

Universidad de La Habana  
Facultad de Matemática y Computación



## **Textos alternativos para las colecciones de imágenes del repositorio digital de la Oficina del historiador de La Habana**

Autores:

**Amanda Cordero Lezcano  
Ana Paula González Muñoz  
Carlos Antonio Bresó Sotto  
Christopher Guerra Herrero  
Dennis Daniel González Durán  
Marian Susana Alvarez Suri**

Proyecto final de la asignatura  
Aprendizaje de Máquinas

Enero 2025

# Resumen

Se presenta un sistema de generación de textos alternativos para imágenes del repositorio digital de la Oficina del Historiador de La Habana. Se exploraron modelos de aprendizaje profundo, incluyendo BLIP, Visual Transformers + GPT-2 y CLIP, para la generación. Se realizó un análisis detallado del estado del arte, identificando modelos de redes recurrentes, convolucionales y basados en Transformers. La metodología combinó la generación de descripciones con un algoritmo de selección basado en similitud semántica. Los resultados fueron evaluados mediante métricas BLEU, METEOR, CIDEr, ROUGE, y SPICE.

# Índice general

<b>1. Introducción</b>	<b>4</b>
<b>2. Estado del Arte</b>	<b>5</b>
2.1. Modelos Basados en Redes Recurrentes . . . . .	5
2.2. Modelos Basados en Redes Convolucionales . . . . .	6
2.3. Atención y Modelos Jerárquicos . . . . .	6
2.4. Transformers y Modelos Multimodales . . . . .	6
2.5. Resumen y Tendencias Actuales . . . . .	7
<b>3. Propuesta</b>	<b>9</b>
<b>4. Desarrollo de la propuesta</b>	<b>10</b>
4.1. Análisis exploratorio de los datos . . . . .	10
4.2. Detalles de los modelos propuestos . . . . .	10
4.2.1. BLIP . . . . .	10
4.2.2. Visual Transformers + GPT2 . . . . .	11
4.2.3. CLIP . . . . .	13
<b>5. Experimentación y resultados</b>	<b>16</b>
<b>6. Conclusiones</b>	<b>19</b>
<b>7. Apéndice</b>	<b>20</b>
<b>Referencias</b>	<b>25</b>

# Capítulo 1

## Introducción

En este proyecto, se desarrolla un sistema de generación de texto alternativo para las imágenes del repositorio digital de la Oficina del Historiador de la Ciudad de La Habana. Creando descripciones precisas y detalladas que mejoren la accesibilidad a la información, permitan una mejor búsqueda y recuperación de imágenes.

El problema de generar texto alternativo para imágenes ha sido abordado mediante el uso de técnicas avanzadas de procesamiento del lenguaje natural (NLP) y visión por computadora. Investigaciones recientes han utilizado redes neuronales convolucionales (CNN) para analizar y comprender el contenido visual de las imágenes, combinadas con redes neuronales recurrentes (RNN) o Transformers para generar descripciones textuales coherentes y contextualmente adecuadas. Estos modelos han demostrado ser efectivos en la tarea de generación de texto alternativo, permitiendo la creación automática de descripciones.

La propuesta de solución implica la combinación de dos modelos pre-entrenados de generación de texto alternativo, con los cuales se generan dos descripciones para cada imagen, y luego se aplica un algoritmo de selección que elige la descripción que mejor se ajusta a cada imagen específica. Esta metodología no solo aprovecha las capacidades avanzadas de cada modelo, sino que también asegura que la descripción final sea la más precisa y relevante.

# Capítulo 2

## Estado del Arte

El campo del *image captioning* ha evolucionado significativamente en la última década, impulsado por el avance de modelos de aprendizaje profundo. En este capítulo, se presentan los principales enfoques y modelos que han marcado hitos en esta área, destacando sus arquitecturas, metodologías y contribuciones.

### 2.1. Modelos Basados en Redes Recurrentes

Los primeros avances en generación de descripciones de imágenes se apoyaron en arquitecturas *encoder-decoder* con redes neuronales recurrentes (RNN). Uno de los primeros modelos destacados fue **Show and Tell** [1], que propuso un enfoque generativo basado en una arquitectura recurrente profunda. Este modelo utilizó una red convolucional (CNN) para la extracción de características visuales y una red LSTM para la generación de texto. Fue entrenado para maximizar la probabilidad de generar una descripción textual dada una imagen, aprendiendo únicamente a partir de descripciones de imágenes.

Posteriormente, **Show, Attend and Tell** [2] introdujo mecanismos de atención visual, permitiendo que el modelo enfocara diferentes regiones de la imagen en cada paso de generación. Este enfoque mejoró la calidad de las descripciones y presentó una formulación matemática más avanzada para el cálculo de la atención.

## 2.2. Modelos Basados en Redes Convolucionales

En un intento por superar las limitaciones de las RNN, **Convolutional Image Captioning** [3] propuso una arquitectura basada en CNNs para la generación de texto. Este modelo demostró que las CNNs pueden superar a las LSTM en tareas de *captioning*, especialmente cuando se combinan con mecanismos de atención, mitigando problemas como el desvanecimiento del gradiente. Se realizó un análisis detallado que proporcionó razones convincentes para preferir los enfoques de generación de lenguaje convolucional.

## 2.3. Atención y Modelos Jerárquicos

Otro enfoque relevante fue el modelo **Bottom-Up and Top-Down** [4], que implementó un mecanismo de atención jerárquico basado en la segmentación de objetos dentro de la imagen. Este modelo propuso una estrategia combinada de atención Bottom-Up y Top-Down, permitiendo un análisis más profundo de la imagen a través de múltiples pasos de razonamiento. En el mecanismo Bottom-Up, basado en Faster R-CNN, se proponen regiones de la imagen, cada una con un vector de características asociado. Por otro lado, el mecanismo Top-Down determina las ponderaciones de estas características, lo que permite una atención más precisa a nivel de objetos y regiones destacadas.

**Knowing When to Look** [5] propuso un mecanismo de atención adaptativo mediante un centinela visual, que decide cuándo prestar atención a la imagen y cuándo confiar en el contexto textual generado previamente. Este modelo abordó una limitación clave de los enfoques tradicionales de atención, que forzaban la atención visual en cada palabra generada, incluso cuando no era necesaria. En lugar de ello, el modelo introdujo un centinela visual que, en cada paso de tiempo, determina si es necesario extraer información de la imagen y, de ser así, selecciona las regiones relevantes. Esto permite al decodificador alternar de manera inteligente entre la información visual y el contexto lingüístico, mejorando la precisión y fluidez de las descripciones generadas.

## 2.4. Transformers y Modelos Multimodales

Con la llegada de los Transformers, los modelos de *captioning* han adoptado arquitecturas más avanzadas. **Multimodal Transformer** [6] exploró la representación visual multi-vista para mejorar la generación de texto, am-

pliendo el éxito del modelo Transformer en traducción automática a la tarea de subtítulos de imágenes. A diferencia de los enfoques tradicionales basados en codificador-decodificador, que utilizan una CNN para extraer características visuales y una RNN con mecanismos de atención para generar texto, este modelo propone un bloque de atención unificado que captura simultáneamente interacciones intra e intermodales. Esto permite un razonamiento multimodal más complejo, integrando tanto la autoatención (interacciones intramodales) como la coatención (interacciones intermodales) en una arquitectura modular y profunda.

**ViT** [7] aplicó Transformers directamente a imágenes dividiéndolas en secuencias de parches, lo que representó un cambio significativo en el procesamiento de información visual. A diferencia de los enfoques anteriores, que combinaban Transformers con redes convolucionales (CNNs) o reemplazaban solo ciertos componentes de las CNNs, ViT demostró que una arquitectura basada únicamente en Transformers puede lograr un rendimiento excepcional en tareas de clasificación de imágenes. Este enfoque elimina la dependencia de las CNNs, procesando las imágenes como secuencias de parches lineales y aplicando mecanismos de atención pura para capturar relaciones globales entre ellos.

**CLIP** [8] introdujo el aprendizaje multimodal a gran escala utilizando correspondencias entre imágenes y texto en internet. A diferencia de los sistemas de visión por computadora tradicionales, que se entrena para predecir un conjunto fijo de categorías de objetos, CLIP aprende representaciones visuales directamente a partir de texto sin procesar, lo que le permite capturar una gama más amplia de conceptos visuales. Este enfoque se basa en una tarea de preentrenamiento simple pero efectiva: predecir qué descripción de texto corresponde a una imagen dada.

**BLIP** [9] amplió la capacidad de los modelos al abordar tanto la generación como la comprensión de imágenes y videos, logrando mejoras significativas en una amplia gama de tareas de visión y lenguaje. A diferencia de los modelos preentrenados existentes, que suelen especializarse en tareas de comprensión o generación, BLIP propone un marco de preentrenamiento unificado que se transfiere de manera flexible a ambas.

## 2.5. Resumen y Tendencias Actuales

La evolución del *image captioning* ha pasado de modelos basados en RNNs con atención visual a enfoques más sofisticados que integran Transformers y aprendizaje multimodal. Modelos recientes como BLIP y CLIP han

demonstrado que la combinación de visión y lenguaje en grandes volúmenes de datos puede llevar a mejoras sustanciales en la generación y comprensión de imágenes.

Las tendencias actuales apuntan a modelos más eficientes y escalables, con capacidades mejoradas en la generación de texto y una mayor comprensión del contexto visual. El impacto de estas tecnologías se extiende más allá del *image captioning*, beneficiando tareas como búsqueda visual, generación de contenido y asistencia en accesibilidad.

# Capítulo 3

## Propuesta

Se propone un sistema para la generación de texto alternativo a partir de las imágenes del repositorio digital del patrimonio cultural de la Oficina del Historiador, combinando múltiples modelos de aprendizaje profundo. La metodología se basa en un enfoque de evaluación comparativa entre diferentes modelos de generación de texto a partir de imágenes, utilizando un criterio de selección basado en la similitud semántica con el contenido visual.

En primer lugar, se emplea el modelo BLIP (Bootstrapped Language-Image Pretraining) para generar una primera descripción de la imagen. Posteriormente, la imagen es procesada por el modelo ViT (Vision Transformer) en conjunto con GPT-2, obteniendo una segunda descripción independiente. Finalmente, el modelo CLIP (Contrastive Language-Image Pretraining) se utiliza como mecanismo de selección, comparando las dos descripciones generadas y eligiendo la que presente una mayor correspondencia semántica con la imagen de entrada.

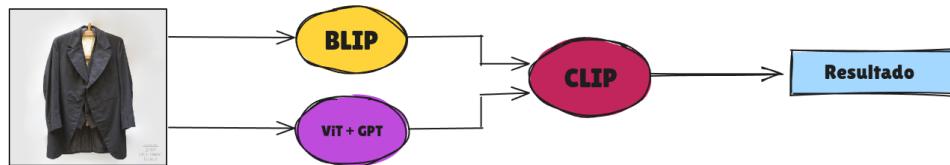


Figura 3.1: Modelo general

## Capítulo 4

# Desarrollo de la propuesta

A continuación se profundiza en los modelos que componen la propuesta.

### 4.1. Análisis exploratorio de los datos

Se realizó una evaluación de los metadatos proporcionados, determinando finalmente que no eran de utilidad para la generación de los textos alternativos de las imágenes. De igual forma, el conjunto de datos no contenía información sobre todas las imágenes a procesar y los datos existentes eran poco consistentes con las imágenes.

### 4.2. Detalles de los modelos propuestos

Para generar los textos alternativos de las imágenes, se dividió el desarrollo del modelo en tres componentes fundamentales.

#### 4.2.1. BLIP

El modelo **BLIP** [9] es un marco de pre-entrenamiento de visión-lenguaje diseñado para mejorar el rendimiento en diversas tareas de comprensión y generación. Utiliza una arquitectura de codificador-decodificador multimodal y un enfoque innovador llamado *Captioning and Filtering* (CapFilt) para mejorar la calidad de los datos.

#### Arquitectura del Modelo

BLIP se basa en la arquitectura de Mezcla Multimodal de Codificador-Decodificador (CDM), la cual opera en tres modos principales:

- **Codificador Unimodal:** Procesa separadamente las representaciones de texto e imagen.
- **Codificador de Texto con Base en Imágenes:** Introduce información visual en el texto mediante mecanismos de atención cruzada.
- **Decodificador de Texto con Base en Imágenes:** Genera descripciones textuales de las imágenes utilizando atención causal.

El modelo BLIP se entrena con tres objetivos clave: El aprendizaje contrastivo imagen-texto (CIT) que alinea representaciones visuales y lingüísticas; las coincidencia imagen-texto (ITM), encargadas de clasificar si un par imagen-texto coincide o no; y el modelado del lenguaje, el cual genera texto condicionado a una imagen.

### Entrenamiento del Modelo

El entrenamiento de BLIP consta de tres fases principales:

**Pre-entrenamiento:** En esta fase, el modelo se entrena en grandes conjuntos de datos que contienen pares de imagen-texto. Se utilizan algunos modelos para la codificación de imágenes y para la codificación de texto.

**Bootstrapping de Datos con CapFilt:** CapFilt aborda el problema del ruido en los datos extraídos de la web mediante los módulos Captioner y Filter, los cuales generan descripciones sintéticas a partir de imágenes web y filtran los textos ruidosos utilizando el clasificador ITM, respectivamente.

**Ajuste Fino (Fine-tuning):** Lo anterior permite que BLIP se ajuste para tareas específicas como la generación de subtítulos (usando el objetivo LM), la recuperación imagen - texto (optimizando CIT e ITM) y la respuesta a preguntas visuales (ajuste basado en el objetivo LM).

### 4.2.2. Visual Transformers + GPT2

El modelo **Vision Transformer (ViT)** [7] está basado en la arquitectura Transformer, ampliamente utilizada en procesamiento de lenguaje natural (NLP), adaptado para tareas de visión por computadora. A diferencia de las redes neuronales convolucionales (CNNs), ViT procesa imágenes como secuencias de parches, permitiendo capturar relaciones espaciales mediante mecanismos de auto-atención.

### Arquitectura del Modelo

El modelo ViT se basa en los siguientes componentes fundamentales:

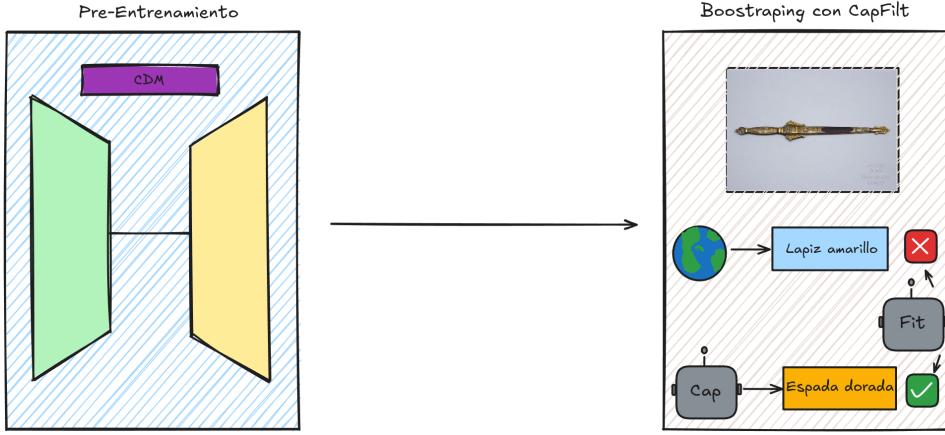


Figura 4.1: BLIP

**División en Parches (Patches)** La imagen de entrada se divide en pequeños parches de tamaño fijo  $p \times p$ . Cada parche actúa como una unidad similar a un token en NLP. Supongamos que la imagen original tiene tamaño  $H \times W \times C$ , donde  $H$  y  $W$  son la altura y el ancho, y  $C$  es el número de canales. Al dividir la imagen en parches de tamaño  $p \times p$ , el número total de parches es  $N = \frac{HW}{p^2}$ .

**Incrustación Lineal** Cada parche se aplana en un vector y se proyecta linealmente a un espacio de menor dimensión mediante una transformación lineal:

$$z_i = WE(x_i) + b, \quad (4.1)$$

donde  $x_i$  representa el  $i$ -ésimo parche,  $W$  es la matriz de pesos de la proyección, y  $b$  es el sesgo.

**Incrustaciones de Posición** Para mantener la información espacial, se añaden incrustaciones posicionales a cada vector de parche:

$$z'_i = z_i + p_i, \quad (4.2)$$

donde  $p_i$  representa la posición del parche  $i$  en la imagen original.

**Token de Clasificación** Se introduce un token especial  $z_{cls}$  al inicio de la secuencia de parches. Este token aprende una representación global de la imagen y es utilizado en la clasificación final.

**Codificador Transformer** La secuencia de parches, junto con el token de clasificación, es procesada mediante un codificador Transformer. Cada

bloque Transformer incluye un mecanismo de auto-atención multicabeza que permite que cada parche preste atención a otros parches en la imagen; un perceptrón multicapa, el cual transforma la representación de cada parche mediante capas densas y activaciones no lineales; y una normalización de capa, encargada de mejorar la estabilidad del entrenamiento. Además, cuenta con conexiones residuales, las cuales facilitan el flujo de información y evitan problemas de desvanecimiento del gradiente.

### Integración con GPT2

El modelo ViT devuelve un conjunto de características de la imagen, las cuales se utilizan como entrada de un decodificador basado en GPT2 para obtener un texto descriptivo en lenguaje natural.

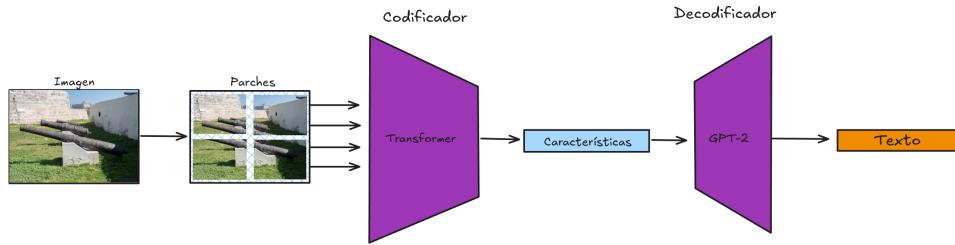


Figura 4.2: ViT

### 4.2.3. CLIP

**CLIP** [8] es un modelo de aprendizaje profundo que permite asociar representaciones de texto e imagen en un espacio multimodal compartido. Su entrenamiento se basa en un esquema contrastivo que optimiza la similitud del coseno entre representaciones correctas y minimiza la similitud de pares incorrectos. Este enfoque permite realizar tareas como clasificación zero-shot y recuperación de imágenes a partir de descripciones textuales.

#### Entrenamiento Contrastivo

El modelo se entrena utilizando un conjunto de datos masivo de pares imagen-texto obtenidos de la web. El procedimiento de entrenamiento sigue los siguientes pasos:

Se presenta un lote de  $N$  pares  $(I, T)$ , donde  $I_i$  representa una imagen y  $T_i$  su descripción textual. Se codifican las imágenes y textos utilizando un codificador de imágenes y un codificador de texto. Se proyectan los embeddings generados a un espacio de representación compartido mediante una transformación lineal. Se normalizan los vectores resultantes utilizando la norma  $L_2$  para asegurar que todos tengan la misma escala. Se calcula la similitud del coseno entre los pares  $(I_i, T_j)$  generando una matriz de similitud de tamaño  $N \times N$ . Se optimiza una función de pérdida de entropía cruzada simétrica para maximizar la similitud de los pares correctos y minimizar la similitud de los pares incorrectos.

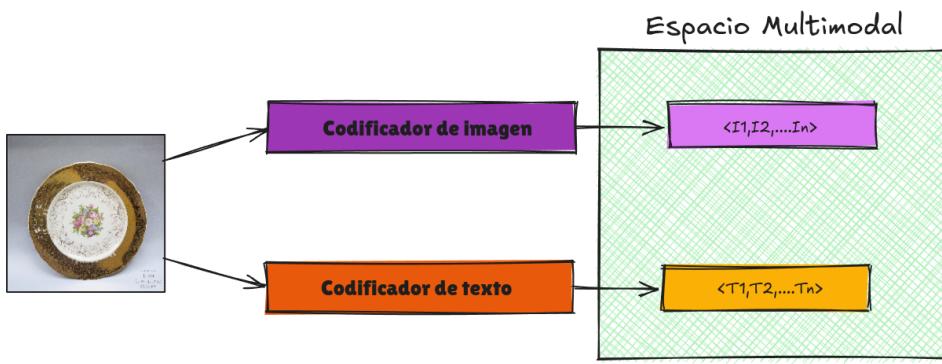


Figura 4.3: Entrenamiento de CLIP

### Codificador de Imágenes

CLIP emplea dos arquitecturas principales para el codificador de imágenes: ResNet modificado y Vision Transformer. El proceso de codificación de una imagen consiste en extraer características visuales relevantes, proyectarlas a un espacio de menor dimensión y normalizarlas para generar un vector representativo.

### Codificador de Texto

El codificador de texto se basa en un Transformer de 12 capas con 512 dimensiones y 8 cabezales de atención. Su procesamiento incluye tokenización del texto usando codificación de pares de bytes, adición de tokens especiales [SOS] y [EOS], pasaje de los tokens por el Transformer, extra-

yendo la activación del token [EOS] como representación del texto, así como normalización y proyección al espacio de embedding multimodal.

### Espacio de Embedding Multimodal

Ambos codificadores transforman sus respectivas entradas en vectores en un espacio compartido. CLIP maximiza la similitud del coseno entre embeddings de pares imagen-texto correctos y minimiza la similitud de los incorrectos. Esto le permite generalizar a nuevas imágenes y descripciones nunca vistas durante el entrenamiento.

### Ejecución del modelo

Para seleccionar entre un conjunto de descripciones el de mayor ajuste dada una imagen, el modelo codifica la imagen y los textos con el objetivo de representar las características de los mismos en un espacio multimodal. Seguidamente, se calcula la similitud de cosenos entre los vectores de texto y el vector de la imagen, seleccionando la descripción que más se ajuste a la imagen.

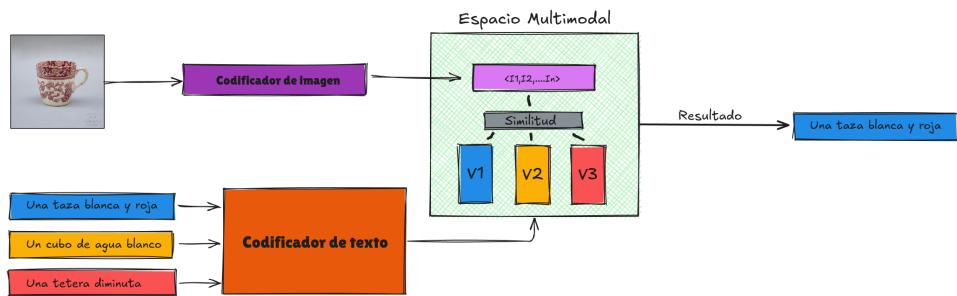


Figura 4.4: Ejecución del modelo CLIP

## Capítulo 5

# Experimentación y resultados

Para evaluar el modelo de subtitulado de imágenes, se seleccionó un conjunto de 391 imágenes de diversas colecciones, garantizando una representación variada de los datos.

Una vez que las imágenes fueron seleccionadas, se crearon textos descriptivos para cada una de ellas, las cuales fueron recopiladas en un documento que se utilizó como conjunto de referencia para la evaluación.

Seguidamente, se utilizó el modelo propuesto para generar descripciones automáticas para las imágenes, que se compararon directamente con las descripciones generadas manualmente mediante las métricas de evaluación BLEU, METEOR, CIDEr y SPICE, ROUGE-1 y ROUGE-L.

### Métricas empleadas

- **CIDEr:** es una forma de evaluar la calidad de las descripciones textuales generadas. Mide la similitud entre una descripción generada y la de referencia. Se basa en la idea de que las buenas descripciones no solo deben ser similares a las descripciones de referencia en términos de elección de palabras y gramática, sino también en términos de significado y contenido.
- **SPICE:** mide la calidad semántica de las descripciones de imágenes generadas por modelos. A diferencia de otras métricas que se centran en la similitud superficial de palabras, evalúa el contenido semántico dividiendo las descripciones en representaciones de gráficos semánticos, las cuales incluyen objetos, atributos y relaciones. Estos se comparan

luego con gráficos de referencia creados a partir de descripciones humanas. Esta métrica se enfoca en capturar la precisión y la integridad semántica, asegurándose de que la descripción generada refleje correctamente los elementos y las relaciones presentes en la imagen.

- **BLEU:** se calcula comparando el texto traducido automáticamente con uno o más textos traducidos por humanos. Se considera que la calidad es la correspondencia entre la salida de la máquina y la de una persona.
- **METEOR:** tiene en cuenta tanto la precisión como la fluidez de la traducción, así como el orden en que aparecen las palabras. El algoritmo compara el texto traducido con la traducción de referencia humana descomponiéndolos en fragmentos y calculando la similitud entre cada fragmento. Finalmente, se utiliza el promedio ponderado de estas medidas para calcular el puntaje METEOR general.
- **ROUGE:** es un conjunto de métricas utilizadas para evaluar la calidad de los modelos de traducción y resumen de documentos. Mide la superposición entre un resumen o traducción generado por el sistema y un conjunto de resúmenes o traducciones de referencia creados por humanos.
- **ROUGE-N:** mide la superposición de n-gramas entre los resúmenes generados por el modelo y los resúmenes de referencia.
- **ROUGE-L:** Se basa en la longitud de la subsecuencia común más larga. Calcula la media armónica ponderada, combinando el puntaje de precisión y el de recall. No requiere coincidencias consecutivas, sino coincidencias en secuencia.

La siguiente tabla representa los resultados obtenidos al evaluar las métricas:

Métricas	SPICE	CIDEr	METEOR	BLEU	ROUGE-1	ROUGE-L
Resultados	0.0469	0.3927	0.2452	0.110	0.2792	0.2506

### Valoración de los resultados

Los resultados indican que el modelo genera descripciones con cierta coherencia léxica pero baja precisión semántica, como lo reflejan los puntajes

bajos en SPICE (0.0469) y CIDEr (0.3927), lo que sugiere dificultades para capturar correctamente la estructura semántica y los términos clave de la imagen. METEOR (0.2452) y ROUGE (0.2792 y 0.2506) muestran una superposición moderada con las descripciones humanas, mientras que BLEU (0.0110) es extremadamente bajo, indicando que la redacción del modelo difiere significativamente de las referencias. Sin embargo, si las descripciones humanas son inconsistentes o imprecisas, estas métricas pueden estar penalizando al modelo más de lo necesario.

Como análisis adicional, se puede destacar que las descripciones generadas por el modelo BLIP, fueron seleccionadas un 88% de las veces, demostrando una mayor preferencia hacia el mismo por el modelo de selección.

# Capítulo 6

## Conclusiones

La evolución de los modelos de generación de descripciones de imágenes ha sido notable en los últimos años, pasando de arquitecturas basadas en RNNs con atención visual a enfoques más avanzados que integran Transformers y aprendizaje multimodal. Modelos recientes, como CLIP y BLIP, han demostrado el potencial de combinar grandes volúmenes de datos de texto e imagen, logrando una comprensión más profunda del contenido visual. Estos avances no solo han mejorado la calidad y precisión de las descripciones generadas, sino que también han abierto nuevas posibilidades en tareas de visión y lenguaje, estableciendo el camino para futuras investigaciones en la intersección entre inteligencia artificial y percepción visual.

Los resultados obtenidos en la evaluación del modelo presentado evidencian que, si bien el sistema logra generar descripciones con coherencia léxica, presenta limitaciones en la captura de relaciones semánticas profundas. Sin embargo, es importante tener en cuenta que la generación de subtítulos manualmente se limita a un conjunto específico de imágenes y no fueron realizadas por expertos. Al no contar con un método supervisado previamente, la evaluación de los resultados constituyó un desafío. Los puntajes moderados en METEOR y ROUGE reflejan que el modelo mantiene cierta similitud con las descripciones humanas, lo que implica que la generación de texto sigue patrones lingüísticos aceptables, aunque con margen de mejora en la exactitud del contenido.

# Capítulo 7

## Apéndice

A continuación se presentan algunos ejemplos de subtítulos generados.



Figura 7.1: Un dibujo de un hombre con un abrigo de piel (Exitoso)



Figura 7.2: Un espejo con un mango de plata (Exitoso)



Figura 7.3: Una portada de un libro con un niño jugando con una pelota (Exitoso)



Figura 7.4: Un grupo de seis japoneses jugando a las cartas (Poco exitoso)

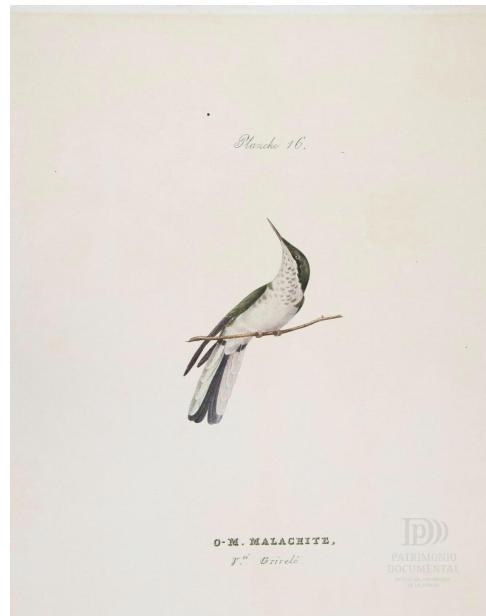


Figura 7.5: Un pájaro sentado sobre una rama con las alas extendidas (Poco exitoso)



Figura 7.6: Una moneda con una foto de un hombre en ella (Poco exitoso)



Figura 7.7: Un pan blanco con un glaseado blanco (No exitoso)



Figura 7.8: Una caja pequeña con un mango de metal y un mango de metal (No exitoso)



Figura 7.9: Un juguete pequeño con un par de zapatos en él

# Referencias

- [1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2015. (Citado en la página 5).
- [2] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. (Citado en la página 5).
- [3] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional image captioning. *arXiv preprint arXiv:1805.09019*, 2018. (Citado en la página 6).
- [4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint arXiv:1707.07998*, 2018. (Citado en la página 6).
- [5] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2017. (Citado en la página 6).
- [6] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *arXiv preprint arXiv:1905.07841*, 2019. (Citado en la página 6).
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021. (Citado en las páginas 7 y 11).

- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. (Citado en las páginas 7 y 13).
- [9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. (Citado en las páginas 7 y 10).