JOHANNES KEPLER
UNIVERSITY LINZ

Dipl.-Ing. Raphael Mosaner, BSc.
Institute for System Software

T +43 732 2468 4347
F +43 732 2468 4345
raphael.mosaner@jku.at

Secretary:
Karin Gusenbauer
Ext 4342
karin.gusenbauer@jku.at

Bachelor Thesis

Low Overhead Neural Network Predictors in a Dynamic Compiler

Student: David Köllner
SKZ/Matr.Nr.: 521 / 11777498

Email: davidkoellner@gmx.at

Advisor: Dipl.-Ing. Raphael Mosaner, BSc

Start date: 15.03.2022

In an ongoing research project, machine learning is leveraged into the domain of dynamic compilers. In contrast to static compilers, compile time overhead due to neural network loading and inference directly impacts overall run time. For this project, a slim ANN predictor should be implemented in the Graal [1] compiler and compared to existing more versatile ML frameworks. The question is, how a slim predictor solely implemented in Java performs against complex frameworks when it comes to prediction only.

The goals of this thesis are:

- Load neural network parameters in Java
  - Preferably stored in the ONNX [2] format
  - The OnnxRuntime [6] library can be used for loading
  - [Write own parser]
- Implement an inference model in Java based on the loaded parameters
  - Supported neural network layers: Linear [3], ReLu [4], [BatchNorm [5]]
  - Search related work regarding arithmetic optimizations to speed up calculations
  - [Write a model generator which hardcodes all inference parameters]
- Write a phase in the Graal compiler, which uses the model for inference
  - Parameterizable, to adapt the number of predictions for evaluation
- Evaluate the predictor against existing frameworks
  - Suggestions are: OnnxRuntime [6] and Tribuo [7]

Implementing a predictor which comes any close to existing frameworks is considered out of scope and is an explicit non-goal. Bullets in brackets are considered secondary objectives, depending on the progress of the project.

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
www.jku.at
DVR 0093696

Modalities

The progress of the project should be discussed at least every two weeks with the advisor. A time schedule and a milestone plan must be set up within the first 3 weeks. It should be continuously refined and monitored to make sure that the thesis will be completed in time. The final version of the thesis must be submitted not later than 15.09.2022.

[1] https://github.com/oracle/graal
[2] https://github.com/onnx/onnx
[3] https://pytorch.org/docs/stable/generated/torch.nn.Linear.html
[4] https://pytorch.org/docs/stable/generated/torch.nn.ReLU.html
[5] https://pytorch.org/docs/stable/generated/torch.nn.BatchNorm1d.html
[6] https://onnxruntime.ai/
[7] https://tribuo.org/

JOHANNES KEPLER UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
www.jku.at
DVR 0093696