Propuesta Final de Proyecto

Centro de Investigación en Matemáticas Unidad Monterrey
Materia: Procesamiento de Texto e Imágenes con Deep Learning

1. Título Tentativo del Proyecto

Modelos Multimodales para Visual Question Answering en Imágenes Histopatológicas

2. Resumen General

Este proyecto propone el desarrollo de un sistema multimodal avanzado de **Visual Question Answering (VQA)** aplicado a imágenes histopatológicas, utilizando exclusivamente el dataset **PathVQA** y las métricas recomendadas por el profesor (BLEU y CIDEr), complementadas con métricas adicionales necesarias para preguntas binarias (Accuracy y F1-score).

El trabajo se inspira en la metodología del paper compartido por el profesor, *Evaluating Low-Cost Multimodal (Visual and Textual) Language Models for Automated Image Understanding in Computational Pathology*, pero propone una mejora sustancial mediante el uso de modelos de visión-lenguaje modernos, específicamente **LLaVA 1.5 (SigLIP + LLaMA-3) con LoRA**.

Esta versión elimina por completo las etapas de OCR y generación manual de datasets de la propuesta original, reduciendo significativamente la carga de trabajo y enfocándose en el núcleo multimodal: **imagen + pregunta → respuesta generada**.

3. 1. Objetivo General

Desarrollar, implementar y evaluar un sistema moderno de **Visual Question Answering** basado en modelos multimodales recientes (LLaVA 1.5), utilizando el dataset PathVQA para producir respuestas clínicas plausibles ante preguntas derivadas de imágenes histopatológicas.

4. 2. Descripción del Problema

El problema de VQA consiste en recibir como entrada:

- Una **imagen histopatológica**, y
- Una **pregunta en lenguaje natural**

y generar una **respuesta precisa**, ya sea binaria (Sí/No), categórica o descriptiva.

El dataset PathVQA contiene preguntas de distintos tipos: Sí/No, identificación anatómica, descripción de hallazgos patológicos, conteo de elementos, etc.

Ejemplo simulado:

```
Imagen: corte histologico de rinon
Pregunta: "Se observan membranas basales engrosadas?"
Respuesta: "Si"
```

5. 3. Dataset: PathVQA

Fuente: HuggingFace → `flaviagiammarino/path-vqa`

Características principales:

- ∼5,000 imágenes histológicas.
- 32,799 pares **Pregunta–Respuesta**.
- Tipos de pregunta:
  - Sí/No
  - What / Where / How
  - Conteo
  - Identificación de estructuras anatómicas
- División estándar: train/validation/test.

El dataset es **ideal para VQA multimodal** y está listo para usar sin preprocesamiento adicional.

Ejemplos del dataset:

```
Pregunta: "where are liver stem cells (oval cells) located?"
Respuesta: "in the canals of hering"
```

```
Pregunta: "is embolus derived from a lower-extremity venous thrombus lodged in
    a pulmonary.."
Respuesta: "yes"
```

6. 4. Metodología

6.1. 4.1. Modelo principal propuesto: LLaVA 1.5

El modelo seleccionado para este proyecto es **LLaVA 1.5**, una arquitectura moderna de visión-lenguaje basada en:

- **SigLIP**: encoder visual SOTA.
- **MLP multimodal**: proyector imagen→lenguaje.
- **LLaMA-3**: modelo de lenguaje de alto desempeño.
- **LoRA**: técnica de fine-tuning eficiente.

El pipeline se resume como:

$$\text{Imagen} \rightarrow \text{SigLIP} \rightarrow \text{Proyector MLP} \rightarrow \text{LLaMA-3} \rightarrow \text{Respuesta}$$

Este modelo supera ampliamente la arquitectura del paper original (CLIP + GPT-2 + prefix tuning), la cual presenta limitaciones de compatibilidad, capacidad de razonamiento y desempeño.

6.2. 4.2. Baseline opcional

Se considerará, únicamente como referencia mínima:

- Encoder visual: **CLIP** o **OpenCLIP**
- Decoder: **GPT-2** o similar
- Fusión: MLP lineal

El baseline sirve únicamente para contrastar la mejora que aportan los modelos modernos.

7. 5. Métricas de Evaluación

El conjunto final de métricas es:

Preguntas tipo Sí/No

- **Accuracy**
- **F1-score** (macro)

Preguntas abiertas (What/Where/How)

- **BLEU**
- **CIDEr**
- (Opcional) ROUGE-L, BERTScore

Se evalúa por tipo de pregunta para evitar distorsiones (p.ej., BLEU no es adecuado para respuestas cortas).

8. 6. Diseño Experimental

Experimento 1 — Baseline (opcional)

CLIP + GPT-2 o CLIP + MLP. Establece un piso de desempeño mínimo.

Experimento 2 — Modelo Moderno

Implementación de **LLaVA 1.5 + LoRA**. Evaluación completa en PathVQA.

Experimento 3 — Comparación Final

- Baseline vs LLaVA 1.5
- Resultados segmentados por tipo de pregunta
- Análisis de errores

9. 7. Resultados Esperados

- Mejora clara en BLEU y CIDEr frente al baseline.
- Alto rendimiento en preguntas de tipo Sí/No (porcentaje de Accuracy).
- Razones interpretables basadas en la arquitectura moderna LLaVA.

10. 8. Conclusión

Esta propuesta presenta un diseño moderno, eficiente y viable para implementar un sistema multimodal de VQA sobre el dataset PathVQA, cumpliendo con las indicaciones del profesor y extendiendo la metodología del paper original mediante el uso de modelos recientes (LLaVA 1.5). El enfoque permite evaluar métricas textuales estándar, así como métricas propias de preguntas binarias, ofreciendo un análisis sólido y completo del desempeño de modelos multimodales actuales.