Universidad Alfonso X el Sabio

Representación en números de coma flotante

Germán Llorente y Carlos Puigserver

October 16, 2023

1 Introducción

La representación de números en coma flotante es un método utilizado para almacenar números reales en computadoras. En el caso de nuestro grado, el año pasado tuvimos la oportunidad de estudiar más este ámbito de las matemáticas en la asignatura Métodos Numéricos. Los formatos FP32, FP16 y BF16 son ejemplos de cómo se representan estos números. A continuación, se detalla cada formato y se exploran sus ventajas y desventajas en términos de precisión, rango y uso de memoria.

2 FP32 (Floating-Point 32-bit)

FP32 utiliza 32 bits para representar un número en coma flotante. Esto proporciona aproximadamente 7 dígitos de precisión decimal. FP32 puede representar números en el rango de aproximadamente $\pm 1.18 \times 10^{-38}$ a $\pm 3.4 \times 10^{38}$. Requiere 4 bytes por número en memoria.

2.1 Ventajas de FP32

- Buena precisión para una amplia gama de aplicaciones.
- Rango adecuado para cubrir una variedad de valores.

2.2 Desventajas de FP32

- Insuficiente para aplicaciones que requieren alta precisión.
- Uso relativamente alto de memoria.

3 FP16 (Half-Precision Floating-Point)

FP16 utiliza 16 bits para representar un número en coma flotante, ofreciendo aproximadamente 3-4 dígitos de precisión decimal. Puede representar números en el rango de aproximadamente $\pm 6.1 \times 10^{-5}$ a $\pm 6.5 \times 10^{4}$. Requiere 2 bytes por número en memoria.

3.1 Ventajas de FP16

- Usa la mitad de memoria en comparación con FP32, útil para aplicaciones con limitaciones de memoria.
- Adecuado para ciertas aplicaciones de aprendizaje profundo y redes neuronales donde la precisión moderada es suficiente.

3.2 Desventajas de FP16

- Precisión limitada, no apto para todas las aplicaciones, especialmente aquellas que requieren cálculos intensivos y precisión crítica.

4 BF16 (Brain Floating-Point 16-bit)

BF16 utiliza 16 bits para representar un número en coma flotante, con una precisión similar a FP32. Este formato se ha utilizado en aplicaciones de aprendizaje automático.

4.1 Ventajas de BF16

- Ofrece una mayor precisión que FP16, lo que lo hace adecuado para ciertas aplicaciones de aprendizaje automático y redes neuronales.
- Usa menos memoria que FP32, lo que lo convierte en una opción eficiente para aplicaciones que requieren un equilibrio entre precisión y consumo de memoria.

4.2 Desventajas de BF16

- Aunque mejor que FP16, la precisión sigue siendo limitada para algunas aplicaciones que requieren alta precisión numérica.

5 Conclusiones

En conclusión, la elección del formato de coma flotante adecuado depende de las necesidades específicas de la aplicación. En general, a mayor número de bits, mayor es la precisión y menor es el error, pero también supone un peso computacional mayor. FP32 ofrece una buena precisión y rango, pero a costa de un mayor uso de memoria. FP16, por otro lado, reduce el uso de memoria a la mitad, pero con una precisión limitada, lo que lo hace adecuado para ciertas aplicaciones de aprendizaje profundo. BF16, con su equilibrio entre precisión y uso de memoria, se ha utilizado en aplicaciones de aprendizaje automático donde se requiere una buena precisión numérica sin incurrir en un uso excesivo de memoria.