Energy-efficient adaptive keyword spotting using slimmable convolutional neural networks

By Octavian Machidon and Veljko Pejović

University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

Overview

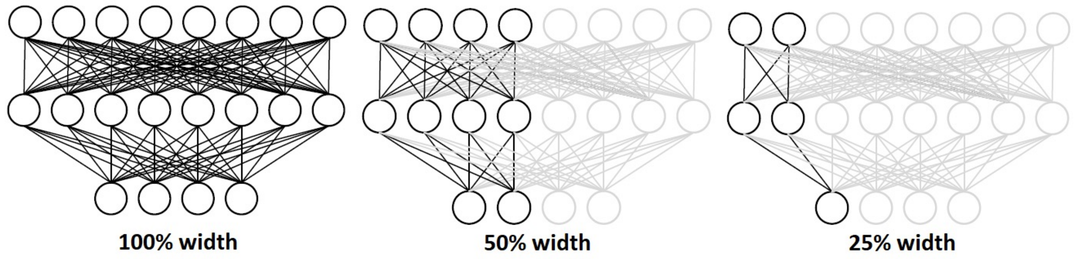
- We trained a **MobileNet-V1 SNN** to perform **KWS** and showed that the inference difficulty of each input is directly impacted by the noise level (SNR)
- For training the model and performing the evaluation we use a subset of the **Google Speech Commands v0.02** dataset
- We evaluated our implementation on the **UDOO Neo Full** single-board computer, an IoT-ready embedded computing platform
- The system-wide energy measurements were conducted using a **Monsoon Power Monitor**
- Classifying "easy-to-classify" samples (higher SNR) using a quarter-width network **saves up to 60% energy**
- Classifying "not-so-difficult" samples (lower SNR) using a three-quarter-width networks **saves up to 40% energy**

Acknowledgement

The research presented in this paper was funded by the project "Bringing Resource Efficiency to Smartphones with Approximate Computing" (project number: N2-0136) and "Context-Aware On-Device Approximate Computing", both funded by the Slovenian National Research Agency (ARRS).

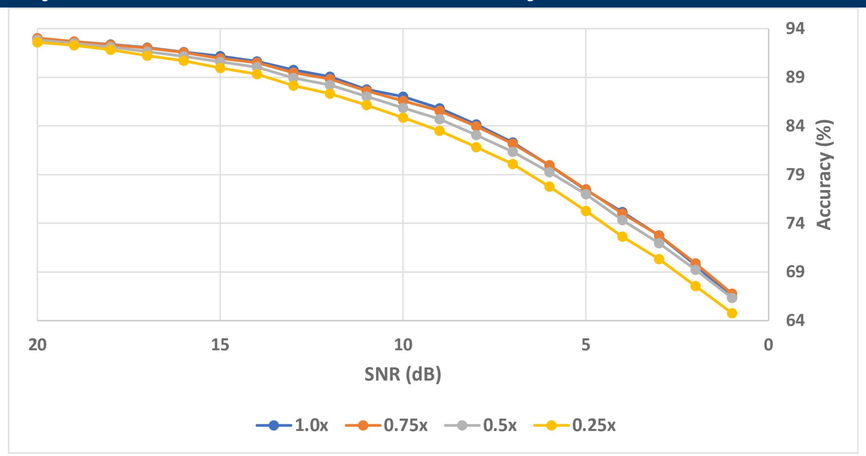
Slimmable Neural Networks for KWS

- The Slimmable Neural Network (SNN) concept enables reducing the number of active network parameters on the fly, during runtime, without the need for re-training;
- The SNN approach enables reducing the number of active network parameters on the fly to a fraction of the network's width selected from a pre-defined subset;



- We train a SNN MobileNet-V1 network (widths 100%, 75%, 50% and 25%) on a sub-set of the Google Speech Commands v0.02 dataset, containing 6 keywords: go, stop, up, down, right, left;
- For the test set, we add additive white gaussian (AWG) noise to each test input, for a SNR ranging from 1 to 30dB and compute the average accuracy of the network in each case, across all 4 widths.

SNR impact on classification accuracy for each network width

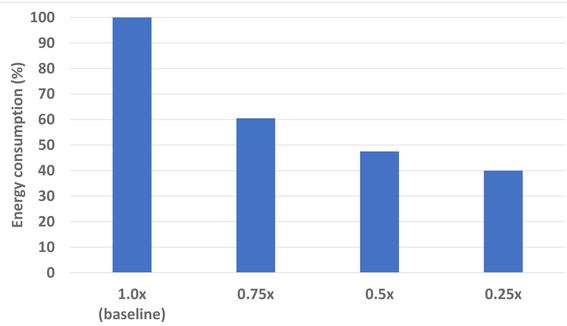


- No AWG noise added: all network widths scored over 96% average accuracy
- Adding AWG: for SNR above 20dB, the 25% width networks scores similarly to the 100% one
- For a SNR < 8 dB, the 75% width network scores on par with the 100% one

Evaluation on the UDOO Neo Full

NXP™ i.MX 6SoloX applications processor with an embedded ARM Cortex-A9 core and a Cortex-M4 Core

- Implemented our solution on the UDOO Neo Full board running Ubuntu 18.04
- Used the Monsoon Power Monitor to measure the system-wide energy consumption across the different network widths



Next steps

- Implementing an Intelligent, real-time adaptation framework of the KWS SNN based on the SNR
- Enabling real-time adaptation to the context: always use the lowest network width that is necessary to achieve a correct classification