ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVII Cycle

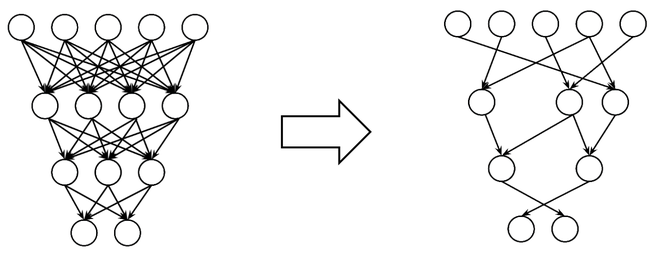
Pruning for DNN simplification with non-conventional structures

Philippe Bich

Supervisor: Prof. Gianluca Setti

Research context and motivation

- **Deep Neural Networks** (DNNs) are structures capable of solving complex tasks with the **use** of **a massive number of trainable parameters**. They are so many **that** their number **results to be greatly redundant**; hence it is possible to prune the unnecessary parameters by removing interconnections between neurons that do not appreciably influence the accuracy of the DNN task.



Before pruning After pruning

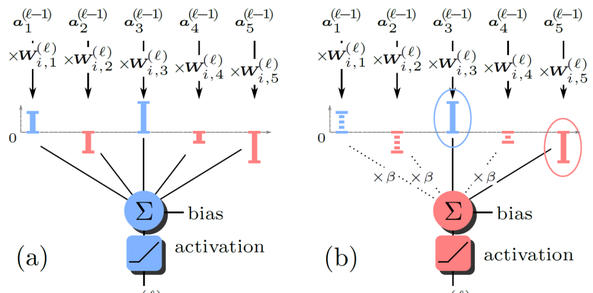
- **Pruning is a necessary operation to achieve low-power, low-latency, and lightweight machine learning inference**, which is fundamental for the implementation of neural networks on mobile devices with limited battery and computational performance. The use of DNNs on mobile Internet of Things (IoT) devices is of high interest as many are the possibilities it unlocks.
- Many works on DNNs applied to Internet of Things (IoT) and mobile computing can be found for example in the fields of Augmented Reality, Natural Language Processing, Computer Vision and also Compressed Sensing for biomedical signals.

Addressed research questions/problems

- The problem of DNN pruning has been addressed in multiple ways and many solutions have been proposed in the literature. **The general approach for pruning neural networks is to score parameters of a DNN and prune the ones with the lowest scores**. The magnitude of the parameters is typically the most used score.
- Pruning can be divided into **structured** and **unstructured** pruning: the goal of the former is to remove single parameters from the network while the latter tries to prune entire neurons/filters/channels to simplify the hardware implementation of the DNN.
- Some works leverage on the possibility of predicting the interconnections to be pruned before training while in others pruning is performed after training using **one-shot techniques**.
- Finally, in the attempt of further reducing the size of DNNs, **iterative approaches** are proposed where the network undergoes several cycles of training and pruning.

Novel contributions

- **The problem of pruning** has been addressed **from the point of view of the** actual **DNN architecture**.
- An alternative to the Multiply and Accumulate (MAC) map-reduce paradigm is introduced: the **Multiply and Max&Min** (**MAM²**, MAM-squared) map-reduce paradigm.
- **This paradigm is found to be naturally prone to pruning** as each neuron typically learns to select always the same few interconnections when performing the maximum and minimum reduction operations. At the same time, the pruning techniques already available in the literature that are being applied to MAC-based layers can be applied to MAM²-based layers as well with little to no changes in the algorithms obtaining better results.
- The computational complexity of Max/Min is similar to the accumulate operations.



Submitted and published works

- Boretti, C., Bich, P., Zhang, Y., and Baillieul, J., "*Visual Navigation Using Sparse Optical Flow and Time-to-Transit*", 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, 2022, pp. 9397-9403
- Bich, P., Prono, L., Mangia, M., Pareschi, P., Rovatti, R., and Setti, G., "Aggressively prunable MAM²-based Deep Neural Oracle for ECG acquisition by Compressed Sensing", accepted to 2022 IEEE Biomedical Circuits and Systems Conference (BIOCAS), Taipei, 2022

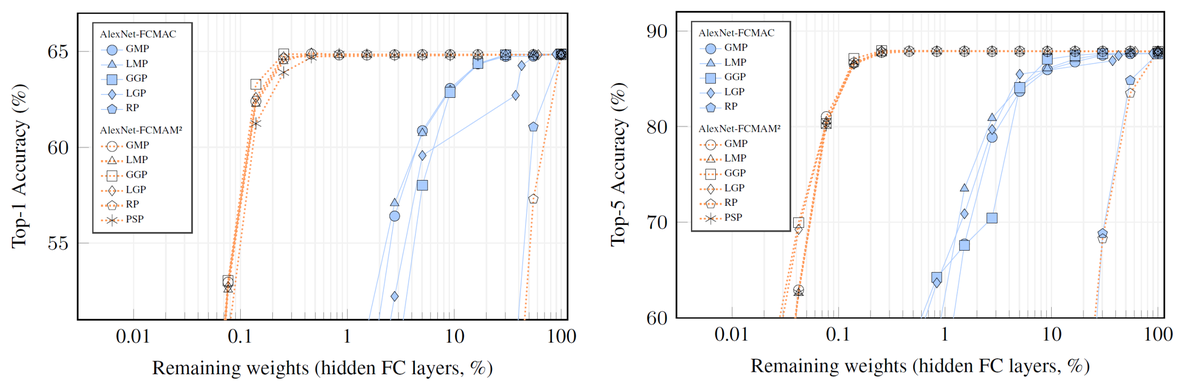
Adopted methodologies

- Several algorithms for performing unstructured pruning are used to prune both MAC-based and MAM²-based layers:

➢ ***Global Magnitude Pruning*** (GMP): the magnitude of the weights is the score and scores are compared globally.
➢ ***Layerwise Magnitude Pruning*** (LMP): as GMP but scores are compared layer-wise and not globally over the whole DNN.
➢ ***Global Gradient Magnitude Pruning*** (GGMP): the score is the absolute value of the product between the parameter value and its mean gradient.
➢ ***Layerwise Gradient Magnitude Pruning*** (LGMP): layer-wise version of GGMP.
➢ ***Random Pruning*** (RP): weights are pruned randomly.
➢ ***Probability of Selection Pruning*** (PSP): a method that can be applied to MAM²-based layers only as it leverages on its specific properties.

Results

- To analyze the behavior of MAM²-based layers, we test them as substitutes of some layers in classical computer vision DNNs, that are structures able to solve image classification tasks.
- We will refer to AlexNet based only on MAC-based layers as AlexNet-FCMAC, while we will call AlexNet-FCMAM² the version of AlexNet where the two hidden fully connected layers containing 91% of the total number of weights are substituted with two MAM²-based layers containing the same number of parameters of the original layers.
- **It can be shown that 99.97% of the weights of the MAM²-based layers can be pruned without any accuracy loss** with respect to the accuracy of AlexNet-FCMAC while only the 80% of weights contained in MAC-based layers can be removed.



Future work

- Implementation on hardware (FPGA, MCU) of DNNs using MAM²-based layers and comparison with standard neural networks relying only on MAC-based layers (memory footprint, inference time, etc.) .
- Theoretical demonstration that DNNs based on MAM² layers are universal approximators.
- Extension of the MAM² map-reduce paradigm to convolutional layers.

List of attended classes

- 01UJBRV – Adversarial training of neural networks (06/06/2022, 3)
- 01QTEIU – Data mining concepts and algorithms (03/02/2022, 4)
- 01SHMRV – Entrepreneurial Finance (28/02/2022, 1)
- 01UJUIU – Human-Ai Interaction (09/02/2022, 4)
- 01DNMIU – Optimized execution of neural networks at the edge (02/08/2022, 5)
- 01UNYRV – Personal branding (23/12/2021, 1)
- 01RISRV – Public speaking (20/12/2021, 1)
- 01SYBRV – Research integrity (06/12/2021, 1)
- 01SWQRV – Responsible research and innovation, the impact on social challenges (24/12/2021, 1)
- 02QUBRS – Statistical data processing (04/02/2022, 4)
- 02RHORV – The new Internet Society: entering the black-box of digital innovations (28/02/2022, 1)
- 01UNXRV – Thinking out of the box (16/11/2021, 1)
- 01SWPRV– Time management (23/12/2021, 1)
- 01QORRV – Writing Scientific Papers in English (24/03/2022, 3)

Politecnico di Torino
1859

PhD program in
Electrical, Electronics and Communications Engineering