AI — Tekniken, Juridiken, Etiken (Paniken)

Fredrik Öhrström (ohrstrom@viklauverk.com)
13 februari 2020

Vad är AI?

April 2018 EU-kommissionens oberoende expert-grupp (AI HLEG).[1]

Artificiell intelligens avser system uppvisar intelligent beteende genom att analysera sin miljö och vidta åtgärder – med viss grad av självständighet – för att uppnå särskilda mål.

AI-baserade system kan vara helt programvarubaserade och fungera i den virtuella världen (t.ex. röstassistenter, bildanalysprogram, sökmotorer, tal- och ansiktsigenkänningssystem), eller inbäddade i hårdvaruenheter, (t.ex. avancerade robotar, självkörande bilar, drönare eller applikationer för sakernas internet).

[1] AI-HLEG (2018). *Artificiell intelligens för Europa.* URL: https://eur-lex.europa.eu/legal-content/SV/TXT/HTML/?uri=CELEX:52018DC0237&from=EN.

AI i EU

Man kan tycka att detta är en generell och intetsägande definition, men Stefan Larsson noterar[2] att definitionen framför allt tar fasta på autonomin, och att exemplen ger en fingervisning om vad de etiska riktlinjerna har för styrningsobjekt.

D.v.s. om en flygande drönare använder AI-relaterade tekniker för att bestämma sin position genom att känna igen geografiska landmärken så är detta ett navigationssystem, inte nödvändigtvis en AI.

Men om drönaren använder AI-relaterade tekniker för att känna igen en stulen bil och därför släppa en röd paintball-kula på dess tak, så är det ett autonomt beslut styrt av en AI.

Det finns en paradox i.o.m. att ordet AI ofta används för ej utvecklad teknik. När tekniken väl kommer i drift får den ett annat namn.

---

[2] Stefan Larsson (2020). "AI i EU". I: *EU och teknologiskiftet*. Santérus Förlag, s. 89–120, s. 93-94.

AI-relaterade tekniker

I denna presentation använder jag termen AI-relaterade tekniker istället för AI, för att poängtera att man använder en teknik för att uppnå ett mänskligt planerat mål/syfte.

(Termen AI är mycket svår att definiera och många försök har gjorts. Jag tänker inte försöka idag.)

En mycket vanlig AI-relaterad teknik är Artificial Neural Networks (ANN).

Tillförlitlig AI — Juridiken/Etiken/Tekniken

De etiska riktlinjer[3] som AI-HLEG sedan publicerade 2019 beskriver tillförlitlig AI som: **lagenlig**, **etisk** och **robust**.

De etiska och robusta kraven skapas utifrån:

- Respekt för människans autonomi
- Förhindrande av skada
- Rättvisa
- Förklarbarhet

Hela dokumentet är intressant, men i denna presentation fokuserar jag på förklarbarhet.

D.v.s. jag vill förklara varför förklarbarhet inom AI är svårt.

---

[3] AI-HLEG (2019). *Ethics guidelines for trustworthy AI.* URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Tillförlitlig AI

Man kan också tala om FAT för att uppnå tillförlitlig AI.

- Fairness
- Accountability
- Transparency

EU-kommissionens white-paper om AI har precis läckt. Men vi kan nog förvänta oss en slutgiltig version inom en snar framtid.

Det kanske blir så att nära förestående AI-specifik lagstiftning innefattar idéer om riskbaserade krav på AI-användare. (accountability)

Mycket har dock redan reglerats i form av GDPR (fairness/transparency)

Men vad *är* AI?

Tekniken för att skapa AI som levererar relevanta resultat baserar sig på logik och statistik.

Logisk, i den mening att mänsklig slutledningsförmåga skapar regler som sedan programvara använder för att härleda slutsatser utifrån premisser och indata.

Resonemang och beslutsfattande

Statistisk, i den mening att man samlar in data och sedan låter programvara skapa regler och kategoriseringar utifrån det insamlade datat. Dessa nyskapade regler används sedan av programvara för att härleda slutsatser från ej-tidigare observerade data.

Maskininlärning

Logik

Om deklarationen lämnas in efter 31 juli, så måste aktiebolaget betala en straffavgift på 5000 kr.

```
if (datetime > 2020-07-31) then penalty_fee = 5000
```

Om antalet bilar på bron är 10 så måste trafikljuset visa rött.

`INVARIANT` $num\_cars \geq 10 \Rightarrow traffic\_light = RED$

Om det juridiska avtalet berör immaterialrättsfrågor så måste avtalet godkännas av vår chefsjurist.

```
if (HAS_PROPERTY(deal, IP_RIGHTS)) then
gc_approval_required = true
```

Men hur ska vi programmera HAS_PROPERTY?

Begränsningar i logiska regler

Vi vet/tror att det inom ett område finns regler som styr något, men vi kan inte skriva ned dessa logiska regler.

- De är för komplexa.
- De är för många.
- Vi vet inte ens hur vi ska formulera dem.

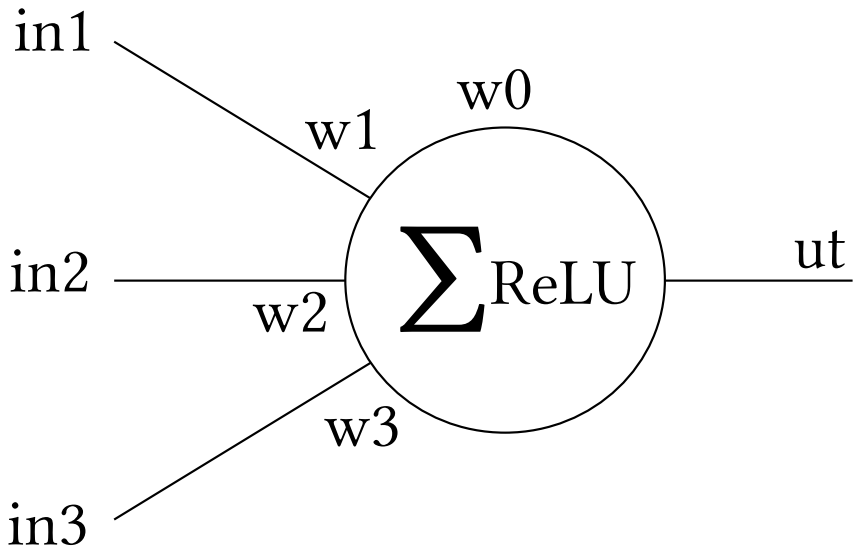
Använd maskininlärning

- Supervised learning: Träna m.h.a. (väl utvalda, etiketterade och helst många) exempel. (hitta tecken på cancer i röntgenbilder, d.v.s. kategorisera bilder i normal vs förhöjd risk)
- Unsupervised learning: Exempel finns, men de är inte etiketterade. (hitta relevanta kategorier och gruppera indata i dessa kategorier, t.ex. word2vec)
- Reinforcement learning: Det finns en återkoppling (som dock kan vara långsam och otydlig) som styr lärandet. (spela dataspel och schack)

Det finns många olika maskininlärningstekniker

- Regressioner: OLSR, linjära, logistic, stepwise, MARS, LOESS
- Instansbaserade: kNN, LVQ, SOM, LWL, SVM
- Regularisering: Ridge regression, LASSO, Elastic Net, LARS
- Beslutsträd: CART, ID3, CHAID, Decision stump, M5, Conditional
- Baysian: Naive, Gaussian, Multinominal, AODE, BBN, BN
- Clustering: k-Means, k-Medians, EM, Hierarchical
- Association: Apriori, Eclat
- Neurala nätverk: Perceptron, MLP, Back-prop, Hopfield, RBFN
- Deep learning: CNN, RNN, LSTM, DBM, DBN
- Dimension reduce: PCA, PCR, PLSR, MDS, LDA, MDA, QDA, FDA
- Ensembles: Boosting, Bagging, Blending, Stacking, Random forest

Artificiella Neurala Nätverk



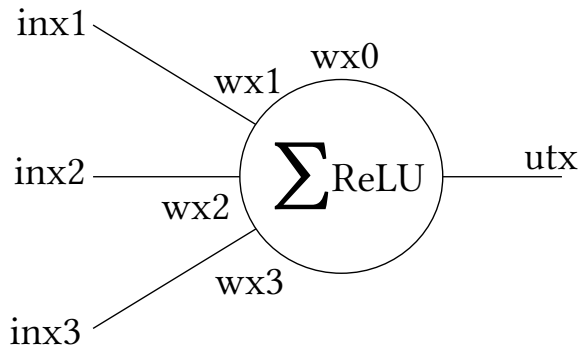
$summa = w_0 + in_1 * w_1 + in_2 * w_2 + in_3 * w_3$
$ut = ReLU(summa)$

```
summa = 0.5 + 7*0.1 + -3*0.5 + 5*0.9
if (summa > 0) then ut = summa else ut = 0;
```

$ut = 0.5 + 0.7 - 1.5 + 4.5 = 4.2$ genom ReLU ger 4.2

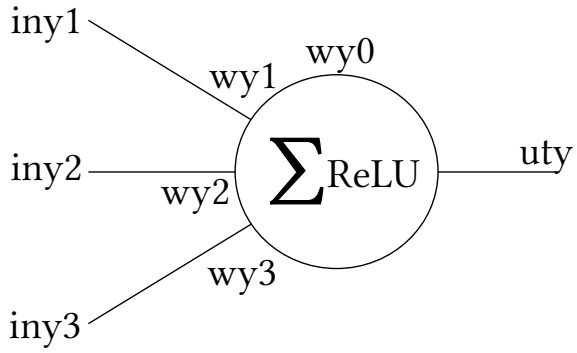
Mål: träna det neurala nätverket genom att justera vikterna w_0 till w_3 så att man får det önskade utdata för givna indata.

Känna igen handskrivna siffror



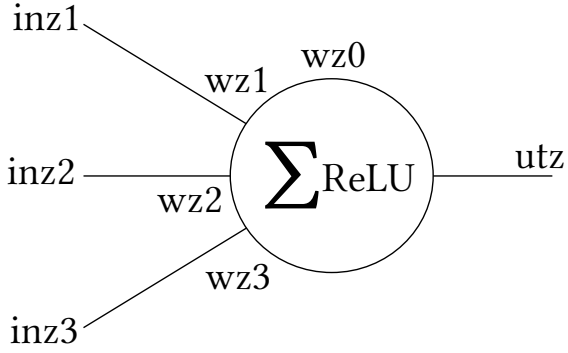
0

AI ser en nolla om utx är nära 1.



1

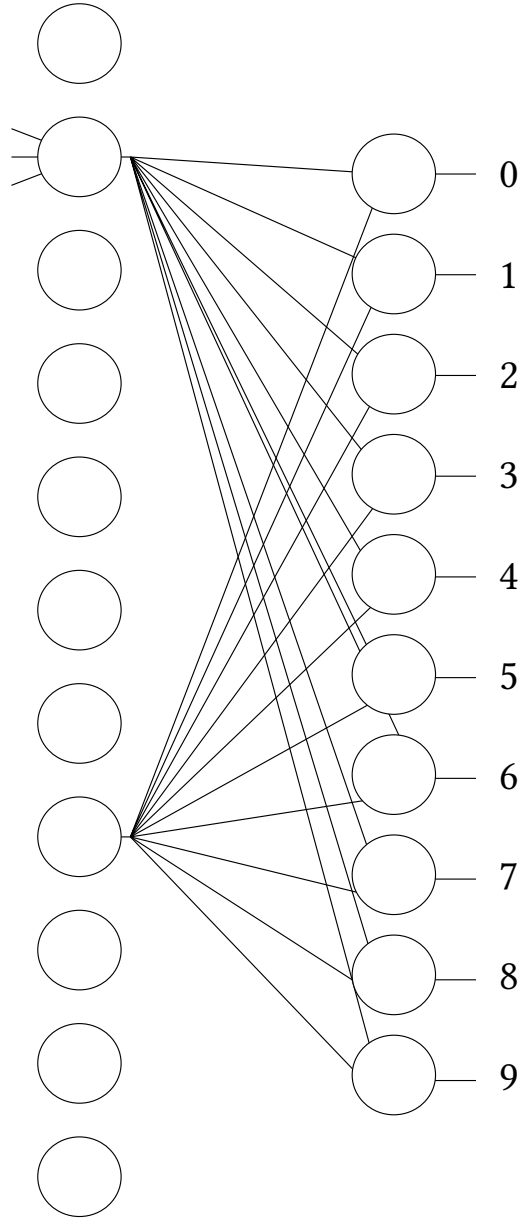
AI ser en etta om uty är nära 1.



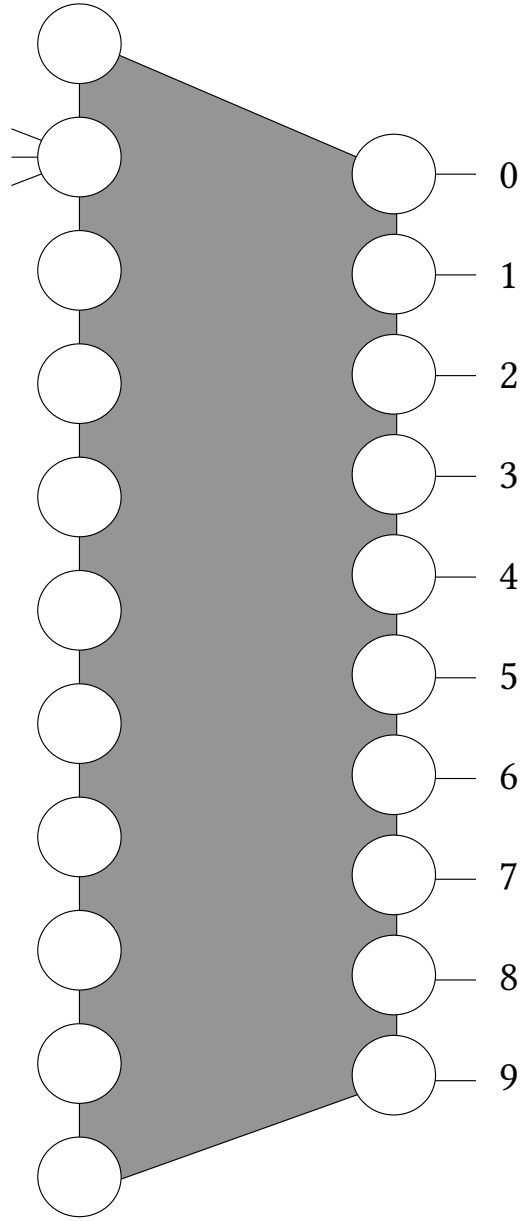
2

AI ser en två om utz är nära 1.

Känna igen handskrivna siffror

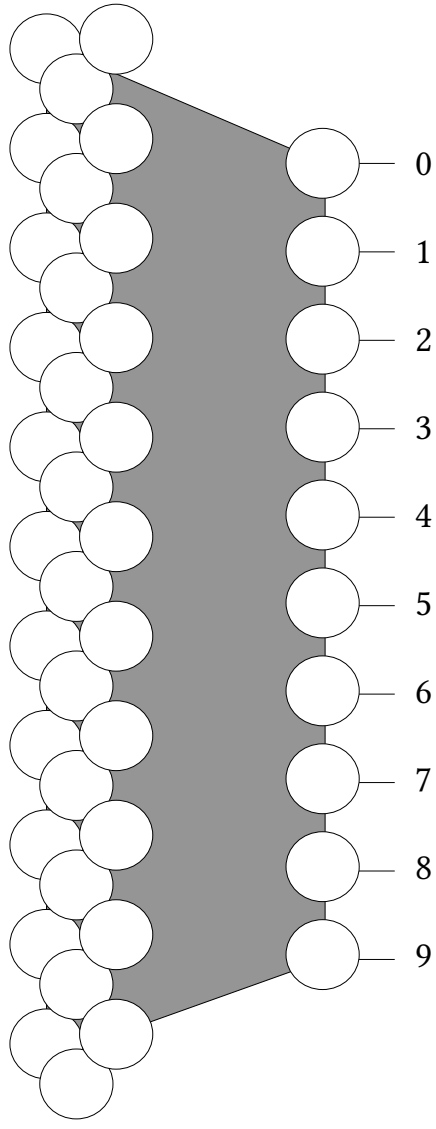


Känna igen handskrivna siffror



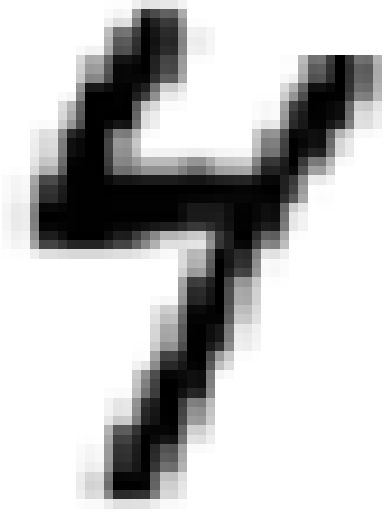
Nu har vi $10 * (11 + 1) = 120$ vikter att träna.

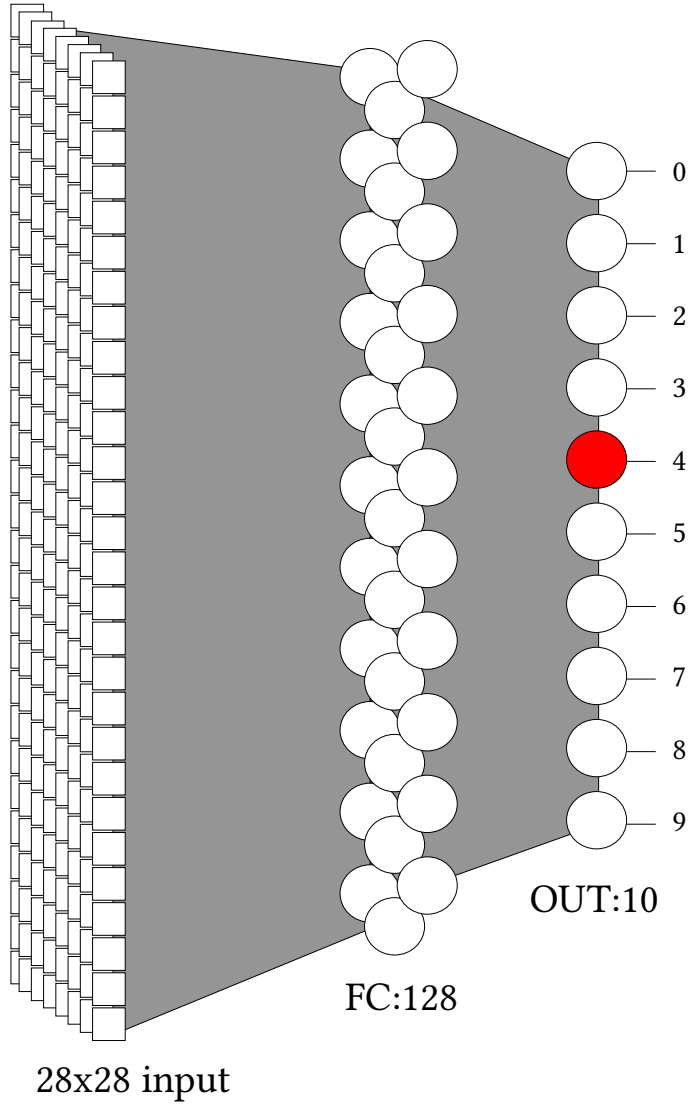
Känna igen handskrivna siffror



$10 * (128 + 1) = 1290$ vikter

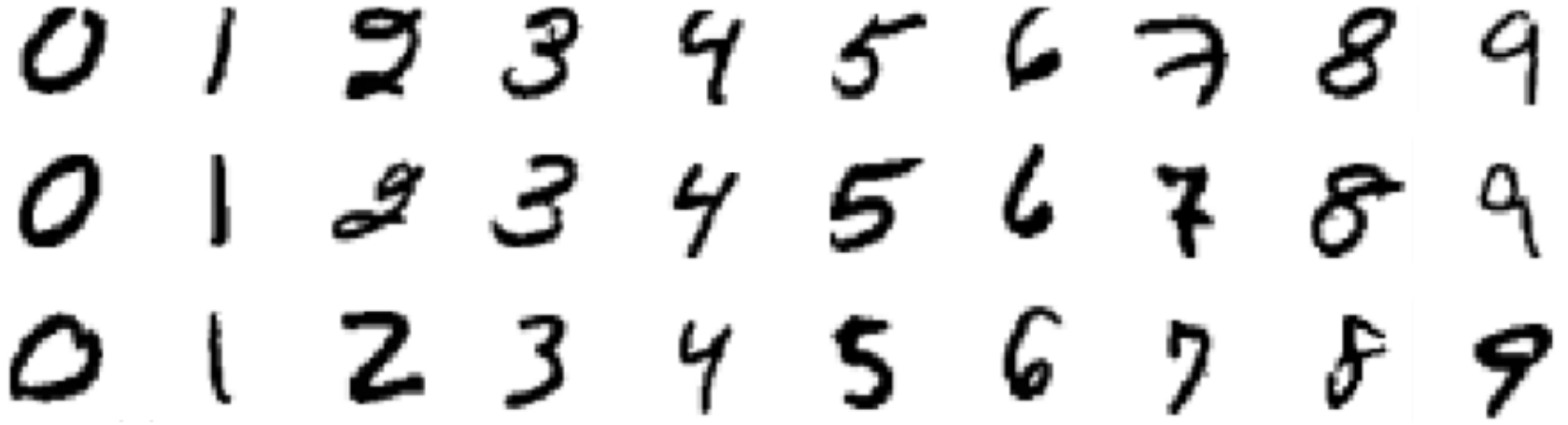
Känna igen handskrivna siffror





$$128 * (1 + 28 * 28) + 10 * (128 + 1) = 101770 \text{ vikter}$$

Träna nätverket



Databas med handskrivna siffror från MINST

Börja med slumpmässiga vikter!

Lägg på en visualisering av en siffra som indata, säg 4.

Önskat utdata är 1 för noden som identifierar fyror.
Önskat utdata är 0 för övriga noder.

Gå baklänges genom nätverket och anpassa vikterna med små steg för att korrigera felet i utdata.

Mycket CPU/GPU-intensivt!

Träna nätverket

Det finns många strategier för att träna nätverket. T.ex. back-propagation med:

- Batch gradient descent
- Stokastic gradient descent
- Mini batch gradient descent
- m.m.

Som i sin tur kombineras med:

- Momentum
- Nesterov accelerated gradient
- Adagrad (adaptive gradient)
- m.m.

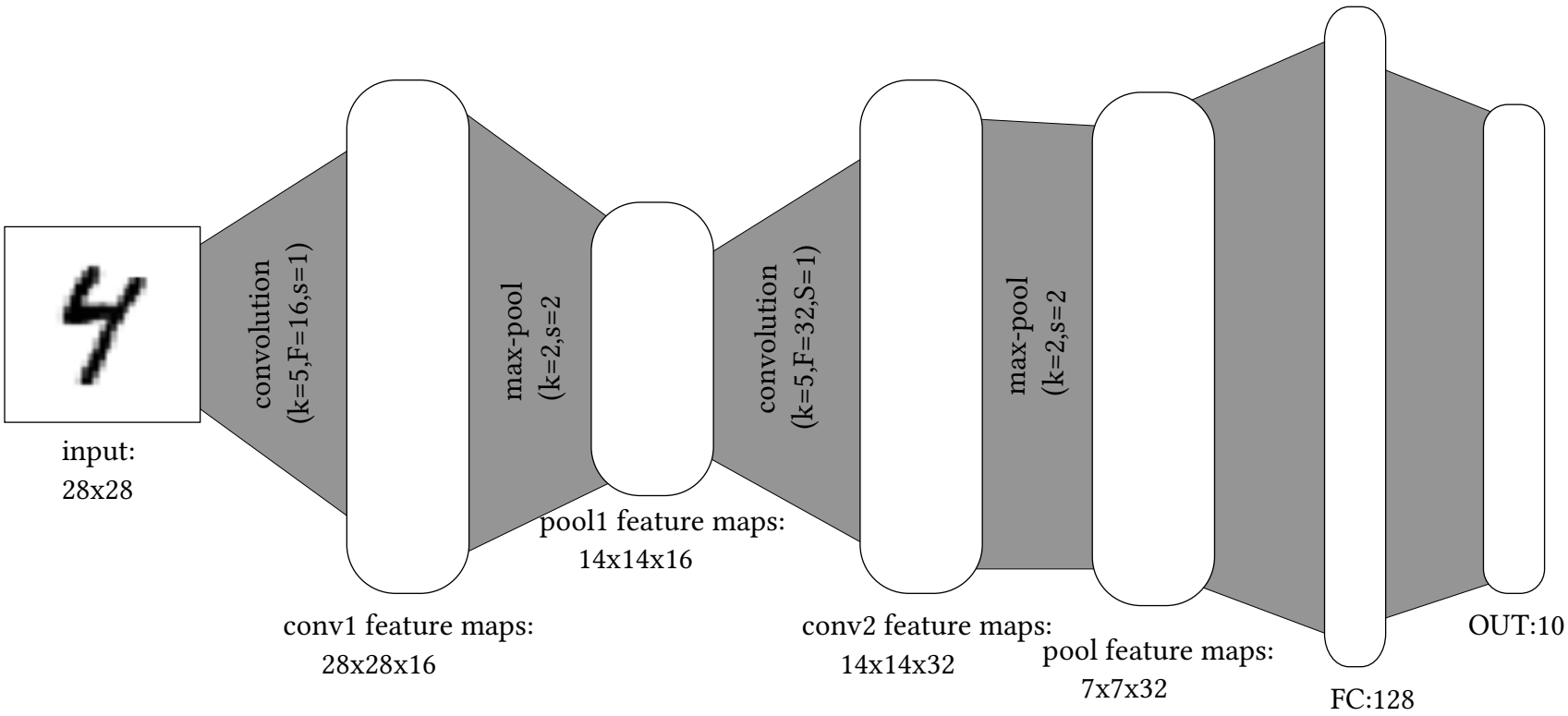
Känna igen handskrivna siffror

Tyvärr klarar inte det tidigare neurala nätverket av skalningar, translationer och rotationer av siffrorna.

Vi behöver mer lager av neuroner.

Lager som kan få specifika uppgifter att identifiera konturer, delar av siffrorna och till slut hela siffran.

Känna igen handskrivna siffror



Vad betyder deep learning?

Ungefär att det finns flera lager i det artificiella neurala nätverket.

D.v.s. olika lager tar ansvar för förståelse av detaljer på olika nivåer.

Hur förklarar vi slutledningen?

Mängden vikter överstiger alla mänskliga försök att få en förståelse för varför den neurala nätverket ger rätt svar.

Vi kan i viss mån förstå vad som händer i de olika lagren därför att vi skapade lagren för olika syften.

Enter Deep Dream.

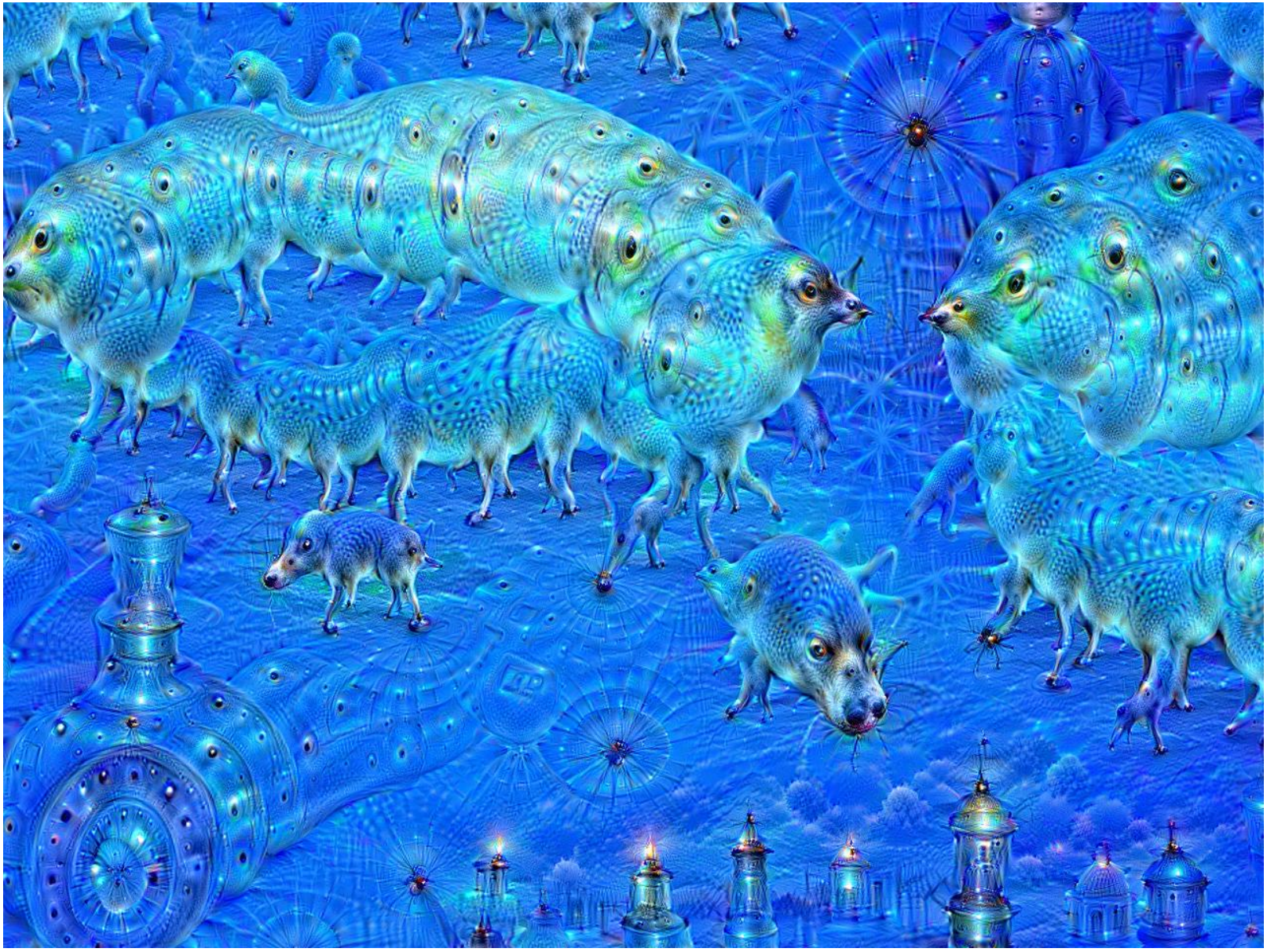
Börja med en bild och modifiera bilden för att få ett redan tränat neuralt nätverket att ge bättre och bättre respons.

D.v.s. om det neurala nätverket är tränat att se hundar, så kommer bilden successivt att innehålla mer och mer hundar.

Deep Dream

av MartinThoma - CC0 wikipedia

Deep Dream



av MartinThoma - CC0 wikipedia

Deep Dream

av MartinThoma - CC0 wikipedia

Deep dream kan i viss mån förklara ett ANN.

Vi kan se att det neurala nätverket har tränats att tro att en människas arm alltid ingår i en hantel.[4]

[4] Alexander Mordintsev, Christoper Olah och Mike Tyka (2015). *Inceptionism: Going Deeper into Neural Networks.* URL: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Var kommer bilderna från?

Vems upphovsrätt ligger bakom bilderna?

- Miljontals bilder på hundar har tränat the neurala nätverket.
- En ursprungsbild har ibland startat inception-processen, men det kan vara slumpmässigt brus.
- En människa har provat olika inställningar t.ex. hur fort modifieringen ska ske, vilka lager i nätverket som ska influera mest.

Varför är det så många hundar?

Det finns färdigklassificerade dataset på hundar och katter.

GAN — Generative Adversial Network

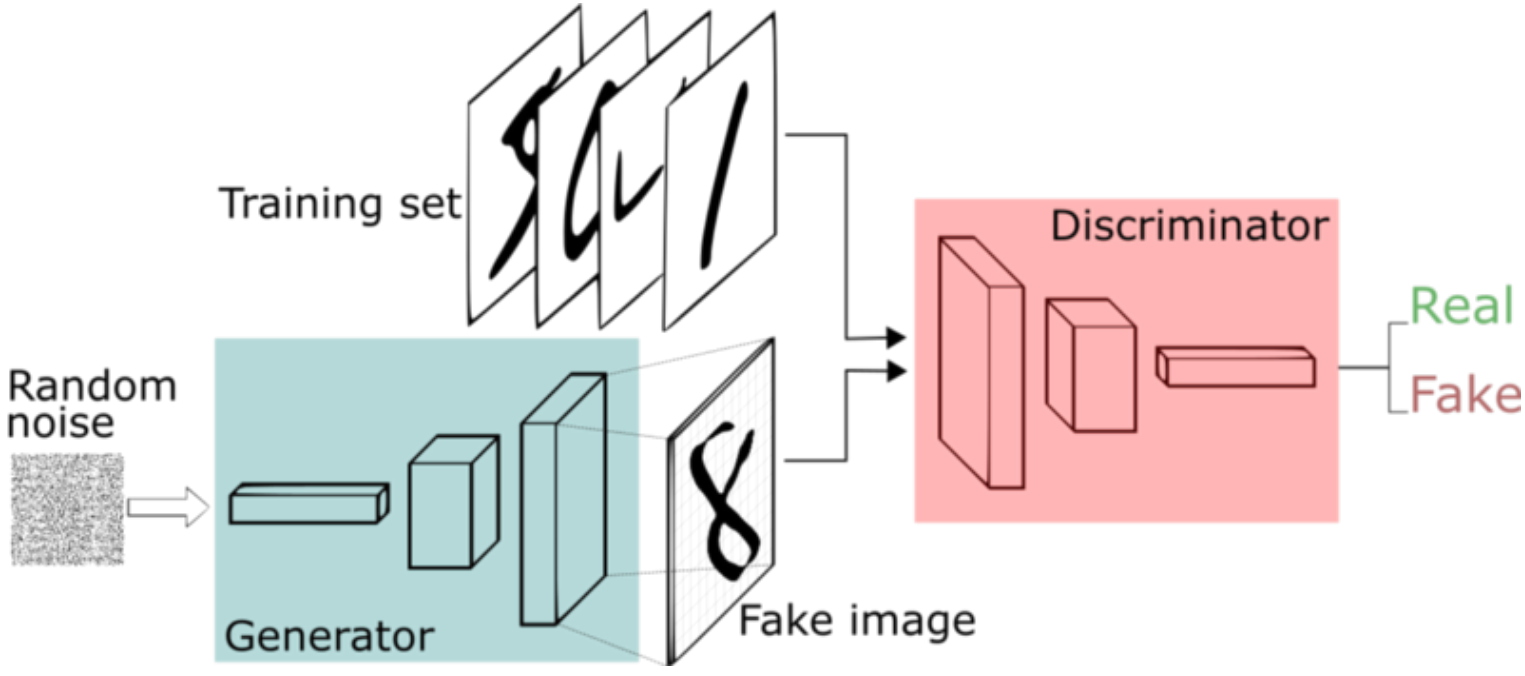


illustration Thalles Silva `https://www.freecodecamp.org`

GAN — Generative Adversial Network

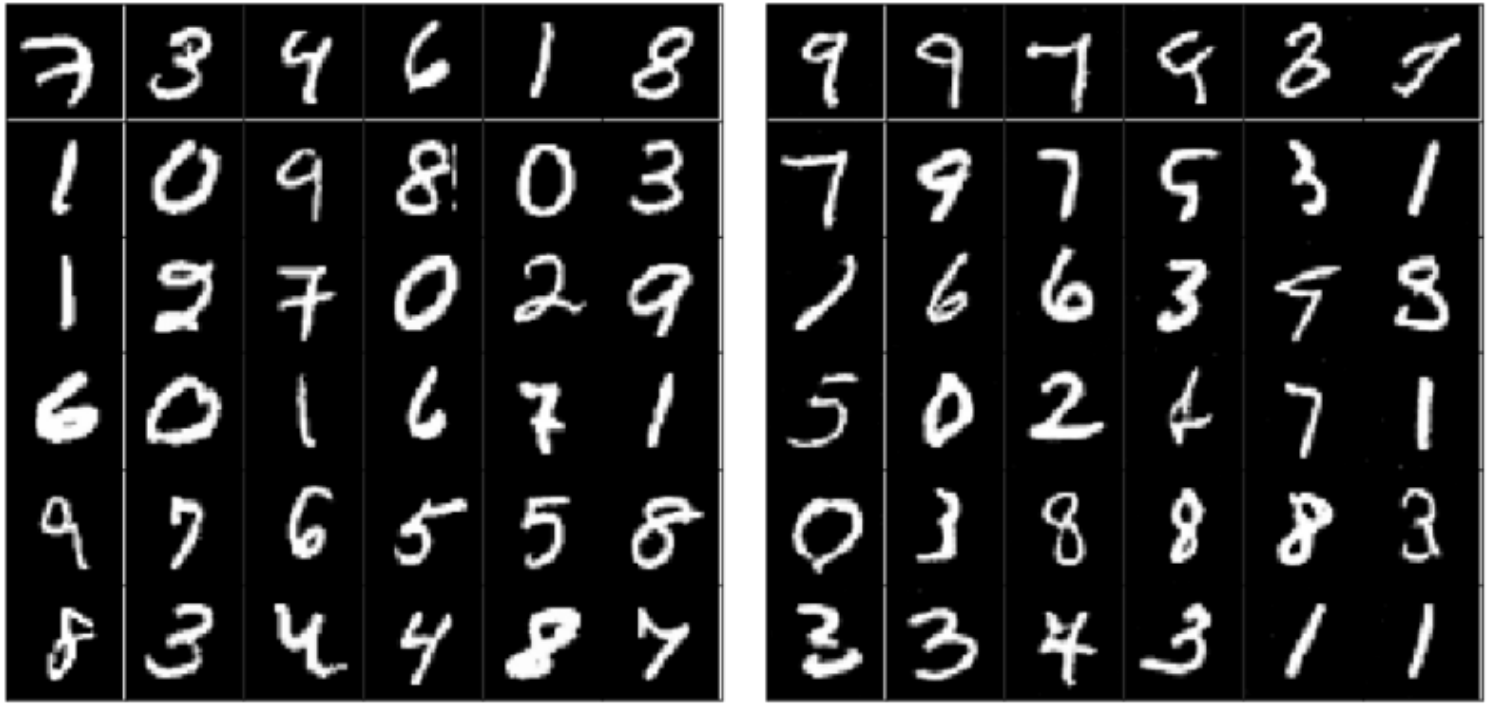


illustration Thalles Silva `https://www.freecodecamp.org`

GAN

https://thiscatdoesnotexist.com/

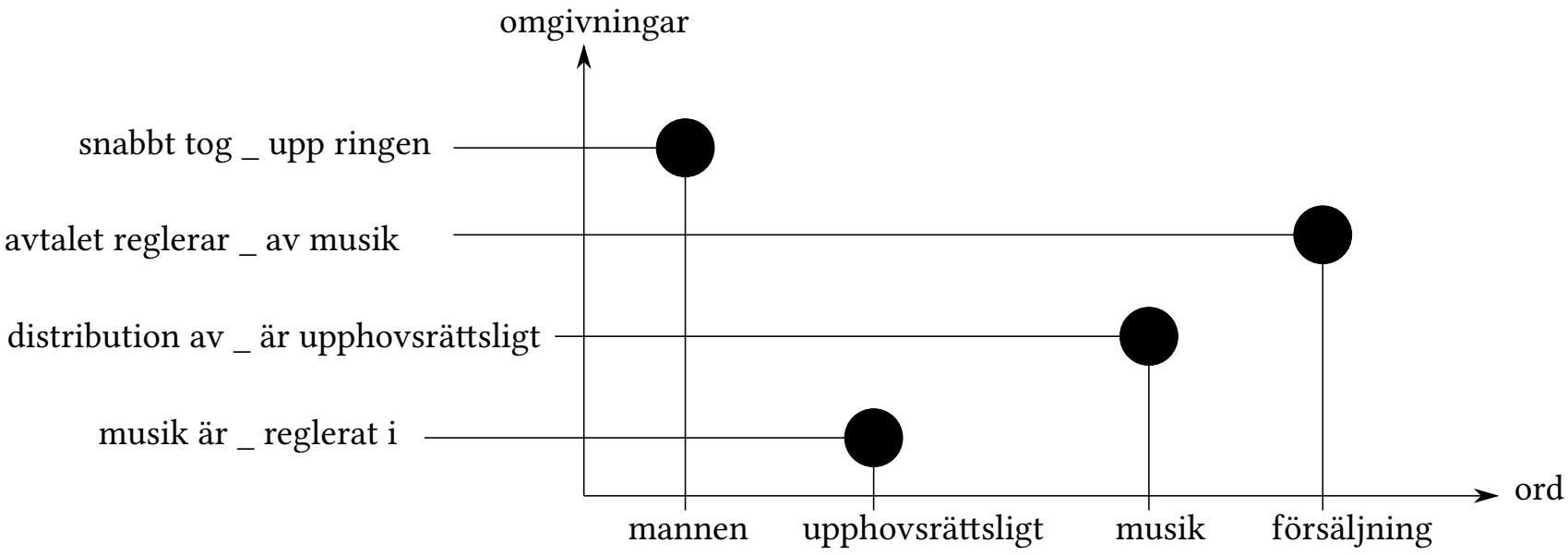
GAN (Paniken?)

`https://www.thispersondoesnotexist.com/`

Källkoden `https://github.com/NVlabs/stylegan2`

Tillbaka till avtal som berör immaterialrätter

Hur ska vi programmera `HAS_PROPERTY(deal, IP_RIGHTS)` ?



Vi har tillgång till en mycket stor korpus med avtal.

Word2Vec

Word2Vec använder ett neuralt nätverk för att skapa vektorer (grupper av siffror) för varje omgivning och för varje ord.

snabbt tog _ upp ringen	(0.84, … ,0.15)
avtalet reglerar _ av musik	(0.16, … ,0.74)
distribution av _ är upphovsrättsligt	(0.17, … ,0.32)
musik är _ reglerat i	(0.30, … ,0.88)
försäljning	(0.46, … ,0.76)
mannen	(0.31, … ,0.31)
musik	(0.27, … ,0.92)
upphovsrättsligt	(0.42, … ,0.99)

Man väljer hur många dimensioner som önskas innan word2vec påbörjar arbetet, t.ex. mellan 40 och 100 dimensioner.

En form av unsupervised learning. Vi vet inte hur vi ska kategorisera orden/omgivningarna i förhand men träningen av ANN:et skapar en kategorisering.

Word2Vec

Den kategorisering som görs m.h.a. av en AI-relaterad teknik visar sig skapa sådana kategoriseringar i vektorerna, så att vanlig aritmetik kan göra analogier!

vector('Paris') - vector('France') + vector('Italy')
resulterar i en vektor som är nära[5] vector('Rome')

vector('king') - vector('man') + vector('woman')
resulterar i en vektor som är nära vector('queen')

`https://code.google.com/archive/p/word2vec/`

[5] Word Cosine Difference

Word2Vec

Kan vi med word2vec skapa `HAS_PROPERTY(deal, IP_RIGHTS)` ?

Ta orden i avtalet och studera deras Word Cosine Difference till uppenbart immaterialrättsliga ord, såsom upphovsrätt, mönsterskydd, patent o.s.v.

Om en vi ser att vissa ord ligger nära, kan vi då dra slutsatsen att avtalet har med immaterialrätt att göra?

Eller ska vi använda ord-vektorerna (som skapats ur vår stora mängd av avtal) som indata till en annat ANN?

De AI-relaterade teknikerna blir snabbt mer och mer komplicerade.

AutoML

AutoML innebär att man med AI-relaterade tekniker (t.ex. genetiska algoritmer) skapar de lager i ANN som vi tidigare manuellt skapade.

AutoML kanske kan användas för att automatiskt skapa den kedja av AI-relaterade tekniker som till slut verkligen kan genomföra `HAS_PROPERTY(deal, IP_RIGHTS)`

Tekniken — Förklarbarheten

- Ibland enkla grunder, men komplexiteten ökar snabbt.
- Slump är delaktig i träningen.
- Träningsdatat kan ibland återskapas från det tränade systemet.
- Är träningsdatat inte fair, så blir kanske inte systemets beslut fair.
- Teknikerna kombineras till kedjor av AI-relaterade tekniker.
- Kedjorna kan också skapas av AI-relaterade tekniker.

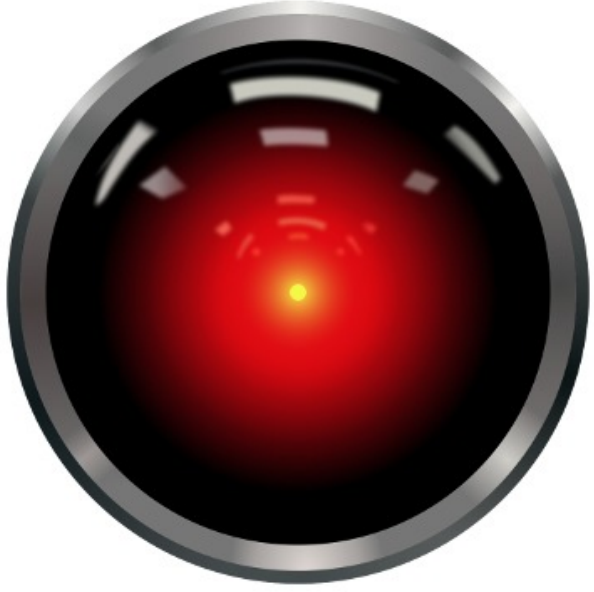
Jag anser att förklarbarheten lägger grunden till tillförlitlig AI.

Än så länge så kan inte de AI-relaterade teknikerna ge en förståelig förklaring till varför de tog ett visst beslut.

Paniken — Förklarbarheten

Två populärkulturella referenser:

Computer says no!



Computer says no!

Patent

- Patent på AI-relaterade tekniker. (CII)
- Patent på tillämpningar av AI-relaterade tekniker. (jmfr simuleringar G 1/19)
- Patent skapade av AI?[6]

[6] Ryan Abbott (2019). *The Artificial Inventor Project.* URL: https://www.wipo.int/wipo_magazine/en/2019/06/article_0002.html.

Avslutande tankar

- Det kan finnas ett stort värde i ett vältränat system. Är vikterna (och strukturen) hos ett tränat ANN skyddat av immaterialrätter eller närstående rättigheter eller inte alls?
- Det som sedan genereras av AI-relaterade tekniker blir det immaterialrättsligt ett kombinerat verk relativt de ingående träningsdata, ett nytt verk eller inte skyddat alls? Blir det också ett kombinerat verk relativt de tränade vikterna? Hur mycket bidrar det mänskliga urvalet och styrningen av den AI-relaterade tekniken till den ev. immaterialrätten?
- Bidrar en eventuell förklarbarhet hos ett ANN till huruvida man kan analysera upphovsrättsliga frågor? D.v.s. kan ett oberoende dubbelskapandet vara synligt i vikterna?

Avslutande tankar

- Vad är definitionen av uppfinnare i patentlagens mening? Liksom slumpmässigt testande av medicinskt relevanta preparat kan leda till en uppfinning, så bör väl slumpmässigt genererande konstruktioner av en AI kunna leda till en uppfinning? Är uppfinnaren den som satt processen i rullning och sedan observerar resultatet och bedömer att det har ett värde?
- Ägaren till ett patent erhåller ett tidsbegränsat monopol i utbyte mot att information görs tillgänglig för samhället. Om informationen är värdefull så spelar det väl ingen roll vem som gjorde uppfinningen?
- Lagstiftningen för upphovsrätt har lämnat exemplarframställning bakom sig och fokuserar på tillgängliggörandet. Framtida AI-relaterad lagstiftning kommer kanske att ha en helt annan infallsvinkel till det skyddsvärda.[7]

[7] Stefan Larsson (2020). "AI i EU". I: *EU och teknologiskiftet*. Santérus Förlag, s. 89–120, s. 97.

Tack för ert intresse!

Fredrik Öhrström (ohrstrom@viklauverk.com)