XIX Jornadas de Ciencias, Tecnologías e Innovación
UNR

IMPLEMENTACIÓN DE REDES NEURONALES MULTICAPA EN HARDWARE RECONFIGURABLE CON FLUJO NUBE-BORDE

VECCHIO, Juan[1]; MARTINEZ, Roberto[1]; CORTI, Rosa[1] ; LARA, Luis[1]

[1]Departamento de Sistemas e Informática – Facultad de Ciencias Exactas, Ingeniería y Agrimensura – U.N.R
Centro de Estudios Interdisciplinarios (CEI) – U.N.R.
jvecchio@fceia.unr.edu.ar

Palabras claves: System on Chip – FPGA – Redes neuronales artificiales

INTRODUCCIÓN

El creciente interés por el procesamiento en el borde (*edge processing*) ha impulsado la búsqueda de soluciones capaces de combinar baja latencia, eficiencia energética y flexibilidad en la implementación de modelos de inteligencia artificial. En este contexto, las redes neuronales artificiales (RNAs) representan una herramienta clave, pero su despliegue en dispositivos embebidos requiere optimizaciones tanto a nivel de diseño como de hardware. Este trabajo aborda esta problemática mediante el desarrollo de un núcleo genérico en VHDL, diseñado para ejecutar redes neuronales del tipo perceptrón multicapa en plataformas SoC (*System-on-Chip*). La propuesta integra el entrenamiento del modelo en la nube, utilizando PyTorch, con la implementación del modelo cuantizado en hardware reconfigurable, constituyendo así un flujo híbrido nube-borde que aprovecha las fortalezas de ambas tecnologías.

OBJETIVOS

El objetivo principal de este trabajo fue el diseño e implementación de un núcleo modular, adaptable y optimizado, capaz de ejecutar redes neuronales multicapa en dispositivos SoC. Objetivos específicos:
-Desarrollar un núcleo reconfigurable en cuanto a número de entradas, cantidad de neuronas por capa, funciones de activación y precisión de los datos.
-Posibilitar la reutilización del diseño en distintas aplicaciones de inteligencia artificial embebida.
-Optimizar el desempeño en términos de uso de recursos, latencia y escalabilidad, de manera que el diseño resulte portable y eficiente.

MATERIALES Y MÉTODOS

La metodología se estructuró en dos etapas fundamentales:
Etapa en la nube: se diseñó y entrenó el modelo en PyTorch, aplicando posteriormente una cuantización lineal de parámetros que permitió su representación en punto fijo, requisito indispensable para su implementación eficiente en hardware.
Etapa en el borde (SoC): el modelo cuantizado fue traducido a descripciones en VHDL, utilizando parámetros genéricos que facilitaron la adaptabilidad del diseño. Además, el núcleo fue empaquetado como un periférico esclavo AXI (*Advanced eXtensible Interface*), lo que posibilitó su integración con los procesadores escalares Cortex-A9 de la plataforma ZYNQ-7020.
La arquitectura propuesta adoptó un diseño jerárquico en dos niveles, lo que permitió implementar diversas topologías de redes neuronales sin modificar la estructura base del núcleo, garantizando flexibilidad y escalabilidad.

RESULTADOS y DISCUSIÓN

Como caso de estudio, se implementó una red neuronal con una capa oculta de dos neuronas con función de activación ReLU y una capa de salida lineal, diseñada para resolver un problema de regresión no lineal múltiple. El modelo entrenado fue validado en hardware, obteniéndose un error relativo medio del 1,2 % respecto a la implementación en software. En términos de rendimiento, la solución desarrollada mostró mejoras significativas en el tiempo de inferencia en comparación con plataformas GPU Tesla T4. El análisis de síntesis evidenció un retardo total de 31,788 ns, acompañado de un balance adecuado entre lógica interna y enrutamiento. Estos resultados demuestran que el diseño propuesto constituye una alternativa eficiente para la implementación de redes neuronales en hardware reconfigurable.

Tabla de métricas representativas

Métrica	GPU (baseline)	SoC (propuesto)	PL (Lógica Reconfig.)
Tiempo de inferencia	10.5 ms	< 1 ms	31.788 ns
PL (Lógica Reconfig.)	**LUTs**	**FFs**	**Potencia estimada**
Uso de recursos	1,92%	1,36%	1,547 W
Diferencia Hardware/Software = 1,2%			

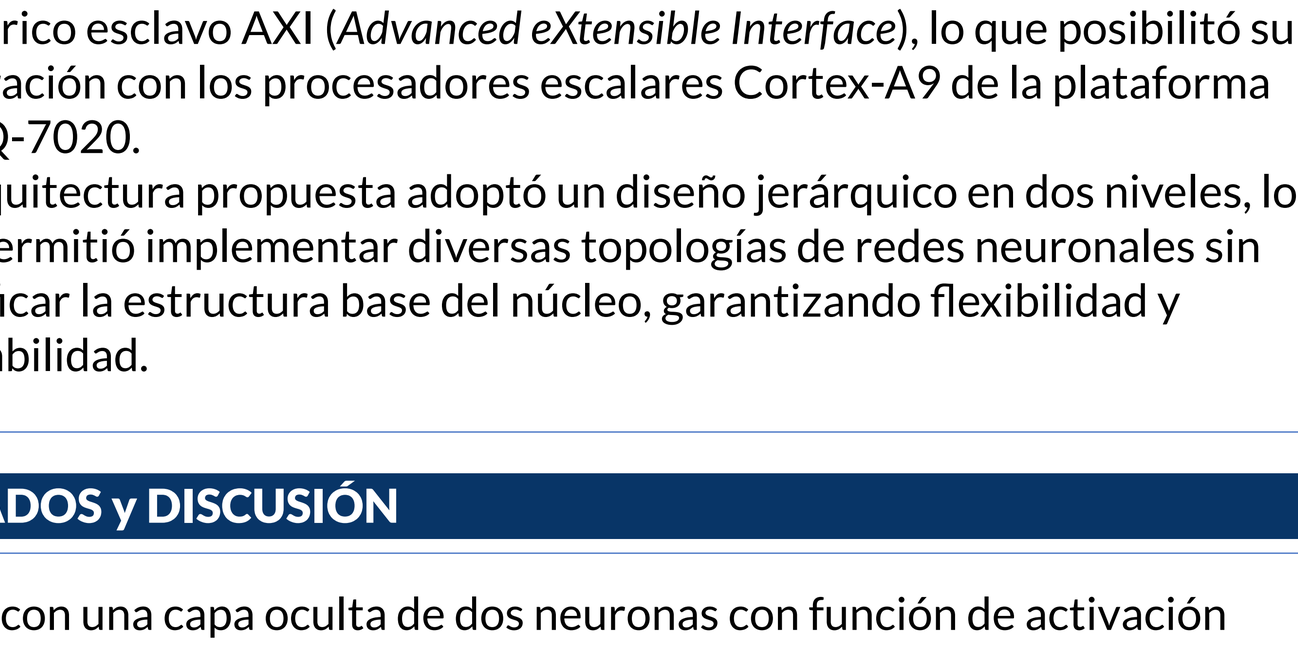

Diagrama en bloques de la aplicación y placa de desarrollo Zybo Z7

CONCLUSIONES

El desarrollo realizado confirma la factibilidad de un flujo de trabajo híbrido nube-borde para la implementación de redes neuronales artificiales en plataformas SoC. La solución propuesta combina las ventajas del entrenamiento en la nube con la ejecución optimizada en hardware reconfigurable, logrando un diseño eficiente, portable y escalable.
Los resultados obtenidos no solo validan el núcleo como una herramienta práctica para aplicaciones actuales, sino que también sientan las bases para el desarrollo de soluciones de inteligencia artificial embebida más complejas y de rápido despliegue, orientadas a escenarios donde la latencia, la eficiencia energética y la flexibilidad resultan determinantes.