



# Proyecto grupal

Grupo 2:

- Kevin Guerra Huamán
- Marco Joel Isidro
- Víctor David Silva

Docentes:

- Abraham Rodriguez
- Oksana Bokhonok



## Índice

<b>1. Introducción</b>	<b>2</b>
<b>2. Selección del modelo</b>	<b>2</b>
<b>3. Selección del conjunto de datos</b>	<b>4</b>
<b>4. EDA (Análisis exploratorio inicial)</b>	<b>4</b>
<b>5. Métricas</b>	<b>6</b>
<b>6. Entrenamiento del modelo</b>	<b>7</b>
6.1 Conjunto de entrenamiento	7
6.2 Hiperparámetros	7
6.3 Hardware	8
<b>7. Resultados</b>	<b>8</b>
7.1 Proceso de entrenamiento	8
7.2 Métricas	9
7.3 Mapas de atención	10
7.4 Pruebas	12
<b>8. Conclusiones</b>	<b>12</b>
<b>Referencias</b>	<b>13</b>

## 1. Introducción

El presente proyecto consiste en la utilización de un modelo multimodal para generación de captions (texto a partir de imágenes). Se presenta la selección del modelo, del dataset, un análisis exploratorio de los datos, las métricas a utilizar, el entrenamiento de modelo, sus resultados y las conclusiones finales.

## 2. Selección del modelo

Bootstrapping Language-Image Pre-training (BLIP) es un Vision-Language Pre-training (VLP) framework, el cual puede llevar a cabo una amplia gama de tareas multimodales. Introduce dos innovaciones principales:

1. Multimodal mixture of Encoder-Decoder (MED): Una arquitectura (Figura 1) preentrenada orientada a múltiples tareas y transfer learning flexible. Sus componentes permiten operar en varios modos:
  - Unimodal Encoder: codifica de manera separada el texto y las imágenes.
    - Image encoder: utiliza un Vision Transformer (ViT) para procesar las imágenes y representarlas como un conjunto de embeddings.
    - Text encoder: similar a BERT, procesa el texto de entrada agregando un token de [CLS] y genera embeddings textuales.
  - Image-grounded text encoder: agrega información visual al agregar una capa de cross-attention entre las capas de self-attention y de la red feed forward (FFN). Genera representaciones del texto en función de la imagen.
  - Image-grounded text decoder: reemplaza las capas de bi-directional self-attention por capas de causal self-attention. Usa un token de [Decode] como inicio de secuencia y usa otro token como end-of-sequence. Genera texto a partir de la combinación de embeddings visuales y textuales.
2. Captioning and Filtering (CapFilt): un nuevo método de generación de datasets para aprender a partir de pares imagen-texto ruidosos. Está compuesto por:
  - Captioner (generador de captions): produce captions sintéticos dadas imágenes web.
  - Filter (filtro): remueve los captions ruidosos, tanto de el texto original de la web como el texto sintético.

Tres diferentes funcionalidades son computadas para determinar las pérdidas:

- Image-Text Contrastive Loss (ITC): alinea representaciones de imágenes y texto. Maximiza la similitud entre pares correctos (imagen-texto) y minimiza la similitud con pares incorrectos
- Image-Text Matching Loss (ITM): clasificar si una imagen y un texto están emparejados correctamente.
- Language Modeling Loss (LM): generar descripciones textuales a partir de imágenes.

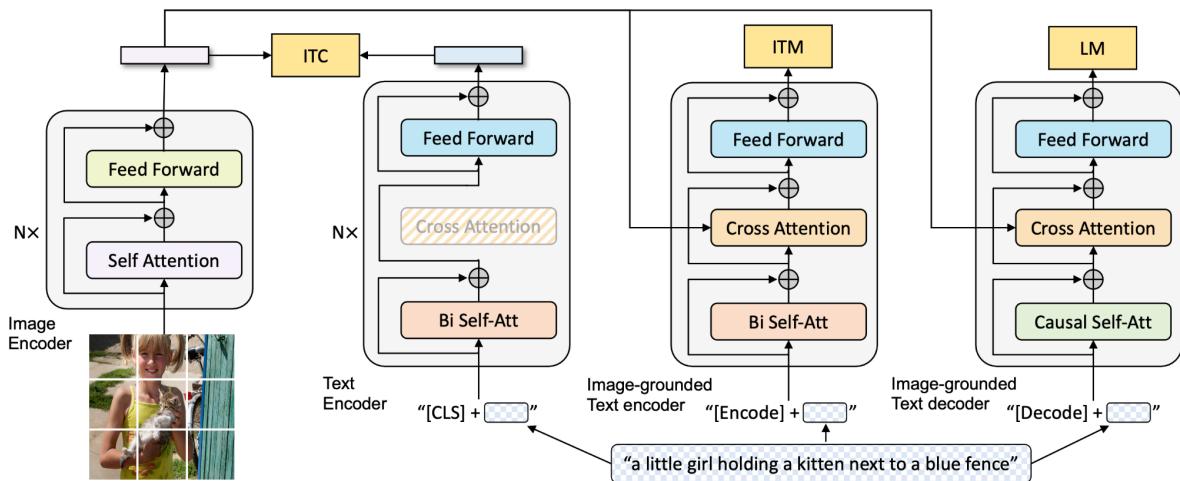


Figura 1. BLIP

Para el problema de image-captioning, se utiliza como modelo Blip-image-captioning-base. La arquitectura se reduce y queda como se observa en la Figura 2.

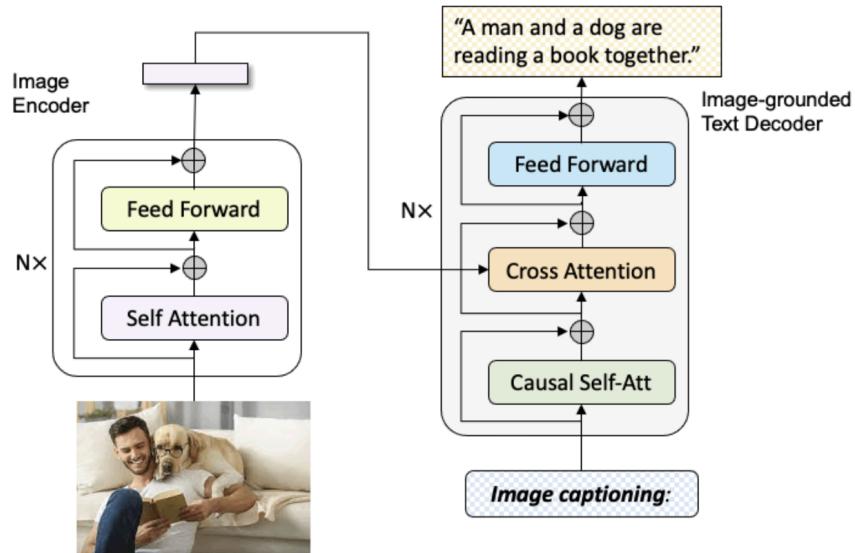


Figura 2. Blip-image-captioning

Las imágenes se dividen en patches que luego se transforman en vectores. La salida del ViT es un conjunto de embeddings que representan las características visuales de la imagen de entrada. Estos se pasan al Image-grounded Text Decoder, particularmente a su capa de Cross Attention, de esta manera el modelo se "fija" en partes específicas de la imagen relevantes para generar el texto, dependiendo del contexto actual de las palabras. La capa de Causal Self-Attention asegura que el modelo genere texto de manera secuencial, prediciendo cada palabra basándose en las palabras generadas previamente y las características visuales. Luego los embeddings pasan por redes feedforward que permiten al modelo capturar patrones no lineales en los datos. Este modelo se obtuvo de Hugging Face.

### 3. Selección del conjunto de datos

Se seleccionó el dataset Flickr30k, ampliamente utilizado en tareas de visión por computadora y procesamiento de lenguaje natural, como image captioning o retrieval multimodal. Este conjunto de datos, creado con fines de investigación avanzada en visión y lenguaje, toma su nombre de la plataforma de fotos Flickr, de la cual se obtuvieron las imágenes.

Flickr30k contiene 31,783 imágenes del mundo real, acompañadas de 158,915 descripciones en inglés (5 captions por imagen), anotadas por humanos. Las imágenes abarcan escenas cotidianas, eventos, personas, paisajes y relaciones entre objetos. Los captions incluyen descripciones detalladas de acciones, objetos y contexto, proporcionando información rica y contextual. El dataset fue obtenido de Kaggle.

### 4. EDA (Análisis exploratorio inicial)

Se realizó un análisis exploratorio inicial de los datos para entender con qué información se está trabajando y cómo usarla. En la Figura 3 se muestra una de las imágenes del dataset con sus dimensiones y canales.

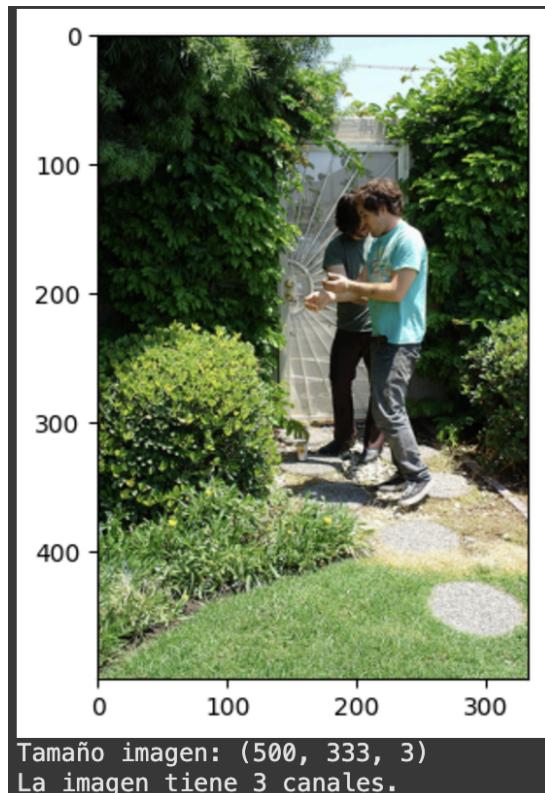


Figura 3. Imagen tomada del dataset y sus dimensiones

Al mostrar algunas filas del dataframe como se aprecia en la Figura 4, se observa que a cada imagen le corresponden 5 filas con diferentes comentarios cada una de ellas. Posteriormente se añadió un token de START al comienzo de cada comment y un token de END al final de cada uno.

```
image_name| comment_number| comment
1000092795.jpg| 0| Two young guys with shaggy hair look at their hands while hanging out in the yard .
1000092795.jpg| 1| Two young , White males are outside near many bushes .
1000092795.jpg| 2| Two men in green shirts are standing in a yard .
1000092795.jpg| 3| A man in a blue shirt standing in a garden .
1000092795.jpg| 4| Two friends enjoy time spent together .
10002456.jpg| 0| Several men in hard hats are operating a giant pulley system .
10002456.jpg| 1| Workers look down from up above on a piece of equipment .
10002456.jpg| 2| Two men working on a machine wearing hard hats .
10002456.jpg| 3| Four men on top of a tall structure .
10002456.jpg| 4| Three men on a large rig .
1000268201.jpg| 0| A child in a pink dress is climbing up a set of stairs in an entry way .
```

Figura 4. Primeras 10 filas del dataframe

Se muestra también la distribución de la longitud de los captions en caracteres en la Figura 5. De la misma manera, se muestra la distribución de palabras por caption, que siguen una distribución similar a la anterior, como se observa en la Figura 6.

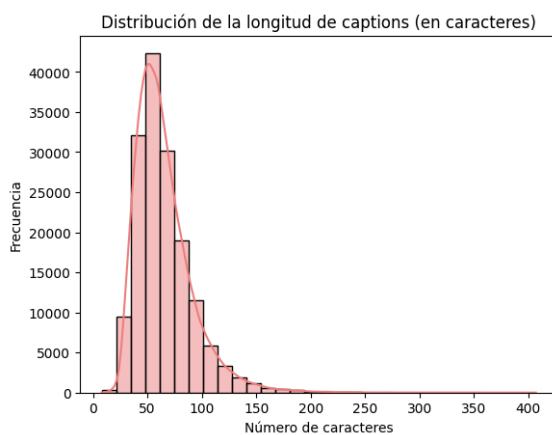


Figura 5. Distribución de la longitud de los captions

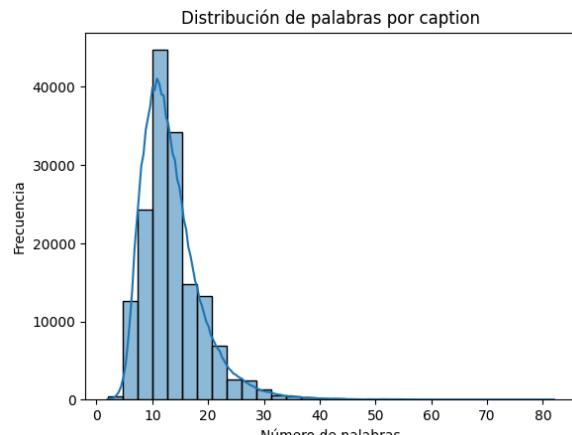


Figura 6. Distribución de palabras por caption

Se visualizan las distribuciones del tamaño de las imágenes en la Figura 7, donde se observa una gran variedad de tamaños.

Además se muestra la distribución de las palabras más usadas en los captions, como se observa en la Figura 8.

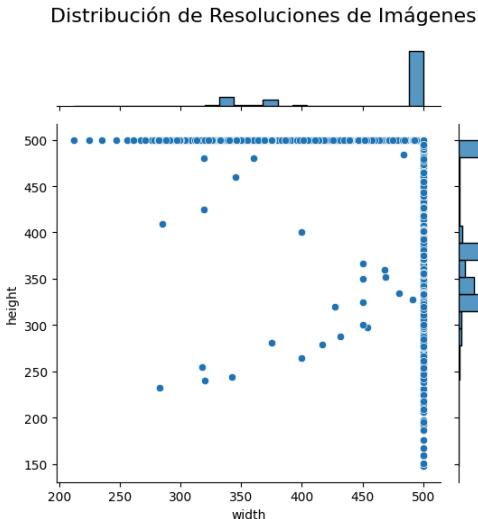


Figura 7. Distribución de resoluciones de imágenes

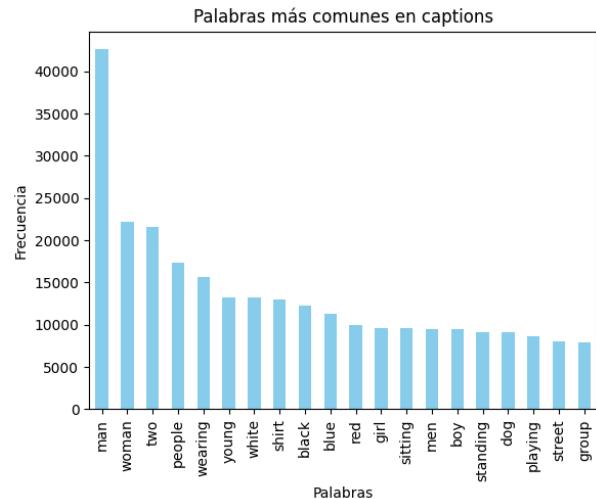


Figura 8. Palabras más comunes en captions

## 5. Métricas

En el proceso de entrenamiento y evaluación del modelo para la generación de captions, fue crucial utilizar métricas que nos permitan medir la calidad y precisión de las descripciones generadas en comparación con las descripciones humanas. Para esta tarea, hemos seleccionado dos métricas ampliamente utilizadas: BLEU (Bilingual Evaluation Understudy) y CIDEr (Consensus-based Image Description Evaluation). Dichas métricas, además de ser las utilizadas en el paper original, aportan la información necesaria para realizar una correcta evaluación del modelo entrenado en tareas de generación de captions. A continuación, explicamos en qué consisten ambas métricas y cómo nos ayudan a evaluar el rendimiento de nuestro modelo.

- BLEU: Es una de las métricas más conocidas y empleadas para evaluar la calidad de la traducción automática y la generación de texto, incluyendo las captions. Esta métrica se basa en la comparación de n-gramas (secuencias de palabras) entre las descripciones generadas por el modelo y las descripciones de referencia humanas. El objetivo de BLEU es medir qué tan similares son las frases generadas a las frases de referencia, tomando en cuenta coincidencias de n-gramas en diferentes niveles.
- CIDEr: Es una métrica diseñada específicamente para evaluar la calidad de las descripciones generadas en el contexto de imágenes. Tiene en cuenta la importancia semántica y la coherencia del contexto, en lugar de limitarse solo a las coincidencias de n-gramas. Utiliza una puntuación basada en el “consenso” de las palabras y frases utilizadas en las descripciones humanas, lo que permite medir qué tan bien el modelo captura los conceptos clave de la imagen.

BLEU y CIDEr son métricas que complementan la evaluación del modelo. BLEU mide qué tan similares son las captions generadas con las de referencia, evaluando la precisión y coherencia de las palabras. CIDEr, por su parte, se enfoca en qué tan bien el modelo entiende el contexto de la imagen y genera descripciones más relevantes y específicas. Juntas, nos dan una visión completa de la precisión y la calidad de las descripciones generadas.

## 6. Entrenamiento del modelo

El modelo BLIP utilizado cuenta con dos variantes: base y large, cada una con distintos niveles de complejidad y robustez, permitiendo adaptar el modelo a diferentes aplicaciones. En este trabajo, se optó por la variante base debido a su menor complejidad, ya que resulta adecuado para recursos computacionales limitados.

### 6.1 Conjunto de entrenamiento

Para la preparación del conjunto de datos, se realizó una reducción en su tamaño original, ya que el procesamiento completo de un dataset original con aproximadamente 30,000 elementos demandaría un tiempo considerable y requeriría mayores capacidades de memoria. Por lo tanto, se redujo el dataset en un 60%, resultando en un nuevo conjunto con alrededor de 12,000 muestras. Es importante destacar que el recorte del dataset se llevó a cabo de forma aleatoria, asegurando una representación uniforme de los datos.

La división final del conjunto de datos se realizó de la siguiente manera:

- 80 % de entrenamiento : 10170 muestras
- 10 % para validación : 1271 muestras
- 10 % para test : 1272 muestras

### 6.2 Hiperparámetros

Los hiperparámetros utilizados en el proceso de entrenamiento se detallan en la Tabla 1. Dentro de estos, se destaca la selección de una tasa de aprendizaje pequeña, lo que facilita un ajuste gradual de los pesos del modelo durante el entrenamiento. Asimismo, se estableció un valor de penalización de pesos (weight\_decay) mayor a 0.01 para abordar problemas de sobreajuste observados en etapas iniciales, ayudando a regularizar el modelo. En cuanto a los tamaños del batch, estos se definieron con el objetivo de equilibrar la eficiencia computacional y las limitaciones de memoria disponibles.

Este conjunto de configuraciones fue seleccionado para optimizar el rendimiento del modelo dentro de las restricciones computacionales y garantizar un proceso de entrenamiento estable y eficiente.

Tabla 1. Principales parámetros definidos

Tasa de aprendizaje	5e-5
Penalización para regularizar los pesos	0.1
Número de épocas	5
Tamaño de batch para entrenamiento	2
Tamaño de batch para evaluación	4
Precisión mixta 16 bits	True

### 6.3 Hardware

El entrenamiento del modelo se realizó mediante la plataforma “Google Colab Pro”, donde se dispone de equipos virtuales con diferentes capacidades. Para el trabajo se utilizó un equipo virtual con las siguientes características:

- GPU Tesla A100
- VRAM de 40 GB

## 7. Resultados

### 7.1 Proceso de entrenamiento

En la Figura 9 se observa la evolución del valor de pérdida en el proceso de entrenamiento y validación. El proceso se realiza con 5 épocas y parámetros definidos anteriormente. Para la gráfica de entrenamiento, la pérdida disminuye de 3.4 a 0.5, mostrando una convergencia estable.

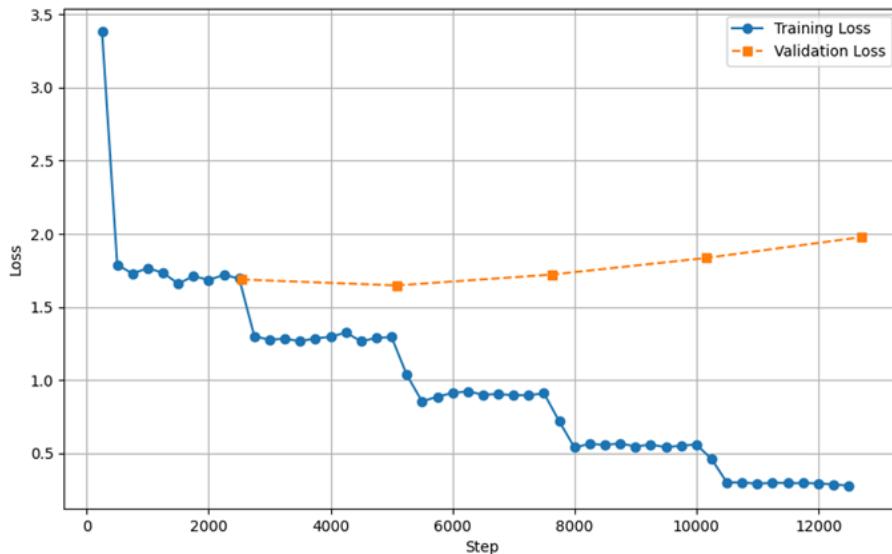


Figura 9. Evolución de la pérdida para dataset de entrenamiento y validación

La gráfica muestra un comportamiento esperado para el fine-tuning de un modelo preentrenado, evidenciando una reducción progresiva de la pérdida tanto en el conjunto de entrenamiento como en el de validación. Este último, presenta una disminución en la primera época, pero no alcanza una estabilización final, lo que sugiere que serían necesarias más épocas para observar un comportamiento más consistente. Cabe destacar que el entrenamiento se llevó a cabo sobre un conjunto de datos reducido, lo que podría influir en la convergencia del modelo y su capacidad de generalización.

## 7.2 Métricas

El modelo fue evaluado utilizando las métricas BLEU y CIDEr, calculadas al final de cada época para medir la calidad de las descripciones generadas en comparación con las etiquetas reales. Los resultados obtenidos se presentan en la Figura 10. La métrica BLEU muestra valores alrededor de 0.225 y para CIDEr valores de 1.17.

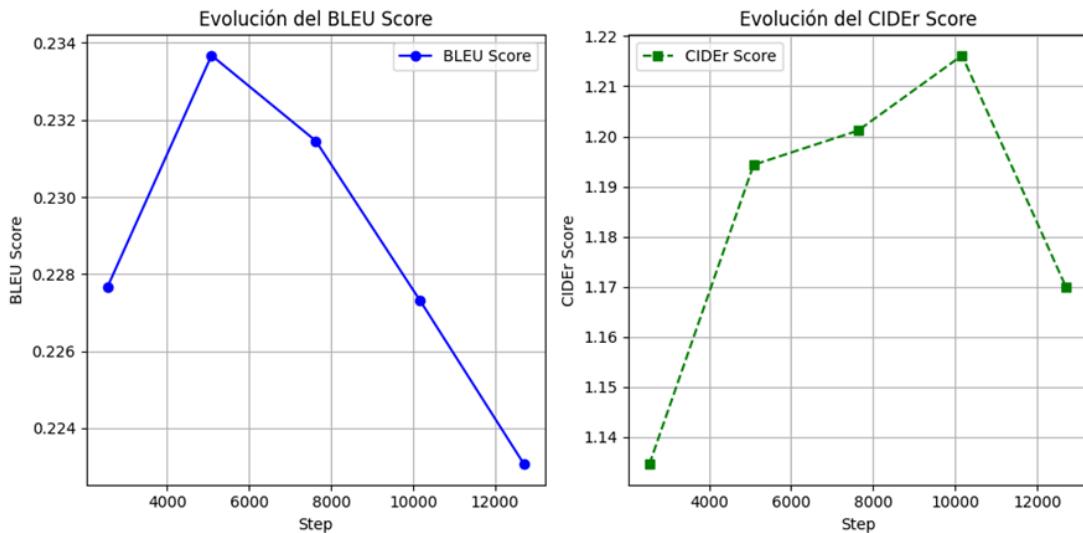


Figura 10. Evolución de las métricas de BLEU y CIDEr

La Figura 11 muestra una comparación de las métricas obtenidas entre los conjuntos de validación y test. Se observa que el modelo muestra un rendimiento ligeramente superior en el conjunto de prueba, lo que podría indicar una adecuada capacidad de generalización a datos no vistos.

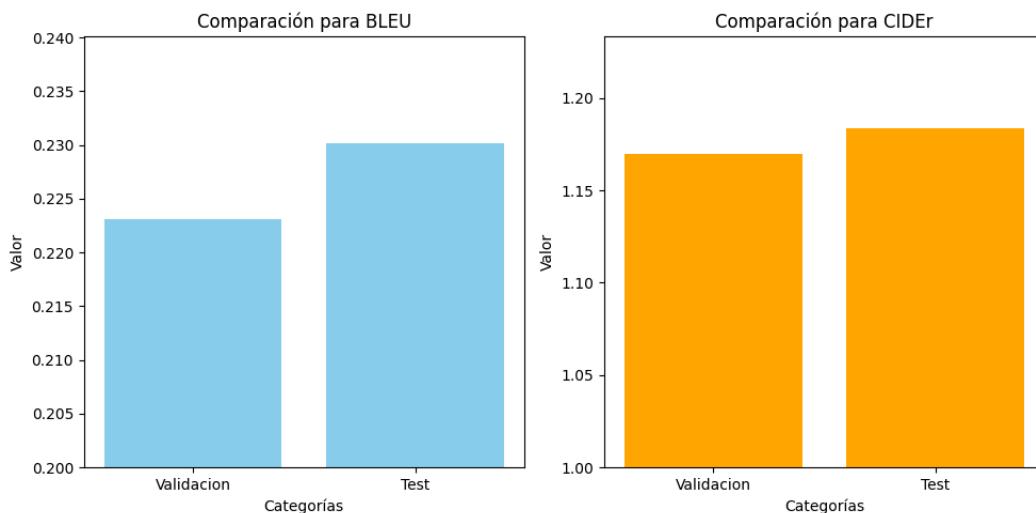


Figura 11. Comparación de las métricas de BLEU y CIDEr en validación y test

### 7.3 Mapas de atención

El modelo BLIP utiliza mapas de atención para focalizarse en diferentes regiones de las imágenes al generar descripciones, identificando las áreas más relevantes para la tarea. Como parte de la implementación, se han visualizado los mapas de atención correspondientes a varias muestras del dataset, proporcionando una representación gráfica del proceso de atención del modelo.

- Descripción original: “a little boy with an orange shirt is riding his blue and white toy car”.
- Predicción: “a photography of a young boy in an orange shirt and blue shorts playing with a radio flyer car on the floor of a house”.



Figura 12. Mapa de atención de imagen 1 del conjunto de test

- Descripción original: “A young female child looks at the photographer while an older woman sits and looks at the ground”.
- Predicción: “A photography of a young child and a young adult in a residential area next to a white van with a child on the back of it, standing next”.



Figura 13. Mapa de atención de imagen 2 del conjunto de test

El mapa de atención mostró un enfoque adecuado, donde corroboran que el modelo utiliza de forma efectiva la información visual para guiar la generación de texto.

#### 7.4 Pruebas

Tras obtener un modelo entrenado con resultados alentadores en la generación de descripciones a partir de imágenes, se llevaron a cabo pruebas adicionales utilizando un par de imágenes diferentes al dataset. A continuación, se presentan dos ejemplos ilustrativos.

Tabla 2. Resultados de inferencia en imágenes no vistas.

Imagen	Descripción
	a man with glasses and a beard is sitting at a desk with his laptop computer on his desk.
	a furry, white a brown dog stands with his hind legs raised.

El modelo demostró un nivel aceptable en las descripciones, con una estructura textual coherente e información relevante. Sin embargo, en algunos casos, las descripciones producidas fueron más generales.

### 8. Conclusiones

El proyecto implementó el modelo multimodal BLIP para la tarea de image-captioning a partir de imágenes, utilizando el dataset Flickr30k. Se realizó un análisis exploratorio de datos para comprender mejor el conjunto, reduciendo su tamaño para adaptarlo a los recursos computacionales disponibles. El modelo fue entrenado en Google Colab Pro con hiperparámetros optimizados y evaluado mediante las métricas BLEU y CIDEr. Finalmente, las visualizaciones de atención indicaron que el modelo utiliza las características visuales para generar el texto correspondiente.



## Referencias

[1] Repositorio de la materia Vision Transformers.

<https://github.com/FIUBA-Posgrado-Inteligencia-Artificial/CEIA-ViT>

[2] Paper de BLIP. <https://arxiv.org/abs/2201.12086>

[3] Modelo de BLIP en Hugging Face.

<https://huggingface.co/Salesforce/blip-image-captioning-large>

[4] Conjunto de datos de Flickr30.

<https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset/data>

[5] CIDEr: Consensus-based Image Description Evaluation, 2015,

<https://arxiv.org/abs/1411.5726>