Real-time Automatic Modulation Classification using RFSoC

Stephen Tridgell (Waymo*) - stridge@waymo.com
David Boland (University of Sydney) – david.boland@sydney.edu.au
Philip H.W. Leong (University of Sydney) - philip.leong@sydney.edu.au
Ryan Kastner (UCSD) - kastner@ucsd.edu
Alireza Khodamoradi (UCSD) - alirezak@eng.ucsd.edu
Siddhartha (University of Sydney) - siddhartha.siddhartha@sydney.edu.au

*work done while at University of Sydney
Automatic Modulation classifier: 488K class/s, 8us latency for CNN
Hard to make fully parallel implementations of a NN on contemporary FPGA due to size

Fit entire DNN on FPGA by exploiting unstructured sparsity and the following techniques:

1. Buffering of streaming inputs in a pipelined manner
2. Ternary weights implemented as pruned adder trees
3. Common subexpression merging
4. 16-bit bit serial arithmetic to minimize accuracy loss with low area
5. Sparsity control

Apply this to AMC
Common Subexpression Elimination

- Weights are ternary
  - Reduces convolution to constructing adder tree
  - Subexpression merged to reduce implementation

\[
\begin{align*}
  a \times (-1) &\quad b \times 0 &\quad c \times 1 \\
  d \times 0 &\quad e \times 1 &\quad f \times 1 \\
  g \times 0 &\quad h \times (-1) &\quad i \times 0 \\
\end{align*}
\]

Computing \(z_0 = c + e + f - (a + h)\) and \(z_1 = c + d - e - f\)
Activations are 16-bit
- Not all layers have same throughput
- Use digit serial to make more compact
- 4-bit digit serial has 1/4 area
- 1-bit bit serial has 1/16 area
O’Shea et. al. Deepsig dataset

![Graph showing accuracy vs. signal-to-noise ratio for different models and architectures]
Use incremental precision activations instead of 16 bit
- Use bit serial adders everywhere
- Adjust precision to match the throughput
- Same area as binary activations
- Almost 5% accuracy gain over binary activations

<table>
<thead>
<tr>
<th>Model</th>
<th>TW-64</th>
<th>TW-96</th>
<th>TW-BA-128</th>
<th>TW-INCRA-128</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLBs</td>
<td>28k (53.5%)</td>
<td>47k (89.3%)</td>
<td>43k (80.7%)</td>
<td>42k (80.2%)</td>
</tr>
<tr>
<td>LUTs</td>
<td>124k (29.1%)</td>
<td>232k (54.7%)</td>
<td>234k (55.1%)</td>
<td>211k (49.6%)</td>
</tr>
<tr>
<td>FFs</td>
<td>217k (25.5%)</td>
<td>369k (43.4%)</td>
<td>333k (39.2%)</td>
<td>324k (38.1%)</td>
</tr>
<tr>
<td>BRAMs</td>
<td>524 (48.5%)</td>
<td>524 (48.5%)</td>
<td>523 (48.4%)</td>
<td>512.2 (48.3%)</td>
</tr>
<tr>
<td>DSPs</td>
<td>1496 (35%)</td>
<td>1207 (28.3%)</td>
<td>1408 (33.0%)</td>
<td>1407 (32.9%)</td>
</tr>
<tr>
<td>Accr</td>
<td>78.7</td>
<td>81.1</td>
<td>75.9</td>
<td>80.2</td>
</tr>
</tbody>
</table>
Implementation

- System implemented on ZCU111 RFSoC
  - [https://github.com/da-steve101/radio_modulation](https://github.com/da-steve101/radio_modulation)
- Open Source Verilog generator
  - [https://github.com/da-steve101/twn_generator](https://github.com/da-steve101/twn_generator)
Conclusion

- Presented an optimized network for AMC which
  - Applies common subexpression elimination and digit serial arithmetic to a fully unrolled ternary network
  - Integrates the entire design on a single chip for a low-latency batch size 1 implementation
- These serve to achieve a level of performance higher than previously reported
- Challenge of achieving state of the art accuracy remains