Nombre: Camila Caicedo

Curso: GR1CC

Fecha de entrega: 28/01/2026

Actividad extracurricular 06b: Factoreo en Transformers

Indicaciones

- Investigue sobre el uso de factoreo de matrices en la arquitectura de redes neuronales *Transformers*.
- Indique las razones y las ventajas de realizar esta operación.

¿Cómo se usa el factoreo de matrices en la arquitectura Transformer?

En las redes neuronales tipo Transformer, el factoreo de matrices se utiliza principalmente para hacer más eficientes los cálculos que se realizan en capas como la atención (attention) y las capas lineales. Estas redes trabajan con matrices muy grandes, especialmente cuando manejan secuencias largas o modelos con millones o miles de millones de parámetros, por lo que reducir el costo computacional se vuelve fundamental.

Uno de los usos más comunes del factoreo es en la descomposición de matrices de peso, como ocurre en técnicas de bajo rango (low-rank factorization), donde una matriz grande se aproxima como el producto de dos matrices más pequeñas. Esto se puede ver en las matrices que generan las consultas, claves y valores (Q, K y V) dentro del mecanismo de atención.

- ### Ventajas de usar factoreo de matrices
  - **Disminuye el costo computacional y reduce número de parámetros.** Cuando se factoriza una matriz grande en dos más pequeñas, el modelo almacena menos pesos, disminuyendo el tamaño del modelo y el consumo de memoria. Esto facilita su uso en dispositivos con recursos limitados.
  - **Se mejora la escalabilidad del modelo.** El factoreo permite entrenar modelos más grandes sin que el costo computacional crezca de forma descontrolada.
  - **Velocidad.** Se aceleran los cálculos, especialmente en tareas con secuencias largas.

- **Facilita la implementación de modelos ligeros**. Es clave en variantes como Transformers comprimidos o eficientes.
- **Regularización implícita.** Las aproximaciones de bajo rango pueden actuar como una forma de regularización, esto ayuda a evitar el sobreajuste al limitar la complejidad de las transformaciones aprendidas por el modelo.

Fuentes de consulta:

- https://www.datacamp.com/es/tutorial/how-transformers-work
- https://blog.tenea.com/matrices-redes-neuronales/