Was kann Künstliche Intelligenz leisten?
CAS Digitale Technologien und Innovation, 13. September 2019
Thilo Stadelmann

Was ist KI?
Warum ist das jetzt aktuell?
Wie funktioniert das?
Was → Warum? → Wie?

1

Was ist Künstliche Intelligenz?
Was ist künstliche Intelligenz?

"The exciting new effort to make computers think... machines with minds, in the full and literal sense." (Haugeland, 1985)

"[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning..." (Bellman, 1978)

"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985)

"The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)

"The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)

"Computational Intelligence is the study of the design of intelligent agents." (Poole et al., 1998)

"AI... is concerned with intelligent behaviour in artefacts." (Nilsson, 1998)
Was gehört zu künstlicher Intelligenz?

Artificial Intelligence (AI)

- deep learning
  - supervised
  - unsupervised
- content extraction
- classification
- machine translation
- question answering
- text generation
- natural language processing (NLP)
- expert systems
- vision
- speech
- text to speech
- planning
- robotics

© 2015 Neota Logic
Was kann KI bereits heute?

1. Play a decent game of **table tennis**  
   \(\text{ok}\)
2. Drive safely along a curving **mountain road**  
   \(\text{ok}\)
3. Drive safely along **Technikumstrasse** Winterthur  
   \(\text{ok} \text{ (only since recently)}\)
4. Buy a week's worth of **groceries on the web**  
   \(\text{ok}\)
5. Buy a week's worth of groceries **at Migros**  
   \(\text{no}\)
6. Play a decent game of **bridge**  
   \(\text{ok}\)
7. Discover and prove a new mathematical **theorem**  
   \(\text{not completely}\)
8. Design and execute a **research program** in molecular biology  
   \(\text{not completely}\)
9. Write an **intentionally funny** story  
   \(\text{no}\)
10. Give competent **legal advice** in a specialized area of law  
    \(\text{ok}\)
11. Translate spoken English into spoken **Swedish** in real time  
    \(\text{ok}\)
12. Converse successfully with another person for an hour  
    \(\text{no}\)
13. Perform a complex **surgical operation**  
    \(\text{not completely}\)
14. Unload any **dishwasher** and put everything away  
    \(\text{no}\)
15. Compete in the game show **Jeopardy!**  
    \(\text{ok}\)
16. Write **clickbait** articles fully automatized  
    \(\text{ok}\)
Beispiel: Machbar vs. gefährlich
Technologie: Computer Vision mit Deep Learning

https://www.cultofmac.com/495088/avoid-potentially-deadly-ai-app/
Beispiel: Markterfolg vs. regulatorische Hürden
Technologie: Recommender Systems
Beispiel: Statistik vs. Bias
Technologie: Machine Learning

See also: Nassim Nicholas Talib, «The Black Swan: The Impact of the Highly Improbable», 2007
Beispiel: künstl. Intelligenz vs. natürl. Dummheit
Technologie: Machine Learning mit nachgelagerten Regeln

Cylance, I Kill You!

Read about our journey of dissecting the brain of a leading AI based Endpoint Protection Product, culminating in the creation of a universal bypass.

TL;DR
AI applications in security are clear and potentially useful, however AI based products offer a new and unique attack surface. Namely, if you could truly understand how a certain model works, and the type of features it uses to reach a decision, you would have the potential to fool it consistently, creating a universal bypass.

By carefully analyzing the engine and model of Cylance's AI based antivirus product, we identify a peculiar bias towards a specific game. Combining an analysis of the feature extraction process, its heavy reliance on strings, and its strong bias for this specific game, we are capable of crafting a simple and rather amusing bypass. Namely, by appending a selected list of strings to a malicious file, we are capable of changing its score significantly, avoiding detection. This method proved successful for 100% of the top 10 Malware for May 2019, and close to 90% for a larger sample of 384 malware.
Gefahren durch KI?

- KI ist per Definition eine "dual use Technology" → siehe Report von Brundage et al., 2018

- Aber: “natürliche Dummheit” ist die grössere Bedrohung

- Algorithmische Ethik und erklärbare KI sind in den letzten Jahren zu einem top Forschungsfeld geworden – nicht wegen der unkalkulierbaren Risiken per se, sondern:
Was → Warum? → Wie?

2

Warum ist das jetzt aktuell?
(Eine kurze Geschichte der letzten Jahre)
Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M

Posted Jan 26, 2014 by Catherine Shu (@catherineshu)

Google will buy DeepMind, reports that the company is in talks to buy a firm, and couldn’t disclose deal terms.

The acquisition was originally confirmed by Google to Re/code.

---

40 days

AlphaGo Zero surpasses all other versions of AlphaGo and, arguably, becomes the best Go player in the world. It does this entirely from self-play, with no human intervention and using no historical data.
Deep neural networks can now transfer the style of one photo onto another

And the results are impressive

by James Vincent | @jvincent | Mar 30, 2017, 1:53pm EDT

A deep neural network can now transfer the style of one image onto another images.

by Emerging Tech

The nature of art of Vincent Van Gogh and Edvard Munch's artworks now recognized by computers.

Original photo | Reference photo | Result

You’ve probably heard of an AI technique known as “style transfer” — or, if you haven’t heard of it, you’ve seen it. The process uses neural networks to apply the look and feel of one image to another, and appears in apps like Prisma and Facebook. These style transfers, however, are stylistic, not photorealistic. They look good because they look like they’ve been painted. Now a group of researchers from Cornell University and Adobe have augmented
What if you could imitate a famous celebrity’s voice or sing like a famous singer? This project started with a goal to convert someone’s voice to a specific target voice. So called, it’s voice style transfer. We worked on this project that aims to convert someone’s voice to a famous English actress Kate Winslet’s voice. We implemented a deep neural networks to achieve that and more than 2 hours of audio book sentences read by Kate Winslet are used as a dataset.

Model Architecture

This is a many-to-one voice conversion system. The main significance of this work is that we could generate a target speaker’s utterances without parallel data like <source’s wav, target’s wav>, <wav, text> or <wav, phone>, but only waveforms of the target speaker. (To make these parallel datasets needs a lot of effort.) All we need in this project is a number of waveforms of the target speaker’s utterances and only a small set of <wav, phone> pairs from a number of anonymous speakers.

A’s Waveforms  Speech Recognition  Speech Synthesis  B’s Waveforms

Train1 lw small parallel dataset

Train2 lw large non-parallel dataset

“My name is Avin!”  “My name is Avin!”
...und die Liste liesse sich fortsetzen!

Nvidia AI Generates Fake Faces Based On Real Celebs

For today’s post, I’ve drawn material not just from one paper, but from five! The subject matter is *word2vec* – the work of Mikolov et al. at Google on efficient vector representations of words (and what you can do with them). The papers are:

- Efficient Estimation of Word Representations in Vector Space – Mikolov et al. 2013
- Distributed Representations of Words and Phrases and their Compositionality – Mikolov et al. 2013
- Linguistic Regularities in Continuous Space Word Representations – Mikolov et al. 2013
- word2vec Parameter Learning Explained – Rong 2014
- word2vec Explained: Deriving Mikolov et al’s Negative Sampling Word-Embedding Method – Goldberg and Levy 2014

From the first of these papers ("Efficient estimation...") we get a description of the *Continuous Bag-of-Words* and *Continuous Skip-gram* models for learning word vectors (we’ll talk about what a word vector is in a moment...). From the second paper we get more illustrations of the power of word vectors, some additional information on optimisations for the skip-gram model (hierarchical softmax and negative sampling), and a discussion of the parameter learning process. From the last two papers we get more information on the negative sampling method.
Was ist passiert?
Der ImageNet Wettbewerb

1000 Kategorien
1 Mio. Beispiele

2015: Computer *haben* “Sehen” gelernt

4.95% Microsoft (06. Februar)
→ Besser als Menschen (5.10%)

4.80% Google (11. Februar)

4.58% Baidu (11. Mai)

3.57% Microsoft (10. Dezember)

A. Krizhevsky verwendet als erster ein sog. «Deep Neural Network» (CNN)
Was → Warum? → Wie?

Wie geht das?
Idee: Mehr «Tiefe» um Merkmale automatisch zu lernen

Automatisierung komplexer Prozesse basierend auf (hoch-dimensionalem) Sensor-Input
Grundlage
Induktives überwachtes Lernen

Annahme
• Ein an genügend viele Beispiele angepasstes Modell…
• …wird auch auf unbekannte Daten generalisieren

Methode
• Suchen der Parameter einer gegebenen Funktion…
• …so dass für alle Beispiele Eingabe (Bild) auf Ausgabe («Auto») abgebildet wird

\[ f(x) = y \]

Suche der Parameter einer Funktion?

Neuron

Merkmale (z.B. Pixel)

Anpassbare Parameter

Entscheidung (Schwellwert)

Ergebnis (z.B. «Auto»)

Neuronales Netz
Ausblick: Innovation aus neuronalen Netzen?

- **Deep Learning** ist der Kern vieler erstaunlicher Anwendungen im Bereich der Automatisierung von Wahrnehmung (Synthese & Analyse)
  ➔ Code & Datensätze laden “Maker” zum Ausprobieren ein

- **Automated Deep Learning** macht in Zukunft vielleicht manche Anwendung einfacher

- Aktuell braucht es ein gutes Verständnis der Details der Methode, um kreativ neue Szenarien zu erdenken (Stichwort “mTrainer” anstatt Programmierer?)
Schlussfolgerungen

• Deep Learning hat zu Paradigmenwechsel in *Mustererkennungsaufgaben* geführt
• Die Zeit vom Grundlagenresultat zur praktischer Anwendung beträgt wenige Monate
• Es gibt Methoden zum Hineinschauen in neuronale Black Boxes (siehe Anhang)
• Spezifische Aufgaben lassen sich sehr gut automatisieren (z.B. Ähnlichkeitssuche)

Zu mir:
• Prof. KI/ML, Scientific Director ZHAW digital
• Email: stdm@zhaw.ch
• Telefon: 058 934 72 08
• Web: https://stdm.github.io/
• Twitter: @thilo_on_data
• LinkedIn: thilo-stadelmann

Mehr zum Thema:
• Data+Service Alliance: www.data-service-alliance.ch
• KI: https://sgaico.swissinformatics.org/
• Zusammenarbeit: datalab@zhaw.ch
ANHANG
Developing for algorithmic fairness
The FAT ML code of conduct
See http://www.fatml.org/resources/principles-for-accountable-algorithms

Purpose
• Help developers to build algorithmic systems in publicly accountable ways
• Accountability: the obligation to report, explain, or justify algorithmic decision-making & mitigate any negative social impacts or potential harms

Premise
• A human ultimately responsible for decisions made/informed by an algorithm

Principles
• Responsibility, Explainability, Accuracy, Auditability, Fairness

Making it actionable
• Publish a Social Impact Statement
• …use above principles as a guiding structure
• …revisit three times during development process: design stage, pre-launch, post-launch

Make available somebody who will take care of adverse individual / societal effects
Explain any algorithmic decision in non-technical terms to end users
Report all sources of uncertainty / error in algorithms & data
Enable 3rd parties to probe & understand system behavior
Ensure algorithmic decisions are not discriminatory w.r.t. to people groups
Suche der Parameter einer Funktion?

- Unser Neuronales Netz: \( f_W(x) = y \)
  mit Bild \( x \), echtem Resultat \( y \) und Parametern \( W \)
  (\( W = \{w_1, w_2, \ldots\} \) anfangs zufällig gewählt)

- Fehlermass: \( l(W) = \frac{1}{N} \sum_{i=1}^{N} (f_W(x_i) - y_i)^2 \)
  Durchschnitt der quadratischen Abweichungen über alle Bilder (Loss)

\[
l(W) = \frac{1}{N} \sum_{i=1}^{N} (f_W(x_i) - y_i)^2
\]

Durchschnitt (über alle Beispiele)

Bestraft grosse Fehler überproportional stärker

Differenz IST – SOLL (Fehler)

\( \leftarrow \) Fehlerlandschaft

Methode: Anpassung der Gewichte von \( f \) in Richtung der steilsten Steigung (abwärts) von \( J \)
Was «sieht» das Neuronale Netz?
Hierarchien komplexer werdender Merkmale

Wie schlussfolgert die Maschine?
«Debugging» für Einblicke in die vermeintliche «Black Box»

Verdeutlichen ein Problem:
• Adversarial Examples

Bieten eine Lösung:
• Saliency Maps

Ruth C. Fong & Andrea Vedaldi, «Interpretable Explanations of Black Boxes by Meaningful Perturbation», 2017
**Trace & detect adversarial attacks**

…using average local spatial entropy of feature response maps

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Adversarial</th>
<th>Original</th>
<th>Adversarial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image:</strong></td>
<td><img src="image1.png" alt="Original Image" /></td>
<td><img src="image2.png" alt="Adversarial Image" /></td>
<td><img src="image3.png" alt="Original Image" /></td>
<td><img src="image4.png" alt="Adversarial Image" /></td>
</tr>
<tr>
<td><strong>Feature response:</strong></td>
<td><img src="image5.png" alt="Feature Response" /></td>
<td><img src="image6.png" alt="Feature Response" /></td>
<td><img src="image7.png" alt="Feature Response" /></td>
<td><img src="image8.png" alt="Feature Response" /></td>
</tr>
<tr>
<td><strong>Local spatial entropy:</strong></td>
<td><img src="image9.png" alt="Entropy Map" /></td>
<td><img src="image10.png" alt="Entropy Map" /></td>
<td><img src="image11.png" alt="Entropy Map" /></td>
<td><img src="image12.png" alt="Entropy Map" /></td>
</tr>
</tbody>
</table>