Study and design of MRAM-based in-memory computing architectures for machine learning applications

Bruno Lovison-Franco, Aymen Romdhane, David Novo, Pascal Benoit, Guillaume Prenat, Lorena Anghel

Action concertée transverse

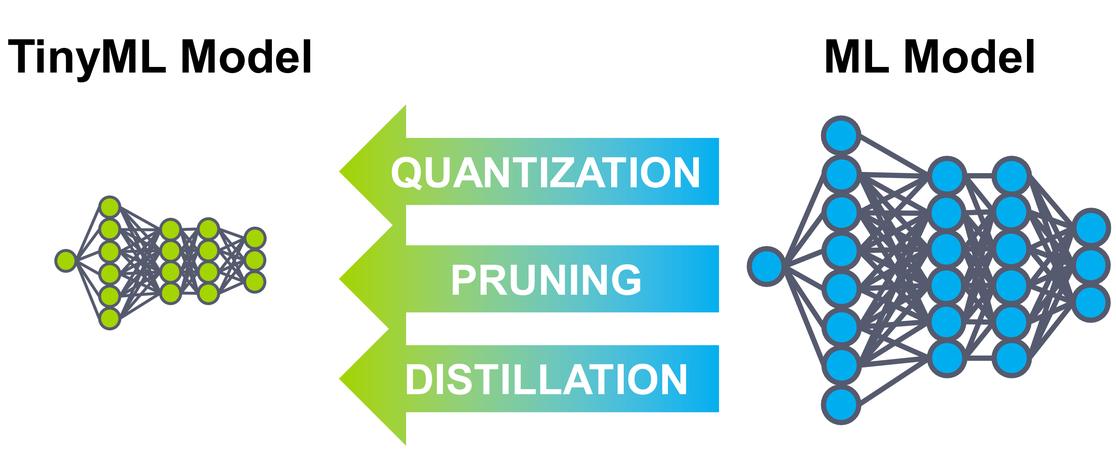
EMCOM

PROGRAMME DE RECHERCHE ÉLECTRONIQUE

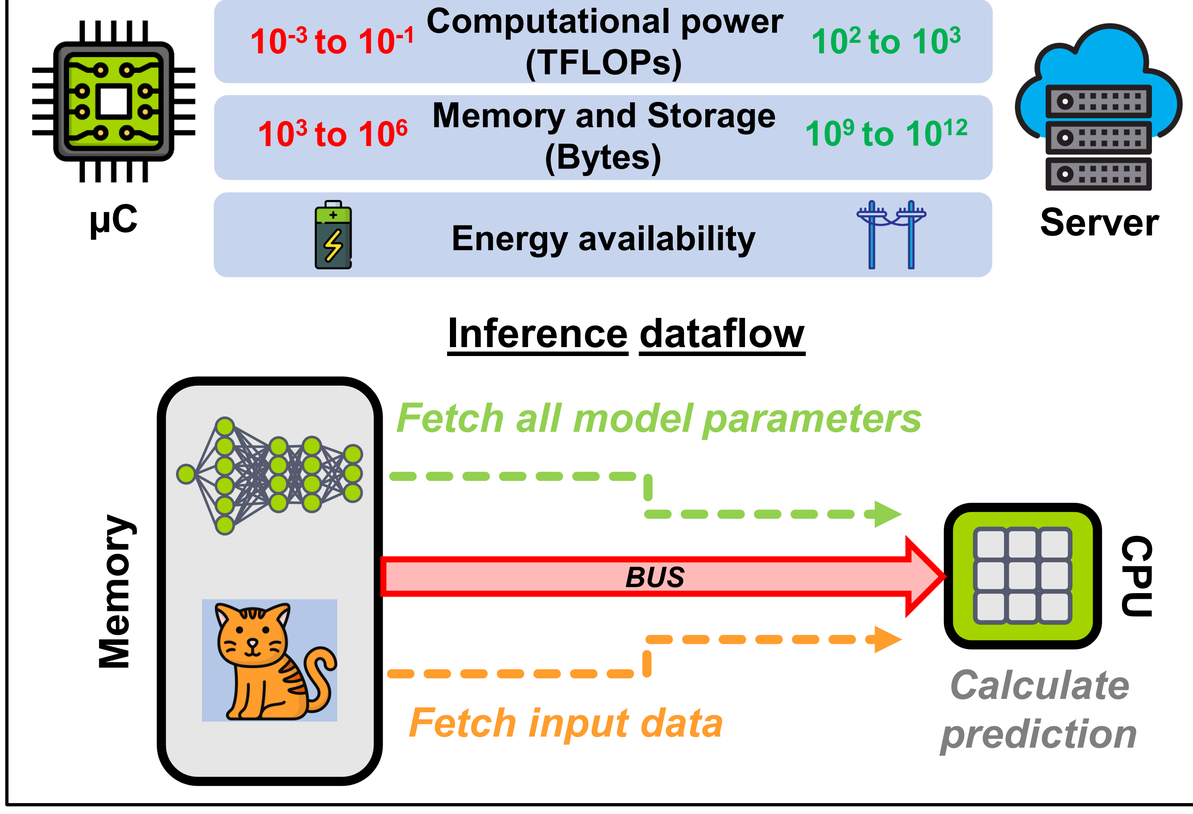
Study and design of MRAM-based in-memory computing architectures for machine learning applications

Bruno Lovison Franco[1], Aymen Romdhane[2], David Novo[1], Pascal Benoit[1], Guillaume Prenat[2], Lorena Anghel[2]

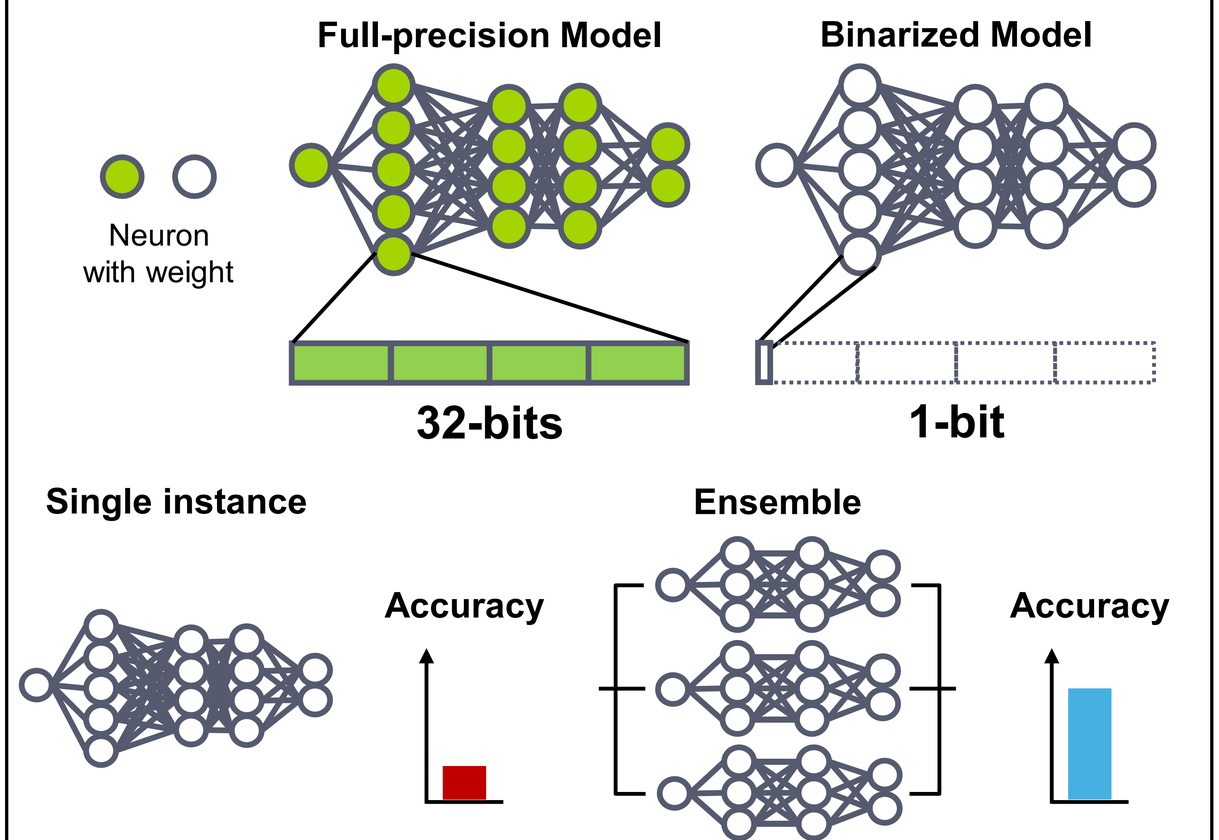
① from regular ML models towards TinyML



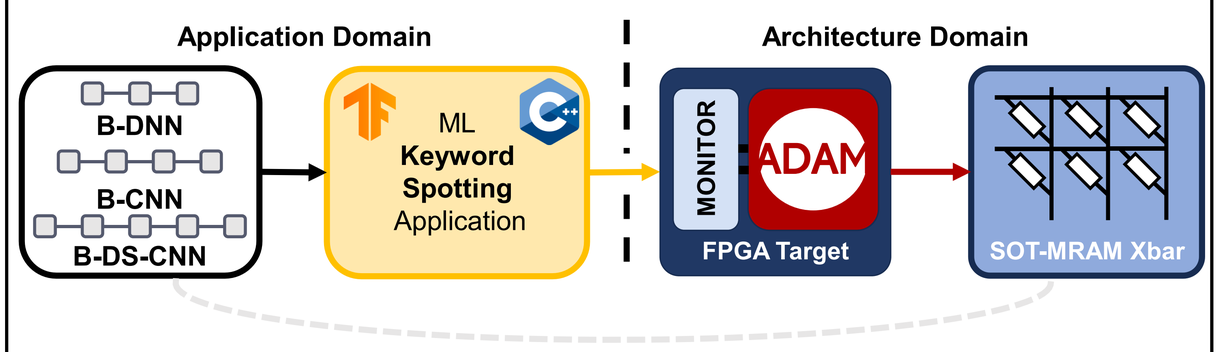
② Model Size, Data Movement and Operation Latency



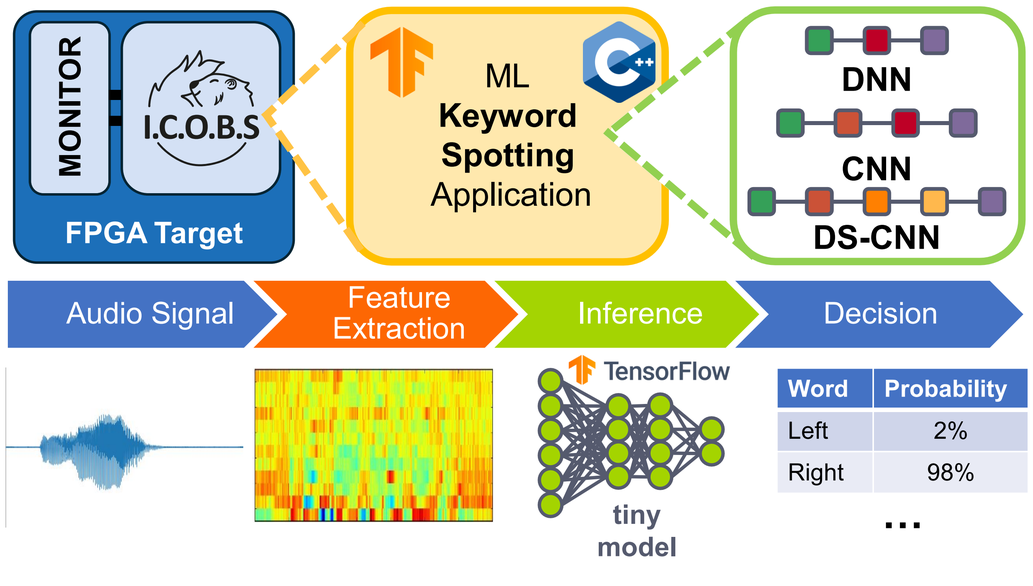
③ Binary Neural Networks & Ensembles

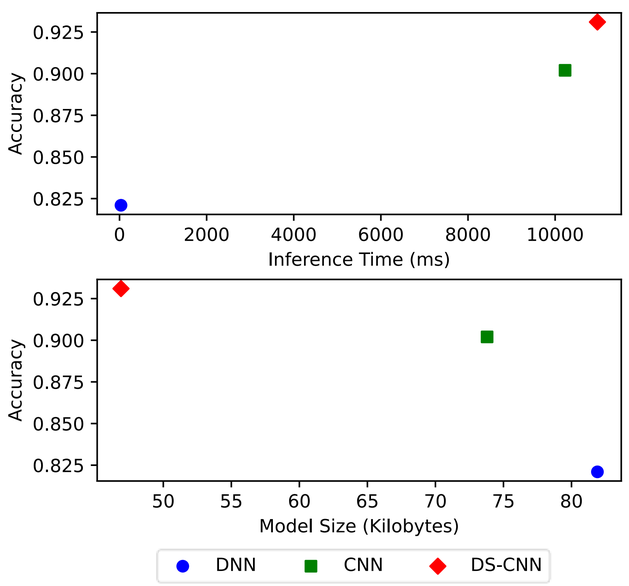


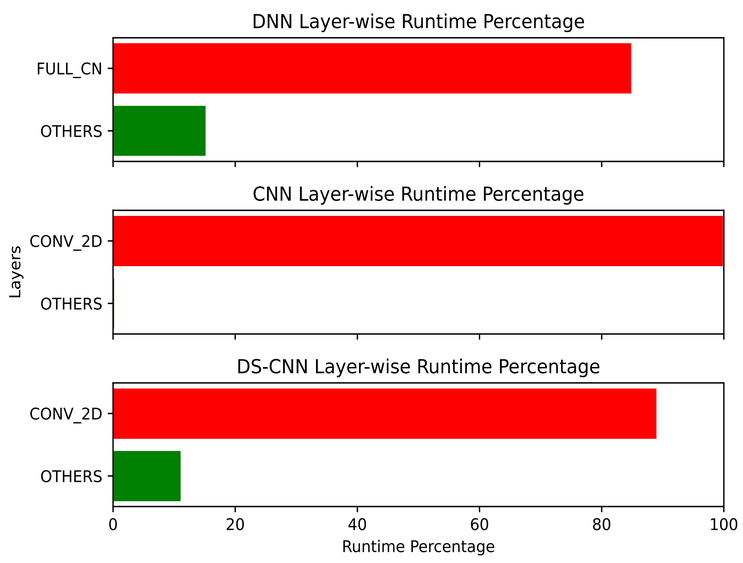
④ Network and Architecture Co-optimization



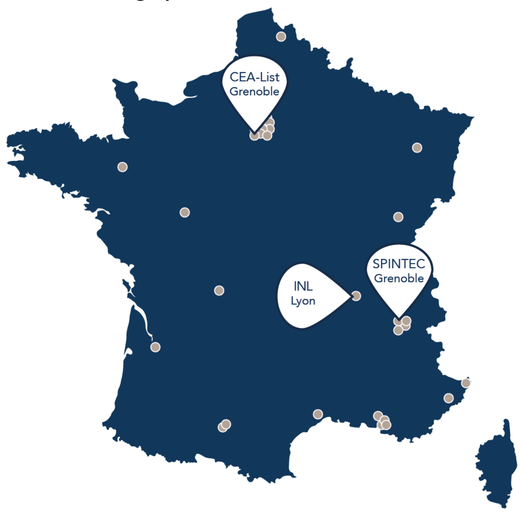
⑤ Keyword Spotting Application and Bottlenecks







1. Tchendjou, Ghislain Takam, et al. "Spintronic Memristor based Binarized Ensemble Convolutional Neural Network Architectures." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (2022).
2. Zhu, Shilin, Xin Dong, and Hao Su. "Binary ensemble neural network: More bits per network or more networks per bit?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
3. Mutlu, Onur, et al. "A modern primer on processing in memory." Emerging Computing: From Devices to Systems: Looking Beyond Moore and Von Neumann. Singapore: Springer Nature Singapore, 2022. 171-243.



(1) Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Université de Montpellier, CNRS -161, rue Ada, 34095 Montpellier cedex 5 - {firstname}.{lastname}@lirmm.fr
(2) Univ. Grenoble Alpes, CEA, CNRS, Grenoble INP, SPINTEC, 38000 Grenoble, France

ANR-22-PEEL-00013