Creación de recursos

Resumen ejecutivo

- **Modelos de IA**: Incluyen grandes modelos lingüísticos (LLM) para texto (GPT, BERT, etc.), modelos de visión (CNN, Diffusion) y modelos multimodales que combinan texto e imagen. Los *modelos fundacionales* son gigantescas redes pre-entrenadas en datos masivos que luego se adaptan a tareas específicas (mediante fine-tuning).
- **Plataformas**: Hay entornos colaborativos en la nube y herramientas locales. Ejemplos destacados son Hugging Face (repositorio de 2M modelos, demos y Spaces de IA), LangChain (framework de encadenado de LLMs con fuentes de datos), Google Colab y Kaggle Notebooks (Jupyter en la nube con GPU/TPU gratuitas), servicios ML empresariales como AWS SageMaker o Azure ML (notebooks integrados y despliegue en producción), y aplicaciones de IA local como Ollama o LM Studio (herramientas gratuitas para correr LLMs open-source en PCs). La librería **vLLM** destaca como motor optimizado para servir LLMs con alta eficiencia de memoria y rendimiento.
- **Librerías**: Las más usadas incluyen **PyTorch** (“biblioteca optimizada de tensores para DL en GPU/CPU”), **TensorFlow** (plataforma de Google para crear modelos ML ejecutables en cualquier entorno) y **Scikit-learn** (biblioteca Python de ML tradicional para regresión, clasificación, clustering, etc. de código abierto).
- **Técnicas clave**: *RAG* (Generación Aumentada con Recuperación) combina LLMs con búsquedas en bases de datos externas para anclar las respuestas en información actualizada. El *fine-tuning* es reentrenar parcialmente un modelo pre-entrenado en datos específicos. *Prompt engineering* consiste en diseñar con cuidado las instrucciones al modelo. Métodos recientes como **LoRA** (Low-Rank Adaptation) añaden matrices de bajo rango a un LLM para adaptarlo con muy pocos parámetros (ganando eficiencia de memoria). En general, las aproximaciones **PEFT** (“Parameter-Efficient Fine-Tuning”) ajustan solo unos pocos parámetros extra del modelo, congelando el resto, lo que reduce drásticamente los costos de cómputo y almacenamiento.
- **Infraestructura y hardware**: Hay aceleradores especializados para IA. **GPUs** (p.ej. NVIDIA H100/A100) son procesadores paralelos versátiles con gran memoria (80–141 GB) y fuerte ecosistema (CUDA, PyTorch). **TPUs** (Google TPU v6e, etc.) son ASICs de tensor optimizados para cargas de inferencia en nube: p.ej. v6e ofrece ~2 PFLOPS FP16 por ~$2.70/h, pero solo están en Google Cloud (Google Drive, Colab). **NPUs** (Unidades de Procesamiento Neural) integradas en SoCs de móviles/PCs (por ejemplo, el *Apple Neural Engine* en iPhones/Mac) ofrecen muy alta eficiencia energética: la 5ª generación de Apple

NPU alcanza ~15.8 TFLOPS FP16. **FPGAs** y **ASICs** (como Google Edge TPU) permiten IA embebida con baja latencia, aunque son menos flexibles y a menudo caros. Los aceleradores neuromórficos (Intel Loihi, IBM TrueNorth) son experimentales para IA inspirada en el cerebro. Cada hardware tiene su costo/rendimiento: las GPUs de consumo cuestan cientos de euros, TPUs cobran por hora en nube, NPUs vienen “incluidas” en dispositivos móviles, FPGAs/ASICs son especializados (educativamente poco comunes), etc.

- **Ejecución de modelos**: Se puede en **la nube**, **en local** o en **el edge**. La nube (Colab, AWS, Azure) ofrece escalabilidad y cero instalación, pero implica costos y dependencia de Internet. En local (PC o servidores propios) hay más control de datos y cero latencia de red, pero requiere disponer de hardware adecuado. En el edge (móviles, IoT) se prioriza rapidez y privacidad (los datos no salen del dispositivo). Por ejemplo, la startup española Multiverse destaca que ejecutar IA en el dispositivo (“edge”) mejora la eficiencia energética y la privacidad.
- **Datos y conjuntos**: La calidad de los datos es crucial. Existen repositorios abiertos (Hugging Face Datasets, Kaggle Datasets, conjuntos académicos como ImageNet, UCI, NOAA, etc.) en múltiples disciplinas. Hay que asegurar la ética (consentimiento de datos, evitar sesgos, transparencia) y cumplir normativas de gobernanza (GDPR en Europa, futuras leyes de IA).

image.png

Revision #4
Created 2025-12-12 13:26:43 CET by Chefo Cariñena
Updated 2026-03-04 17:43:09 CET by Luis Hueso