Universidad de Córdoba

Escuela Politécnica Superior de Córdoba

Anteproyecto de Trabajo Final de Grado

Ingeniería Informática.

Mención en Computadores

Optimización en la toma de decisiones utilizando FPGAs

Autor: **Javier Carmona Tejero**

Correo: i72catej@uco.es

Directora: **Joaquín Olivares Bueno**

9 de marzo de 2023

OLIVARES BUENO JOAQUIN - 26487222P

Firmado digitalmente por OLIVARES BUENO JOAQUIN - 26487222P
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-26487222P, givenName=JOAQUIN, sn=OLIVARES BUENO, cn=OLIVARES BUENO JOAQUIN - 26487222P
Fecha: 2023.03.10 11:36:45 +01'00'

Índice

1. Introducción

Hoy en día se puede apreciar que el uso de tecnologías que se basan en toma de decisiones va en auge. Podemos encontrar herramientas que ayudan a tomar un camino concreto a una empresa, como otras que reconocen a una persona por la calle y hace que un vehículo frene en caso de peligro de colisión.

El fin de estas tecnologías es que una máquina tome la decisión más correcta teniendo en cuenta una serie de parámetros que le indica el desarrollador de dicha máquina. Esta tiene que ser capaz de dar una respuesta correcta frente a una situación de riesgo o no, de la cual no es consciente del mundo real o de lo que en realidad está pasando. Solo se dedica a captar información, procesar dicha información, generar una respuesta en base a los datos captados y dar una respuesta correcta al usuario.

Según del campo en el que este expuesta, el tiempo de respuesta será más o menos importante. Por eso utilizar FPGAs es una muy buena opción, ya que trabajan a frecuencias muy altas, tienen gran capacidad de procesamiento en paralelo y tienen bajos costes en comparación a otras tecnologías.

Las redes neuronales se utilizan para que su uso sea el más parecido al de un cerebro humano. Estas redes son entrenadas para dar el resultado que más se ajuste a unos ciertos parámetros. Existen varios modelos de estas redes neuronales que se ajustan a los objetivos que tengamos, si trabajan con memoria (Dynamic Neural Network), según si todas las neuronas de una capa se conectan a todas de la siguiente capa con el enfoque al reconocimiento de imágenes (Convolutional Neural Network), etc. Para nuestro proyecto usaremos el modelo que más se adapte a nuestros objetivos.

2. Objetivos

El objetivo principal del proyecto es la optimización de algoritmos de toma de decisiones basados en redes neuronales en FPGAs. Esta tecnología puede ser aplicada a varios campos de la vida, en este proyecto desarrollaremos un ejemplo de uso de esta tecnología.

Para conseguir el objetivo final, primero hay que cumplir otros objetivos como:

- Realizar un estudio sobre el funcionamiento de redes neuronales.
- Estudiar los distintos casos de uso y aplicaciones que tiene esta tecnología.
- Aprender a desarrollar e implementar algoritmos para la toma de decisiones en VHDL.
- Implementar el proyecto a una FPGA.
- Probar su correcto funcionamiento.

3. Antecedentes

Como antecedentes podemos fijarnos en el proyecto Catapult que anunció Microsoft. En este proyecto, se demostró con éxito la aceleración que produce el uso de una FPGA a una red neuronal convolucional, además de lograr un rendimiento excelente con un bajo consumo.

Este proyecto tiene como objetivo acelerar la computación de imágenes en la nube mediante una red neuronal convolucional (CNN) acelerada por FPGAs. Siendo capaz de recibir imágenes de entrada, procesar los pixeles de la imagen en el chip y recircular la salida calculada a entradas de otras capas, entrenando así la red neuronal.

En este proyecto se comprobó que el uso de FPGAs es capaz de aportar más de 4 veces su rendimiento y consumiendo cerca de 10 veces menos su energía.

	CIFAR-10 [4]	ImageNet 1K [1]	ImageNet 22K [2]	Max Device Power
Catapult Server + Stratix V D5 **[3]**	2318 images/s	134 images/sec	91 images/sec	25W
Catapult Server + Arria 10 GX1150 **[8]**	-	~233 images/sec (projected)	~158 images/sec (projected)	~25W (projected)
Best prior CNN on Virtex 7 485T **[5]**	-	46 images/sec[3]	-	-
Caffe+cuDNN on Tesla K20 **[6]**	-	376 images/sec	-	235W
Caffe+cuDNN on Tesla K40 **[6]**	-	500-824 images/sec[4]	-	235W

Table 1: Comparison of Image Classification Throughput and Power.

4. Fases de desarrollo del proyecto

Las fases que se llevarán a cabo para el desarrollo del proyecto serán las siguientes:

4.1. Estudio e Investigación

Se realizará una investigación sobre los conceptos teóricos de las redes neuronales y su funcionamiento. Además, se estudiará el entorno de desarrollo del proyecto y se elegirá el más adecuado y eficiente posible.

4.2. Análisis y Definición de Requisitos

Se marcarán los objetivos a conseguir del proyecto, tanto funcionales como no funcionales.

4.3. Diseño

Tras el estudio y análisis del proyecto se hará el diseño de la red neuronal que se usará.

4.4. Implementación

Implementación del diseño.

4.5. Pruebas

Se comprobará con varios test de funcionalidad los resultados obtenidos y se verá si se ajustan o no a lo especificado en los objetivos.

Se comprobará con un exhaustivo test de funcionalidad los resultados obtenidos y si se ajustan a lo especificado en los objetivos.

4.6. Documentación

Se incluirán el manual técnico, el manual de usuario y el código.

5. Recursos

5.1. Recursos Humanos

El proyecto será realizado por Javier Carmona Tejero, alumno del Grado de Ingeniería Informática en mención de Computadores, siguiendo las indicaciones y recomendaciones del director del proyecto Joaquín Olivares Bueno.

5.2. Recursos Hardware

El proyecto será llevado a cabo en el siguiente equipo de trabajo:

Ordenador:

- Memoria: 8 GB RAM
- Procesador: Intel Core i7-6500U 2.5GHz
- GPU: NVIDIA GeForce GTX 950M with 2GB VRAM

FPGA:

- Digilent Nexys 4 Artix-7
- Digilent Arty A7 Artix-7
- O FPGAs similares.

5.3. Recursos Software

- Sistema Operativo: Windows 10 o Ubuntu 20.4
- Entorno de trabajo: Vivado o similares
- Documentación: Word o similares
- Lenguaje de programación: principalmente VHDL con posibilidad de usar algún otro lenguaje complementario para alguna funcionalidad.

Referencias bibliográficas.

[1] Web Consultada el 2023-03-09
Como empezar con Deep Learning HDL Toolbox :
https://es.mathworks.com/help/deep-learning-hdl/get-started-with-deep-learning-hdl-toolbox.html

[2] Web Consultada el 2023-03-09
¿Es posible crear con VHDL una red neuronal profunda?:
https://www.mathworks.com/matlabcentral/answers/743692-is-it-possible-to-create-vhdl-code-of-the-image-classifier-block-from-the-deep-neural-networks-lib

[3] Web Consultada el 2023-03-09
Deep Learning:
https://es.wikipedia.org/wiki/Aprendizaje_profundo

[4] Web Consultada el 2023-03-09
Aplicaciones de las FPGA:
https://www.generatecnologias.es/aplicaciones_fpga.html

[5] Web Consultada el 2023-03-09
Convolutional Neural Network:
https://en.wikipedia.org/wiki/Convolutional_neural_network

[6] Web Consultada el 2023-03-09
Artificial Neural Network:
https://advancedtech.wordpress.com/2007/08/24/redes-neuronales/#:~:text=Las%20redes%20neuronales%20din%C3%A1micas%20se,conversi%C3%B3n%20de%20texto%20a%20voz.

[7] Web Consultada el 2023-03-09
¿Qué son las redes neuronales?:
https://www.xataka.com/robotica-e-ia/las-redes-neuronales-que-son-y-por-que-estan-volviendo

[8] Web Consultada el 2023-03-09
Project Catapult Microsoft:
https://www.microsoft.com/en-us/research/project/project-catapult/

[9] Web Consultada el 2023-03-09
Aceleración de CNN mediante FPGAs:
https://www.microsoft.com/en-us/research/publication/accelerating-deep-convolutional-neural-networks-using-specialized-hardware/