.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

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

DOCTORADO DE LA UNIVERSIDAD DE BUENOS AIRES, ÁREA INGENIERÍA

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)

Jurados:

Dr. Hernán Merlino (FIUBA-CONICET)
Dr. Pedro Julián (UNS-CONICET)
Dr. Gustavo Javier Meschino (UNMDP-CONICET)
Dra. Paola Britos (UNRN-SPU)

Customize and control Google Chrome

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

AGENDA

.UBAfiuba FACULTAD DE INGENIERÍA

AGENDA

Laboratorio de Sistemas Embebidos

1. Introducción a la IA en el mundo real.
2. Mi aporte: una solución en dos partes.
3. Primera parte: Metodología de implementación sistemática.
4. Segunda parte: Cuantización flexible para maximizar eficiencia.
5. Conclusiones y trabajos futuros.

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

HISTORIA

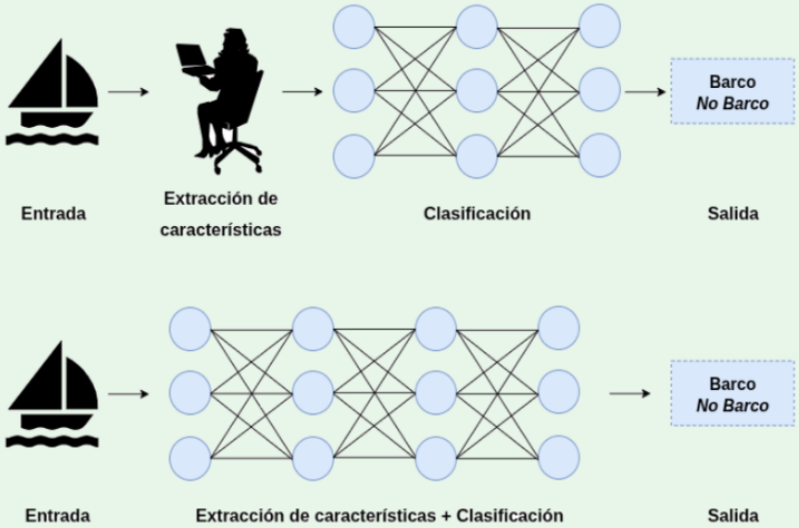
ChatGPT-5 (2025).
Imagen generada con inteligencia artificial.
OpenAI Image Generation.
https://openai.com

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

MODELOS TRADICIONALES Y MODERNOS



.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

APLICACIONES Y MODELOS DE IA

TIPOS DE MODELOS DE IA

- **Redes Neuronales Convolucionales (CNN)**
  - Procesamiento de imágenes
  - Reconocimiento de patrones visuales
  - Clasificación de imágenes
- **Redes Neuronales Recurrentes**
  - Procesamiento de secuencias
  - Lenguaje natural
  - Series temporales
- **Transformers**
  - Modelos de lenguaje (GPT, BERT)
  - Procesamiento paralelo
  - Atención multi-cabeza

APLICACIONES ACTUALES DE IA

- **Visión por Computadora**
  - Reconocimiento facial
  - Autos autónomos
  - Diagnóstico médico por imágenes
- **Procesamiento de Lenguaje**
  - Chatbots y asistentes virtuales
  - Traducción automática
  - Generación de texto
- **Otros Dominios**
  - Recomendaciones (Netflix, Spotify)
  - Juegos (AlphaGo, chess engines)
  - Robótica

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

ALGUNOS TAMAÑOS DE REDES NEURONALES DE REFERENCIA

Tipo de Red	Modelo	Parámetros	FLOPs aprox.	Comentarios
No profunda	MLP 1 capa (784→128→10)	~100,000	~0.1 MFLOPs	Ej. clasificación MNIST
No profunda	MLP 2 capas	~500,000	~0.5 MFLOPs	Más expresividad, aún simple
Profunda	Deep MLP (5 capas, 512 neuronas)	~5M	~5 MFLOPs	Tareas estructuradas, embebidas
Profunda	Deep MLP (10 capas, 1024)	~50M	~50 MFLOPs	No escalan tan bien como CNN/Transformers
CNN	LeNet-5	~60K	~0.06 MFLOPs	Muy liviana, MNIST
CNN	AlexNet	~60M	~720 MFLOPs	Imagenet (2012)
CNN	VGG-16	~138M	~15.5 GFLOPs	Precisa, pero costosa
CNN	ResNet-50	~25.6M	~4 GFLOPs	Más eficiente, uso moderno
CNN	EfficientNet-B0	~5.3M	~0.39 GFLOPs	Muy eficiente, misma precisión que ResNet-50
Transformer	BERT Base	~110M	~22 GFLOPs	12 capas, 768 dim, encoder NLP
Transformer	BERT Large	~340M	~66 GFLOPs	24 capas, más lento pero potente
Transformer	GPT-2 Small	~117M	~20 GFLOPs	Generativo
Transformer	GPT-3	~175B	~350 TFLOPs	Requiere hardware especializado
ViT	ViT-B/16	~86M	~17.6 GFLOPs	Imagenet, imágenes 224x224
ViT	ViT-L/16	~307M	~60 GFLOPs	Más preciso pero costoso

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

MOTIVACIÓN Y OBJETIVOS

Motivación:

- IAs cada vez más grandes
- que solucionana problemas cada vez más complejos
- en diversos campos de aplicación

Objetivos:

- IAs más eficientes a nivel de energía
- y a nivel de costos

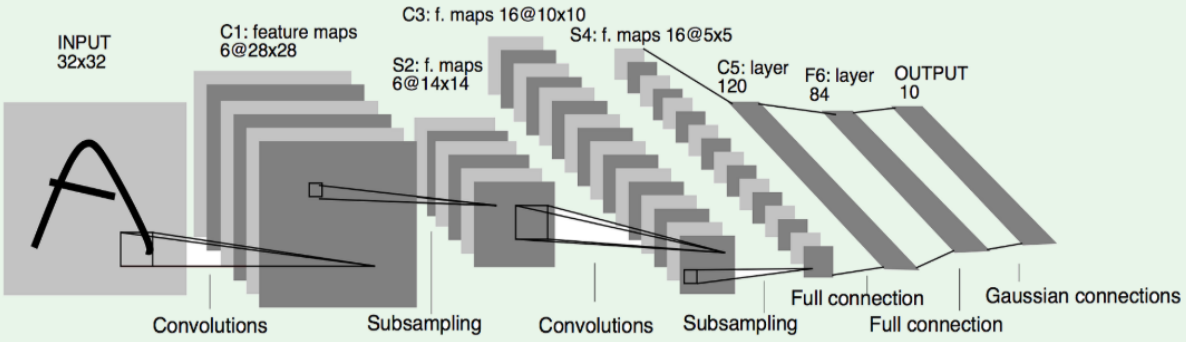
.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

VISIÓN CON IA: REDES NEURONALES CONVOLUCIONALES

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



Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998).
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11), 2278–2324.
https://doi.org/10.1109/5.726791

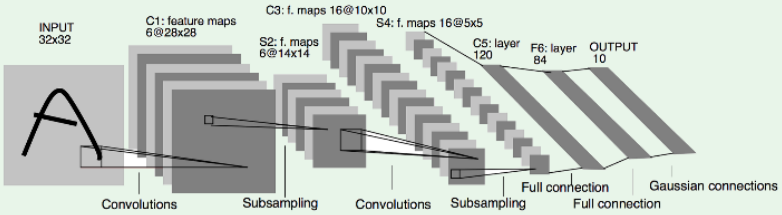
.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

VISIÓN CON IA: REDES NEURONALES CONVOLUCIONALES

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



Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

- Las primeras capas detectan cosas simples (tramas, bordes, colores)
- Las capas más profundas reconocen objetos complejos (caras, autos)
- El conjunto de todas las capas convolucionales extraen las características de la imagen
- Las capas finales son densas y su función varía según el problema

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

LA SOLUCIÓN CLAVE ES LA COMPRESIÓN DE LA RED

Existe un **relación de compromiso** entre costo y calidad o entre complejidad y precisión

Customize and control Google Chrome

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

LA SOLUCIÓN CLAVE ES LA COMPRESIÓN DE LA RED

Existe un **relación de compromiso** entre costo y calidad o entre complejidad y precisión

Tamaño original

.UBAfiuba FACULTAD DE INGENIERÍA

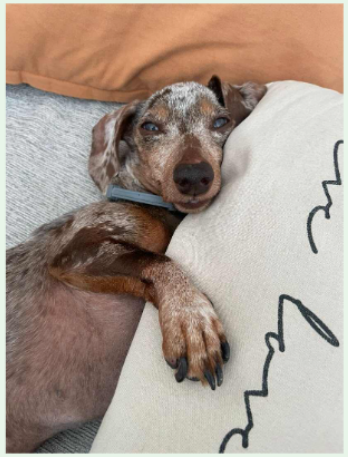
INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

LA SOLUCIÓN CLAVE ES LA COMPRESIÓN DE LA RED

Existe un **relación de compromiso** entre costo y calidad o entre complejidad y precisión

Tamaño original



50%

.UBAfiuba FACULTAD DE INGENIERÍA

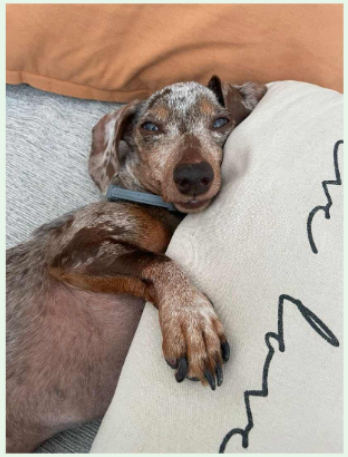
INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

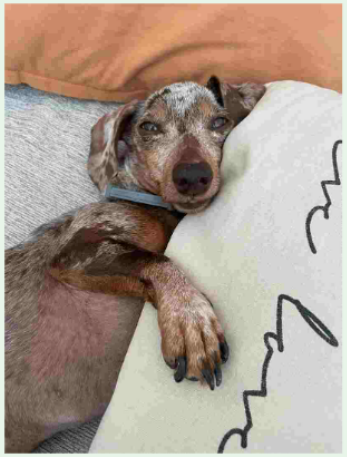
LA SOLUCIÓN CLAVE ES LA COMPRESIÓN DE LA RED

Existe un **relación de compromiso** entre costo y calidad o entre complejidad y precisión

Tamaño original



50%



10%

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

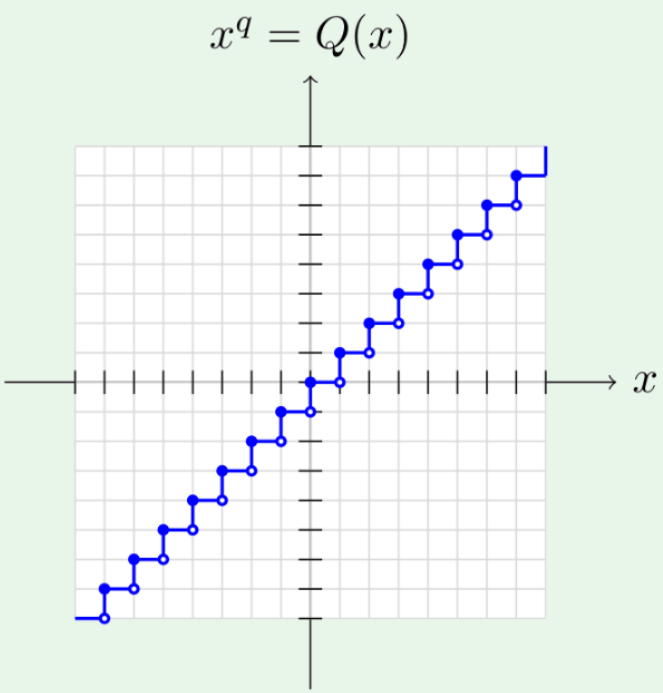
LA SOLUCIÓN CLAVE ES LA COMPRESIÓN DE LA RED

Existe un **relación de compromiso** entre costo y calidad o entre complejidad y precisión

Tamaño original | 50% | 10% | 5%

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

LA COMPRESIÓN A TRAVÉS DE LA CUANTIZACIÓN



A grandes rasgos:

- el tamaño de la red está definido por el número de **pesos o parámetros** que tiene y el *tamaño* de cada uno,
- y la **complejidad de las operaciones** está definido por el *tamaño* de los operandos.

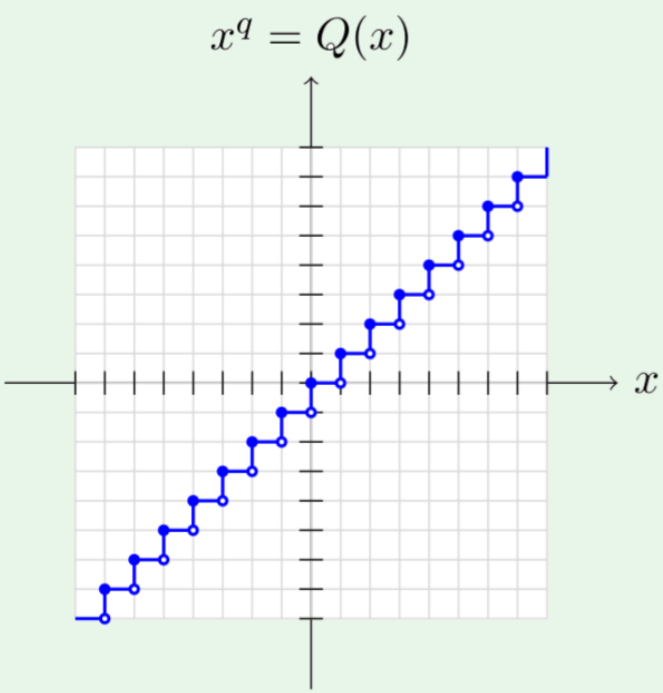
¿Podemos usar números de "baja precisión" para hacer la red más rápida y pequeña?

.UBAfiuba FACULTAD DE INGENIERÍA

INTRODUCCIÓN Y CONCEPTOS FUNDAMENTALES

Laboratorio de Sistemas Embebidos

LA COMPRESIÓN A TRAVÉS DE LA CUANTIZACIÓN



Sí, podemos:

- Cuantización de **pesos**:
  - No se reduce el número de pesos
  - Sí se reduce el *tamaño* de los pesos
  - Sí se reduce la complejidad de las operaciones
- Cuantización de **activaciones**:
  - No se reduce el número de valores de las activaciones
  - Sí se reduce el *tamaño* de las activaciones
  - Sí se reduce la complejidad de las operaciones

MIS APORTES ORIGINALES

.UBAfiuba FACULTAD DE INGENIERÍA

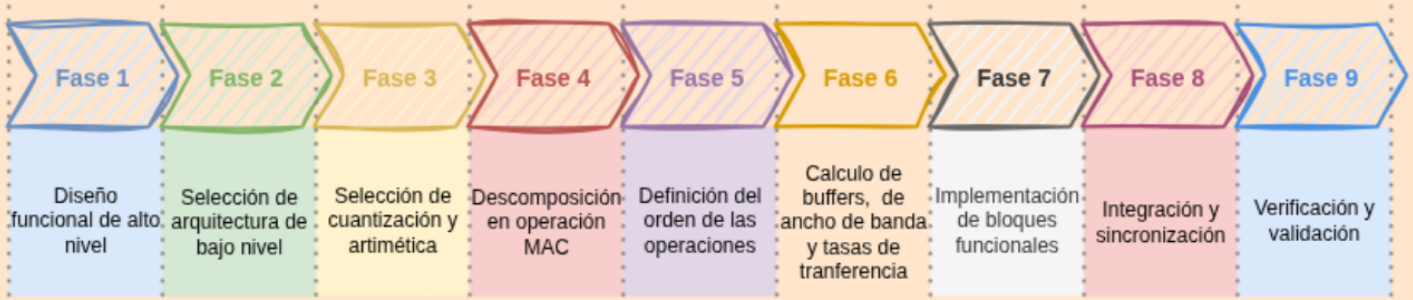
MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

MI PROPUESTA: UN PUENTE INTEGRAL DEL SOFTWARE AL HARDWARE

El núcleo de esta tesis es la implementación de modelos de IA en hardware de forma eficiente. Mis dos aportes pueden verse como ideas complementarias:

1. **La metodología sistemática:** Provee la **estructura** para la construcción del hardware. Responde a la pregunta: *¿cómo lo construimos de forma ordenada?*
2. **La cuantización flexible:** 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?*



.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

.UBAfiuba FACULTAD DE INGENIERÍA

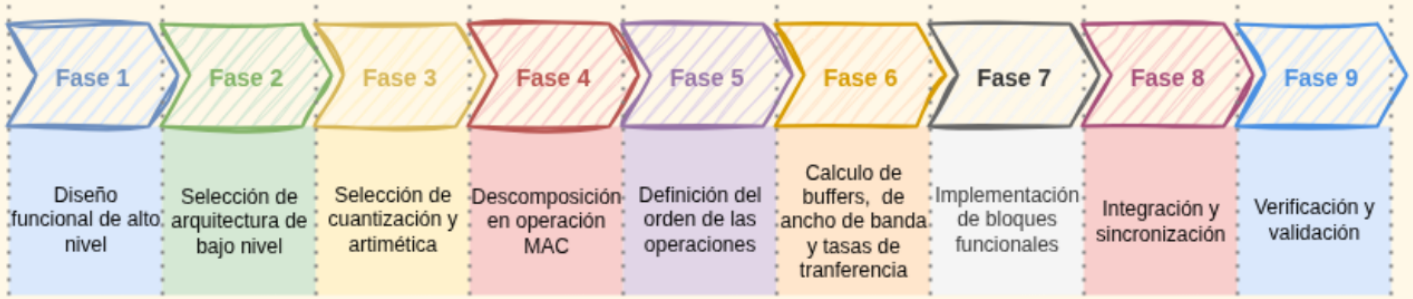
Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

DEL CAOS AL ORDEN: UN PROCESO EN 9 FASES:

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



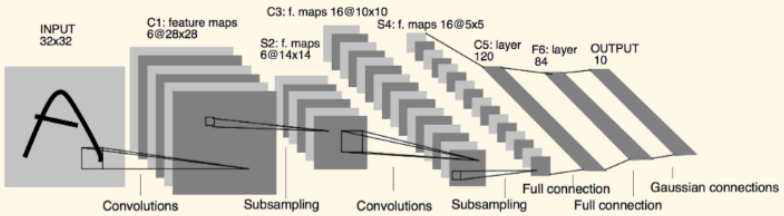
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 1: Diseño funcional de alto nivel



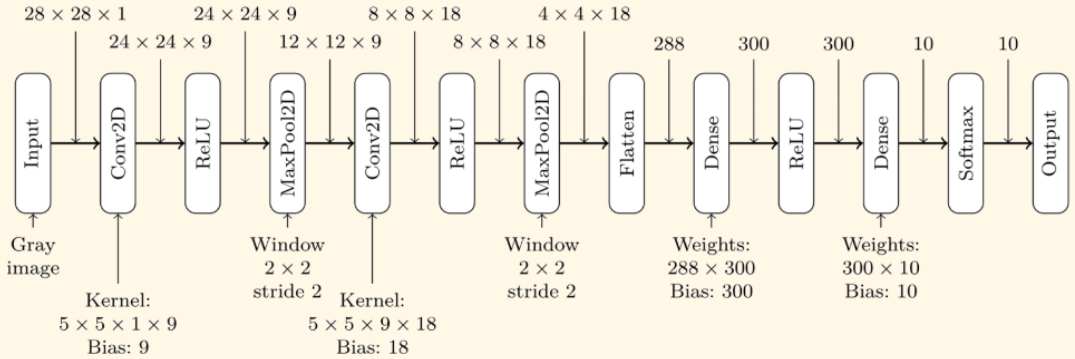
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 2: Selección de arquitectura de bajo nivel



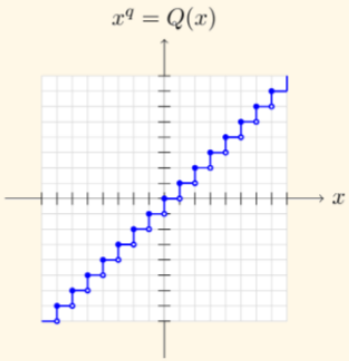
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 3: Selección de cuantización y aritmética



.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 4: Descomposición en operaciones MAC

$$y_{i,j,k} = \left(\sum_{m=1}^{w^{\mathrm{K}}} \sum_{n=1}^{h^{\mathrm{K}}} \sum_{l=1}^{c^{\mathrm{I}}} x_{i+m-1,j+n-1,l} \cdot K_{m,n,l,k} \right) + B_k$$

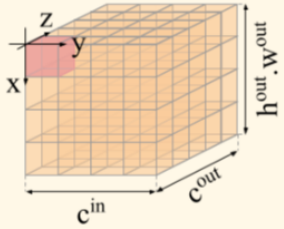
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 5: Definición del orden de operaciones



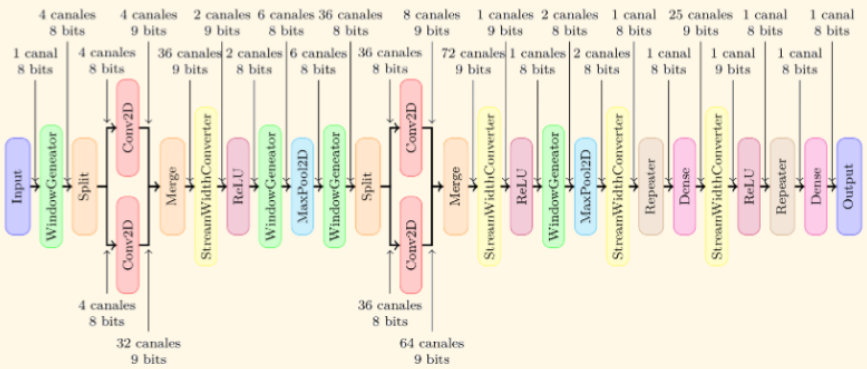
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 6: Diseño del flujo de datos



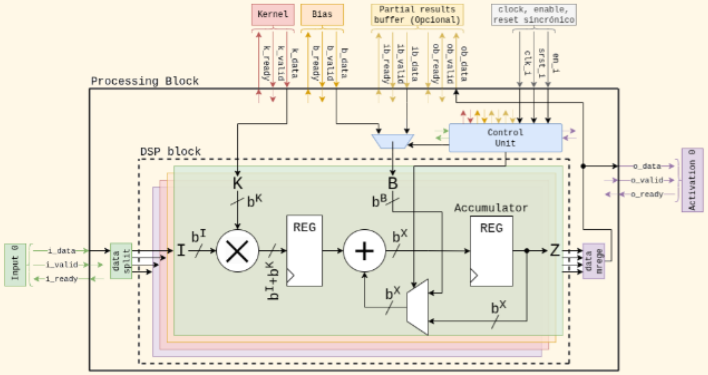
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 7: Implementación de bloques funcionales



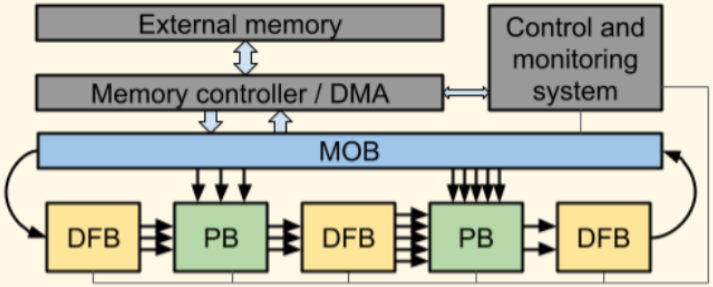
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Fase 8: Integración y sincronización

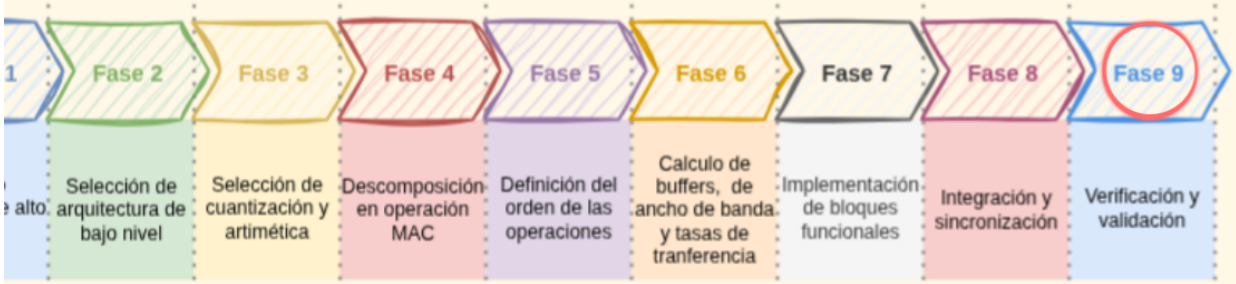


.UBAfiuba FACULTAD DE INGENIERÍA

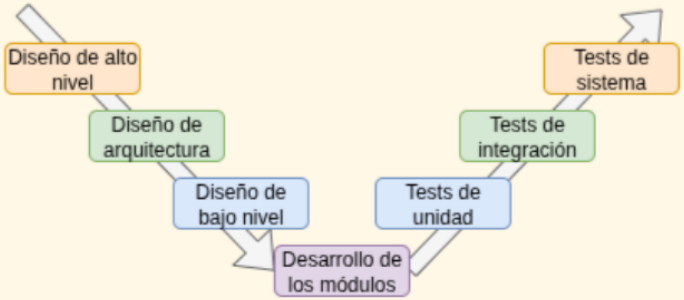
Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN



Fase 9: Verificación y validación



.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

V1: DISEÑO TRADICIONAL

- Un bloque por capa
- Hardware adaptado al modelo
- Restricciones de potencias de 2

V2: METODOLOGÍA PROPUESTA

- Flexibilidad de diseño
- Modelo exacto implementado
- PEs no restringidos a potencias de 2

REFERENCIA: GONG ET AL. (2018)

- Implementación de referencia
- 205 DSPs, 242 BRAMs
- 54.61 GOPS @ 142.8 MHz

L. Gong, C. Wang, X. Li, H. Chen, and X. Zhou, "MALOC: A fully pipelined FPGA accelerator for convolutional neural networks with all layers mapped on chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2601–2612, 2018, doi: 10.1109/TCAD.2018.2857078.

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Rendimiento
11.30 → 36.48 GOPS
×3.23

Imágenes/s
17,371 → 38,141
×2.20

Latencia
99.3 → 70.8 µs
−28.7%

Eficiencia DSP
0.404 → 0.953
×2.36

Eficiencia computacional superior:
V2 logra el 67% del rendimiento de referencia usando solo el 65% de los DSPs (134 vs 205 DSPs)

Cerca del límite teórico:
38.1 kimg/s vs 39.7 kimg/s máximo ideal (96.15% de eficiencia)

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

METODOLOGÍA PARA LA IMPLEMENTACIÓN

Métrica	Gong et al. (2018)	Gong et al. (2018)*	V1	V2
# de DSP	205	205	98	134
# de BRAM	242	242	2	2
Frecuencia [MHz]	200	142.8	142.8	142.8
Imágenes/s	-	-	17,371	38,141
Rendimiento [GOPS]	76.48	54.61	11.30	36.48
Total # de Operaciones [OP]	-	-	650,640	956,400
Latencia [μs]	-	-	99.3	**70.8**
Rend./(# DSP) [MOPS/DSP]	373.07	266.37	115.33	**272.22**
Eficiencia DSP	0.933	0.933	0.404	**0.953**

**Valores escalados para una frecuencia de trabajo de 142.8 MHz.*

L. Gong, C. Wang, X. Li, H. Chen, and X. Zhou, "MALOC: A fully pipelined FPGA accelerator for convolutional neural networks with all layers mapped on chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2601–2612, 2018, doi: 10.1109/TCAD.2018.2857078.

.UBAfiuba FACULTAD DE INGENIERÍA

MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

METODOLOGÍA PARA LA IMPLEMENTACIÓN

RESULTADOS: UNA ARQUITECTURA REAL Y VERIFICADA

- 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).
- Se demostró que el método sistemático permite crear arquitecturas que son modulares, escalables y, lo más importante, verificables.

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

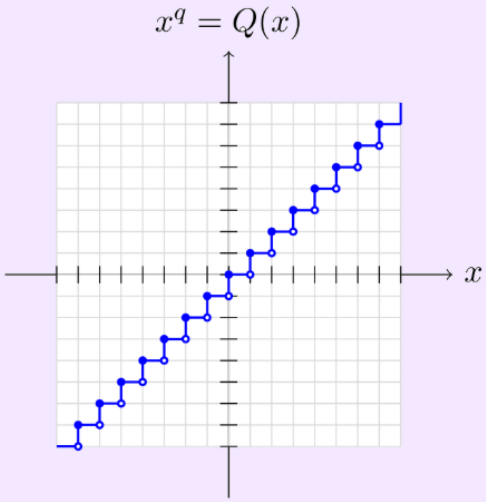
.UBAfiuba FACULTAD DE INGENIERÍA

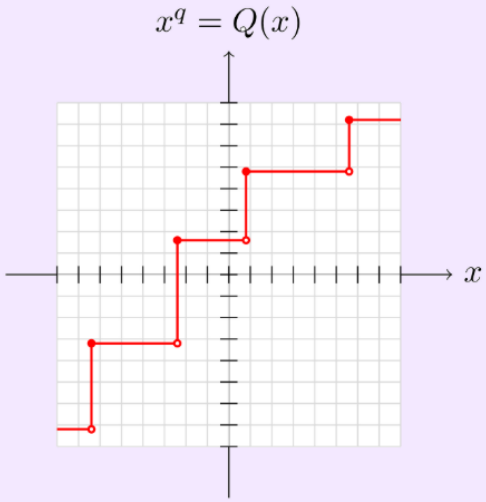
Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

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





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.

.UBAfiuba FACULTAD DE INGENIERÍA

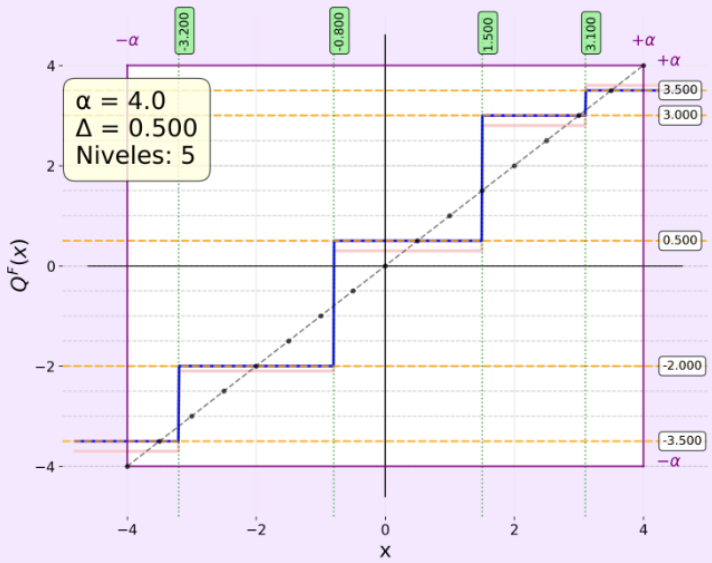
MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

CUANTIZACIÓN FLEXIBLE

¿CÓMO FUNCIONA LA CUANTIZACIÓN FLEXIBLE?

- La idea central es que el número de niveles de cuantización (k) no tiene por qué ser igual a 2^b.
- Además los umbrales, pueden estar en cualquier posición.
- Esto nos da la libertad de elegir los niveles y umbrales de forma óptima (**no uniforme**) pero forzándolos a una grilla que el hardware sí entiende (**uniforme**).
- Se puede interpretar como una aproximación de la función de cuantización no uniforme, compatible con el hardware de punto fijo.

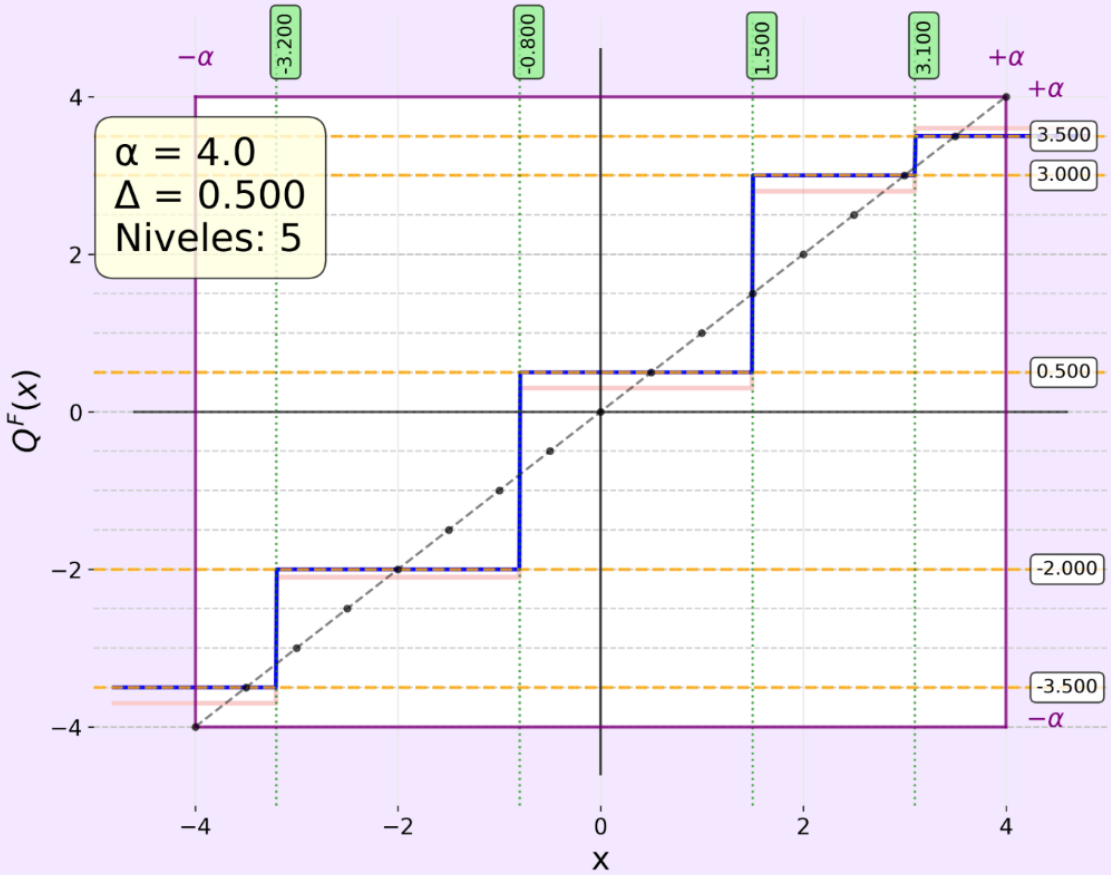


.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE



.UBAfiuba FACULTAD DE INGENIERÍA

MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

CUANTIZACIÓN FLEXIBLE

EN BUSCA DE LA CONFIGURACIÓN ÓPTIMA

Considerando la red VGG16 y cuantización uniforme tenemos 16 capas a las que les tenemos que asignar:

- el número de bits (b): 16 valores
- el valor de *clipping* (α): 16 valores

Son 32 valores en total

.UBAfiuba FACULTAD DE INGENIERÍA

MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

CUANTIZACIÓN FLEXIBLE

EN BUSCA DE LA CONFIGURACIÓN ÓPTIMA

Considerando la red VGG16 y cuantización flexible tenemos 16 capas a las que les tenemos que asignar:

- el número de bits (b): 16 valores
- el valor de *clipping* (α): 16 valores
- el número de nivles (k): 16 valores
- los valores de los niveles (v): $\sum k_i$ valores
- los valores de los umbrales (t): $\sum (k_i-1)$ valores

Asumiendo k entre 2 y 10, tenemos entre 96 y 352 valores en total

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

EL PROCESO DE CUANTIZACIÓN

Se desarrollaron dos métodos para cuantizar a los modelos:

1. **Cuantización post-entrenamiento (PTQ):**
    1. Sin ajuste fino: rápido y sin reentrenamiento.
    2. Con ajuste fino: más lento, pero con mejore resultados. Requiere datos.
2. **Entrenamiento consciente de cuantización (QAT):** más lento pero recupera más precisión. Require datos.

.UBAfiuba FACULTAD DE INGENIERÍA

MIS APORTES ORIGINALES

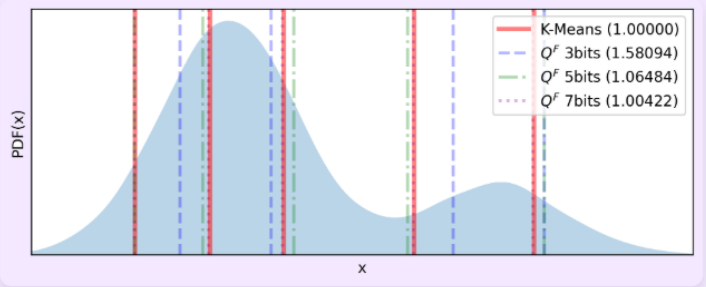
Laboratorio de Sistemas Embebidos

CUANTIZACIÓN FLEXIBLE

EL PROCESO DE OPTIMIZACIÓN PARA PTQ - EXPLORACIÓN Y CASOS DE USO

PROCESO DE OPTIMIZACIÓN:

1. Se fijan α, b, y k, para una capa
2. Se ajustan para cumplir con las restricciones
3. Se calculan v y t utilizando k-means
4. Se calcula la *accuracy*
5. Se realiza el ajuste fino
6. Se recalcula la *accuracy*
7. Se pasa a la siguiente capa
8. Se repite el proceso



.UBAfiuba FACULTAD DE INGENIERÍA

MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

CUANTIZACIÓN FLEXIBLE

EL PROCESO DE OPTIMIZACIÓN PARA PTQ - MAYOR GRANULARIDAD

No se realizó una exploración exhaustiva, sino que se consideraron 3 configuraciones:

- Aritmética uniforme
- Primeras tres y últimas tres capas con el doble número de niveles
- Inversamente proporcional al tamaño de los pesos de la capa

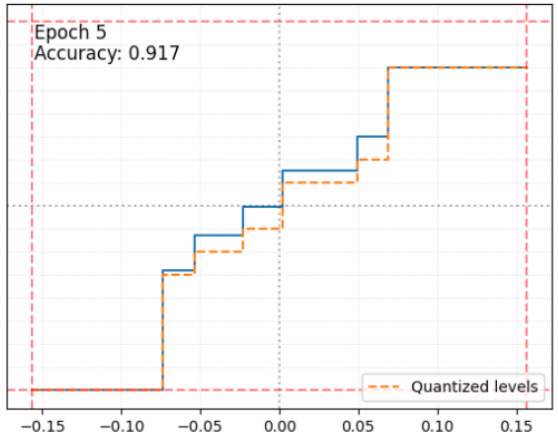
MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

EL PROCESO DE OPTIMIZACIÓN PARA QAT

PROCESO DE OPTIMIZACIÓN:

1. Se fijan α, b, y k, para la red
2. Se entrenan α, v y t, junto con los pesos de la red
3. Se calcula la *accuracy* y la complejidad
4. Se repite el proceso



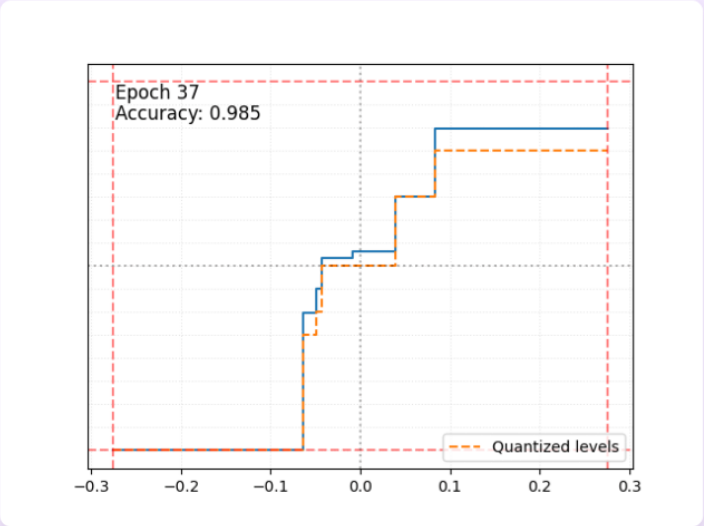
MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

EL PROCESO DE OPTIMIZACIÓN PARA QAT

PROCESO DE OPTIMIZACIÓN:

1. Se fijan α, b, y k, para la red
2. Se entrenan α, v y t, junto con los pesos de la red
3. Se calcula la *accuracy* y la complejidad
4. Se repite el proceso



.UBAfiuba FACULTAD DE INGENIERÍA

MIS APORTES ORIGINALES

Laboratorio de Sistemas Embebidos

CUANTIZACIÓN FLEXIBLE

¿FUNCIONÓ? ¡SÍ!

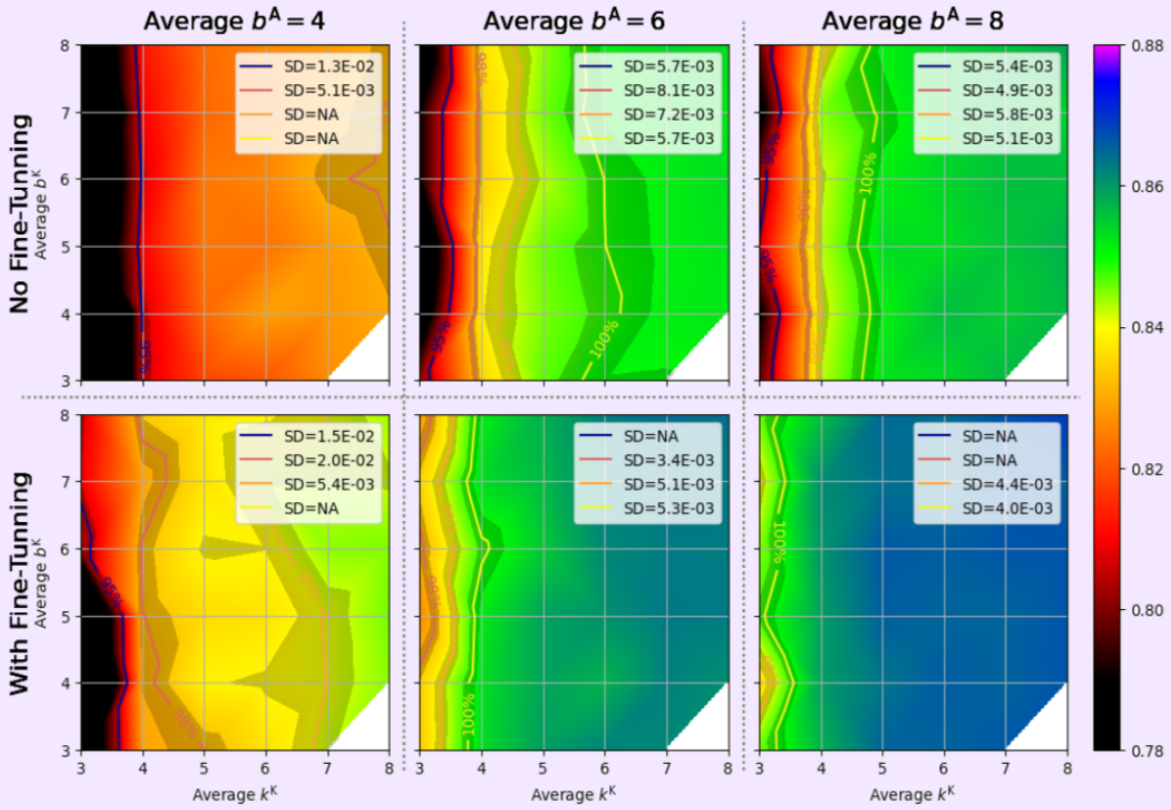
- Se logró comprimir la red hasta **20 veces** (equivalente a ~1.58 bits) con una pérdida de precisión mínima.
- En muchos casos, se recuperó el **100% de la precisión** del modelo original, pero con una fracción del costo computacional.
- Se demostró el potencial de aumentar la granularidad para mejorar la precisión.
- Se demostró el potencial de utilizar QAT para una compresión más profunda.

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE



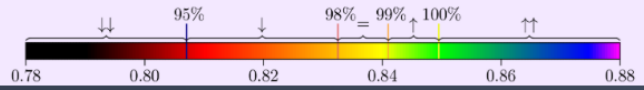
.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

Case			Accuracy (SD)		Complexity		
			Without FT	With FT	BTC	BSC	BAC
OM-FP32			0.852 (0.013)	– (–)	NA	1.0	1.0
OM-FP16			0.852 (0.013)	– (–)	NA	2.0	2.0
UQ-MP-FxP	b^A	b^K	Without FT	With FT	BTC	BSC	BAC
	4	2	0.422 (0.131) ↓↓	0.644 (0.088) ↓↓	2.4	16.0	29.3
		3	0.791 (0.026) ↓↓	0.821 (0.016) ↓	2.2	10.7	29.3
	6	2	0.528 (0.078) ↓↓	0.751 (0.047) ↓↓	2.2	16.0	19.5
		3	0.815 (0.023) ↓	0.844 (0.010) ↑	1.9	10.7	19.5
	8	2	0.567 (0.085) ↓↓	0.765 (0.055) ↓↓	2.0	16.0	14.7
		3	0.825 (0.018) ↓	0.851 (0.009) ↑↑	1.7	10.7	14.7
FQ-MP-FxP	b^A	$b^K(k^K)$	Without FT	With FT	BTC	BSC	BAC
	4	3 (3)	0.710 (0.061) ↓↓	0.770 (0.054) ↓↓	2.2	20.2	29.3
		8 (3)	0.750 (0.050) ↓↓	0.775 (0.094) ↓↓	1.5	20.2	29.3
		3 (4)	0.802 (0.022) ↓↓	0.826 (0.010) ↓	2.2	16.0	29.3
		8 (4)	0.813 (0.008) ↓	0.834 (0.005) =	1.5	16.0	29.3
		4 (8)	0.829 (0.005) ↓	0.844 (0.004) ↑	2.0	10.7	29.3
		8 (8)	0.831 (0.004) ↓	0.844 (0.005) ↑	1.5	10.7	29.3
	6	3 (3)	0.797 (0.019) ↓↓	0.835 (0.017) =	1.9	20.2	19.5
		8 (3)	0.783 (0.041) ↓↓	0.830 (0.017) ↓	1.2	20.2	19.5
		3 (4)	0.834 (0.009) =	0.852 (0.009) ↑↑	1.9	16.0	19.5
		8 (4)	0.835 (0.010) =	0.850 (0.008) ↑↑	1.2	16.0	19.5
		4 (8)	0.851 (0.006) ↑↑	0.860 (0.006) ↑↑	1.7	10.7	19.5
		8 (8)	0.852 (0.005) ↑↑	0.862 (0.005) ↑↑	1.2	10.7	19.5
	8	3 (3)	0.800 (0.024) ↓↓	0.845 (0.011) ↑	1.7	20.2	14.7
		8 (3)	0.796 (0.030) ↓↓	0.835 (0.033) =	1.0	20.2	14.7
		3 (4)	0.840 (0.009) =	0.856 (0.006) ↑↑	1.7	16.0	14.7
		8 (4)	0.838 (0.007) =	0.856 (0.006) ↑↑	1.0	16.0	14.7
		4 (8)	0.856 (0.004) ↑↑	0.866 (0.004) ↑↑	1.5	10.7	14.7
		8 (8)	0.859 (0.003) ↑↑	0.868 (0.005) ↑↑	1.0	10.7	14.7

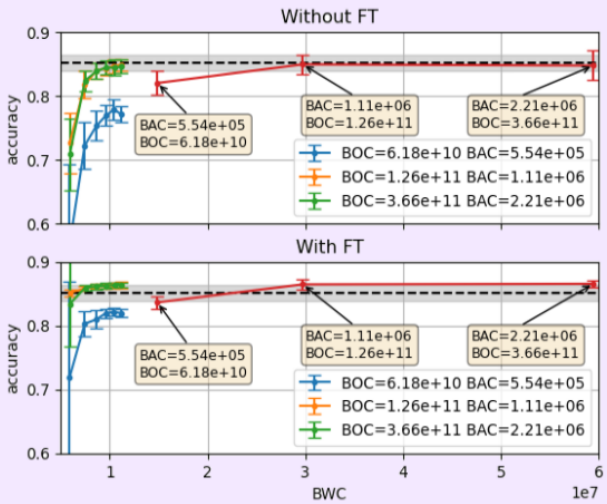


.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

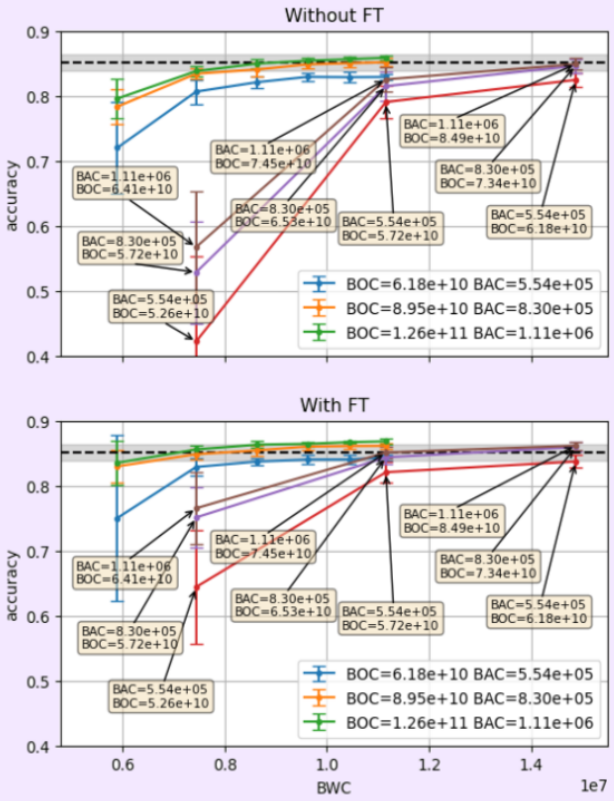


.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

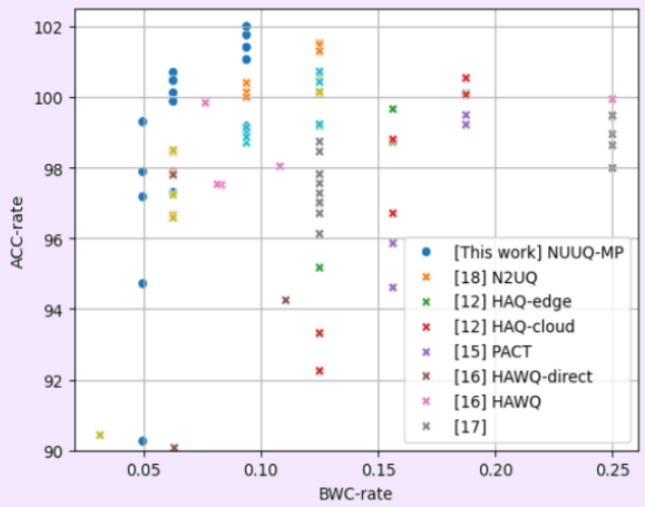


.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

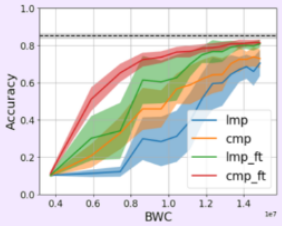


.UBAfiuba FACULTAD DE INGENIERÍA

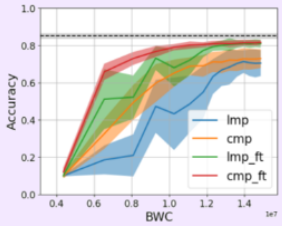
Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

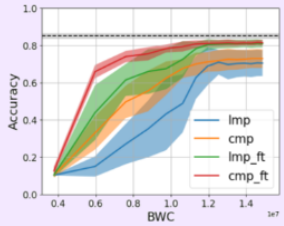
CUANTIZACIÓN FLEXIBLE



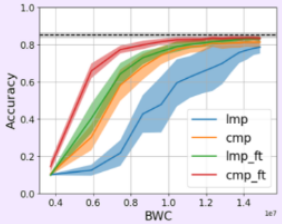
(a) 4 bits, configuración 0



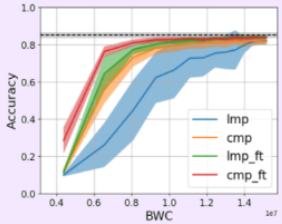
(b) 4 bits, configuración 1



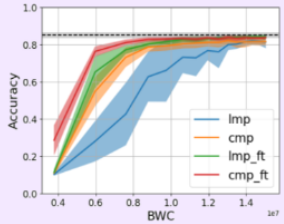
(c) 4 bits, configuración 2



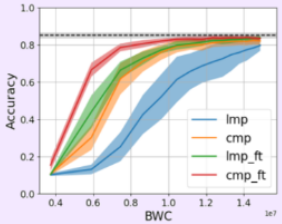
(d) 6 bits, configuración 0



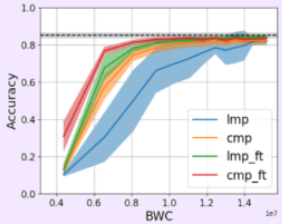
(e) 6 bits, configuración 1



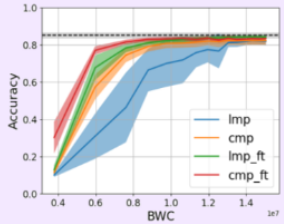
(f) 6 bits, configuración 2



(g) 8 bits, configuración 0



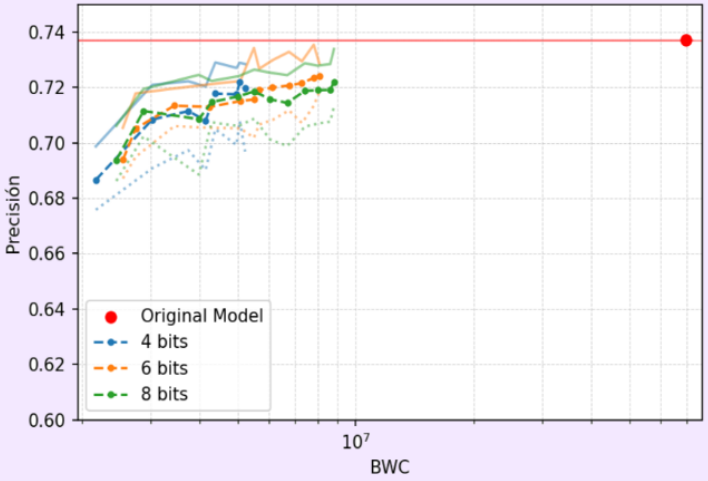
(h) 8 bits, configuración 1



(i) 8 bits, configuración 2

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE

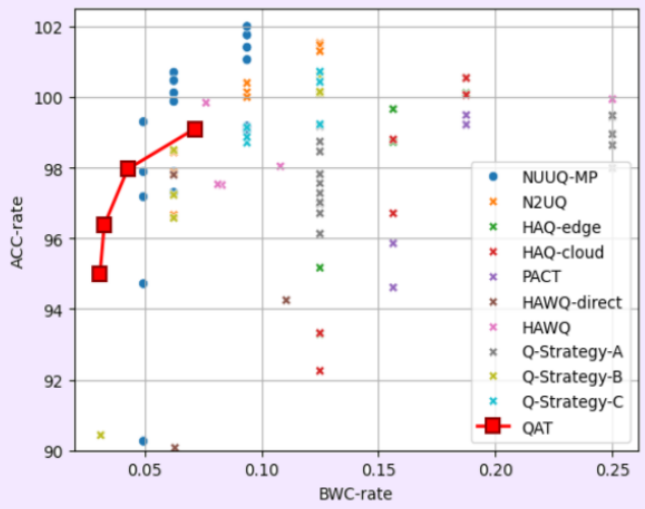


.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

MIS APORTES ORIGINALES

CUANTIZACIÓN FLEXIBLE



CONCLUSIONES Y TRABAJOS FUTUROS

.UBAfiuba FACULTAD DE INGENIERÍA

CONCLUSIONES Y TRABAJOS FUTUROS

Laboratorio de Sistemas Embebidos

RESUMEN DE APORTES

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.

.UBAfiuba FACULTAD DE INGENIERÍA

CONCLUSIONES Y TRABAJOS FUTUROS

Laboratorio de Sistemas Embebidos

PUBLICACIONES

- F. G. Zacchigna, “Methodology for CNN Implementation in FPGA-Based Embedded Systems,” IEEE Embedded Systems Letters, vol. 15, no. 2, pp. 85–88, 2023, doi: 10.1109/LES.2022.3187382.
- F. G. Zacchigna, S. Lew, and A. Lutenberg, “Flexible Quantization for Efficient Convolutional Neural Networks,” *Electronics*, vol. 13, no. 10, Art. no. 1923, 2024. [Online]. Available: https://www.mdpi.com/2079-9292/13/10/1923. doi: 10.3390/electronics13101923
- Los resultados sobre la granularidad por canal todavía no fueron publicados.
- Los resultados de QAT todavía no fueron publicados.

.UBAfiuba FACULTAD DE INGENIERÍA

CONCLUSIONES Y TRABAJOS FUTUROS

Laboratorio de Sistemas Embebidos

¿Y AHORA QUÉ?

- **Algoritmos de optimización mejorados:** Optimización bayesiana y técnicas evolutivas para búsqueda más eficiente de configuraciones de cuantización
- **Cuantización de activaciones flexible:** Extensión del cuantizador flexible a activaciones con hardware especializado
- **Granularidad adaptativa:** Técnicas automáticas para determinar el nivel óptimo de granularidad por canal, grupo o capa
- **Framework completo:** Herramientas de desarrollo, análisis y simulación para cuantización flexible

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

¿PREGUNTAS?

.UBAfiuba FACULTAD DE INGENIERÍA

Laboratorio de Sistemas Embebidos

¡MUCHAS GRACIAS!