DHBW
Duale Hochschule
Baden-Württemberg
Karlsruhe

Entwicklung einer DSL zum Rechnen mit mathematischen Formeln für Anwendungen im Maschinellem Lernen

Studienarbeit

für die Prüfung zum

Bachelor of Science

des Studienganges Angewandte Informatik

an der

Dualen Hochschule Baden-Württemberg Karlsruhe

von

Sebastian Bernauer

Abgabedatum XX.XX.20XX

Bearbeitungszeitraum	XX Wochen
Matrikelnummer	7390071
Kurs	TINF16B5
Ausbildungsfirma	United Internet AG Karlsruhe
Betreuer	Oliver Rettig Marcus Strand

Erklärung

Ich versichere hiermit, dass ich meine Studienarbeit mit dem Thema: „Entwicklung einer DSL zum Rechnen mit mathematischen Formeln für Anwendungen im Maschinellem Lernen“ selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Ort Datum	Unterschrift

Zusammenfassung

TODO

Inhaltsverzeichnis

KAPITEL 1

Motivation

Immer öfter werden Probleme der realen Welt mittels neuronaler Netze gelöst. Allerdings kann man in den meisten Fällen nicht einfach das neuronale Netz auf die gemessenen Daten angewendet werden. Die Daten müssen vorher aufbereitet werden und gegebenenfalls unwichtige Daten entfernt werden. Dies geschieht mittels verschiedener Frameworks in verschiedenen Sprachen. Es soll eine Domain-specific language (DSL) entwickelt werden, die genau auf diesen Anwendungsfall zugeschnitten ist. Durch die DSL soll einen einheitliche Sprache geschaffen werden, welche das Vorprozessieren der Daten vereinfacht. Die Sprache soll plattformübergreifend sein und Central processing unit (CPU)- und Graphics processing unit (GPU)-Berechnungen ermöglichen.

KAPITEL 2

Überblick über Technologien

Das Aufbereiten der Daten für das neuronale Netz kann in mehreren Frameworks in mehreren Sprachen erfolgen. Oft wird für das Vorverarbeiten die gleiche Sprache wie für das neuronale Netz verwendet. Die gängigsten Frameworks sind in Tabelle 2.1 gelistet.

Framework	Sprache
Tensorflow	Python
DL4J	Java

Tabelle 2.1: Die gängigsten Frameworks für maschinelles Lernen

2.1 Unterschied Compiler, Transpiler und Interpreter

Compiler

Übersetzt von einer höheren Sprache in eine niedrigere Sprache.

Transpiler

Übersetzt zwischen zwei Sprachen mit ungefähr gleichem Abstraktionsgrad.

Interpreter

Führt Code einer höheren Sprache direkt aus ohne den Code in eine andere Sprache zu übersetzen.

2.2 ND4J

ND4J ist eine Framework für die Sprache Java, in welchem effiziente Matrizenoperationen durchgeführt werden können. Es kann auf der CPU oder GPU ausgeführt werden. In ND4J ist größtenteils nur das Konstrukt einer Matrix bekannt, Vektoren oder Skalare sind nur ein Spezialfall einer Matrix.

KAPITEL 3

Aufgabenstellung

3.1 Zu Grunde liegendes Framework

In dem Umfeld der Arbeit hat sich das Framework ND4J in der Programmiersprache Java für den Praxiseinsatz durchgesetzt. Daher soll die in dieser Arbeit entwickelte DSL auf diesem Framework aufbauen.

3.2 Compiler oder Interpreter?

Die DSL ist für das Vorprozessieren von Daten für maschinelles Lernen. Für die DSL kommt ein Compiler oder Interpreter in Frage. Ein Transpiler ist nicht möglich, da es keine in der Praxis verwendete Sprache für das Vorprozessieren der Daten gibt, in die übersetzt werden kann. In der Tabelle 3.1 werden die Implementierungsmöglichkeiten mittels Compiler und Interpreter gegenüber gestellt.
Wegen der Vorteile (vornehmlich das Debugging) von Interpretern im Vergleich zu Compiler soll PrePro als Interpreter implementiert werden.

Implementierung	**Vorteile**	**Nachteile**
Compiler	Generierter Java-Code kann auf jeder Java virtual machine (JVM) ausgeführt werden, es wird kein Interpreter benötigt.	Debugging ist nur in dem generierten Java-Code möglich.
Interpreter	Debugging leichter möglich	Möglicherweise nicht so performant

Tabelle 3.1: Vor- und Nachteile einer Implementierung mittels Compiler oder Interpreter

3.3 Verwendete Technologien

Die DSL wird mittels einem Interpreter ausgeführt. Dieser baut auf folgenden Technologien auf:

ND4J

Matrizen-Berechnungen werden mittels dem ND4J-Framework durchgeführt.

Java

Das ND4J-Framework ist in der Programmiersprache Java verfügbar. Damit der Interpreter es verwenden kann, wird in dieser Sprache geschrieben.

Groovy

Groovy ist eine Sprache, die auf Java aufbaut und kompatibel ist. Sie unterstützt zum Beispiel dynamic dispatching[1].

Antlr

Antlr in der Version 4 wird für das Parser der eingegebenen Programme verwendet. Für das Netbeans-Plugin wird Antlr in der Version 3 verwendet.

[1] `https://en.wikipedia.org/wiki/Dynamic_dispatch`

KAPITEL 4

Design der DSL

4.1 Funktionen

In Programmiersprachen werden meist manche Codezeilen häufig benötigt. Anstatt diese Zeilen mehrfach zu kopieren, kann man diese Zeilen in eine sogenannte Funktion packen. Diese Funktionen können an beliebiger Stelle im Code aufgerufen werden. Auf diese Weise kürzt man den entstandenen Code, erhöht die Lesbarkeit und verhindert Kopierfehler.
Daher soll die zu entwickelnde DSL auch Funktionen unterstützen.

Parameter

Funktionen können auch parametrisiert werden, was bedeutet, dass bei dem Aufruf der Funktion Werte mitgegeben werden können. Jede Funktion hat ihren eigenen Variablen-Gültigkeitsbereich, das bedeutet, dass Funktionen ihren Variablen den gleichen Namen geben können, aber unterschiedliche Variablen verwenden. Wenn eine Funktion Werte übergeben bekommen möchte, so muss sie diese mitsamt ihrem Typ angeben.

Rückgabetyp

Funktionen können eine Wert zurückgeben. Dieser muss einen bestimmten Typ haben. In der DSL ist ein Rückgabetyp möglich und muss mittels “returns <Typ>” gekennzeichnet sein. Ist keine Angabe gemacht ist keine Rückgabe vorhanden.

Überladen von Funktionen

Eine Funktion bezeichnet man als überladen, wenn es mehrere Funktionen mit gleichen Namen, aber unterschiedlicher Zahl oder Art von Parametern gibt. In der DSL ist ein Überladen von Funktionen nicht vorgesehen, könnte aber nachträglich noch implementiert werden.

Ein Beispiel von verschiedenen Funktionen befindet sich in Codefragment 4.1 auf der nächsten Seite.

4.1.1 Main-Funktion

Jedes prozedurale Programm benötigt einen Einstiegspunkt, wo das Programm gestartet wird. Da die DSL ein Framework verwendet, welches in Java geschrieben ist, ist es nicht unwahrscheinlich, dass die zukünftigen Nutzer vorher in Java programmiert haben. Daher wurde als Einstiegspunkt des Programms - wie in Java - eine Main-Funktion gewählt. In der DSL besitzt sie keine Parameter. Um Daten in sein Programm zu laden wurde der Ansatz eines DataSets gewählt, mehr dazu in Kapitel 4.4 auf Seite 11.

4.1.2 Weglassen der Main-Funktion

Es gibt mehrere Gründe, warum die Definition einer Main-Funktion unnötig ist:

- Es soll nur ein einziger arithmetischer Ausdruck ausgewertet werden.
- Kompatibilität mit bisher bestehenden anderen Tools, die keine Funktionen bieten, sondern nur eine Liste von Anweisungen entgegennehmen.

Deshalb wird die Main-Funktion in PrePro als optional gesehen und muss nicht deklariert werden. Es reicht aus die Befehle untereinander zu schreiben.
Trotzdem wird es als guter Stil erachtet, eine Main-Funktion zu deklarieren.

```
function main() {
  import vec3 p1, vec3 p2, vec3 p3;

  vec3 x = calculateDifference(p1, p2);
  vec3 s = calculateDifference(p1, p3);
  vec3 y = s X x;
  vec3 z = y X x;


  printResults(x, y, z);

  export x, y, z;
}

function calculateDifference(vec3 p1, vec3 p2) returns vec3 {
```

```
  return p2 - p1;
}

function printResults(vec3 x, vec3 y, vec3 z) {
  print x;
  print y;
  print z;
}
```

Codefragment 4.1: Beispiel Funktionen

4.2 Variablen

Die DSL ist für den Einsatz auf Zeitreihenberechnungen ausgelegt. Daher stellt in der DSL jede Variable eine Zeitreihe dar. Die Operationen der DSL sind immer auf Zeitreihen definiert.
Beispielhaft wird der Ausdruck `x = a - b;` angenommen. In diesem Fall sind `a` und `b` gemessene Zeitreihen von Sensordaten. Die entstehende Variable `x` ist wiederum eine Zeitreihe, welche durch elementweise Subtraktion jedes Zeitelements entstanden ist.
Der Vorteil liegt darin, dass der simple Ausdruck `x = a - b;` sehr leicht les- und wartbar ist. Wenn jede Variable keine Zeitreihe, sondern ein einzelner Messpunkt wäre, müsste man eine Schleife verwenden oder sich eigene Methoden definieren bzw. (falls in der Sprache möglich) die Operatoren überschreiben.
Variablen haben in der DSL immer einen Typ. Bei dem Anlegen einer Variablen muss dieser auch immer definiert werden. Ein Typ ist zum Beispiel ein Vector3 (vec3) oder eine Matrix (mat). Ein Vector3 ist eine Zeitreihe von Vektoren mit der Länge 3, eine Matrix eine Zeitreihe von Matrizen. Ein dem Programm zur Verfügung gestellter Vektor der Länge 3 kann nun als Vector3 oder auch als Matrix aufgefasst werden. Daher muss dem Interpreter beim Anlegen der Variablen immer der Typ mitgeteilt werden. Wenn die Variable schon existiert, muss der Typ nicht erneut angegeben werden.
Ein Beispiel befindet sich in Codefragment 4.2.
Das Import-Statement wird in Kapitel 4.4 auf der nächsten Seite erläutert, relevant ist an dieser Stelle nur, dass mit dem Import Daten aus einem DataSet geladen werden.

```
import vec3 p1, vec3 p2, vec3 p3;

vec3 x = p2 - p1;
vec3 s = p3 - p1;
vec3 y = s X x;
vec3 z = y X x;
```

Codefragment 4.2: Beispiel Variablenzuweisung

4.3 DataSet

Ein DataSet ist in der DSL eine Sammlung von Variablen der DSL. In das DataSet können beliebig viele Variablen unter ihrem Namen gespeichert werden. Es ist nicht möglich zwei Variablen mit dem gleichen Namen abzulegen. Eine Variable kann bequem aus dem DataSet ausgelesen werden.

4.4 Import & Export

Ein Programm, das nur Berechnungen anstellt erscheint auf den ersten Blick sinnlos. Das Programm muss die Möglichkeit haben, Daten zu lesen und zu schreiben. Im Falle der DSL wird ein eigenes DataSet definiert. Das Programm erhält bei der Ausführung ein DataSet und gibt als Ergebnis wieder ein DataSet zurück. In das Eingabe-DataSet werden alle Variablen gespeichert, die für die Berechnungen benötigt werden. In dem Ausgabe-DataSet sind anschließend alle Variablen gespeichert, die berechnet wurden.
Die Verwendung eines DataSet hat gegenüber dem Hereingeben mittels Parametern in die Main-Funktion folgende Vorteile:

- Es gibt nur einen Rückgabetyp (DataSet). Andernfalls müsste ein Konstrukt ersonnen werden, mehrere Variablen von der Main-Funktion zurückgeben zu lassen.

- Einfacher Aufruf der Main-Funktionen (ab 4 Parametern wird der Funktionsaufruf unübersichtlich[1]). Statt 20 Parameter zu übergeben kann übersichtlich das DataSet zusammengebaut werden und als einziges Argument übergeben werden.

- Einfaches “Weiterschleifen” von DataSets zwischen mehreren PrePro-Programmen. Falls mehrere PrePro-Programme nacheinander ausgeführt werden kann bequem das Ausgabe-DataSet des ersten Programms als Eingabe-DataSet des zweiten Programms genommen werden.

4.5 Kommentare

Kommentare sollen in der DSL möglich sein.
Einen Zeilen-Kommentar wird ein “//” vorangestellt.
Ein Block-Kommentar wird mir “/*” und “*/” umschlossen.

KAPITEL 5

Implementierung der DSL

5.1 Lexer & Parser

Lexer und Parser werden beide von einer ANTLR4 Grammatik erzeugt.

5.2 Typsystem

Das Typsystem von PrePro ist in Abbildung 5.1 auf der nächsten Seite dargestellt. Alle Variablen erben von er abstrakten Klasse `Variable`. Es gibt die Untertypen `Vector`, `Marix`, `Scalar` und `Constant`. Die Unterklassen `Vector` und `Matrix` haben wiederum Unterklassen für drei- und vierelementige Varianten. Diese Unterklassen sind wichtig, da z.B. eine `Matrix3` mit einem `Vector3` multipliziert werden kann, allerdings nicht mit einem `Vector4`.
In der abstrakten Klasse `Variable` sind die Funktionen `add`, `sub`, `mul` und `div` definiert. Werden die Funktionen auf der abstrakten Klasse aufgerufen, werfen sie eine Exception, dass die mathematische Operation nicht definiert sei.
Die Unterklassen haben nun die Möglichkeit, Operationen mit anderen Typen zu definieren. Mittels Polymorphie und dynamic dispatching wird bei einer arithmetischen Operation die passende Funktion gesucht. Falls keine passende Funktion wird die allgemeine - in der abstrakten Klasse definierte - Funktion verwendet, und daraufhin eine Exception geworfen, dass die mathematische Operation nicht definiert sei.

5.2.1 Darstellung der Typen in Java

Alle Typen werden als `INDArray` von `ND4J` dargestellt.
Die Unterklassen besitzen einen Konstruktor, der ein `INDArray` übergeben bekommt. In den Konstruktoren wird jeweils geprüft, ob die Dimensionen dem entsprechenden Typ entsprechen (z.B eine Matrix4 muss eine 4x4-Matrix übergeben bekommen). Die einzige Ausnahme bildet der Typ “Constant”, dieser bietet zusätzlich einen Konstruktur, dem ein `double` übergeben werden kann. Intern wird aus dem `double` ein `INDArray` erzeugt.

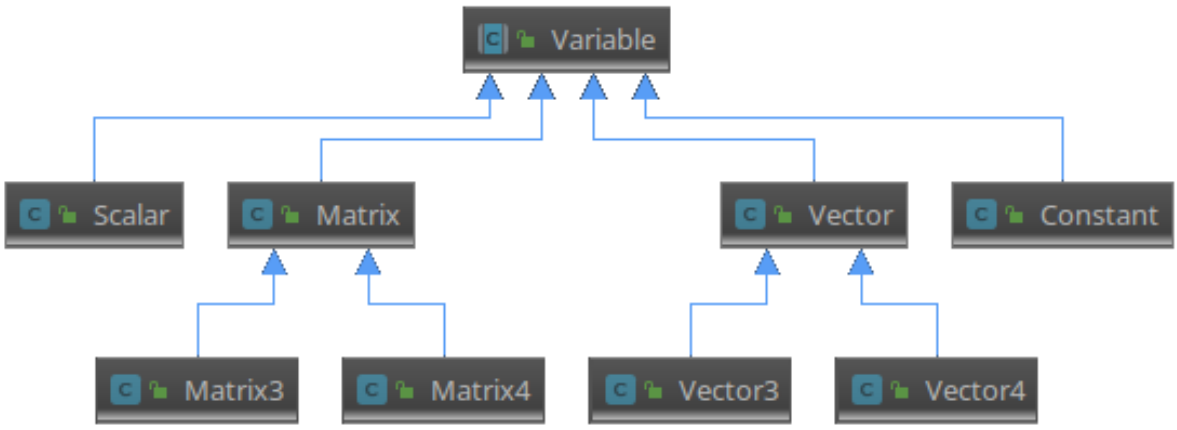

Abbildung 5.1: Typsystem von PrePro

5.3 Grundrechenarten

Bei der Darstellung der Zahlen im Speicher wird für jede Zahl ein Double verwendet. Das erhöht die Genauigkeit gegenüber einem float und erspart Konvertierungen zwischen Ganz- und Fließzahlen. Grundrechenarten sind die Addition, Subtraktion, Multiplikation und Division von Matrizen. Sie Operationen sind elementweise und werden direkt von ND4J zur Verfügung gestellt und können aufgerufen werden.

5.4 Kreuzprodukt

Anders als die vier Grundarten aus dem vorherigen Abschnitt wird das Kreuzprodukt von ND4J nicht als Operation angeboten. Das Kreuzprodukt wird in der Praxis allerdings zu Beispiel für das Aufspannen von Vektoren im dreidimensionalen Raum benötigt. Daher wird an dieser Stelle das Kreuzprodukt zweier Vektoren selber implementiert. Falls ND4J in Zukunft den Operator Kreuzprodukt bereit stellt, kann in zukünftigen Varianten auf ihre Implementierung zugegriffen werden, da diese wahrscheinlich effizienter sein wird. Die eigene Implementierung des Kreuzprodukts kann auf folgende Arten geschehen:

1. Implementierung in Plain Java mittels einer for-Schleife.

2. Implementierung in Plain Java mittels Subarrays

5.4.1 Implementierung Kreuzprodukt mittels For-Schleife

Bei der Implementierung mittels der For-Schleife werden die eigentlichen Berechnungen in Java durchgeführt. Als erstes wird ein Double-Array als Zwischenspeicher für das Ergebnis angelegt. Es besitzt (Anzahl der Zeitelemente in den Eingabe-Vektoren) * drei Elemente. Die Anzahl der Elemente entspricht so der Anzahl der Elemente der Ergebnismatrix, die Daten können in dem Double-Array effizient gespeichert werden. Das Array wird im Anschluss in eine Matrix konvertiert. Für die Berechnung der Elemente wird mittels einer For-Schleife über alle Zeilen der Matrix iteriert. In jeder Zeile werden die 3 Werte des entstehenden Ergebnisvektors berechnet und in das Double-Array gespeichert. Abschließend wir das Double-Array in eine Matrix mit den Dimensionen [<Anzahl Zeitelemente> x drei] konvertiert und durch eine Vector3-Wrapper-Klasse als Vector mit 3 Werten gekennzeichnet. Der Algorithmus ist in Codefragment 5.1 auf der nächsten Seite dargestellt.

```
private Vector3 crossProduct(Vector3 left, Vector3 right) {
    INDArray a = left.getNdArray();
    INDArray b = right.getNdArray();

    int size = a.shape()[0];
    double[] result = new double[size * 3];

    for (int i = 0; i < size; i++) {
        result[i * 3 + 0] = a.getDouble(i, 1) *
           b.getDouble(i, 2) - a.getDouble(i, 2) *
           b.getDouble(i, 1);
        result[i * 3 + 1] = a.getDouble(i, 2) *
           b.getDouble(i, 0) - a.getDouble(i, 0) *
           b.getDouble(i, 2);
        result[i * 3 + 2] = a.getDouble(i, 0) *
           b.getDouble(i, 1) - a.getDouble(i, 1) *
           b.getDouble(i, 0);
    }
    return new Vector3(Nd4j.create(result, new int[]{size,
       3}));
}
```

Codefragment 5.1: Implementierung Kreuzprodukt mittels for-Schleife

5.4.2 Implementierung Kreuzprodukt mittels Subarrays

```
private Vector3 crossProductSubArray(Vector3 left, Vector3
   right) {
    INDArray a = left.getNdArray();
    INDArray b = right.getNdArray();

    INDArray a1 = a.getColumn(0);
    INDArray a2 = a.getColumn(1);
    INDArray a3 = a.getColumn(2);

    INDArray b1 = b.getColumn(0);
    INDArray b2 = b.getColumn(1);
    INDArray b3 = b.getColumn(2);

    INDArray c1 = (a2.mul(b3)).sub(a3.mul(b2));
    INDArray c2 = (a3.mul(b1)).sub(a1.mul(b3));
    INDArray c3 = (a1.mul(b2)).sub(a2.mul(b1));

    int size = a.shape()[0];
    INDArray result = Nd4j.create(size, 3);
    result.putColumn(0, c1);
    result.putColumn(1, c2);
    result.putColumn(2, c3);

    return new Vector3(result);
}
```

Codefragment 5.2: Implementierung Kreuzprodukt mittels for-Schleife

5.5 Vergleich der Implementierungen für Kreuzprodukt

Beide Impelemtierungsmöglichkeiten haben Vor- und Nachteile. Diese sind in Tabelle 5.1 auf der nächsten Seite aufgeführt.

Implementierungs-möglichkeit	**Vorteile**	**Nachteile**
Mittels For-Schleife	Leichter verständlich.	Wird direkt in Java ausgeführt. Mögliche Optimierungen von ND4J können nicht verwendet werden. Berechnungen finden nur auf der CPU statt!
Mittels Subarray	Durch die Verwendung von ND4J können die Optimierungen verwendet werden. Wenn ND4J so konfiguriert ist, dass es auf der GPU läuft, kann die eigentliche Berechnung weiterhin auf der GPU erfolgen.	Schwerer verständlich.

Tabelle 5.1: Vor- und Nachteile einer Implementierung mittels Compiler oder Interpreter

Für große Zeitreihen müsste sich die Implementierung mittels dem Subarray als effizienter erweisen, besonders wenn die Berechnungen auf der GPU durchgeführt werden. Als Nachweis und für das Effizienzverhalten bei kleinen Zeitreihen wurde ein Benchmark durchgeführt. Die Ergebnisse sind in Tabelle 5.2 festgehalten.

Anzahl Datensätze	**For-Schleife**	**Subarray**
100	6 (342)	4 (13)
1.000	20 (489)	7 (22)
10.000	53 (527)	5 (21)
100.000	333 (1159)	5 (15)
1.000.000	3205 (4129)	47 (64)
10.000.000	32358 (33594)	442 (453)

Tabelle 5.2: Benchmark-Ergebnisse in ms der verschiedenen Implementierungsmöglichkeiten für das Kreuzprodukt. Die Messung ist nach 5 Durchläufen gemessen, die Zahl in Klammern gibt die Zeit des ersten Durchlaufs an.

3 Varianten

Benchmarks! Klassifizierung CPU oder GPU-Workload.

Möglicherweise lohnt sich die eher GPU-betonte Variante erst ab gewisser Größe. => Dann mit konstantem Aufwand entscheiden, welches Verfahren. Vom Nutzer (während Laufzeit) auswählbar?

KAPITEL 6

Integration in NetBeans-IDE

Glossar

Inhalt

Daten, welche auf den Portal-Homepages[1] angezeigt werden, z.B. Lottodaten, Wetterdaten, Bundesliga-Liveticker und das Horoskop.

Portal-Homepage

Die Startseite einer der Portale web.de, gmx.net, gmx.ch, gmx.at und home.1und1.de.

[1] Die Startseite einer der Portale web.de, gmx.net, gmx.ch, gmx.at und home.1und1.de

Abkürzungsverzeichnis

DSL Domain-specific language 4

JVM Java virtual machine 6

CPU Central processing unit 4

GPU Graphics processing unit 4

Abbildungsverzeichnis

Literatur

[1] Rober C. Martin. *Robert C. Martin's Clean Code Tip of the Week #10: Avoid Too Many Arguments*. 2009. URL: http://www.informit.com/articles/article.aspx?p=1375308 [besucht am 21. 02. 2019] [siehe S. 11].