Ejercicio 2: Cuantización

Jordi Blasco Lozano

Enunciado

Calcular la cuantización simétrica uniforme para un vector de entrada dado, convirtiendo valores de punto flotante (Float32) a enteros de 8 bits (INT8).

Datos iniciales:

- Entrada (valores en punto flotante): $x = [1{,}23, -0{,}87]$
- Número de bits: $n = 8$
- Los valores cuantizados son enteros: $q \in \mathbb{Z}$

Fórmulas utilizadas:

- Valor mínimo cuantizado: $q_{min} = -2^{n-1}$
- Valor máximo cuantizado: $q_{max} = 2^{n-1} - 1$
- Valor absoluto máximo del vector: $\alpha = \text{máx}(|x|)$
- Factor de escala: $s = \frac{\alpha}{q_{max}}$
- Cuantización: $q = \text{round}\left(\frac{x}{s}\right)$

1. Paso 1: Calcular el rango de valores cuantizados

Para una cuantización de $n = 8$ bits, calculamos el rango de valores enteros que podemos representar.

Cálculo de q_{min}:

$$q_{min} = -2^{n-1} \tag{1}$$
$$q_{min} = -2^{8-1} \tag{2}$$
$$q_{min} = -2^{7} \tag{3}$$

$$\boxed{q_{min} = -128}$$

Cálculo de q_{max}:

$$q_{max} = 2^{n-1} - 1 \tag{4}$$
$$q_{max} = 2^{8-1} - 1 \tag{5}$$
$$q_{max} = 2^{7} - 1 \tag{6}$$
$$q_{max} = 128 - 1 \tag{7}$$

$$\boxed{q_{max} = 127}$$

Interpretación: Con 8 bits en representación con signo (complemento a dos), podemos representar valores enteros desde -128 hasta 127, lo que nos da un total de $2^8 = 256$ valores posibles.

2. Paso 2: Calcular el valor absoluto máximo (α)

El valor α determina el rango dinámico de los datos originales. Tomamos el máximo de los valores absolutos del vector de entrada.

Cálculo de α:

$$\alpha = \text{máx}(|x|) \tag{8}$$
$$\alpha = \text{máx}(|x_1|, |x_2|) \tag{9}$$
$$\alpha = \text{máx}(|1{,}23|, |-0{,}87|) \tag{10}$$
$$\alpha = \text{máx}(1{,}23, 0{,}87) \tag{11}$$

$$\boxed{\alpha = 1{,}23}$$

Interpretación: El valor máximo absoluto en nuestros datos es 1,23. Este valor se mapeará al límite del rango cuantizado ($q_{max} = 127$), y su negativo ($-1{,}23$) se mapearía aproximadamente a q_{min}.

3. Paso 3: Calcular el factor de escala (s)

El factor de escala s relaciona los valores originales en punto flotante con los valores cuantizados enteros.

Cálculo de s:

$$s = \frac{\alpha}{q_{max}} \tag{12}$$
$$s = \frac{1{,}23}{127} \tag{13}$$
$$s = 0{,}009685... \tag{14}$$

$$\boxed{s \approx 0{,}00969}$$

Interpretación: Cada unidad en el espacio cuantizado representa aproximadamente 0,00969 en el espacio original. Este es el "paso.º resolución de nuestra cuantización.

4. Paso 4: Cuantizar cada valor del vector

Ahora aplicamos la fórmula de cuantización a cada elemento del vector x.

4.1. Cuantización de $x_1 = 1{,}23$

Paso 4.1: Calcular q_1

$$q_1 = \text{round}\left(\frac{x_1}{s}\right) \tag{15}$$
$$q_1 = \text{round}\left(\frac{1{,}23}{0{,}00969}\right) \tag{16}$$
$$q_1 = \text{round}(126{,}93...) \tag{17}$$
$$q_1 = \text{round}(127) \tag{18}$$

$$\boxed{q_1 = 127}$$

Verificación: El valor 1,23 es el máximo absoluto (α), por lo que se mapea exactamente a $q_{max} = 127$, como era de esperar en la cuantización simétrica.

4.2. Cuantización de $x_2 = -0{,}87$

Paso 4.2: Calcular q_2

$$q_2 = \text{round}\left(\frac{x_2}{s}\right) \tag{19}$$

$$q_2 = \text{round}\left(\frac{-0{,}87}{0{,}00969}\right) \tag{20}$$

$$q_2 = \text{round}(-89{,}78...) \tag{21}$$

$$q_2 = \text{round}(-90) \tag{22}$$

$$\boxed{q_2 = -90}$$

Verificación: El valor está dentro del rango válido $[-128, 127]$, por lo que no hay necesidad de recorte (clipping).

5. Paso 5: Vector cuantizado final

Vector cuantizado: $q = [127, -90]$

6. Paso 6: Decuantización (Reconstrucción)

Para verificar la precisión de la cuantización, podemos reconstruir los valores originales aproximados usando la operación inversa:

Fórmula de decuantización:

$$\hat{x} = q \cdot s$$

6.1. Reconstrucción de x_1

$$\hat{x}_1 = q_1 \cdot s \tag{23}$$

$$\hat{x}_1 = 127 \cdot 0{,}00969 \tag{24}$$

$$\hat{x}_1 = 1{,}2306 \tag{25}$$

$$\boxed{\hat{x}_1 \approx 1{,}23}$$

6.2. Reconstrucción de x_2

$$\hat{x}_2 = q_2 \cdot s \tag{26}$$

$$\hat{x}_2 = (-90) \cdot 0{,}00969 \tag{27}$$

$$\hat{x}_2 = -0{,}8721 \tag{28}$$

$$\boxed{\hat{x}_2 \approx -0{,}87}$$

7. Paso 7: Cálculo del Error de Cuantización

El error de cuantización mide la diferencia entre los valores originales y los reconstruidos.

Error absoluto para cada elemento:

$$\epsilon_1 = |x_1 - \hat{x}_1| = |1{,}23 - 1{,}2306| = 0{,}0006 \tag{29}$$

$$\epsilon_2 = |x_2 - \hat{x}_2| = |-0{,}87 - (-0{,}8721)| = 0{,}0021 \tag{30}$$

$$\boxed{\epsilon_1 = 0{,}0006, \quad \epsilon_2 = 0{,}0021}$$

Error cuadrático medio (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{31}$$

$$\text{MSE} = \frac{1}{2}\left[(0{,}0006)^2 + (0{,}0021)^2\right] \tag{32}$$

$$\text{MSE} = \frac{1}{2}\left[0{,}00000036 + 0{,}00000441\right] \tag{33}$$

$$\text{MSE} = \frac{0{,}00000477}{2} \tag{34}$$

$$\boxed{\text{MSE} \approx 2{,}39 \times 10^{-6}}$$

8. Resumen Final de Resultados

Parámetros de cuantización (8 bits):

- Rango cuantizado: $q \in [-128, 127]$
- Valor absoluto máximo: $\alpha = 1{,}23$
- Factor de escala: $s = 0{,}00969$

Valores originales y cuantizados:

Elemento	**Original (x)**	**Cuantizado (q)**	**Reconstruido ($\hat{x}$)**	**Error**
x_1	1.23	127	1.2306	0.0006
x_2	$-0{,}87$	-90	$-0{,}8721$	0.0021

Resultados finales:

- Vector original: $x = [1{,}23, -0{,}87]$
- Vector cuantizado: $q = [127, -90]$
- Error cuadrático medio: MSE $\approx 2{,}39 \times 10^{-6}$

Ventaja de la cuantización:
Reducción de memoria de $\frac{32 \text{ bits}}{8 \text{ bits}} = 4\times$ menos espacio de almacenamiento.