.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

ESTUDIO Y DISEÑO DE ARQUITECTURAS DIGITALES PARA LA IMPLEMENTACIÓN EFICIENTE DE REDES NEURONALES CONVOLUCIONALES

TESIS DOCTORAL

3 DE OCTUBRE DE 2025

Autor: Ing. Federico Giordano Zacchigna

Director: Dr. Ing. Ariel Lutenberg (FIUBA-CONICET)

Co-Director: Dr. Ing. Sergio Lew (FIUBA-CONICET)

AGENDA

AGENDA

HOJA DE RUTA

COMPLETAR

1. El Desafío: La IA en el mundo real.
2. Mi Aporte: Una solución en dos partes.
3. Aporte 1: Metodología de implementación sistemática.
4. Aporte 2: Cuantización flexible para máxima eficiencia.
5. Resultados y Conclusiones.

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

APRENDIENDO A "VER" COMO LOS HUMANOS

¿QUÉ ES UNA RED NEURONAL CONVOLUCIONAL (CNN)?

Concepto: Una CNN es un modelo inspirado en el cerebro que procesa imágenes en capas:

- Las primeras capas detectan cosas simples (bordes, colores)
- Las capas más profundas reconocen objetos complejos (caras, autos)

Visual: Usar una figura como `figures/general_cnn_simple.png` o `figures/lenet5.png` para ilustrar las capas.

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

EL BALANCE ENTRE COSTO Y CALIDAD

EL DESAFÍO TÉCNICO: COMPLEJIDAD Y PRECISIÓN

Concepto: La "magia" de las redes neuronales está en sus números (los "pesos"). Usar números de alta precisión (como `float32`) da buenos resultados, pero es lento y consume mucha memoria.

Pregunta Clave: ¿Podemos usar números de "baja precisión" para hacer la red más rápida y pequeña?

Fuente: `02-main_matter/2-specific.tex` *- Sección 2.2 y 2.3*

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

COMPRIMIENDO LA RED: LA IDEA DE LA CUANTIZACIÓN

LA SOLUCIÓN CLAVE: LA CUANTIZACIÓN

Concepto: La cuantización es el proceso de reducir la precisión de los números. Es como redondear. Esto reduce drásticamente el tamaño del modelo y la complejidad de las operaciones.

Visual: Usar la figura `figures/specific_quantization_function_uniform.png` para mostrar cómo un rango continuo se mapea a niveles discretos.

MIS APORTES ORIGINALES

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

MI PROPUESTA: UN PUENTE INTEGRAL DEL SOFTWARE AL HARDWARE

Concepto: El núcleo de esta tesis es la creación de un flujo de trabajo completo para llevar modelos de IA desde el software al hardware de forma eficiente. Mis dos aportes no son ideas aisladas, sino que forman las dos mitades de este puente:

1. **La Metodología Sistemática (Aporte 1):** Provee el **mapa** y la **estructura** para la construcción del hardware. Responde a la pregunta: *¿Cómo lo construimos de forma ordenada?*
2. **La Cuantización Flexible (Aporte 2):** Provee la **herramienta de optimización** para que esa estructura sea lo más eficiente posible. Responde a la pregunta: *¿Cómo lo hacemos pequeño y rápido sin perder calidad?*

Visual: Un diagrama de flujo muy simple que muestre:

`Modelo IA (Software)` → `Metodología (Aporte 1)` → `Cuantización (Aporte 2)` → `Implementación (Hardware)`

MIS APORTES ORIGINALES PARA LA METODOLOGÍA

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES PARA LA METODOLOGÍA

DEL CAOS AL ORDEN: UN PROCESO EN 9 FASES

APORTE 1: METODOLOGÍA DE IMPLEMENTACIÓN SISTEMÁTICA

Concepto: Se propone un flujo de trabajo claro que descompone el problema en 9 fases manejables: desde el diseño en alto nivel hasta la verificación final en hardware.

Beneficio: Esto hace que el proceso sea reproducible, escalable y más fácil de optimizar.

Fuente: `02-main_matter/3-architectures.tex` *- Sección 3.2.1*

MIS APORTES ORIGINALES PARA LA METODOLOGÍA

ORGANIZANDO EL HARDWARE: BLOQUES ESPECIALIZADOS

APORTE 1: ARQUITECTURA MODULAR (LOS 3 TIPOS DE BLOQUES)

Concepto: La arquitectura se organiza en 3 tipos de bloques: de **Procesamiento (PB)**, de **Flujo de Datos (DFB)** y de **Organización de Memoria (MOB)**. Cada uno tiene una función clara.

Visual: Un diagrama simple que muestre estos tres tipos de bloques interconectados.

Fuente: `02-main_matter/3-architectures.tex` *- Sección 3.5.7*

MIS APORTES ORIGINALES PARA LA METODOLOGÍA

DE LA TEORÍA A LA PRÁCTICA: EL HARDWARE FUNCIONANDO

RESULTADO DEL APORTE 1: UNA ARQUITECTURA REAL Y VERIFICADA

Concepto: La metodología no fue solo un ejercicio teórico. Se aplicó para construir una arquitectura de hardware completa y funcional en un dispositivo real (una FPGA).

Beneficio Clave: Se demostró que el método sistemático permite crear arquitecturas que son modulares, escalables y, lo más importante, verificables. Esto sienta las bases para todo el trabajo de optimización que viene después.

Visual: Se puede usar la figura `figures/architecture.png` o `figures/HW_structure.png` para mostrar un esquema del sistema implementado.

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

EL DILEMA DE LA CUANTIZACIÓN: ¿UNIFORME O NO UNIFORME?

APORTE 2: CUANTIZACIÓN FLEXIBLE (FQ)

Concepto:

- **Uniforme:** Fácil de implementar en hardware, pero imprecisa.
- **No Uniforme:** Muy precisa, pero incompatible con hardware eficiente.

Mi Solución: Cuantización Flexible, que combina lo mejor de ambos mundos.

Fuente: `02-main_matter/4-flexible_quantization.tex` *- Sección 4.1.3*

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

LA CLAVE: DESACOPLAR BITS Y NIVELES

¿CÓMO FUNCIONA LA CUANTIZACIÓN FLEXIBLE?

Concepto: La idea central es que el número de niveles de cuantización (`K`) no tiene por qué ser una potencia del número de bits (`b`). Esto nos da la libertad de elegir los niveles de forma óptima (como en la no uniforme) pero forzándolos a una grilla que el hardware sí entiende (como en la uniforme).

Visual: Usar la figura `figures/flex_ptq.drawio.png` que es perfecta para explicar esto.

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

ENCONTRANDO LA CONFIGURACIÓN ÓPTIMA

APORTE 2: EL PROCESO DE OPTIMIZACIÓN (PTQ Y QAT)

Concepto: Se desarrollaron dos métodos para encontrar los mejores parámetros de cuantización:

1. **PTQ (Post-Entrenamiento):** Rápido y sin reentrenamiento.
2. **QAT (Consciente del Entrenamiento):** Más lento pero recupera más precisión.

Mencionar: "Estos algoritmos buscan la mejor configuración respetando siempre las restricciones de hardware, como se explica en detalle en el Capítulo 4".

Fuente: `02-main_matter/4-flexible_quantization.tex` *- Secciones 4.3 y 4.4*

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

¿FUNCIONÓ? ¡SÍ!

RESULTADOS CLAVE

Concepto: Mostrar los resultados más impactantes de forma directa.

- Se logró comprimir la red hasta **20 veces** (equivalente a ~1.58 bits) con una pérdida de precisión mínima.
- En muchos casos, se recuperó el **100% de la precisión** del modelo original, pero con una fracción del costo computacional.

Fuente: `02-main_matter/5-conclusions.tex` *- Sección 5.2.3*

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

MÁS PRECISIÓN CON LA MISMA COMPLEJIDAD

GRÁFICO DE RESULTADOS: FQ VS. CUANTIZACIÓN UNIFORME

Visual: Mostrar un gráfico de "Precisión vs. Complejidad". La curva de Cuantización Flexible (FQ) debe estar por encima de la de Cuantización Uniforme (UQ), demostrando que para un mismo nivel de complejidad, FQ es más precisa.

Idea: Puedes generar una versión simplificada de las figuras de `flex_ptq_results.tex` o los heatmaps de `figure_heatmap_sep_...png`.

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES PARA LA CUANTIZACIÓN

CONTRIBUCIONES VALIDADAS POR LA COMUNIDAD CIENTÍFICA

VALIDACIÓN ACADÉMICA: PUBLICACIONES

Contenido: Listar las publicaciones derivadas de la tesis.

- *"Methodology for CNN Implementation in FPGA-Based Embedded Systems", IEEE Embedded Systems Letters, 2023*
- *(Agregar otras publicaciones si existen)*

RESULTADOS Y CONCLUSIONES

RESULTADOS Y CONCLUSIONES

RESUMEN DE APORTES

CONCLUSIONES PRINCIPALES

Contenido: Reforzar los dos mensajes clave.

1. Se propuso una **metodología sistemática** que ordena y eficientiza la implementación de CNNs en hardware.
2. Se creó la **cuantización flexible**, un esquema que logra un balance superior entre precisión y eficiencia, manteniendo la compatibilidad con hardware real.

Fuente: `02-main_matter/5-conclusions.tex` *- Sección 5.1*

RESULTADOS Y CONCLUSIONES

¿Y AHORA QUÉ?

IMPACTO Y TRABAJOS FUTUROS

Contenido:

- **Impacto:** Este trabajo facilita el despliegue de IA avanzada en dispositivos de borde, con aplicaciones en tiempo real y de bajo consumo.
- **Futuro:** Extender la metodología a arquitecturas más complejas (ResNet), automatizar más el flujo y explorar nuevas aplicaciones.

Fuente: `02-main_matter/5-conclusions.tex` *- Secciones 5.4 y 6*

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

PREGUNTAS

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

¡MUCHAS GRACIAS!