

ABSTRAK

Komputasi inferensi Convolutional Neural Network (CNN) membutuhkan bandwidth memori yang tinggi dan kemampuan komputasi yang besar. Pada implementasi akselerator perangkat keras, performa sistem sering dibatasi oleh tingginya intensitas akses ke memori eksternal selama proses konvolusi. Oleh karena itu, pada penelitian ini dirancang sistem antarmuka komunikasi dan subsistem buffering untuk mendukung akselerator CNN berbasis Diagonal Cyclic Array (DCA) menggunakan metode Hardware/Software Co-design pada platform FPGA untuk menjalankan inferensi citra angka berukuran 28×28 piksel. Komunikasi antara Processing System (PS) dan Programmable Logic (PL) menggunakan protokol AXI-Stream untuk transfer data citra dan AXI-Lite untuk konfigurasi parameter jaringan. Subsistem buffering terdiri atas data-recycle FIFO untuk menahan dan mendistribusikan data piksel serta weight buffer untuk mendistribusikan bobot sesuai kebutuhan aliran data pada arsitektur DCA. Hasil pengujian menunjukkan bahwa AXI-Stream mampu mentransmisikan 784 data piksel dengan laju transfer sebesar 100 MB/s dan throughput sebesar 1 byte/clock. Hasil implementasi pada FPGA menunjukkan tingkat kesesuaian hasil inferensi terhadap model Python dengan akurasi klasifikasi CNN sebesar 91,5%.

Kata Kunci: *Akselerator, Convolutional Neural Network, FPGA, AXI-Stream, FIFO*