

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Resumen

Escuela de doctorado

Doctor

Desarrollo Flexible Mediante Herramientas de Alto Nivel de Sistemas SoPC de Inteligencia Artificial en la Periferia

Por Jorge Deolindo FÉ

La presente tesis doctoral aborda el diseño, el desarrollo y la implementación de sistemas SoPC (*System-on-Programmable-Chip*) para el procesamiento eficiente de algoritmos de inteligencia artificial en la periferia, centrándose en la inferencia de redes neuronales profundas. El motivo del presente trabajo es la creciente necesidad de procesar datos localmente en dispositivos con recursos limitados, con el objetivo de reducir la latencia, el consumo energético y la dependencia de conexiones remotas a la nube, sin comprometer la capacidad de decisión autónoma en tiempo real.

La investigación propone una arquitectura híbrida *hardware/software* que permite acelerar funciones críticas de redes neuronales mediante *hardware* reconfigurable, mientras mantiene la flexibilidad y capacidad de adaptación que ofrece el procesamiento *software*. Para ello, se emplea una herramienta de síntesis de alto nivel, específicamente Vitis™ HLS, que permite convertir algoritmos escritos en lenguajes de alto nivel (C/C++) a módulos *hardware* optimizados (RTL), facilitando la implementación eficiente de operaciones como vector-matriz, funciones de activación y transferencia de datos mediante DMA.

La metodología adoptada sigue una serie de etapas claramente definidas: el entrenamiento de la red neuronal en Python, el diseño y la optimización de funciones con Vitis™ HLS, la integración en la arquitectura SoPC y, finalmente, la validación del sistema mediante casos de uso de redes neuronales profundas. Se evalúan distintas configuraciones del sistema, al tener en cuenta el ancho de bus de datos, la topología de la red neuronal y el tipo de funciones implementadas en *hardware*, en punto fijo o punto flotante, con el objetivo de caracterizar el impacto en el rendimiento, el uso de recursos lógicos del SoPC y el consumo energético.

Los resultados obtenidos muestran que, para determinadas configuraciones de red neuronal, la aceleración por *hardware* permite reducir significativamente los tiempos de procesamiento y mejorar la eficiencia energética. Asimismo, se demuestra que el sistema desarrollado es reconfigurable por *software* y adaptable a diferentes topologías de redes neuronales profundas, confirmando su aplicabilidad en entornos reales de inteligencia artificial en la periferia.

Finalmente, se discute el aporte original de este trabajo, que consiste en una plataforma flexible y eficiente basada en SoPC para el procesamiento de redes neuronales profundas, utilizando flujos de desarrollo ágiles con herramientas como Vitis™ HLS. Se proponen además líneas para futuras investigaciones, incluyendo la optimización dinámica entre *hardware* y *software* y la exploración de formatos numéricos de baja precisión.