# Apuntes Semana 12

Apuntes del 23 de octubre

Juan Pablo Rodríguez Cano
*IC-6200 Inteligencia Artificial*
Tecnológico de Costa Rica
jp99@estudiantec.cr

***Abstract*—La cuantización en una técnica en redes neuronales para reducir el tamaño de los parámetros de los modelos, principalmente transformando los datos de punto flotante a enteros, lo cual además reduce el tiempo de computación de operaciones. Esta técnica es esencial para distribuir modelos en sistemas comerciales y ampliar la cantidad de plataformas que puedan correr estos modelos.**

***Index Terms*—cuantización, punto flotante, reducción de parámtetros.**

## I. Actividad de IEEE

Es un evento anual que se dará esta vez en noviembre en la sabana. Es una oportunidad para conocer sobre temas innovadores en inteligencia artificial y biología molecular. Es una oportunidad para crear contactos dentro de la industria ya que los presentadores suelen ser receptivos al público y disponen de tiempo para hablar.

## II. Quantization

Una vez que entrenado un modelo de redes neuronales, se debe colocar en un sistema para la distribución de este. Para esto existen varias técnicas, entre ellas, una opción común es utilizar el framework ONNX, que toma modelos escritos en diferentes lenguajes y bibliotecas y se crea una versión que maximiza la eficiencia de recursos y computación utilizando C++.

El mecanismo por el cual se disminuye es la cuantización y se enfoca en el hecho que los parámetros de los modelos son representados con tipos de datos de punto flotante, se reducen para hacer los modelos más densos con técnicas especiales para no afectar mucho la precisión de la inferencia. Aunque no es posible no introducir error, es necesario asumir esta desventaja para desplegar los modelos.

LLaMA 2 es un modelo muy popular y notorio por tener un tamaño muy grande, tiene 70 mil millones de parámetros, cada uno está representado por un punto flotante de 32 bits, lo que resulta en 28GB que deberían estar en memoria si se quisiera utilizar en una máquina local. Esto claramente no es viable porque la mayoría de máquinas comerciales cuentan con una capacidad menor a eso. Además, las operaciones que se se hacen con datos de punto flotante son muy lentas en comparación a datos representados por enteros. La cuantización hace una reducción de los bits requeridos para representar cada parámetro y lo convierte a enteros, que se pueden representar en las siguientes configuraciones: 8, 5, 2 y hasta 1 bit. La cuantización resulta en un menor tiempo de inferencia y menor consumo de energía, además de facilitar la opción de correr estos modelos en sistemas pequeños como dispositivos móviles o sistemas embebidos.

### A. Representación de números

Se suelen utilizar números en bloques de 8 bits para los enteros, para representar números negativos se utiliza el complemento a2 en los computadores. En contraste, para los punto flotantes se utiliza el ieee-754, cuyo tamaño de representación es de 32 bits, se utiliza la siguiente fórmula.

$$v = (-1)^{sign} \times 2^{E-127} \times (1 + \sum_{i=1}^{23} b_{23-i} 2^{-i})$$

Para no perder tanta información se tiene el siguiente mecanismo:

1) Antes de que las entradas lleguen a la siguiente capa se cuantizan los pesos
2) Estos pesos se limitan a ciertos rangos, dependiendo de la cantidad de bits de la cuantización. Lo que se quiere es que la distribución sea equivalente.
3) Se hacen las operaciones con los datos de tipo entero.
4) Al salir de la capa, se de-cuantizan los pesos para que las siguientes capas operen con números de punto flotante, sin ”saber” que fueron cuantizados.

## III. Tipos de cuantización

1) Asimétrica $\rightarrow$ el valor de 0 corresponde al valor menor y el máximo es el peso máximo
2) Simétrica $\rightarrow$ el cero es el peso 0, el valor absoluto máximo de los pesos se mapea a un extremo, si es negativo se mapea al valor más negativo dentro de los valores posibles con los bits

### A. Cuantización Asimétrica

$x_q = clamp(\frac{x_f}{s} + z; 0; 2^n - 1)$
$x_f = valor flotante$
$z = -1 \times \frac{\beta}{s}$
*s es el parámetro de escalado
$s = \frac{\alpha - \beta}{2^b - 1}$
$x_f = s(x_q - z)$ $\rightarrow$ permite volver al valor original con un grado de error

### B. Cuantización Simétrica

rango: $[-(2^n - 1), (2^n - 1)]$
$s = \frac{abs(\alpha)}{2^{n-1}-1}$
$x_f = sx_q$

## IV. Estrategias de selección del rango

- Cuantización Dinámica
  - Cálculo estadístico de cuál será el valor de esa capa
  - se utiliza en la etapa post-training quantization
- Post training quantization
  - hay que tratar los pesos atípicos porque puede confinar los demás pesos en un rango muy pequeño e introduce más error
  - Se puede utilizar el percentil en vez del min y max
  - agregamos observers que se encargan de hacer la estadísticas, calibran todas las salidas de la capa
  - se hace con los datos de prueba
- Quantization Aware Training (QAT)
  - insertar módulos irreales en la computación de grafo del modelo para similar el efector de cuantización durante el entrenamiento.
  - La función de perdida es usada para actualizar los pesos que constantemente sufren.

-