Methods and Applications for Low-power Deep Neural Networks on Edge Devices

Summary of the dissertation:

Methods and Applications for Low-power Deep Neural Networks on Edge Devices

Ph.D. in Electrical, Electronics and Telecommunication Engineering – Cycle XXXV

Ph.D. Candidate: *Luciano Prono*

Ph.D. Coordinator: *Stefano Grivet Talocia*

Supervisor of the Ph.D Candidate: *Gianluca Setti*

Machine Learning (ML) introduces the possibility of solving complex tasks requiring humongous amounts of data. In particular, Deep Neural Networks (DNNs) are structures that provide excellent results with the use of a massive number of trainable parameters. At the same time, the implementation of DNNs on cheap and low-power devices with high resource constraints, referred to as tinyML, would unlock a large number of new applications possible in the fields of Internet of Things (IoT) and Edge Computing. In this dissertation, we focus on the design of low-power, low-memory DNN structures and we implement them with novel Edge Computing approaches such as an innovative Compressed Sensing (CS) DNN-based decoder.

Part I of the dissertation revolves around the model compression techniques and DNN architectures that can be used to implement neural network structures on low-end, low-power devices with high resource constraints.

In Chapter 1 state-of-the-art model compression techniques, namely quantization of the parameters and pruning of the interconnections, are introduced.

A novel Sum-and-Max based map-reduce paradigm, used to build multiplier free DNN layers, is introduced in Chapter 2. The Sum-and-Max (SAM) map-reduce paradigm for DNNs has been researched in the attempt of developing a structure that can be entirely implemented without the use of expensive and power-hungry multipliers. This structure is used as a substitute of the common MAC paradigm. For each neuron, the weights are summed to the inputs, then the maximum value and the minimum value are selected and added together. The structure is intrinsically non-linear and does not require the use of an activation function. Also, this kind of structure is naturally prone to pruning, i.e., many weights can be removed as not necessary, with a big reduction in memory requirements.

Following the example of the Sum-and-Max paradigm, in Chapter 3 the novel Multiply-and-Max&Min paradigm is introduced, with performance comparable to standard Multiply and Accumulate structures but highly resistant to the pruning process. As SAM is difficult to train using Stochastic Gradient Descent (SGD) and backpropagation due to the difficulty of propagating the gradient (gradient propagates only through the interconnections that are selected through maximum and minimum reduce operations), MAM² is trained with a hybrid approach, i.e., a MAM²-based DNN layer is first trained as a standard MAC-based layer and then gradually converted to a MAM² layer with the *vanishing contributes* technique. Additionally, MAM² can be pruned with any state-of-the-art pruning methods with great results, as it is intrinsically much more resistant to pruning compared to classic MAC structures. In Figure 1 is shown a comparison of a MAC-based VGG-16 against a MAM²-based VGG-16 on the ImageNet task.

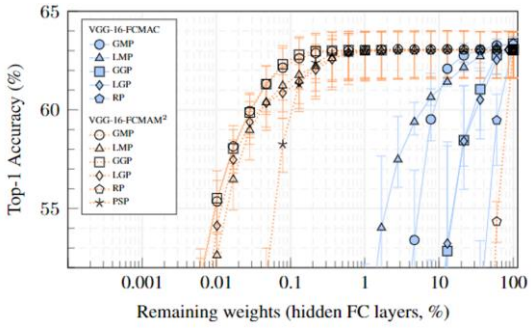

Figure 1: top-1 accuracy of VGG-16 on ImageNet dataset. Layers are pruned using several different one-shot pruning methods, both for the MAC-based structure and for the MAM²-based structure.

For many different non-structured one-shot pruning methods, MAM² map-reduce paradigm guarantees far better result with a very low number of remaining parameters. This structure is tested on multiple big-size standard image-vision datasets and networks. Additionally, an on-device implementation of a MAM²-based pruned DNN autoencoder is tested and analyzed.

In Chapter 4 is presented an on-device Recurrent Spiking Neural Network (RSNN), whose intrinsic sparse characterization both in space and time allows a big reduction in terms of energy consumption of the structure. The structure has been implemented on an ARM Cortex-M4 microcontroller and tested on the Spiking MNIST dataset. Results show that the maximum delay for each time step is 40.6 µs with a clock frequency of 180 MHz, while the average energy per time-step is 4.1 µJ.

Part II introduces a novel approach to the decoding phase of the CS paradigm which uses a DNN structure. The aim of the part is the presentation of the novel CS decoder and the assessment of its reconstruction performance on bio-signals, even when compressed by means of parameters quantization and pruning.

In Chapter 5, the novel decoder, composed of two distinct working blocks, namely a DNN-based Support Oracle and a pseudoinverse-based reconstruction block, is described and its performance is assessed with synthetic bio-signal datasets. In particular, we have a signal $\boldsymbol{x} \in \mathbb{R}^n$ that is *sparse*, i.e., when projected over an adequate orthogonal sparsity basis $\boldsymbol{D}$, we obtain a sparse vector $\boldsymbol{\xi} = \boldsymbol{D}\boldsymbol{x}$ of which only $k < n$ components are non-zero. Then, given a *sensing* matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with $m < n$ and an acquired *measurement vector* $\boldsymbol{y} \in \mathbb{R}^m$, the compression operation is the projection of $\boldsymbol{x}$ over the rows of $\boldsymbol{A}$, i.e., $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. At this point, it is possible to retrieve the compressed signal by using a two-step approach:

- a DNN-based Support Oracle divines the *support* of $\boldsymbol{\xi}$ from the measured signal $\boldsymbol{y}$, i.e., the positions of the non-zero contributes;
- using the retrieved support from the oracle, the typically ill-conditioned problem of retrieving $\boldsymbol{\xi}$ becomes overdetermined and can be easily solved by solving the Least Mean Square problem (e.g., with the use of Moore-Penrose pseudoinverse operation).

This CS decoder has been proposed in two versions, namely Trained CS with Support Oracle (TCSSO), which uses SGD both for optimizing the DNN and the sensing matrix $\boldsymbol{A}$, and Trained Support Oracle for Compressed signals (TSOC), that uses the Rakeness approach to adapt $\boldsymbol{A}$ but is optimized also for non-perfectly sparse signals.

The two-step CS decoders have been tested both with EEG and ECG signals using antipodal sensing matrices and compared with other state-of-the-art reconstruction methods. Results for TCSSO are shown in Figure 2 (for ECG).

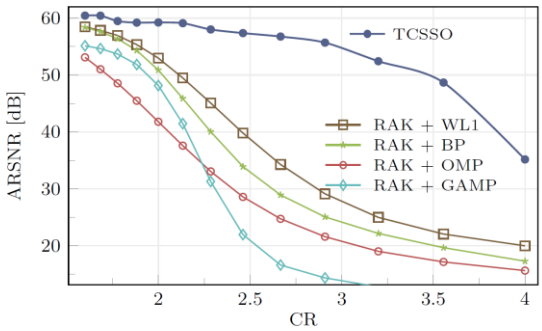

Figure 2: Average Reconstruction Signal to Noise Ratio (ARSNR) with sparse ECG signals for different Compression Ratios (CS).

In Chapter 6, the techniques and special architectures introduced in Part I are applied to the CS decoder structure, and the performance with the compressed models is assessed. In particular, MAM² layers are used to prune the support oracle in TCSSO, a preliminary post-training quantization analysis is performed on the TCSSO structure and finally the improved TSOC structure is fully quantized using quantization-aware techniques. Results show that all these compression methods allow a considerable reduction of the system size minimizing the effects on its reconstruction performance.