IBM Research –Zurich

Contact: Dr. Abbas Rahimi
Email: abr@Zurich.ibm.com

The thesis will be performed at the IBM Research-Zurich in Rüschlikon.

Master Projects

Accelerating Transformers with Computational Memory

Introduction

Transformers are a type of neural network architecture that have been gaining popularity. Transformers were developed to solve the problem of sequence transduction, or neural machine translation. That means any task that transforms an input sequence to an output sequence. This includes speech recognition, text-to-speech transformation, and recently expanded the to image domain as well. Nevertheless, their computational complexity is massive. Feed-forward layers constitute two-thirds of a transformer model's parameters opening up opportunities for efficient mapping on non-von Neumann architectures such as computational memories [1].

Goal

We aim at efficiently mapping transformer models on computational memory as an emerging hardware fabric

Tasks and Type

There are several challenges that need to be overcome at algorithmic (40%), and hardware (60%) levels to realize such efficient transformers including developing novel methods for quantization, transformations, and hardware-aware retraining. We are inviting applications from students to conduct their Master's thesis work on this exciting new topic. The work performed could span high-level algorithmic developments all the way to efficient realization on emerging hardware using phase-change material devices at scale. It also involves interactions with several researchers across IBM research focusing on various aspects of the project. The ideal candidate should have a multi-disciplinary background, strong mathematical aptitude and programming skills. Prior knowledge on machine learning, and architectures is a bonus but not necessary.