Integration of Deep Learning Backends Using Collage

The thesis will contribute to the research on Collage for integration of Deep Learning (DL) backends and provide insights into the challenges in this field.

The strong demand for efficient and performant deployment of DL applications prompts the rapid development of a rich DL ecosystem. To keep up with this fast advancement, it is crucial for modern DL frameworks to efficiently integrate a variety of optimized tensor algebra libraries and runtimes as their backends and generate the fastest possible executable using these backends. However, current DL frameworks require significant manual effort and expertise to integrate every new backend while failing to unleash its full potential. Given the fast-evolving nature of the DL ecosystem, this manual approach often slows down continuous innovations across different layers; it prevents hardware vendors from the fast deployment of their cutting-edge libraries, DL framework developers must repeatedly adjust their hand-coded rules to accommodate new versions of libraries, and machine learning practitioners need to wait for the integration of new technologies and often encounter unsatisfactory performance.

Collage, a DL framework that offers seamless integration of DL backends. Collage provides an expressive backend registration interface that allows users to precisely specify the capability of various backends. By leveraging the specifications of available backends, Collage automatically searches for an optimized backend placement strategy for a given workload and execution environment.

Your work:

1. Conduct a comprehensive literature review on Collage and similar frameworks
2. Conduct an experiment on Collage including a heterogenous system including UMA-integrated backends in TVM

Requirements:

- Fundamental understanding of Deep Learning and embedded systems
- Basic understanding of TVM compiler
- Experience in programming C\C++ and Python
- Self-motivation and ability to work independently

References:


[UMA] https://tvm.apache.org/docs/tutorial/uma.html

If you are interested in this topic, please contact me at samira.ahmadifarsani@tum.de.