Benchmarking and improving NN execution on DSP vs. custom accelerator for hearing instruments

Zuzana Jelčicová, Demant

Hearing instruments are supported by multicore processor platforms, that include several DSPs. These resources can be used to implement NNs, however, execution time and energy consumption are prohibitive to do so. This work benchmarks NN workloads relevant for hearing aids on Oticon's DSP-based platform. Next, a custom NN processing engine (NNE) is developed to achieve further power optimizations, exploiting following mutually dependent techniques:

- *Reduced wordlength* – default 24 bits are unnecessary for NN inference. We reduced the parameter wordlength to 8 bits with insignificant loss in accuracy.
- *Several MACs in parallel* – reduced wordlength enables to process more data at once. MAC design in the current DSP can process only 1 neuron at a time. Our NNE processes 12 neurons in parallel.
- *Two-step scaling* – reduced wordlength causes overflows to happen more frequently. The two-step scaling technique eliminates the need to reload and scale already computed outputs to maintain the ratio. It is done i) *within a vector* of 12 neurons (when writing the results) ii) *across a layer* (when reading the results in a subsequent layer). This technique makes our NNE always execute in a deterministic number of cycles.

A fully connected feedforward NN (250x144x144x144x12 [1]) was used as a benchmark model to run a keyword spotting application using speech command dataset [2].

The baseline 24-bit implementation in the DSP needed at least 39,744 96-bit memory accesses (inputs, weights) and 888 24-bit accesses (biases, results), corresponding to 480kB transferred. This number might grow depending on how many times the already computed results need to be fetched and scaled additionally as a result of an overflow. Our NNE requires only 7,250 96-bit memory accesses in total, summing up to 87kB transferred. Furthermore, the 24-bit NN parameters (inputs, weights, biases) occupy 240kB of memory, while the 8-bit NNE implementation only 80kB. The estimated energy costs (memory accesses and MACs) using a 45 nm process are 709nJ and 129nJ for 24-bit DSP and 8-bit NNE, respectively [3]. The accuracy using the pretrained DNN was 81.24% and 80.28% for the DSP and NNE, respectively.

[1] https://github.com/ARM-software/ML-KWS-for-MCU

[2] https://ai.googleblog.com/2017/08/launching-speech-commands-dataset.html

[3] Horowitz, M. 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (Feb 2014), pp. 10–14