SCIENCE · PASSION · TECHNOLOGY

Open Thesis / Project
Model Compression through Iterative Constrained Knapsack Optimization

Embedded Learning and Sensing Systems Group

Motivation

Deep neural network models deployed on edge devices frequently encounter resource variability due to fluctuating energy levels, timing constraints, or the prioritization of other critical tasks within the system. State-of-the-art (SOTA) machine learning pipelines produce resource-agnostic models, which are not capable of adapting at runtime. We propose Resource-Efficient Deep Subnetworks (REDS) to address model adaptation to variable resources. REDS identifies a set of nested subnetworks by formulating and solving a global iterative knapsack problem. REDS surpasses SOTA in neural architecture search (μNAS: Constrained Neural Architecture Search for Microcontrollers) and subnetwork methods (DRESS: Dynamic REal-time Sparse Subnets) regarding search time and memory overhead. We aim to further investigate the knapsack solver formulation proposed in REDS when applied iteratively to each layer of the model progressively.

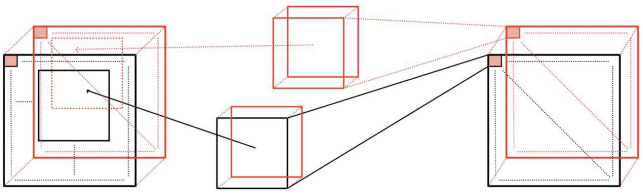
Interested? Please contact us for more details!

Target Group

Students in ICE, Computer Science or Software Engineering.

Thesis Type

Master Project / Master Thesis.



Convolutional neural network model compression

Goals and Tasks

The goal of this work is to extend our novel model compression knapsack solver to support layer-wise iterative structural pruning and fine-tuning. The implementation will be tested on computer vision datasets, and its performance will be analyzed through TensorFlow Lite and Edge Impulse benchmarks. The project includes the following tasks:

- In depth understanding of our previous work REDS and its subnetworks structure;
- Familiarize yourself with model compression, *i.e.*, structural pruning, and the knapsack solver formulation;
- Implement the algorithm and analyze its performance;
- Summarize the results in a written report.

Requirements:

- Eager to learn about deep neural network compression and on-device adaptation to dynamic resource constraints;
- Programming skills in Python;
- Prior experience of machine learning framework (TensorFlow) and linear programming framework (OR-Tools) is recommended but not mandatory.

Used Tools & Equipment

- A laptop (GPU infrastructure will be provided if needed).

Contact Persons

- Francesco Corti (francesco.corti@tugraz.at)
- Assoc. Prof. Olga Saukh (saukh@tugraz.at)

Institute for Technical Informatics
Embedded Learning and Sensing Systems Group