ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

GNNs for TPU Code Graphs

An AI model can be represented as a graph, where a node is a tensor operation (e.g. matrix multiplication, convolution, etc), and an edge represents a tensor. A compilation configuration controls how the compiler transforms the graph for a specific optimization pass. In particular, there are two types of configurations/optimizations:

- A layout configuration control how tensors in the graph are laid out in the physical memory, by specifying the dimension order of each input and output of an operation node.
- A tile configuration controls the tile size of each fused subgraph.

Being able to predict an optimal configuration for a given graph will not only help to improve the compiler's heuristic to select the best configuration without human's intervention. This will make AI models run more efficiently, consuming less time and resources overall!

In this thesis, your aim is to take part in the Google Kaggle competition train a machine learning model based on the runtime data provided to you in the training dataset and further predict the runtime of graphs and configurations in the test dataset.

Requirements: Strong motivation, knowledge in neural networks and machine learning, as well as good coding skills. Prior practical experience with neural networks is a big advantage. We will have weekly meetings to discuss open questions and determine the next steps.

Interested? Please contact us for more details!

Contact

In a few short sentences, please tell us why you are interested in the project and about your coding and machine learning background (i.e., your own projects or courses).

- Florian Grötschla: fgroetschla@ethz.ch, ETZ G63
- Joël Mathys: jmathys@ethz.ch, ETZ G63