UNIVERSIDAD DE CHILE

Inteligencia Artificial Generativa

Let's talk about hype stuff

Valentin Barriere // Clemente Henriquez

Universidad de Chile – DCC

Diplomado de Postítulo en Inteligencia Artificial, Primavera 2025

Laboratorio 5

Laboratorio 5: Large Multimodal Models

En este laboratorio, exploraremos las capacidades y limitaciones de modelos multimodales modernos.

- Trabajaremos con **SmolVLM-500M** (2024) y **BLIP-2-OPT-2.7B** (2023).
- Experimentaremos con **image captioning** y **Visual Question Answering (VQA)**.
- Identificaremos límites sistemáticos: conteo, texto, razonamiento espacial.

El Desafío: Entender Imágenes con IA

Nuestro Objetivo

Evaluar qué tan bien los modelos multimodales pueden **describir imágenes** y **responder preguntas** sobre ellas.

El Plan del Laboratorio

1. **Warm-up:** Captioning básico con SmolVLM para familiarizarnos con el modelo.
2. **Prompting Dirigido:** Experimentar con diferentes instrucciones para la misma imagen.
3. **Visual Question Answering (VQA):** Hacer preguntas específicas sobre imágenes.
4. **Identificar Límites:** Probar casos difíciles (conteo, texto en imágenes, razonamiento espacial).
5. **Comparación:** SmolVLM (2024) vs BLIP-2 (2023) - ¿ha mejorado la tecnología?

Qué vas a hacer? (Los Ejercicios)

Tu rol será explorar y evaluar sistemáticamente las capacidades de estos modelos.

- **Ejercicio 1-2 (Prompting):**
  - Completarás prompts para experimentar con diferentes tipos de instrucciones.
  - Observarás cómo cambia la respuesta del modelo según el prompt.
- **Ejercicio 3-4 (VQA y Límites):**
  - Escribirás preguntas factuales, inferenciales y de razonamiento.
  - Probarás casos problemáticos: conteo de objetos, lectura de texto, razonamiento espacial.
- **Ejercicio 5 (Comparación):**
  - Compararás las respuestas de SmolVLM y BLIP-2 en los mismos casos.
  - Evaluarás si el modelo más nuevo realmente es mejor.

Resultados Esperados

Los modelos multimodales tienen fortalezas claras pero también limitaciones sistemáticas.

- **Fortalezas:** Descripciones generales coherentes, detección de objetos principales.
- **Limitaciones:**
  - Conteo impreciso (¿cuántas frutas hay?)
  - Lectura de texto problemática (OCR limitado)
  - Razonamiento espacial débil (¿qué está a la izquierda?)
- **Comparación:** SmolVLM más moderno pero no necesariamente mejor en todo.

Questions?

References i