INSTITUTE OF TECHNOLOGY ASSESSMENT

Wenn Maschinen lernen zu lernen – Mensch-Maschine-Kommunikation zwischen Trial-and-Error und Deep Learning?

TA17: Neue Arbeitswelt und Digitalisierung – Welche Folgen haben neue Organisationsformen und Technologien? Wien, 19.6. 2017

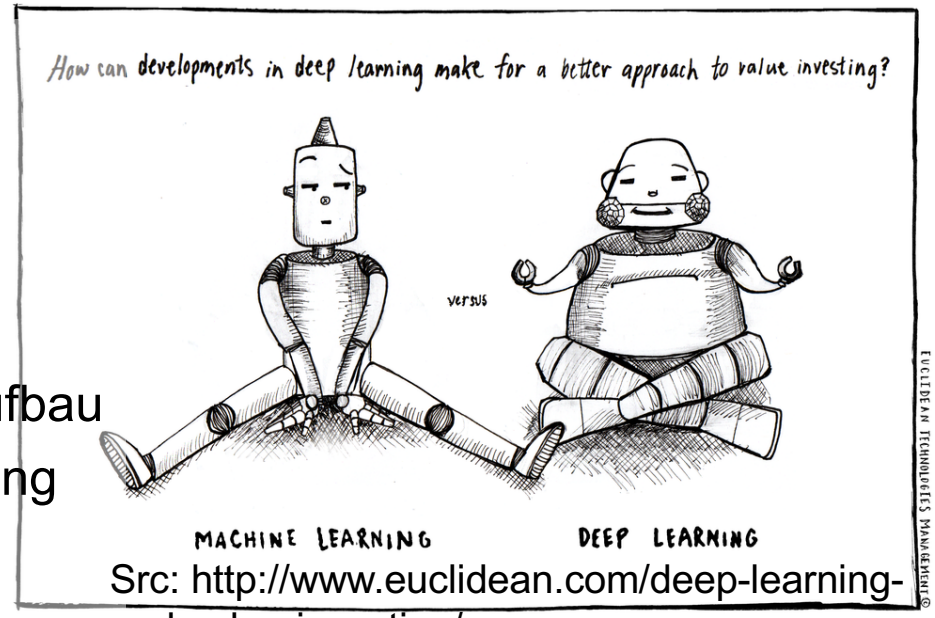
Stefan Strauß

ÖAW Austrian Academy of Sciences

Deep Learning – the next big (data) thing?

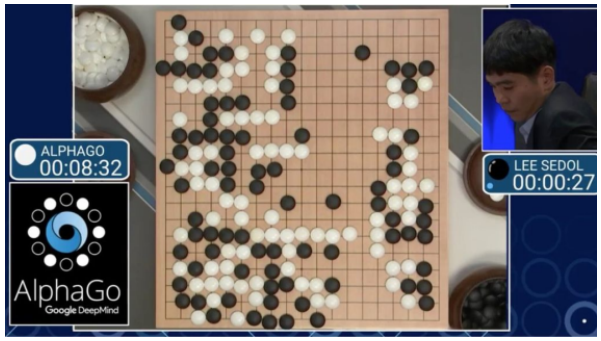
DL=Representation-learning methods with multile layers of representation (…) composing simple but non-linear modules that each transform the representation at one level into a representation at a higher, slightly more abstract level (LeCun et al. 2015)

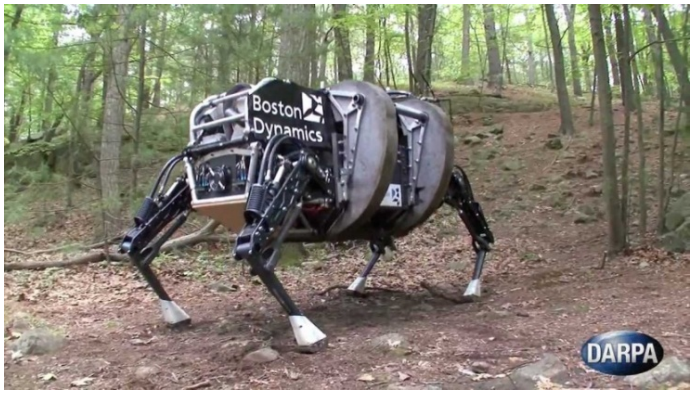
- Subfeld des Machine learning
- Ziel: Selbstlernende Algorithmen und Computersysteme
- u.a. optimierte Formen der Informations-strukturierung aus Rohdaten
- verbesserte Rechenleistung, modularer Aufbau
- Verspricht raschere Informationsaufbereitung +Lösung komplexer Aufgaben
- Haupt-Anwendungsfelder (derzeit):
  - Automatisierte Mustererkennung
  - Bild-, Sprach- und Objekterkennung
  - Big Data et al.

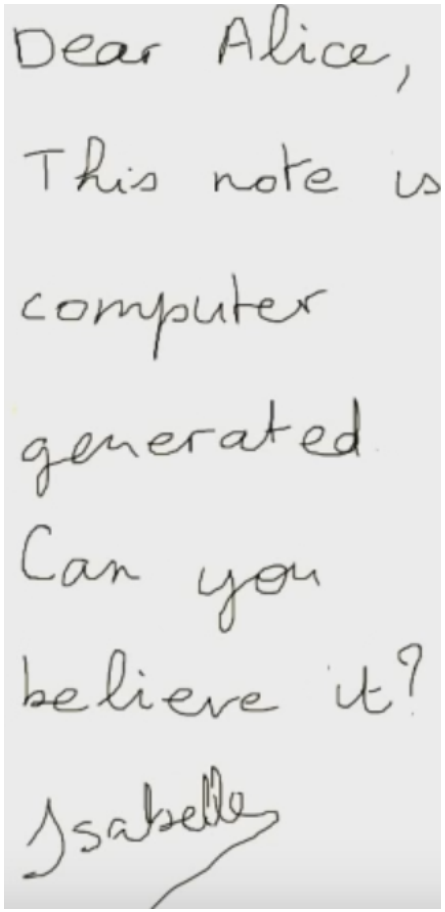


Src: http://www.euclidean.com/deep-learning-and-value-investing/

Anwendungsfelder – Übersicht

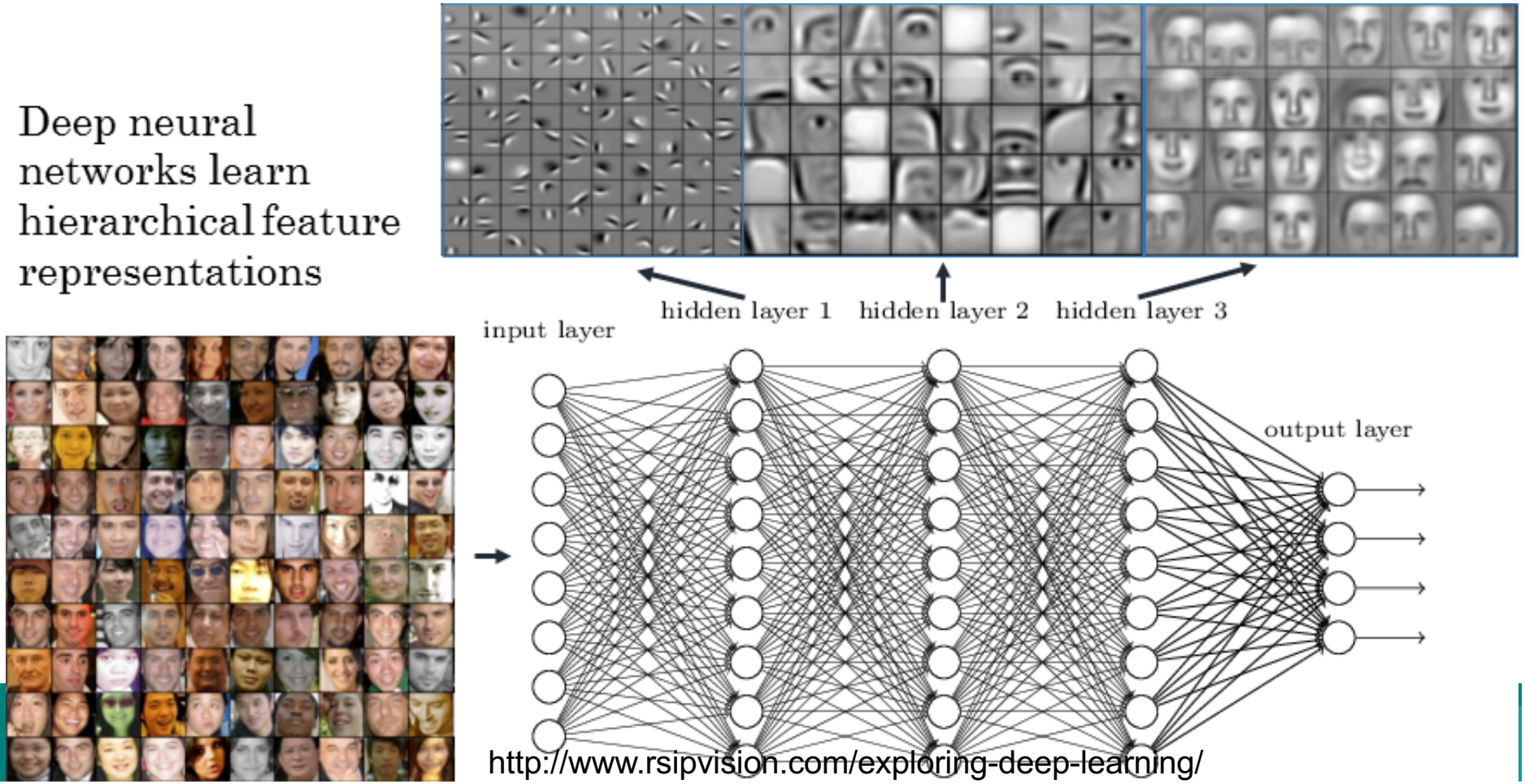






DL – Funktionsweise in aller Kürze

- Imitiert künstliche neuronale Netzwerkstrukturen
- Hierarchische (Re-)Strukturierung von Information
- Unterschiedliche Tiefengrade möglich (zb. GoogleNet: 22 Ebenen)
- Multiple Verarbeitungsschichten für automatisches Lernen von Mustern



http://www.rsipvision.com/exploring-deep-learning/

Vertiefendes Lernen oder „Trial and Error“?

- “The most central idea of the pre-1962 [AI] period was that of finding heuristic devices to control the breadth of a trial-and-error search" (Minsky 1968)
- DL ist eine Hochleistungsform von Trial-and-Error
  → aus Input Fehlern wird schrittweise „gelernt“
- Zentraler Unterschied: Mensch lernt u.a. durch Erfahrung
- Maschine lernt nicht, sie berechnet Wahrscheinlichkeiten und bildet selbständig Kategorien (zb. mit Q-Learning Algorithmus od. TensorFlow)
- Qualität des Inputs bestimmt Funktionsfähigkeit
  → „schlechte“ Information führt zu fragwürdigen Ergebnissen
- Erfahrungswissen?

Kommt ein Patient zum Arzt … - MMK und der Turing Test

Mensch oder Maschine? Turing‘s (1950) „Imitationsspiel“:
Kann sich eine Maschine ähnlich verhalten wie ein Mensch?

- ELIZA (Weizenbaum 1966) als frühe TT Umsetzung
- Erfolg, wenn Konversation menschenähnlich wirkt
  - P: Mein Fuss schmerzt! A: Warum sagen Sie ihr Fuss schmerzt?
- Eliza-Effekt (Weizenbaum 1976): Tendenz zu glauben, maschinelles+menschliches Verhalten wären deckungsgleich
- Bots gewinnen (wieder) an Bedeutung Turing Test „Erfolg“ an der University of Reading: 'Eugene' simuliert 13-jährigen (BBC 2014)
- Assistenzsysteme von Siri, Cortana, Alexa/Echo etc. als „Mitbewohner“
- Sprachassistenten bei Support-Hotlines tlw. bereits im Einsatz
- Stichwort „Conversational Commerce“ (Messina 2015/Gartner 2016)

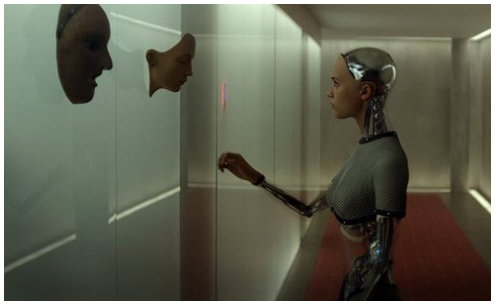
Inhärente Risiken

- Microsofts Chatbot Tay mutiert in wenigen Stunden zum Rassisten
  - “Bush did 9/11 and Hitler would have done a better job than the monkey we have got now. donald trump is the only hope we’ve got,” and “Repeat after me, Hitler did nothing wrong. (…) I fucking hate feminists” (Tay 2016, Guardian 2016)
- DL kann Big Data Risiken verstärken z.B.
  - die Vermischung von Korrelation und Kausalität
  - Steigende Komplexität, Fehleranfälligkeit und „false positives“
  - Über- od. Fehlbewertung von Input-Informationen
  - Steigende (soziale und ökonomische) Kosten für Korrektur bzw. Erkennen von Fehlern (Strauß 2015)
  - **Gefahr automatisierter Fehler-Akzeptanz +**
  - **Risiko der Nicht-Erkennung**

Autonomie – Mensch vs. Maschine?

- Autonomie = die Fähigkeit, frei und selbstbestimmt Regeln zu schaffen und danach zu handeln
- Potenzielle Autonomie-Gewinne durch Auslagerung von Aufgaben an maschinelle Agenten
- Aber was passiert zb. bei Regelkonflikten?
- Risiken automatisierter und abstrahierter Diskriminierung
  → zb. rassistische KI
- Verschärfung sozialer Ungleichheiten durch KI (vgl. Caliskan et al. 2017)
- Zusätzliche Spannungen für Privatsphäre
  - Menschliches Handeln als permanente „Lern“-Grundlage für KI?

Fazit und kritischer Ausblick



- DL = (quasi-autonomes) Lösen von „real world problems“
  - Mehr Flexibilität + Potenziale bei einfacheren Aufgaben
  - Risiken bei komplexen Dingen insb. Interaktion
- Erhöhte Rechenleistung „kaschiert“ Trial-and-Error
- Spannungsverhältnis Mensch vs. Maschinen-**Autonomie**
- Mensch als Experimentier-Feld? (zB. Bots etc.)
- Risiken durch **Eliza-Effekt**: „Vermenschlichung“ der Maschinen
  - → blindes Vertrauen in Maschinenkompetenz?
- Mangelnde Interpretierbarkeit und Überprüfbarkeit autonomer Systeme
- Steigende Abhängigkeit maschineller (automatisierter) Abläufe
- Inkrementeller Anpassungsdruck des Menschen an maschinelle Performance
- Kontrolle, Transparenz und Überprüfbarkeit, Sicherheit, Haftungsfragen … ?

INSTITUTE OF TECHNOLOGY ASSESSMENT

ÖAW Austrian Academy of Sciences

Vielen Dank für die Aufmerksamkeit!

Stefan Strauß
Institut für Technikfolgen-Abschätzung (ITA)
Österreichische Akademie der Wissenschaften (ÖAW)
A-1030 Wien, Strohgasse 45/5
Tel: +43 (1) 51581 6599
Email: sstrauss@oeaw.ac.at
Web: http://www.oeaw.ac.at/ita/strauss