Deep Neural Network IP Core for FPGA

DeepIP is a deep learning IP for Xilinx FPGAs that allows you to focus on training your AI model rather than writing FPGA code. DeepIP is a fully customizable IP core that accepts trained machine learning models from most commercial machine learning tools and enables importing into your Vivado FPGA design. DeepIP reduces development time, and complexity to create your own embedded deep learning applications.

Off the Shelf IP

- Low power (~1Watt)
- Low latency (microseconds)
- Flexible neural network architecture
- Focus on training not coding
- Iterate quickly through training models
- Enables rapid development
- Evaluation licenses
- 16-bit fixed point format

Train - Import To Vivado - Deploy to FPGA

Easy to Use - Easy to Configure - Easy to Deploy

www.i-a-i.com
contact: contact@i-a-i.com

Copyright © Intelligent Automation Inc. (IAI) All rights reserved. IAI retains ownership of all other rights to the material expressed in this document. Any reproduction of the content on this document without prior written permission from IAI is prohibited.
Use Cases

- Computer vision
- Natural language processing
- Wireless communications
- Embedded security

Resource Utilization

Image classification on Xilinx Zync UltraScale+

<table>
<thead>
<tr>
<th>Resource</th>
<th>Utilization</th>
<th>Available</th>
<th>Utilization %</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUT</td>
<td>60281</td>
<td>274080</td>
<td>21.99</td>
</tr>
<tr>
<td>LUTRAM</td>
<td>41820</td>
<td>144000</td>
<td>29.04</td>
</tr>
<tr>
<td>FF</td>
<td>12505</td>
<td>548160</td>
<td>2.28</td>
</tr>
<tr>
<td>DSP</td>
<td>266</td>
<td>2520</td>
<td>10.56</td>
</tr>
<tr>
<td>IO</td>
<td>10</td>
<td>328</td>
<td>3.05</td>
</tr>
<tr>
<td>BUFG</td>
<td>5</td>
<td>404</td>
<td>1.24</td>
</tr>
</tbody>
</table>

Performance Comparison

<table>
<thead>
<tr>
<th>Measures</th>
<th>Device</th>
<th>VC707 FPGA</th>
<th>Xavier GPU</th>
<th>ARM</th>
<th>Measures of effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>94%</td>
<td>95%</td>
<td>95%</td>
<td></td>
<td>Classification accuracy</td>
</tr>
<tr>
<td>Latency</td>
<td>29.62 µs</td>
<td>0.22 ms</td>
<td>0.301 ms</td>
<td></td>
<td>Time requires to process/classify single input</td>
</tr>
<tr>
<td>Energy</td>
<td>30 µJ</td>
<td>3.2 mJ</td>
<td>0.36 mJ</td>
<td></td>
<td>Energy consumption to process single input</td>
</tr>
</tbody>
</table>

What's Included?

- Conversion script
- Flexible neural network architecture
- User guide
- Demo application
- Trial license

Additional Services

- Deploy software
- Hardware and software design services
- R&D services