Daniel Sonntag, DFKI
10/10/2019

Wie funktionieren neuronale Netze eigentlich?

Viele Ideen und Prinzipien im Gebiet der neuronalen Netze entstammen der Hirnforschung. Die Neuro-Kognitionswissenschaft hat die Funktion natürlicher neuronaler Netze in den letzten 50 Jahren untersucht. In der Künstlichen Intelligenz wird versucht, diese Funktionen zu approximieren (oder zu modellieren, zu simulieren). Diese Approximation suggeriert die Behandlung der intelligenten Funktionen als "schwachen KI", als problemorientierte KI, die eine Teildisziplin der Informatik darstellt und direkt an Anwendungsfälle wie das Klassifizieren von Bildinhalten geknüpft ist. Hierbei stellt sich jedem KI-Forscher die spannende Herausforderung und wissenschaftliche Challenge, die Ergebnisse mit der Leistungsfähigkeit von menschlichen Experten zu vergleichen, wie beispielsweise Radiologen bei der medizinischen Bildverarbeitung.

Das mathematische Modell geht ähnlich dem biologischen Modell von Neuronen und Synapsen aus, die die Verbindungen zwischen den Neuronen herstellen. Die Simulation durch ein mathematisches Modell wird durch eine sogenannte Aktivierungsfunktion implementiert, die jedes Neuron aufgrund der Verbindungen zu anderen Neuronen berechnet. Dadurch ist es möglich, die zentrale Frage des Lernens in neuronalen Netzen zu beantworten: es werden die Gewichte zwischen Neuronen gestärkt, die zu einem positiven Ergebnis führen, und die Gewichte abgeschwächt, die zu einem negativen Ergebnis führen. Gesucht wird die Gewichtung, die bei einem bestimmten Eingangsdatensatz den geringsten Fehler auf diesen Trainingsdaten erreicht. Dadurch wird erreicht, dass das neuronale Netz bei ähnlichen Eingaben ähnliche Ergebnisse bei der Klassifikation liefert; das Netz hat gelernt, wie die Eingabe zu klassifizieren ist.

Die Methode, komplexe Netze auf diese Weise zu trainieren (Backpropagation) ist schon 30 Jahre alt, wird aber immer noch verwendet und hat sich in der vor allem in der Musterkennung (Bilderkennung) und Robotersteuerung bewährt. Bei großen Netzen treten Konvergenzprobleme auf. Oft dauert es viel zu lange, bis ein verwertbarer Trainingseffekt einsetzt. Hier besteht noch Forschungspotential, ebenso beim Versuch, gelernte neuronale Netze zu verstehen. Die verteilte Repräsentation des in neuronalen Netzen gespeicherten Wissens hat aber den entscheidenden Nachteil, dass es schwierig ist, Informationen in den tausenden (oder Millionen!) von Gewichten in tausenden

von Neuronen zu lokalisieren. Es ist praktisch nicht möglich, die Gewichte zu analysieren und zu verstehen. Allerdings kann man beispielsweise versuchen zu verstehen, wo das Netz gut funktioniert, und wo nicht, was zum Interaktivem Maschinellem Lernen (iml.dfki.de) führt. Oder man versucht bei der Bilderkennung zu visualisieren, welche Bereiche in einem Bild bei der Klassifikation verantwortlich waren (http://dfki.de/IML/XAI/xai.html). Generell wird dem XAI (explainable AI) eine große Bedeutung zugemessen, die inneren Abläufe zu verstehen und einem Regelwerk gleichzusetzen, das vom Menschen verstanden und bewertet werden kann.