LICDIA
Laboratorio de Investigación en Ciencia de Datos & Inteligencia Artificial

Línea de I+D - Modelos de lenguaje abiertos | Trabajo distinguido · Escuela NLP 2026

JORNADAS 2026 DE CIENCIA, TECNOLOGÍA Y EXTENSIÓN — DEPARTAMENTO DE CIENCIAS BÁSICAS

PEFT vs. FFT: ¿Entrenar menos parámetros es entrenar más rápido?

Malek Camilo[1] - Juan Manuel Fernández[1,2] - Marcelo Errecalde[3]

[1] LICDIA, Departamento de Ciencias Básicas, Universidad Nacional de Luján - [2] ITBA - [3] LIDIC, Universidad Nacional de San Luis

mcamilo@unlu.edu.ar - jmfernandez@{unlu, itba}.edu.ar - merreca@unsl.edu.ar

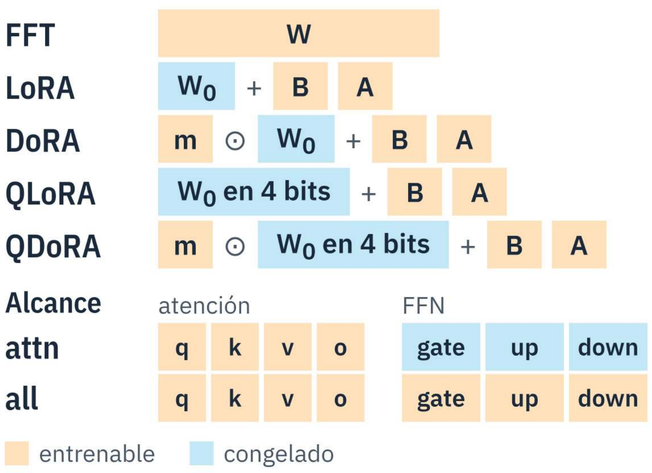
Entrenar menos parámetros no garantiza entrenar más rápido: DoRA r = 32 logra el mejor F1-Macro (0,657) con el 1,05 % de los parámetros, pero tarda 2,6 veces lo que el ajuste fino completo (FFT).

1 Problema

Adaptar modelos de lenguaje a un dominio con ajuste fino completo (FFT) es costoso. El ajuste fino eficiente en parámetros (PEFT) promete bajar ese costo. Este trabajo compara LoRA, DoRA, QLoRA y QDoRA frente a FFT sobre LLaMA 3.1 8B Instruct, y suma paralelismo de datos (DDP) para acelerar DoRA.

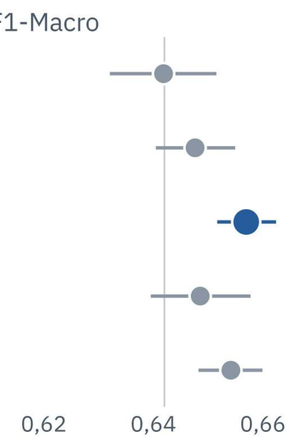
2 Enfoque

Clasificación generativa del tipo documental (6 clases) sobre 19.974 registros de Informática de SEDICI, con título y resumen como entrada. Son 31 configuraciones con 3 semillas: 93 entrenamientos y 771 h en Clementina XXI. Métrica principal: F1-Macro, por el fuerte desbalance del corpus (78,18 % de objetos de conferencia).



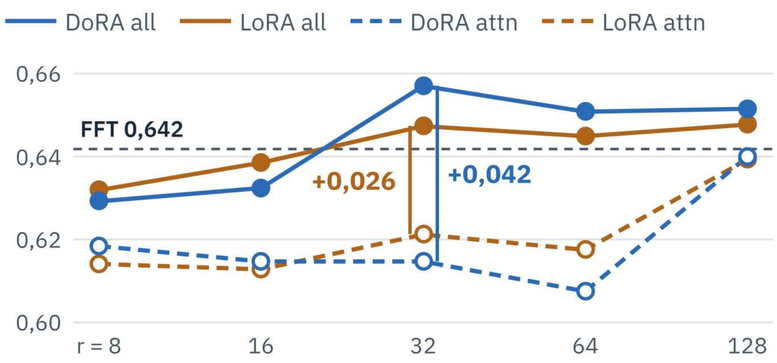
3 La mejor configuración de cada familia

	Entrenables	F1-Macro	Tiempo (h)	Memoria (GB)	Accuracy	Exact Match
FFT referencia	100 % 8.030 M	0,6418 ± 0,0097	6,50 ± 0,36	61,20 ± 0,00	0,8490 ± 0,0030	0,8418 ± 0,0046
LoRA r = 128	4,01 % 335,5 M	0,6477 ± 0,0073	6,17 ± 0,07	25,68 ± 0,00	**0,8595** ± 0,0064	**0,8592** ± 0,0065
DoRA r = 32	**1,05 %** 85,3 M	**0,6570** ± 0,0053	16,59 ± 0,07	22,86 ± 0,01	0,8507 ± 0,0014	0,8436 ± 0,0020
QLoRA r = 128	4,01 % 335,5 M	0,6486 ± 0,0091	**5,62** ± 0,04	**14,14** ± 0,00	0,8576 ± 0,0025	0,8572 ± 0,0025
QDoRA r = 128	4,03 % 336,9 M	0,6541 ± 0,0058	9,46 ± 0,03	**14,14** ± 0,00	0,8561 ± 0,0056	0,8529 ± 0,0055



El orden cambia según qué se mida. Todas las PEFT adaptan todas las capas lineales (all). Media ± desvío de 3 semillas en test (n = 2.997).

4 Alcance antes que rango



Adaptar todas las capas lineales rinde más que subir el rango: en attn, ir de r = 32 a 128 suma +0,018 y +0,025 con 4 veces más parámetros, apenas debajo de FFT.

5 DoRA en varias GPU

	F1-Macro	Tiempo (h) y aceleración	
1 GPU sin DDP	**0,657** ± 0,005	16,59 ± 0,07	×1
4 GPU 1 nodo	0,640 ± 0,006	4,42 ± 0,01	×3,8
8 GPU 2 nodos	0,631 ± 0,017	**2,52** ± 0,24	**×6,6**

DoRA r = 32, all. Trazo: tiempo ideal (1 GPU / n.º de GPU).

El F1-Macro baja porque, con 8 ejemplos por GPU, el lote global pasa de 8 a 32 y 64, y el modelo realiza 4 y 8 veces menos actualizaciones, respectivamente.

6 En síntesis: eficiencia en parámetros ≠ eficiencia en tiempo

Ningún método gana en todo

CALIDAD
DoRA all, r = 32
Mejor F1-Macro (0,657), aunque es el más lento: 16,6 h.

MEMORIA
QDoRA all, r = 128
F1-Macro 0,654 con 14,1 GB: casi toda la ganancia de DoRA.

MEMORIA Y TIEMPO
QLoRA all, r = 128
5,6 h y 14,1 GB, con F1-Macro 0,649.

SIMPLICIDAD
LoRA all, r = 32
F1-Macro 0,647 en 6,0 h; subir a r = 128 casi no suma.

Sigue: recuperar la calidad con DDP (fijar el lote global en 8 o escalar la tasa de aprendizaje), extender a otras técnicas y a 4 o más nodos, y medir el olvido catastrófico al adaptar a varios dominios.

Referencias

1. Camilo et al. Advanced PEFT methods in domain-specific LLM adaptation. Enviado a JCST.
2. Hu et al. LoRA: Low-rank adaptation of large language models. ICLR 2022.
3. Liu et al. DoRA: Weight-decomposed low-rank adaptation. ICML 2024.
4. Dettmers et al. QLoRA: Efficient finetuning of quantized LLMs. NeurIPS 2023.
5. Grattafiori et al. The Llama 3 herd of models. arXiv 2024.
6. Li et al. PyTorch Distributed: Experiences on accelerating data parallel training. VLDB 2020.

Agradecimientos

Los autores agradecen al Repositorio Institucional de la UNLP (SEDICI) por brindar acceso a los datos académicos, y al Sistema Nacional de Computación de Alto Desempeño (SNCAD), a Clementina XXI (IPAC, proyecto PCI-146), a la SICyT y a su personal técnico por su apoyo. Parte de la tesis de maestría Estrategias de adaptación eficiente para modelos de lenguaje abiertos (UNLP, 2026).

Código y datos | Modelos | Demo

Líneas de I+D - Investigación aplicada en ciencia de datos e IA, con fuerte impronta en los problemas regionales

Modelos de lenguaje abiertos · IA en educación · Diálogo y RAG · Gobernanza y transformación digital · Blockchain y credenciales · Imágenes y datos geoespaciales

licdia.unlu.edu.ar