Aceleración de redes neuronales bayesianas en procesadores de bajo consumo RISC-V

Samuel Pérez Pedrajas, Javier Resano Ezcaray, Darío Suárez Gracia

Afiliación: Grupo de Arquitectura de Computadores de Zaragoza (GAZ)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
E-mail: *samuel.perez@unizar.es*

Resumen

La fiabilidad es un aspecto crítico en las predicciones de inteligencia artificial no cubierto por la mayoría de redes neuronales. Al contrario, las redes neuronales bayesianas (BNN) ofrecen una solución permitiendo calcular métricas de incertidumbre junto a sus predicciones, a cambio de aumentar el coste de la inferencia. Este trabajo optimiza dicho proceso desarrollando una extensión personalizada de bajo coste y consumo para la arquitectura RISC-V.

Motivación

La Figura 1 muestra las principales diferencias entre una red neuronal (NN) clásica y una bayesiana (BNN). Considerando que ambas redes han sido entrenadas para clasificar imágenes de perros y gatos, cuando son expuestas a un dato anómalo como la imagen de un tigre solo la BNN es capaz de reportar este suceso utilizando las métricas de incertidumbre.

En la actualidad ninguna de las bibliotecas para trabajar con NN en dispositivos de bajas prestaciones tiene soporte para BNN, por lo que en este trabajo hemos desarrollado un motor de inferencia en C junto con un conversor de modelos para poder ejecutar modelos entrenados en TensorFlow.

El algoritmo de inferencia de las BNN requiere muestrear pesos de distribuciones gaussianas. Utilizando un algoritmo de muestreo sencillo basado en el teorema central del límite (CLT) este proceso ocupa el 80% de los ciclos de ejecución.

Para acelerar este algoritmo se proponen tanto optimizaciones software como aceleración hardware mediante una extensión al repertorio de instrucciones RISC-V.

Optimización software

Partiendo de Awano *et al.* [1] se propone transformar los parámetros de las distribuciones gaussianas en parámetros de distribuciones uniformes sin alterar la esperanza ni la varianza, de forma que se muestreen distribuciones uniformes, cuyo algoritmo de muestreo tiene un coste mucho menor.

Esta optimización reduce el porcentaje de ciclos dedicados al muestreo lo que permite obtener tiempos de ejecución 5 veces menores. Sin embargo, esta optimización en modelos grandes causa pérdidas de precisión y altera las métricas de incertidumbre, como se muestra en la Figura 2.

Extensión RISC-V

Para reducir el impacto del muestreo, este trabajo propone un nuevo generador de números pseudoaleatorios gaussianos (GRNG) basado en el CLT. Este GRNG genera 12 muestras de distribuciones uniformes mediante un lookahead linear-feedback-shift-register de 151 bits y las acumula con un árbol de sumadores segmentado en 4 ciclos, permitiendo obtener una muestra cada ciclo sin afectar a la frecuencia del reloj del diseño original, como se muestra en a Figura 4.

Esta unidad funcional (UF) se ha añadido a un procesador RISC-V de bajas prestaciones como una unidad funcional extra junto con 2 nuevas instrucciones para poder interactuar con él.

Las nuevas instrucciones y la UF reducen los ciclos aún más, resultando en tiempos de ejecución 8 veces menores sin afectar negativamente a la precisión o métricas de incertidumbre de ningún modelo. Añadir esta UF solamente aumenta el consumo del procesador en un 0.65%, por lo que prácticamente se obtiene un consumo 8 veces menor. La Figura 3 muestra un resumen de los diferentes porcentajes de ciclos dedicados al muestreo obtenidos con las diferentes optimizaciones.

Conclusiones

Este trabajo desarrolla un conjunto de herramientas y bibliotecas para poder ejecutar inferencia de BNN en entornos de bajas prestaciones. Además ofrece 2 métodos de optimización diferentes con diferentes ventajas y desventajas. Uno solamente basado en software con un rendimiento 5 veces mejor pero con pérdida de precisión en modelos grandes y uno basado en una extensión RISC-V con un

REFERENCIAS

[1]. AWANO, Hiromitsu y HASHIMOTO, Masanori. B2N2: Resource efficient Bayesian neural network accelerator using Bernoulli sampler on FPGA. *Integration*. 2022. ISSN 0167-9260. Disponible en: doi:10.1016/j.vlsi.2022.11.005

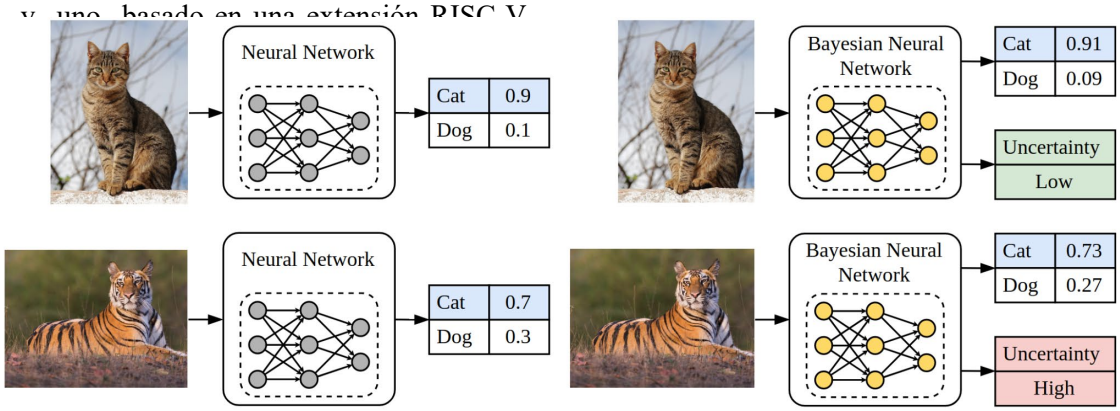

Figura 1. Comparación entre tipos de redes.

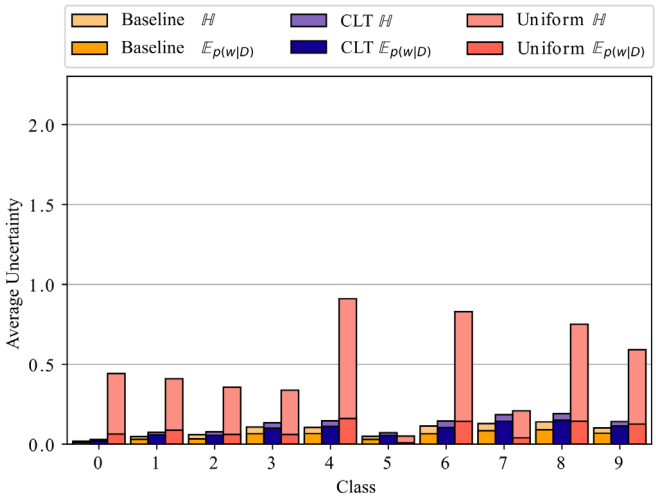

Figura 2. Comparación de métricas de incertidumbre.

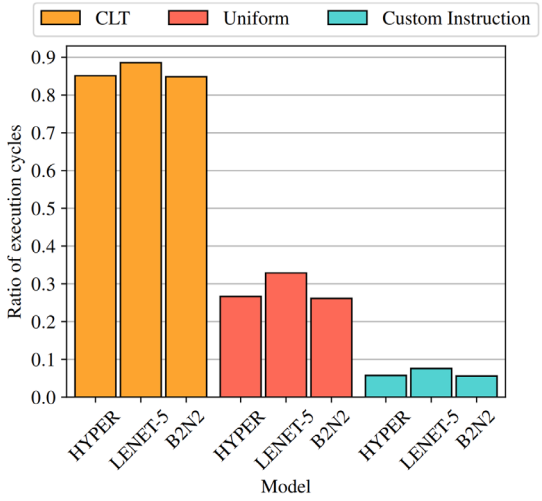

Figura 3. Porcentaje de ciclos dedicados al muestreo.

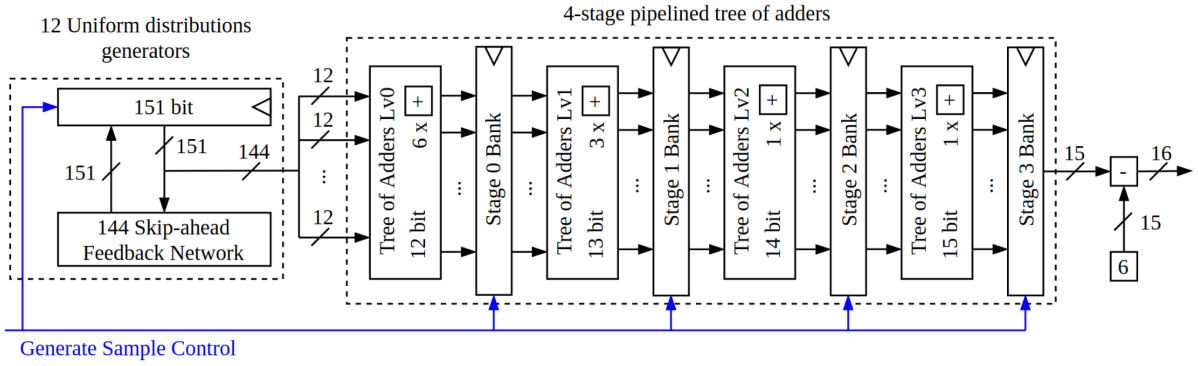

Figura 4. Diagrama del GRNG.