Introduction

DSPs in general
• Datapaths with 16 and 24-bit elements
Neural network (NN) inference
• 8 bits are usually sufficient
  • More frequent overflows → additional memory accesses, increased energy costs, and unpredictable timing behavior.
xDSP vs NN processing engine (NNE)
xDSP → Oticon’s DSP-based platform used for obtaining baseline results
NNE achieves further power optimizations by exploiting three mutually dependent techniques:
1. Reduced wordlength – 24 to 8 bits with insignificant loss in accuracy
2. Several MACs in parallel – reduced wordlength → processing of more data at once
3. Two-step scaling
  • eliminates the need to reload and scale already computed outputs to maintain the ratio
  • makes our NNE always execute in a deterministic number of cycles

Keyword Spotting Application (KWS)
1. Feedforward fully connected DNN model (250x144x144x144x12) [1]
   • trained on 12 words (first two categories represent silence and unknown words): “yes”, “no”, “up”, “down”, “left”, “right”, “on”, “off”, “stop”, “go”
2. RelU and softmax (output layer) activation functions
3. Dataset: 65,000 one-second utterances of 30 words [2]

Digital Signal Processor (xDSP)
1. In input to DNN is the flattened feature matrix
2. Input signal (length L) is framed into overlapping frames (length f=40ms) with a stride (s=40ms), giving 25 frames as input/inference
   \[ T = \frac{L}{s} + 1 \]
3. 10 frequency bins → 250 inputs

Neural Network Engine (NNE)

NNE includes three optimizations:
1. Reduced wordlength
   • 24 to 8-bit parameters (inputs, weights, biases)
   • On-the-fly symmetric quantization (each layer individually)

   ![Insignificant loss in accuracy](image)

3.1 Within a vector (when writing results)

2. Several MACs in parallel
   • Input & output stationary processing

3. Two-step scaling
   • Solves inefficient individual scaling of activations

Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of shifts/group</th>
<th>Additional shifts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Blue</td>
<td>1</td>
<td>3 - 2 - 1</td>
</tr>
<tr>
<td>Orange</td>
<td>3</td>
<td>3 - 1 - 2</td>
</tr>
</tbody>
</table>

Table 1: Number of read operations (inputs and weights) needed for inference using 8-bit parameters.

Digital Processing

• Generic audio DSP → baseline for comparisons
• 24 bits for representing data (Q5.19)
• SIMD4 (96-bit memory interface)

References