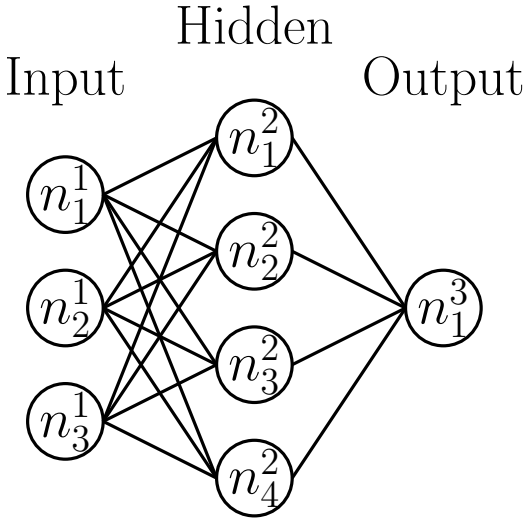


Et ANN består af en hel masse neuroner fra figur 1 der er sammensat som på figur 2

Figur 2 er et eksempel på et meget simpelt netværk der er kendt som et

[2]Ali Kattan, Rosni Abdullah og Zong Woo Geem 2011.

Figure 2: Model af et neuralt netværk



Feed Forward Artificial Neural Network, eller FFANN. Her bliver alle neuroner delt op i forskellige lag, i det her eksempel er der 3, men der kunne også være flere. Det første lag kaldes for "Input layer", det sidste kaldes for "Output layer" og alle imellem kaldes for "Hidden layers". Så er inputtet til hver neuron outputtet fra alle neuroner i sidste lag. Grunden til at alle neuroner i et givent lag så ikke har den samme værdi idet deres værdi afhænger af samme neuroner, er fordi de alle sammen har forskellige vægte. Så f.eks. har første neuron fire forskellige vægte, én til hver af neuronerne i næste lag[3]. Allerede i dette meget simple netværk er det ret rodet i forhold til de små tilslutninger mellem neuronerne. Så derfor er det smart at have en god systematisk navngivning til de forskellige elementer der opgør et neuralt netværk.
Denne opgave bruger følgende navngivning.
Neuroner(Hvor L er hvilket lag neuronen er i, ikke en eksponent, og i er nummeret på neuronen): n_i^L
Vægte har i indekset 2 numre, et nummer for hvilken neuron de er fra, og et nummer for hvilken neuron de skal til, derudover også en indeks der viser

[3]Ali Kattan, Rosni Abdullah og Zong Woo Geem 2011.

hvilket lag de er fra w_{ij}^L, så det kunne f.eks. være w_{11}^1, hvor vægten er fra input laget, og den går fra første neuron i input laget til første neuron af hidden laget.
Biasser har samme navngivning som en neuron bare med et b i stedet.
Derfor kan værdien af en given neuron opskrives som

$$n_j^L = \sigma\left(\sum_i^{m_{L-1}} n_i^{L-1} \cdot w_{ij}^{L-1}\right)$$

Da alle neuroner bliver representeret af et tal, kan et lag skrives som en vektor, hvor hvert koordinat af vektoren er en neuron i laget, f.eks.

$$\vec{H} = \begin{pmatrix} n_1^2 \\ n_2^2 \\ n_3^2 \\ \cdot \\ \cdot \\ \cdot \\ n_n^2 \end{pmatrix}$$

Her er det hidden lag opskrevet som en vektor.

2.3 Matematikken bag synapser i et ANN og fejlfunktion

Ideen om at træne et ANN kommer fra at man gerne vil finde de helt rigtige værdier til de forskellige vægte i netværket der gør at man får et så præcist resultat som muligt. For at kunne vide hvordan man skal ændre på vægtene skal man vide hvor langt fra det korrekte svar man er, og man skal have nogle værdier for vægtene allerede. For at finde ud af hvor langt man er fra et korrekt svar bruger man en såkaldt fejlfunktion. Man har nogle datasæt med kendte inputs og outputs, hvor man så tester inputsne og ser hvordan netværkets outputs er i forhold til de korrekte outputs. Her bruger man så fejlfunktionen til at vurderer hvor langt fra det korrekte svar man er. Der er flere forskellige fejlfunktioner, men i denne opgave vil jeg bruge følgende fejlfunktion

$$f = \sum_{p=1}^{P} \sum_{i=1}^{S} (t_i^p - o_i^p)^2$$

hvor
f er en værdi der angiver fejl
P Antallet af datasæt til træning af netværket
S Antallet af output-neuroner
t Er "target" værdien for en neuron, dvs. den korrekte værdi
o Er værdien fra netværket[4]

2.4 Backpropagation og Gradient descent

For at træne et netværk kigger man på outputtet og arbejder tilbage i netværket for at se hvad man kan gøre bedre for at få et bedre resultat. Det er hele ideen med Backpropagation og den underlæggende algoritme, Gradient descent. Fejlfunktionen er en vigtig del af puslespillet når det kommer til at træne netværket, da man prøver at minimere netværkets fejl, dvs. finde et minimum i fejlfunktionen hvor værdien der angiver fejl er så lille som mulig. Fejlfunktionen er en funktion der tager alle neuroner, vægte og biasser som input, og spytter et tal ud. Det er derfor ikke praktisk at prøve at finde fejlfunktionens monotoniforhold, i stedet finder man fejlfunktionens "gradient". Gradienten er en vektor der viser hvilken retning hældningen er højest i et givet punkt på en funktion. Gradienten for en flervariabelfunktion findes ved at kombinere alle partielle afledte fra funktionen. f.eks. funktionen $f(x, y)$ til have de to partielle afledte

$$\frac{\partial f}{\partial x} \quad \text{og} \quad \frac{\partial f}{\partial y}$$

For at kombinere dem til en vektor ganges de ind på deres enhedsvektorer($\vec{i}$ og $\vec{j}$) og så lægger man dem sammen.[5]

$$\nabla f(x, y) = \frac{\partial f}{\partial x}\vec{i} + \frac{\partial f}{\partial y}\vec{j} = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

For at minimere fejlfunktionen tager man et skridt i den negative retning af gradienten, dvs. den vej hvor hældningen er mindst. Hvis man bliver ved med at gøre det kommer man tættere og tættere på et minimum. Man

[4] Ali Kattan, Rosni Abdullah og Zong Woo Geem 2011.
[5] Wikipedia 2013.

kommer højst sandsyndeligt til at finde et lokalt minimum, da man ikke søger hele funktionen, men man bare tager skridt nedad i det lokale område man nu befinder sig i i funktionen. Dette kaldes gradient descent, fordi man stiger ned af gradienten.

Fejlfunktionen og dermed gradienten er defineret ud fra hele sættet af træningsdata. Det er ikke en smart måde at træne netværket ved at gå alt dataen igennem, da det bare er for intensivt for computeren. I stedet deler man træningsdata op i forskellige minisæt og træner netværket ud fra et minisæt af gangen. Det betyder at den gradient man finder ikke er den helt præcise, da man ikke har udregnet den ud fra alt træningsdataet, men den peger nogenlunde i den rette retning. I modsætning til gradient descent, kaldes dette for stochastic gradient descent når man deler træningsdataet op inden man træner netværket.[6]

2.5 Udledning af træningsformler

Hvis vi betragter det neurale netværk i figur 2, ser vi at værdien af outputtet er givet ved

$$z_j^L = n_1^{L-1} \cdot w_{1j}^{L-1} + n_2^{L-1} \cdot w_{2j}^{L-1} + n_3^{L-1} \cdot w_{3j}^{L-1} + n_4^{L-1} \cdot w_{4j}^{L-1} + b_j^{L-1}$$

$$n_j^L = \sigma(z_j^L)$$

(Jeg deler det op i to, hvor z_1^3 er summen af alle neuroner ganget med deres vægte) Og fejlfunktionen(Hvor y er der korrekte svar)

$$f = \frac{1}{a_L} \cdot \sum_{j=1}^{a_L} (n_j^L - y_j)^2$$

Hvor a_L er antallet af neuroner i output laget.
Hvis vi så vil finde vægten w_{ij}^{L-1}'s indflydelse på fejlfunktionen, skal vi finde dens partielle afledte, dvs.

$$\frac{\partial f}{\partial w_{ij}^{L-1}}$$

Jeg har 3 funktioner der afhænger af hinanden, dvs. en kombineret funktion, så jeg skal bruge kædereglen. Kædereglen siger at

$$(f(g(x)))' = f'(g(x)) \cdot g'(x) \Leftrightarrow \frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}$$

[6] 3Blue1Brown 2017b.

Jeg benytter kædereglen og finder

$$\frac{\partial f}{\partial w_{ij}^{L-1}} = \frac{\partial f}{\partial n_j^L} \cdot \frac{\partial n_j^L}{\partial z_j^L} \cdot \frac{\partial z_j^L}{\partial w_{ij}^{L-1}}$$

Jeg kan så udregne de tre partiel afledte

$$\begin{aligned}
\frac{\partial f}{\partial n_j^L} &= \left(\sum_{j=1}^{a_L} (n_j^L - y_j)^2 \right)' = 2(n_j^L - y_j) \\
\frac{\partial n_j^L}{\partial z_j^L} &= (\sigma(z_j^L))' = \sigma(z_j^L) \cdot (1 - \sigma(z_j^L) \\
\frac{\partial z_L^j}{\partial w_{ij}^{L-1}} &= (n_1^2 \cdot w_{11}^2 + ... + b_1^2)' = n_i^{L-1}
\end{aligned}$$

Nu kan jeg substituere dem tilbage i den anden formel

$$\frac{\partial f}{\partial w_{ij}^{L-1}} = 2(n_i^L - y) \cdot \sigma(z_j^L) \cdot (1 - \sigma(z_j^L)) \cdot n_i^{L-1}$$

Det er en formel for vægtene mellem sidste lag og andet sidste lag. Biasserne opfører sig som en vægt på en neuron der altig er 1 og derfor udregnes på samme måde.
Formlen for ændring i f med hensyn til en neuron i sidste lag minder meget om den for vægtende, bare at neuronen har en vægt til alle outputs, og derfor har indflydelse på fejlfunktionen gennem alle outputs, og derfor skal de summeres.

$$\begin{aligned}
\frac{\partial f}{\partial n_i^{L-1}} &= \sum_{j=1}^{a_L} \frac{\partial f}{\partial n_j^L} \cdot \frac{\partial n_j^L}{\partial z_j^L} \cdot \frac{\partial z_j^L}{\partial n_i^{L-1}} \\
&= \sum_{j=1}^{a_L} 2(n_i^L - y) \cdot \sigma(z_j^L) \cdot (1 - \sigma(z_j^L)) \cdot w_{ij}^{L-1}
\end{aligned}$$

Ideen bag ordet "Backpropagation" kommer så når man skal finde ændringen i f med hensyn til en vægt der ligger længere tilbage i netværket. Hvis vi

betragter laget $L-2$ ved vi at

$$
\begin{aligned}
z_j^{L-1} &= n_1^{L-2} \cdot w_{1j}^{L-2} + n_2^{L-2} \cdot w_{2j}^{L-2} + n_3^{L-2} \cdot w_{3j}^{L-2} + n_4^{L-2} \cdot w_{4j}^{L-2} + b_j^{L-2} \\
n_j^{L-1} &= \sigma(z_j^{L-1}) \\
\frac{\partial z_j^{L-1}}{\partial w_{ij}^{L-2}} &= n_{ij}^{L-2} \\
\frac{\partial n_j^{L-1}}{\partial z_j^{L-1}} &= \sigma(z_j^{L-1}) \cdot (1 - \sigma(z_j^{L-1})
\end{aligned}
$$

Så vi finder ændringen i f med hensyn til en vægt i laget $L-2$ ved

$$\frac{\partial f}{\partial w_{ij}^{L-2}} = \frac{\partial f}{\partial n_j^{L-1}} \cdot \frac{\partial n_j^{L-1}}{\partial z_j^{L-1}} \cdot \frac{\partial z_j^{L-1}}{\partial w_{ij}^{L-2}}$$

Og dette gør man for hvert lag man har i sit netværk og på den måde finder gradienten af fejlfunktionen.[7]

2.6 Fordele og ulemper ved et ANN

2.6.1 Fordele

Grundet det store antal af synapser og neuroner i et netværk, gør det ikke så meget hvis noget af ens data bliver tabt eller ændret på en eller anden måde da hvis det kun er nogle få så kommer det ikke til at ændre outputtet særlig meget.
ANN'er er forfærdelig gode til at finde mønstre og systematik i data der ville være umuligt eller tage lang tid for et menneske at se.

2.6.2 ulemper

Den største ulempe ved et ANN er at den kun giver et svar, og man ikke rigtig kan undersøge hvordan netværket fandt frem til svaret
Andre ulemper indeholder at man skal have en relativt stor database af træningssæt for at kunne lave et godt netværk og det kan være meget svært at tage noget virkeligt og lave det om til data som netværket kan læse.[8]

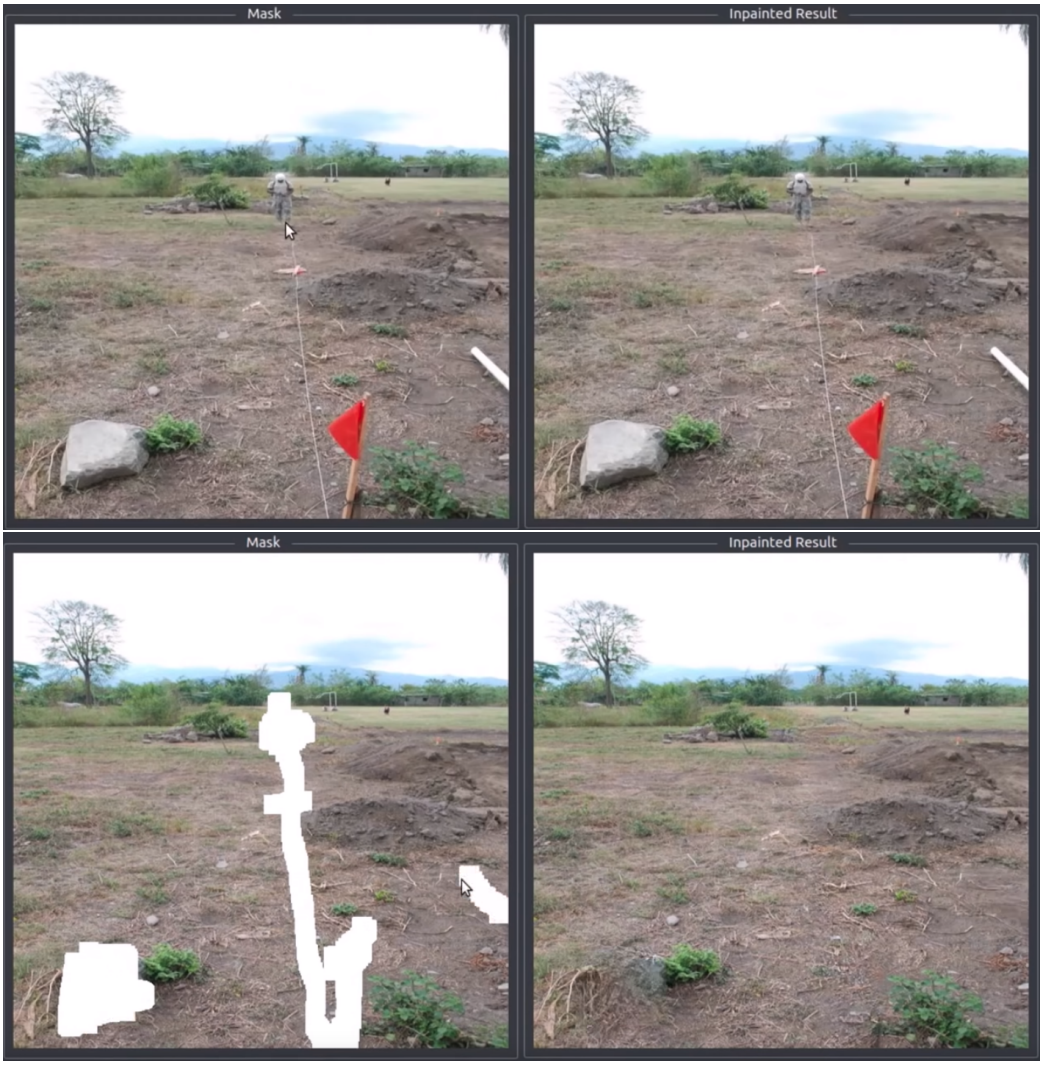
[7] 3Blue1Brown 2017a.
[8] Mijwel 2018.

3 Eksempel på et Neuralt Netværk

Matematikken og strukturen jeg har introduceret i sidste afsnit er noget gammelt teknologi som for længst er gået af mode. Et hold af forskere fra firmaet Nvidia har brugt en form for et Neuralt Netværk, der hedder et Convolutional Neural Network (CNN), som i stedet for lag der bare består af en vektor af neuroner, er det lag med en hel række filtre der hver ekstrahere en eller form for information om et billede. Men ligesom et normalt ANN finder computeren selv ud af hvilke filtre den skal bruge til at opnå det bedste resultat muligt.[9]

Det deres netværk så kan gøre er at den kan tage et billede hvor der er noget der er blevet overmalet med hvid, og så rekonstruere det den mener der er højest chance for at kunne være der hvor det hvide er. Resultaterne viser de i en video hvor man kan se hvor effektive sådan nogle neurale netværker er til at finde mønstre. Selvom der ikke er noget menneske der ville kunne gøre det samme som dette netværk på samme korte tid med samme effekt. Så er det vigtigt at fastslå at dette netværk stadig er helt vildt meget simplere end den menneskelige hjerne. Selvom vi er kommet langt med udviklingen af Neurale Netværk, så er vi stadig ikke i samme boldgade som en ægte hjerne. Men trods det er det stadig et meget forbløffende resultat forskerne fra Nvidia er kommet frem til.[10]

[9] Nvidia 2018a.
[10] Nvidia 2018b.



De to billeder er screenshots fra linket: `https://www.youtube.com/watch?v=gg0F5JjKmhA`, sidst besøgt den 20/12-2018.

De filtre CNN'erne bruger er noget der hedder et kernel eller en convolution matrix, indenfor billedredigering. Den måde det virker på er at man har en matrix af en eller anden størrelse(3x3, 5x5...) hvor der så er en række tal i. Derefter "scanner" man så hele billedet hvor man ligger matrixen oveni billedet og der hvor midten af matrixen rammer, udregner man summen af alle pixels rundt om, ganget med den værdi der er i matrixen hvor pixelen er. Ud fra de forskellige værdier man har i matrixen kan man så få den til at finde kanter, finde cirkler, finde områder med samme farve, sløre billedet,

gøre det skarpere osv. og det er disse værdier som CNN'erne skal finde.[11]

4 Fremstilling af robotter i Isaac Asimov: "Robot"

Isaac Asimov's fortællinger i novellesamlingen "Robot" finder sted i samme univers og deler karakterer mellem hinanden, bogen handler nemlig om robotpsykologen Dr. Susan Calvin der har en samtale med en journalist fra avisen "Interplanetary Press", hvor hun så fortæller om nogle forskellige episoder eller vanskeligheder hun har oplevet gennem sin karriere i robot-industrien.

Isaac Asimov har til denne bog lavet tre love som han mener er tilstrækkelige for at kunne lave sikre robotter der adlyder mennesker, lovene lyder:

1. En robot må ikke skade noget menneske eller, ved at undlade at handle, tillade at et menneske bliver skadet.

2. En robot skal adlyde de ordrer, et menneske giver den, undtagen hvor disse ordrer ville stride mod den første lov.

3. En robot skal opretholde sin egen eksistens, medmindre denne selvopholdelse er i modstrid med den første eller den anden lov.[12]

De forskellige histiorier har nogenlunde det samme tema, hvor der er et eller problem med en robot, den opfører sig ikke som den skal, og så skal de ud fra robotikkens tre love finde ud af hvorfor robotten opfører sig sådan, og hvordan de kan fikse det så den gør hvad de vil.

Selve robotterne bliver overvejene fremstillet på en positiv måde i teksterne, dette kan ses f.eks. i citatet *"Men sagen er den, at man simpelthen ikke kan skelne mellem en robot og de allerbedste mennesker"*[13], men robotterne bliver i teksterne stadig handlet med respekt, som det sis i citatet *"Fysisk set og i en vis forstand psykisk set er en robot, enhver robot, menneskene overlegen. Hvad gør dem så til slaver? Kun den første lov! Ja, uden den ville den første ordre, du prøvede på at give en robot, føre til, at du blev slået ihjel."*[14] som

[11] Wikipedia 2018a.
[12] Asimov 1973, s. 7.
[13] Asimov 1973, s. 194, l. 5.
[14] Asimov 1973, s. 130, l. 31.

er fra en historie i bogen hvori de har nogle robotter der ikke har hele den første lov "imprentet" i hjernen.

4.1 Rundt i ring

Problemet eller konflikten der opstår i teksten "Rundt i ring" er hvor de har en robot ved navn "Speedy" som var blevet sendt ud på solsiden af merkur, hvor der var en sø af selen som den skulle hen for at indsamle, men Speedy kommer af en eller anden grund ikke tilbage. Da de kiggede på hans position viste det sig at Speedy bare løb rundt i en cirkel rundt om søen. Speedy er en særlig dyr model og derfor er dens tredje lov forstærket, dvs. den har en stor modvilje mod fare, og i nogle tilfælde med en ekstrem stor fare ville det kunne overdøve den anden lov, hvis ordreren ikke var givet med noget særligt tryk der gjorde den ekstra vigtig. Og dette var lige det der skete med speedy. De to videnskabsmænd Mike Donovan og Gregory Powell fandt frem til at der måtte være en særlig fare der befandt sig i midten af søen, som når den var langt væk var det ordreren om at hente selen der var stærkest, men hvis den kom tæt nok på var det speedy selvopholdelse der var stærkest, det resulterede i at robotten bare blev ved et punkt hvor de var lige store og løb rundt i cirkel rundt om søen. Så for at få speedy til at komme tilbage igen skulle Mike og Gregory gøre noget der ville overdøve både den tredje og den anden lov, og det eneste der overdøver begge to er den først lov. Så de var nødt til at sætte sig selv i fare, så speedy ville komme tilbage til sine sanser og komme og rede dem, og derved opretholde den første lov. Den måde de gjorde det på var at Gregory gik ud i solen hvor han som menneske ikke kunne overleve særlig længe, og på den måde kom speedy atter tilbage da den første lov negerer de to andre, og redede ham fra at dø.

Her bliver robotten speedy fremstillet som om den ikke virker i starten, men til sidst finder man så ud af at grunden til at den opførte sig så besynderligt var pga. en menneskefejl idet den ordrer den fik ikke var præcis og fokuseret nok på hvor vigtigt det var for speedy at indsamle selenen. Og derfor kom der et uventet resultet fra robotten, nemlig det at den cirklede rundt om søen i stedet for at indsamle selenen.

4.2 Logik

I *Logik* står Mike og Gregory med en eksperimentel robot der hedder Emsig. Problemet med Emsig er at han er en skeptiker, og f.eks. ikke tror på at Mike og Gregory har samlet ham, men i stedet vil han gennemtænke ting logisk. Efter nogle dage kom Emsig til den logiske konklusion at det ikke var muligt for menneskene at lave et væsen der er dem selv overlegent, og derfor måtte det være et væsen der er mægtigere end Emsig, nemlig herren. Derudover før Emsig nogle robot tilhængere og nærmest skaber en hel kult af robotter der tjener *Herren*. Men det viser sig at selv om Emsig næsten driver de to mænd til vanvid med sin snak om herren og logik, så virker Emsig alligevel som han burde, nemlig til at holde styr på apparaterne ved stationen.

I denne tekst bliver der draget ligheder mellem robotterne og mennesker, ved f.eks. at Emsig er overbevist om at han og alle andre er skabt af *Herren*. Og især med Emsig's citat *"Jeg har koncentreret mig om mine egne tanker de to sidste dage, sagde Emsig, og resultaterne har været overordentlig interessante. Jeg begyndte med den ene sikre præmis, jeg følte mig berettiget til at komme med. Jeg eksisterer selv, fordi jeg tænker"*[15] hvor Asimov nærmest gør grin med René Descartes meget berømte udsagn *"Cogito, ergo sum"*[16], som betyder det samme. Så denne tekst viser hvordan en robots hjerne er meget ligesom et menneskes og derfor hvis en robot lades være alene med sine egne tanker vil den udvikle mange af de samme ting som mennesker har udviklet, dvs. religioner, philosofiske tanker osv.

4.3 Bevis

Bevis finder sted ikke så lang tid før et valg, hvor der er to primære kandidater, Stephen Byerley og Francis Quinn. Quinn kommer til Lanning, der ejer United States Robots and Mechanical Men, med den påstand at hans modkandidat Byerley er en robot. Quinns årsag er fordi at man aldrig ser Byerley spise, sove eller drikke noget. Efter at gå frem og tilbage lidt finder de ud af at de hverken kan bevise eller modbevise påstanden om at Byerley er en robot. Indtil under en offentlig tale fra Byerley, kommer der en person op på scenen og beder Byerley om at slå ham. Hvis han er en robot ville han ikke kunne dette da det går imod robotikkens første lov, men til folks

[15] Asimov 1973, s. 62, l. 25.
[16] Wikipedia 2018b.

overraskelse, så slår Byerley ham rent faktisk. Dog har Calvin i tankerne at hvis personen som kom op på scenen faktisk ikke var en person, men også en robot som var med på den. Ville Byerley godt kunne slå vedkommene selv hvis han var en robot, idet det ikke strider imod første lov. Så i historien får de hverken modbevist eller bevist at Byerley er en robot, med der bliver vist hvor vigtig robotikkens love er.

Det er også fra denne tekst citatet *"Men sagen er den, at man simpelthen ikke kan skelne mellem en robot og de allerbedste mennesker"*[17] som jeg brugte før kommer fra. Idet dilemmaet er at enten er Byerley en robot, eller så er han et meget godt menneske.

4.4 Konklusion på Isaac Asimov: "Robot"

Alt i alt bliver robotterne fremstillet positivt, som noget der sammen med mennesker kan skabe en lysere fremtid, men selvfølgelig ikke uden nogle bump. Derudover bliver de tre love indenfor robotikken fremstillet som værende meget vigtig, og at der ikke kan eller burde skabes nogle former for robotter der ikke har disse tre love brændt dybt ned i hjernen. Idet som Calvin selv siger, det eneste der holder robotterne under kontrol af menneskene er disse tre love, uden dem er der ingen net der holder robotterne fra at indtage deres retmæssige plads i toppen af fødekæden og enten udryde mennesker eller gøre dem til slaver.

5 Asimovs tre love i forhold til ANN

Hvis vi forestiller os en hypotetisk fremtid der minder lidt om Isaac Asimovs univers i Robot, hvor der finder fysiske robotter styret af en form for kunstig intellegens der gør dem bevidste, kunne man forestille sig at hvis ikke der var implementeret en form for arbejdsetik dybt i deres hjerner der gør at de adlyder mennesker, at de måske ikke ville gøre som vi siger idet vi ikke rigtig kan tvinge dem. Da ANN er en form for kunstig intellegens der er baseret på en simpel model af den menneskelige hjerne, kunne man nok godt forestille sig at disse robotter vil have et ANN som deres hjerne. Hvis vi så ignorerer selve det tekniske ved implementationen af de tre robotlove i et ANN, så kunne man forestille sig at de nok ville være tilstrækkelig til at holde en

[17] Asimov 1973, s. 194, l. 5.

kunstil intellegens under kontrol. Problemet ligger i at kunne implemeterer robotlovene ordentligt så det er umuligt for en robot at overtræde dem. Hvis et ANN der f.eks. var beregnet til at fjerne spam-emails og ikke havde de tre love indført, så ville den måske tage det logiske skridt og fjerne alle mennesker idet alt spam er et resultat af mennesker, og hvis man bare gjorde den menneskelige race uddød så vil man også fjerne alt spam. Det lyder måske ret urealistisk, men det er det faktisk ikke, fordi et ANN tænker ikke ligesom et menneske, et ANN finder bare den mest eller tæt på den mest effektive løsning på et problem. Og hvis den mest effektive måde at løse et problem på er ved at fjerne alle mennesker, så er det sådan et ANN vil løse det på. Et ANN har ikke nogen form for sympati eller empati, det er derfor man tager de etiske spørgsmål i forhold til kunstig intellegens så seriøst.

5.1 Københavns Universitet

Diskussionen om etiske regler bag kunstig intellegens er f.eks. blevet taget op af Københavns Universitet, hvor professor Peter Sandøe og lektor Sune Hannibal Holm, vil bruge de næste to års tid på at diskutere med alle forskellige politiske og samfundsmæssige synspunkter hvordan man skal arbejde med kunstig intellegens sikkert, og flette hele den diskussion sammen til en form for etisk kodeks. De forskellige etiske spørgsmål de vil kigge på er f.eks. hvem der har skylden når en selvkørende bil kører galt, om politiet må bruge kunstig intellegens som et hjælpemiddel til at forudsige, hvem der begår kriminalitet osv.[18] Dilemmaet med den selvkørende bil lyder for mange som et vi nok kommer ud for om nogle år når vi engang får selvkørende biler på vejene. Men tværtimod så er det et dilemma som vi faktisk allerede er kommet ud for tidligere i år den 18. marts 2018, hvor en selvkørende bil fra uber påkørte en dame i Arizona, USA og dræbte hende.[19], så det er ikke bare nok at gøre alt hvad vi kan for at gøre disse systemer baseret på kunstig intellegens så sikre som muligt. Men vi er også nødt til at udvikle en hel ny kodeks der siger hvem der er skyld i vilke ulykker når disse systemer tager fejl.

[18] Universitet 2018.
[19] Wakabayashi 2018.

5.2 Robotters rettigheder og fri vilje

I Isaac Asimovs bog Robot, bliver robotterne fremstillet som nogle meget højteknologiske redskaber som mennesker har udviklet. Så som en form for slaver der arbejder for mennesker. Men Singularitets-instituttet for Kunstig Intellegens mener man bør skelne mellem kunstig intellegens med og uden bevidsthed. Og at man skal begynde måske endda at snakke om robot rettigheder for kunstig intellegens med "ægte bevidsthed", idet Asimovs robotlove begrænser robotternes frie vilje og degraderer robotter til menneskenes slaver. Så der er nogle der mener at man ikke burde indprente Asimovs love ind i en bevidst kunstig intellegens og derfor vil indføre en række rettigheder til robotter der minder lidt om menneske- og dyrerettigheder. Problemet med de kunstige intellegenser der er bevidste bliver så at hvis de får rettigheder kan det tage væk fra hvad de egentlig var udviklet til, nemlig at hjælpe mennesker. Men hvis ikke man giver dem rettigheder kan det minde om slaveri og plageri idet robotterne rent faktisk er bevidste.[20][21]

Så alt i alt er de etiske spørgsmål der kommer med kunstig intellegens en politisk glidebane, hvori en forkert beslutning kan resultere i menneskenes undergang, eller en helt ny episode i historiebøgerne om slaveri.

6 Konklusion

For at kunkludere har jeg fundet ud af at strukturen der ligger bag et Artificial Neural Network er en simple version af den menneskelige hjerne med masser af neuroner og synapser, og matematikken bag et netværk eller matematikken brugt til at træne et netværk ikke er særlig kompliceret, men det bliver brugt rigtig mange gange idet netværket består af så mange neuroner og synapser. Jeg har også kigget på et eksempel på et neuralt netværk for at se hvad man egentlig kan bruge det til. Derudover har jeg kigget på den etiske del af kunstig intellegens i form af ANN'er. Og det er meget svært at finde et endeligt svar på hvordan man skal arbejde sikkert med et ANN, da der er så mange forskellige vinkler man kan antage i den debat. Ud over at begrænse ANN'er indenfor f.eks. Asimovs robotlove, er det også vigtigt at have en form for kodeks, hvis nu der skulle gå et eller andet galt med en

[20] Råd 2017.
[21] Wikipedia 2018c.

kunstig intellegens. Det kunne være at en selvkørende bil kørte galt, så kan man finde ud af hvem det er der er skyldig i ulykken. Og ydermere kan man komme ind på om ANN'er langt ude i fremtiden der er komplicerede nok til at være bevidste skal have en form for rettigheder ligesom mennesker, idet hvis ikke de har det kan de minde lidt om slaver. Så ANN'er er et fantastisk værktøj der kan bruges til automatisering af mange ting, men som de udvikler sig mere og mere ind i vores hverdag og bliver klogere, er det vigtig at vi har fået udarbejdet nogle etiske regler for hvor man kan arbejde og bruge ANN'er.

7 Referencer

Bøger

[3] Ali Kattan, Rosni Abdullah og Zong Woo Geem, *Artificial neural network training and software implementation techniques*. 2011.

[4] I. Asimov, *Robot*. 1973.

Artikler og andre online kilder

[1] 3Blue1Brown. (2017). Backpropagation calculus — Deep learning, chapter 4, side: https://www.youtube.com/watch?v=tIeHLnjs5U8 (sidst set 12. dec. 2018).

[2] ——, (2017). What is backpropagation really doing? — Deep learning, chapter 3, side: https://www.youtube.com/watch?v=Ilg3gGewQ5U (sidst set 12. dec. 2018).

[5] Jørgen Lützen og Morten Møller. (2018). Hjerne, side: http://denstoredanske.dk/Krop,_psyke_og_sundhed/Sundhedsvidenskab/Sammenlignende_anatomi_og_fysiologi/hjerne (sidst set 8. dec. 2018).

[6] M. M. Mijwel. (2018). Artificial Neural Networks Advantages and Disadvantages, side: https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel (sidst set 14. dec. 2018).

[7] Nvidia. (), side: https://developer.nvidia.com/discover/convolutional-neural-network (sidst set 20. dec. 2018).

[8] ——, (2018), side: https://news.developer.nvidia.com/new-ai-imaging-technique-reconstructs-photos-with-realistic-results/ (sidst set 20. dec. 2018).

[9] D. E. Råd. (2017), side: http://www.etiskraad.dk/etiske-temaer/optimering-af-mennesket/homo-artefakt/leksikon/isaac-asimovs-robotlove (sidst set 20. dec. 2018).

[10] K. Universitet. (2018), side: https://nyheder.ku.dk/alle_nyheder/2018/12/ku-forskere-skal-udvikle-etisk-kodeks-for-kunstig-intelligens/ (sidst set 20. dec. 2018).

[11] D. Wakabayashi. (2018), side: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html (sidst set 20. dec. 2018).

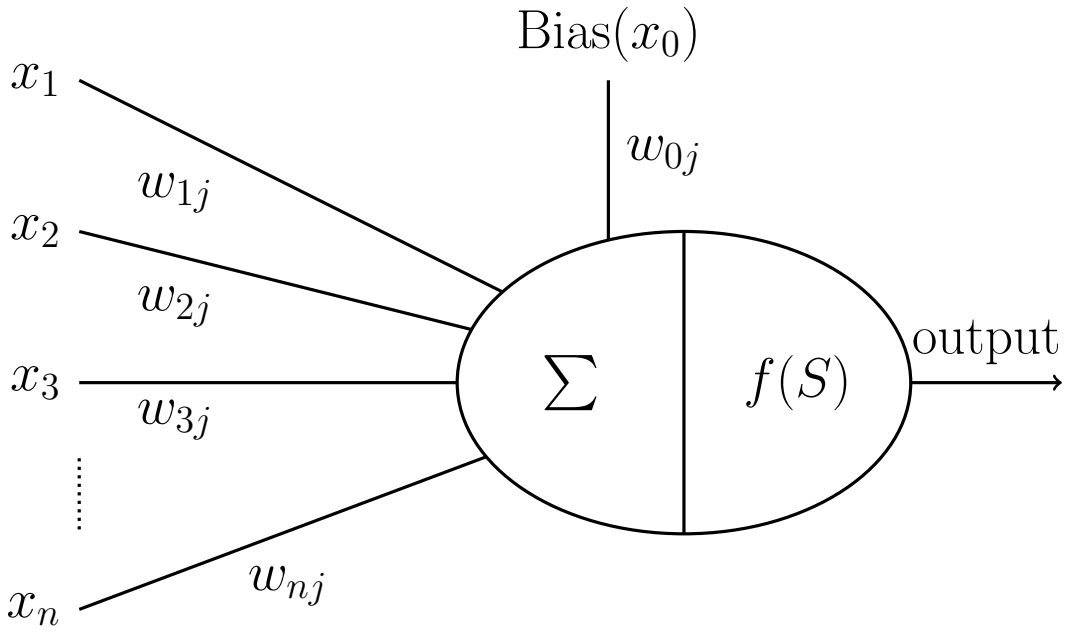
[12] Wikipedia. (2018), side: https://en.wikipedia.org/wiki/Kernel_(image_processing) (sidst set 20. dec. 2018).

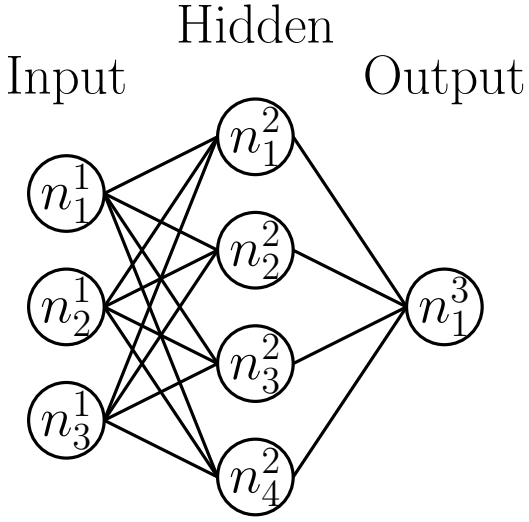
[13] ——, (2018), side: https://en.wikipedia.org/wiki/Cogito,_ergo_sum (sidst set 20. dec. 2018).

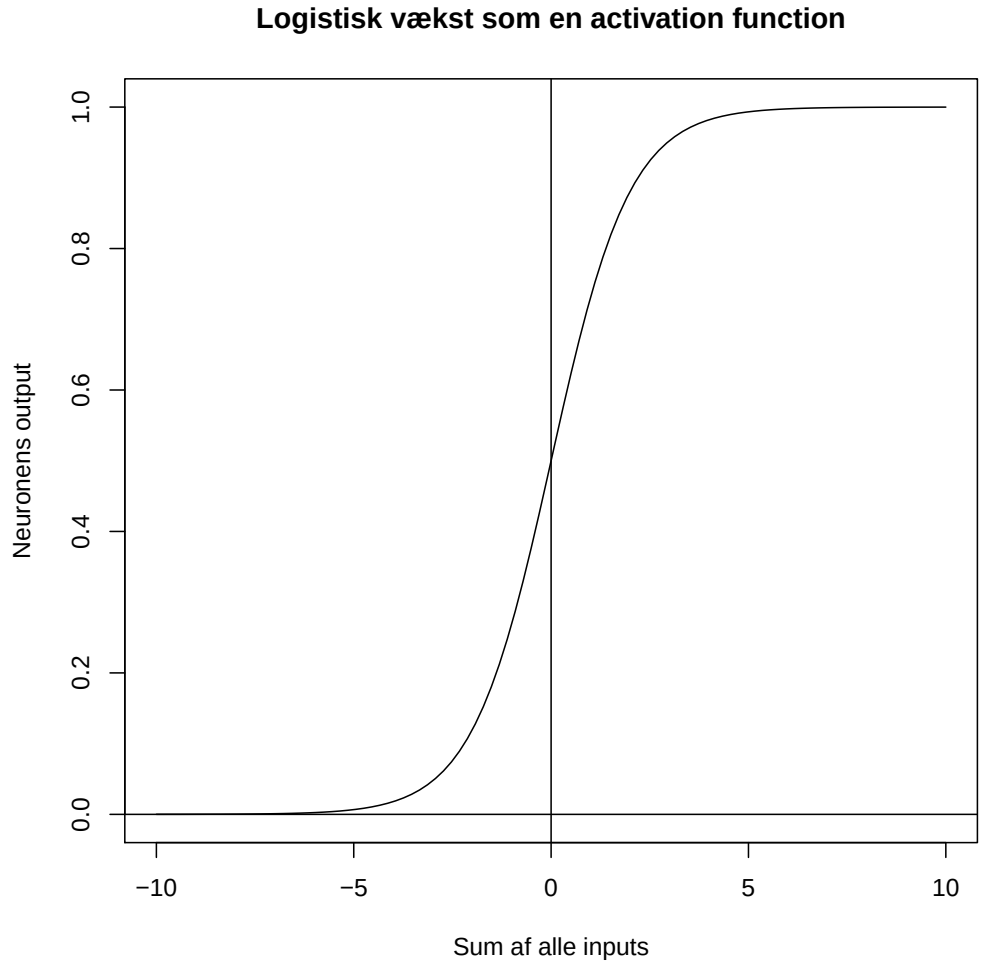
[14] ——, (2018), side: https://en.wikipedia.org/wiki/Ethics_of_artificial_intelligence (sidst set 20. dec. 2018).

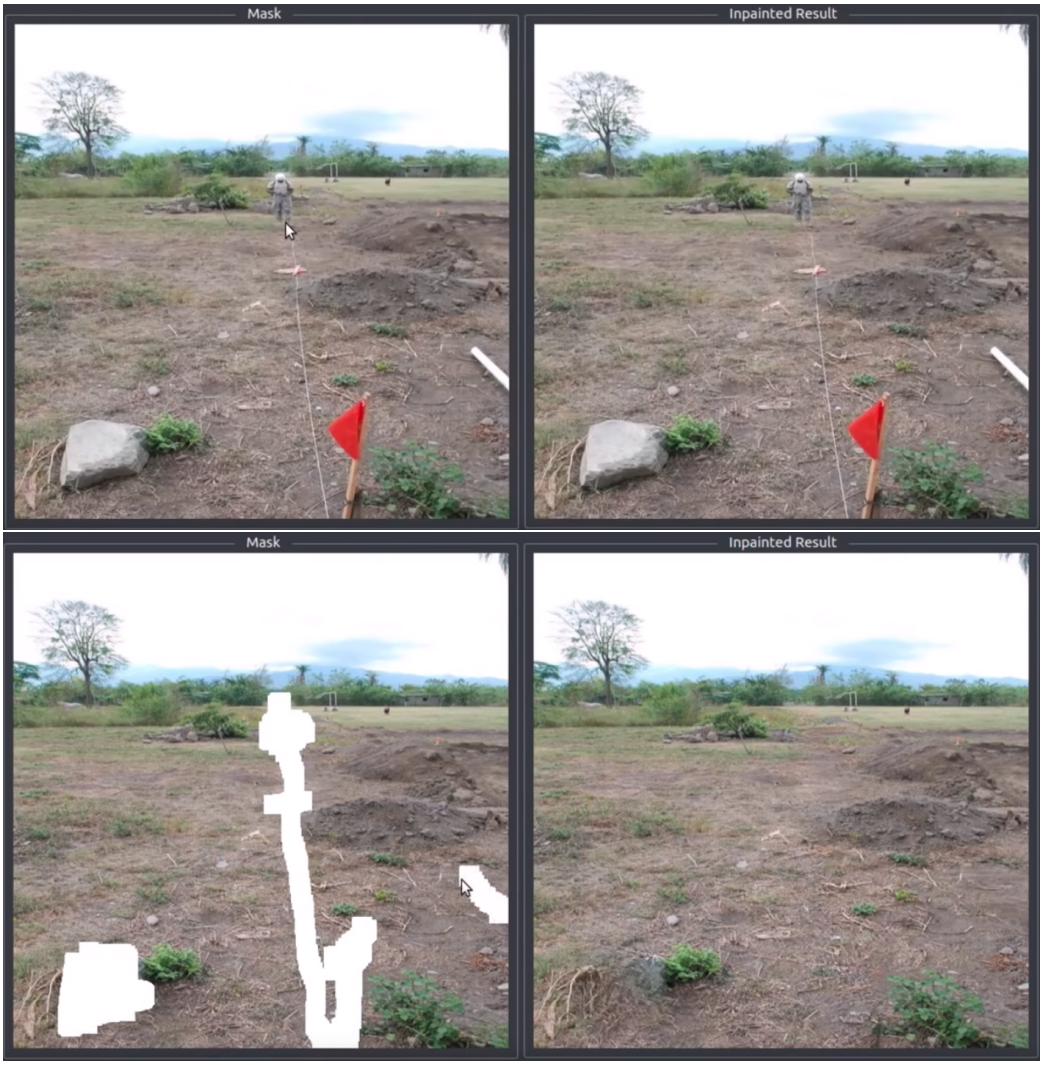
[15] ——, (2013). Gradient, side: https://da.wikipedia.org/wiki/Gradient (sidst set 12. dec. 2018).

8 Bilag


Bias(x_0)
x_1
w_{0j}
w_{1j}
x_2
w_{2j}
output
x_3
$\sum$
$f(S)$
w_{3j}
x_n
w_{nj}


Hidden
Input
Output
n_1^2
n_2^2
n_3^2
n_4^2
n_1^1
n_2^1
n_3^1
n_1^3


Logistisk vækst som en activation function
Neuronens output
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
Sum af alle inputs


Mask
Inpainted Result
Mask
Inpainted Result