PhD proposal : Dynamic Neural Network compression

Context:

Research on machine learning and Deep Neural Networks (DNN) has made considerable progress in the past decades. State-of-the-art DNN models usually require large amounts of data to be trained and contain a tremendous number of parameters leading to overall high resource requirements, in terms of computation and memory and thus energy. In the past years, this gave rise to approaches to reduce these requirements, where, for example, during or after training, parts of the model are removed (pruning) or stored with lower precision (quantisation) or surrogate models are trained (knowledge distillation) or where the best configuration is searched by testing different parameters (Neural Architecture Search, NAS). Also, concerning the hardware, many optimisations have been proposed to accelerate the inference of DNNs on different architectures.

But these accelerators are usually specific to a given hardware and are optimised to satisfy certain static performance criteria. However, for many applications, the performance requirements of a DNN model deployed on a given hardware platform are not static but evolving dynamically as its operating conditions and environment change. In the context of this ANR-funded project we propose an original interdisciplinary approach that allows DNN models to be dynamically configurable at run-time on a given reconfigurable hardware accelerator architecture, depending on the external environment, following an approach based on feedback loops and control theory.

Objectives:

At software (SW) level, for a given DNN model, different variants with incremental precision levels can be obtained by setting parameters along different dimensions: (i) data precision or quantization (increasing/decreasing bit-width of activations and/or weights), (ii) degree of sparsification (e.g., pruning, tensor decomposition), (iii) depth of the NN (number and type of network layers to execute). Depending on the chosen SW precision level, the mean output accuracy will change, as well as energy consumption and timing. The key observation is that, for some particularly “easy” inputs, using high-precision energy-hungry computations is an “overkill”. Conversely, for “hard” inputs, low-precision energy-efficient computations are not enough. Therefore, being able to dynamically change the SW precision is key to enable energy-efficient and accurate NN computations.

At the same time, at the hardware (HW) level, the DNN accelerator needs to be configurable at runtime to satisfy SW processing requirements. Different degrees of existing HW strategies have an impact on energy consumption and runtime, which need to be taken into account when designing NN architectures and SW compression schemes.

This PhD thesis will concentrate on the SW (i.e. machine learning) side. There are two major challenges that need to be addressed. First, an effective method is required to train different DNN model variants for the same task but responding to different performance criteria, i.e., providing different levels of classification accuracy, throughput, energy consumption etc. Also, to reduce memory requirements, the more complex variants should be able to mutualize as much parameters as possible with the simpler models. Based on existing approaches and results on structured pruning and multi-exit models, one objective is to develop new algorithms and DNN architectures that can be parameterized at inference time to execute only the parts of the model that are really needed. For some layers, different parameter or activation precisions (8, 16, 32-bit) may be necessary and should be able to be switched at run time depending on the sensitivity of the different parts of the model, that should be quantified. The second challenge concerns the influence of these architectural changes on the performance. The impact of different possible SW configurations must be assessed in terms of final accuracy of the obtained results and in terms of the corresponding effort (e.g., in terms of number of operations, their datatype, potential computational over-head etc.).

Environement:

The PhD is funded by the ANR project RADYAL starting as soon as possible. It will be conducted at INSA Lyon and the LIRIS laboratory in Lyon (Doua campus) under the supervision of Stefan Duffner (MCF HDR) and Christophe Garcia (PR).
Teaching activities may be done in parallel at INSA Lyon.

Requirements:

- A Master degree in computer science or applied mathematics or similar.
- A strong background in machine learning and in particular neural networks
- Capacity to work autonomously and within a research team
- Scientific curiosity and creativity
- Very good English proficiency

Contact:

To apply please send your CV, grade records and cover letter (and potentially recommendation letters) to Stefan Duffner stefan.duffner@insa-lyon.fr