Zuverlässige und vertrauenswürdige Künstliche Intelligenz

Prof. Dr. Marco Huber, Universität Stuttgart und Fraunhofer IPA

Viele Verfahren des maschinellen Lernens wie etwa Deep Learning werden oftmals als "Black Box" betrachtet, weil es für Menschen unmöglich ist, die internen Entscheidungsprozesse der Verfahren zu verstehen. Für einige Anwendungsfälle, etwa beim autonomen Fahren oder in der Medizin, ist jedoch nicht nur die Genauigkeit, sondern auch die Sicherheit und das Vertrauen in die Algorithmen von enormer Bedeutung. In solchen Anwendungen ist es wichtig, dass kritische algorithmische Entscheidungen abgesichert werden.

In diesem Vortrag werden Aspekte einer zuverlässigen und vertrauenswürdigen KI betrachtet. Im Fokus steht der Aspekt der Absicherung maschinell gelernter Modelle, die auf tiefen künstlichen neuronalen Netzen basieren. Dabei werden drei Fragestellungen detailliert:

(1) *Erklärbarkeit:* Es wird der Frage nachgegangen, wie ein Mensch die Entscheidung eines gelernten Modells nachvollziehen kann.
(2) *Verifikation:* Bei der Verifikation wird formal geprüft, ob das Modell zugesicherte Eigenschaften auch bei beliebigen Eingaben erfüllt.
(3) *Unsicherheitsquantifizierung:* Mittels Bayes'scher Verfahren wird aufgezeigt, ob das Modell überhaupt weiß was es nicht weiß.

Der Vortrag gewährt somit Einblick in konkrete Ansätze zur Absicherung gelernter Modelle. Der Nutzen wird anhand praktischer Anwendungsfälle demonstriert.