Zusammenfassung der Dissertation von Dipl.-Math. Sven Arnhold

**Data Restructuring as Formal Preprocessing for Machine Learning with Neural Networks**

Künstliche neuronale Netze werden im Bereich des maschinellen Lernens zur Nachahmung von Expertenwissen eingesetzt. Sie können als feedforward Netze zwischen Daten mit fester Struktur abbilden, als rekurrente Netze auf Daten mit sequentiellem Charakter wie z.B. Zeitreihen und als rekursive Netze zum Lernen auf Datenstrukturen wie chemischen Strukturformeln verwendet werden.

In der Praxis gestaltet sich das Training, also die Anpassung der freien Parameter, meistens schwierig. Ständiger Gegenstand der Forschung ist daher unter anderem, spezielle Netzarchitekturen zu entwickeln, die sich für einen praktischen Einsatz gut eignen. Die Netzarchitektur Long Short-Term Memory (LSTM) wurde z.B. gezielt konstruiert, um dem Effekt des „fading gradient“ zu begegnen. Dieser Effekt verhindert das praktikable Training rekurrenter Netze mittels Gradientenabstieg.

Die Netze verarbeiten die Daten in fester Durchlaufrichtung. Ist eine zu lernende Ausgabe an der konkreten Stelle aber von nachfolgenden Punkten abhängig, kann dieser Sachverhalt nicht gelernt werden. Der Kompromiss, ein Fenster statischer Größe aus Eingabedaten zu verwendet, lässt sich für rekursive Netze nur schwer einsetzen.

Es gibt daher nicht-kausale Netzwerkarchitekturen, die den Kontext, also auch punktuelle Nachfolger, berücksichtigen. Weiterhin wurden bidirektionale rekurrente Netze (BRN) definiert, die eine bereits gegebene Netzarchitektur verwenden und die Sequenz in zwei Durchlaufrichtungen gleichzeitig verarbeiten. Kontextuelle Netze erfordern Einschränkungen an ihre interne Struktur. Beide Netzarchitekturen, kontextuelle und bidirektionale, lassen die Form der Eingabedaten unangetastet und erhalten den sequentiellen Charakter der Datenverarbeitung.

In dieser Arbeit wird gezeigt, dass eine Sequenz derart in Baumstrukturen abgebildet werden kann, dass ein rekurrentes Elman-BRN auf der Sequenz dasselbe leistet wie ein rekursives Elman-Netz (auch: Simple Recurrent Network) auf den Baumstrukturen. Diese Sequenz-zu-Baum-Abbildung wird auf Baumstrukturen verallgemeinert, sodass auch sie bidirektional restrukturiert werden können. Diese Restrukturierung wird als Form-bezogene Vorverarbeitung der Eingabedaten interpretiert.

Es werden neue Restrukturierungsverfahren definiert, also Algorithmen zur Abbildung sequentieller Daten in Baumstrukturen. Das Resultat ist unter anderem ein schnelles Verfahren zur Klassifikation translationsinvarianter Sequenzen. Weiterhin ergibt sich die Möglichkeit, eine nicht-kausale Sequenz-zu-Sequenz-Abbildung zu definieren, die unter gewissen Umständen invertierbar ist. Ein sehr einfach zu implementierendes Verfahren wird vorgestellt. Dieses verwirklicht das Konzept des „teile und herrsche“ und wird zusätzlich mit der bidirektionalen Restrukturierung kombiniert.

Alle vorgestellten Verfahren werden anhand verschiedener Klassifikationsprobleme mit dem rekurrenten Standard, basierend auf LSTM und Elman-Netzen, verglichen. Dazu werden Netze mit nur drei bis fünf Neuronen trainiert. Um ein breites Spektrum an Verwendungsszenarien abzudecken, werden synthetische und Real-world-Daten von diskreter und kontinuierlicher Natur als Eingabedaten verwendet. Die Güte des Trainings wird untereinander verglichen. Für Datensätze mit unausgewogenem Verhältnis zwischen positiven und negativen Mustern wird eine automatisch ausbalancierende Variante des Gradientenabstiegs vorgestellt. Weiterhin wird eine spezielle Initialisierung für das Trainingsverfahren Resilient Backpropagation angegeben.

Es stellt sich heraus, dass die Restrukturierungsverfahren den rekurrenten Standard übertreffen und auch dort erfolgreich sein können, wo rekurrente Netze fehlschlagen, und sie daher unbedingt zwecks Optimierung in Betracht gezogen werden sollten.