Schlussbericht

19-109-R "Understanding Deep Learning" **CHF 88'740.-**

Dr. Tim Räz, Prof. Claus Beisbart, Universität Bern

Das Ziel des Projekts war es, eine Verbindung zwischen Wissenschaftsphilosophie und Deep Learning (DL) herzustellen. DL-Modelle sind ein neue Technologie zur automatischen Mustererkennung, mit vielen Anwendungen in Wissenschaft und Gesellschaft. Allerdings ist die Funktionsweise dieser Technologie nicht sehr gut verstanden, wie Informatiker regelmässig betonen. Doch was genau ist gemeint, wenn wir sagen, dass DL nicht oder ungenügend verstanden ist oder dass wir Aspekte dieser Modelle nicht erklären können? In unserer Untersuchung wollten wir klären, wie man den Verstehens- und den Erklärungsbegriff im Kontext von Deep Learning auffassen kann, indem wir auf die entsprechenden Debatten in der Wissenschaftsphilosophie zurückgriffen und Verbindungen zwischen Philosophie und Informatik herstellten. Das Projekt umfasste zwei Teilprojekte sowie zusätzliche Forschungen, die nachfolgend kurz beschrieben werden.

Teilprojekt 1 "*From Deep Learning to Philosophy*" (bearbeitet von Tim Räz):
Darin wurde eine theoretische Methode aus der Informatik untersucht, die sogenannte "Information Bottleneck" (IB) Methode, mit der Aspekte von DL-Modellen erklärt werden sollen. Die IB-Methode zeigt auf, dass neuronale Netzwerke implizit ein bestimmtes Optimierungsproblem lösen. Dieses Optimierungsproblem ist ein Kompromiss zwischen genauen Vorhersagen und dem "Vergessen" von irrelevanter Information. Das Optimierungsproblem der IB-Methode verallgemeinert sogenannte minimal hinreichende Statistiken, ein Konzept aus der klassischen Statistik. Unsere Arbeit zeigt, dass minimal hinreichende Statistiken eine formale Entsprechung in Wesley Salmons philosophischem Erklärungsmodell der statistischen Relevanz haben. Diese formale Beziehung ermöglicht einen Transfer von Ideen zwischen Philosophie und Informatik; so entsprechen etwa zwei Lernphasen von neuronalen Netzwerken, Fehlerminimierung und Kompression, zwei Desideraten von mathematischen Erklärungen.

Teilprojekt 2 "*From Philosophy to Deep Learning*" (bearbeitet von Tim Räz und Claus Beisbart):
Hier stand die Frage im Vordergrund, wie das mangelnde Verständnis von DL-Modellen die Möglichkeit beschränkt, empirische Phänomene mit Hilfe solcher Modelle zu verstehen. Dazu wurde die Position von Emily Sullivan kritisch analysiert, welche sie in einer neuen Publikation zu dieser Frage eingenommen hat. Sullivan argumentiert, dass man die DL-Modelle nicht besser verstehen müsse, als dies gegenwärtig der Fall sei, um sie als Instrumente zum Verstehen empirischer Phänomene zu verwenden. Gemäss Sullivan ist es wichtiger, die Relation zwischen den Modellen und der Welt besser zu verstehen. In unserem Beitrag argumentierten wir, dass wir DL-Modelle durchaus besser verstehen müssten, um sie zum Verstehen empirischer Phänomene verwenden zu können. Für diese Argumentation unterschieden wir verschieden starke Verstehensbegriffe, um zu zeigen, dass die DL-Modelle keine Erklärungen liefern können, weil der prognostische Erfolg dieser Modelle nur ungenügend verstanden wird. Ausserdem kann man keine strikte Trennlinie zwischen dem Verstehen eines Modells selbst und der Relation zwischen Modell und Welt ziehen.

Zusätzliches Projekt (bearbeitet von Tim Räz gemeinsam mit Julie Jebeile und Vincent Lam):
Neben den oben beschriebenen Teilprojekten wurde ein zusätzliches Forschungsprojekt lanciert. Dabei stand die Frage im Zentrum, wie der Einsatz von DL-Modellen in der Klimamodellierung unsere Fähigkeit beeinflusst, Klimaphänomene zu verstehen. Zu diesem Zweck formulierten wir fünf Kriterien, die erfassen sollen, wann man mit Hilfe eines Modells Klimaphänomene verstehen kann. Die Kriterien umfassen Verstehen durch Manipulierbarkeit; empirische Genauigkeit; Verstehen des physikalischen Prozesses; physikalische Konsistenz; und den Gültigkeitsbereich. Mit diesen fünf Kriterien untersuchten wir zwei Fallstudien aus den Klimawissenschaften: zum einen die Verwendung von statistischer Skalierung mit Quantilabbildungen in der Modellierung des regionalen Klimas in der Schweiz, zum anderen die Verwendung von neuronalen Netzwerken in einem globalen Zirkulationsmodell. Es zeigte sich, dass sich das Verstehen entlang der genannten Kriterien graduell und zum Teil gegenläufig verhält. Gleichzeitig ist aber die Veränderung der fünf Kriterien in beiden Fallstudien, also im Fall eines eher traditionellen statistischen Verfahrens, und im Fall von DL-Modellen, qualitativ gleich. Dies spricht dafür, Verstehen als mehrdimensionalen und gradierten Begriff aufzufassen.

Allgemeine Schlussfolgerungen:

1. Philosophische Erkenntnisse zu Erklären und Verstehen können helfen zu klären, wie DL-Modelle verständlicher werden können.
2. Es erweist sich als fruchtbar, konzeptuelle Verbindungen zwischen Wissenschaftsphilosophie und Deep Learning herauszuarbeiten; solche Verbindungen sind für beide Forschungsfelder erhellend.
3. Für die Philosophie, die sich mit Deep Learning beschäftigt, führt kein Weg um eine genaue Analyse technischer Arbeiten aus der Informatik herum.
4. Fallstudien (wie etwa zu den Klimawissenschaften) sind ein fruchtbares Mittel, um das Potential, aber auch die Grenzen von Deep Learning in den Wissenschaften zu verstehen.