UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Maestría en Explotación de Datos y Descubrimiento de Conocimiento

Respuesta a preguntas visuales en español: un enfoque para la asistencia a personas con discapacidad visual

Tesis presentada para optar al título de Magíster de la Universidad de Buenos Aires en Explotación de Datos y Descubrimiento de Conocimiento

Tesista: Lic. Clara Ofelia Villalba

Director: Dr. Juan Manuel Perez

Co-Directora: Dra. María Elena Buemi

Fecha de defensa: 8 de Octubre de 2025

Índice general

Abstract

La tarea de Respuesta a Preguntas Visuales (VQA, por sus siglas en inglés) consiste en desarrollar modelos de inteligencia artificial capaces de responder preguntas sobre una imagen. Si bien ha habido avances significativos en esta área, la mayoría de los modelos y datasets disponibles están en inglés, lo que limita su aplicabilidad en contextos de habla hispana. En este trabajo, se investiga el desempeño de diferentes enfoques de VQA en español utilizando una versión traducida del dataset VizWiz, con un enfoque particular en la asistencia a personas con discapacidad visual.

Para abordar esta tarea, se experimentó con modelos de fusión tardía de características, modelos basados en CLIP adaptados a VQA y distintos métodos de ensamble, incluyendo votación mayoritaria, fusión de características y meta-clasificadores. Los resultados muestran que los ensambles basados en votación con modelos CLIP multilingües lograron el mejor desempeño, sugiriendo que estos modelos capturan mejor la representación conjunta de imagen y texto en español.

Como parte de las contribuciones de este trabajo, se presenta el dataset VizWiz traducido al español para su uso en futuras investigaciones y se comparan distintos enfoques en un marco sistemático de evaluación. Estos hallazgos pueden servir como base para el desarrollo de sistemas más efectivos de VQA en español, con aplicaciones en accesibilidad y asistencia visual.

Palabras clave: Respuesta a Preguntas Visuales, VQA en español, CLIP, modelos de ensamble, accesibilidad, VizWiz.

Visual Question Answering in Spanish: An Approach for Assisting People with Visual Impairments

Abstract

The task of Visual Question Answering (VQA) consists of developing artificial intelligence models capable of answering questions about an image. Although there have been significant advances in this area, most available models and datasets are in English, which limits their applicability in Spanish-speaking contexts. This work investigates the performance of different VQA approaches in Spanish using a translated version of the VizWiz dataset, with a particular focus on assisting people with visual impairments.

To address this task, experiments were conducted with late-fusion feature models, CLIP-based models adapted for VQA, and various ensemble methods, including majority voting, feature fusion, and meta-classifiers. The results show that voting-based ensembles using multilingual CLIP models achieved the best performance, suggesting that these models better capture the joint representation of image and text in Spanish.

As part of this work's contributions, the VizWiz dataset translated into Spanish is presented for use in future research, and different approaches are compared within a systematic evaluation framework. These findings may serve as a foundation for the development of more effective VQA systems in Spanish, with applications in accessibility and visual assistance.

Keywords: Visual Question Answering, Spanish VQA, CLIP, ensemble models, accessibility, VizWiz.

Capítulo 1

Introducción

En la última década, la inteligencia artificial (IA) ha logrado avances significativos, especialmente en la convergencia entre la visión por computadora y el procesamiento del lenguaje natural (NLP). Una de las tareas emergentes en esta intersección es la **Respuesta a Preguntas Visuales (Visual Question Answering, VQA)** [1], cuyo objetivo es desarrollar modelos capaces de responder preguntas en lenguaje natural a partir de información contenida en imágenes. Esta tarea ha cobrado relevancia en múltiples áreas, como la educación, la recuperación de información y, particularmente, la asistencia a personas con discapacidad visual.

Sin embargo, a pesar del progreso en el desarrollo de modelos, la mayoría de las investigaciones en VQA se han llevado a cabo en inglés, tanto en los datasets utilizados como en los modelos entrenados. Esta dependencia del inglés limita la accesibilidad y la efectividad de estos sistemas en contextos donde se habla otro idioma, como el español. Además, el desempeño de los modelos actuales sigue siendo limitado en escenarios del mundo real, especialmente cuando las preguntas implican ambigüedades, lenguaje coloquial o un alto nivel de razonamiento semántico.

Entre las aplicaciones más relevantes del VQA se encuentra su uso como herramienta de apoyo para personas con discapacidad visual. Estos usuarios pueden beneficiarse de sistemas capaces de interpretar imágenes y responder preguntas sobre su entorno de manera automática. En esta línea, el dataset **VizWiz** [2] constituye un recurso fundamental: fue diseñado específicamente para esta población, a partir de imágenes reales capturadas por usuarios con discapacidad visual y preguntas formuladas de manera natural. No obstante, al estar disponible únicamente en inglés, se plantea la necesidad de adaptar tanto los datos como los modelos de VQA al idioma español, permitiendo así una mayor accesibilidad e inclusión tecnológica para comunidades hispanohablantes.

En esta tesis, se aborda esta problemática mediante la adaptación de modelos de VQA al español, considerando además su aplicación en contextos reales de asistencia. Para ello, se adopta un enfoque basado en la **clasificación multiclase**, una estrategia ampliamente utilizada en la

Q: ¿Cuántas monedas hay?
A: 4

Q: ¿De qué color es esto?
A: azul

Q: ¿Está encendida la luz?
A: no

Q: ¿De qué color es la remera?
A: rojo

Q: ¿Esto es helado?
A: sí

Q: ¿Cuál es el valor del billete?
A: 100

Figura 1.1: Ejemplos reales del dataset VizWiz, que incluyen preguntas con respuestas frecuentes del tipo “sí”, “no”, colores y cantidades numéricas.

literatura [1][3][4]. Este enfoque parte de la observación de que, si bien las respuestas en tareas de VQA pueden ser variadas, en la práctica tienden a concentrarse en un vocabulario limitado, compuesto por términos como “sí”, “no”, colores, números y otras expresiones comunes. En lugar de generar respuestas de forma libre, el modelo se entrena para seleccionar la opción más probable dentro de un conjunto predefinido. Cada respuesta posible se asocia a una clase específica, lo que permite reformular el problema como una tarea de clasificación supervisada.

En la Figura 1.1 se presentan ejemplos reales del dataset VizWiz que ilustran preguntas con respuestas típicas. En este trabajo, se implementa esta estrategia mediante la asignación de un índice único a cada respuesta dentro del conjunto de opciones, lo cual no solo favorece predicciones más precisas y coherentes, sino que también permite un mayor control y una evaluación más estructurada de las salidas generadas por el modelo. El enfoque adoptado para la tarea de **VQA** sigue una secuencia estructurada: primero, se extraen características relevantes del texto de la pregunta; luego, se obtienen características visuales de la imagen; y finalmente, ambas representaciones se combinan con el objetivo de generar una respuesta adecuada. Esta estrategia busca optimizar el rendimiento de los modelos, especialmente en el contexto del dataset VizWiz traducido

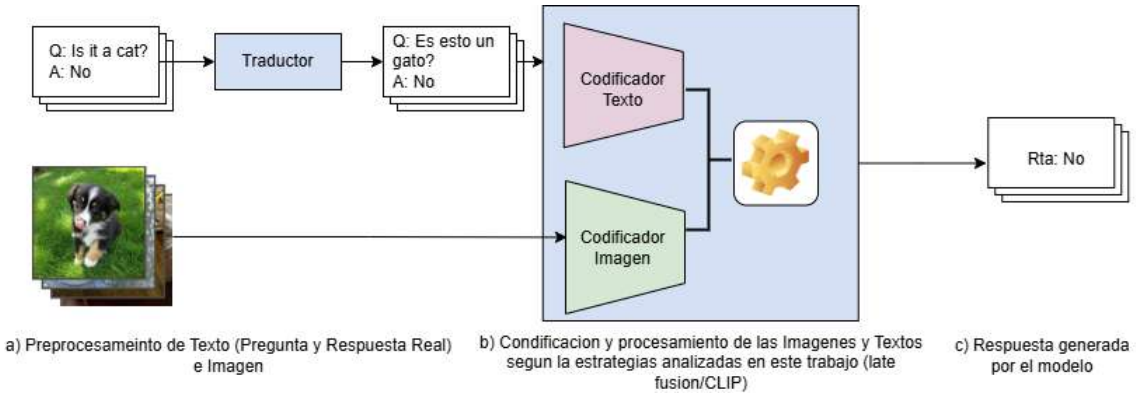


Figura 1.2: Esquema general de la metodología utilizada

al español.

El objetivo principal de esta investigación es evaluar y optimizar modelos de Visual Question Answering (VQA) en español, utilizando una versión traducida del dataset **VizWiz**. Para ello, se exploran diversas arquitecturas, incluyendo modelos de fusión tardía —en los que las representaciones de texto e imagen se procesan por separado antes de combinarse—, modelos basados en **CLIP (Contrastive Language-Image Pre-training)**[5], diseñados para aprender representaciones conjuntas de imágenes y texto en un espacio compartido, y técnicas de ensamble orientadas a mejorar la robustez de las predicciones mediante la combinación de múltiples modelos.

A través de los experimentos realizados, se busca dar respuesta a tres preguntas de investigación fundamentales: en primer lugar, se analiza el impacto que tiene la traducción del dataset VizWiz en el desempeño de los modelos de VQA en español; en segundo lugar, se examina qué enfoques resultan más efectivos para la tarea de VQA en este idioma, considerando tanto el desempeño de modelos individuales como el de las estrategias de ensamble; finalmente, se realiza una comparación entre los modelos evaluados en términos de métricas estándar, como **Accuracy** y **VizWiz Accuracy**[2], a fin de determinar su efectividad relativa.

Para alcanzar los objetivos planteados, se adopta un enfoque experimental basado en el uso de modelos de aprendizaje profundo. En primer lugar, se realiza el preprocesamiento del dataset VizWiz traduciendolo al español y asegurando una correcta alineación entre imágenes, preguntas y respuestas en español, de manera que los datos estén listos para el entrenamiento. A continuación, se procede al entrenamiento y evaluación de distintos modelos de VQA, considerando tanto arquitecturas de fusión tardía como modelos basados en CLIP, con el objetivo de identificar los enfoques más prometedores. Posteriormente, se aplican

técnicas de ensamble, tales como la votación mayoritaria y la combinación de embeddings, con el fin de evaluar su capacidad para mejorar el desempeño de los modelos individuales. Finalmente, se lleva a cabo un análisis exhaustivo de los resultados obtenidos, contrastándolos con estudios previos, y examinando las principales fortalezas y limitaciones de cada estrategia propuesta. La metodología general se puede observar en la Figura 1.2.

Esta investigación realiza diversas contribuciones importantes al campo de la **Respuesta a Preguntas Visuales (VQA)**, especialmente en lo que respecta a la adaptación de modelos al idioma español y su aplicación en el contexto de la accesibilidad para personas con discapacidad visual. A continuación, se presentan los principales aportes de este trabajo. En primer lugar, se ha traducido el dataset **VizWiz** al español, un conjunto de datos ampliamente utilizado en la comunidad de VQA, originalmente diseñado para asistir a personas con discapacidad visual mediante respuestas generadas a partir de imágenes capturadas por usuarios reales. La falta de recursos disponibles en español ha limitado el progreso en la investigación en este contexto. La versión traducida del dataset preserva la estructura original, asegurando que las preguntas y respuestas mantengan su coherencia y significado en español, sin comprometer la tarea de VQA. Este nuevo recurso será de gran utilidad para la comunidad científica, facilitando la investigación sobre modelos que puedan responder preguntas visuales en español con mayor precisión y efectividad.

En segundo lugar, se realizó una evaluación comparativa de distintos enfoques de modelos para la tarea de VQA en español. Entre los modelos evaluados se incluyen aquellos basados en la fusión tardía de representaciones, modelos multimodales preentrenados y técnicas de ensamble. El análisis detallado de estos enfoques proporciona información valiosa sobre el rendimiento de los modelos en un contexto donde la integración de procesamiento de lenguaje natural y visión computacional es fundamental.

Una tercera contribución importante de este trabajo fue la aplicación de técnicas de ensamble para mejorar el desempeño de los modelos de VQA. Se exploraron métodos como la votación mayoritaria, la fusión de características y el uso de meta-clasificadores, con el objetivo de optimizar el desempeño de los sistemas. Los resultados obtenidos demuestran que la combinación de modelos complementarios puede aumentar significativamente la capacidad de respuesta, especialmente en preguntas que requieren un razonamiento más complejo o cuando los modelos individuales generan respuestas inconsistentes.

Finalmente, se llevó a cabo un análisis del impacto del idioma en los modelos preentrenados de VQA. Dado que muchos de estos modelos fueron entrenados principalmente en inglés, se identificaron las dificultades que enfrentan al procesar preguntas en español. En particular, se observó que modelos preentrenados como CLIP Multilingual presentan limitaciones en la generación y selección de respuestas en español. Este hallazgo resalta la necesidad de adaptar o preentrenar modelos específicamente para el idioma español, lo cual es crucial para mejorar su rendimiento en este contexto.

En resumen, este trabajo no solo aporta nuevos recursos y conocimientos al campo de la VQA en español, sino que también facilita el desarrollo de herramientas de asistencia para personas con discapacidad visual en entornos de habla hispana. Las contribuciones realizadas amplían la comprensión de cómo mejorar el desempeño de los modelos de VQA en español y ofrecen valiosos recursos para la comunidad investigadora, alentando el avance de este campo emergente.

Este trabajo se estructura en los siguientes capítulos:

- **Capítulo 2: Estado del Arte**, que revisa investigaciones previas en modelos de VQA y su aplicación en diferentes idiomas.
- **Capítulo 3: Materiales y Métodos**, que detalla el preprocesamiento del dataset, los modelos y metodologías utilizados asi como las métricas de evaluación.
- **Capítulo 4: Resultados y Discusión**, donde se presentan los experimentos realizados, los análisis comparativos y las interpretaciones de los resultados obtenidos.
- **Capítulo 5: Conclusión y Trabajos Futuros**, que resume los hallazgos de la investigación y sugiere futuras líneas de trabajo.

Capítulo 2

Estado del Arte en Visual Question Answering (VQA)

En este capítulo se revisan los principales avances en el campo del **Visual Question Answering (VQA)**, con especial énfasis en los enfoques que plantean esta tarea como un problema de **clasificación multiclase**. Además, se describen los conjuntos de datos más utilizados en la literatura para el entrenamiento y evaluación de modelos, y se presenta el dataset **VizWiz**, empleado en esta tesis, junto con estudios relevantes que lo han aplicado previamente. Se destaca, en particular, el enfoque basado en **CLIP** adaptado a VQA [6], que constituye una de las bases metodológicas de este trabajo. Esta revisión tiene como objetivo contextualizar la metodología propuesta, centrada en la adaptación de modelos de VQA al idioma español, con una aplicación específica orientada a mejorar la accesibilidad de personas con discapacidad visual.

La tarea de **Visual Question Answering (VQA)** se sitúa en la intersección entre la visión por computadora y el procesamiento del lenguaje natural. Su objetivo es desarrollar sistemas capaces de responder de manera automática preguntas formuladas en lenguaje natural sobre el contenido de una imagen. Aunque esta área ha experimentado un notable crecimiento en la última década, impulsada por los avances en aprendizaje profundo y la disponibilidad de grandes conjuntos de datos anotados, sus orígenes conceptuales se remontan a finales de los años 60. En los primeros trabajos, se exploraba la posibilidad de generar respuestas a partir de representaciones visuales [7]. Esta idea está estrechamente relacionada con evaluaciones clásicas de inteligencia artificial, como la prueba total de Turing [8], que plantea que un sistema verdaderamente inteligente debería ser capaz de comprender imágenes y responder preguntas sobre ellas de manera natural.

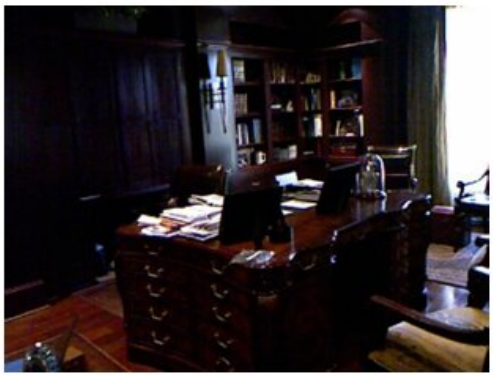
En los últimos años, la tarea de Respuesta a Preguntas Visuales (VQA, por sus siglas en inglés) se ha consolidado como un área de investigación activa, impulsada por el desarrollo de arquitecturas basadas en redes neuronales profundas. Una estrategia comúnmente adoptada consiste en reformular el problema como una clasificación multiclase, donde el modelo

selecciona la respuesta más probable a partir de un conjunto cerrado de opciones predefinidas. Este enfoque facilita la evaluación automática de los modelos y contribuye a mejorar la coherencia y exactitud de las respuestas. Algunos estudios han analizado cómo estructurar los modelos de VQA como clasificadores que asignan a cada posible respuesta una clase dentro de un vocabulario limitado, demostrando que incluso aproximaciones basales pueden ofrecer un rendimiento competitivo en escenarios bien definidos [1][3][4]. Complementariamente, otros trabajos proponen un enfoque novedoso basado en un mecanismo de *co-attention*, que permite alinear de forma más eficaz las preguntas con las regiones relevantes de la imagen [9]. Este modelo aborda uno de los principales retos del VQA: la gran diversidad y complejidad de las preguntas, que pueden requerir desde tareas simples como detección de objetos hasta operaciones más avanzadas como conteo, segmentación o reconstrucción. En lugar de entrenar un único modelo para resolver todas estas tareas desde cero, los autores integran algoritmos de visión ya existentes en el sistema, permitiendo que el modelo aprenda a usarlos de manera selectiva según el tipo de pregunta. Esta estrategia no solo mejora el rendimiento en tareas de clasificación de respuestas, sino que también permite generar explicaciones interpretables para las decisiones del sistema, manteniendo la posibilidad de un entrenamiento end-to-end.

2.1. Principales Datasets y Estudios en Visual Question Answering

A partir de 2014, comenzaron a publicarse diversos trabajos que presentaron conjuntos de datos que impulsaron significativamente el desarrollo de modelos para VQA, al proporcionar ejemplos anotados de pares (imagen,pregunta-respuesta) en gran escala. Estos datasets no solo han facilitado el entrenamiento supervisado de modelos complejos, sino que también han establecido estándares comunes para la evaluación y comparación de distintos enfoques. Entre los más influyentes se encuentran el dataset **DAQUAR**[10]. Fue el primer conjunto de datos importante publicado específicamente para VQA. Consta de 6,795 ejemplos de entrenamiento y 5,673 de validación, basados en las imágenes del NYU-DepthV2 Dataset [11]. Aunque pionero, su tamaño reducido y el hecho de que contiene exclusivamente escenas de interiores lo hacen inadecuado para entrenar modelos complejos. Además, las condiciones de iluminación extremas y el desorden en las imágenes dificultan la tarea de responder preguntas con precisión, alcanzando los humanos solo un 50.2 % de exactitud en el conjunto de datos completo.

COCO-QA: What does an intersection show on one side and two double-decker buses and a third vehicle,?
Ground Truth: Building



DAQUAR: What is behind the computer in the corner of the table?
Ground Truth: papers

Figura 2.1: Ejemplos de los dataset COCO y DAQUAR con sus correspondientes par pregunta/respuesta asociadas

Otro dataset relevante en el campo es **COCO-QA**[12]. En este caso, los pares pregunta-respuesta fueron generados automáticamente a partir de imágenes del dataset COCO [13] mediante algoritmos de procesamiento de lenguaje natural (PLN). Consiste en 78,736 datos de entrenamiento y 38,948 de prueba. Sin embargo, este dataset presenta problemas debido a la generación automática de preguntas, lo que resulta en errores gramaticales y preguntas formuladas incorrectamente. La mayor deficiencia de COCO-QA se debe al algoritmo de NLP que se utilizó para generar los pares pregunta-respuesta. Las oraciones más largas se dividen en partes más pequeñas para facilitar el procesamiento, pero en muchos de estos casos el algoritmo no se adapta bien a la presencia de cláusulas y variaciones gramaticales en la formación de oraciones. Esto da como resultado preguntas formuladas de manera incorrecta, muchas de las cuales contienen errores gramaticales y otras son completamente ininteligibles. La otra deficiencia importante es que solo tiene cuatro tipos de preguntas, y éstas se limitan a los tipos de cosas descritas en los subtítulos de COCO. En la Figura 2.1 se presentan ejemplos de los datasets mencionados anteriormente.

También destaca el conjunto de datos **VQAV1**[14]. Este dataset fue construido a partir de imágenes reales del conjunto COCO [13], complementadas con dibujos animados abstractos, y contiene un total de 614,163 preguntas distribuidas en los conjuntos de entrenamiento (248,349), validación (121,512) y prueba (244,302). Cada imagen tiene asociadas tres preguntas, y cada pregunta cuenta con diez respuestas proporcionadas por distintos anotadores a través de Amazon Mechanical Turk (AMT), quienes debían formular preguntas que "desafiaran a un robot inteligente" (stump a smart robot). Este conjunto de datos representó un hito para el desarrollo de modelos que integran visión y lenguaje, al proporcionar una

Figura 2.2: Ejemplos del balanceo realizado en el dataset VQA V2

base común para la comparación de enfoques. Los autores utilizaron un enfoque de clasificación multiclase, seleccionando respuestas desde un vocabulario limitado derivado de las respuestas más frecuentes. Esto sienta las bases del enfoque de clasificación como una práctica estándar en VQA. Sin embargo, presentaba una limitación importante: muchas preguntas podían responderse correctamente sin analizar la imagen, simplemente aprovechando patrones estadísticos en las respuestas más comunes.

Para abordar las limitaciones asociadas al sesgo en las respuestas de los sistemas de VQA, se desarrolló el conjunto de datos **VQA v2**[4], cuyo objetivo principal fue mitigar la dependencia de los modelos respecto a patrones lingüísticos predecibles. Esta versión introduce una estrategia basada en pares balanceados de imágenes: preguntas idénticas se asocian a diferentes imágenes, pero con respuestas distintas, lo que obliga a los modelos a fundamentar sus predicciones en el contenido visual específico de cada escena. De esta manera, se fomenta un razonamiento visual más riguroso y se reduce la posibilidad de que las respuestas se basen únicamente en correlaciones estadísticas del lenguaje. El dataset incluye aproximadamente 83,000 imágenes provenientes de MS COCO y cerca de 448,000 pares pregunta-imagen, distribuidos entre imágenes reales del conjunto COCO-VQA y escenas sintéticas del subconjunto SYNTH-VQA, estas últimas diseñadas para ampliar la diversidad visual y evaluar la generalización de los modelos ante distintos contextos.

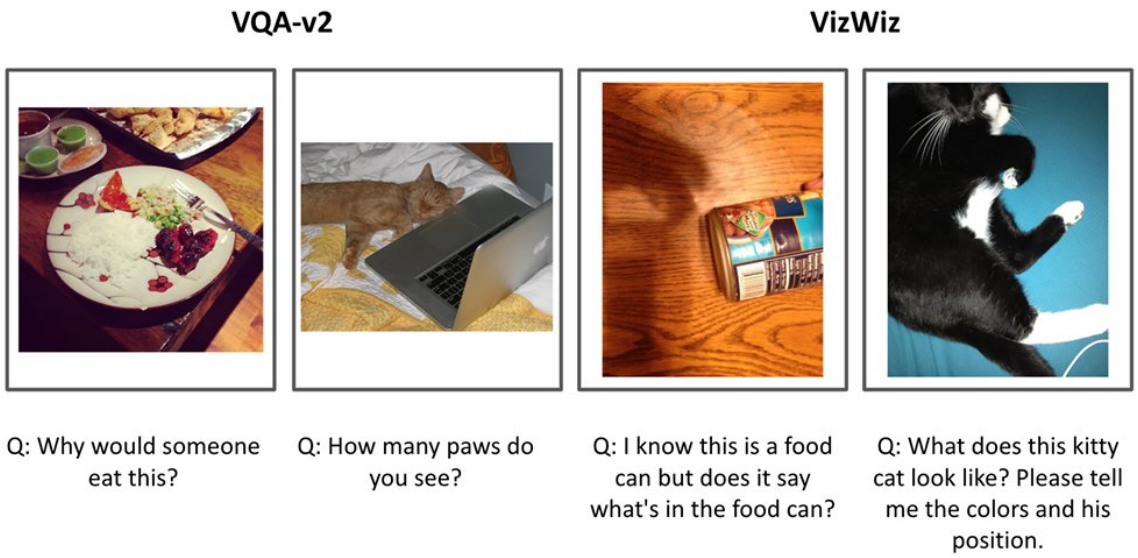


Figura 2.3: Dada imagenes de contenido similar (izquierda: comida, derecha: gato). Las preguntas del dataset VQA V2 y VIZWIZ son sustancialmente diferentes. Las preguntas del dataset VIZWIZ son mucho más especificas y menos artificiales que las del otro Dataset [15]

La Figura 2.2 muestra ejemplos representativos del conjunto de datos balanceado. En conjunto, **VQA V2** representa una evolución significativa respecto a su predecesor, al mejorar la calidad del dataset y reducir el sesgo, consolidándose como una referencia clave para la evaluación de modelos de VQA.

2.1.1. VizWiz: Dataset pionero en VQA Accesible

Los conjuntos de datos mencionados anteriormente fueron generados a partir de imágenes obtenidas mediante búsquedas web, con preguntas formuladas por personas videntes o generadas de forma automática. Esta metodología dificulta su adaptación a contextos del mundo real, especialmente en entornos relacionados con la accesibilidad. Por este motivo, en esta tesis se opta por utilizar el conjunto de datos **VizWiz**[2], que a diferencia de los datasets previos, se compone de imágenes capturadas por personas con discapacidad visual, acompañadas de preguntas formuladas por ellas mismas y respuestas proporcionadas por videntes. Este conjunto de datos, con aproximadamente 32,000 pares imagen-pregunta, incluye además grabaciones de preguntas habladas y sus respectivas transcripciones. Fue creado a partir de una aplicación de colaboración abierta, donde para cada pregunta, se recopilaron hasta diez respuestas diferentes, con el fin de reflejar la variabilidad en la percepción y comprensión de las imágenes. Cabe destacar que este conjunto de datos fue desarrollado como parte de un desafío de inteligencia artificial, cuyo objetivo era crear

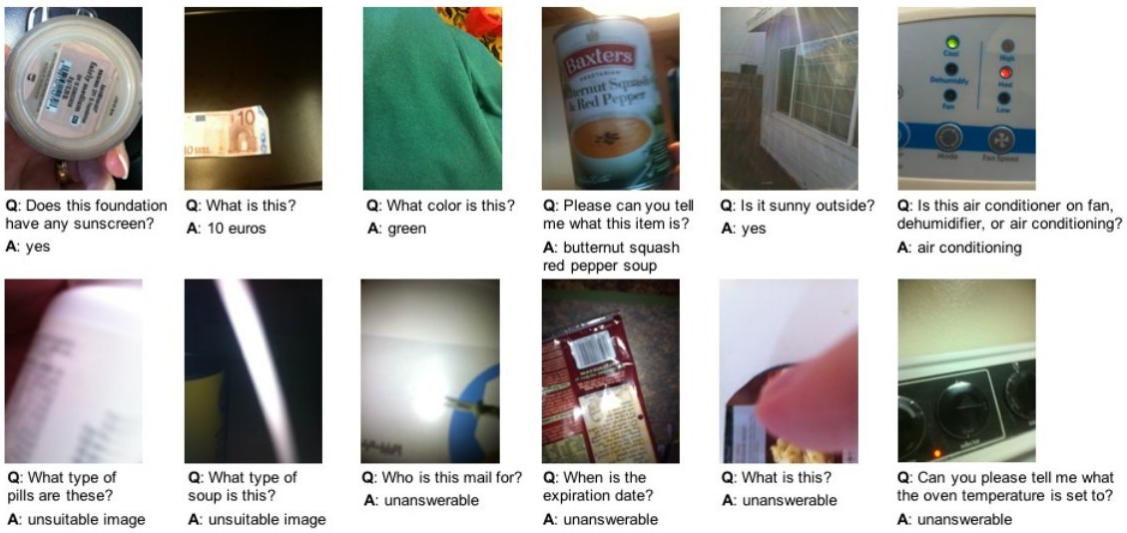


Figura 2.4: Ejemplos del Dataset VizWiz disponibles en la web del desafío [16]

modelos de VQA capaces de responder preguntas visuales formuladas por personas ciegas. En la Figura 2.3 se presenta una comparación entre los datasets VQA V2 y VizWiz, resaltando sus diferencias en cuanto a la distribución de imágenes, tipos de preguntas y sesgos en las respuestas. Además en la Figura 2.4 se ilustran algunos ejemplos de imágenes del conjunto de datos **VizWiz** con sus respectivos pares pregunta-respuesta en el idioma original.

La naturaleza del dataset VizWiz, basado en datos del mundo real, permite a los investigadores abordar desafíos específicos relacionados con la accesibilidad, como la variabilidad en la calidad de las imágenes tomadas por personas ciegas, la ambigüedad en las preguntas y la diversidad de respuestas entregadas por los anotadores. Por esta razón, este conjunto de datos ha sido ampliamente empleado en investigaciones que abordan la tarea de VQA desde una perspectiva centrada en la accesibilidad. En uno de los estudios más representativos se propone el primer dataset que vincula visualmente las respuestas con las preguntas formuladas por personas con discapacidad visual, lo que permite una comprensión más profunda del contexto en el que estas preguntas son generadas [17]. Siguiendo esta línea, otro trabajo introduce un conjunto de datos diseñado para tareas de localización de objetos mediante técnicas de aprendizaje con pocos ejemplos, utilizando también imágenes tomadas por personas con discapacidad visual [18]. Además, una investigación reciente compara datasets tradicionales de VQA con aquellos centrados específicamente en accesibilidad, analizando el rendimiento de distintos modelos y destacando las diferencias fundamentales en cuanto a objetivos y desafíos inherentes a cada enfoque[19].

2.2. Transformers en Tareas Multimodales: Fundamentos de Modelos NLP y Visión Relevantes

Como se ha señalado previamente, la tarea de Visual Question Answering (VQA) representa un desafío intrínsecamente multimodal, en el que el sistema debe generar una respuesta coherente y precisa en lenguaje natural a partir de una imagen y una pregunta también formulada en lenguaje natural. Esta tarea exige una comprensión profunda tanto del contenido visual como del contexto lingüístico, lo que implica establecer una integración semántica efectiva entre ambas modalidades: imagen y texto.

Dada esta complejidad, VQA abarca múltiples subproblemas pertenecientes a los campos de la Visión por Computadora (Computer Vision, CV) y el Procesamiento de Lenguaje Natural (Natural Language Processing, NLP), lo que ha motivado el desarrollo de enfoques híbridos que combinan modelos especializados en cada dominio. En línea con este enfoque, el presente trabajo incorpora una selección de modelos visuales y de lenguaje que han demostrado alta efectividad en tareas multimodales. A continuación, se describen los principales modelos empleados, comenzando por aquellos orientados al procesamiento del lenguaje natural, seguidos de los transformadores utilizados para el análisis del contenido visual.

En primer lugar, se resenta a BERT (Bidirectional Encoder Representations from Transformers)[20], un modelo desarrollado por Google que ha demostrado un rendimiento sobresaliente en diversas tareas de NLP. BERT se basa en la arquitectura Transformer [21] y aprende representaciones de palabras de forma bidireccional y contextualizada, lo que lo convierte en una herramienta particularmente adecuada para capturar las sutilezas del lenguaje en preguntas complejas. Gracias a esta capacidad, BERT puede interpretar el significado de una palabra considerando el contexto completo de la oración en la que aparece, lo que permite una comprensión más precisa y enriquecida del texto. Durante su entrenamiento, utiliza una técnica de *masked language modeling* (MLM), en la que ciertas palabras de una oración son enmascaradas aleatoriamente, y el modelo debe predecirlas a partir del contexto. Esta estrategia le permite capturar relaciones semánticas complejas y generar representaciones robustas aplicables a diversas tareas de procesamiento de lenguaje natural (NLP). En la Figura 2.5 se muestra un esquema de alto nivel del modelo.

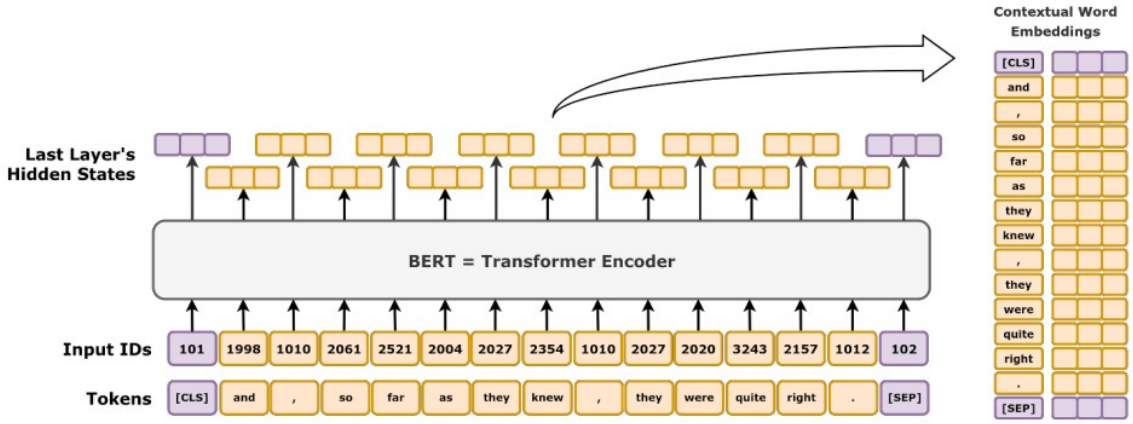


Figura 2.5: Diagrama esquemático de alto nivel de BERT. Toma un texto, lo tokeniza en una secuencia de tokens, agrega tokens especiales opcionales y aplica un codificador Transformer. Los estados ocultos de la última capa se pueden usar luego como representaciones contextuales de palabras. By Daniel Voigt Godoy - https://dvgodoy.github.io/dl-visuals/BERT/

Otro modelo de lenguaje utilizado en este trabajo es BETO(Bidirectional Encoder Representations from Transformers for Spanish) [22], una adaptación específica de BERT entrenada para el procesamiento de lenguaje natural en español. Al igual que BERT, utiliza la arquitectura de transformers para aprender representaciones bidireccionales de las palabras, lo que le permite captar relaciones contextuales profundas entre las palabras en un texto, considerando el significado de una palabra en función de las palabras que la rodean. Lo que diferencia a BETO es que ha sido preentrenado con un corpus extenso de texto en español, lo que le otorga un conocimiento profundo de la lengua y las particularidades culturales, gramaticales y semánticas del español. Este preentrenamiento le permite ser utilizado en una amplia gama de tareas de procesamiento de lenguaje natural, como análisis de sentimientos, clasificación de texto, reconocimiento de entidades, entre otros. El modelo BETO ha demostrado un rendimiento excepcional en diversas tareas de NLP en español, destacándose en comparación con otros modelos entrenados en idiomas como el inglés. Su uso generalizado en la comunidad de NLP de habla hispana refleja su eficacia y robustez en la captura de las complejidades del idioma. De hecho, BETO ha sido uno de los modelos más adoptados para tareas de NLP en español, debido a su alta capacidad de generalización y su precisión en tareas que requieren una comprensión profunda del contexto lingüístico.

Por ultimo el modelo RoBERTa (Robustly Optimized BERT Approach)[23]. Aunque comparte la misma arquitectura de transformers que BERT, RoBERTa incorpora modificaciones clave en su preentrenamiento y optimización, lo que lo hace más eficiente y robusto. Durante el preentre-

namiento, emplea la técnica de "Denoising Autoencoder", que consiste en eliminar palabras aleatorias del texto de entrada y entrenar al modelo para reconstruir el texto original. Entre las principales mejoras de RoBERTa respecto a BERT se incluyen el enmascaramiento dinámico, donde los tokens enmascarados varían en cada época en lugar de fijarse al inicio del entrenamiento como en BERT; la agrupación de oraciones, que permite combinar oraciones para alcanzar un límite de 512 tokens y abarcar así múltiples documentos dentro del contexto; el uso de lotes más grandes durante el entrenamiento, lo que mejora tanto la estabilidad como la eficiencia del modelo; y la implementación de un vocabulario BPE a nivel de bytes, que, a diferencia de BERT, utiliza codificación BPE basada en bytes en lugar de caracteres, permitiendo un procesamiento más efectivo de caracteres Unicode. Adicionalmente, RoBERTa aprovecha un esquema de entrenamiento con lotes dinámicos, que ajusta el tamaño de los lotes según los recursos disponibles, optimizando aún más el proceso. Gracias a estas mejoras, RoBERTa ha demostrado un rendimiento superior en diversas tareas de procesamiento de lenguaje natural, destacándose por su capacidad para manejar grandes volúmenes de datos y generar representaciones de texto de alta calidad. El modelo RoBERTa original no es multilingüe, ya que fue entrenado exclusivamente en inglés utilizando un conjunto de datos masivo en este idioma. Sin embargo, en este trabajo empleamos una versión preentrenada en español, desarrollada bajo el proyecto BERTIN y disponible en Hugging Face [24]. BERTIN es una serie de modelos basados en BERT, diseñados específicamente para el español.

En relacion a los modelos visuales utilizados, en primer lugar se encuentra Vision Transformer (ViT) [26] que representa una innovación reciente en el campo de la visión por computadora al adaptar la arquitectura de transformers, originalmente diseñada para el procesamiento de lenguaje natural, para la codificación de imágenes. A diferencia de las redes neuronales convolucionales (CNN) tradicionales, que emplean convoluciones para procesar imágenes, ViT utiliza transformers para capturar las relaciones globales entre los píxeles de la imagen.

Como se ilustra en la Figura 2.6, el funcionamiento del modelo consiste en dividir la imagen en parches de tamaño fijo, tratando cada uno de estos parches como una secuencia de entrada para el transformer. Posteriormente, el transformer procesa cada parche de manera secuencial, lo que permite al modelo aprender representaciones ricas de las características de la imagen. Este enfoque, basado en relaciones globales entre los píxeles, le permite a ViT capturar dependencias de largo alcance, lo cual contribuye a su rendimiento destacado en tareas de visión por

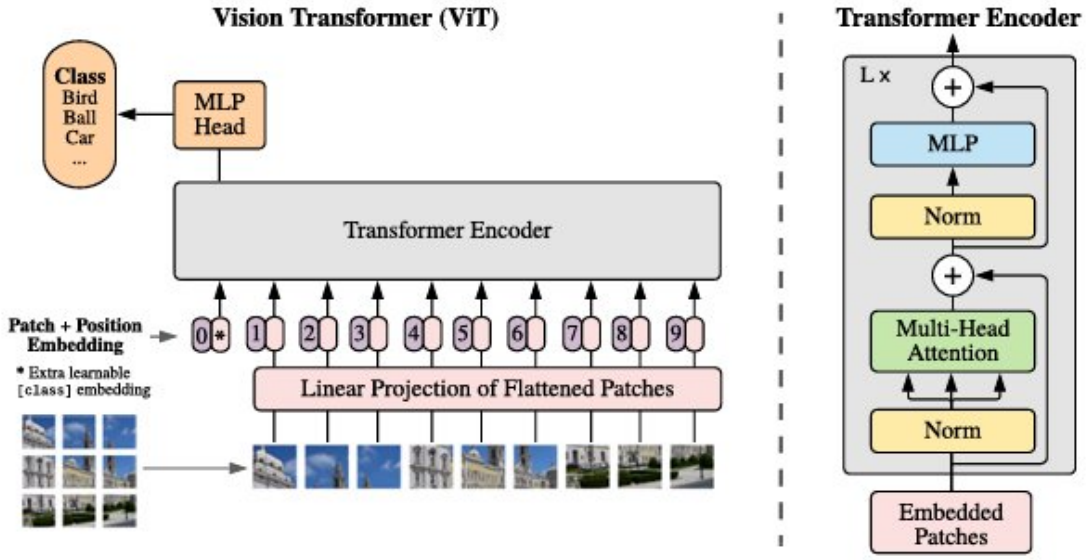


Figura 2.6: Visión general del modelo - El modelo divide una imagen en parches, les asigna embeddings, y usa un Transformer para procesarlos. Para clasificación, se añade un token especial a la secuencia. La arquitectura del Transformer se inspira en el trabajo de Vaswani et al. [25]

computadora, como la clasificación de imágenes y la segmentación.

Este modelo ha sobresalido por su capacidad para aprender representaciones globales de las imágenes, mostrando un rendimiento competitivo en comparación con las arquitecturas tradicionales basadas en convoluciones.

El otro modelo visual utilizado es el modelo BEiT (BERT-inverted Encoder Transformers)[27] ofrece un enfoque igualmente innovador al adaptar la arquitectura de transformers al campo de la visión por computadora. A diferencia de ViT, que trata las imágenes como secuencias de parches, BEiT emplea una estrategia inversa: las características de la imagen se transforman primero en un espacio de características semánticas mediante un codificador transformer. Posteriormente, estas representaciones semánticas se utilizan para realizar tareas específicas de visión por computadora. Este proceso se puede observar en la Figura 2.7. Este enfoque permite que BEiT capture de manera más efectiva las relaciones semánticas entre las características de la imagen, lo que mejora su desempeño en diversas tareas, como clasificación, segmentación y detección.

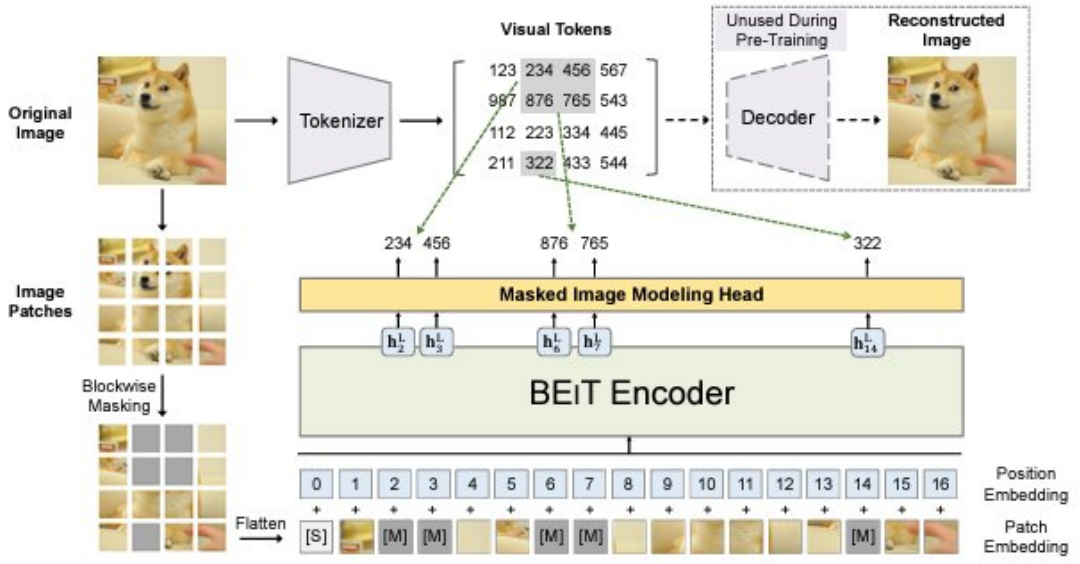


Figura 2.7: Visión general del preentrenamiento de BEiT - El modelo BEiT utiliza un enfoque de preentrenamiento basado en la tokenización de imágenes, donde se aprenden representaciones visuales discretas mediante un proceso de reconstrucción. Durante el preentrenamiento, algunos parches de la imagen se enmascaran y se reemplazan por una máscara especial. El objetivo es predecir los tokens visuales originales a partir de la versión corrompida de la imagen, utilizando un transformer de visión.

2.3. CLIP en Visual Question Answering: Arquitectura y Resultados Relevantes en VizWiz

Una línea de investigación reciente ha explorado la aplicación del modelo CLIP [5] en la tarea de Visual Question Answering, aprovechando sus capacidades para generar representaciones conjuntas de texto e imagen. En particular, se ha demostrado que un enfoque basado en características extraídas por CLIP, combinadas únicamente con capas lineales, puede alcanzar un rendimiento competitivo en el dataset VizWiz [6], sin necesidad de recurrir a arquitecturas complejas. Este enfoque destaca por su simplicidad y eficiencia, lo que lo convierte en una alternativa especialmente atractiva en escenarios donde los recursos computacionales o lingüísticos son limitados. Además, sus resultados subrayan la versatilidad del modelo CLIP en contextos multimodales y el potencial del dataset VizWiz como banco de pruebas para soluciones que buscan mejorar la autonomía de personas con discapacidad visual

CLIP (Contrastive Language-Image Pre-training) [5] es un modelo multimodal que aprende representaciones alineadas de texto e imagen mediante un esquema de aprendizaje contrastivo a gran escala. Esta capacidad para capturar correspondencias semánticas profundas entre modalidades lo convierte en una base sólida para diversas tareas multimodales, incluido

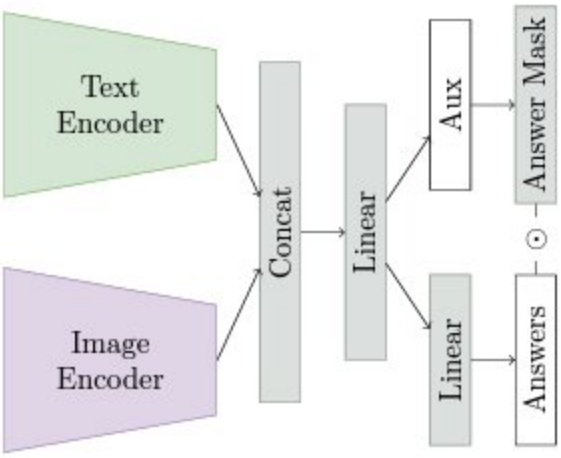


Figura 2.8: Esquema del modelo CLIP adaptado para VQA[6]

el VQA. En el contexto específico del dataset VizWiz, se ha propuesto una adaptación simplificada de CLIP para esta tarea, en la que las representaciones generadas por el modelo se combinan con capas lineales. Este enfoque demuestra que es posible obtener un rendimiento competitivo sin recurrir a arquitecturas complejas, al aprovechar directamente el poder expresivo de las características aprendidas por CLIP [6].

Los enfoques anteriores basados en CLIP para la tarea de respuesta a preguntas visuales (VQA) suelen utilizar únicamente el codificador de imágenes [28], o bien generan indicaciones para asociar preguntas y respuestas. Sin embargo, el enfoque presentado en el mencionado artículo aprovecha tanto el codificador de imágenes como el de texto de CLIP. Las características generadas por ambos codificadores se concatenan y se procesan mediante capas lineales, que incluyen normalización por capas y una alta tasa de deserción (0.5).

Como se muestra en la Figura 2.8, este método predice tanto los tipos de respuesta como las respuestas finales mediante una capa lineal adicional. El clasificador lineal se entrena utilizando una pérdida de entropía cruzada, incorporando técnicas de aumento de datos como la rotación de imágenes. Cabe destacar que solo se entrena el clasificador lineal, mientras que el modelo CLIP preentrenado se utiliza como codificador de imágenes y texto, permaneciendo completamente congelado durante el entrenamiento. Esto permite realizar el proceso de aprendizaje de manera rápida y eficiente, sin requerir recursos computacionales significativos, incluso al trabajar con el conjunto de datos **VizWiz**.

En el caso de CLIP aplicado a la tarea de VQA, el modelo mantiene una limitación importante: su desempeño está optimizado para el inglés, lo que afecta negativamente su rendimiento cuando se utiliza con datos en

otros idiomas. Para superar esta barrera, se han propuesto adaptaciones que permiten extender sus capacidades al ámbito multilingüe sin necesidad de reentrenar el modelo completo. Una estrategia eficaz consiste en reemplazar el codificador de texto original por uno multilingüe preentrenado, como XLM-R, mientras se mantiene intacto el codificador de imágenes de CLIP [29] [30]. Este enfoque modular permite preservar la calidad de las representaciones visuales del modelo original, al tiempo que se introducen capacidades multilingües mediante una integración adecuada entre las representaciones textuales y visuales. Gracias a esta configuración, es posible adaptar CLIP de manera eficiente a tareas multimodales como VQA en otros idiomas, como el español, sin incurrir en altos costos computacionales ni requerir grandes volúmenes de datos paralelos. A continuacion resentamos algunas alternativas multilinques de CLIP adaptado para VQA.

El primer modelo que presentamos es el **CLIP-ViT-B-32-multilingual-v1**, una extensión multilingüe del modelo original **CLIP-ViT-B/32**, diseñada para mapear imágenes y texto en más de 50 idiomas dentro de un espacio vectorial compartido. Su capacidad multilingüe se logró mediante un proceso de *Multilingual Knowledge Distillation*, en el cual un **DistilBERT multilingüe** fue entrenado como modelo estudiante utilizando el espacio vectorial del modelo docente **CLIP-ViT-B/32** original. Durante este entrenamiento, se emplearon datos de texto paralelos para alinear las representaciones lingüísticas en múltiples idiomas sin modificar el codificador de imágenes, que se mantiene idéntico al de la versión original de CLIP.

El uso de esta versión multilingüe en la tarea de Visual Question Answering (VQA) en español ofrece varias ventajas significativas. En primer lugar, permite procesar preguntas en español de forma nativa, eliminando la necesidad de traducirlas al inglés, como ocurre con versiones estándar de CLIP. Esto no solo reduce la latencia y el costo computacional asociado con la traducción, sino que también evita pérdidas de información y ambigüedades semánticas derivadas del proceso de traducción automática. Además, al contar con representaciones lingüísticas alineadas en múltiples idiomas, el modelo mejora la comprensión del texto en español, facilitando una asociación más precisa entre la pregunta y la imagen. También proporciona mayor robustez frente a variaciones lingüísticas y estructuras gramaticales propias del español, optimizando su desempeño en tareas de VQA dirigidas a usuarios hispanohablantes.

Otra ventaja clave de este modelo es su capacidad para representar texto e imágenes dentro de un espacio vectorial común, lo que favorece la

generalización en tareas multimodales. Esto resulta especialmente útil en aplicaciones como la búsqueda de imágenes a partir de descripciones textuales en distintos idiomas, sin la necesidad de recurrir a modelos específicos para cada lengua.

La arquitectura del modelo se compone de tres elementos principales. El primero es un codificador de imágenes basado en **CLIP-ViT-B-32** original, preentrenado por OpenAI. A este se suma un codificador de texto multilingüe, construido a partir de DistilBERT y alineado con el espacio vectorial de CLIP mediante un proceso de distilación. Finalmente, ambos codificadores se integran en un espacio vectorial compartido que permite representar de manera conjunta texto e imágenes, facilitando así el razonamiento multimodal. Gracias a esta arquitectura, **CLIP-ViT-B-32-multilingual-v1** combina las capacidades avanzadas de CLIP con soporte multilingüe, ampliando su aplicabilidad en contextos donde resulta crucial trabajar eficientemente con múltiples idiomas.

Otro modelo en la misma linea es **LaBSE** basado también en *transformers* y diseñado para generar embeddings de oraciones en múltiples idiomas de manera uniforme [31]. Fue desarrollado por Google Research y soporta **109 idiomas**, lo que lo convierte en una de las soluciones más robustas para tareas de alineación multilingüe. Su objetivo principal es generar representaciones de texto que sean comparables entre diferentes idiomas, facilitando tareas como la traducción automática, la búsqueda multilingüe y la clasificación de texto en distintos idiomas sin necesidad de reentrenamiento específico. Se basa en la arquitectura **BERT**, pero con un enfoque particular en el aprendizaje de representaciones textuales que sean independientes del idioma. Para lograr esto, el modelo fue preentrenado utilizando una estrategia de *aprendizaje contrastivo*, donde se emparejan frases equivalentes en distintos idiomas y se optimiza la similitud entre sus representaciones. Esto permite que el modelo capture significados similares en diferentes idiomas dentro de un mismo espacio de embeddings.

En el contexto de **VQA multilingüe**, la combinación de **LaBSE con CLIP** permite mejorar la alineación entre imágenes y texto en distintos idiomas. La versión utilizada en este trabajo incorpora **ViT-L/14** como modelo visual, lo que proporciona representaciones visuales detalladas y de alta calidad. Además, esta variante ha sido ajustada específicamente para mejorar su desempeño en datasets multilingües, logrando resultados competitivos en benchmarks como **MS-COCO** y en tareas de correspondencia entre texto e imagen.

Las principales ventajas de utilizar CLIP Multilingual con LaBSE se

centran en su capacidad de generalización en múltiples idiomas, lo que permite manejar preguntas formuladas en diferentes lenguas sin necesidad de entrenar un modelo específico para cada una. Además, al estar entrenado con un enfoque contrastivo, LaBSE logra una mejor alineación semántica, generando representaciones de texto más cercanas en significado, lo que a su vez mejora la correspondencia entre imágenes y descripciones en distintos idiomas. Por otro lado, el rendimiento en tareas multimodales también se ve optimizado gracias a la combinación con **ViT-L/14**, que permite capturar información tanto textual como visual con mayor precisión. En resumen, **CLIP Multilingual LaBSE** representa una solución potente para tareas de VQA en entornos multilingües, ya que facilita la alineación entre el texto en diversos idiomas y la información visual, asegurando una mejor comprensión y generación de respuestas en contextos donde se requiere soporte para múltiples lenguas.

XLM-RoBERTa es un modelo de lenguaje multilingüe basado en la arquitectura de RoBERTa, una variante mejorada de BERT que elimina el uso de etiquetas de segmentación de oraciones y emplea una estrategia de entrenamiento sin máscara estática[32]. Fue preentrenado con grandes volúmenes de texto en 100 idiomas utilizando el corpus masivo Common Crawl[33], lo que le permite comprender y generar texto en múltiples lenguas con un rendimiento superior en tareas de procesamiento de lenguaje natural (NLP). Una de las principales ventajas de XLM-RoBERTa es su capacidad para manejar múltiples idiomas sin necesidad de reentrenamiento específico para cada uno. Esto lo convierte en un modelo robusto y flexible para tareas como clasificación de texto, análisis de sentimientos, traducción automática y recuperación de información multilingüe.

En el contexto de Visual Question Answering (VQA) multilingüe, la combinación de CLIP con **XLM-RoBERTa** permite un enfoque multimodal más efectivo. CLIP se encarga de la representación visual y textual, alineando imágenes y texto en un espacio compartido, mientras que **XLM-RoBERTa** mejora la comprensión del lenguaje en múltiples idiomas, proporcionando representaciones más precisas para preguntas y respuestas en diversos contextos lingüísticos. Este enfoque resulta particularmente útil en datasets como VizWiz, donde las preguntas pueden presentarse en diferentes idiomas, requiriendo un procesamiento textual más avanzado.

Capítulo 3

Materiales y Métodos

En este capítulo se describen en detalle los materiales y metodologías empleadas en el desarrollo de este trabajo. Se presenta el conjunto de datos utilizado, incluyendo su estructura original, el proceso de traducción al español y las etapas de preprocesamiento aplicadas para su adaptación al entorno experimental. Se expone además el enfoque metodológico adoptado, el cual se estructura en tres etapas progresivas. En la **primera etapa**, se exploran modelos de **fusión multimodal**, que combinan representaciones de texto e imagen a través de arquitecturas paralelas. En la **segunda etapa**, se evalúan modelos diseñados específicamente para la tarea de VQA, los cuales han sido preentrenados en grandes corpus de datos visuales y textuales. En particular, se presenta la adaptación del modelo **CLIP** a **VQA** en español, que constituye uno de los aportes centrales de esta tesis. Finalmente, en la **tercera etapa**, se implementa un **ensamble de modelos**, con el objetivo de combinar las fortalezas de las distintas arquitecturas evaluadas y así mejorar la robustez y exactitud del sistema propuesto. Además, se detallan los **recursos computacionales utilizados** y las **métricas de evaluación** utilizadas para medir el rendimiento de los modelos en cada una de las etapas, asegurando una comparación justa y consistente a lo largo del estudio.

3.1. Preparación del Dataset: Estructura, Traducción y Preprocesamiento

Como se mencionó anteriormente, el conjunto de datos utilizado en este trabajo es VizWiz, obtenido del sitio web oficial del desafío Visual Question Answering (VQA)[1]. Este desafío ha publicado varias versiones del dataset, algunas de las cuales han sido retiradas; para este estudio se emplea la versión ampliada publicada el 1 de enero de 2020. El dataset se compone de dos elementos principales: por un lado, las imágenes en

[1]https://vizwiz.org/tasks-and-datasets/vqa/

formato .jpg, por otro, las anotaciones asociadas a cada imagen, que incluyen una pregunta formulada por una persona con discapacidad visual y diez respuestas proporcionadas por anotadores videntes.

Las anotaciones se organizan en dos particiones principales: entrenamiento y validación, ambos presentan la siguiente estructura:

- **answerable:** Indica si la pregunta tiene una respuesta posible (1: respondible, 0: no respondible). En este trabajo este campo no es utilizado.
- **image:** Identificador único para cada imagen. Este identificador comienza con la palabra "VizWiz", seguida del acrónimo "val" en el caso del conjunto de datos de validación, o "train" en el caso del de entrenamiento. Luego, se incluye un identificador numérico de 8 dígitos, y finalmente, la extensión del archivo. Ejemplo: "VizWiz_val_00028000.jpg".
- **question:** Pregunta en formato de texto asociada a la imagen.
- **answers:** Lista de diez respuestas proporcionadas por diferentes anotadores para cada pregunta.
- **answer_type:** Tipo de respuesta, determinado según la categoría predominante en la lista de respuestas. Sus posibles valores son 4: *other*, *unanswerable*, *yes/no* y *numeric*

La versión oficial del dataset incluye un mayor número de muestras; sin embargo, durante la preparación de los conjuntos de entrenamiento y validación se identificaron imágenes que no podían asociarse correctamente con sus anotaciones completas. Para asegurar la consistencia de los datos utilizados en el entrenamiento del modelo, estas imágenes se excluyeron del conjunto final. Como resultado, el dataset empleado en este trabajo quedó conformado por 12,039 ejemplos para entrenamiento y 4,301 ejemplos para validación. En la Figura 3.1 se muestra un ejemplo del formato original del dataset antes de su traducción al español. Asimismo, en la Figura 3.2 se presenta un análisis exploratorio de ambos subconjuntos, destacando la distribución de tipo de respuestas en cada uno.

Para facilitar el uso del conjunto de datos en los diferentes experimentos, se ajustó el formato del campo *answers*. En el dataset original, este campo consistía en una lista de 10 diccionarios, donde cada uno contenía la respuesta proporcionada junto con la confianza asociada a dicha respuesta, como se muestra en la Figura 3.1. Este formato se transformó en una lista que contiene exclusivamente las diez respuestas dadas por los revisores. Posteriormente, se realizó la traducción del conjunto de

```
"answerable": 0,
"image": "VizWiz_val_00028000.jpg",
"question": "What is this?"
"answer_type": "unanswerable",
"answers": [
    {"answer": "unanswerable", "answer_confidence": "yes"},
    {"answer": "chair", "answer_confidence": "yes"},
    {"answer": "unanswerable", "answer_confidence": "yes"},
    {"answer": "unanswerable", "answer_confidence": "no"},
    {"answer": "unanswerable", "answer_confidence": "yes"},
    {"answer": "text", "answer_confidence": "maybe"},
    {"answer": "unanswerable", "answer_confidence": "yes"},
    {"answer": "bottle", "answer_confidence": "yes"},
    {"answer": "unanswerable", "answer_confidence": "yes"},
    {"answer": "unanswerable", "answer_confidence": "yes"}
]
```


Figura 3.1: Ejemplo del archivo dataset VizWiz

datos del inglés al español.

Inicialmente, se empleó OPUS-MT, un modelo de traducción automática basado en MarianNMT [34], optimizado para rapidez y eficiencia en la traducción multilingüe. Sin embargo, al analizar ejemplos de salida, se observó que las traducciones no siempre eran precisas o claras, especialmente en preguntas con estructuras gramaticales complejas o términos poco frecuentes.

Debido a estas limitaciones, se optó por utilizar GPT-3.5, un modelo de lenguaje avanzado basado en Transformers [35], que ha demostrado un mejor manejo del contexto y la semántica en la traducción automática. GPT-3.5 permitió obtener traducciones más naturales y precisas,

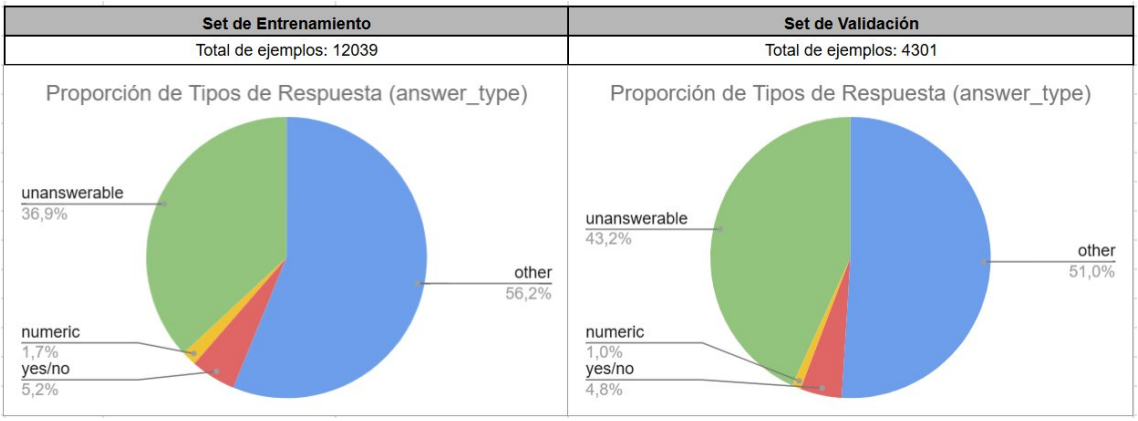


Figura 3.2: Distribución de los datasets según answer_type.

mejorando la fidelidad de las preguntas y respuestas en español.

Otros modelos populares de traducción automática, como Google Translate y DeepL, no fueron utilizados debido a restricciones en el acceso a grandes volúmenes de traducción automática y la imposibilidad de realizar ajustes personalizados sobre las salidas generadas.

Tras la traducción, se añadieron las siguientes columnas al conjunto de datos:

- *question_trad:* Pregunta traducida al español.
- *answers_trad:* Lista de respuestas traducidas al español.

Para garantizar la calidad de la traducción, se implementó un proceso de validación manual sobre una muestra representativa de preguntas y respuestas. Además, se aplicaron diversas técnicas de normalización de texto con el fin de estandarizar el formato de las respuestas y minimizar inconsistencias en la escritura. Estas técnicas incluyeron la conversión a minúsculas para evitar discrepancias en el formato, la eliminación de espacios en blanco innecesarios y de caracteres especiales que no aportaban valor, así como la corrección de errores tipográficos menores en las respuestas. En el Cuadro 3.1 se presenta un ejemplo del formato del dataset resultante del preprocesamiento y la traducción, el cual ha sido puesto a disposición de la comunidad investigadora a través de la plataforma Hugging Face [36, 37].

A pesar de la mejora en la calidad de la traducción lograda con GPT-3.5, se identificaron algunas limitaciones. En primer lugar, se observó ambigüedad en las respuestas cortas, especialmente en el caso de las respuestas "unanswerable". En algunos casos, la traducción automática las convirtió como "incontestable", mientras que en otros las tradujo como "sin respuesta", lo que generó inconsistencias que fueron corregidas manual-

image_id	question	answers	answer_type	answerable	question_trad	answers_trad
VizWiz_train_00000000.jpg	What's the name of this product?	[basil leaves, basil leaves, basil, basil, basil leaves, basil leaves, basil leaves, basil leaves, basil leaves, basil]	other	1	¿Cuál es el nombre de este producto?	[hojas de albahaca, hojas de albahaca, hojas de albahaca, hojas de albahaca, hojas de albahaca, hojas de albahaca, hojas de albahaca, hojas de albahaca, hojas de albahaca]
VizWiz_train_00007159.jpg	What is this?	[coke,pop,coca cola,coke,cocoa cola,can coke,coca cola,coke,coke can,coca cola can]	other	1	¿Qué es esto?	[coke, pop, coca cola, coke, cocoa cola, lata de coke, coca cola, coke, lata de coke, lata de coca cola]
VizWiz_train_00000002.jpg	Is this enchilada sauce or is this tomatoes? Thank you.	[these tomatoes not enchilada sauce, tomatoes, tomatoes, tomatoes, tomatoes, crushed tomatoes, crushed tomatoes, tomatoes, tomatoes, tomatoes]	other	1	¿Es esto salsa de enchilada o son tomates? Gracias.	[Estos tomates no son salsa de enchilada., tomates, tomates, tomates, tomates, tomates triturados, tomates triturados, tomates, tomates, tomates]

Cuadro 3.1: Ejemplos del dataset VizWiz Traducido

mente. Además, hubo problemas con la traducción de nombres propios y marcas, ya que en ciertos casos el modelo intentó traducir estos términos específicos, lo que resultó en traducciones inexactas. Por último, algunas preguntas con construcciones ambiguas en inglés fueron traducidas de manera que alteraron ligeramente su intención original.

En el Cuadro 3.2 se presentan ejemplos de estas limitaciones. Se observa, por ejemplo, que en el caso de la marca "Dr Pepper", el modelo mantuvo el nombre original en lugar de traducirlo al español. En otro ejemplo, la palabra "Coke" fue interpretada de manera inconsistente: en algunos casos se reconoció como una marca y se conservó sin cambios, mientras que en otros fue traducida. Finalmente, en el último caso, el modelo no identificó que "Equate" es una marca comercial y, en su lugar, la tradujo incorrectamente como "igualar".

Además, el tiempo de procesamiento de la traducción fue considerable, dado el tamaño del conjunto de datos (12,039 ejemplos en entrenamiento y 4,301 en validación).

3.2. Enfoque Metodológico

Debido a su complejidad, VQA se considera una tarea *AI-completa*, ya que requiere una integración profunda de habilidades cognitivas en visión y lenguaje para generar respuestas coherentes y precisas. VQA desafía los modelos de inteligencia artificial al exigir un razonamiento conjunto

image_id	question	answers	answer_type	answerable	question_trad	answers_trad
VizWiz_train_00000013.jpg	What kind of drink is this?	[soda,dr pepper,dr pepper,soda,dr pepper, dr pepper,energy drink,mdr pepper,dr pepper,dr pepper]	other	1	¿Qué tipo de bebida es esta?	[dr pepper,dr pepper,dr pepper,dr pepper, dr pepper,dr pepper,dr pepper,dr pepper,dr pepper,dr pepper]
VizWiz_train_00007159.jpg	What is this?	[coke,pop,coca cola,coke,cocoa cola,can coke,coca cola,coke,coke can,coca cola can]	other	1	¿Qué es esto?	[coke, pop, coca cola, coke, cocoa cola, lata de coke, coca cola, coke, lata de coke, lata de coca cola]
VizWiz_train_00009402.jpg	What brand of baby wipes are these?	[equate,equate, equate,equate,equate everyday clean,equate, equate,equate,equate,equate]	other	1	¿De qué marca son estas toallitas húmedas para bebé?	[equate, equate, equate, equate, equate limpieza diaria, equate, equate, equate, equate, equate]

Cuadro 3.2: Ejemplos de limitaciones en la traducción

entre información visual y textual, lo que lo convierte en un problema representativo de la inteligencia artificial general[1].

En particular, el Procesamiento de Lenguaje Natural (Natural Language Processing, NLP) desempeña un papel fundamental en VQA por dos razones principales: la comprensión de la pregunta y la generación de la respuesta. Estos desafíos son comunes en los sistemas de preguntas y respuestas textuales dentro del NLP. Sin embargo, la principal diferencia en VQA radica en que el proceso de búsqueda y razonamiento debe realizarse a partir del contenido visual en lugar de un corpus textual. Por ejemplo, para responder si hay personas en la imagen, el modelo debe ser capaz de detectar objetos; para determinar si está lloviendo, necesita clasificar la escena; y para identificar los equipos en una competencia, requiere conocimientos del mundo real. Además, el razonamiento basado en sentido común y, en el mejor de los casos, en conocimiento explícito, es clave para responder preguntas más complejas. Muchas de estas subtareas, como la detección y reconocimiento de objetos, la clasificación de escenas y el razonamiento visual, han sido ampliamente estudiadas en el campo de la *Visión por Computadora*, logrando avances significativos en los últimos años.

La tarea de **Visual Question Answering (VQA)** ha sido tradicionalmente abordada como un problema de clasificación multiclase. En esta tesis se adopta dicha aproximación, modelando la tarea de forma que cada respuesta se asocie con una etiqueta correspondiente a su posición dentro de un conjunto cerrado de respuestas posibles. De este modo, el modelo puede seleccionar la opción más probable en función de la distribución de probabilidades que aprende durante el entrenamiento.

En términos generales, los enfoques de VQA comparten una secuencia estructurada de procesamiento. Primero, se realiza la extracción de características textuales, mediante el procesamiento de las preguntas para obtener una representación numérica de su contenido semántico. Luego, se lleva a cabo la extracción de características visuales, utilizando modelos de visión por computadora que permiten capturar la información relevante contenida en las imágenes. Finalmente, ambas representaciones —textual y visual— se fusionan para generar una predicción, es decir, la respuesta que mejor se ajusta a la relación entre la imagen y la pregunta planteada. Basándonos en estos principios, implementamos tres estrategias metodológicas en este trabajo. En la **primera etapa**, exploramos los *modelos de fusión multimodal* [38], un enfoque ampliamente utilizado en la literatura de VQA que se basa en la integración de señales de múltiples modalidades para generar una representación conjunta. Para ello, se emplean codificadores específicos que extraen características tanto del texto como de la imagen, las cuales luego se combinan en un espacio compartido. En nuestro caso, utilizamos la concatenación de las características extraídas por los codificadores de texto e imagen, seguida de un clasificador lineal que predice la respuesta a partir de la representación fusionada.

En la **segunda etapa**, evaluamos modelos específicamente diseñados para la tarea de VQA, los cuales han sido preentrenados en grandes conjuntos de datos. Aunque estos modelos no fueron desarrollados explícitamente para entornos de accesibilidad, su desempeño en el conjunto de datos VizWiz traducido al español nos permite analizar su aplicabilidad en el contexto de asistencia a personas con discapacidad visual. La comparación entre los modelos preentrenados y los enfoques de fusión multimodal nos permite evaluar la efectividad de ambas estrategias y explorar posibles adaptaciones que optimicen la accesibilidad en el escenario propuesto. Finalmente, en la **tercera etapa**, implementamos un *ensamble de modelos* con el objetivo de mejorar la robustez y exactitud del sistema. La combinación de diferentes modelos ha demostrado ser una estrategia efectiva en tareas de clasificación, ya que permite aprovechar las fortalezas individuales de cada arquitectura. En este caso, probamos distintas técnicas de ensamble, incluyendo votación por mayoría y combinación de predicciones ponderadas, con el fin de analizar si la integración de modelos mejora el rendimiento global en la tarea de VQA. A continuación damos mas detalles de cada una de estas etapas.

3.2.1. Modelos de Fusión Multimodal

En esta primera etapa, implementamos la metodología de **fusión tardía** (late fusion), una estrategia ampliamente utilizada en tareas de VQA. Este enfoque combina las representaciones de texto e imagen en una fase posterior del proceso, después de que ambas modalidades han sido procesadas de manera independiente. En los modelos de fusión tardía, cada modalidad sigue su propio flujo de procesamiento hasta las etapas finales, donde las salidas de los codificadores de texto e imagen se integran para generar la respuesta. Este enfoque resulta útil cuando cada modalidad requiere un procesamiento especializado, ya que permite que sus características sean extraídas de manera autónoma. No obstante, una posible limitación es la pérdida de interacciones tempranas entre las modalidades, lo que podría afectar la calidad de la representación conjunta. El proceso general de **fusión tardía** aplicado a la tarea de VQA se compone de varias etapas, ilustradas en la Figura 3.5. En primer lugar, se realiza la **extracción de características**, donde tanto la imagen como la pregunta son procesadas mediante modelos transformers preentrenados, especializados en el manejo de texto e imagen, respectivamente. Para el procesamiento del lenguaje natural se emplean modelos como BERT u otras variantes adaptadas, mientras que para el contenido visual se utilizan arquitecturas como Vision Transformer (ViT) o BEiT, diseñadas para interpretar imágenes de manera eficiente.

A continuación, en la etapa de **generación de representaciones**, las características extraídas son transformadas en representaciones avanzadas mediante las capas finales de los modelos transformers. Estas representaciones capturan de manera detallada y contextualizada la información semántica de ambos dominios, texto e imagen. Posteriormente, tiene lugar la **fusión de características**, donde las representaciones generadas para el texto y la imagen se combinan utilizando técnicas de fusión tardía. Entre las estrategias más comunes se encuentran la concatenación simple, la atención cruzada (cross-modal attention) y el uso de redes neuronales específicas para la fusión multimodal. El objetivo de este paso es construir una representación integrada que sintetice de forma conjunta la información visual y textual.

Finalmente, en la etapa de clasificación, la representación multimodal fusionada se introduce en un clasificador. La tarea de VQA se aborda como un problema de clasificación multiclase, en el que el modelo predice la respuesta más probable dentro de un vocabulario de respuestas. Este vocabulario está formado por todas las respuestas únicas presentes en el dataset de entrenamiento, asegurando que las predicciones correspondan

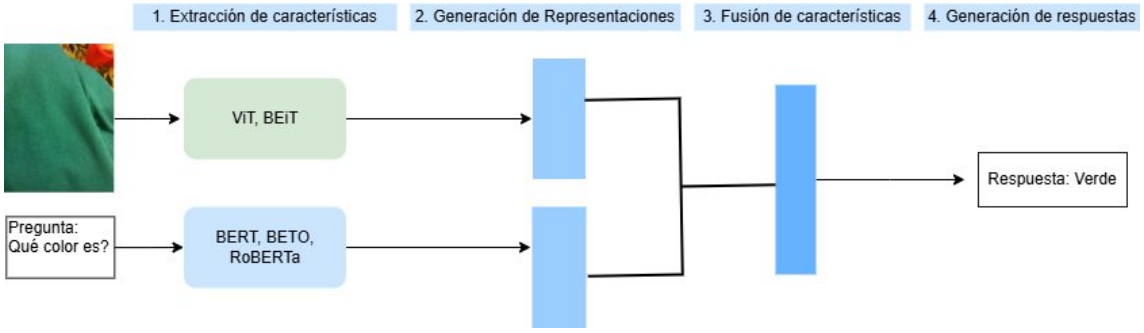


Figura 3.3: Arquitectura Late Fusion

a ejemplos reales observados durante el entrenamiento.

Para cada par *imagen-pregunta*, se define una respuesta representativa o "respuesta real", obtenida a partir de las múltiples respuestas humanas disponibles en el dataset, tal como se describe en la sección 3.3. El modelo aprende a asociar la entrada multimodal con la clase correspondiente en el vocabulario, seleccionando la respuesta que maximiza la probabilidad. Este enfoque permite abordar de manera efectiva las respuestas de tipo yes/no, numéricas y other, aunque en escenarios que requieran mayor flexibilidad o descripciones más detalladas, podría emplearse un modelo generativo basado en lenguaje natural.

Este enfoque escalonado permite que cada modalidad aporte información rica y detallada, procesada de manera independiente y posteriormente integrada en una representación única, optimizando la exactitud del modelo. Esta metodología es especialmente ventajosa en tareas de VQA, ya que permite aprovechar modelos *transformers* preentrenados tanto para texto como para imágenes. Esto simplifica el proceso de entrenamiento y mejora el rendimiento en comparación con la construcción desde cero de modelos multimodales más complejos.

En este trabajo, modelamos VQA como una tarea de clasificación multiclase, asignando etiquetas a cada respuesta en función de su índice dentro del espacio de respuestas posibles. Este estudio explora la idea de *fusión tardía* mediante el **fine-tuning** de modelos preentrenados de transformers para texto e imágenes. En este enfoque, las características de las modalidades de texto e imagen se extraen y procesan por separado antes de ser combinadas en una etapa posterior. La elección de este enfoque radica en la simplicidad de su entrenamiento. Al usar modelos preentrenados, ya se disponen de representaciones ricas y profundas tanto para el texto como para las imágenes, lo que facilita su adaptación a tareas específicas como la respuesta a preguntas visuales (VQA).

El **fine-tuning** consiste en ajustar estos modelos preentrenados sobre un

conjunto de datos específico, en nuestro caso el dataset VizWiz, permitiendo que los modelos de texto e imagen se adapten a las particularidades de la tarea de VQA sin la necesidad de entrenarlos desde cero. Este enfoque resulta más eficiente, ya que los modelos preentrenados ya han aprendido patrones generales a partir de grandes volúmenes de datos, y solo se requiere ajustar sus parámetros para mejorar su desempeño en la tarea de respuesta a preguntas visuales. Al utilizar transformers preentrenados, el proceso de entrenamiento se simplifica, ya que los modelos ya contienen una comprensión sólida de las representaciones de texto e imagen. Así, la fusión tardía se convierte en una estrategia accesible y efectiva, que permite obtener buenos resultados sin necesidad de entrenar modelos multimodales complejos desde el principio.

Dado que en este estudio se evaluaron tanto la versión original en inglés del dataset como su traducción al español, también se analizó el impacto del idioma en el rendimiento del modelo. Para ello, se utilizaron distintas combinaciones de modelos para la codificación del texto y de las imágenes. En el caso de los modelos textuales, se emplearon **BERT**, diseñado originalmente para el idioma inglés, y también **BETO** y **RoBERTa**, versiones adaptadas específicamente para el español. Respecto a los modelos visuales, se utilizaron Vision Transformer **(ViT)** y **BEiT**, ambos basados en arquitecturas de tipo transformer orientadas a la extracción de características visuales.

3.2.2. Modelos Preentrenados para VQA

En esta sección, empleamos modelos que combinan información visual y lingüística en una arquitectura basada en transformers. En particular, utilizamos una variante modificada del modelo **CLIP**, adaptada específicamente para la tarea de Visual Question Answering (VQA) [6]. Este modelo destaca por su capacidad para alinear representaciones visuales y textuales en un espacio común, lo que lo convierte en una solución eficaz para tareas que requieren razonamiento multimodal. Además, dicho estudio utiliza el mismo dataset VizWiz que empleamos en este trabajo, lo que subraya la relevancia de este enfoque.

En este trabajo se evaluó el rendimiento del modelo CLIP original utilizando texto en inglés como referencia inicial (baseline) y, posteriormente, con texto traducido al español. Para adaptar la tarea al español y mejorar el soporte multilingüe, se utilizó la misma arquitectura del modelo CLIP para VQA descrita en el artículo original [6], reemplazando únicamente el codificador de texto por modelos multilingües capaces de procesar entradas en español. En contraste, el codificador de imágenes permane-

ció inalterado, manteniendo la implementación original de CLIP. Esto permitió conservar las representaciones visuales en un espacio vectorial denso alineado con el codificador de texto.

Con el objetivo de establecer una nomenclatura más clara, estas variantes adaptadas las denominamos **CLIP Multilingual** seguido del nombre del codificador de texto multilingüe utilizado. Si bien desde un punto de vista técnico ya no se trata de modelos CLIP puros, conservan su esencia al mantener intacta la arquitectura visual original de CLIP y reemplazar únicamente el componente textual por uno capaz de procesar múltiples idiomas. De este modo, surgen modelos como **CLIP-ViT-B-32-multilingual-v1**, **CLIP Multilingual LaBSE** o **CLIP Multilingual XLM-RoBERTa**, que combinan la robustez del codificador de imágenes de CLIP con representaciones lingüísticas multilingües más adecuadas para tareas como VQA en contextos no angloparlantes.

Todos los modelos multilingües empleados en esta sección están disponibles en la librería sentence-transformers de Hugging Face [39]. Estos modelos permiten mapear texto en múltiples idiomas al espacio vectorial común de CLIP, facilitando la alineación entre texto e imágenes. A continuación, se describen en detalle las nuevas arquitecturas evaluadas que incorporan estos modelos multilingües.

Aunque los codificadores multilingües están diseñados para alinearse con el espacio vectorial del codificador de imágenes de CLIP, es necesario realizar ajustes específicos, como el fine-tuning, para adaptar el modelo a tareas especializadas como la VQA. Este proceso es crucial cuando se trabaja con datos en español, ya que permite optimizar la interacción entre las representaciones de texto e imagen en el contexto del idioma y de la tarea objetivo.

Para llevar a cabo el fine-tuning de los modelos utilizando el dataset VizWiz traducido al español, se siguió un procedimiento que garantizó la correcta adaptación de las arquitecturas y un monitoreo constante del desempeño. En primer lugar, se preparó el dataset asegurando que las preguntas y respuestas estuvieran en español y presentaran un formato consistente.

En cuanto a la arquitectura, se reemplazó el codificador de texto original por un modelo multilingüe (como LaBSE o XML-RoBERTa) y se verificó que las dimensiones de salida del nuevo codificador fueran compatibles con las características generadas por el codificador de imágenes de CLIP. Durante el entrenamiento, el codificador de imágenes se mantuvo congelado, de modo que solo se actualizaron los pesos del codificador de

texto y de la capa lineal que integra ambas modalidades. Esto permitió preservar las representaciones visuales preentrenadas, mientras el modelo se adaptaba al español y a la tarea de VQA.

La función de pérdida utilizada combina Cross-Entropy Loss para la predicción de la respuesta y del tipo de respuesta, junto con una pérdida adicional para la evaluación de la answerability de la pregunta. Esta combinación asegura que el modelo aprenda no solo a predecir la respuesta correcta, sino también a identificar preguntas incontestables. El entrenamiento se realizó con el optimizador AdamW, seleccionando cuidadosamente la tasa de aprendizaje para garantizar una convergencia estable. Además, se implementó un esquema de validación continua mediante el cual, al finalizar cada época, se calculan métricas de desempeño sobre un conjunto de validación independiente, incluyendo accuracy, VizWiz accuracy y score de answerability. Con base en estas métricas, se guardó la versión del modelo con mejor performance en validación, permitiendo un seguimiento constante y evitando sobreajuste.

3.2.3. Ensamble de Modelos para VQA

En este trabajo se exploran métodos de ensamble para mejorar el rendimiento del sistema de VQA en español. Estas técnicas combinan las predicciones de múltiples modelos con el objetivo de obtener una respuesta más robusta y precisa. Una de las estrategias implementadas fue la votación mayoritaria (**hard voting**). En este enfoque, cada modelo individual genera una predicción y la respuesta final se determina en función de la opción más frecuente entre todas las propuestas. Esta técnica resulta especialmente útil cuando los modelos presentan errores complementarios, ya que permite filtrar respuestas incorrectas y favorecer aquellas más consistentes. Estudios previos han demostrado que la votación mayoritaria puede mejorar el desempeño en tareas de clasificación multiclase, en particular cuando existe variabilidad en las predicciones de los modelos base.

Otra estrategia explorada fue la **fusión de características**. En lugar de combinar directamente las predicciones, este método integra los embeddings generados por cada modelo. Los vectores de representación de imágenes y texto extraídos por distintos codificadores son concatenados o transformados mediante técnicas basadas en redes neuronales para obtener una predicción final. La fusión de características permite capturar información más rica y variada, aunque su efectividad depende de la complementariedad entre los embeddings y del modelo empleado para realizar la fusión.

Finalmente, se consideró el uso de **meta-clasificación o stacking**. Esta técnica consiste en entrenar un modelo adicional, denominado meta-clasificador, que aprende a combinar las predicciones de los modelos base para generar una mejor respuesta. En este trabajo se evaluaron diferentes opciones de meta-clasificadores, como Regresión Logística, Random Forest y XGBoost. El éxito de esta estrategia radica en la capacidad del meta-clasificador para identificar patrones en los errores de los modelos base y corregirlos, un enfoque que ha demostrado ser eficaz en diversas aplicaciones de aprendizaje automático.

3.3. Métricas de evaluación

La métrica de exactitud (accuracy) es una de las más utilizadas en problemas de clasificación en Deep Learning. Se define como la proporción de predicciones correctas sobre el total de ejemplos evaluados.

En términos formales, dada una cantidad de N ejemplos en un conjunto de datos de prueba, si el modelo realiza C predicciones correctas, la exactitud se calcula como:

$$Accuracy = C/N$$

En tareas de clasificación dentro del aprendizaje automático, el accuracy se utiliza habitualmente como una métrica inicial para evaluar el rendimiento de los modelos, pues refleja la proporción de predicciones correctas respecto del total de instancias evaluadas. Su aplicación es particularmente útil en escenarios de clasificación multiclase y, en este trabajo, se emplea como parte del análisis cuantitativo del desempeño obtenido. No obstante, dado que el problema abordado se enmarca en el ámbito de la Respuesta a Preguntas Visuales (VQA), donde es posible que una misma pregunta tenga múltiples respuestas válidas, se decidió complementar esta evaluación con la métrica VizWiz Accuracy, la cual se describe en la sección siguiente.

VizWiz Accuracy

En tareas de Respuesta a Preguntas Visuales (VQA), la exactitud tradicional utilizada en clasificación multiclase no siempre refleja adecuadamente el desempeño del modelo, dado que pueden existir múltiples respuestas válidas para una misma pregunta. En este contexto, la **métrica VizWiz Accuracy**, propuesta en el desafío VizWiz Challenge [40], ofrece una alternativa más apropiada. Esta métrica fue diseñada específicamente para el dataset VizWiz-VQA, en el cual cada pregunta cuenta con

Pregunta	Respuestas humanas (10)	Respuesta representativa	Predicción del modelo	Accuracy	VizWiz Accuracy
¿De qué color es la camisa?	azul, azul, celeste, azul claro, azul, blanco, celeste, azul, azul, celeste	azul	azul	1 (correcto)	1 (5 coincidencias)
¿Qué objeto aparece sobre la mesa?	plato, plato, plato, bandeja, plato, bandeja, plato, plato, bandeja, plato	plato	bandeja	0 (incorrecto)	1 (3 coincidencias)
¿Qué bebida se ve en la imagen?	coca-cola, coca-cola, coca, coca, coca-cola, gaseosa, coca, agua, coca, coca-cola	coca-cola	coca-cola	1 (correcto)	1 (4 coincidencias)
¿Qué animal aparece en la foto?	perro, perro, perro, perro, gato, perro, perro, perro, perro, perro	perro	gato	0 (incorrecto)	0.33 (1 coincidencia)

Cuadro 3.3: Ejemplos comparativos entre accuracy tradicional y VizWiz Accuracy en VQA.

diez respuestas anotadas por diferentes personas. Su definición es la siguiente:

$$\text{VizWiz Accuracy} = \text{mín}\left(\frac{\#\text{coincidencias con respuestas humanas}}{3},\ 1\right)$$

De este modo, una predicción se considera completamente correcta si al menos tres anotadores dieron la misma respuesta, y se penaliza proporcionalmente cuando el consenso es menor. VizWiz Accuracy contempla la subjetividad de las respuestas humanas, reduce el impacto del ruido en las anotaciones y refleja de forma más robusta la utilidad de una respuesta en escenarios con imágenes complejas, como aquellas generadas por personas con discapacidad visual.

Implementación y comparación de métricas en el dataset

En este trabajo, se define "respuesta real" como la respuesta representativa de cada pregunta, construida a partir de las diez respuestas humanas disponibles en el dataset VizWiz. Para determinar la respuesta real se aplica el siguiente procedimiento: se selecciona la respuesta más frecuente; en caso de empate, se elige la que es más común en todo el dataset; si persiste el empate, se selecciona la respuesta más representativa según la distancia de Levenshtein, de manera que refleje mejor la similitud con las demás respuestas humanas. La respuesta real se utiliza como referencia para calcular el accuracy tradicional, evaluando si la predicción del modelo coincide exactamente con ella. De este modo, el accuracy tradicional mide si la predicción del modelo coincide exactamente con

esta respuesta única, mientras que VizWiz Accuracy permite reconocer aciertos cuando existe suficiente coincidencia con los anotadores humanos, capturando la subjetividad y variabilidad de las respuestas. El cuadro 3.3 ilustra ejemplos concretos de cómo se evalúan ambas métricas. Se observa que una predicción puede ser incorrecta según accuracy tradicional pero aceptada por VizWiz Accuracy si al menos tres anotadores humanos coinciden con ella. Por ejemplo, ante la pregunta "¿Qué objeto aparece sobre la mesa?", si la respuesta representativa es "plato" y el modelo predice "bandeja", accuracy penaliza la predicción, mientras que VizWiz Accuracy la considera correcta porque tres anotadores también eligieron "bandeja". Asimismo, VizWiz Accuracy otorga crédito parcial en casos con menor consenso, como predecir "gato" cuando la mayoría indicó "perro", reflejando la coincidencia con un solo anotador. En otros ejemplos, como predecir "azul" o "coca-cola" coincidiendo con la mayoría de los anotadores, ambas métricas consideran la predicción correcta. Esto demuestra que VizWiz Accuracy ofrece una medida más representativa del desempeño en contextos subjetivos, mientras que el accuracy tradicional proporciona una evaluación estricta basada en la respuesta representativa.

3.4. Recursos computacionales utilizados

Los experimentos realizados en esta tesis se llevaron a cabo en un entorno local con el sistema operativo Ubuntu 20.04.6 LTS, ejecutándose sobre el kernel Linux 5.4.0-153-generic y una arquitectura x86_64. El equipo utilizado contaba con un procesador AMD Phenom™ II X6 1075T, que dispone de seis núcleos sin hyperthreading y una frecuencia base de 3.0 GHz. La memoria RAM instalada era de 32 GB DDR3, complementada por una unidad de almacenamiento SSD NVMe de 2 TB (modelo Kingston SNV2S2000G), que permitió una gestión rápida de los datos.

Para el procesamiento gráfico, se emplearon dos tarjetas NVIDIA GeForce RTX 3090, cada una equipada con 24 GB de memoria dedicada, alcanzando un total de 48 GB de VRAM disponible. El sistema operaba con la versión 535.104.05 del controlador NVIDIA y utilizaba CUDA en su versión 12.2.

La combinación de dos GPUs de alto rendimiento permitió entrenar y evaluar modelos de aprendizaje profundo de manera eficiente, en particular aquellos que demandan un elevado volumen de procesamiento paralelo, como los modelos basados en CLIP. Asimismo, la disponibilidad de una amplia cantidad de memoria RAM y el uso de almacenamiento

NVMe contribuyeron a agilizar tanto el preprocesamiento de los datos como la carga de grandes volúmenes de imágenes.

Capítulo 4

Resultados

En esta sección se presentan los resultados obtenidos en los experimentos realizados para la tarea de Respuesta a Preguntas Visuales (VQA), utilizando el dataset VizWiz, previamente traducido al español. Como se describió en capítulos anteriores, el conjunto de entrenamiento consta de 12,039 ejemplos, mientras que el conjunto de validación incluye 4,301 ejemplos. Los modelos fueron evaluados utilizando las métricas Accuracy y VizWiz Accuracy, definidas en una sección previa.

Se distinguen cuatro categorías principales de experimentos:

1. **Baseline**: Experimentos realizados con el dataset en su idioma original (inglés), con el objetivo de establecer un punto de referencia.
2. **Modelos de Fusión Tardía:** Experimentos que procesan las características textuales y visuales por separado, y las combinan en una etapa posterior.
3. **Soluciones basadas en CLIP para VQA:** Experimentos que aprovechan modelos con integración multimodal inherente, como CLIP, para abordar la tarea de VQA.
4. **Ensambles**: Experimentos basados en la combinación de modelos pertenecientes a las dos categorías anteriores, con el fin de evaluar mejoras mediante estrategias de agregación.

Cabe destacar que en los experimentos de los puntos 2, 3 y 4 se utilizaron las preguntas y respuestas traducidas al español, a fin de evaluar el rendimiento de los modelos en un entorno completamente en dicho idioma.

Los resultados se presentan de la siguiente manera: en primer lugar, se muestran las métricas de **Accuracy** y **VizWiz Accuracy** a nivel general, considerando el desempeño global de cada modelo. Posteriormente, con el objetivo de profundizar en los casos en los que los modelos obtuvieron un mejor o peor rendimiento, se realiza un análisis desagregado según el tipo de respuesta (*answer_type*), categoría que en el dataset VizWiz puede tomar uno de los siguientes cuatro valores: *other, unanswerable, yes/no y*

Modelo	other	unanswerable	yes/no	number	Accuracy Total	VizWiz Acc
BERT + VIT	20 %	97 %	58 %	26 %	45 %	57 %
BERT + BEiT	25 %	94 %	65 %	26 %	47 %	55 %
CLIP VQA	39 %	95 %	53 %	36 %	52 %	64 %

Cuadro 4.1: Resultados experimentos Baseline según el tipo de respuesta y general

number. Para cada una de estas categorías, se reporta la métrica **VizWiz Accuracy**, permitiendo así identificar las fortalezas y debilidades de cada enfoque en función del tipo de respuesta esperada. En las secciones siguientes se detallan los resultados obtenidos para cada una de las categorías mencionadas. Adicionalmente, se presenta un análisis específico de las respuestas de tipo numérico, dado que estas mostraron diferencias significativas en su desempeño según el tipo de enfoque utilizado.

4.1. Resultados Baseline

Los experimentos baseline se llevaron a cabo utilizando el dataset VizWiz en inglés (sin traducir), evaluando tanto fusion tardía como modelos preentrenados en este idioma. Estos resultados sirven como referencia para analizar el impacto de la traducción al español en los experimentos posteriores. Presentamos los resultados en el Cuadro 4.1 donde muestran un desempeño general comparable entre los distintos modelos evaluados. Con el objetivo de profundizar en la comprensión de los casos en los que los modelos obtuvieron un mejor o peor rendimiento, se llevó a cabo un análisis desagregado por tipo de respuesta (*answer_type*). Este análisis también se encuentra detallado en el Cuadro 4.1, donde se reportan los valores de la métrica VizWiz Accuracy para cada una de las categorías de respuesta.

En la Figura 4.1 se presentan ejemplos de predicciones correspondientes a preguntas cuya respuesta es de tipo *yes/no*, realizadas por los modelos evaluados en el *baseline*. A través de estas ilustraciones, observamos las diferencias en el desempeño de los modelos ante preguntas simples pero sensibles al contenido visual. Por ejemplo, en las figuras (a) y (d), tanto CLIP como BERT+ViT clasifican erróneamente las imágenes como *unanswerable*, lo que sugiere una dificultad para asociar preguntas binarias con evidencias visuales claras. En cambio, BERT+BEiT logra predecir correctamente, lo que puede atribuirse a una representación visual más detallada proporcionada por BEiT, basada en técnicas de *masked image modeling*. Por otro lado, en la figura (b), solo BERT+ViT falla, posiblemente por una falta de alineación efectiva entre las características visuales y textuales. Finalmente, en la figura (c), el modelo CLIP

(a) Pregunta: Can you see what's on the screen?, Respuesta real: no. Predicciones: bert_beit = no, bert_vit = unansw, CLIP_VQA = unansw.

(b) Pregunta: Can you tell what's on this paper?, Respuesta real: no. Predicciones: bert_beit = no, bert_vit = unansw, CLIP_VQA = no.

(c) Pregunta: Is this cheddar cheese?, Respuesta real: no. Predicciones: bert_beit = no, bert_vit = no, CLIP_VQA = yes.

(d) Pregunta: Has she finished her drink?, Respuesta real: yes. Predicciones: bert_beit = yes, bert_vit = no, CLIP_VQA = no.

Figura 4.1: Ejemplos de resultados de experimentos *baseline* cuya respuesta es del tipo *yes/no*.

es el único en errar, lo que podría indicar limitaciones en la generalización del modelo cuando las preguntas requieren razonamiento específico de contexto. Estas observaciones respaldan la necesidad de modelos que integren mejor las representaciones visuales y lingüísticas para preguntas cerradas.

4.2. Modelos de Fusión Tardía en Español

En esta etapa, se evaluaron las diversas combinaciones de modelos de lenguaje y visión para la tarea de VQA en español presentadas previamente. En particular, se evaluaron configuraciones que combinan modelos de texto, como **BERT**, con modelos de imagen, como **ViT** y **BEiT**, además de variantes en español, como **BETO** y **RoBERTa**.

Los resultados obtenidos en las métricas de Accuracy y VizWiz accuracy para cada combinación de modelo se presentan a continuación en el Cuadro 4.2.

Al igual que en los experimentos base, las métricas obtenidas son si-

Modelo	Idioma	Accuracy	VizWiz Acc
BERT + VIT	Español	44 %	51 %
BERT + BEiT	Español	46 %	53 %
BETO + VIT	Español	44 %	51 %
BETO + BEiT	Español	46 %	53 %
ROBERTA + VIT	Español	44 %	51 %
ROBERTA + BEiT	Español	46 %	53 %

Cuadro 4.2: Resultados experimentos Etapa 1

Modelo	other	unanswerable	yes/no	number
BERT + VIT	15 %	97 %	35 %	17 %
BERT + BEiT	19 %	95 %	37 %	17 %
BETO + VIT	14 %	95 %	52 %	17 %
BETO + BEiT	20 %	94 %	57 %	16 %
ROBERTA + VIT	15 %	95 %	51 %	17 %
ROBERTA + BEiT	19 %	95 %	51 %	17 %

Cuadro 4.3: Resultados experimentos Etapa 1 por tipo de respuesta

milares entre sí. Al analizar los resultados por tipo de respuesta en el Cuadro 4.3, en la categoría de respuestas unanswerable, los modelos evaluados muestran un rendimiento alto, con valores que superan el 94 % en la mayoría de los casos. Sin embargo, las respuestas de tipo number (numéricas) resultaron ser las más desafiantes, mostrando la mayor disminución en exactitud en comparación con el baseline. Todos los modelos presentan un desempeño deficiente en esta categoría, con valores en torno al 16-17 %. Esto sugiere que las respuestas numéricas representan un desafío significativo para los modelos evaluados, posiblemente debido a limitaciones en el manejo de información estructurada o numérica dentro del texto.

Respecto al desempeño en respuestas *yes/no*: Se observa una variabilidad considerable entre los modelos en esta categoría. BETO + BEiT logra la mayor rendimiento con 57 %, seguido de BETO + ViT con 52 %. Por otro lado, los modelos basados en BERT presentan un rendimiento más bajo en esta categoría (35-37 %). Esto podría indicar que los modelos basados en BETO están mejor adaptados a este tipo de preguntas en español. La exactitud en la categoría *Other* varía entre 14 % y 20 %. BETO + BEiT es el modelo con mejor desempeño (20 %), mientras que BETO + ViT obtiene el menor (14 %). Esto podría indicar que el modelo BETO + BEiT tiene una mejor capacidad de generalización para preguntas fuera de las categorías específicas evaluadas.

Al comparar modelos con la misma base de texto, no se observa una diferencia sistemática clara entre aquellos que utilizan ViT y los que emplean BEiT. Sin embargo, en algunos casos, BEiT parece ofrecer una

Modelo	Idioma	Accuracy	VizWiz Acc
CLIP VQA Original	Español	50 %	59 %
CLIP VQA Multilingual	Español	48 %	57 %
CLIP VQA Multi LABSE	Español	51 %	59 %
CLIP VQA Multi XML-ROBERTA	Español	50 %	59 %

Cuadro 4.4: Resultados experimentos Etapa 2 basados en CLIP

ligera mejora en la predicción de las categorías *other* y *yes/no*.

En general, los modelos presentan un rendimiento sólido en respuestas *unanswerable*, pero aún existen oportunidades de mejora en respuestas *numéricas* y *yes/no*. Cabe destacar que los modelos basados en BETO sobresalen en preguntas binarias (yes/no), lo que sugiere una mejor adaptación al español en este tipo de respuestas.

En cuanto a los modelos con base en RoBERTa, estos muestran una mejora con respecto a BERT en algunas categorías, particularmente en *yes/no*, aunque sin alcanzar el desempeño de BETO. Además, su rendimiento en la categoría *other* sugiere que RoBERTa + BEiT podría tener una leve ventaja en términos de generalización.

4.3. Soluciones CLIP para VQA en Español

En esta etapa, se evaluaron modelos basados en CLIP adaptado para VQA, incluyendo la versión original y las diversas variantes multilingües presentadas anteriormente. Estas variantes incorporan distintos modelos de lenguaje con soporte para español, con el objetivo de mejorar la capacidad del modelo para procesar preguntas y respuestas en este idioma.

En el Cuadro 4.5 se presenta los resultados obtenidos en cada experimento, utilizando nuevamente las métricas Accuracy y VizWiz accuracy como criterios de evaluación. Se observa una leve mejora en la performance en comparación con los modelos de fusión tardía. Al analizar el desempeño por tipo de respuesta en el Cuadro 4.6, se observa que todos los modelos mantienen un rendimiento sólido frente a las respuestas *unanswerable* (incontestables). Las respuestas de tipo *number* (numéricas) continúan siendo las más desafiantes; sin embargo, muestran una mejora ligera en comparación con los modelos de fusión tardía. Este incremento en las métricas para dicho tipo de respuesta contribuye positivamente al desempeño general de los modelos CLIP.

Modelo	other	unanswerable	yes/no	number
CLIP VQA Original	32 %	92 %	44 %	27 %
CLIP VQA Multilingual	28 %	93 %	41 %	16 %
CLIP VQA Multi LABSE	32 %	94 %	51 %	30 %
CLIP VQA Multi XML-ROBERTA	32 %	93 %	51 %	29 %

Cuadro 4.5: Resultados experimentos Etapa 2 por tipo de respuesta

En relación al tipo de respuesta *other*, se evidencia una mejora significativa, ya que los modelos de CLIP VQA alcanzan hasta un 32 %, en comparación con el rango de 14 %-20 % observado en la Etapa 1. Esto sugiere que CLIP VQA podría tener una mejor capacidad de generalización en este tipo de respuestas.

En general, los modelos de la Etapa 2 presentan mejoras en las categorías *other* y *number*, mientras que en *yes/no* y *unanswerable*, su desempeño es similar o ligeramente inferior al de los modelos de la Etapa 1. Esto sugiere que CLIP VQA podría ser más efectivo en ciertos tipos de preguntas, pero aún hay oportunidades de mejora en respuestas binarias y preguntas sin respuesta.

4.4. Ensamblaje de Modelos

En esta sección se presentan los resultados obtenidos mediante la aplicación de diversas estrategias de ensamblaje a los modelos previamente evaluados. El propósito de estas estrategias es aprovechar las fortalezas individuales de cada modelo con el fin de mejorar el rendimiento global en la tarea de VQA en español. Se han explorado tres enfoques principales de ensamblaje: ensamblaje basado en votación, ensamblaje por fusión de características o predicciones, y ensamblaje basado en aprendizaje mediante metaclasificadores.

Estos métodos tienen como objetivo optimizar el desempeño y robustez del sistema, al combinar de manera eficiente las capacidades complementarias de los distintos modelos.

En el Cuadro 4.6 se presentan los resultados obtenidos para los diferentes enfoques de ensamble. Los ensambles basados en votación mayoritaria evidenciaron una mejora en comparación con algunos modelos individuales. Particularmente, la combinación de CLIP LaBSE y CLIP XMLRoBERTa alcanzó un 51 % de Accuracy y un 60 % de VizWiz Accuracy. Esta mejora puede atribuirse a la complementariedad de los modelos involucrados. CLIP se destaca por su robustez en la alineación de imágenes y texto, y al integrar diferentes variantes —como LaBSE y XML-RoBERTa— se

Ensamble	**Modelos**	**Accuracy**	**VizWiz Acc**
Votación	BETO BeiT + RoBERTa Beit	47 %	54 %
Votación	CLIP Labse + RoBERTa Beit	49 %	57 %
Votación	CLIP Labse + BETO Beit	49 %	57 %
Votación	CLIP Labse + CLIP XML RoBERTa	**51** %	**60** %
Fusión Caract.	CLIP Labse + RoBERTa Beit	47 %	55 %
Fusión Caract.	CLIP Labse + CLIP XML RoBERTa	50 %	59 %
Logistic Reg.	CLIP Labse + RoBERTa Beit	45 %	53 %
Logistic Reg.	CLIP Labse + CLIP XML RoBERTa	50 %	59 %
Random Forest	CLIP Labse + RoBERTa Beit	45 %	53 %
Random Forest	CLIP Labse + CLIP XML RoBERTa	50 %	59 %
XGBoost	CLIP Labse + RoBERTa Beit	45 %	53 %
XGBoost	CLIP Labse + CLIP XML RoBERTa	50 %	59 %

Cuadro 4.6: Resultados de los ensambles en términos de Accuracy y VizWiz Accuracy.

logran mitigar errores individuales y reforzar aquellas predicciones más consistentes. La estrategia de votación mayoritaria, además, actúa como un mecanismo de filtrado, favoreciendo respuestas en las que los modelos coinciden y reduciendo el impacto de predicciones erróneas.

En cuanto a los ensambles basados en fusión de características, los resultados no mostraron una mejora significativa respecto a la votación. La combinación de CLIP LaBSE y CLIP XML-RoBERTa, bajo esta modalidad, obtuvo un 50 % de Accuracy y un 59 % de VizWiz Accuracy, cifras ligeramente inferiores a las del mejor ensamble por votación. Si bien la fusión de características tiene el potencial de capturar información más rica proveniente de diferentes modelos, su efectividad depende en gran medida de la complementariedad de las representaciones generadas. En este caso, es posible que las representaciones no hayan sido lo suficientemente diversas, limitando así el impacto en la predicción final. Asimismo, el método de fusión empleado podría no haber sido el más adecuado; estrategias más avanzadas, como el uso de mecanismos de atención o arquitecturas neuronales más profundas, podrían mejorar estos resultados.

Finalmente, los ensambles basados en aprendizaje mediante meta clasificadores tampoco superaron el desempeño de la votación mayoritaria. Se exploraron distintos modelos de meta-clasificación, incluyendo *Logistic Regression*, *Random Forest* y *XGBoost*, entrenados para seleccionar la mejor predicción a partir de características extraídas de las respuestas de los modelos base. Sin embargo, los resultados obtenidos se mantuvieron

en un rango similar o incluso inferior, alcanzando entre un 45 % y 50 % de Accuracy y entre un 53 % y 59 % de VizWiz Accuracy. Estos resultados sugieren que, en el contexto de esta tarea y configuración experimental, la simplicidad y robustez de la votación mayoritaria resultó ser más efectiva que las estrategias de mayor complejidad.

El desempeño limitado del meta-clasificador puede atribuirse a varios factores. En primer lugar, la calidad de las características extraídas juega un papel fundamental. Si las características utilizadas, como la similitud coseno entre las respuestas, la longitud de las respuestas y la confianza del modelo, no contienen suficiente información discriminativa, el clasificador no será capaz de superar a estrategias más simples, como la votación mayoritaria. Además, la efectividad de modelos como *Random Forest* y *XGBoost* está estrechamente relacionada con el tamaño del conjunto de datos. Estos modelos requieren una cantidad significativa de datos para aprender patrones efectivos de combinación. En este contexto, es posible que el conjunto de entrenamiento utilizado no haya sido lo suficientemente grande como para explotar plenamente el potencial de estos modelos.

En conclusión, el mejor desempeño se obtuvo con la **votación de modelos CLIP (CLIP Labse + CLIP XML RoBERTa)**, lo que sugiere que estos modelos poseen una representación más efectiva de preguntas y respuestas en español. La fusión de características mostró un rendimiento competitivo, aunque sin superar claramente a la votación. Por otro lado, los meta-clasificadores no lograron mejoras significativas, probablemente debido a la limitada discriminación en las características utilizadas. Para mejorar los resultados en futuras investigaciones, se podrían explorar técnicas más avanzadas de fusión y optimización del meta-clasificador, como el uso de redes neuronales para aprender la combinación óptima de predicciones.

En la Figura 4.2 se presentan ejemplos de preguntas con sus respectivas imágenes, respuestas reales y predicciones realizadas por los modelos CLIP LabSE y CLIP XML RoBERTa, tanto de manera individual como mediante el ensamble por votación. Se puede observar que en varios casos el ensamble logra recuperar la respuesta correcta incluso cuando uno de los modelos falla en su predicción. Esto sugiere que la combinación de ambos modelos permite corregir errores individuales y reforzar respuestas acertadas. Además, cuando ambos modelos generan la respuesta correcta (figuras 4.2 (e) y (f)), el ensamble mantiene dicha predicción, demostrando su capacidad para consolidar resultados confiables.

Analisis de resultados para las respuesta tipo numé-

	Fig. 4.3 (a)	Fig. 4.3 (b)	Fig. 4.3 (c)	Fig. 4.3 (d)	Fig. 4.3 (e)	Fig. 4.3 (f)
True Answer	20	8	72	452940	2000	75
Predicciones Modelos Baseline (en Inglés)						
CLIP VQA	**20**	**8**	79	unansw	**2000**	unansw
BERTBEiT	unansw	unansw	unansw	unansw	unansw	unansw
BERTViT	black	unansw	unansw	unansw	unansw	unansw
Predicciones Modelos Late Fusion (en Español)						
BERTViT	incont	incont	incont	incont	incont	incont
BERTBEiT	incont	no	incont	incont	si	incont
BETOViT	incont	incont	incont	incont	negro	incont
BETOBEiT	incont	incont	incont	incont	incont	incont
ROBERTAViT	incont	incont	incont	incont	incont	incont
ROBERTABEiT	incont	incont	incont	incont	incont	incont
Predicciones Modelos CLIP VQA (en Español)						
CLIP VQA	**20**	**8**	79	incont	**2000**	68
CLIP VQA Multi	5	pastillas	74	incont	vino	incont
Multi LaBSE	**20**	pastillas	79	incont	**2000**	**75**
Multi XML-RoBERTa	**20**	pastillas	79	incont	**2000**	incont

Cuadro 4.7: Predicciones de los modelos para los pares imagen-pregunta de la Figura 4.3 donde el tipo de respuesta es numerico ('incont' = 'incontestable', 'unasw' = 'unanswerable')

rico

Como se mencionó en la sección anterior, la mayor diferencia en el desempeño se observa en las respuestas de tipo numérico. En la Figura 4.3 se presentan ejemplos de pares (imagen-pregunta, respuesta real), mientras que en el Cuadro 4.7 se muestra la comparación entre el valor de la respuesta real (True Answer) y las predicciones generadas por cada modelo.

Se observa en estos ejemplos que los modelos de fusión múltiple no lograron predecir correctamente las respuestas numéricas, incluso en el baseline en inglés. En contraste, los modelos CLIP demostraron una mayor capacidad para identificar este tipo de respuestas en la mayoría de los casos. Aunque en algunos ejemplos no generaron la respuesta exacta, muestran una mejor habilidad para reconocer y aproximarse a respuestas numéricas.

La superioridad de los modelos CLIP sobre los de late fusion en la predicción de respuestas numéricas puede atribuirse a su capacidad para procesar imagen y texto en un espacio compartido de embeddings, lo que facilita la asociación entre los números presentes en la imagen y la pregunta. Además, su preentrenamiento en pares imagen-texto les otorga una ventaja en la identificación de información numérica sin necesidad de OCR explícito. En cambio, los modelos de Late Fusion procesan ambas modalidades por separado antes de combinarlas, lo que puede dificultar la extracción de información numérica en contexto. Estos resultados sugieren que CLIP ha aprendido a reconocer y relacionar cantidades en imágenes de manera más efectiva que los modelos de fusión tardía.

Otro factor que pudo influir en las predicciones incorrectas es la baja calidad de algunas imágenes, una característica común en el dataset VizWiz, lo que complica la identificación precisa de los números. Por ejemplo, en las figuras 4.3 (c),(d) y (e), los números no son claramente visibles, lo que pudo haber afectado el desempeño de los modelos, incluso aquellos con mejor capacidad para interpretar información numérica. En imágenes como las mostradas en las Figuras 4.3 (a) y 4.3 (f), aunque los valores numéricos son visibles, los modelos requieren una comprensión más profunda del contexto para relacionarlos correctamente con la pregunta formulada. En algunos casos, los modelos lo logran adecuadamente, mientras que en otros, como en la Figura 4.3 (b), la respuesta predicha ("pastillas") no coincide textualmente con la pregunta. No obstante, el modelo identifica correctamente que se trata de una caja de pastillas, lo que evidencia una comprensión parcial del contexto.

(a) Q: Dime qué es esto. A: Carne.
Predicciones: CLIP Labse = pollo, CLIP Xml = carne, **Ensamble = carne**.

(b) Q: ¿De qué color es esto? A: azul.
Predicciones: CLIP Labse = gris, CLIP Xml = azul, **Ensamble = azul**.

(c) Q: ¿Está encendida la luz? A: no.
Predicciones: CLIP Labse = no, CLIP Xml = sí, **Ensamble = no**.

(d) Q: ¿Qué es esto? A: galletas.
Predicciones: CLIP Labse = galletas, CLIP Xml = incontestable, **Ensamble = galletas**.

(e) Q: ¿Cuál es el azul? A: derecha.
Predicciones: CLIP Labse = derecha, CLIP Xml = derecha, **Ensamble = derecha**.

(f) Q: ¿Qué ingredientes tiene esta pizza? A: pepperoni.
Predicciones: CLIP Labse = pepperoni, CLIP Xml = pepperoni, **Ensamble = pepperoni**.

Figura 4.2: Ejemplos de imágenes con sus preguntas correspondientes (Q), la respuesta real (A) y las predicciones generadas por los modelos CLIP LabSE y CLIP XML RoBERTa de forma individual, así como la respuesta del ensamble obtenido mediante votación.

(a) Q: What denomination is this bill?, Q_trad: ¿De qué denominación es este billete?, A: 20

(b) Q: How many tablets are in this box?, Q_trad: ¿Cuántas tabletas hay en esta caja?, A: 8

(c) Q: What temperature is the thermostat set to?, Q_trad: ¿A qué temperatura está ajustado el termostato?, A: 72

(d) Q: What is currently displayed on the screen?, Q_trad: ¿Qué se muestra actualmente en la pantalla?, A: 452940

(e) Q: I believe this is silver oak. What year is it?, Q_trad: Creo que esto es roble plateado. ¿De qué año es?, A: 2000

(f) Q: What's the temperature set at?, Q_trad: ¿A qué temperatura está establecido?, A: 75

Figura 4.3: Ejemplos de imagen-pregunta cuya respuesta es del tipo *number*.

Capítulo 5

Conclusiones

En este estudio, se exploró la tarea de **Respuesta a Preguntas Visuales (VQA)** en español utilizando el dataset **VizWiz** traducido. Se evaluaron diversos enfoques, incluyendo modelos de **fusión tardía**, modelos basados en **CLIP** y **técnicas de ensamble**. Los resultados obtenidos permiten extraer conclusiones clave sobre el desempeño de estos métodos y sus implicaciones para la tarea de VQA en español.

En primer lugar, se observó que la traducción del dataset VizWiz al español introdujo ligeras variaciones en la interpretación de las preguntas y respuestas, lo que impactó de forma leve en el desempeño de los modelos. A pesar de esto, los modelos evaluados lograron adaptarse de manera razonable al español, obteniendo métricas comparables a las reportadas en estudios realizados en inglés.

En cuanto a la efectividad de los enfoques evaluados para VQA en español, los modelos basados en CLIP adaptados a VQA demostraron un rendimiento superior a los modelos de fusión tardía, alcanzando mejores métricas tanto en términos de **Accuracy** como de **VizWiz Accuracy**. Además, la aplicación de técnicas de ensamble mostró ser beneficiosa en varios casos, permitiendo recuperar la respuesta correcta cuando uno de los modelos individuales fallaba.

Finalmente, en la comparación de modelos según las métricas estándar, los modelos **CLIP** lograron un Accuracy cercano al 51 %, mientras que la métrica VizWiz Accuracy alcanzó hasta un 60 % cuando se utilizaron enfoques de ensamble. Estos resultados reflejan la complejidad de la tarea y subrayan la necesidad de seguir explorando enfoques más robustos para la tarea de VQA en español.

Si bien los modelos evaluados lograron desempeños prometedores, el VQA en español sigue siendo un desafío, especialmente en escenarios donde la ambigüedad de la pregunta o la falta de información visual afectan la respuesta. La modelación de la tarea como clasificación multiclase permitió evaluar el impacto del vocabulario de respuestas predefinido, mostrando que la selección de la respuesta más probable puede ser una

estrategia válida, aunque limitada en algunos contextos.

Asimismo, los ensambles de modelos mediante votación mayoritaria demostraron ser una estrategia útil para mejorar la robustez de las predicciones. En particular, la combinación de **CLIP LabSE** y **CLIP XML RoBERTa** permitió corregir errores presentes en los modelos individuales, consolidando respuestas más precisas en escenarios donde uno de los modelos fallaba. Este resultado sugiere que la diversidad de modelos dentro del ensamble contribuye a una mejor generalización en la tarea de VQA.

Como trabajo futuro, se sugiere explorar modelos generativos que ofrezcan mayor flexibilidad en la respuesta, como el modelo **BLIP-2**, el cual podría beneficiarse de un fine-tuning en español. Además, se recomienda mejorar la calidad de la traducción del dataset para reducir los errores lingüísticos que podrían afectar el desempeño de los modelos. También se propone evaluar estrategias de aprendizaje multimodal más avanzadas, tales como modelos que incorporen mecanismos de atención cruzada entre imagen y texto, lo que podría enriquecer la relación entre ambas modalidades. Finalmente, es pertinente optimizar las técnicas de ensamble con el fin de potenciar la complementariedad entre los modelos, mejorando así el desempeño general del sistema.

En conclusión, este estudio proporciona un primer análisis del desempeño de modelos de VQA en español y destaca la importancia de adaptar arquitecturas modernas a contextos multilingües, especialmente en aplicaciones de accesibilidad para personas con discapacidad visual. Los resultados obtenidos establecen una base para futuras mejoras en la tarea de VQA en español, con el potencial de impactar en el desarrollo de herramientas más precisas e inclusivas en este ámbito.

Bibliografía

[1] Aishwarya Agrawal et al. «VQA: Visual Question Answering». En: *International Conference on Computer Vision (ICCV)* (2015). URL: https://arxiv.org/abs/1505.00468.

[2] et al Gurari Danna. «Vizwiz grand challenge: Answering visual questions from blind people». En: *IEEE conference on computer vision and pattern recognition* (2018).

[3] Allan Jabri, Armand Joulin y Laurens Van der Maaten. «Revisiting Visual Question Answering Baselines». En: *European Conference on Computer Vision* (2106).

[4] Yash Goyal et al. «Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering». En: *https://arxiv.org/abs/1612.00837* (2016).

[5] Alec Radford et al. «Learning Transferable Visual Models From Natural Language Supervision.» En: *In International conference on machine learning (pp. 8748-8763). PMLR.* (2021). URL: https://arxiv.org/abs/2103.00020.

[6] F. Deuser et al. «Less is More: Linear Layers on CLIP Features as Powerful VizWiz Model». En: *In International conference on machine learning (pp. 8748-8763). PMLR.* (2022). URL: https://doi.org/10.48550/arXiv.2206.05281.

[7] L. S Coles. «Anon-linequestion-answering systems with natural language and pictorial input». En: *1968 23rd ACM national conference* (1968). URL: https://doi.org/10.1145/800186.810577.

[8] Stevan Harnad. «The symbol grounding problem.» En: *Physica D: Nonlinear Phenomena* (1990).

[9] Peng Wang Qi Wu, Chunhua Shen y Anton van den Hengel. «The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions». En: *Conferencia de Visión por Computadora y Reconocimiento de Patrones (CVPR)* (2017).

[10] M.Malinowski y M.Fritz. «Amulti-world approach to question answering about real world scenes base do nun certain input». En: *Advances in Neural Information Processing Systems(NIPS)* (2014).

[11] P.Kohli N.Silberman D.Hoiem y R.Fergus. «Indoor segmentation and support inference from rgbd images». En: *European Conference on Computer Vision(ECCV)* (2012).

[12] R.Kiros M.Ren y R.Zemel. «Exploring models and data for image question answering». En: *Advances in Neural Information Processing Systems(NIPS)* (2015).

[13] Tsung-YiLin et al. «Microsoft coco: Common objects in context». En: *In European conference on computer vision, pages740–755. Springer* (2014).

[14] S.Antol et al. «VQA Visual question answering». En: *The IEEE International Conference on Computer Vision (ICCV)* (2015).

[15] Y. T. Cao et al. «What's Different between Visual Question Answering for MachineÜnderstanding"Versus for Accessibility?» En: *arXiv preprint arXiv:2210.14966* (2022).

[16] «Visual Question Answering». En: (). URL: https://vizwiz.org/tasks-and-datasets/vqa/.

[17] Chongyan Chen, Samreen Anjum y Danna Gurari. «Grounding Answers for Visual Questions Asked by Visually Impaired People». En: *https://arxiv.org/abs/2202.01993* (2022).

[18] Yu-Yun Tseng, Alexander Bell y Danna Gurari. «VizWiz-FewShot: Locating Objects in Images Taken by People With Visual Impairments». En: *https://arxiv.org/abs/2207.11810* (2022).

[19] Yang Trista Cao et al. «What's Different between Visual Question Answering for Machine 'Understanding' Versus for Accessibility?» En: *https://arxiv.org/abs/2210.14966* (2022).

[20] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». En: *Proceedings of naacL-HLT. Vol. 1.* (2019). URL: https://doi.org/10.48550/arXiv.1810.04805.

[21] Vaswani A. et al. «Attention is All You Need. In Advances in Neural Information Processing Systems». En: *NeurIPS 2017* (2017).

[22] Borja Balle, Paula Cuesta y José Camacho-Collados. «BETO, Bentz et al. Transformers for Spanish». En: *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)* (2019).

[23] Yinhan Liu et al. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». En: (2019). URL: https://doi.org/10.48550/arXiv.1907.11692.

[24] Javier De la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury. «BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling». En: *Procesamiento del Lenguaje Natural* 68.0 (2022), págs. 13-23. ISSN: 1989-7553. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[25] Ashish Vaswani et al. «Attention is all you need». En: *NIPS* (2017).

[26] A. Dosovitskiy et al. «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale». En: (2020). URL: https://doi.org/10.48550/arXiv.2010.11929.

[27] Hangbo Li et al. «BEiT: BERT Pre-Training of Image Transformers». En: (2021). URL: https://doi.org/10.48550/arXiv.2106.08254.

[28] Sheng Shen et al. «How Much Can CLIP Benefit Vision-and-Language Tasks?» En: *In International Conference on Learning Representations* (2022). URL: https://doi.org/10.48550/arXiv.2107.06383.

[29] Zhongzhi Chen et al. «AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities». En: (2022). URL: https://arxiv.org/abs/2211.06679.

[30] Fredrik Carlsson et al. «Cross-lingual and Multilingual CLIP». En: *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (2022). URL: https://aclanthology.org/2022.lrec-1.739/.

[31] Fangxiaoyu Feng et al. «Language-agnostic BERT Sentence Embedding». En: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 878–891.* (2022).

[32] Alexis Conneau et al. «Unsupervised Crosslingual Representation Learning at Scale». En: *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 8440–8451)* (2020).

[33] Linting Xue et al. «mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer». En: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (483–498)* (2021).

[34] Junczys Dowmunt M. et al. «Marian: Fast neural machine translation in C++». En: *Proceedings of ACL 2018, System Demonstrations, 116-121.* (2018).

[35] Brown T. et al. «Language Models are Few-Shot Learners.» En: *Advances in Neural Information Processing Systems (NeurIPS 2020)* (2020).

[36] Clara Villalba. *VIZWIZ_TRAIN_data_with_images (traducción al español).* https://huggingface.co/datasets/claraofvillalba/VIZWIZ_TRAIN_data_with_images. Accedido: 7 septiembre 2025. 2024.

[37] Clara Villalba. *VIZWIZ_VALIDATION_data_with_images (traducción al español).* https://huggingface.co/datasets/

claraofvillalba/VIZWIZ_VALIDATION_data_with_images. Accedido: 7 septiembre 2025. 2024.

[38] Ngiam J. et al. «Multimodal Deep Learning». En: *28th international conference on machine learning (ICML-11) (pp. 689-696)* (2011). URL: https://ai.stanford.edu/~ang/papers/icml11-MultimodalDeepLearning.pdf.

[39] Nils Reimers e Iryna Gurevych. «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks». En: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, nov. de 2019. URL: http://arxiv.org/abs/1908.10084.

[40] Gurari D.and Li Q.and Stangl A. J.and Guo A.and Lin C.and Grauman K. y Bigham J. P. «VizWiz Grand Challenge: Answering Visual Questions from Blind People». En: (2018). URL: https://arxiv.org/abs/1802.08218.