# Actividad extracurricular 06b

## Factoreo en Transformers

Isaac González

**Metodos Númericos**

---

Los **Transformers** son una arquitectura de redes neuronales ampliamente utilizada en tareas de procesamiento de lenguaje natural, visión por computadora y modelos generativos. Su funcionamiento se basa principalmente en operaciones matriciales de gran escala, especialmente en el mecanismo de *self-attention*.
Debido al alto costo computacional que implican estas operaciones, el factoreo de matrices se ha convertido en una técnica importante para mejorar la eficiencia de los modelos sin afectar significativamente su desempeño.

---

## ¿Qué es el factoreo de matrices?

El factoreo de matrices consiste en descomponer una matriz grande en el producto de dos o más matrices de menor dimensión. Matemáticamente, una matriz de pesos $ W \in \mathbb{R}^{m \times n} $ puede aproximarse como:

$W \approx A \cdot B $

donde:

- $ A \in \mathbb{R}^{m \times k} $
- $ B \in \mathbb{R}^{k \times n} $
- $ k \ll \min(m, n) $

Esta descomposición permite representar la misma transformación lineal utilizando menos parámetros.

---

## Uso del factoreo en la arquitectura Transformers

### Factoreo en las proyecciones lineales

En la arquitectura Transformer, las matrices asociadas a *Query*, *Key* y *Value* suelen ser de gran tamaño. El factoreo de estas matrices permite reducir el número de operaciones necesarias durante el cálculo de la atención, mejorando la eficiencia del modelo.

### Atención de bajo rango

El mecanismo de atención estándar presenta una complejidad cuadrática respecto al tamaño de la secuencia. Mediante aproximaciones de bajo rango, es posible disminuir esta complejidad, lo que resulta especialmente útil al trabajar con secuencias largas.

### Compresión del modelo

El factoreo de matrices también se emplea como técnica de compresión, ya que reduce el número de parámetros del modelo y facilita su ejecución en entornos con recursos computacionales limitados.

---

## Razones para aplicar factoreo de matrices en Transformers

1. **Reducción del costo computacional**
   Al reemplazar una multiplicación matricial grande por varias más pequeñas, se disminuye el número total de operaciones aritméticas.

2. **Menor uso de memoria**
   Al almacenar matrices de menor dimensión, se reduce el consumo de memoria RAM y VRAM, lo cual es crítico en modelos grandes.

3. **Mejor escalabilidad**
   El factoreo permite entrenar y ejecutar Transformers más profundos o con secuencias más largas sin un crecimiento excesivo del costo computacional.

---

## Ventajas del factoreo en Transformers

- Disminuye el número de parámetros del modelo
- Acelera el entrenamiento y la inferencia
- Reduce el consumo de memoria
- Facilita el despliegue en hardware limitado