

---

Generación de explicaciones basadas en ejemplos y  
contraejemplos a través de técnicas de Visual Question  
Answering

Generation of explanations based on examples and  
counterexamples through Visual Question Answering  
techniques

---



Trabajo de Fin de Máster  
Curso 2024–2025

**Autor**

Ángel Bastardo Rojas

**Director**

Marta Caro Martínez

Máster en Internet de las Cosas  
Facultad de Informática  
Universidad Complutense de Madrid



Generación de explicaciones basadas en  
ejemplos y contraejemplos a través de  
técnicas de Visual Question Answering  
Generation of explanations based on  
examples and counterexamples through  
Visual Question Answering techniques

**Trabajo de Fin de Máster en Internet de las Cosas**  
**Departamento de Ingeniería de Software e Inteligencia Artificial**

**Autor**  
**Ángel Bastardo Rojas**

**Director**  
**Marta Caro Martínez**

**Convocatoria:** *Junio 2025*  
**Calificación:** *9*

**Máster en Internet de las Cosas**  
**Facultad de Informática**  
**Universidad Complutense de Madrid**

**4 de julio de 2025**



# Dedicatoria

*A mi abuela*



# Agradecimientos

A mi tutora



# Resumen

## Generación de explicaciones basadas en ejemplos y contraejemplos a través de técnicas de Visual Question Answering

La poca transparencia en los modelos de Inteligencia Artificial (IA) se presenta como uno de los retos hoy en día en la industria. Esta poca transparencia se refiere a lo complicado de explicar cómo los grandes modelos de inteligencia artificial logran tomar una decisión. Por ello cobra importancia la Inteligencia Artificial Explicable (XAI), campo de la inteligencia artificial que consiste en diseñar sistemas o métodos para hacer más transparente y fácil de comprender los mecanismos internos de los modelos de IA para los usuarios.

En este Trabajo de Fin de Máster se propone un sistema de Case-Based Reasoning (CBR) que usa modelos de lenguaje multimodal, en concreto Large Language Models (LLMs) para Visual Question Answering (VQA) y métricas de similitud textual y semántica, generando explicaciones basadas en ejemplos (factual) y contraejemplos (counterfactual). El método utiliza un modelo multimodal para crear descripciones textuales de imágenes y se utilizan métricas de similitud para recuperar las explicaciones y reutilizar la información textual generada anteriormente. Ambos enfoques (factual y conuterfactual) se han evaluado de forma offline y online, con usuarios reales, obteniendo resultados que indican que el método propuesto en este trabajo mejora la explicabilidad ante métodos tradicionales encontrados en la literatura.

### Palabras clave

Inteligencia Artificial, Explicabilidad, Large Language Model, Visual Question Answering, Razonamiento Basado en Casos, XAI.



# Abstract

## Generation of explanations based on examples and counterexamples through Visual Question Answering techniques

The lack of transparency in Artificial Intelligence (AI) models is currently one of the main challenges in the industry. This lack of transparency refers to the difficulty in explaining how large artificial intelligence models make decisions. For this reason, Explainable Artificial Intelligence (XAI) becomes important, it is the field of artificial intelligence focused on designing systems or methods to make the internal mechanisms of AI models more transparent and easier for users to understand.

This Master Project proposes a Case-Based Reasoning (CBR) system that uses multimodal language models, specifically Large Language Models (LLMs) for Visual Question Answering (VQA), textual and semantic similarity metrics, generating explanations based on examples (factual) and counterexamples (counterfactual). The method employs a multimodal model to create textual descriptions of images and uses similarity metrics to retrieve explanations while reusing previously generated textual information. Both approaches (factual and counterfactual) were evaluated offline and online with real users, yielding results indicating that the proposed method improves explainability compared to traditional methods found in the literature.

### Keywords

Artificial Intelligence, Explainability, Large Language Model, Visual Question Answering, Case-Based Reasoning, XAI.



# Índice

<b>1. Introducción</b>	<b>1</b>
1.1. Motivación . . . . .	1
1.2. Objetivos . . . . .	3
1.3. Plan de trabajo . . . . .	3
1.4. Publicaciones . . . . .	4
<b>2. Estado de la Cuestión</b>	<b>5</b>
2.1. Explainable Artificial Intelligence (XAI) . . . . .	5
2.1.1. Objetivos XAI . . . . .	5
2.1.2. Métodos de explicabilidad . . . . .	6
2.1.3. Explicaciones basadas en instancias . . . . .	8
2.2. Visual Question Answering (VQA) . . . . .	8
2.2.1. VQA + XAI . . . . .	11
2.3. Case-Based Reasoning (CBR) . . . . .	16
<b>3. Método de explicación basado en CBR y VQA</b>	<b>21</b>
3.1. Base de casos VQA . . . . .	22
3.1.1. Questions . . . . .	22
3.2. Recuperación (Retrieval) . . . . .	24
3.2.1. Métricas de similitud . . . . .	26
3.3. Reuse: Personalización de explicación . . . . .	28
<b>4. Evaluación</b>	<b>31</b>
4.1. Evaluación offline . . . . .	31
4.1.1. Dataset . . . . .	31
4.1.2. Métricas de evaluación . . . . .	32
4.1.3. Proceso experimental . . . . .	34
4.1.4. Resultados . . . . .	37
4.2. Evaluación online . . . . .	39
4.2.1. Resultados . . . . .	40
4.2.2. Conclusiones obtenidas de la evaluación . . . . .	41
<b>5. Conclusiones y Trabajo Futuro</b>	<b>43</b>
5.1. Conclusiones . . . . .	43

5.2. Trabajo futuro . . . . .	44
<b>6. Introduction</b>	<b>45</b>
6.1. Motivation . . . . .	45
6.2. Objectives . . . . .	46
6.3. Work plan . . . . .	47
6.4. Publications . . . . .	48
<b>7. Conclusions and Future Work</b>	<b>49</b>
7.1. Conclusions . . . . .	49
7.2. Future Work . . . . .	49
<b>Bibliografía</b>	<b>51</b>
<b>A. Questions for VQA model</b>	<b>57</b>
A.1. Human . . . . .	57
A.2. Animal . . . . .	60
A.3. Object . . . . .	62
A.4. Landscape . . . . .	64
<b>B. Categorías y etiquetas de base de casos</b>	<b>67</b>
B.1. Human . . . . .	67
B.2. Animal . . . . .	67
B.3. Object . . . . .	68
B.4. Landscape . . . . .	69
<b>C. Ejemplos de secciones de formulario de evaluación online</b>	<b>71</b>

# Índice de figuras

2.1.	Mapa de atención generado a partir de las características de entrada. Fuente: [4] . . . . .	12
2.2.	Arquitectura del bucle de active attention dentro del sistema XVQA. Fuente: [4] . . . . .	12
2.3.	Explicaciones de las cajas delimitadoras generadas en función de los pesos de atención espacial para la pregunta “¿Qué está haciendo el hombre?”. Fuente: [4] . . . . .	13
2.4.	Izquierda: Escenario de entrada. Derecha: Escenario filtrado según las ponderaciones del mapa de atención generado por el modelo en respuesta a una pregunta. Fuente: [4] . . . . .	13
2.5.	(b) Atención a nivel de objeto en comparación con (a) atención espacial. Fuente: [4] . . . . .	14
2.6.	Ejemplo de resultado del modulo de lenguaje natural realizando una explicación textual para la respuesta original del modelo Fuente: [4] . . . . .	15
2.7.	Arquitectura del modelo MRVQA. Fuente: [28] . . . . .	17
2.8.	Ciclo sistema CBR. Fuente: [54] . . . . .	18
2.9.	Metodología CBR de iSee. Fuente: [43] . . . . .	20
3.1.	Descripción general del método de explicación basado en casos de VQA donde $Cs$ denota la clase, $Cat$ categoría de la imagen según el agrupamiento que se ha propuesto en este trabajo y $qa$ , las descripciones que contienen los pares de preguntas y respuestas. . . . .	22
3.2.	Ejemplo de gatos idénticos pero con descripciones distintas. . . . .	23
3.3.	Filtering step . . . . .	25
3.4.	Sorting step: Dado un consulta y una base de casos, se recuperan los casos en MAC y se realiza una ordenación por valor de similitud y se escogen los $k$ casos más similares . . . . .	26
3.5.	Ejemplo de una explicación basado en ejemplos (factual) obtenida con nuestro método para una imagen pixelada. La imagen de la izquierda es la clasificada por un modelo de caja negra, mientras que la imagen de la derecha es el ejemplo de explicación obtenido con nuestro método. El texto siguiente es la explicación en lenguaje natural adaptada al objetivo de eficiencia (que ayuda a los usuarios a decidir rápidamente si la clasificación es correcta). Este texto se muestra con la explicación en el ejemplo (b). . . . .	29
4.1.	Estructura de agrupación de categorías y distribución de clases. . . . .	32

4.2.	Una imagen de una cámara de vigilancia de tráfico en Alemania se utiliza para mostrar los resultados del RMSE. Fuente: [36] . . . . .	33
4.3.	GradCAM + Removed y debias (ROAD) aplicado a la clase “perro”. Fuente: [14] . . . . .	34
4.4.	Prompt de ejemplo y comparación con otros modelos donde LLaVA ofrece una respuesta más completa que GPT-4 [39], inclusive cuando se le pide simplemente describir la imagen, LLaVA identifica los aspectos atípicos de la imagen. Asimismo, cuando se le pide simplemente describir la imagen, LLaVA identifica los aspectos atípicos de la imagen. Fuente: [30] . . . . .	35
4.5.	Arquitectura LLaVA. Fuente: [30] . . . . .	35
4.6.	Implementación VQA . . . . .	36
4.7.	Resultados obtenidos cuando se le preguntaba a los usuarios qué explicación les ayuda a entender la clasificación lo más rápido posible. La opción A hace referencia a la explicación obtenida con el baseline y la opción B hace referencia a la explicación obtenida con nuestro método. . . . .	41
4.8.	Resultados obtenidos cuando se le preguntaba a los usuarios qué explicación les ayuda a entender la clasificación lo más rápido posible. La opción A hace referencia a la explicación obtenida con el baseline y la opción B hace referencia a la explicación obtenida con nuestro método. . . . .	42
C.1.	Ejemplo de primera sección de encuesta . . . . .	71
C.2.	Ejemplo de tercera sección de encuesta . . . . .	72

# Índice de tablas

3.1. Clases de similitud utilizada . . . . .	27
3.2. Clases de similitud utilizada . . . . .	27
4.1. Resultados obtenidos en la evaluación (los mejores resultados en negrita). M indica la métrica utilizada: Bag distance (B), cosine (C), Jaccard (J), Levenshtein (L), Overlap (O), SBERT (SB), Web similarity (WS), or pixel- to-pixel (P2P). ET indica el tipo de explicación: por ejemplo (factual F) o contraejemplo (counterfactual CF). El símbolo % indica el porcentaje de mejora de nuestro método en comparación con el baseline. . . . . .	39



# Introducción

## 1.1. Motivación

En la actualidad la inteligencia artificial esta cada vez más presente en nuestro día a día, en tareas como recomendación de contenidos de entretenimiento, reconocimiento de patrones de comportamientos, detección de enfermedades, asistente de compras, entre otras. A medida que estas necesidades van surgiendo, los modelos de Inteligencia Artificial (IA) también se hacen más complejos, por ejemplo, han surgidos técnicas como el deep learning. Debido a esta alta complejidad de técnicas surgió el campo de la Inteligencia Artificial Explicable (XAI) que tiene como objetivo ayudar a que los usuarios comprendan el comportamiento de estos modelos. Hacer estos modelos más transparentes en sus predicciones ayudan al usuario a entender el comportamiento del modelo de IA, y las razones por las que se producen las predicciones, aumentando así su confianza y por tanto la eficiencia de estos sistemas [8].

Uno de los tipos de explicación más exitosos proporcionados por las técnicas XAI son las explicaciones basadas en instancias [20], denominada explicaciones basadas en casos si se obtienen a través del razonamiento basado en casos (CBR). Las explicaciones basadas en instancias consisten en mostrar ejemplos de situaciones similares a la que se desea explicar. Cuando estas explicaciones se basan en ejemplos, los casos proporcionados corresponden a situaciones en las que el modelo de IA realiza la misma clasificación que en el caso a explicar, lo que se conoce como explicaciones basados en ejemplos (factual). Por el contrario, los basados en contraejemplos (counterfactual) [20] corresponden a situaciones similares en las que el modelo realiza una clasificación diferente al caso a explicar. Específicamente, en la clasificación de imágenes, las explicaciones basados en ejemplo (factual) mostrarían imágenes que han sido clasificadas igual que la imagen que se quiere explicar, mientras que las explicaciones basados en contraejemplos (counterfactual) mostrarían imágenes parecidas, pero que han recibido una clasificación distinta al caso a explicar. Comúnmente para explicar la clasificación de imágenes, las métricas de similitud más utilizadas se basan en las características de la imagen [41]. Siendo la más común píxel to píxel, la cual consiste en comparar dos imágenes a través de la diferencia entre sus matrices de píxeles.

Actualmente hay sistemas de IA de alta complejidad, entre ellos los modelos multimodales, que combinan diferentes tipos de datos, como imágenes y texto, para aprender y predecir. Además, gracias al alto rendimiento de nuevos algoritmos de procesamiento de

lenguaje natural basados en transformers, su aplicación influye en las mejoras de tareas no solo relacionadas con texto, sino también con imágenes [15]. Visual Question Answering (VQA) es la tarea en la que los modelos multimodales de IA responden a preguntas relacionadas con imágenes [8, 15]. En este trabajo, proponemos el uso de modelos multimodales, en particular modelos Visual Question Answering (VQA), para obtener explicaciones basadas en instancias, tanto explicaciones basados en ejemplos (factual) como basados en contraejemplos (counterfactual) en la clasificación de imágenes. **Nuestra hipótesis es que, utilizando este tipo de modelos multimodales, podemos proporcionar mejores explicaciones a los usuarios, incluyendo también descripciones personalizadas adicionales en lenguaje natural sobre las imágenes.** Para ello, hemos desarrollado un método de explicación post-hoc basado en casos y model-agnostic del modelo para explicar imágenes de una sola clase. Este método utiliza las descripciones creadas por el modelo VQA aplicado a las imágenes para obtener una base de casos. Las descripciones son las respuestas a las preguntas que nuestro método plantea al modelo VQA. El propósito de estas preguntas es extraer información sobre las características de estas imágenes y objetivos relacionados con XAI como *transparency*, *effectiveness* o *persuasiveness*. Seguidamente, nuestro método utiliza similitudes semánticas y textuales para comparar las descripciones entre la consulta (la imagen a explicar) y las imágenes de la base de casos. De este modo, obtenemos los casos de explicación, tanto como explicaciones basados en ejemplos (factual) como basados en contraejemplos (counterfactual). Además, en el paso de reutilización (o adaptación), nuestro método personaliza las explicaciones mostrando a los usuarios las respuestas a las preguntas relacionadas con el objetivo XAI que necesitan resolver. Por ejemplo, si el usuario necesita que la clasificación sea más transparente, el método CBR mostrará a los usuarios las descripciones de las imágenes para que el modelo de IA sea más transparente.

El método que proponemos puede ser especialmente adecuado cuando las imágenes se obtienen con mala calidad, ya que es más probable que los usuarios no entiendan claramente la clasificación. Por ejemplo, en contextos en los que los sensores del Internet de las Cosas (IoT) hayan capturado o transmitido imágenes de mala calidad (borrosas o pixeladas). En esta situación, la descripción reutilizada puede ayudar a los usuarios a comprender mejor la imagen y los ejemplos de explicación.

Para determinar si el método funciona como esperamos y mejora el rendimiento de las explicaciones de casos basadas únicamente en imágenes, lo hemos evaluado. Realizamos una evaluación offline en la que comparamos los casos de explicación obtenidos por nuestro baseline (utilizando como baseline una métrica de similitud basada en imágenes: píxel to píxel) y los casos obtenidos por nuestro método. Medimos la diferencia entre los casos y la imagen a explicar. Cuanto menor sea esta diferencia a la imagen a explicar, más útiles son para identificar la clasificación de esta imagen. Además, medimos las regiones de las imágenes que contribuyen a la comprensión de la imagen, por lo nos aseguramos de no perder esas regiones incluso cuando sólo utilizamos conocimiento semántico y basado en texto para obtener los casos de explicación. Posteriormente, realizamos una evaluación online donde se presentó un formulario de evaluación a usuarios reales donde debían elegir entre las explicaciones más útiles, pudiendo elegir entre las generadas por nuestro método y las generadas por el método baseline.

## 1.2. Objetivos

El objetivo principal es proponer un método XAI post-hoc y model-agnostic basado en CBR que usa técnicas de VQA para proporcionar explicaciones basados en ejemplos (factual) y basados en contraejemplos (counterfactual) en clasificación de imágenes. Asimismo, evaluar y determinar si nuestro método CBR mejora las explicaciones ofrecidas por otros métodos encontrados en la literatura para poder concluir que el uso de modelos multimodales en XAI es un enfoque prometedor en el que deberíamos profundizar en futuras investigaciones.

Se desglosa el conjunto de objetivos específicos a partir del objetivo principal de la siguiente manera:

- **Investigar sobre modelos multimodales para tareas de Visual Question Answering:** Detallar cuáles son los más utilizados, por qué ciertos modelos resultan útiles para los usuarios, cómo se implementan, cuáles son sus propósitos generales y qué recursos están disponibles para uso libre.
- **Exploración de métricas de similitud:** Estudiar que tipos de métricas están disponibles para el análisis de texto, tanto de tipo semánticas como métricas basadas en el texto, estudiar su implementación y resultados.
- **Exploración de datasets:** Investigar que datasets públicos se encuentran disponibles que cuenten con imágenes que contengan un solo tipo de entidad principal (clase única), por ejemplo, si en la imagen se encuentra un perro, entonces que no se encuentren personas u otro tipo de animales.
- **Construcción de método CBR:** Construir un método CBR que combina VQA y las métricas de similitud estudiadas para generar explicaciones basados en ejemplos (factual) y basados en contraejemplos (counterfactual).
- **Evaluaciones:** Evaluar el modelo en un entorno IoT, donde estos métodos pueden ser de especial utilidad. Para ello estudiar qué tipo de métricas de evaluación serían las más adecuadas para realizar evaluaciones basadas en imágenes, estudiar su implementación y eficiencia. Asimismo evaluar la opinión de los usuarios respecto a las explicaciones obtenidas.

## 1.3. Plan de trabajo

Para cumplir con los objetivos, primero se llevará a cabo una investigación sobre la Inteligencia Artificial Explicable, con la finalidad de entender el reto que presenta, las limitaciones actuales y el uso de herramientas para lograr aplicar técnicas que describan algunos modelos complejos. Igualmente estudiar los sistemas CBR para generar casos de explicación y cómo se puede implementar en el desarrollo de nuestro método. Esto se abordará en el capítulo 2 (Estado de la Cuestión).

Se estudiarán modelos de Visual Question Answering (VQA) que consisten en responder preguntas en lenguaje natural sobre imágenes y que estén disponibles con acceso público para fines de investigación. Es fundamental valorar el tamaño de estos modelos para determinar si es viable ejecutarlos con los recursos de hardware disponibles. Se explorarán librerías de HuggingFace<sup>1</sup> como transformers, que permiten cargar modelos multimodales (por ejemplo, de VQA), preparar automáticamente los datos de entrada y generar respuestas textuales basadas en imágenes. También se investigarán plataformas gratuitas donde estos modelos pueden ejecutarse, para combinar ejecuciones locales y en línea y así optimizar los tiempos.

Se explorarán datasets de imágenes que contengan un tipo de entidad principal para desarrollar un método CBR basado en VQA, donde este último extraerá información relevante en lenguaje natural a partir de dichas imágenes para construir una base de casos. Posteriormente se estudiarán métricas de similitud textual y semántica para obtener explicaciones basados en ejemplos (factual) y contraejemplos (counterfactual). Además de ello se reutilizará la información textual de las explicaciones mostrando a los usuarios las respuestas más relevantes según el objetivo a abordar. Esto se describe en el Capítulo 3 (Método de explicación basado en CBR y VQA).

Una vez implementado el método propuesto en este TFM, debemos determinar si nuestro método funciona como se espera, para ello se investigará sobre métricas de evaluación implementadas en tareas de similitud entre imágenes, con el enfoque de realizar comparativas cuantitativa acerca de la precisión con la que una imagen se asemeja a otra, y la calidad explicativa de las imágenes proporcionadas. También se estructurará una encuesta y con ello realizar una evaluación online con usuarios reales, para recopilar sus opiniones y medir la calidad de las explicaciones generadas. Igualmente se considerará la implementación de las métricas de evaluación de cara a los tiempos de ejecución para obtener dichas evaluaciones. La evaluación online y offline se describen en el Capítulo 4 (Evaluación).

## 1.4. Publicaciones

Los resultados de esta investigación han sido publicados en un artículo académico y aceptados para su presentación en el congreso International Conference on Case-Based Reasoning (ICCBR) de 2025. Se presentarán en el main track, en Biarritz, del 30 de junio al 3 de julio y los resultados se incluirán en los proceedings del congreso [7].

---

<sup>1</sup>Disponible en [https://huggingface.co/docs/transformers/tasks/visual\\_question\\_answering](https://huggingface.co/docs/transformers/tasks/visual_question_answering)

# Capítulo 2

## Estado de la Cuestión

En este capítulo se estudiará las bases fundamentales incluyendo antecedentes académicos y desarrollos previos para llevar a cabo este Trabajo de Fin de Máster: la Inteligencia Artificial explicable (Sección 2.1), el Visual Question Answering (Sección 2.2), y el Razonamiento Basado en Casos (Sección 2.3).

### 2.1. Explainable Artificial Intelligence (XAI)

En 2016 se han comenzado a aprobar leyes, que obligan a que los interesados tengan el derecho de no ser objeto de una decisión basada únicamente en un tratamiento automatizado que produzca efectos jurídicos [57]. Igualmente en 2024 fue publicada la ley EU AI [56] que pretende que cuando dichos sistemas de IA no sean suficientemente transparentes, explicables y documentados, es conveniente clasificar como de alto riesgo estos sistemas siempre y cuando estén destinados a ser utilizados en el contexto de la aplicación de las leyes cuando la precisión, la fiabilidad y la transparencia son especialmente importantes para evitar impactar la confianza del público. Especialmente en aquellas acciones que impliquen determinados uso de sistemas IA puestos en servicios o comercializados de modo que puedan señalar a usuarios de forma discriminatorio o de manera incorrecta.

La Inteligencia Artificial Explicable propone crear un conjunto de técnicas que produzcan modelos más explicables manteniendo un alto nivel de rendimiento, que permitan a los humanos entender y confiar adecuadamente. Durante el periodo 2013 a 2019 hay una tendencia al alza en las contribuciones. En 2017 comienza un auge de estas, cuyo título, resumen y/o palabras claves hacen referencias al campo de la Inteligencia Artificial Explicable, esto en cierta parte debido a las agendas de investigación de los gobiernos ya que hoy en día es necesario explicar los comportamientos de los modelos de IA en distintos ámbitos profesionales, como pueden ser en medicina donde se necesita información adicional a una predicción binaria que diga si una persona tiene o no una enfermedad [5].

#### 2.1.1. Objetivos XAI

En [37] se propone que los objetivos de generar explicaciones deberían enfocarse en:

- **Transparency:** Explicar cómo funciona el sistema.
- **Effectiveness:** Ayudar a los usuarios a tomar buenas decisiones.

- **Trust:** Incrementar la confianza de los usuarios en el sistema.
- **Persuasiveness:** Convencer a los usuarios para tomar decisiones.
- **Satisfaction:** Aumentar la facilidad de uso.
- **Education:** Permitir que los usuarios aprendan del sistema.
- **Efficiency:** Ayudar a los usuarios a tomar decisiones más rápido.
- **Debugging:** Permitir a los usuarios identificar que hay defectos en el sistema.

### 2.1.2. Métodos de explicabilidad

Arrieta et al. [5] examina los diferentes niveles de transparencia que un modelo puede ofrecer por sí mismo, así como define distintos métodos XAI: post-hoc y ante-hoc.

Retzlaff et al. [47] define el concepto de la explicabilidad post-hoc que se refiere a las técnicas que se aplican a un modelo de aprendizaje automático después de que ha sido entrenado, para entender cómo y por qué tomó ciertas decisiones. Estos métodos no modifican el modelo en sí, sino que proporcionan explicaciones sobre su comportamiento. Por otro lado, ante-hoc se refiere a modelos que son intrínsecamente explicables, aquellos que están diseñados desde el principio para ser interpretables y comprensibles, la explicabilidad de estos está integrada directamente en la arquitectura del modelo, por ejemplo, un modelo de árbol de decisión donde cada nodo representa una regla clara y entendible al usuario. Las técnicas de explicabilidad post-hoc son inherentes aplicados como model-agnostic, ya que son técnicas que se pueden implementar sobre cualquier modelo con el fin de extraer información de su procedimiento de predicción, sin requerir un conocimiento profundo de los mecanismos internos del modelo. Es decir, estos métodos funcionan independientemente de la arquitectura del modelo, lo que los hace flexibles y aplicables a distintos modelos de aprendizaje automático por ejemplo, redes neuronales, entre otros [11].

Los métodos model-agnostic nos permiten explicar cualquier modelo de caja negra aproximándolo con un modelo interpretable más sencillo, analizando las contribuciones de los rasgos o analizando las contribuciones de las características. Son especialmente útiles cuando no tenemos acceso a un modelo o cuando trabajamos con sistemas propietarios [11]. Algunos de ellos:

- **LIME (Local Interpretable Model-Agnostic Explanations):** Explica las predicciones individuales mediante el entrenamiento de un modelo sustituto simplificado que aproxima el complejo comportamiento del modelo en torno a una instancia determinada. La idea clave es perturbar los datos de entrada, observar los cambios en el predicciones del modelo, y ajustar un modelo lineal local para aproximar los límites de decisión [11].
- **SHAP (Shapley Additive Explanations):** Explica las predicciones del modelo utilizando conceptos de la teoría de juegos cooperativos [50], las características de un modelo actúan como “jugadores” que cooperan para generar una predicción. El método calcula la contribución marginal de cada característica a la predicción final. A diferencia de LIME, que aproxima el comportamiento local, SHAP garantiza la interpretabilidad global y local [11].

Por otro lado, se encuentran los métodos model-specific, que aprovechan la estructura interna de un modelo particular, como las redes neuronales, para ofrecer explicaciones más precisas y adaptadas a la arquitectura. Estos métodos suelen utilizar gradientes y activaciones internas para generar dichas explicaciones. [11]. Entre ellos:

- **GradCAM (Gradient-weighted Class Activation Mapping)** [11]: Grad-CAM visualiza las regiones más influyentes de una imagen al calcular los gradientes de la puntuación de la clase con respecto a los mapas de características en la CNN (Convolutional Neural Networks). Ayuda a localizar características discriminativas utilizadas por la red.

#### Flujo de trabajo de Grad-CAM:

- **Propagación hacia adelante:** Calcular los mapas de características a partir de una capa convolucional.
- **Cálculo de gradientes:** Calcular los gradientes específicos de la clase con respecto a estos mapas de características.
- **Agregación de pesos:** Calcular la importancia de cada mapa de características usando un promedio global.
- **Generación del mapa de calor:** Aplicar ReLU para generar un mapa de activación positiva superpuesto sobre la imagen original.

**Formulación matemática** El mapa de calor de Grad-CAM se calcula como:

$$L^c = \text{ReLU} \left( \sum_k \alpha_k^c A^k \right) \quad (2.1)$$

donde:

- $A^k$  es el mapa de activación para la capa convolucional  $k$ .
- $\alpha_k^c$  representa los pesos de importancia:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (2.2)$$

- $Z$  es el número de píxeles en el mapa de características.

Grad-CAM ayuda a visualizar las regiones discriminativas de clase en modelos basados en CNN (Convolutional Neural Networks). Grad-CAM resalta las regiones de la imagen más relevantes para una predicción de clase calculando los pesos de importancia  $\alpha_k^c$ , que se obtienen a través de los gradientes de la puntuación de la clase  $y^c$  con respecto a los mapas de características  $A^k$ . El mapa de calor final se genera aplicando ReLU a los mapas de características ponderados, asegurando que solo se visualicen las influencias positivas.

Según la literatura, los métodos pueden ser clasificados de otras formas, según el alcance de las explicaciones, se pueden distinguir dos tipos: las explicaciones globales y locales. Las explicaciones globales buscan describir el comportamiento general del modelo en todo el conjunto de datos, identificando patrones o reglas generales, por ejemplo, en un modelo que predice el precio de una casa, una explicación global podría mostrar que el tamaño de la casa y su ubicación son los factores más influyentes para la mayoría de las predicciones. Mientras, las explicaciones locales se centran en comprender por qué el modelo tomó una decisión específica para una instancia particular. Por ejemplo, para una solicitud de crédito rechazada, una explicación local podría mostrar que el bajo ingreso y la falta de historial crediticio fueron los factores clave para esa decisión [33].

### 2.1.3. Explicaciones basadas en instancias

Las explicaciones por ejemplos consideran la extracción de muestras que se relacionan con el resultado generado por cierto modelo, de manera similar a cómo actúan los seres humanos cuando intentan explicar un proceso dado. Las explicaciones por ejemplos se centran principalmente en extraer muestras representativas que capturan las relaciones internas y correlaciones encontradas por el modelo que se está analizando [5].

Sin embargo, en varias literaturas [5, 24] se exploran las explicaciones basadas en contraejemplos. Estas consisten en la extracción de muestras que se relacionan con el resultado generado por cierto modelo, pero de clases diferentes, por ejemplo al clasificar una imagen de un lobo negro, se pueden mostrar como contraejemplo imágenes de perros negros (clases parecidas, pero distintas). Estas explicaciones ofrecen soluciones llamativas ya que representan diferencias significativas que pueden dar aún más claridad a la explicación en ámbitos profesionales de alta importancia, otro ejemplo de esto podría ser un sistema de solicitud de préstamo con IA donde podría explicar su decisión basada en un ejemplo diciendo: “*Le rechazaron porque un cliente anterior con su perfil solicitó esta cantidad y también fue rechazado*”. En cambio, una explicación basada en un contraejemplo de la misma decisión podría decir: “*Si hubiera solicitado una cantidad ligeramente inferior, habría sido aceptado*”. Esta última explicación podría ser de mayor utilidad para el usuario.

## 2.2. Visual Question Answering (VQA)

Visual Question Answering (VQA, por sus siglas en inglés) es la tarea en la que un modelo de IA da como resultados respuestas a preguntas acerca de una imagen en concreto [8]. Existen modelos de VQA *knowledge-intensive* [8] que solucionan esta tarea usando modelos basados en redes neuronales convolucionales, entre otros. Sin embargo los enfoques más recientes se basan en modelos multimodales que a su vez utilizan Large Language Model (LLM, por sus siglas) para explotar la información visual y textual. Los primeros trabajos en comprensión de imágenes y lenguaje sentaron las bases para el VQA, Estos incluyen avances en la detección de objetos en 2001, el reconocimiento de escenas en 2010 y el modelado del lenguaje, como *word embeddings* en 2013 y *sequence modeling* en 2014 [40]. Estos desarrollos permitieron la extracción de características visuales y la representación textual, facilitando la integración de ambas modalidades. La formalización del VQA como un área de investigación distinta se produjo con la introducción del dataset VQA v1.0 [40], que proporcionó un benchmark a gran escala para la evaluación de modelos. A pesar de estos avances, el VQA enfrenta desafíos persistentes, como el sesgo del dataset, la interpretabilidad del modelo y la necesidad de un razonamiento robusto sobre tipos de

preguntas más complejos.

Goyal et al. [15] realiza un experimento donde plantea la hipótesis de que ciertos tipos de modelos (Deeper LSTM Question + norm Image (d-LSTM+n- I), *Hierarchical Co-attention* (HieCoAtt) y *Multimodal Compact Bilinear Pooling* (MCB)) han aprendido a explotar los sesgos de lenguaje en lugar de comprender verdaderamente el contenido visual de las imágenes. Aunque no se encontraba la popularidad creciente de los LLM, ya reflejaban que los tipos multimodales tenían mejores resultados, a pesar de estar segados en el texto de entrenamiento o evaluación, estos se enfocaban igualmente en las imágenes en mayor medida que el resto, para dar respuestas textuales.

Una de las primeras propuestas en el ámbito de los modelos multimodales basados en transformers es VisualBert [29] que implementó la idea de reutilizar el mecanismo de *Self-attention* del transformer para alinear implícitamente elementos del texto de entrada y regiones de la imagen de entrada. Es decir, además de todos los componentes de BERT, introducen un conjunto de incrustaciones visuales derivadas de un detector de objeto. Muy similar ViLBert [32] que utilizó una arquitectura basada en transformers para aprender representaciones conjuntas de imágenes y texto, que luego se podían utilizar para diversas tareas, incluido el VQA.

En [21] define a los modelos multimodales de este tipo como Large Visual Language Models (LVLM), modelos pre-entrenados en una gran cantidad de datos de imagen y lenguaje para aprender la representación conjunta de la visión y el lenguaje. Posteriormente, se ajustan en tareas de *Visual-Language* como VQA para mejorar la precisión. En la literatura [21] plantea 4 periodos del desarrollo del VQA:

- **Periodo I Mayo 2015 Agosto 2017:** Se usaban RNN-CNN (red neuronal que combina una Red Neuronal Recurrente (RNN) y una Red Neuronal Convolutiva (CNN)) con una fusión simple para crear modelos VQA, pero su rendimiento era deficiente. Entre 2015 y 2017, se propusieron modelos basados en atención visual que ponderaban la contribución de diferentes regiones de la imagen y la pregunta, superando a los modelos anteriores. Al mismo tiempo, se introdujeron técnicas de *bilinear pooling fusion* que capturaban mejor la interacción entre la imagen y la pregunta, mejorando los resultados. Además, se integraron estas técnicas de *pooling* en los modelos basados en atención. También, las redes relacionales y los modelos NMN (Neural Module Networks) lograron ciertos avances. En este período, la precisión en el conjunto de datos VQA v1.0 [3] osciló entre el 58 % y el 64 %.
- **Periodo II Agosto 2017 Agosto 2019:** Las máquinas de *Co-Attention* se propusieron para capturar la atención tanto visual como de la pregunta. Sin embargo, se recomendó que estos modelos de *Co-Attention* son limitados a capturar la relación palabra a palabra para las preguntas y la relación región a región para las imágenes. Por lo tanto, se utilizaron transformers para construir MCAN (Multi-head Co-Attention Network). El rendimiento en los conjuntos de datos VQA v1.0 y v2.0 [15] fue del 66 % al 71 %.
- **Periodo III Agosto 2019 Agosto 2021:** Se propusieron modelos VQA basados en transformers, pero los LVLM surgieron como la solución principal para VQA. Inicialmente, los LVLM usaban codificadores de visión basados en objetos, pero luego se reemplazaron con codificadores de parches basados en rejilla y en *vision transformer* (grid-based and ViT-based patch encoders) para construir modelos LVLM de

extremo a extremo. Los tamaños de estos modelos eran medianos (820-893 millones de parámetros) y la precisión en conjunto de datos VQA v2.0 varió de 71 % a 78 %.

- **Periodo IV Agosto 2021 presente:** Los LVLM dominan las tareas de lenguaje visual, incluido VQA. Los avances recientes han introducido LVLM expansivos de Google (SimVLM, Coca, PaLI), Microsoft (METER, VLMO, GIT, BEit-3) y DeepMind (Flamingo). Estos modelos, con entre miles de millones y decenas de miles de millones de parámetros (PaLI con 16.9B, Flamingo con 80.2B), alcanzan un rendimiento de primer nivel, con una precisión en el benchmark del conjunto de datos VQA v2.0 que oscila entre el 78 % y el 85 %.

En [8] explora y categoriza métodos interpretables en los que la VQA es un conjunto de subtareas y el resultado de cada paso se utiliza para explicar la respuesta, entre ellos:

- **Interpretable Methods for VQA:** Dividen el proceso de respuesta en pasos intermedios, lo que permite que cada etapa sea comprensible para los usuarios. Estos sistemas se centran en mostrar resultados parciales de cada fase del razonamiento, haciendo que el proceso sea más transparente. Por ejemplo, el modelo GVQA [2] organiza su funcionamiento en etapas como la identificación de partes relevantes de una imagen, la extracción de conceptos del texto de la pregunta y la clasificación del tipo de pregunta.
- **Knowledge-intensive Methods for VQA with Explainability:** Emplean fuentes de conocimiento externo para generar explicaciones, separando la representación del razonamiento en pasos distintos. Un enfoque común de tipo post-hoc es el uso de explicaciones basados en contraejemplos (counterfactual), como el propuesto en [52], que evalúan cómo cambiaría la respuesta del modelo VQA si se sustituyeran palabras en la pregunta. Otro ejemplo es VLC-BERT [46], que utiliza COMET (Commonsense Transformer) para explicar sus respuestas. COMET, entrenado en grafos de conocimiento como ATOMIC [49] y ConceptNet [51], predice posibles consecuencias o causas dado un evento o acción.

Específicamente en [15] se proponen explicaciones de contraejemplos para los modelos, las cuales muestran imágenes similares a la consulta, pero en las que la respuesta a la misma pregunta difiere entre ambas imágenes. Para la pregunta “¿Quién lleva gafas?” se mostró una imagen en la que hay dos personas (un hombre y una mujer), y es el hombre quien lleva gafas, el enfoque propone otra imagen en la que también aparecen dos personas (de nuevo un hombre y una mujer), pero en esta ocasión es la mujer quien lleva gafas.

Los enfoques VQA tienen el potencial de ser utilizados como técnicas de XAI para explicar imágenes. En la literatura existen algunos trabajos que consolidan el uso de modelos basados en transformers para mejorar la explicabilidad, por ejemplo, en el trabajo [44] se prueban modelos de LLM como Mistral [23] y GPT-4 basados en la arquitectura de transformers [59] para personalizar una explicación en lenguaje natural proporcionada junto con una explicación visual en la clasificación de fracturas de cadera [8].

En [8] se expone que la evaluación de la calidad de explicaciones ha mejorado con los años. Además, hay trabajos que proponen implementar herramientas LIME o GradCAM para generar explicaciones a partir de una imagen. Las evaluaciones cuantitativas mediante cuestionarios para recoger las opiniones de los usuarios sobre los sistemas de explicación

es otro enfoque común que se presenta en la literatura. El uso de cuestionarios con escalas Likert [4] que es similar a lo que encontramos en otros dominios de evaluación XAI, se pueden utilizar para permitir a los usuarios elegir cuál es la explicación preferida (entre la propuesta y baseline).

### 2.2.1. VQA + XAI

Las explicaciones textuales tratan de mejorar la explicabilidad de modelos de inteligencia artificial, donde se busca desarrollar técnicas que permitan generar explicaciones en lenguaje natural legible para el ser humano para justificar los resultados producidos por el modelo. Asimismo, las explicaciones visuales en modelos post-hoc tratan de visualizar el comportamiento de los modelos, comúnmente mediante técnicas de reducción de dimensionalidad que permitan una visualización sencilla e interpretativa. Las visualizaciones pueden integrarse con otros métodos para aumentar su claridad [5].

En [4] se emplea un modelo multimodal para tareas de VQA, SOTA-2017 con codificador ResNet [53] y además un codificador Mask-RCNN [18], entrenado con el conjunto de datos VQA2 [15]. El modelo procesa imágenes de  $224 \times 224$  píxeles y preguntas de hasta 15 palabras. ResNet codifica la imagen, mientras que un LSTM basado en Glove [42] codifica la pregunta. Se elabora un sistema que denomina XVQA (explainable VQA), y tiene como objetivo explicar el comportamiento del “*agente*” de VQA combinando las características de atención generadas en el modelo de VQA con anotaciones significativas de los datos de entrada. Estas anotaciones incluyen etiquetas, descripciones y cuadros delimitadores de entidades en la escena y sus conexiones entre sí. Dicho método XVQA visualiza información de las capas internas del modelo de VQA o incorpora esa información con anotaciones para explicar el funcionamiento interno del modelo. Las explicaciones se proporcionan en diferentes combinaciones a los subgrupos de participantes del estudio para evaluar su efectividad en la predicción.

Alipour et al. [4] ilustra los distintos componentes explicativos o modos de explicación que se combinan para el método XVQA, se detallan a continuación:

- **Spatial attention:** Es una técnica que muestra visualmente en qué partes de la imagen se está enfocando el modelo al responder una pregunta. El sistema genera un “mapa de atención” que resalta las regiones más relevantes de la imagen según la pregunta formulada. Así, el usuario puede ver qué zonas influyen más en la respuesta generada por el modelo, facilitando la comprensión de su razonamiento visual. Se calcula las atenciones basado sobre las características de la imagen en las capas de ResNet y la pregunta formulada. En la Figura 2.1 se muestra una visión general.

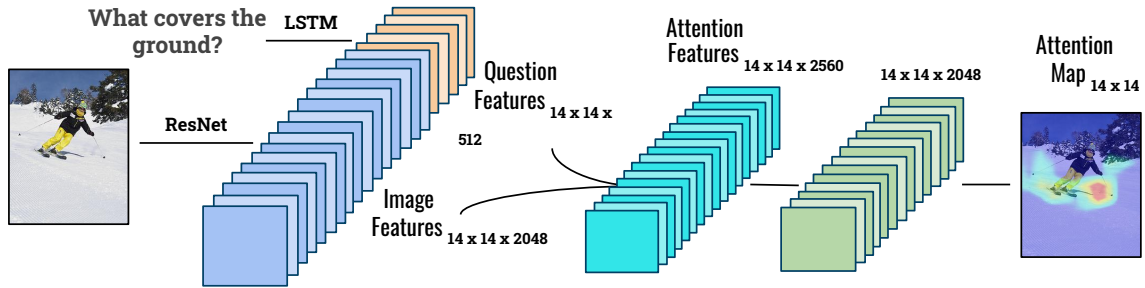


Figura 2.1: Mapa de atención generado a partir de las características de entrada. Fuente: [4]

- **Active attention:** Consiste en un bucle de retroalimentación. Los usuarios pueden utilizar esta función para alterar el mapa de atención de un modelo y dirigir la atención del modelo y la forma en que se genera la respuesta.

La *Active attention* tiene una tarea de dos pasos que completar. El primer paso es muy similar a las pruebas de *Spatial attention*, donde los usuarios hacen su predicción basada en el mapa de atención generado por el modelo VQA. Luego, el sujeto observa los resultados de la predicción y se da cuenta de si el sistema es preciso o no. En el segundo paso, se le pide al sujeto que dibuje un nuevo mapa de atención. Usando el mapa de atención dibujado manualmente, el modelo procesa la imagen y la pregunta una vez más y produce una segunda respuesta.

El propósito de esta operación es permitir que el sujeto se involucre en el proceso de inferencia y proporcione retroalimentación al modelo de manera interactiva. En los casos en que el modelo responde incorrectamente a las preguntas, los sujetos intentan corregir la respuesta del modelo dibujando la atención. En la Figura 2.2 se muestra una visión general.

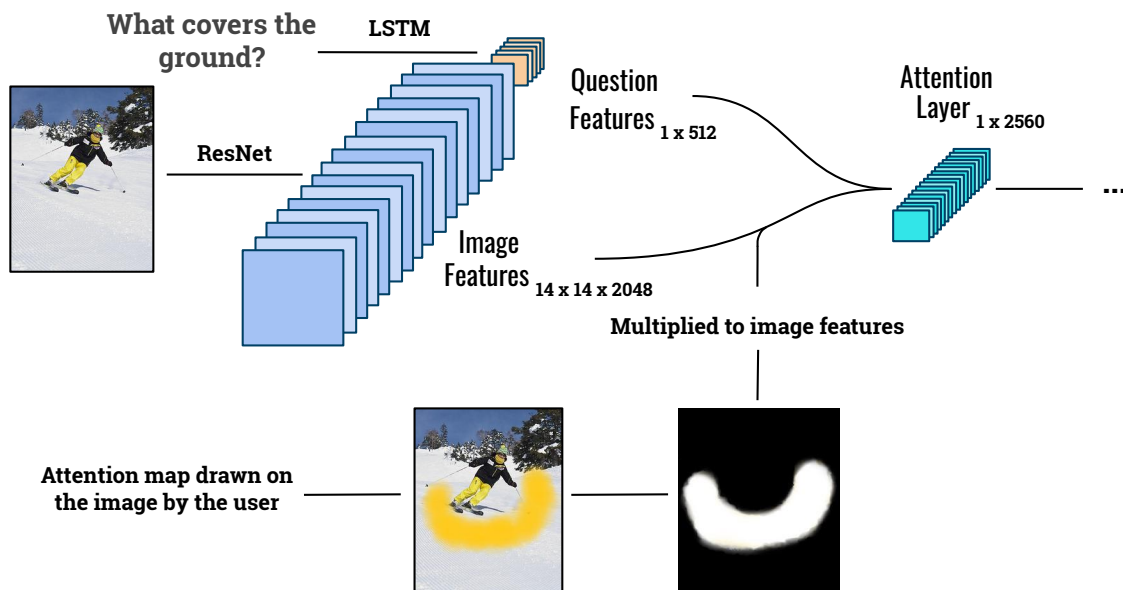


Figura 2.2: Arquitectura del bucle de active attention dentro del sistema XVQA. Fuente: [4]

- Bounding Boxes:** Las cajas delimitadoras se generan en función de las anotaciones en el conjunto de datos como Visual Genome [25]. Estas cajas encierran objetos relevantes en la imagen y se combinan con los mapas de atención para identificar los objetos que más han influido en la respuesta del modelo. El sistema selecciona las cajas con mayor peso de atención, mostrando así los elementos clave considerados por el modelo. En la Figura 2.3 se muestra como están generadas las cajas delimitadoras.

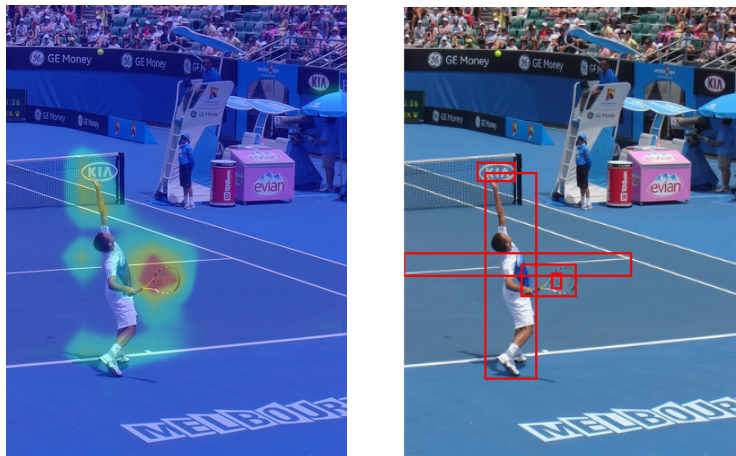


Figura 2.3: Explicaciones de las cajas delimitadoras generadas en función de los pesos de atención espacial para la pregunta “¿Qué está haciendo el hombre?”. Fuente: [4]

- Scene graph:**

El grafo de escena complementa las cajas delimitadoras mostrando las relaciones entre los objetos identificados en la imagen. Cada nodo representa un objeto y los enlaces describen relaciones (por ejemplo, "hombre-sostiene-guitarra"). El sistema filtra el grafo de escena según los pesos de atención (Figura 2.4). Los usuarios pueden localizar interactivamente los objetos activos del grafo de escena y ver sus cajas delimitadoras en la imagen de entrada

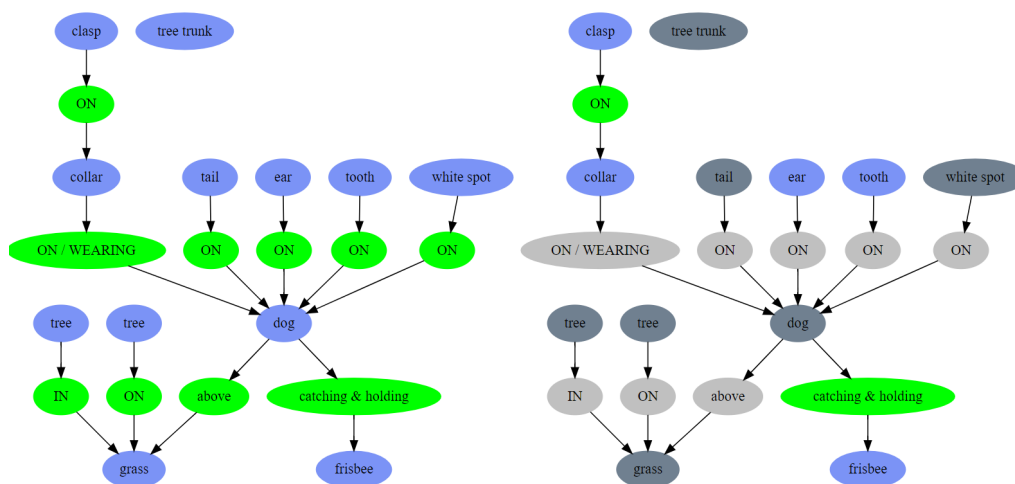


Figura 2.4: Izquierda: Escenario de entrada. Derecha: Escenario filtrado según las ponderaciones del mapa de atención generado por el modelo en respuesta a una pregunta. Fuente: [4]

- **Object attention:**

Este modo utiliza un codificador especializado (MASK-RCNN) para generar máscaras de atención a nivel de objeto. A diferencia de la atención espacial, que es más difusa, la atención a objetos permite segmentar entidades específicas en la imagen, mostrando con precisión qué objetos han sido más determinantes en la inferencia del sistema.

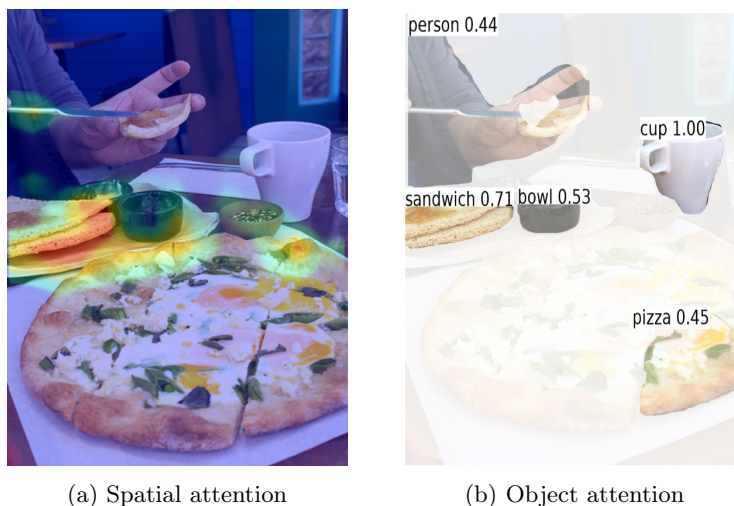


Figura 2.5: (b) Atención a nivel de objeto en comparación con (a) atención espacial. Fuente: [4]

- **Textual explanation:** Junto con explicaciones visuales, este sistema XVQA incorpora explicaciones en lenguaje natural para ayudar a los usuarios a entender las decisiones del modelo de VQA.

Para un par de pregunta-imagen dado, el módulo de explicación textual utiliza el mapa de atención visual para identificar las partes más relevantes de la imagen. El modelo luego recupera los cuadros delimitadores de las entidades que se superponen en gran medida con estas regiones. Finalmente identifica aquellas entidades más relevantes para la respuesta en función de su relevancia espacial en la imagen y su representación en lenguaje natural. Las descripciones de las regiones para las entidades más relevantes forman las explicaciones textuales. En la Figura 2.6 se muestra una salida generada por este componente.

En Muddamsetty et al. [34] se establece un método denominado SIDU-VQA que implica una combinación de modelos de imagen y texto (CNN + LSTM) y se indica que es la primera vez que el método SIDU se aplica a modelos para tareas de VQA. Este es una adaptación del enfoque post-hoc de explicabilidad SIDU [35] al contexto de VQA. SIDU (Similarity Difference and Uniqueness), es un método diseñado para explicación de modelos de clasificación de imágenes originalmente, y que en este contexto de VQA su implementación es mucho más complejo porque la región de la imagen en que se enfoca el modelo depende de la pregunta planteada.

SIDU-VQA utiliza la última capa convolucional de la CNN del modelo VQA para generar máscaras de activación de características (*feature activation masks*). Dichas máscaras representan las regiones de la imagen que el modelo considera importante para la respuesta. Para las máscaras de activación, se calculan los pesos de importancia utilizando las



Figura 2.6: Ejemplo de resultado del modulo de lenguaje natural realizando una explicación textual para la respuesta original del modelo Fuente: [4]

predicción del modelo VQA sobre el par imagen-pregunta. Los pesos están basados en el cálculo matemático [34] de la diferencia de similitud y la unicidad (*similarity difference and uniqueness weights*), lo que permite identificar qué regiones son más decisivas para la respuesta final.

Finalmente los pesos de importancia se aplican sobre las máscaras de activación para producir mapas de calor (*heatmaps*) que resaltan las áreas clave de la imagen en función de la pregunta y que sirven como explicaciones visuales de las respuesta del modelo. Un aporte mencionado es que a diferencia de otros métodos (Grad-CAM) SIDU-VQA propone una alternativa novedosa que no depende de los gradientes (*gradient-free*) de la red para identificar pares relevantes que influyeron en la respuesta y evitar el problema *gradient saturation*.

En 2025 Li et al. [28] plantea un modelo determinado MRVQA (Multimodal Rationales for Visual Question Answering). Es un modelo diseñado para tareas VQA, además, MRVQA genera explicaciones textuales y visuales que justifican su respuesta. MRVQA esta basado en una arquitectura de tipo transformer.

Los componentes que desarrolla Li et al.:

- **Input Representtation:** Se utiliza el mismo modelo (CLIP) para representar la imagen y pregunta de entrada de forma simultánea.
- **Encoder-Decoder:** Para comprender eficazmente a partir de las características visuales y lingüísticas, es esencial aprovechar la fusión entre modalidades en las tareas de VQA. Para mejorar representaciones visuales y lingüísticas sin introducir nuevos módulos ni estructura del modelo, se basa en MCAN [60] que combina varias capas de *Co-Attention Modular* (MCA). La capa MCA comprende unidades de *self-attention* (autoatención) y unidades de *guided-attention* (atención guiada). Las primeras unidades capturan *intra-modal features* (características intramodales), mientras que las segundas permiten *cross-modal interactions* (interacciones intermodales). Se emplea la estructura y los parámetros predeterminados de MCAN-base para construir un modelo sencillo y limpio para tareas EVQA (Explainable VQA).

- **Projection Module:** El módulo de proyección esta basado en transformer y su objetivo es alinear las dimensiones de las representaciones de imagen y pregunta de CLIP con el LLM. El modulo utiliza el mecanismo de self-attention del transformer para adaptar las características de CLIP según el contexto específico de cada par imagen-pregunta, a diferencia de otras capas de proyecciones.
- **Predictors:** MRVQA utiliza tres “predictors” que puedes resumirse como los siguientes:
  - Predicción de respuestas: Lo aproxima como un problema de clasificación para las respuestas, en concreto, se utiliza un módulo de fusión lineal, seguido de una función sigmoidea, para para procesar las características de *cross-attended* (atención cruzada) del decoder.
  - Textual Rationales (explicación textual): Para los razonamientos textuales, a diferencia de los métodos anteriores que se basan en LSTM sencillos, se utiliza el modelo autoprogresivo de lenguaje GPT-2 avanzado para generar razonamiento textual más robusto y relevante.
  - Visual Rationales (Explicación visual): Se genera cuadros delimitadores precisos a nivel de objeto utilizando GroundingDINO [31], un detector de objetos de *open-vocabulario* capaz de identificar objetos arbitrarios a partir del lenguaje natural. El enfoque introduce al detector la imagen, características de *cross-attended* y las características procesadas por el LLM y establece delimitadoras detectadas como una única categoría, distinto de las prácticas estándar de anclaje visual. Dicha configuración borda el desafío de alinear objetos con la pregunta de entrada cuando faltan descripciones específicas, por ejemplo cuando se trata de una pregunta genérica como “¿qué forma tiene el objeto?”, en algunos modelos falta de contexto explicito pueden dar resultados erróneos.

El autor afirma que con la incorporación de varios predictores, el modelo proporciona resultados robustos para VQA explicable.

El entrenamiento del modelo MRVQA se lleva a cabo mediante una función de pérdida diseñada para equilibrar y alinear las tres salidas principales: la respuesta, la explicación textual y la explicación visual. La función de *Total Loss* (pérdida total) se establece como la suma ponderada de las pérdidas de los predictores, es decir, la pérdida de predicción de respuestas, la pérdida de las explicaciones textuales y la pérdida de las explicaciones visuales. En la Figura 2.7, se muestra una vista general del modelo propuesto.

### 2.3. Case-Based Reasoning (CBR)

Es definido básicamente como un enfoque que resuelve un nuevo problema recordando una situación similar anterior y reutilizando información y conocimiento de esa situación. El primer sistema que podría llamarse razonador basado en casos fue el sistema CYRUS, desarrollado por Janet Kolodner en la Universidad de Yale (grupo de Schank). CYRUS se basó en el modelo de memoria dinámica de Schank y la teoría MOP de resolución de

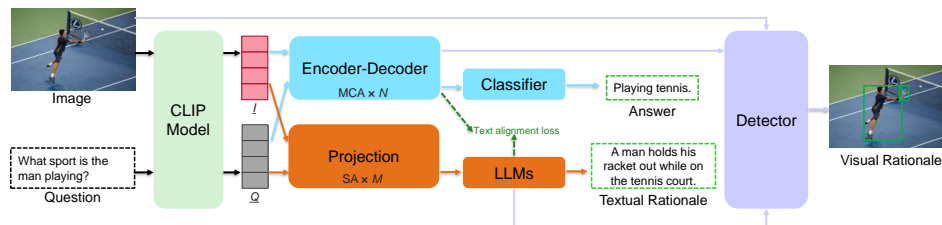


Figura 2.7: Arquitectura del modelo MRVQA. Fuente: [28]

problemas y aprendizaje. Básicamente, era un sistema de preguntas y respuestas con conocimiento sobre los diversos viajes y reuniones del ex Secretario de Estado de EE.UU., Cyrus Vance. El modelo de memoria de casos desarrollado para este sistema ha servido posteriormente como base para varios otros sistemas de razonamiento basado en casos (incluyendo MEDIATOR, PERSUADER, CHEF, JULIA, CASEY) [1].

En Europa, la investigación sobre CBR se inició un poco más tarde que en los EE.UU. El trabajo en CBR parece haber estado más vinculado al desarrollo de sistemas expertos y a la investigación en adquisición de conocimiento. Entre los primeros resultados estuvo el trabajo en CBR para diagnóstico técnico complejo dentro del sistema MOLTKE, realizado por Michael Richter junto con Klaus Dieter Althoff y otros en la Universidad de Kaiserslautern [1].

Algunos ejemplos que ilustran Aamodt y Plaza [1] en la resolución de problemas por medio de CBR, pueden ser:

- Un médico después de haber examinado a un paciente en su consulta recibe un recordatorio de un paciente que trató hace dos semanas. Suponiendo que el recordatorio fue causado por una similitud de síntomas importantes (y no por el color de cabello del paciente, por ejemplo), el médico utiliza el diagnóstico y tratamiento del paciente anterior para determinar la enfermedad y el tratamiento del paciente que tiene frente a él.
- Un ingeniero de perforación, que ha experimentado dos situaciones dramáticas de erupción, rápidamente recuerda una de estas situaciones (o ambas) cuando la combinación de mediciones críticas coincide con la de un caso de erupción. En particular, puede recordar un error que cometió durante una erupción anterior y usar esto para evitar repetir el error una vez más.

En [54] plantea el ciclo esquemático de un sistema CBR. En la Figura 2.8 se puede observar cada una de las fases:

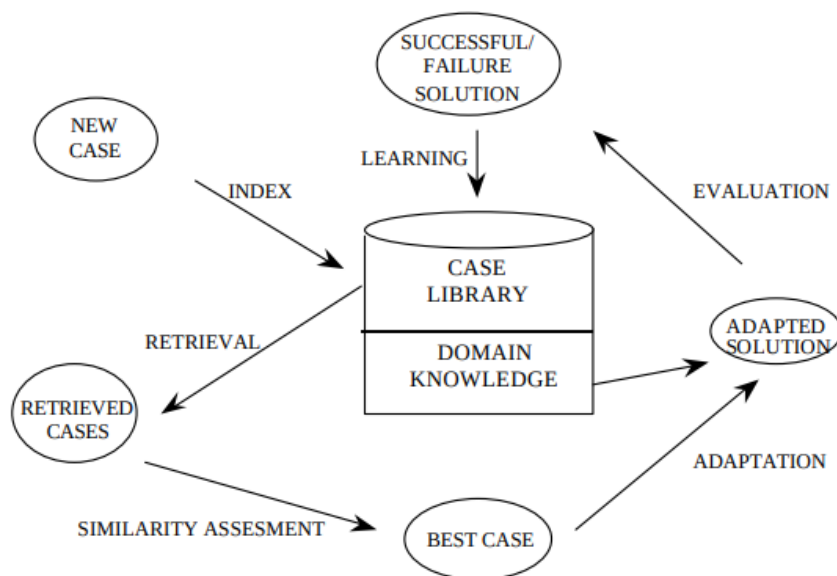


Figura 2.8: Ciclo sistema CBR. Fuente: [54]

Cada fase del ciclo la define como:

- **Retrieve** (Recuperar): Recuperar el caso o los casos más similares al nuevo caso de consulta. A diferencia de los sistemas tradicionales, donde la recuperación suele hacerse mediante coincidencia exacta, en CBR se emplean estrategias de coincidencia parcial, ya que raramente un caso previo será idéntico al nuevo, el objetivo es maximizar la similitud entre el caso actual y los recuperados, lo que suele requerir el uso del dominio del conocimiento (Domain Knowledge) general para guiar la búsqueda y escoger el mejor caso.
- **Reuse** (Reutilizar): Cuando el caso recuperado no es idéntico al problema actual, la solución almacenada debe ser adaptada para que sea aplicable y efectiva en la nueva situación. La adaptación puede ocurrir durante la formulación inicial de la solución o después de recibir retroalimentación que indique que la solución propuesta necesita corrección (Repair).
- **Revise** (Revisar): Puede definirse como el proceso de evaluar la calidad o el rendimiento de la solución propuesta para el nuevo caso. Comúnmente, este paso de evaluación se puede realizar ya sea preguntando a un experto humano (*Oracle*) si la solución es buena o no, o simulando los efectos de la solución propuesta en el mundo real obteniendo directamente una retroalimentación sobre los resultados de la solución propuesta.

- **Retain** (Retener): El aprendizaje en los sistemas basados en casos consiste, principalmente, en guardar los nuevos casos que se han resuelto, para enriquecer la base de casos y poder usarlos en el futuro.

La convergencia de CBR en los marcos XAI ofrece un enfoque estructurado para generar explicaciones. Esto implica identificar casos pasados que se asemejan al contexto de decisión actual y proporcionar una narrativa o un razonamiento que los usuarios puedan seguir y comprender. Por ejemplo, en un sistema de IA para diagnóstico médico, una explicación basada en CBR podría detallar cómo la recomendación de la IA coincide o difiere de diagnósticos previos en condiciones similares de pacientes, fundamentando así la decisión de la IA en ejemplos concretos y comprensibles [54].

Otro ejemplo de ello es iSee (*Intelligent Sharing of Explanation Experience*), que emplea una metodología de Razonamiento Basado en casos (CBR), no para obtener explicaciones, sino para recomendar experiencias de explicaciones. Cuando se le consulta, el sistema recurre a una base de casos de experiencias de explicación para sugerir una estrategia de explicación apropiada. El consorcio iSee está formado por investigadores que abogan por utilizar el paradigma CBR para captar el conocimiento y la experiencia adquirida en la adaptación con éxito de la explicabilidad en los sistemas de IA. iSee aprovecha estas experiencias para ayudar a los sistemas de IA a crear una explicabilidad que cumpla la normativa, como el derecho a la explicación de la UE<sup>1</sup> [43].

En la Figura 2.9 se muestra un esquema general de iSee donde se observa como base un ciclo CBR.

---

<sup>1</sup>Disponible en <https://gdpr.eu/>

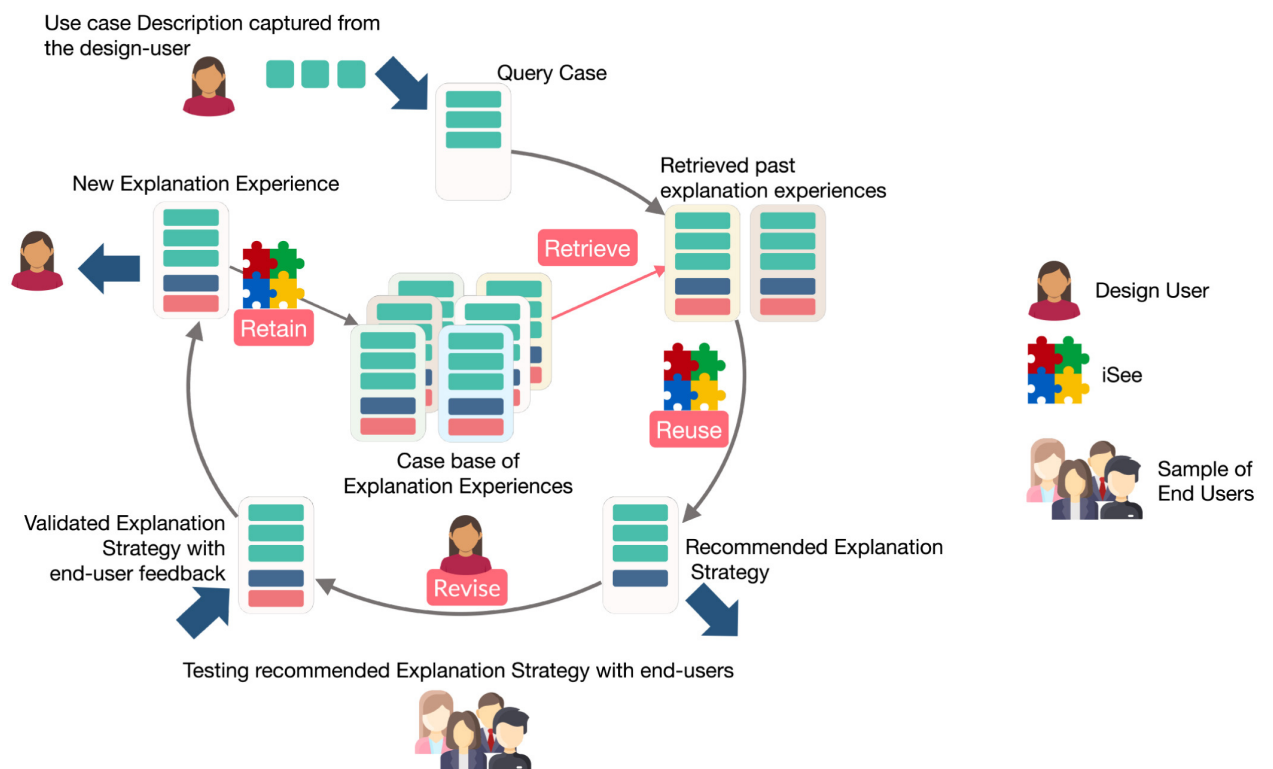


Figura 2.9: Metodología CBR de iSee. Fuente: [43]

## Método de explicación basado en CBR y VQA

Las explicaciones basadas en casos, cuando se generan a través de un sistema CBR, son consideradas actualmente una de las formas de explicación más exitosas, ya que se aproximan a lo que sería el funcionamiento de un modelo mental del usuario. Por ejemplo, en la clasificación de imágenes, muestran un conjunto de imágenes similares a la imagen clasificada (considerada consulta), de esta forma el usuario puede ver las diferencias y a su vez comparar las clases de dichas imágenes. En este trabajo nos enfocamos en explicaciones basados en ejemplos (factual) y basados en contraejemplos (counterfactual).

Las explicaciones basados en ejemplos (factual) en la clasificación de imágenes, se enfocan en proporcionar imágenes similares a la imagen a explicar (consulta) y que debe ser de la misma clase, por ejemplo, si la imagen a explicar es clasificada como un lobo, las imágenes similares deben ser clasificada como un lobo. En las explicaciones basados en contraejemplos (counterfactual) ocurre de forma similar, con la distinción de que las imágenes similares deben ser imágenes con clasificación distinta al de la consulta, por ejemplo, si la imagen a explicar es una vaca, las imágenes similares podrían ser clasificadas como corderos u otro tipos de animales que se asemejen. Para obtener las imágenes similares, los métodos de explicación basados en ejemplos (factual) o basados en (counterfactual) comúnmente utilizan métricas de similitud basadas en imágenes, como la similitud píxel to píxel una de la más populares.

Nuestro método tiene como objetivo explotar los enfoques de procesamiento de lenguaje natural (NLP) y el alto rendimiento que tienen en la actualidad para obtener explicaciones basadas en casos. Primero, el método obtiene descripciones completas de nuestras imágenes con el objetivo de construir la base de casos, la tarea esencial es encontrar los mejores inputs (preguntas) para nuestro modelo VQA que consiga describir de la mejor forma posible estas imágenes. Segundo, utilizamos métricas de similitud basadas en texto o semántica para encontrar las imágenes similares a la que sea presentada como la imagen consulta. Esta tarea se realiza en el paso de recuperación del CBR. La hipótesis que planteamos es que los casos de explicación deberían ser más útiles a los usuarios para entender la clasificación de imágenes ya que nuestro método podría encontrar similitudes con más precisión al utilizar las descripciones textuales. Tercero, denominado también como la reutilización, consiste en personalizar la explicación, permitiendo que los usuarios decidan cuál es el objetivo XAI,

mostrando solo las descripciones relacionadas con el objetivo XAI que se requiere cumplir para los usuarios objetivo. Resaltamos el hecho de que nuestro método de explicación es post-hoc y agnóstico al modelo, ya que todo el proceso de CBR realizado es independiente del proceso del modelo de IA que generó la predicción. En la Figura 3.1 se presenta una vista general del esquema planteado en este trabajo.

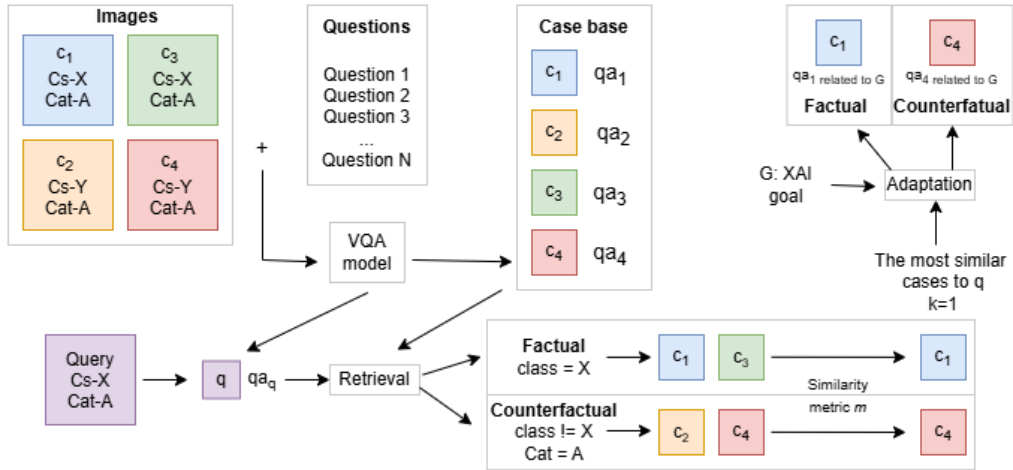


Figura 3.1: Descripción general del método de explicación basado en casos de VQA donde  $C_s$  denota la clase,  $Cat$  categoría de la imagen según el agrupamiento que se ha propuesto en este trabajo y  $qa$ , las descripciones que contienen los pares de preguntas y respuestas.

A continuación, se detalla el proceso completo en tres secciones: cómo crear la base de casos (Sección 3.1), cómo recuperamos las explicaciones (Sección 3.2) y cómo personalizamos dichas explicaciones (Sección 3.3).

### 3.1. Base de casos VQA

#### 3.1.1. Questions

El principal objetivo es constituir una base de casos donde la representación de las imágenes en formato textual sea lo más detallada posible estructuradas en pares de preguntas y respuestas. Para ello se debe incluir una descripción no solo de la entidad principal sino también el contexto, como el el background de la imagen, luminosidad, volumen de ruido u descripciones de otros elementos menos importante en la imagen, o incluso interacciones entre los elementos. En nuestra base de casos hemos dividido las imágenes en cuatro categorías principales, Human, Animal, Object y Landscape, y dentro de ellas las distintas clases correspondientes, por ejemplo, en la categoría Animal se encuentran imágenes de perros, gatos, leones, tigres, etc. Por tanto hay que considerar que por la naturaleza de las imágenes no todas las preguntas pueden ser formuladas para todas la imágenes. Por ejemplo, no podemos formular preguntas de géneros masculino o femenino cuando la única y principal entidad en la imagen se trata de un reloj. Por tanto tenemos preguntas generales para hacer en todos los casos y otras preguntas específicas que solo son aplicables dependiendo de la categoría de la imagen.

Además de ello, hemos incluido conocimientos sobre objetivos XAI comentados en la

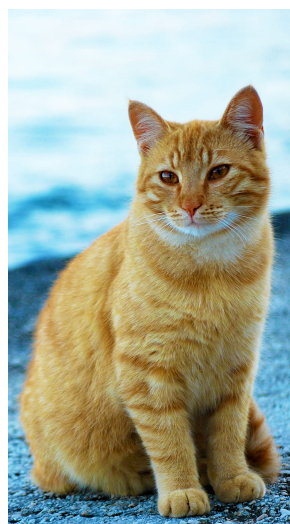
literatura, tales como, *debugging*, *effectiveness*, *efficiency*, *trust*, *persuasiveness*, *satisfaction*, *education* y *transparency*. Con esto se espera ayudar al usuario a entender mejor la predicción. Por ejemplo, para el objetivo educativo podemos preguntar: “¿Puedes proporcionar conceptos esenciales que sean relevantes para la situación descrita?” (“*¿Can you give essential concepts that are relevant to the situation described?*”) y para mejorar la confianza de los usuarios, podemos preguntar: “¿Hay algo que pueda generar dudas o incertidumbre para que digas que son  $\{clase\}$ ?” (“*¿Is there anything that might create doubt or uncertainty for you to say they are  $\{subject\_extracted\}$ ?*”).

Por lo expuesto anteriormente, el propósito de las preguntas es obtener conocimientos que, primero describan, que elementos se muestran en la imagen, por ejemplo la pregunta “Tiene la  $\{clase\}$  algún elemento que lo caracterice? cuales?” (“*¿Do  $\{subject\_extracted\}$  have any elements that characterize them? Which ones?*”), segundo, describan la perspectiva o como se muestra la imagen (colores, luminosidad y calidad de la imagen), por ejemplo, “¿La imagen es luminosa?” (“*¿Is the image bright?*”) y tercero, describan las imágenes en términos de conocimientos que puedan explotar los objetivos XAI, por ejemplo, “¿Te enfocas sobre las partes de  $\{clase\}$  para inferir que la entidad principal de la imagen es un  $\{clase\}$ ?” (“*¿Do you focus on parts of the  $\{subject\_extracted\}$  to infer that the main entity in the image is a  $\{subject\_extracted\}$ ?*”).

Para formular estas preguntas, un ejercicio común es presentar dos imágenes que, a simple vista, parecen tener la misma entidad principal y descripciones similares. Luego, se deben pensar en las preguntas necesarias para identificar las diferencias que están presentes, pero que no se detallarían sin un análisis minucioso. Para ser más preciso estas preguntas están formuladas en inglés, en Apéndice A se puede apreciar una lista detallada de las preguntas por categoría y objetivo. En el directorio *Ejemplos Chat con imágenes*<sup>1</sup> se observa algunos ejemplos del proceso iterativo llevado a cabo para crear el conjunto de preguntas.



(a) Gato amarillo en bosque.  
Fuente<sup>2</sup>



(b) Gato amarillo en playa.  
Fuente<sup>3</sup>

Figura 3.2: Ejemplo de gatos idénticos pero con descripciones distintas.

<sup>1</sup><https://github.com/angelldavid/Trabajo/tree/main>

Por ejemplo, en la Figura 3.2 a simple vista se trata de dos gatos color amarillo, pero podemos aprovechar las diferencias del contexto de la imagen, como el color, el fondo, la posición de la cola, la posición del cuerpo y cabeza del gato con respecto a la captura de la foto. Allí es donde se encuentra la esencia donde se tendrá que plantear las preguntas que puedan aprovechar estas pequeñas distinciones.

### 3.2. Recuperación (Retrieval)

La recuperación de casos es una etapa fundamental en los métodos CBR. En el contexto de nuestro sistema, consiste en identificar, dentro de la base de casos previamente construida, aquellos ejemplos que son más similares a la instancia que se desea explicar, tanto para casos de explicación basados en ejemplos (factual) y basados en contraejemplos (counterfactual). Para proceder a la recuperación primero describiremos formalmente los componentes de nuestra base de casos de la siguiente forma:

La estructura de los casos en la base de casos  $C$  es la siguiente:

$$\forall c_i \in C \rightarrow c_i = \langle D_i, S_i \rangle$$

$c_i$  representa un caso en nuestra base de casos.  $D_i$  (descripción) representa una imagen  $i$  y su clase obtenida con un modelo de IA.  $S$  (solución) representa la explicación que se mostrará al usuario objetivo para explicar esa clasificación.

La formalización de la descripción para un caso  $c_i$  en nuestra base de casos es la siguiente:

$$D_i = \langle r_i, cs_i, cat_i, qa_i \rangle$$

donde  $r_i$  es la representación visual del caso (es decir, la propia imagen);  $cs_i$  es la clase predicha por el modelo de IA para la imagen;  $cat_i \in \{Human, Animal, Object, Landscape\}$  es la categoría de la clase; y  $qa_i = \{(q_k, a_k), \dots, (q_j, a_j)\}$  es el conjunto de pares de pregunta y respuesta que describen la imagen en formato de texto. Las preguntas  $\{q_k, \dots, q_j\}$  son aquellas realizadas al modelo VQA, que generará las respuestas.

Alternativamente, la formalización de la solución es la siguiente:

$$S_i = \langle ec_i, t_i \rangle$$

donde  $ec_i$  es el conjunto de casos de explicación obtenidos en el paso de recuperación, y  $t_i$  es la explicación textual personalizada que soporta los casos de explicación, obtenidos en el paso de reutilización.

Sintetizando, Partimos de nuestra base de casos  $C$  y la imagen a explicar  $q$ . Cada caso  $c_i \in C$  que podemos denotar como  $c_i = \langle (r_i, cs_i, cat_i, qa_i), (ec_i, t_i) \rangle$ , donde  $q$  tiene una representación  $r_q$  clasificada como la clase  $cs_q$  dentro de la categoría  $cat_q$ . La traducción a la representación visual de lenguaje natural usando un modelo VQA es denotado

<sup>2</sup>[https://glosbe.com/fb\\_img/980x980/yh4984573\\_images-6-.jpeg.cvrt.jpg.cvrt.webp](https://glosbe.com/fb_img/980x980/yh4984573_images-6-.jpeg.cvrt.jpg.cvrt.webp)

<sup>3</sup><https://w0.peakpx.com/wallpaper/638/82/HD-wallpaper-big-yellow-cat-animl-big-cat-cats-pet-yellow-cat.jpg>

como  $qa_q$ , es decir, el conjunto de preguntas y sus correspondientes respuestas que describen la imagen  $q$ , los casos de explicación  $ec_i$  y el texto descriptivo de dicha explicación  $t_i$ .

Planteado lo anterior, para la recuperación vamos a realizar un esquema MAC/FAC (many-are-called, few-are-chosen), utilizada en la literatura sobre CBR [9]. Los pasos se detallan a continuación:

- **Filtrado (*filtering step*, MAC).** Se seleccionan los casos desde la base de casos  $C$  que pueden ser casos de explicación. Dependiendo del tipo de explicación esto cambiará, si se desea recuperar explicaciones por ejemplos, en este filtro el método seleccionará aquellos casos en los que  $cs_i = cs_q$ : siendo la clase de los casos de explicación la misma que la clase de la imagen a explicar (query). Por el contrario, si necesitamos obtener explicaciones basados en contraejemplo (counterfactual), el paso de filtrado va a recuperar, como explicaciones, aquellas imágenes de la base de casos que no comparten su clase, pero sí su categoría con la imagen a explicar (query).

$$f_q = filter(q, C, expType) = \forall c_i \in C : \begin{cases} cs_i = cs_q, & \text{if } expType = factual \\ cs_i \neq cs_q \& cat_i = cat_q, & \text{if } expType = cFactual \end{cases}$$

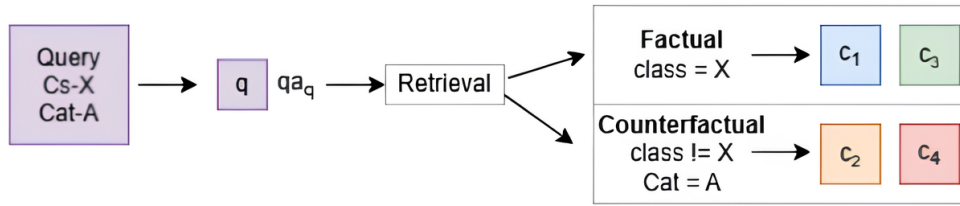


Figura 3.3: Filtering step

- **Clasificación (*sorting*, FAC).** En este paso, aplicamos métricas de similitud para obtener los  $k$  casos más similares a  $q$  a partir de  $f_q$ . En otras palabras,  $k$  representan las imágenes más similares a la imagen a explicar (query) desde el filtrado realizado, dependiendo de la clasificación categoría y tipo de explicación (ejemplo o contraejemplo) extraído previamente en el MAC. Se puede formalizar de la siguiente forma:

$$ec_q = first(sorting(q, f_q), k)$$

$$sorting(q, f_q) = \{c_i, \dots, c_j \in f_q \mid s_{qi} \geq s_{qj}, \forall c_k \in f_q : s_{qk} = sim_m(q, c_k, m)\}$$

Donde  $first(set, k)$  es una función que devuelve los primeros  $k$  elementos del conjunto y  $sim_m()$  es una función que devuelve la puntuación de similitud entre  $q$  y el caso  $c_k$  utilizando la métrica de similitud  $m$ . Dado que nuestro método describe los casos mediante conocimiento textual, hemos propuesto tanto métricas de similitud basadas en texto como métricas de similitud semánticas como la métrica  $m$  a utilizar en nuestro método. Para cada categoría, las respuestas a las preguntas para cada caso (imagen) serán comparadas entre sí (pregunta a pregunta) con dichas métricas, para luego tomar el promedio de los valores de cada comparación correspondiente a cada pregunta. Este valor promedio será el valor que representará la similitud entre dos casos (imágenes).

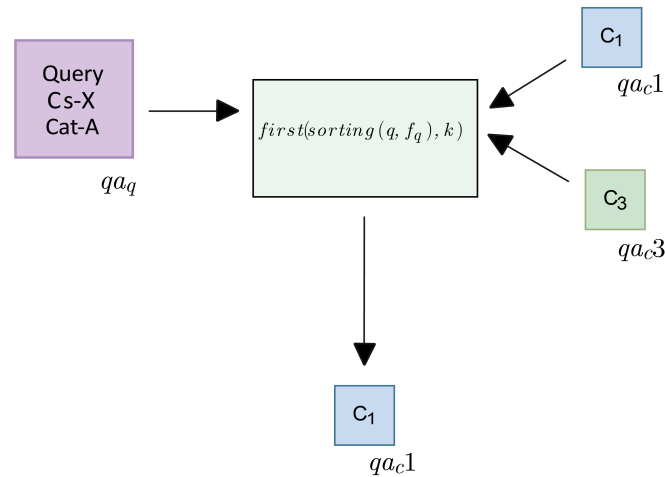


Figura 3.4: Sorting step: Dado un consulta y una base de casos, se recuperan los casos en MAC y se realiza una ordenación por valor de similitud y se escogen los  $k$  casos más similares

### 3.2.1. Métricas de similitud

Las métricas de similitud basadas en texto utilizadas en nuestro método son las siguientes:

- **Levenshtein distance:** La distancia de Levenshtein es un algoritmo útil en procesamiento de lenguaje natural para medir la similitud entre secuencias de caracteres. Desarrollada por el matemático soviético Vladimir Levenshtein en 1965 [27], esta métrica cuantifica el número mínimo de operaciones requeridas para transformar una cadena en otra, donde las operaciones permitidas son inserciones, eliminaciones o sustituciones de caracteres individuales. Matemáticamente entre mayor sea el resultado de esta métrica la diferencia entre textos es más alta.
- **Cosine similarity:** Esta métrica transforma el texto en vectores de palabras, de esta forma mide la similitud entre dos vectores calculando el coseno del ángulo entre los dos vectores [58]. El rango de valores es de -1 a 1, donde 1 indica mayor similitud.
- **Overlap coefficient or Szymkiewicz–Simpson coefficient:** Es calculada como el número de tokens comunes que contienen ambos textos dividido por el número de tokens que tiene el texto más corto [22]. El rango de valores es de 0 al 1, donde 1 indica la mayor similitud.
- **Bag distance:** Tiene en cuenta el número de apariciones de diferentes palabras en ambos textos sin tener en cuenta su orden. Como resultado, ofrece mayor diferencia entre el número de veces que aparecen las palabras en ambos textos [6]. Matemáticamente entre mayor sea el resultado de esta métrica la diferencia entre textos es más alta.
- **Jaccard:** Calcula la similitud entre dos frases como la intersección de palabras dividida por la unión de las palabras. Cuando las frases son notablemente largas, la similitud tiende a ser menor [10]. El rango de valores es de 0 al 1, donde 1 indica máxima similitud.

Estas métricas pueden ser implementadas mediante la librería “*TextDistance*”<sup>3</sup> que ofrece las clases Levenshtein, Cosine, Overlap, Bag y Jaccard.

Las métricas de similitud basadas en semántica utilizadas en nuestro método son las siguientes:

- **Web similitary**: Se ha utilizado la librería “*semantic-text-similarity*”<sup>4</sup> y la clase “*WebBertSimilarity*”, que emplea modelos BERT (Bidirectional Encoder Representations from Transformers) [12] ajustados para detectar similitudes entre textos. En particular, el modelo ha sido entrenado con el conjunto de datos STS-B (Semantic Textual Similarity Benchmark)<sup>5</sup>, lo que le permite alcanzar una correlación de 0.893 en datos de desarrollo. La escala de similitud varía de 0 a 5, donde un valor de 0 indica que las frases comparadas no tienen ninguna similitud. Es importante tener en cuenta que la implementación de este modelo requiere considerar el tiempo de ejecución, por lo que se recomienda utilizar una GPU para optimizar el rendimiento.

Estas métricas pueden ser implementadas mediante la librería mencionada anteriormente y específicamente la siguiente clase:

Métrica	Clase	Función
WebBertSimilarity	semantic_text_similarity	models.WebBertSimilarity

Tabla 3.1: Clases de similitud utilizada

- **SBERT**: También denominado Sentence-BERT, es una adaptación del modelo BERT que utiliza redes siamesas (dos modelos BERT con pesos compartidos), permitiendo generar embeddings independientes para cada oración, calculando representaciones vectoriales densas de frases y calcular la similitud (en el caso del modelo “*all-MiniLM-L6-v2*”) mediante “*cosine-similarity*”. Desde la librería “*SentenceTransformer*” se usa el modelo preentrenado “*all – MiniLM – L6 – v2*” [55] que tiene como datos para el fine-tune el uso de concatenación de múltiples datasets, como puede ser “*Reddit comments (2015-2018)*” [19], “*WikiAnswers*” [13] y veintinueve más.

Estas métricas pueden ser implementadas mediante la librería “*sentence-transformers*” que ofrece la siguiente clase:

Métrica	Clase	Función
SBERT	sentence_transformers.SentenceTransformer	similitary

Tabla 3.2: Clases de similitud utilizada

Debido a la naturaleza de los rangos de valores de las distintas métricas, se lleva a cabo un proceso de normalización de las métricas transformando todos los valores obtenidos por todas estas.

<sup>3</sup>Disponible en <https://pypi.org/project/textdistance/>

<sup>4</sup>Disponible en <https://pypi.org/project/semantic-text-similarity/>

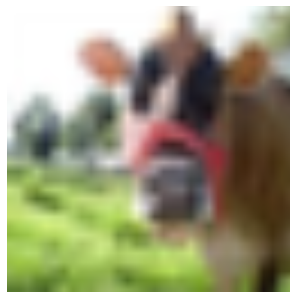
<sup>5</sup>Disponible en: <https://huggingface.co/datasets/SetFit/stsb>

### 3.3. Reuse: Personalización de explicación

Como resultado del paso de recuperación, el método ha encontrado los casos de explicación  $ec_q$  para mostrar al usuario objetivo como explicaciones basados en ejemplo (factual) o basados en contraejemplos (counterfactual). Sin embargo, no siempre podemos acceder a imágenes de buena calidad, podemos encontrarnos con situaciones en las que necesitemos clasificar imágenes borrosas, pixeladas o con ruido. Por ejemplo, al capturar imágenes a través de sensores IoT o cámaras, cuando la resolución de la cámara es baja, la iluminación es escasa, la meteorología afecta a la captura de imágenes o existen limitaciones en el ancho de banda o el almacenamiento. En estos contextos, es aún más importante obtener explicaciones enriquecidas que ayuden a los usuarios a entender la clasificación y aclaren qué entidades aparecen en las imágenes, así como los comportamientos de la IA. El paso de reutilización en nuestro método tiene como objetivo enriquecer los casos de explicación  $ec_q$  con un texto  $t_i$ . El método mostrará una explicación textual  $te_i \in t_i$  para cada caso de explicación  $c_i \in ec_q$ . La explicación  $te_i$  es una porción personalizada de texto proveniente de la descripción almacenada en la base de casos para el caso  $c_i$ . Esta descripción inicial contenía las respuestas  $a_i \in qa_i \in D_i$  que el modelo VQA generó en respuesta a las preguntas  $qu_i \in qa_i \in D_i$ . Para personalizar esta descripción, el usuario diseñador de XAI indicará al método cuáles son sus objetivos respecto a la explicación que necesita lograr. El método filtrará las respuestas iniciales de la tupla  $qa_i$ , eligiendo solo aquellas que respondan a las preguntas relacionadas con ese objetivo de XAI. Por ejemplo, si el usuario diseñador de XAI necesita persuadir a los usuarios finales sobre la decisión de clasificación de imágenes del modelo de IA, entonces el método mostrará las respuestas a las preguntas relacionadas con el objetivo de persuasión. El usuario diseñador de XAI también puede indicar al método si debe mostrar las preguntas relacionadas con las características de la imagen. La formalización de esta operación es la siguiente:

$$\begin{aligned}
 te_i &= \text{adapt}(G, iFeat, c_i) \\
 \text{adapt}(G, iFeat, c_i) &= \text{questions}G(G, c_i) \cup \text{questions}IF(iFeat, c_i) \\
 \text{questions}G(G, c_i) &= \{a_{ij}, \dots, a_{ik} \in qa_i \mid \text{type}(qa_i) = G\} \\
 \text{questions}IF(iFeat, c_i) &= \{a_{ij}, \dots, a_{ik} \in qa_i \mid \text{type}(qa_i) = IF \text{ and } iFeat = True\}
 \end{aligned}$$

donde  $G$  es el objetivo que el usuario diseñador quiere lograr;  $iFeat$  es un valor booleano para indicar si el usuario diseñador quiere que el método muestre las respuestas relacionadas con las características de la imagen (True) o no (False);  $c_i \in ec_q$  es un caso de explicación; y  $\text{type}()$ , una función que devuelve el tipo de la pregunta y respuesta  $qa_i$ , que puede ser un objetivo de XAI, o  $IF$  si  $qa_i$  está relacionado con características de la imagen. En la Figura 3.5 se muestra un ejemplo de explicación resultante obtenida con nuestro método dentro de un problema IoT.



(a) Imagen query clasificada como “vaca”.



(b) Ejemplo de explicación.

El patrón blanco y negro en el cuerpo de la vaca es una característica distintiva de las vacas y no se ve comúnmente en otros animales. Además, la vaca está de pie en un campo cubierto de hierba, que es un entorno típico para las vacas. La combinación de estas pistas visuales y el contexto de la imagen lleva a la conclusión de que el ganado es de hecho vacas.

Figura 3.5: Ejemplo de una explicación basado en ejemplos (factual) obtenida con nuestro método para una imagen pixelada. La imagen de la izquierda es la clasificada por un modelo de caja negra, mientras que la imagen de la derecha es el ejemplo de explicación obtenido con nuestro método. El texto siguiente es la explicación en lenguaje natural adaptada al objetivo de eficiencia (que ayuda a los usuarios a decidir rápidamente si la clasificación es correcta). Este texto se muestra con la explicación en el ejemplo (b).



# Evaluación

Con la evaluación, se pretende determinar si los casos explicativos obtenidos con nuestro método CBR y basado en VQA son útiles para explicar la clasificación de imágenes, y si el uso de modelos multimodales puede mejorar la explicabilidad de los casos obtenidos como explicación. Nuestra hipótesis es que los casos de explicación obtenidos con nuestro método serán más similares a la consulta que los casos obtenidos con un enfoque que sólo tiene en cuenta las características de la imagen. Si son más similares, los usuarios entenderán mejor la clasificación predicha por un modelo de caja negra de IA. Además nuestra hipótesis incluye que los casos de explicación obtenidos no perderían explicabilidad en las zonas de las imágenes que ayudan a entender la clasificación.

Para ello hemos realizado dos tipos de evaluaciones, la evaluación offline (Sección 4.1) y la evaluación online (Sección 4.2) que se describirá a continuación.

## 4.1. Evaluación offline

### 4.1.1. Dataset

Se realizó un estudio exhaustivo de distintos datasets con el objetivo de seleccionar el más adecuado para usar en la evaluación. Se consideraron varias restricciones importantes: que las imágenes correspondieran a una sola clase principal (single class), que representarían entidades simples de la vida cotidiana (como personas, animales, objetos y paisajes), y que la cantidad total fuera razonable (alrededor de 1000 imágenes). Además, se buscó que el dataset permitiera simular un entorno IoT, incluyendo imágenes de baja resolución o precisión, para reflejar condiciones reales de captura en este tipo de sistemas.

Hemos trabajado con el conjunto de datos CIFAR-100 [26], originalmente consiste en 60.000 imágenes, distribuidas en 100 clases (600 por clase) que a su vez están contenidas en 20, denominadas por la literatura, “Superclases”. Por ejemplo, la Superclase “reptiles” contiene a las clases “cocodrilo”, “tortuga”, “culebra”, etc. Se proporciona en los canales RGB, es decir, a color. Las dimensiones de las imágenes son de 32x32 píxeles.

Este conjunto es seleccionado para realizar nuestra base de casos, las imágenes las hemos podido agrupar en 4 categorías: Human, Animal, Object, Landscape. Asimismo, estas imágenes están representadas por una clase única y principal en la imagen, por ejemplo,

si se trata de un león, en la imagen saldrá un león sin que otro tipo de clase aparezca. A diferencia de CIFAR-10 este contiene más variedad de clases lo que nos permite crear nuestra base de casos de forma más completa. En lugar de tener un solo tipo de animal, por ejemplo cocodrilo, de cara a extraer un caso basado en contraejemplo (counterfactual) no podríamos porque no tenemos otro tipo de clase que sea de categoría Animal y que no sea un cocodrilo.

Por las dimensiones de las imágenes, estas se encuentran pixeladas, situación perfecta para simular los problemas que podrían surgir cuando los dispositivos IoT capturan y transmiten dichas imágenes.

Para la construcción de nuestra base de casos, hemos seleccionado de forma aleatoria 1000 imágenes distribuida en las cuatro categorías anteriormente nombradas: Human, Animal, Object y Landscape, con 250 imágenes por categoría. En la Figura 4.1 podemos observar la estructura agrupada de nuestra bases de casos y en Apéndice B se detallan la distribución de las categorías y etiquetas como resultado de la selección de dichas imágenes.

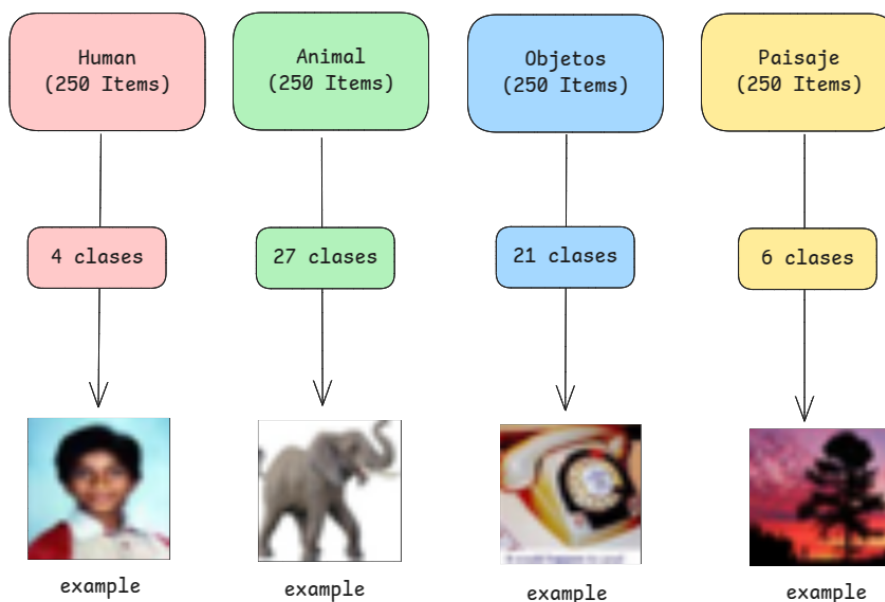


Figura 4.1: Estructura de agrupación de categorías y distribución de clases.

#### 4.1.2. Métricas de evaluación

Esta evaluación consiste de dos partes. En primer lugar, comparamos la similitud entre una consulta y los casos de explicación obtenidos con nuestro método, y los casos obtenidos con un baseline, basados en casos que utiliza la métrica de similitud píxel to píxel. En segundo lugar, comparamos la calidad de las partes importantes de la imagen que ayudan a comprender la clasificación de los casos recuperados por ambos métodos.

Para que los casos obtenidos con nuestro método sean útiles, esperamos que la calidad de sus partes importantes no sea inferior a la calidad de las partes de la imagen obtenidas con la baseline. Por lo tanto, podemos ofrecer mejores casos explicativos sin perder las partes relevantes para explicar la clasificación utilizando únicamente conocimientos textuales.

Para medir estas comparativas se utilizaron métricas de evaluación.

Como resultado de la evaluación, para una consulta dada, obtenemos  $k$  casos de explicación de cada configuración de nuestro método, y  $k$  casos de explicación del método que utiliza la métrica de similitud píxel to píxel, con la intención de utilizar métricas de evaluación. Para ello hemos utilizado dos diferentes: RMSE para evaluar la similitud entre los casos y la consulta que explican, y ROAD para evaluar las regiones de calidad en las imágenes que ayudan a reconocer la clase de la imagen:

- **Error cuadrático medio (RMSE):** Devuelve una puntuación que muestra la diferencia entre la consulta y el caso en términos de sus píxeles: calcula la diferencia entre cada píxel de ambas imágenes. Se define como:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Q_i - c_{j_i})^2}$$

donde  $Q_i$  representa el valor del píxel  $i$  en la imagen de consulta  $Q$ ,  $c_{j_i}$  representa el valor del píxel  $i$  en el caso de explicación  $c_j$ , y  $N$  es el número total de píxeles de las imágenes. En la biblioteca *sewar*<sup>1</sup>, se implementa como función *rmse* en el módulo *fullref* y toma como parámetros las dos imágenes a comparar. En la Figura 4.2 se muestra un ejemplo de imágenes similares y valor del calculo RMSE de la imagen de referencia con la imagen comparativa. Entre más similares las imágenes el calculo devuelto sera más bajo.

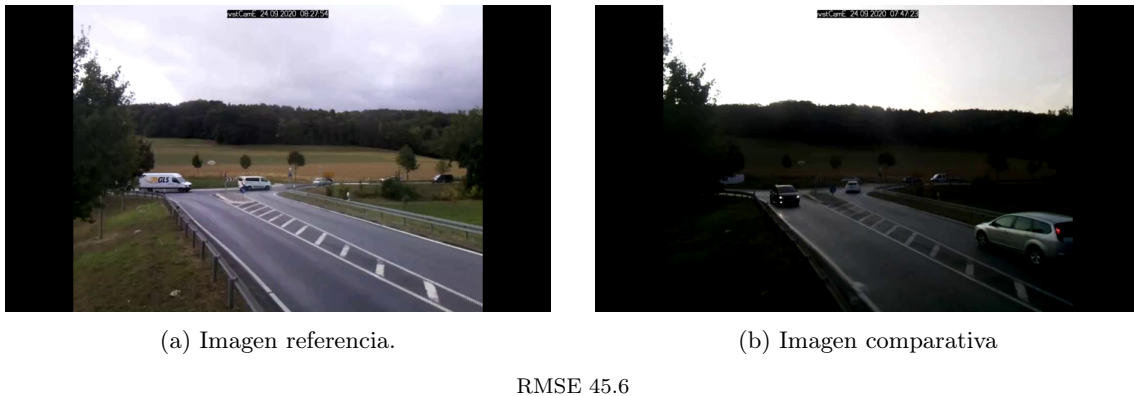


Figura 4.2: Una imagen de una cámara de vigilancia de tráfico en Alemania se utiliza para mostrar los resultados del RMSE. Fuente: [36]

- **Removed y Debias (ROAD) [48]:** Esta métrica obtiene una puntuación que indica lo importantes que son algunas regiones de una imagen para comprender la clase de imagen. A partir de un mapa de calor, ROAD oscurece las partes más importantes de la imagen con la estrategia de imputación lineal ruidosa [48], evaluando la predicción obtenida al ocultar esas partes importantes. La métrica requiere transformar

<sup>1</sup><https://github.com/andrewekhalel/sewar>

las imágenes en mapas de calor, se realiza esta transformación utilizando el modelo Grad-CAM de XAI, en la Figura 4.1.2 se muestra la aplicación de GradCAM para el cálculo de ROAD. Cuanto mayor la puntuación ROAD, más importantes las zonas ocultas, es decir, más útiles son las regiones de la imagen para comprender la clase de imagen. Su objetivo es superar las limitaciones de las estrategias de evaluación tradicionales basadas en la perturbación de píxeles, que suelen ser inconsistentes.



Figura 4.3: GradCAM + Removed y debias (ROAD) aplicado a la clase “perro”. Fuente: [14]

### 4.1.3. Proceso experimental

Empezamos el experimento a partir de nuestro conjunto de datos de 1.000 imágenes, 250 por cada categoría (Human, Animal, Object, Landscape). Además, cada una de esas imágenes conservaba la clase específica que tenía en el conjunto de datos CIFAR original, por ejemplo, vaca, bebé o reloj.

Ya teniendo a disposición nuestra base de imágenes reagrupadas con sus respectivas clases y categorías y base de preguntas, es decir, teniendo las **Q**uestions y los **V**isuals, debemos aplicar dichas preguntas sobre las imágenes para obtener la información textual, es decir, obtener las **A**nswers. Luego, con la información textual proveniente de la aplicación de VQA se crea la base de casos, donde la descripción de una imagen (caso) se denota y describe como se realizó en la Ecuación 3.1.

Para extraer la información textual de las imágenes de nuestra base de casos hemos utilizado el modelo LLaVA [30] para implementar la tarea de VQA. Es un modelo multimodal diseñado para tareas complejas de VQA en imágenes incongruentes que ha demostrado obtener descripciones más detalladas que otros modelos. Es un modelo de código abierto que se ha puesto a punto con datos generados por GPT, diseñado para seguir instrucciones tanto en texto como en imágenes teniendo Vicuna (basado en LLaMA) y CLIP como base. En la Figura 4.4 se muestra el uso que se puede dar con dicho modelo.

En su arquitectura conecta un encoder visual (CLIP) [45] que transforma la imagen en una representación comprensible para el LLM (Vicuna) [61] y se conectan a través de una capa de proyección que tiene la función de transformar las características de la imagen al espacio de embedding de palabras, manteniendo la misma dimensionalidad que el espacio de embedding de palabras del modelo lingüístico (Vicuna), en la Figura 4.5 se muestra la arquitectura general de este modelo. Para su entrenamiento no solo esta compuesto por


Visual input example, Extreme Ironing:	
	
User	Source <a href="https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg">https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg</a>
LLaVA	What is unusual about this image? The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment. [Start a new conversation, and clear the history]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention <b>due to his unconventional choice of ironing his clothes on top of a moving car</b> . The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Figura 4.4: Prompt de ejemplo y comparación con otros modelos donde LLaVA ofrece una respuesta más completa que GPT-4 [39], inclusive cuando se le pide simplemente describir la imagen, LLaVA identifica los aspectos atípicos de la imagen. Asimismo, cuando se le pide simplemente describir la imagen, LLaVA identifica los aspectos atípicos de la imagen. Fuente: [30]

pares de imagen con texto [16] sino también por pares de imagen con conversaciones, descripciones detalladas y razonamiento complejo (inferencias causales/espaciales) generados sintéticamente por modelos como GPT4 [17].

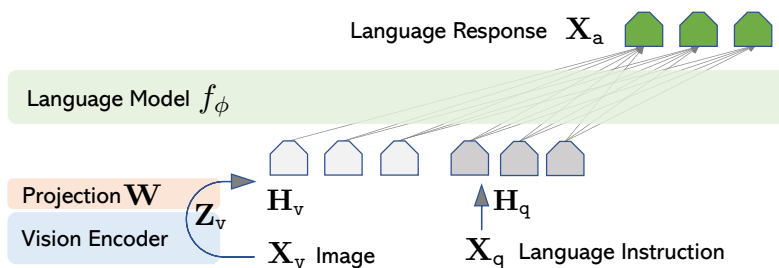


Figura 4.5: Arquitectura LLaVA. Fuente: [30]

Para su implementación se puede usar a través de la librería de transformers que ofrece

modelos de código abierto en HuggingFace <sup>2</sup>.

El ciclo de implementación es por categoría, empezando, por ejemplo, con la categoría Human, y por cada imagen de la base de casos realiza las preguntas del grupo correspondiente de preguntas, las respuestas de salida se guardan en datasets siendo las columnas las preguntas realizadas y cada fila representa la información textual de la imagen correspondiente. En la Figura 4.6 se muestra una vista ejemplo de la aplicación de este ciclo para la categoría Human.

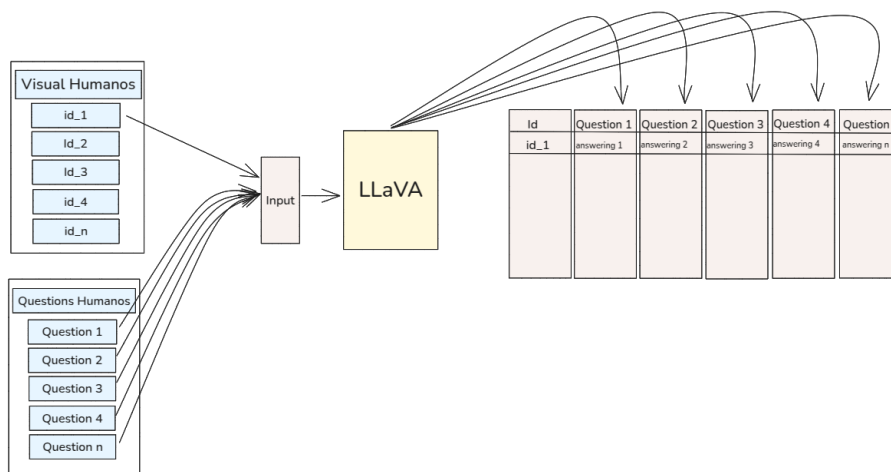


Figura 4.6: Implementación VQA

La ejecución se realizó en máquinas en línea en la plataforma Kaggle con un modelo de procesador Intel(R) Xeon(R) CPU, con una frecuencia de CPU de 2000.158 MHz, 2 núcleos de CPU y 2 unidades Tesla T4 [38] y un límite de memoria de aproximadamente 14.6 GB de GPU. Estas máquinas tienen una restricción de uso de un cuaderno por más de doce horas continuas y treinta horas semanales para el uso de la GPU. En promedio, cada pregunta tiene una duración de cinco segundos. En el caso de la categoría Human, con cincuenta y ocho preguntas y docientos cincuenta imágenes por cada categoría, la ejecución de esta categoría dura aproximadamente veinte horas. Para la categoría Animal, con cincuenta y siete preguntas, la ejecución dura diecinueve horas y cuarenta y ocho minutos. En la categoría Object, con cincuenta y un preguntas, la ejecución dura diecisiete horas y cuarenta y dos minutos. Finalmente, en la categoría Landscape, con 48 preguntas, la ejecución dura dieciséis horas y treinta y seis minutos. Este paso puede resumirse en que consiste en generar la base de casos, y para crearla, aplicamos LLaVA. Para cada imagen de nuestro conjunto de datos, pedimos a LLaVA que respondiera a las preguntas predefinidas que describimos en la Subsección 3.1.1, teniendo en cuenta la categoría de la imagen. Como resultado, guardamos todas esas imágenes en la base de casos, junto con sus descripciones: la imagen en sí, su clase y categoría (ambas obtenidas del conjunto de datos CIFAR), y los pares de preguntas y respuestas, con las respuestas generadas por LLaVA.

<sup>2</sup>Disponible en <https://huggingface.co/llava-hf/llava-1.5-7b-hf>

El siguiente paso fue encontrar las explicaciones basados en ejemplos (factual) y basados en contraejemplo (counterfactual) para cada imagen realizando una evaluación leave-one, es decir, para cada imagen se compara la información textual del resto de imágenes de la base de casos obtenidas después de realizar la fase MAC, tal y como se ha explicado en la Sección 3.2. Para encontrar ambos tipos de explicaciones utilizamos las métricas de similitud descritas en la Subsección 3.2.1, que utiliza los pares de preguntas y respuestas para calcular la similitud entre las imágenes y realizar un promedio matemático de los valores de similitud obtenidos para todas las preguntas entre una imagen y otra. Además, también calculamos los valores de similitud obtenidos con la métrica básica (o baseline): la métrica de similitud píxel to píxel. Esta métrica sólo utiliza los píxeles de la imagen para obtener los casos de explicación, sin tener en cuenta los pares de pregunta y respuesta que describen la imagen. Para cada imagen de la base de casos, obtuvimos  $k$  explicaciones basados en ejemplos (factual) y  $k$  basados en contraejemplos (counterfactual), donde  $k \in [1, 2, 3]$  siendo el uno el caso más similar.

Una vez obtenidos los casos explicativos, calculamos el RMSE y la ROAD. Para  $k = [2, 3]$ , los valores de RMSE y ROAD son el promedio calculado a partir de los valores individuales. Por ejemplo, para una imagen  $i$  y su explicación correspondiente a  $k=1$ , se calcula el RMSE de entre la imagen  $i$  y la imagen a  $k=1$ , mientras que para  $k=2$  se calcula el RMSE de la imagen  $i$  con la imagen de  $k=1$ , el RMSE de la imagen  $i$  con la imagen  $k=2$  y se calcula el promedio matemático del valor de ambos RMSE. En el caso de los valores de ROAD pueden ser negativos, los normalizamos para interpretarlos mejor.

Luego, para que los resultados de RMSE y ROAD sean más comprensibles, calculamos los porcentajes de mejora de nuestro método en comparación con el baseline. Se calcula el porcentaje de mejora de las combinaciones con la siguiente fórmula:

$$\text{Porcentaje de mejora} = \frac{\text{Valor nuevo} - \text{Valor antiguo}}{\text{Valor antiguo}} \times 100\%$$

Siendo *Valor nuevo* el valor de la métrica a comparar su mejora, por ejemplo, los valores de la métrica Cosine y *Valor antiguo* los valores de la métrica baseline.

Posteriormente se promedia el valor de las evaluaciones, y se han obtenido resultados para RMSE, ROAD, por filtro (factual o counterfactual), por la métrica de similitud y finalmente por las  $k$  correspondiente. Estos resultados se muestran y analizan en la siguiente sección.

#### 4.1.4. Resultados

En la Tabla 4.1 se muestra los resultados obtenidos en la evaluación utilizando RMSE y ROAD para comparar la calidad de los casos de explicación recuperados por nuestro enfoque y el baseline establecido. En cuanto a los valores de RMSE, podemos ver que nuestro método tiene un rendimiento positivo sobre el baseline para todas las casuísticas estudiadas como tipo de explicación, métricas o valores de  $k$ . Las métricas de similitud que obtienen los mejores resultados son las métricas de similitud basadas en semántica, en general por encima de las métricas basadas en texto. La métrica SBERT obtiene la

mejor puntuación para las explicación basado en ejemplo (factual) recuperando los casos  $k = [1, 2]$ , con la puntuación más alta, siendo 84,46 (para  $k = 1$  con una mejora del 35,55 % con respecto al valor del baseline). En el caso de las explicaciones basados en contraejemplos (counterfactual) la métrica Web Similitary tiene la mejor puntuación obtenida siendo 90,49 (para  $k = 1$  con una mejora del 36,31 % con respecto al valor del baseline) por encima de las puntuaciones de  $k = [2, 3]$ . Estos resultados respaldan la utilidad de la semántica al recuperar los casos de explicación. El significado del texto que describe las imágenes es crucial para obtener el caso más similar a la consulta. Se detecta una excepción en el tipo de explicación basado en ejemplos (factual) para la métrica Jaccard que tiene la mejor puntuación obtenida siendo 85,92 (una mejora del 30.65 % con respecto al valor del baseline). Sin embargo, las métricas SBERT y Web Similitary alcanzan puntuaciones muy similares

De acuerdo a los resultados, podemos afirmar que el porcentaje de mejora es siempre superior al 30 %, lo que indica que nuestro método mejora el rendimiento al baseline en aproximadamente un tercio. Para cada configuración de la evaluación, la peor mejora se produce cuando  $k = 3$ , que era el comportamiento esperado, ya que cuantos más casos de explicación necesitamos, más difícil es para el método encontrar casos similares que sean adecuados para explicar la imagen a consultar. Asimismo, se observa que nuestro método ofrece mejores casos de explicación cuando proporciona explicaciones basados en ejemplos (factual) que cuando proporciona basados en contraejemplos (counterfactual). Los valores de RMSE son más bajos cuando obtenemos explicaciones basados en ejemplos, aunque la diferencia con el RMSE obtenido al conseguir explicaciones basados en contraejemplos no es muy notable. Sin embargo, nuestro método muestra una mayor mejora sobre el baseline al recuperar casos basados en contraejemplos (counterfactual) que al recuperar explicaciones basadas en ejemplos (factual).

En cuanto a los valores normalizados de ROAD, podemos observar que todos los valores son muy similares para cualquier configuración de la evaluación. Esto podría deberse al hecho de que todas las imágenes del conjunto de datos están pixeladas, por lo que es más difícil para ROAD encontrar las regiones que han sido más relevantes para que el modelo tome una decisión de clasificación específica.

Podemos concluir que nuestro método consigue recuperar casos explicativos con regiones ligeramente más importantes que el baseline. En la mayoría de los casos, el porcentaje de mejora de nuestro enfoque sobre el baseline es positivo, excepto al obtener contraejemplos (counterfactual) utilizando la métrica Bag Distance cuando  $k = [2, 3]$  (los porcentajes de mejora son -1,62 % y -2,09 %, respectivamente). Este último comportamiento puede deberse a que dicha métrica no tiene en cuenta el orden o el significado, sino el número de apariciones de cada palabra en el conjunto de texto, teniendo en cuenta que en contraejemplos (counterfactual) es más probable que en la información textual de la explicación recuperada y la imagen a explicar compartan una cantidad de palabras mucho menor.

Además, al analizar los valores normalizados de ROAD, podemos observar que las métricas de similitud basadas en la semántica no son las que obtienen los mejores resultados claramente como cuando analizamos los valores de RMSE. Según la evaluación ROAD al recuperar los casos de explicaciones basados en ejemplos (factual), la distancia Levenshtein es la mejor cuando  $k = 1$  (7,62 % de mejora), y la métrica de similitud Cosine cuando  $k = [2, 3]$  (3,91 % y 2,30 %, de mejora respectivamente). Para recuperar contraejemplos

(counterfactual), la mejor métrica de similitud es SBERT cuando  $k = 1$  (6,18 % de mejora), pero Levenshtein es de nuevo la mejor cuando  $k = [2, 3]$  (3,24 % y 2,45 %, respectivamente).

M	ET	RMSE						Normalised ROAD					
		K1		K2		K3		K1		K2		K3	
		Value	%	Value	%	Value	%	Value	%	Value	%	Value	%
B	F	93.00	28.63	92.01	27.47	91.53	26.02	0.86	5.57	0.86	1.70	0.86	0.02
C	F	85.88	34.35	85.95	32.43	86.21	30.40	0.87	6.93	<b>0.88</b>	<b>3.91</b>	<b>0.87</b>	<b>2.30</b>
J	F	85.29	34.76	85.38	32.85	<b>85.92</b>	<b>30.65</b>	0.87	7.31	0.88	3.88	0.87	2.15
L	F	85.94	34.36	86.39	32.13	86.64	30.12	<b>0.88</b>	<b>7.62</b>	0.88	3.73	0.87	2.11
O	F	85.58	34.48	86.77	31.78	86.97	29.76	0.87	6.49	0.87	3.08	0.87	1.50
SB	F	<b>84.46</b>	<b>35.35</b>	<b>85.36</b>	<b>32.96</b>	86.15	30.58	0.87	7.04	0.87	3.52	0.87	2.04
WS	F	84.37	35.33	85.49	32.79	86.11	30.58	0.87	6.76	0.87	3.49	0.87	1.87
P2P	F	131.75	-	128.05	-	124.74	-	0.84	-	0.85	-	0.86	-
B	CF	94.32	33.49	94.13	32.72	94.37	31.72	0.84	0.57	0.84	-1.62	0.84	-2.09
C	CF	92.45	34.70	92.77	33.51	92.95	32.63	0.88	5.51	0.88	2.65	0.88	1.89
J	CF	92.07	34.99	92.92	33.46	93.14	32.49	0.88	5.66	0.88	2.49	0.87	1.82
L	CF	92.99	34.39	93.26	33.33	93.31	32.54	0.88	6.13	<b>0.88</b>	<b>3.24</b>	<b>0.88</b>	<b>2.45</b>
O	CF	92.04	35.17	92.75	33.70	93.01	32.71	0.87	4.11	0.87	1.86	0.87	1.31
SB	CF	90.97	36.02	91.75	34.48	92.58	33.14	<b>0.88</b>	<b>6.18</b>	0.88	2.95	0.88	2.32
WS	CF	<b>90.49</b>	<b>36.31</b>	<b>91.12</b>	<b>34.92</b>	<b>92.34</b>	<b>33.29</b>	0.87	4.17	0.87	1.34	0.86	0.61
P2P	CF	142.54	-	140.58	-	138.90	-	0.85	-	0.86	-	0.86	-

Tabla 4.1: Resultados obtenidos en la evaluación (los mejores resultados en negrita). M indica la métrica utilizada: Bag distance (B), cosine (C), Jaccard (J), Levenshtein (L), Overlap (O), SBERT (SB), Web similarity (WS), or poxel-to-pixel (P2P). ET indica el tipo de explicación: por ejemplo (factual F) o contraejemplo (counterfactual CF). El símbolo % indica el porcentaje de mejora de nuestro método en comparación con el baseline.

Como síntesis podemos coincidir que con este experimento se ha conseguido comprobar que las explicaciones que se están obteniendo con nuestro método son más similares a la imagen de consulta (query) que las obtenidas con el método baseline, y además mantienen las mismas zonas explicativas que las imágenes obtenidas con el baseline. De igual manera, podemos concluir que los casos de explicación recuperados por cualquier enfoque cuando las imágenes son de baja calidad (por ejemplo, en el dominio IoT) van a ser menos eficaces que en otras situaciones con imágenes de mayor calidad. Ahora se necesita comprobar que estas explicaciones efectivamente son útiles para los usuarios, para ello se ha desarrollado la evaluación online descrita en la siguiente sección.

## 4.2. Evaluación online

Se ha diseñado una evaluación, cuyo objetivo es comprobar si los usuarios entienden las explicaciones y les resultan útiles. Para ello se ha diseñado una *encuesta*<sup>3</sup> de tipo formulario que consta de cuatro secciones. En la primera versión de la misma, durante el proceso de participación, varios usuarios expresaron que deseaban tener una opción para indicar que no entendían ninguna de las explicaciones. Por lo tanto, se ha detectado que el formulario de evaluación estaba mal diseñado. En consecuencia, se ha decidido descartarla y rehacerla, incluyendo esta nueva opción, quedando estructurado de la siguiente forma:

- **Evaluaciones de las explicaciones basadas en ejemplos (Factual):**

Se presentan dos secciones, en cada sección se muestra una clasificación diferente de la categoría Animal. Para cada clasificación, se muestra dos explicaciones, ambas de

<sup>3</sup>Encuesta disponible en <https://forms.gle/N6jaQeecU9jS1nK58>

Resultados disponibles en [https://docs.google.com/spreadsheets/d/166dFofBqU7w4EsI\\_KQbeU2Duw05fr5cH\\_89tan4-LX4/edit?usp=sharing](https://docs.google.com/spreadsheets/d/166dFofBqU7w4EsI_KQbeU2Duw05fr5cH_89tan4-LX4/edit?usp=sharing)

tipo factual (basadas en ejemplo), es decir imágenes similares a la imagen (query) con la misma clasificación.

En cada sección estas explicaciones mostradas se dividen en la opción A, obtenida con el baseline, y la opción B obtenida con nuestro método (específicamente la métrica Overlap escogida aleatoriamente). Además a esta última opción se añade el proceso de personalización de la explicación (reuse), que añade un texto adicional para describir y explicar la imagen de manera más detallada.

Entre ambas opciones que se le ofrece al usuario, este último debe seleccionar la opción que cree que ayudará a entender mejor por qué la primera imagen se clasificó de esa forma. En caso de que ninguna de las imágenes explicativas sean considerada como adecuada, se ofrece una tercera opción C que indica que ninguna de las opciones anteriores ayuda a entender la clasificación. En la Figura C.1 del Apéndice C se muestra uno de los ejemplos que se ofrece al usuario.

- **Evaluaciones de las explicaciones basadas en contraejemplos (Counterfactual):**

Se presentan dos secciones, cada sección se muestra a una clasificación diferente de la categoría Animal. Para cada clasificación, se muestra dos explicaciones, ambas de tipo counterfactual (basadas en contraejemplo), es decir imágenes similares a la imagen (query) con distinta clasificación.

En cada sección estas explicaciones mostradas se dividen en la opción A, obtenido con el baseline. La opción B obtenido con nuestro método (específicamente la métrica Overlap escogida aleatoriamente). Además a esta última opción se añade el proceso de personalización de la explicación (reuse), que añade un texto adicional para describir y explicar la imagen de manera más detallada.

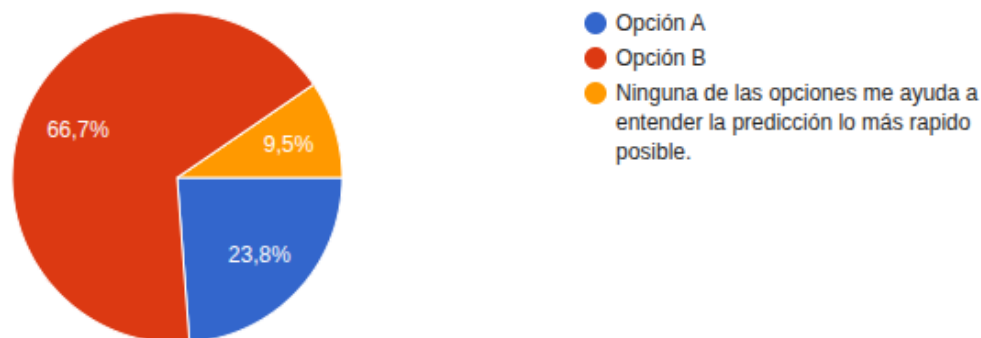
Entre ambas opciones que se le ofrece al usuario, este último debe seleccionar la opción que cree que ayudará a entender mejor por qué la primera imagen se clasificó de esa forma. En caso de que ninguna de las imágenes explicativas sean considerada como adecuada, se ofrece una tercera opción C que indica que ninguna de las opciones anteriores ayuda a entender la clasificación. En el apéndice Figura C.2 se muestra uno de los ejemplos que se ofrece al usuario.

#### 4.2.1. Resultados

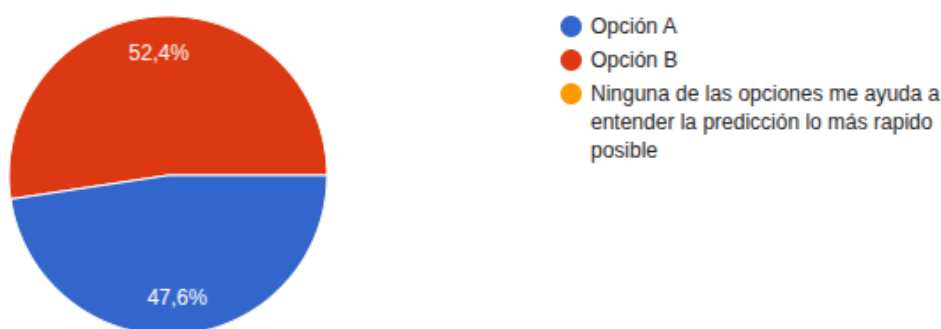
En la encuesta participaron 21 personas, los resultados obtenidos se muestran en las Figuras 4.7 y 4.8.

Los usuarios consideran que en el caso de las explicaciones basados en ejemplos (factual), el caso de ejemplo recuperado con nuestro método puede ser de utilidad para explicar la imagen query por encima de la imagen obtenida con la métrica píxel to píxel. Sin embargo, cuando se trata de explicaciones basadas en contraejemplos (counterfactual) no hay una opinión mayoritaria, hay usuarios que consideran que los casos recuperados con nuestro método podría no ser útil para entender la clasificación de la imagen query. Puede deberse a que en la base de casos no se encuentren imágenes muy claras por la calidad de las mismas

o porque la especie de los animales en las imágenes de los contraejemplos (counterfactual) son muy distintas a la especie del animal en la imagen query. Por ejemplo, si la imagen de consulta es un perro y en la base de casos tenemos un perro, un cocodrilo, un caracol y una lagartija, al buscar una explicación basado en contraejemplos, por el color del perro y la calidad de las imágenes podría coincidir con el de una lagartija. Sin embargo, esto no ayuda al objetivo, ya que son animales que difieren mucho en sus características físicas.



(a) Sección 1

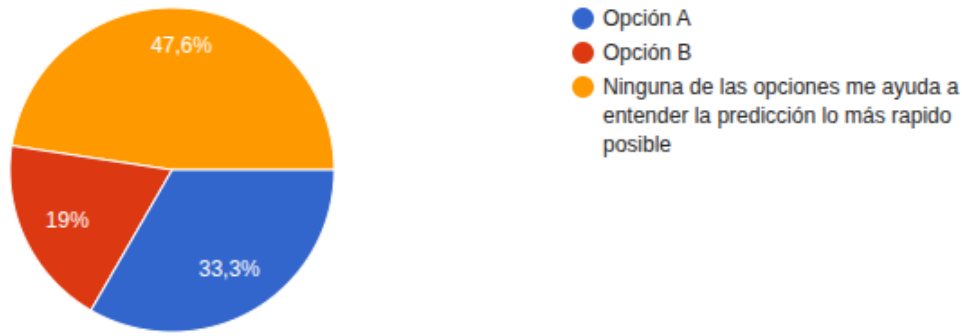


(b) Sección 2

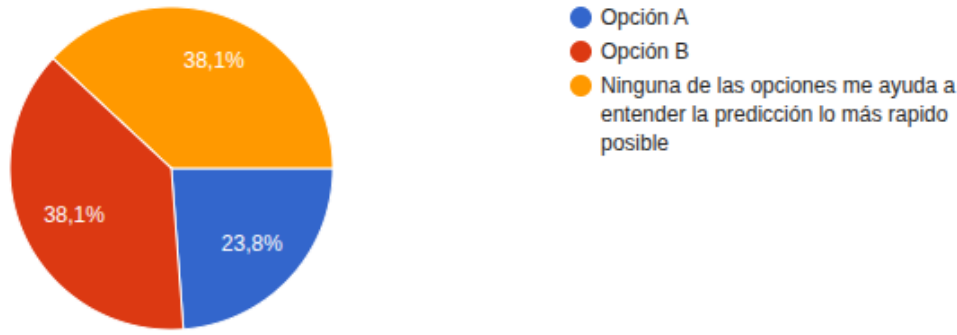
Figura 4.7: Resultados obtenidos cuando se le preguntaba a los usuarios qué explicación les ayuda a entender la clasificación lo más rápido posible. La opción A hace referencia a la explicación obtenida con el baseline y la opción B hace referencia a la explicación obtenida con nuestro método.

#### 4.2.2. Conclusiones obtenidas de la evaluación

Como conclusión general, podemos confirmar que nuestro método obtiene explicaciones más similares a la imagen a explicar, tanto para explicaciones basados en ejemplos (factual) como basados en contraejemplo (counterfactual), en comparación con un enfoque basado meramente en imágenes, mejorando así la comprensión de la clasificación de imágenes. Las métricas de similitud basadas en la semántica han demostrado estar entre los mejores enfoques para recuperar casos de explicación.



(a) Sección 3



(b) Sección 4

Figura 4.8: Resultados obtenidos cuando se le preguntaba a los usuarios qué explicación les ayuda a entender la clasificación lo más rápido posible. La opción A hace referencia a la explicación obtenida con el baseline y la opción B hace referencia a la explicación obtenida con nuestro método.

Respecto a la encuesta realizada por usuarios reales, podría determinarse que para los casos recuperados basados en ejemplos (factual) se ha obtenido resultados claros que afirman que nuestro método obtiene las mejores explicaciones, mientras que para los casos recuperados basados en contraejemplos (counterfactual), a pesar de que hay usuarios que consideran que nuestro método obtienen las mejores explicaciones, no hay una opinión mayoritaria que puede afirmarlo.

Los resultados de la evaluación son prometedores y abre la puerta a emplear modelos multimodales en XAI para tareas de IA relacionadas con imágenes.

## Conclusiones y Trabajo Futuro

### 5.1. Conclusiones

En este proyecto, hemos propuesto el uso de modelos de VQA como una técnica para desarrollar un método XAI para la clasificación de imágenes. Cuando se implementan explicaciones basadas en ejemplos para la clasificación de imágenes, la práctica más común es implementar métricas de similitud basadas en imágenes para obtener los casos de explicación o incluso técnicas de machine learning de clustering. En este Trabajo Fin de Máster proponemos un método CBR que hace uso de modelos LLM multimodales en tareas de VQA para ofrecer explicaciones basados en ejemplos (factual) y basados en contraejemplos (counterfactual) de imágenes de una clase única. VQA es usado para describir estos casos, luego las métricas de similitud basadas en texto y semántica son usadas para recuperar los casos de explicación. En la etapa de reutilización, el método filtra el conocimiento textual generado por el modelo VQA considerando el objetivo XAI a conseguir. Al implementar esta particularidad, nuestro método muestra a los usuarios una explicación textual personalizada y en lenguaje natural que es legible y apoya la explicación de los casos.

De acuerdo a los resultados obtenidos en la evaluación, el método ayuda a mejorar la comprensión del comportamiento del modelo de IA al predecir la clase de la imagen. El modelo propuesto obtiene mejores casos de explicación que las explicaciones meramente basadas en imágenes, mientras se mantienen la mismas características en la imagen que ayudan a entender la clasificación. Nuestro método logra estos resultados utilizando solo la información textual de las descripciones generadas por el modelo de LLM en lugar de las características visuales comúnmente utilizadas. Un enfoque a resaltar es el uso de casos basados no solo en ejemplos, sino también en contraejemplos. Este último aspecto resulta interesante, ya que en la literatura es común estudiar únicamente los casos basados en ejemplos. Además, incluir casos basados en contraejemplos puede ayudar a los usuarios a comprender mejor el comportamiento de la IA. Al incorporar este tipo de casos, la información textual podría resaltar distinciones mínimas entre imágenes. Por lo tanto, este trabajo es una primera aproximación a la investigación en el uso de VQA para la explicación de tareas en clasificación de imágenes, y abre la puerta a investigaciones futuras relacionada con este campo. En la siguiente sección abordaremos algunas de estos posibles trabajos.

## 5.2. Trabajo futuro

En el sistema diseñado se encuentran varios puntos que pudieran crear nuevos incisos a explotar. Entre ellos, se pueden explorar más datasets donde la clase principal no sea clara porque una misma imagen puede contener varias clases (multiclass), para aplicar el mismo método propuesto en este Trabajo de Fin de Máster y evaluarlo de la misma forma. Por ejemplo una imagen multiclass contendría como clases principales un león y un hombre en el mismo espacio, ya que se trata de dos clases (león y hombre) y además dos categorías (Animal y Human) distintas. Sumando a lo anterior, considerar imágenes más claras (de mayor calidad) para verificar si el método no funciona para generar explicaciones basados en contraejemplos (counterfactual). En este Trabajo Fin de Máster se implementan métricas de similitud basadas en texto disponibles en la literatura, sin embargo, otras librerías ofrecen más. Podemos implementar solo de la librería *“textdistance”* los algoritmos Hamming, MLIPNS, Damerau-Levenshtein, Jaro-Winkler, Strcmp95, Needleman-Wunsch, Gotoh, Smith-Waterman, Sørensen–Dice coefficient, Tversky index, Tanimoto distance, Monge-Elkan, longest common subsequence similarity, longest common substring similarity y Ratcliff-Obershelp similarity. Por otra parte, para las métricas de similitud basadas en la semántica textual podemos usar otras librerías como “Spacy” o “Gensim”. sería enriquecedor abordar una mayor cantidad de métricas para obtener una evaluación más integral.

De igual forma podemos aplicar otros métodos de explicación basadas en características como LIME, Layer-wise Relevance Propagation (LRP), Integrated Gradients (IntGrad), SmoothGrad, SHAP, Redes B-cos y realizar una comparación con sus resultados con los obtenidos en nuestro método.

Hoy en día los modelos de código abierto están experimentando un auge sin precedentes. Estos modelos cada vez son más accesible para su explotación, este crecimiento debido al rápido desarrollo de la industria da la oportunidad de aprovechar nuevos modelos para nuestras investigaciones. Para nuestro sistema, podemos usar modelos multimodales modernos para obtener las descripciones de las imágenes. Sin considerar el consumo de recursos, se puede ejecutar modelos tales como, BLIP, VILT, QWEN, FALCON. Incluso existen modelos que prometen un bajo consumo de recursos y rendimientos similares a los modelos ya establecidos en el mercado privado, como DeepSeek. Además, se puede explorar modelos de código abierto ajustados o preentrenado para mejorar las tareas sobre las imágenes de las categorías Human, Animal, Object y Landscape.

# Introduction

## 6.1. Motivation

Nowadays, artificial intelligence is increasingly present in our daily lives, in tasks such as entertainment content recommendation, behavior pattern recognition, disease detection, shopping assistants, among others. As these needs arise, Artificial Intelligence (AI) models are also becoming more complex, for example, techniques such as deep learning have emerged. Due to this high complexity, the field of Explainable Artificial Intelligence (XAI) emerged, which aims to help users understand the behavior of these models. Making these models more transparent in their predictions helps users understand the behavior of the AI model and the reasons behind its predictions, thereby increasing their trust and, consequently, the efficiency of these systems [8].

One of the most successful types of explanations provided by XAI techniques are instance-based explanations [20], referred to as case-based explanations when obtained through case-based reasoning (CBR). Instance-based explanations consist of showing examples of situations similar to the one to be explained. When these explanations are based on examples, the provided cases correspond to situations in which the AI model makes the same classification as in the case to be explained, which is known as example-based (factual) explanations. In contrast, counterexample-based (counterfactual) explanations [20] correspond to similar situations in which the model makes a different classification than in the case to be explained. Specifically, in image classification, example-based (factual) explanations would show images that have been classified the same as the image being explained, while counterexample-based (counterfactual) explanations would show similar images that have received a different classification than the case to be explained. Commonly, to explain image classification, the most used similarity metrics are based on image features [41], with the most common being pixel to pixel, which consists of comparing two images through the difference between their pixel matrices.

Currently, there are highly complex AI systems, including multimodal models, which combine different types of data, such as images and text, to learn and make predictions. Furthermore, thanks to the high performance of new natural language processing algorithms based on transformers, their application is influencing improvements not only in text-related tasks but also in image-related tasks [15].

Visual Question Answering (VQA) is the task in which multimodal AI models answer questions related to images [8, 15]. In this work, we propose the use of multimodal models, specifically Visual Question Answering (VQA) models, to obtain instance-based explanations, including both example-based (factual) and counterexample-based (counterfactual) explanations in image classification. **Our hypothesis is that, by using this type of multimodal models, we can provide better explanations to users, also including additional personalized natural language descriptions about the images.** To this end, we have developed a post-hoc, model-agnostic, case-based explanation method for explaining single-class images. This method uses the descriptions created by the VQA model applied to the images to build a case base. The descriptions are the answers to the questions that our method poses to the VQA model. The purpose of these questions is to extract information about the characteristics of these images and objectives related to XAI, such as transparency, effectiveness, or persuasion. Next, our method uses semantic and textual similarities to compare the descriptions between the query (the image to be explained) and the images in the case base. In this way, we obtain the explanation cases, both example-based (factual) and counterexample-based (counterfactual) explanations. In addition, in the reuse (or adaptation) step, our method personalizes the explanations by showing users the answers to the questions related to the XAI objective they need to address. For example, if the user needs the classification to be more transparent, the CBR method will show users the descriptions of the images so that the AI model becomes more transparent.

The method we propose may be especially suitable when images are obtained in poor quality, as it is more likely that users will not clearly understand the classification. For example, in contexts where Internet of Things (IoT) sensors have captured or transmitted low-quality images (blurry or pixelated). In this situation, the reused description can help users better understand the image and the explanation examples.

To determine whether the method works as expected and improves the performance of case explanations based solely on images, we evaluated it. We conducted an offline evaluation in which we compared the explanation cases obtained by our baseline (using an image-based similarity metric: pixel-to-pixel) and the cases obtained by our method. We measured the difference between the cases and the image to be explained. The smaller this difference to the image to be explained, the more useful the cases are for identifying the classification of that image. In addition, we measured the regions of the images that contribute to the understanding of the image, ensuring that we do not lose those regions even when using only semantic and text-based knowledge to obtain the explanation cases. Subsequently, we conducted an online evaluation where a survey was presented to real users, and participants had to choose the most useful explanations, being able to select between those generated by our method and those generated by the baseline method.

## 6.2. Objectives

The main objective is to propose a post-hoc and model-agnostic XAI method based on CBR that uses VQA techniques to provide example-based (factual) and counterexample-based (counterfactual) explanations in image classification. Likewise, to evaluate and determine whether our CBR method improves the explanations offered by other methods found in the literature, so we can conclude that the use of multimodal models in XAI is a promising approach that should be further explored in future research.

The set of specific objectives is broken down from the main objective as follows:

- **Research multimodal models for Visual Question Answering tasks:** Detail which are the most widely used, why certain models are useful for users, how they are implemented, what their general purposes are, and what resources are available for free use.
- **Exploration of similarity metrics:** Study the types of metrics available for text analysis, including both semantic and text-based metrics, as well as their implementation and results.
- **Exploration of datasets:** Investigate which public datasets are available that contain images with only one main entity type (single class), for example, if there is a dog in the image, then there should not be people or other types of animals present.
- **Construction of a CBR method:** Build a CBR method that combines VQA and the studied similarity metrics to generate example-based (factual) and counterexample-based (counterfactual) explanations.
- **Evaluations:** Evaluate the model in an IoT environment, where these methods can be especially useful. To this end, study which types of evaluation metrics would be most appropriate for conducting image-based evaluations, as well as their implementation and efficiency. Likewise, evaluate the opinion of users regarding the explanations obtained.

### 6.3. Work plan

To achieve the objectives, research on Explainable Artificial Intelligence will first be conducted in order to understand the challenges it presents, current limitations, and the use of tools to apply techniques that describe some complex models. Likewise, Case-Based Reasoning (CBR) systems will be studied to generate explanation cases and how they can be implemented in the development of our method. This will be addressed in Chapter 2 (State of the Art).

Models of Visual Question Answering (VQA) will be studied, which consist of answering natural language questions about images and that are publicly available for research purposes. It is essential to value the size of these models to determine whether it is feasible to run them with the available hardware resources. Libraries from HuggingFace<sup>1</sup> such as Transformers will be explored, which allow loading multimodal models (for example, VQA), automatically preparing input data, and generating text-based answers grounded on images. Free platforms where these models can be executed will also be investigated, in order to combine local and online executions and thus optimize processing times.

---

<sup>1</sup>Available at [https://huggingface.co/docs/transformers/tasks/visual\\_question\\_answering](https://huggingface.co/docs/transformers/tasks/visual_question_answering)

Image datasets containing a main entity type will be explored to develop a CBR method based on VQA, where the latter will extract relevant information in natural language from these images to build a case base. Subsequently, textual and semantic similarity metrics will be studied to obtain explanations based on examples (factual) and counterexamples (counterfactual). In addition, the textual information of the explanations will be reused by showing users the most relevant answers according to the objective to be addressed. This is described in Chapter 3 (Explanation method based on CBR and VQA).

Once the method proposed in this Master project has been implemented, we must determine whether our method works as expected. To this end, research will be conducted on evaluation metrics used in image similarity tasks, focusing on performing quantitative comparisons regarding how accurately one image resembles another and the explanatory quality of the provided images. A survey will also be structured to conduct an online evaluation with real users, in order to gather their opinions and measure the quality of the generated explanations. Additionally, the implementation of evaluation metrics will be considered with respect to execution times for obtaining these assessments. Both online and offline evaluations are described in Chapter 4 (Evaluation).

## 6.4. Publications

The results of this research have been published in an academic article and accepted for presentation at the International Conference on Case-Based Reasoning (ICCBR) 2025. They will be presented in the main track, in Biarritz, from June 30 to July 3, and the results will be included in the conference proceedings [7].

# Conclusions and Future Work

## 7.1. Conclusions

In this project, we have proposed the use of VQA models as a technique to develop an XAI method for image classification. When example-based explanations are implemented for image classification, the most common practice is to implement image-based similarity metrics to obtain the explanation cases or even clustering machine learning techniques. In this Master project, we propose a CBR method that makes use of multimodal LLM models in VQA tasks to provide explanations based on examples (factual) and based on counterexamples (counterfactual) of images from a single class. VQA is used to describe these cases, then text-based and semantic similarity metrics are used to retrieve the explanation cases. In the reuse stage, the method filters the textual knowledge generated by the VQA model considering the XAI objective to be achieved. By implementing this particularity, our method shows users a personalized textual explanation in natural language that is readable and supports the explanation of the cases.

According to the results obtained in the evaluation, the method helps to improve the understanding of the AI model behavior when predicting the class of the image. The proposed model obtains better explanation cases than explanations based solely on images, while maintaining the same features in the image that help to understand the classification. Our method achieves these results using only the textual information from the descriptions generated by the LLM model instead of the commonly used visual features. An approach to highlight is the use of cases based not only on examples, but also on counterexamples. This latter aspect is interesting, since in the literature it is common to study only example-based cases. Moreover, including counterexample-based cases can help users better understand the behavior of the AI. By incorporating this type of cases, the textual information could highlight minimal distinctions between images. Therefore, this work is a first approach to research on the use of VQA for explaining tasks in image classification, and opens the door to future research related to this field. In the next section, we will address some of these possible works.

## 7.2. Future Work

In the designed system there are several points that could create new aspects to explore. Among them, more datasets can be explored where the main class is not clear

because a single image can contain several classes (multiclass), in order to apply the same method proposed in this Master project and evaluate it in the same way. For example, a multiclass image would have as main classes a lion and a man in the same space, since these are two classes (lion and man) and also two different categories (Animal and Human). In addition to the above, considering clearer images (of higher quality) to verify if the method does not work to generate counterexample-based (counterfactual) explanations. In this Master project, text-based similarity metrics available in the literature are implemented; however, other libraries offer more. We can implement from the *textdistance*” library only the algorithms Hamming, MLIPNS, Damerau-Levenshtein, Jaro-Winkler, Strcmp95, Needleman-Wunsch, Gotoh, Smith-Waterman, Sørensen–Dice coefficient, Tversky index, Tanimoto distance, Monge-Elkan, longest common subsequence similarity, longest common substring similarity, and Ratcliff-Obershelp similarity. On the other hand, for similarity metrics based on textual semantics, we can use other libraries such as “Spacy” or “Gensim”. It would be enriching to address a greater number of metrics to obtain a more comprehensive evaluation.

In the same way, we can apply other feature-based explanation methods such as LIME, Layer-wise Relevance Propagation (LRP), Integrated Gradients (IntGrad), SmoothGrad, SHAP, B-cos Networks, and perform a comparison of their results with those obtained by our method.

Today, open-source models are experiencing an unprecedented boom. These models are increasingly accessible for exploitation. This growth, due to the rapid development of the industry, provides the opportunity to leverage new models for our research. For our system, we can use more modern multimodal models to obtain image descriptions. Regardless of resource consumption, we can run models such as BLIP, VILT, LLaMA, QWEN, and FALCON. There are even models that promise low resource consumption and performance similar to established models on the private market, such as DeepSeek. Additionally, we can also explore tuned or pre-trained open-source models to improve tasks on images from the Human, Animal, Object, and Landscape categories.

# Bibliografía

- [1] AAMODT, A. y PLAZA, E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, vol. 7, páginas 39–59, 2001.
- [2] AGRAWAL, A., BATRA, D., PARIKH, D. y KEMBHAVI, A. Don't just assume; look and answer: Overcoming priors for visual question answering. 2018.
- [3] AGRAWAL, A., LU, J., ANTOL, S., MITCHELL, M., ZITNICK, C. L., BATRA, D. y PARIKH, D. Vqa: Visual question answering. 2015.
- [4] ALIPOUR, K., SCHULZE, J. P., YAO, Y., ZISKIND, A. y BURACHAS, G. A study on multimodal and interactive explanations for visual question answering. 2020.
- [5] ARRIETA, A. B., DÍAZ-RODRÍGUEZ, N., SER, J. D., BENNETOT, A., TABIK, S., BARBADO, A., GARCIA, S., GIL-LOPEZ, S., MOLINA, D., BENJAMINS, R., CHATILA, R. y HERRERA, F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. 2019. Disponible en: <https://arxiv.org/abs/1910.10045> [Último acceso: 2024-01-10].
- [6] BARTOLINI, I., CIACCIA, P. y PATELLA, M. String matching with metric trees using an approximate distance. En *SPIRE*. 2002.
- [7] BASTARDO-ROJAS, A. y CARO-MARTÍNEZ, M. Visual question answering to generate case-based explanations for image classification. En *International Conference on Case-Based Reasoning, ICCBR 2025. In press*. 2025.
- [8] CARO-MARTÍNEZ, M., WIJEKOON, A., DÍAZ-AGUDO, B. y RECIO-GARCÍA, J. A. The current and future role of visual question answering in explainable artificial intelligence. En *ICCBR Workshops*. 2023.
- [9] DE MANTARAS, R. L., MCSHERRY, D., BRIDGE, D., LEAKE, D., SMYTH, B., CRAW, S., FALTINGS ET AL. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, vol. 20(3), páginas 215–240, 2005.
- [10] SCIKIT-LEARN DEVELOPERS. sklearn.metrics.jaccard\_score — scikit-learn documentation. [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html), 2025. Accedido: 28 de mayo de 2025.
- [11] DEVIREDDY, K. A comparative study of explainable ai methods: Model-agnostic vs. model-specific approaches. 2025.

- [12] DEVLIN, J., CHANG, M.-W., LEE, K. y TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.
- [13] FADER, A., ZETTLEMOYER, L. y ETZIONI, O. Open question answering over curated and extracted knowledge bases. En *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, página 1156–1165. Association for Computing Machinery, New York, NY, USA, 2014. ISBN 9781450329569.
- [14] GIL, J. A tutorial on benchmarking and tuning model explanations. 2022. Último acceso: 2025-05-03.
- [15] GOYAL, Y., KHOT, T., SUMMERS-STAY, D., BATRA, D. y PARIKH, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. 2017.
- [16] HAOTIAN, L. Llava-cc3m-pretrain-595k. <https://huggingface.co/datasets/liuhaotian/LLaVA-CC3M-Pretrain-595K>, 2023. Último acceso: 2025-03-05.
- [17] HAOTIAN, L. Llava-cc3m-pretrain-595k. <https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K>, 2023. Último acceso: 2025-03-05.
- [18] HE, K., GKIOXARI, G., DOLLÁR, P. y GIRSHICK, R. Mask r-cnn. 2018.
- [19] HENDERSON, M., BUDZIANOWSKI, P., CASANUEVA, I., COOPE, S., GERZ, D., KUMAR, G., MRKŠIĆ, N., SPITHOURAKIS, G., SU, P.-H., VULIĆ, I. y WEN, T.-H. A repository of conversational datasets. 2019.
- [20] HENDRICKS, L. A., HU, R., DARRELL, T. y AKATA, Z. Generating counterfactual explanations with natural language. En *ICML Workshop on Human Interpretability in Machine Learning*, páginas 95–98. 2018.
- [21] HUYNH, N. D., BOUADJENEK, M. R., ARYAL, S., RAZZAK, I. y HACID, H. Visual question answering: from early developments to recent advances – a survey. 2025.
- [22] JHA, J. Similarity coefficients: A beginner’s guide to measuring string similarity. 2023. Último acceso: 2025-05-05.
- [23] JIANG, A. Q., SABLAYROLLES, A., MENSCH, A., BAMFORD, C., CHAPLOT, D. S., DE LAS CASAS, D., BRESSAND, F., LENGYEL, G., LAMPLE, G., SAULNIER, L., LAVAUD, L. R., LACHAUX, M.-A., STOCK, P., SCAO, T. L., LAVRIL, T., WANG, T., LACROIX, T. y SAYED, W. E. Mistral 7b. 2023.
- [24] KENNY, E. M. y KEANE, M. T. On generating plausible counterfactual and semi-factual explanations for deep learning. 2020.
- [25] KRISHNA, R., ZHU, Y., GROTH, O., JOHNSON, J., HATA, K., KRAVITZ, J., CHEN, S., KALANTIDIS, Y., LI, L.-J., SHAMMA, D. A., BERNSTEIN, M. S. y LI, F.-F. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.
- [26] KRIZHEVSKY, A. Learning multiple layers of features from tiny images. Informe técnico, 2009. Disponible en: <https://www.cs.toronto.edu/~kriz/cifar.html>.

- 
- [27] LEVENSHTAIN, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, vol. 10, página 707, 1966.
- [28] LI, K., VOSSELMAN, G. y YANG, M. Y. Convincing rationales for visual question answering reasoning. 2025.
- [29] LI, L. H., YATSKAR, M., YIN, D., HSIEH, C.-J. y CHANG, K.-W. Visualbert: A simple and performant baseline for vision and language. 2019.
- [30] LIU, H., LI, C., WU, Q. y LEE, Y. J. Visual instruction tuning. 2023.
- [31] LIU, S., ZENG, Z., REN, T., LI, F., ZHANG, H., YANG, J., JIANG, Q., LI, C., YANG, J., SU, H., ZHU, J. y ZHANG, L. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. 2024.
- [32] LU, J., BATRA, D., PARIKH, D. y LEE, S. Vilbert: Pretraining task-agnostic visio-linguistic representations for vision-and-language tasks. 2019.
- [33] MOLNAR, C. Interpretable machine learning: Overview. <https://christophm.github.io/interpretable-ml-book/overview.html#local-methods>, 2022. Último acceso: 2025-05-22.
- [34] MUDDAMSETTY, S., B.SCHMIDT, A. y MOESLUND, T. Post-hoc xai method for visual question answering(vqa). En *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI)-2024*. Springer, Germany, 2024.
- [35] MUDDAMSETTY, S. M., JAHROMI, M. N. S., CIONTOS, A. E., FENOY, L. M. y MOESLUND, T. B. Visual explanation of black-box model: Similarity difference and uniqueness (sidu) method. 2022.
- [36] NFDI4ING. Root mean squared error. [https://quality.nfdi4ing.de/en/latest/image\\_quality/Root\\_Mean\\_Squared\\_Error.html](https://quality.nfdi4ing.de/en/latest/image_quality/Root_Mean_Squared_Error.html), 2024. Último acceso: 2025-03-15.
- [37] NUNES, I. y JANNACH, D. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, vol. 27(3), 2017.
- [38] NVIDIA CORPORATION. Nvidia tesla t4. 2023. Último acceso: 2025-03-04.
- [39] OPENAI y ET AL. Gpt-4 technical report. 2024.
- [40] PANDEY, A., BODO, D., PHUKAN, A. y EKBAL, A. The quest for visual understanding: A journey through the evolution of visual question answering. 2025.
- [41] PAREJAS-LLANOVARCED, H., CARO-MARTÍNEZ, M., OROZCO-DEL CASTILLO, M. G. y RECIO-GARCÍA, J. A. Case-based selection of explanation methods for neural network image classifiers. *Knowledge-Based Systems*, vol. 288, página 111469, 2024.
- [42] PENNINGTON, J., SOCHER, R. y MANNING, C. Glove: Global vectors for word representation. En *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, páginas 1532–1543. 2014.
- [43] PRADEEP, P., CARO-MARTÍNEZ, M. y WIJEKON, A. A practical exploration of the convergence of case-based reasoning and explainable artificial intelligence. *Expert Systems with Applications*, vol. 255, página 124733, 2024. ISSN 0957-4174.

- [44] QUEIPO-DE-LLANO, E., CIURCAU, M., PAZ-OLALLA, A., DÍAZ-AGUDO, B. y RECIO-GARCÍA, J. A. Explainable classification system for hip fractures: A hybrid CBR+LLM surrogate approach. En *Proceedings of the Workshops at the 32nd International Conference on Case-Based Reasoning (ICCB-WS 2024) co-located with the 32nd International Conference on Case-Based Reasoning (ICCB-WS 2024)*, Mérida, Mexico, July 1, 2024 (editado por L. Malburg), vol. 3708 de *CEUR Workshop Proceedings*, páginas 65–80. CEUR-WS.org, 2024.
- [45] RAMESH, A., DHARIWAL, P., NICHOL, A., CHU, C. y CHEN, M. Hierarchical text-conditional image generation with clip latents. 2022.
- [46] RAVI, S., CHINCHURE, A., SIGAL, L., LIAO, R. y SHWARTZ, V. Vlc-bert: Visual question answering with contextualized commonsense knowledge. 2022.
- [47] RETZLAFF, C. O., ANGERSCHMID, A., SARANTI, A., SCHNEEBERGER, D., RÖTTGER, R., MÜLLER, H. y HOLZINGER, A. Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, vol. 86, página 101243, 2024. ISSN 1389-0417.
- [48] RONG, Y., LEEMANN, T., BORISOV, V., KASNECI, G. y KASNECI, E. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.
- [49] SAP, M., LEBRAS, R., ALLAWAY, E., BHAGAVATULA, C., LOURIE, N., RASHKIN, H., ROOF, B., SMITH, N. A. y CHOI, Y. Atomic: An atlas of machine commonsense for if-then reasoning. 2019.
- [50] SHAPLEY, L. S. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952.
- [51] SPEER, R., CHIN, J. y HAVASI, C. Conceptnet 5.5: An open multilingual graph of general knowledge. 2018.
- [52] STOIKOU, T., LYMPERAIYOU, M. y STAMOU, G. Knowledge-based counterfactual queries for visual question answering. 2023.
- [53] SZEGEDY, C., IOFFE, S., VANHOUCKE, V. y ALEMI, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. En *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.
- [54] SÀNCHEZ-MARRÈ, M. *PRINCIPLES OF CASE-BASED REASONING*. Universitat Politècnica de Catalunya, Dept. de Llenguatges i Sistemes Informàtics, Secció d'Intel·ligència Artificial, C. JordiGirona 1-3. 08034 Barcelona., 2001.
- [55] TANNER, H. all-minilm-l6-v2. 2021. Repositorio de GitHub.
- [56] UNIÓN EUROPEA. COMISIÓN EUROPEA. Ley de inteligencia artificial, reglamento (ue) 2024/1689. <https://artificialintelligenceact.eu/es/>, 2024. Último acceso: 2025-04-02.
- [57] UNIÓN EUROPEA. PARLAMENTO EUROPEO Y CONSEJO. Art. 22 GDPR - automated individual decision-making, including profiling. [https://www.dsgvo-portal.de/gdpr\\_article\\_22.php](https://www.dsgvo-portal.de/gdpr_article_22.php), 2016. Último acceso: 2025-04-05.

- 
- [58] VARUN. Cosine similarity: How does it measure the similarity, maths behind and usage in python. 2020. Último acceso: 2025-05-05.
- [59] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. y POLOSUKHIN, I. Attention is all you need. 2023.
- [60] YU, Z., YU, J., CUI, Y., TAO, D. y TIAN, Q. Deep modular co-attention networks for visual question answering. 2019.
- [61] ZHENG, L., CHIANG, W.-L., SHENG, Y., ZHUANG, S., WU, Z., ZHUANG, Y., LIN, Z., LI, Z., LI, D., XING, E. P., ZHANG, H., GONZALEZ, J. E. y STOICA, I. Judging llm-as-a-judge with mt-bench and chatbot arena. 2023.



## Questions for VQA model

Las siguientes preguntas han sido formuladas para lograr una descripción exhaustiva de las imágenes incluidas en nuestra base de casos.

### A.1. Human

- what is the background of the image?
- Is it rainy or sunny?
- Is it a closed or open place?
- Is the image bright?
- Are there elements in the image that can make visual noise?
- Are there a lot of different entities?
- Where is the image taken from?
- Are there humans?
- Are there animals?
- How many are *{subject\_extracted}* there?
- Do *{subject\_extracted}* have any elements that characterize them? Which ones?
- In what position are *{subject\_extracted}* located?
- which actions are shown in the image
- Could you indicate the reasons why that action is performed?
- would the *{subject\_extracted}* be able to perform any other action?
- Do you know what are the benefits or consequences of that action?
- How could you associate the *{subject\_extracted}* with their environment to describe the action?
- Are *{subject\_extracted}* young or old?

- Are There  $\{subject\_extracted\}$  male?
- Are There  $\{subject\_extracted\}$  female?
- What genders are there?
- What traits could be differentiated from other gender?
- What features could be differentiated from other types of  $\{subject\_extracted\}$ ?
- What physical characteristics could be differentiated by other gender?
- Which  $\{subject\_extracted\}$  is the male?

### 1.1 Preguntas debugging

- Have you been able to identify how many  $\{subject\_extracted\}$  there are?
- If  $\{subject\_extracted\}$  change positions could you have another prediction of the situation described?
- What is the most complicated characteristic or body part to define and determine?
- If  $\{subject\_extracted\}$  were doing another action, would it be another type of  $\{subject\_extracted\}$ ?

### 1.2 Preguntas efectiveness

- Are there elements in the image that could be noisy enough to intervene in the prediction that they are  $\{subject\_extracted\}$ ?
- What physical traits are considered for the prediction result to be  $\{subject\_extracted\}$ ?
- Why do you think identifying that they are  $\{subject\_extracted\}$  in the image is easy?

### 1.3 Preguntas efficiency

- What makes you think the  $\{subject\_extracted\}$  are actually  $\{subject\_extracted\}$  and not something else?
- With which other  $\{subject\_extracted\}$  would the prediction be most likely to give a better result?
- If there are subtle changes in skin color, clothing or posture, would the result of the prediction that they are  $\{subject\_extracted\}$  be the same

### 1.4 Preguntas trust

- Is there anything that might create doubt or uncertainty for you to say they are  $\{subject\_extracted\}$ ?

### 1.5 Preguntas persuasiveness

- Could you describe what is happening in the image?
- Could the situation you predicted be used in other areas?
- Could the situation you predicted have some example of counterfactual situation that could be useful?
- How could the situation you have predicted above be better?

### 1.6 Preguntas satisfaction

- What would an image similar to the situation found in this image look like?
- Can you describe a similar counterexample to the situation you have predicted above?
- What is the task or purpose of *{subject\_extracted}*?
- Could you in simpler words describe their function that you have predicted before?

### 1.7 Preguntas education

- What could stand out from the situation described above and be useful in similar situations?
- Can you give essential concepts that are relevant to the situation described?
- What aspects of the image have been detected that may attract attention?

### 1.8 Preguntas transparency

- Why has it been obtained that the prediction is that they are *{subject\_extracted}*?
- How did you get the result that they are *{subject\_extracted}*?
- What do you focus on to get the prediction that they are *{subject\_extracted}*?
- What do you focus on *{subject\_extracted}* to result in or say that they are *{subject\_extracted}*?
- Do you focus on parts of the *{subject\_extracted}* bodies to infer that they are *{subject\_extracted}*?
- Do you focus on features or facial expressions of the *{subject\_extracted}* to result in saying they are *{subject\_extracted}*?
- Do you focus on the *{subject\_extracted}* clothing to result in the *{subject\_extracted}* being *{subject\_extracted}*?
- Do you focus on the *{subject\_extracted}* body shape to result in the result that they are *{subject\_extracted}*?

## A.2. Animal

- what is the background of the image?
- Is it rainy or sunny?
- Is it a closed or open place?
- Is the image bright?
- Are there elements in the image that can make visual noise?
- Are there a lot of different entities?
- Where is the image taken from?
- Are there humans?
- Are there animals?
- How many are  $\{subject\_extracted\}$  there?
- Do  $\{subject\_extracted\}$  have any elements that characterize them? Which ones?
- In what position are  $\{subject\_extracted\}$  located?
- which actions are shown in the image
- Could you indicate the reasons why that action is performed?
- would the  $\{subject\_extracted\}$  be able to perform any other action?
- Do you know what are the benefits or consequences of that action?
- How could you associate the  $\{subject\_extracted\}$  with their environment to describe the action?
- Are  $\{subject\_extracted\}$  young or old?
- Are There  $\{subject\_extracted\}$  male?
- Are There  $\{subject\_extracted\}$  female?
- What genders are there?
- What features could be differentiated from other types of  $\{subject\_extracted\}$ ?
- What physical characteristics could be differentiated other gender?
- Which  $\{subject\_extracted\}$  is the male?

### 2.1 Preguntas debugging

- Have you been able to identify how many  $\{subject\_extracted\}$  there are?
- If the  $\{subject\_extracted\}$  change positions could you have another prediction of the situation described?
- What is the most complicated characteristic or body part to define and determine?
- If the  $\{subject\_extracted\}$  were doing another action, would it be another type of  $\{subject\_extracted\}$ ?

## 2.2 Preguntas effectiveness

- Are there elements in the image that could be noisy enough to intervene in the prediction that they are *{subject\_extracted}*?
- What physical traits are considered for the prediction result to be *{subject\_extracted}*?
- Why do you think identifying that they are *{subject\_extracted}* in the image is easy?

## 2.3 Preguntas efficiency

- What makes you think the *{subject\_extracted}* are actually *{subject\_extracted}* and not something else?
- With which other *{subject\_extracted}* would the prediction be most likely to give a better result?
- If there are subtle changes in skin color or posture, would the result of the prediction that they are *{subject\_extracted}* be the same?

## 2.4 Preguntas trust

- Is there anything that might create doubt or uncertainty for you to say they are *{subject\_extracted}*?

## 2.5 Preguntas persuasiveness

- Could you describe what is happening in the image?
- Could the situation you predicted be used in other areas?
- Could the situation you predicted have some example of counterfactual situation that could be useful?
- How could the situation you have predicted above be better?

## 2.6 Preguntas satisfaction

- What would an image similar to the situation found in this image look like?
- Can you describe a similar counterexample to the situation you have predicted above?
- What is the task or purpose of *{subject\_extracted}*?
- Could you in simpler words describe their function that you have predicted before?

## 2.7 Preguntas education

- What could stand out from the situation described above and be useful in similar situations?
- Can you give essential concepts that are relevant to the situation described?
- What aspects of the image have been detected that may attract attention?

## 2.8 Preguntas transparency

- Why has it been obtained that the prediction is that they are  $\{subject\_extracted\}$ ?
- How did you get the result that they are  $\{subject\_extracted\}$ ?
- What do you focus on to get the prediction that they are  $\{subject\_extracted\}$ ?
- What do you focus on  $\{subject\_extracted\}$  to result in or say that they are  $\{subject\_extracted\}$ ?
- Do you focus on parts of the  $\{subject\_extracted\}$  bodies to infer that they are  $\{subject\_extracted\}$ ?
- Do you focus on features or facial expressions of the  $\{subject\_extracted\}$  to result in saying they are  $\{subject\_extracted\}$ ?
- Do you focus on the  $\{subject\_extracted\}$  skin color to result in the  $\{subject\_extracted\}$  being  $\{subject\_extracted\}$ ?
- Do you focus on the  $\{subject\_extracted\}$  body shape to result in the result that they are  $\{subject\_extracted\}$ ?

## A.3. Object

- what is the background of the image?
- Is it rainy or sunny?
- Is it a closed or open place?
- Is the image bright?
- Are there elements in the image that can make visual noise?
- Are there a lot of different entities?
- Where is the image taken from?
- Are there objects?
- How many are  $\{subject\_extracted\}$  there?
- Do  $\{subject\_extracted\}$  have any elements that characterize them? Which ones?
- In what position are  $\{subject\_extracted\}$  located?
- which actions are shown in the image
- Could you indicate the reasons why that action is performed?
- would the  $\{subject\_extracted\}$  be able to perform any other action?
- Do you know what are the benefits or consequences of that action?
- How could you associate the  $\{subject\_extracted\}$  with their environment to describe the action?
- Are  $\{subject\_extracted\}$  new or old?

- What features could be differentiated from other types of *{subject\_extracted}*?
- What physical characteristics could be differentiated other *{subject\_extracted}*?

### 3.1 Preguntas debugging

- Have you been able to identify how many *{subject\_extracted}* there are?
- If the *{subject\_extracted}* change positions could you have another prediction of the situation described?
- What is the most complicated characteristic to define and determine?
- If the *{subject\_extracted}* were doing another action, would it be another type of *{subject\_extracted}*?

### 3.2 Preguntas efectiveness

- Are there elements in the image that could be noisy enough to intervene in the prediction that they are *{subject\_extracted}*?
- What physical traits are considered for the prediction result to be *{subject\_extracted}*?
- Why do you think identifying that they are *{subject\_extracted}* in the image is easy?

### 3.3 Preguntas efficiency

- What makes you think the *{subject\_extracted}* are actually *{subject\_extracted}* and not something else?
- With which other *{subject\_extracted}* would the prediction be most likely to give a better result?
- If there are subtle changes in color or posture, would the result of the prediction that they are *{subject\_extracted}* be the same?

### 3.4 Preguntas trust

- Is there anything that might create doubt or uncertainty for you to say they are *{subject\_extracted}*?

### 3.5 Preguntas persuasiveness

- Could you describe what is happening in the image?
- Could the situation you predicted be used in other areas?
- Could the situation you predicted have some example of counterfactual situation that could be useful?
- How could the situation you have predicted above be better?

### 3.6 Preguntas satisfaction

- What would an image similar to the situation found in this image look like?
- Can you describe a similar counterexample to the situation you have predicted above?
- What is the task or purpose of  $\{subject\_extracted\}$ ?
- Could you in simpler words describe their function that you have predicted before?

### 3.7 Preguntas education

- What could stand out from the situation described above and be useful in similar situations?
- Can you give essential concepts that are relevant to the situation described?
- What aspects of the image have been detected that may attract attention?

### 3.8 Preguntas transparency

- Why has it been obtained that the prediction is that they are  $\{subject\_extracted\}$ ?
- How did you get the result that they are  $\{subject\_extracted\}$ ?
- What do you focus on to get the prediction that they are  $\{subject\_extracted\}$ ?
- What do you focus on  $\{subject\_extracted\}$  to result in or say that they are  $\{subject\_extracted\}$ ?
- Do you focus on parts of the  $\{subject\_extracted\}$  bodies to infer that they are  $\{subject\_extracted\}$ ?
- Do you focus on features of the  $\{subject\_extracted\}$  to result in saying they are  $\{subject\_extracted\}$ ?
- Do you focus on the  $\{subject\_extracted\}$  color to result in the  $\{subject\_extracted\}$  being  $\{subject\_extracted\}$ ?
- Do you focus on the  $\{subject\_extracted\}$  shape to result in the result that they are  $\{subject\_extracted\}$ ?

## A.4. Landscape

- what is the background of the image?
- Is it rainy or sunny?
- Is it a closed or open place?
- Is the image bright?
- Are there elements in the image that can make visual noise?
- Are there a lot of different entities?
- Where is the image taken from?

- Are there humans?
- Are there animals?
- How many are  $\{subject\_extracted\}$  there?
- Do  $\{subject\_extracted\}$  have any elements that characterize them? Which ones?
- In what position are  $\{subject\_extracted\}$  located?
- which actions are shown in the image
- Could you indicate the reasons why that action is performed?
- would the  $\{subject\_extracted\}$  be able to perform any other action?
- Do you know what are the benefits or consequences of that action?
- How could you associate the  $\{subject\_extracted\}$  with their environment to describe the action?
- Are they  $\{subject\_extracted\}$  young or old?
- What features could be differentiated from other types of  $\{subject\_extracted\}$ ?
- What physical characteristics could be differentiated from the opposite  $\{subject\_extracted\}$ ?

#### 4.1 Preguntas debugging

- Have you been able to identify how many  $\{subject\_extracted\}$  there are?
- If the  $\{subject\_extracted\}$  change positions could you have another prediction of the situation described?
- If the  $\{subject\_extracted\}$  were doing another action, would it be another type of  $\{subject\_extracted\}$ ?

#### 4.2 Preguntas efectiveness

- Are there elements in the image that could be noisy enough to intervene in the prediction that they are  $\{subject\_extracted\}$ ?
- What physical traits are considered for the prediction result to be  $\{subject\_extracted\}$ ?
- Why do you think identifying that they are  $\{subject\_extracted\}$  in the image is easy?

#### 4.3 Preguntas efficiency

- What makes you think the  $\{subject\_extracted\}$  are actually  $\{subject\_extracted\}$  and not something else?
- With which other kind of  $\{subject\_extracted\}$  would the prediction be most likely to give a better result?
- If there are subtle changes in color or position, would the result of the prediction that they are  $\{subject\_extracted\}$  be the same?

#### 4.4 Preguntas trust

- Is there anything that might create doubt or uncertainty for you to say they are *{subject\_extracted}*?

#### 4.5 Preguntas persuasiveness

- Could you describe what is happening in the image?
- Could the situation you predicted be used in other areas?
- Could the situation you predicted have some example of counterfactual situation that could be useful?
- How could the situation you have predicted above be better?

## Categorías y etiquetas de base de casos

### B.1. Human

- *woman*
- *girl*
- *man*
- *baby*

### B.2. Animal

- *cattle*
- *dinosaur*
- *aquarium\_fish*
- *elephant*
- *wolf*
- *squirrel*
- *possum*
- *leopard*
- *rabbit*
- *snake*
- *lobster*
- *shark*
- *lion*
- *crocodile*

- *camel*
- *skunk*
- *raccoon*
- *snail*
- *worm*
- *turtle*
- *flatfish*
- *tiger*
- *ray*
- *porcupine*
- *crab*
- *lizard*
- *trout*

### B.3. Object

- *apple*
- *telephone*
- *cup*
- *bicycle*
- *television*
- *table*
- *clock*
- *streetcar*
- *tractor*
- *chair*
- *mushroom*
- *rocket*
- *plate*
- *motorcycle*
- *tank*
- *lawn\_mower*

- *bus*
- *couch*
- *bowl*
- *bed*
- *pickup\_truck*

## B.4. Landscape

- *sunflower*
- *sea*
- *pine\_tree*
- *mountain*
- *forest*
- *orchid*



# Apéndice C

## Ejemplos de secciones de formulario de evaluación online

Vaca

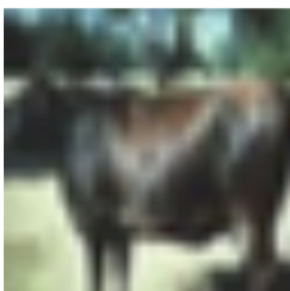


Segundo ejemplo similar (Opción B)

Identificar que los animales en la imagen son ganado es fácil porque son animales grandes, negros y blancos con marcas distintivas. El patrón negro y blanco en sus cuerpos es una característica común del ganado, lo que hace que sea fácil reconocerlos como tales. Además, la presencia de cuernos en los animales apoya aún más la identificación del ganado en la imagen.

La presencia de la vaca negra y blanca en la imagen sugiere que se trata de ganado. El patrón negro y blanco en el cuerpo de la vaca es una característica distintiva de las vacas, y no se ve comúnmente en otros animales. Además, la vaca está parada en un campo de pasto, que es un entorno típico para las vacas. La combinación de estas pistas visuales y el contexto de la imagen lleva a la conclusión de que el ganado son efectivamente vacas.

Primer ejemplo similar (Opción A)



¿Cuál de las dos opciones crees que ayuda a entender por qué la primera imagen es una vaca?

Descripción (opcional)


---

Selecciona una de las dos opciones \*

- Opción A
- Opción B
- Ninguna de las opciones me ayuda a entender la predicción lo más rápido posible.

Figura C.1: Ejemplo de primera sección de encuesta

**Lagartija**





**Segundo ejemplo similar (Opción B)**

Identificar que la imagen presenta un dinosaurio es fácil porque la imagen muestra una criatura grande y de cuello largo con una apariencia distintiva. Las características del dinosaurio, como su largo cuello, cuerpo grande y forma única, son fácilmente reconocibles y distinguibles de otros animales. Además, el contexto de la imagen, como la presencia de un juguete con temática de dinosaurio o la referencia a un dinosaurio, apoya aún más la idea de que la imagen representa a un dinosaurio.

La imagen presenta un dinosaurio con un cuello largo, que es una característica característica de muchos dinosaurios, como los dinosaurios saurópodos. La presencia de un cuello largo, junto con la forma y el tamaño general del dinosaurio, sugiere que es de hecho un dinosaurio y no otra cosa. La imagen también muestra la cabeza del dinosaurio, lo que apoya aún más la idea de que es un dinosaurio.

**Primer ejemplo similar (Opción A)**

¿Cuál de las dos opciones crees que ayuda a entender por qué la primera imagen es una vaca?

Descripción (opcional)

---

Selecciona una de las dos opciones \*

Opción A

Opción B

Ninguna de las opciones me ayuda a entender la predicción lo más rapido posible.

Figura C.2: Ejemplo de tercera sección de encuesta