INFERENCE OPTIMIZATION FOR EDGE AI DEPLOYMENT

Context

Neural networks are generally task-specific over parametrized models, excelling at several applications. Direct deployment on less powerful hardware is not a trivial task.

As developments in AI mature and move from concept to application, industries look to integrate this intelligence into their tooling and machinery. Hence intelligent systems are as close as possible to individual processes allowing for better insight or automated decision-making. However, the resources available when developing these models differ significantly from those present at the edge. Limits on processing power, memory, or inference time require consideration.

A possible solution is modeling with the target hardware in mind. However, it would require adapting the entire machine learning pipeline every time the hardware changes. A different approach is to employ techniques that allow pre-trained models to work within the target hardware constraints. In this manner, a hardware revision may only require an adaptation of the existing model.

Internship overview

- Master Student
- Internship
- Mathware
- Location: Eindhoven

Technologies

- Deep Learning
- Convolutional neural networks
- Image processing
- Edge AI

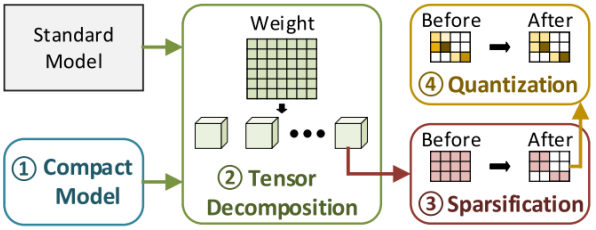
SIOUX TECHNOLOGIES

Sioux Technologies
Eindhoven | Mijdrecht | Delft | Apeldoorn

+31 (0)40 2677 100
jobs@sioux.eu
jobs.sioux.eu

Assignment

Researching and summarizing recent trends in model compression for edge deployment. Selecting a hardware platform for simulated deployment. Identifying the most promising techniques for efficient model deployment on the edge.

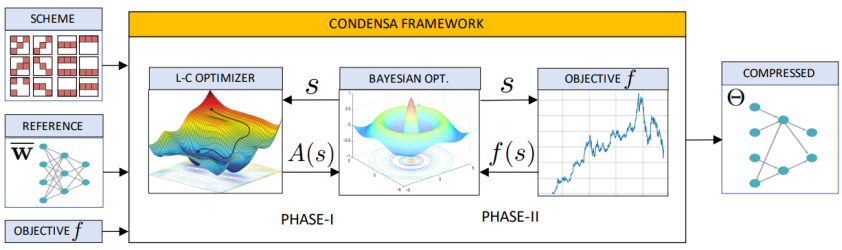


Implementing the chosen model compression techniques and deploying on the hardware platform. Studying and comparing the performance of the identified techniques across different metrics (e.g. memory consumption, inference time, performance degradation).

Analyzing the model performance before and after inference optimization for different computer vision tasks.

Activities

- A preliminary study of state-of-the-art techniques and frameworks for model compression.
- Study how do the selected techniques impact task-specific performance as well as deployment-specific performance (memory footprint, inference time).
- Study the applicability of these techniques to diverse computer vision tasks (classification, detection, segmentation).



Why choose Sioux?

- Working on innovative technology
- Challenging, dynamic and varied work
- A comfortable and personal work environment
- Plenty of opportunities for personal development
- Great carreer opportunities
- Contributing to a safe, healthy and sustainable society

Get in touch!
Would you like to know more about this student assignment?

Contact:
Raffaele Imbriaco
+31 (0)40 751 61 16
werving_mathware@sioux.eu