



Museumcollecties & AI: technische achtergrond

Maarten Schermer

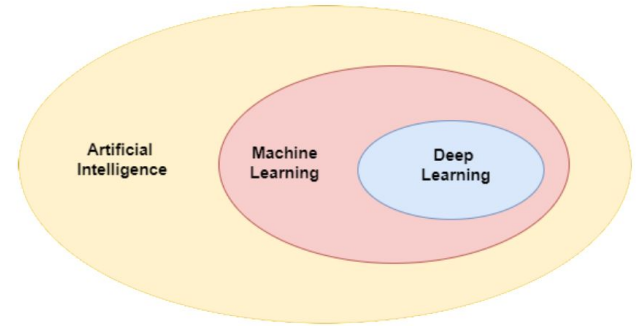
16 mei 2022

Naturalis
Biodiversity
Center

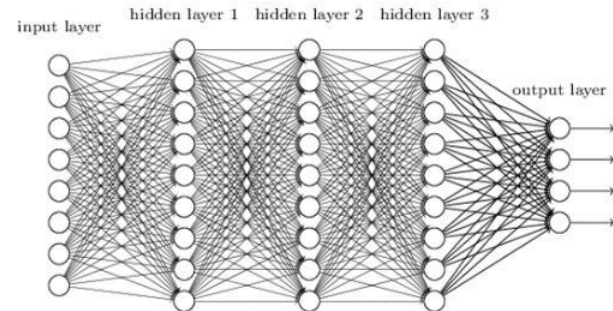


AI ML DL ...

- **Artificiële Intelligentie:** programmatuur die iets kan waar normaal menselijke intelligentie voor nodig is (computerspellen, Spotify recommendations)
- **Machine Learning:** technieken die het computers mogelijk maken te leren (adaptieve spam filters)
- **Deep Learning:** *machine learning* met behulp van 'diepe neurale netten'
- **Neuraal Net:** een netwerk van een heleboel onderling verbonden neuronen, zeer simpele imitaties van een hersencel



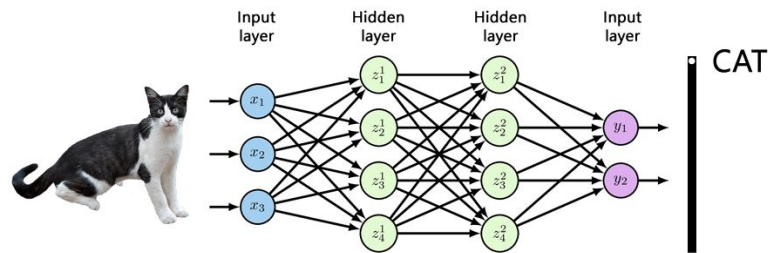
Deep neural network



Wat doet een neuraal net?

“Een neuraal netwerk is een functie die uit trainingsdatasets de verwachte output voor een bepaalde input leert”

- *Trainingsdatasets*: verzameling gelabelde afbeeldingen
- *Verwachte output*: correct label
- *Input*: een onbekende afbeelding

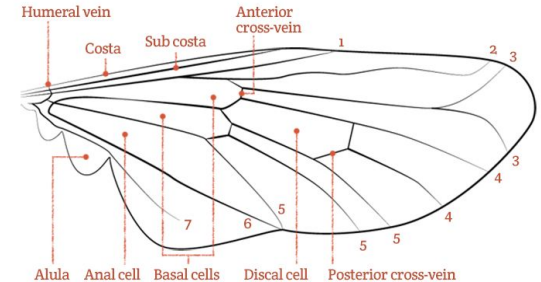
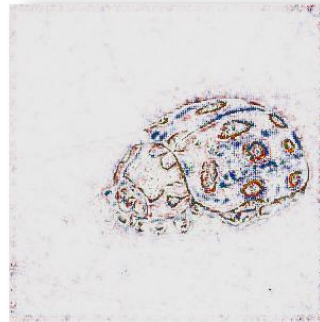


Supervised learning: door tijdens training correct gelabelde afbeeldingen te bekijken, leert een neuraal net afbeeldingen te classificeren die het niet eerder heeft gezien, door de labels toe te passen die het heeft geleerd.

Wat doet een neuraal net?

Wat ziet het (en wat niet)?

- Kijkt naar de hele afbeelding (inclusief achtergrond, labels, etc.)
- Herkent geen ‘klassieke onderdelen’ (zoals onderzoekers dat kunnen)
- Doorslaggevende aspecten zelden intuïtief
- “Als een mens het (niet) kan zien, dan een model ook (niet)”



Voorbeeld implementatie

Waarneming.nl

- 1.5M gelabelde afbeeldingen van observaties (v1)
- Soortidentificatie als *classifier* (ca. 9,000 soorten) (v1)
- Implementatie in app (ObsIdentify)
- Implementatie in validatie-workflow v/d site

Automatisch gevalideerde identificatie voor **236 soorten**.

Vooral dagvlinders, aantal nachtvlinders en vogels (situatie Augustus 2018).



Waarneming.nl
Add - My waarneming.nl - Observations - Species - Pics & sounds - Geography - This site - PNL

Automatische herkenning door Naturalis / Cosmonio / Observation.org
Als de foto(s) ook gps informatie bevat(ter) wordt gelijk een waarneming aangemaakt.
Nadat dit het met jouw zijn wordt u overgeleid naar het invoerscherm met de soort al ingevuld, de foto(s) zal u dan nog een keer invoeren, uploaden
Zorg dat de soort goed vingeren op de foto staat.
Bij gebruik van onze mobiele app is nodig: Soortcode, zoom-en-bidraathoeking.
The records need to be from the same species and location - In het geval een admin het toch niet oens is met de uitkomst dan geldt altijd het meest recente.

Use control (ctrl) key and mouse
Choose Files | submit.jpg

New Observation

fields with ! are required
Use all form fields

Date: 2018-08-07

Area: [empty]

Species group: Mammals

Species: Bank Vole - Myodes glareolus

Number: 1

unknown male female certain

Appearance: unknown

Activity: present

remarks: [empty]

OK



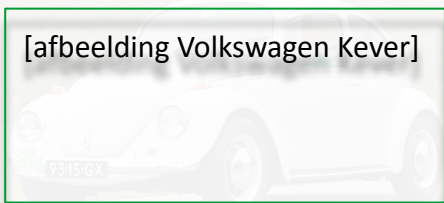
Waarneming.nl

Wat doet een neurale net?

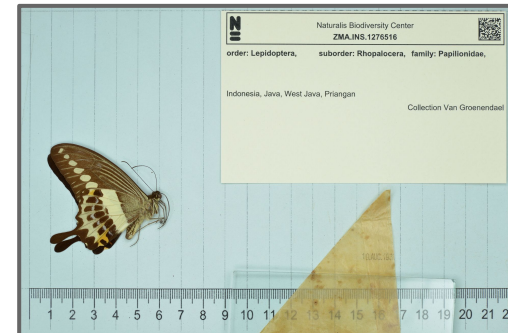
Beperkingen

- Accepteert trainingslabels blind als waarheid
- “Tunnelvisie”: kan alleen herkennen wat het ooit gezien heeft
- Weet niet wat het niet weet (geen “geen van bovenstaande”)
- Geeft altijd antwoord (*‘open world’*-probleem)
- Is soms zeer overtuigd van ‘foute’ antwoorden

[afbeelding Volkswagen Kever]



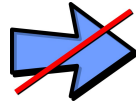
```
{  
  "predictions": [  
    {  
      "class": "Papilio demolion Cramer, 1776",  
      "prediction": 0.22217419743537903  
    },  
    ...  
  ]  
}
```



Trainingsdata

Selectie van data

- Welke vraag moet een model beantwoorden?
- Bevat de data het antwoord?
- Hoe wordt het voltooide model gebruikt? → vergelijkbaar met de trainingsdata (model heeft geen context)



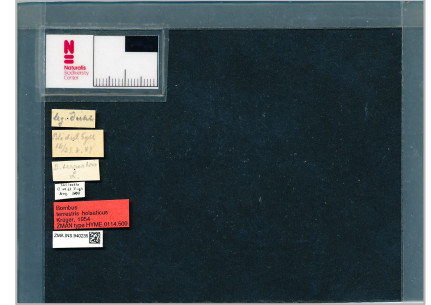
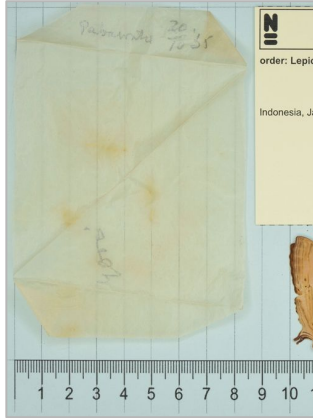
Graphium sarpedon (Linnaeus, 1758)

Trainingsdata

- Hoe meer beelden per klasse hoe beter (tot een paar duizend)
...mits verschillende beelden
 - Absolute minimum is 2 per klasse: split training & validatie
 - Gebalanceerd: ideaal als iedere klasse ongeveer even groot is
...lieft meer data (vaak moeilijk), anders trucs als upsampling
 - Uniformiteit labels
 - “Cyanistes caeruleus (Linnaeus, 1758)” ≠ “Pimpelmees”
 - “Cyanistes caeruleus (Linnaeus, 1758)” ≠ “Cyanistes caeruleus”
 - “Cyanistes caeruleus (Linnaeus, 1758)” ≠ “Parus caeruleus Linnaeus, 1758”
 - “Cyanistes caeruleus (Linnaeus, 1758)” ≠ “Cyanistes caeruleus (Linnaeus 1758)”
 - “Cyanistes caeruleus (Linnaeus, 1758)” ≠ “cyanistes caeruleus (linnaeus, 1758)”
 - “Cyanistes caeruleus (Linnaeus, 1758)” ≠ “Cyanistes caeruleus (Linnaeus, 1758)”
- GBIF *name resolver*
...vertegenwoordigt ook een taxonomische opvatting



Trainingsdata...



Beoordelingscriteria model

Verschillende maten:

Recall (gevoeligheid)

Verhouding tussen afbeeldingen die correct geïdentificeerd zijn als X en alle afbeeldingen die X zijn.
 $\text{true positives} / (\text{true positives} + \text{false negatives})$

Precision (precisie)

Verhouding tussen afbeeldingen geïdentificeerd als X dat ook echt X is.
 $\text{true positives} / (\text{true positives} + \text{false positives})$

Accuracy

Verhouding tussen aantal correct voorspellingen en alle voorspellingen.
 $(\text{true positives} + \text{true negatives}) / (\text{true \& false positives} + \text{true \& false negatives})$



- Accuracy alleen voor hele model, Recall en Precision ook per klasse.
- Kan zijn dat een model goed werkt voor sommige klassen, slechter voor anderen → specifiek gebruik

The proof of the pudding ...is in an API

Trainingssoftware gepubliceerd als open source via GitLab

<https://gitlab.com/naturalis/bii/beeldherkenning-museumproject>

Modellen in principe ook open source (maar niet gepubliceerd)

Reeks modellen project live gezet als API

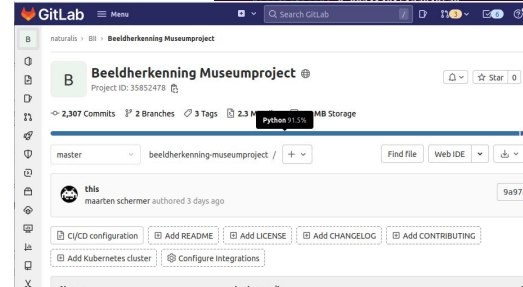
<https://museum.identify.biodiversityanalysis.nl/api>

Beschikbaar via de projectsite

<https://museum.identify.biodiversityanalysis.nl>



```
naartenghaarten-Latitude=7400!-/Pictures$ curl -s  
{  
  "predictions": [  
    {"class": "Conus pennaceus Bonn, 1778",  
     "prediction": 0.6182044744491577  
    },  
    {"class": "Conus textile Linnaeus, 1758",  
     "prediction": 0.3457694351673126  
    },  
    {"class": "Conus praerelatus Hass, 1792",  
     "prediction": 0.016861671581864357  
    },  
    {"class": "Conus purus Reuss, 1863",  
     "prediction": 0.008889370920062065  
    },  
    {"class": "Conus natalis G.B.Sowerby II, 1858"  
  }  
]
```



Huisjes van kegelslakken (Conidae)

Model ID: 20220428-141146

[switch to English](#)

[classes_link](#)

Het herkenningsmodel voor de huisjes van kegelslakken is getraind met foto's van 797 soorten Conidae. Het beeldmateriaal van de gedetermineerde soorten waarmee het model getraind is, is afkomstig uit vier museale collecties en één privé-collectie.

Beelden trainingsmodel

Het model is getraind met foto's van Conidae tegen zowel een donkere als lichte achtergrond. Het aantal specimen op de foto varieert; veel foto's bevatten één of twee exemplaren, maar de trainingsset bevat ook foto's met meerdere exemplaren.

Afhankelijk van de collectie bevatten de foto's behalve de specimen(s) ook een of meerdere labels.

Trainingsdata

Beelden en identificaties zijn afkomstig van:

- Naturalis Biodiversity Center
- Naturhistorisch Museum Rotterdam
- National History Museum London (collectie GBIF)
- Muséum national d'Histoire naturelle, Paris (collectie GBIF)
- Privécollectie van Fabrice Prugnaud (Frankrijk)

Voor dit model is een minimum van twee afbeeldingen per soort (klasse) gehanteerd.



Bedankt voor uw aandacht!



Naturalis
Biodiversity
Center



Bedankt voor uw aandacht!



Naturalis
Biodiversity
Center