ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

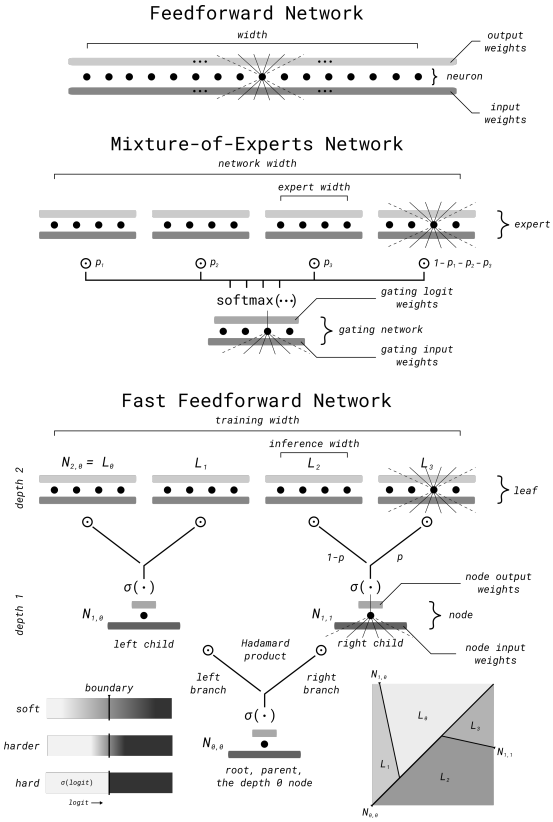
Prof. R. Wattenhofer

Evaluating Fast FeedForward Networks

Over recent years, language and vision models have shown themselves to be excellent at many tasks. However, this excellence has come at a large computational cost, which methods such as Mixture-of-Experts try to address.

Another approach that has been proposed is Fast Feedforward Networks, which speed up networks by breaking up the linear layers to activate only a subset of the models' neurons. Some work has been done to evaluate the benefits of Fast Feedforward Networks when applied to language models; however, these evaluations have been incomplete.

This project focuses on doing a proper and thorough evaluation of Fast Feedforward Networks when applied to large language and vision models, thus assessing whether using Fast Feedforward Networks is beneficial over using Mixture-of-Experts.



Requirements

Good programming skills (Python, C / C++, etc.) and a good knowledge of machine learning and machine learning libraries. Knowledge of CUDA programming and HPC concepts is an advantage.
We will have weekly meetings to address questions together, discuss progress, and think about future ideas.

Contact

In a few short sentences, please explain why you are interested in the project and about your coding and machine learning background (i.e., your own projects or relevant courses you have taken at ETH or elsewhere).

- Andreas Plesner: aplesner@ethz.ch, ETZ G95