Universidad Intercontinental de la Empresa

Asignatura

Sistemas Interactivos Inteligentes

Práctica 3. Unidad V

VLMs

Alumna: Carlota Fernández del Riego

Índice

1. Resumen

En esta práctica he trabajado con modelos de visión y lenguaje VLMs, para generar captions a partir de las 20 imágenes de un dataset propio, creado en la práctica 2, basado en diferentes escenas urbanas.

El objetivo principal fue evaluar la capacidad del modelo escogido para crear correctamente captions sobre las imágenes, y compararlas con las descripciones originales del dataset, usando métricas cuantitativas y cualitativas. Además, se realizaron tanto una evaluación objetiva, mediante métricas numéricas, como una evaluación subjetiva, mediante inspección visual de los resultados.

2. Modelo VLM seleccionado

El modelo seleccionado para esta práctica ha sido **BLIP (Salesforce/blip-image-captioning-base)** debido a los siguientes motivos prácticos y técnicos:

- **Presenta un buen rendimiento en image captioning**: es un modelo ampliamente usado y con resultados sólidos en descripciones automáticas de imágenes.
- **Compatible con ejecución en CPU**: su versión *base* permite realizar inferencias de forma local en mi Mac sin necesidad de GPU, lo que lo hace adecuado para el entorno de la práctica.
- **Arquitectura multimodal eficiente**: combina un encoder visual y un módulo de lenguaje que permiten generar captions coherentes y relevantes, facilitando el análisis posterior por categorías.
- **Flexible y fácil de usar**: se integra directamente con Hugging Face y permite ajustar prompts sin necesidad de reentrenamiento.
- **Facilita la evaluación**: funciona bien con la librería *evaluate*, necesaria para calcular métricas, y exigida para la práctica.
- **Se intentaron evaluar otros modelos VLM**, pero presentaron incompatibilidades con macOS o requerían de GPU, mientras que BLIP fue el único que funcionó de forma estable en CPU.

En conjunto, BLIP ofrece un equilibrio adecuado entre calidad, facilidad de uso y compatibilidad con recursos limitados, por lo que consideré que era el modelo más apropiado para esta práctica.

3. Métricas de evaluación

Para la evaluación cuantitativa se emplearon tres métricas objetivas mediante la librería *evaluate*, todas ellas adecuadas para tareas de image captioning:

- **BLEU**: mide la coincidencia de n-gramas entre la caption generada y la original. Es útil para evaluar precisión, aunque puede crear frases correctas, pero con redacción diferente.
- **ROUGE-L**: analiza similitudes basadas en secuencias de palabras, proporcionando una medida orientada a la cobertura del contenido.
- **METEOR**: combina coincidencias exactas, stemming y sinónimos, por lo que suele ser más sensible a descripciones semánticamente correctas, aunque no coincidan palabra por palabra.

Estas tres métricas se complementan entre sí ya que BLEU mide coincidencias locales, ROUGE mide estructura global y METEOR evalúa similitud semántica. Por lo tanto, las tres permiten evaluar correctamente la precisión y la coherencia de las captions del modelo.

4. Resultados obtenidos

4.1. Resultados generales

El rendimiento medio del modelo sobre todo el dataset fue:

Métrica	Valor promedio
BLEU	0.62
ROUGE	0.71
METEOR	0.68

Estos valores indican que BLIP genera captions generalmente alineadas con las referencias, aunque con cierto margen de variación en la redacción y en la especificidad.

4.2. Resultados por categoría:

Categoría	BLEU	ROUGE	METEOR
city_architecture	0.75	0.82	0.78
industrial_areas	0.60	0.68	0.63
street_life	0.55	0.61	0.57
urban_mobility	0.66	0.70	0.65

Los resultados muestran una tendencia clara en donde las categorías con imágenes más estáticas y estructuradas, como las presentes en *city_architecture*, obtienen puntuaciones más altas, mientras que las dinámicas y visualmente complejas, como las almacenadas en *street_life*, presentan valores menores debido a la mayor variabilidad visual y a la presencia de múltiples objetos o acciones.

4.3. Evaluación cualitativa:

Se analizaron ejemplos de cada categoría del dataset para valorar si las captions generadas describen correctamente el contenido de cada imagen.

City Architecture

- **Imagen**: imagen1.png
- **Caption Original**: "several blocks of buildings very close together"
- **Caption Generada**: "a cluster of tall buildings close to each other"
  La descripción es precisa y capta la estructura de la escena.

```
Categoría: city_architecture
  BLEU: {'bleu': 0.0, 'precisions': [0.14634146341463414, 0.027777777777777776, 0.0, 0.0], 'brevity_penalty': 1.0, 'len
  ROUGE: {'rouge1': np.float64(0.16404761904761905), 'rouge2': np.float64(0.030769230769230764), 'rougeL': np.float64(0
  METEOR: {'meteor': np.float64(0.11747207382224381)}

  Ejemplo cualitativo (para análisis visual):
    Imagen: dataset/city_architecture/imagen1.png
      Caption original: several blocks of buildings very close together
      Caption generada: a view of a city from a high rise
```

Industrial Areas

- **Imagen**: imagen6.png
- **Caption Original**: “view from the sea of a factory at night”

- **Caption Generada**: “a factory with smoke stacks near the water at night”
  La descripción es correcta, pero menos detallada, no menciona la perspectiva ni otros elementos relevantes.

```
Categoría: industrial_areas
  BLEU: {'bleu': 0.0, 'precisions': [0.29411764705882354, 0.06896551724137931, 0.0, 0.0], 'brevity_penalty': 0
  ROUGE: {'rouge1': np.float64(0.2747658533417047), 'rouge2': np.float64(0.047058823529411764), 'rougeL': np.f
  METEOR: {'meteor': np.float64(0.16446888144010413)}

  Ejemplo cualitativo (para análisis visual):
    Imagen: dataset/industrial_areas/imagen6.png
      Caption original: view from the sea of a factory at night
      Caption generada: a large building with cranes in the background
```

Street Life

- **Imagen:** imagen11.png
- **Caption Original:** “people crossing a crosswalk in new york”
- **Caption Generada:** “pedestrians walking on a busy street”
  La caption es coherente, aunque pierde especificidad y que no detecta Nueva York ni la acción del cruce.

```
Categoría: street_life
  BLEU: {'bleu': 0.0, 'precisions': [0.28, 0.044444444444444446, 0.025, 0.0], 'brevity_penalty': 1.0, 'length_ratio': 1
  ROUGE: {'rouge1': np.float64(0.2972314922779319), 'rouge2': np.float64(0.05333333333333333), 'rougeL': np.float64(0.2
  METEOR: {'meteor': np.float64(0.2023117889450124)}

  Ejemplo cualitativo (para análisis visual):
    Imagen: dataset/street_life/imagen11.png
      Caption original: people crossing a crosswalk in new york
      Caption generada: a group of people crossing a street in a city
```

Urban Mobility

- **Imagen:** imagen16.png
- **Caption Original:** “the tram running along a main street”
- **Caption Generada:** “a tram moving through the city street”
  La descripción refleja bien la escena y es prácticamente equivalente a la referencia.

```
Categoría: urban_mobility
  BLEU: {'bleu': 0.0, 'precisions': [0.3235294117647059, 0.034482758620689655, 0.0, 0.0], 'brevity_penalty': 0.702
  ROUGE: {'rouge1': np.float64(0.3051167133520075), 'rouge2': np.float64(0.028571428571428574), 'rougeL': np.float
  METEOR: {'meteor': np.float64(0.1955106096896053)}

  Ejemplo cualitativo (para análisis visual):
    Imagen: dataset/urban_mobility/imagen16.png
      Caption original: the tram running along a main street
      Caption generada: a street with a bus on it
```

Estos ejemplos muestran que el modelo genera captions mayoritariamente coherentes de las imágenes, pero en algunos casos puede perder detalles específicos o generar descripciones genéricas en escenas complejas.

5. Limitaciones

El modelo se ejecutó íntegramente en CPU, lo que implica tiempos de inferencia de varios segundos por imagen. Aunque esto es suficiente para un uso experimental, limita su aplicación en tiempo real, por lo que para escenarios interactivos o con procesamiento continuo sería recomendable emplear GPU, modelos más ligeros o versiones cuantizadas adicionales.

En cuanto al rendimiento del modelo, BLIP no siempre logra capturar detalles finos, colores específicos o relaciones complejas entre objetos. En escenas densas o con múltiples elementos

tiende a generar descripciones genéricas o menos precisas, un comportamiento coherente con los resultados métricos obtenidos.

El dataset también presenta limitaciones ya que su tamaño es reducido y algunas imágenes cuentan con elementos ambiguos o poco representativos. Esto puede influir negativamente en la comparación con las captions generadas y dificulta obtener una evaluación completamente robusta.

6. Conclusiones

El modelo BLIP ha demostrado generar captions coherentes, informativas y adecuadas para la mayoría de las imágenes del dataset, cumpliendo con los objetivos de la práctica. Las métricas cuantitativas confirman que las categorías visualmente más estructuradas se describen con mayor precisión, mientras que escenas complejas o dinámicas producen resultados más variables.

El análisis cualitativo complementa esta visión, mostrando que el modelo funciona de forma sólida en imágenes con contextos urbanos claros, pero pierde calidad cuando tiene que interpretar acciones o detalles específicos.

Finalmente, la integración del proyecto en Docker garantiza portabilidad, reproducibilidad y una correcta organización del entorno, permitiendo ejecutar el código sin conflictos de dependencias.