ENTRENAMIENTO DEL SISTEMA

HARDWARE NECESARIO Y COSTE APROXIMADO DEL ENTRENAMIENTO

Entrenar hoy (en 2025) un sistema multimodal capaz de generar audiodescripciones en tiempo real pensado para integrarse en unas gafas para personas ciegas implica utilizar hardware actual de alto rendimiento.

Partiendo de la base de que partimos de un modelo multimodal ya existente (tipo CLIP+LLM o LLaVA-like) al que hacemos fine-tuning, trabajaríamos con un modelo compacto para correr en las gafas de aproximadamente **1.3 billones de parámetros** (elegimos estos números porque son suficientemente grandes como para generar lenguaje fluido y comprender imágenes, pero también suficientemente pequeño como para poder ejecutarse en un dispositivo edge, como podemos ver en modelos ya existentes similares como gpt-4o-mini).

Con esos números, el entrenamiento completo supondría alrededor de **$2{,}3\times10^{20}$ FLOPs**, lo que equivale a unas **200 horas de GPU A100** en un escenario ideal. Pero en un proyecto real se repite el proceso varias veces, así que lo razonable es asumir un factor ×3, llegando a unas **624 GPU·h.**

- **Precio por hora:** alrededor de **3 USD/h (A100 -> 1.79USD/h, H100 -> 2.99USD/h, B200 -> 4.99USD/h)** en proveedores cloud.

- **Coste total estimado:** entre **2000 y 3.000 USD**, dependiendo de cuántas iteraciones y pruebas se hagan. Aquí hemos tenido en cuenta también un margen de gastos por almacenamiento de datos, tráfico, etc.

- **Consumo energético:** aprox. **400 kWh** para todo el entrenamiento, teniendo en cuenta la eficiencia real del datacenter.

- **Huella de carbono:** en torno a **100 kg de CO_2e**, usando el mix eléctrico europeo.

DATASETS NECESARIOS PARA EL ENTRENAMIENTO

Para cubrir todas las capacidades como pueden ser descripción visual general, lectura de texto, manejo de escenas de personas ciegas y narración continua, es necesario combinar distintos repositorios de datos públicos.

– Datos principales existentes:

- **MS COCO Captions** (CC-BY 4.0): base para aprender a describir imágenes generales.
- **LAION-5B** (CC-BY / Apache): para pre-training multimodal amplio.
- **TextCaps** y **COCO-Text**: fundamentales para aprender a leer texto dentro de la imagen (carteles, números, señales).

- **VizWiz (Captions, VQA y Classification)**: datos reales capturados por personas ciegas, con imágenes imperfectas y necesidades reales.
- **Ego4D**: vídeo en primera persona para comprender movimiento, manos y cambios de escena.
  Cityscapes / Mapillary: escenas urbanas muy importantes para obstáculos y señales.
- **Datasets de audiodescripción de películas (LSMDC, MAD)**: útiles para aprender narrativa continua.
-

– Datos faltantes y que podría ser interesante crear:

- Anotaciones específicas de obstáculos, prioridad de información, texto relevante.
- Ejemplos de situaciones de seguridad donde es crítico no fallar.

La forma de obtenerlos sería a través de colaboraciones con asociaciones de personas ciegas, grabaciones consentidas y un proceso riguroso de anonimización.

EVALUACIÓN DEL MODELO Y CRITERIOS DE CALIDAD

Evaluar un sistema de audiodescripción no se limita a medir la calidad del texto generado. Requiere medir precisión, seguridad, utilidad real y rendimiento en el dispositivo final.

– Métricas automáticas:

- **CIDEr, SPICE, BLEU, METEOR**: para comparar la calidad de las descripciones con otros modelos.
- **OCR**: para verificar que el sistema lee correctamente números y textos importantes.
- **Seguridad**:
  - Recall de obstáculos críticos (bordillos, escaleras, tráfico).
  - Cero falsos negativos en escenarios de prueba críticos.
- **Rendimiento**: latencia

– Evaluación con usuarios ciegos:

- Pruebas guiadas (en interiores o entornos controlados).
- Tareas como identificar un autobús, orientarse, o encontrar un producto.
- Percepción de facilidad de uso y confianza.
- Cuestionarios estándar de utilidad.

– Criterios go/no-go:

Para poder avanzar a pruebas reales en la calle, el sistema debería cumplir:

- **CIDEr ≥ 90–100** en VizWiz-Captions, acercándose al estado del arte
- lectura fiable del texto en imágenes
- **99 % o más de recall** en detección de obstáculos importantes
- **0 fallos críticos**
- latencia muy baja y estable
- ≥80 % de usuarios valorando el sistema como útil (≥4/5) y un **SUS ≥ 80**.

REFERENCIAS UTILIZADAS

Precios de las GPUs:
https://lambdalabs.com/service/gpu-cloud

Cálculos de flops y horas de entrenamiento:
https://chatgpt.com/share/691b62aa-b83c-8012-ae45-c0cbe574c7bf

Datasets:

https://cocodataset.org/#captions-2015

https://laion.ai/blog/laion-5b/

https://textvqa.org/textcaps

https://bgshih.github.io/cocotext/

https://vizwiz.org/

https://www.cityscapes-dataset.com/

https://research.netflix.com/research-area/lsmdc