Final project topic:

A comparison between different hardware architecture for tinyML(Pruning)
— Paper Survey

Motivation:

TinyML is efficient due to hardware and computing resource constraints. In tinyML we focus on pruning techniques. We want to improve across the board, both at the algorithmic level and at the hardware level.

Introduction:

Due to the limitations of hardware and computing resources, we pursue models with smaller memory usage and faster computation with little performance loss. The idea of Pruning was proposed in 1990.

Expected results (paper titles):

We want to improve across the board, both at the algorithmic level and at the hardware level.

[1] Molchanov, Pavlo, et al. "Importance estimation for neural network pruning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[2] Jang, Jeonggyu, et. al. "A SIMD-aware pruning technique for convolutional neural networks with multi-sparsity levels: work-inprogress." Proceedings of the International Conference on Hardware/Software Codesign and System Synthesis Companion. 2019.

[3] Lee, Kwangbae, et al. "Flexible group-level pruning of deep neural networks for on-device machine learning." 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2020.

[4] Liberis, Edgar, and Nicholas D. Lane. "Differentiable Network Pruning for Microcontrollers." arXiv preprint arXiv:2110.08350