UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

FACULTAD DE INGENIERÍA

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

MODELO *VISUAL QUESTION ANSWERING* PARA LA INTERPRETACIÓN DE IMÁGENES MÉDICAS

POR:

JOSÉ ALBERTO PAYÁN MARTA

CHIHUAHUA, CHIH,. MÉXICO **2 de diciembre de 2024**

Modelo Visual Question Answering para la Interpretación de Imágenes Médicas. Tesis, presentada por José Alberto Payán Marta como requisito parcial para obtener el grado de Maestro en Ingeniería en Computación, ha sido aprobado y aceptado por:

M.I. Fabián Vinicio Hernández Martínez
Director de la Facultad de Ingeniería

Dr. Fernando Martínez Reyes
Secretario de Investigación y Posgrado

Karina Rocío Requena Yáñez
Coordinadora Académica

Graciela Ramirez A.

Graciela María de Jesús Ramírez Alonso
Directora de Tesis

Noviembre 2024

30 de noviembre de 2024

COMITÉ
Dra. Graciela María de Jesús Ramírez Alonso
Dr. Abimael Guzmán Pando
Dr. Jesús Roberto López Santillán
Dra. Olanda Prieto Ordaz

UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

14 de enero de 2025.

ING. JOSE ALBERTO PAYAN MARTA
Presente. -

En atención a su solicitud relativa al trabajo de tesis para obtener el grado de Maestría en Ingeniería en Computación, nos es grato transcribirle el tema aprobado por esta Dirección, propuesto y dirigido por la directora **Dra. Graciela María de Jesús Ramírez Alonso** para que lo desarrolle como Tesis, con el título **"Modelo visual question answering para la interpretación de imágenes médicas"**.

Índice de Contenido

FACULTAD DE INGENIERÍA
Circuito No. 1, Campus Universitario 2
Chihuahua, Chih., México. C.P. 31125
Tel. (614) 442-95-00
www.fing.uach.mx

Medical Vision Language Learner

M2I2 (Masked language modeling, Masked image modeling, Image text matching and Image text contrastive learning)

MUMC (Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses for Medical Visual Question Answering)

Metodología

Adquisición, descripción y análisis de la base de datos

Tipos de preguntas

Tipos de respuestas

Análisis de las imágenes, preguntas y respuestas

Preprocesamiento de datos

Filtrado y clasificación de las entradas

Limpieza del texto

Tokenizacion, codificación, y enmascaramiento

Procesamiento de imágenes y etiquetas

Modelo BioMedViLL

Métrica de evaluación

Evaluación

UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

Resultados cuantitativos

Desempeño en Preguntas Abiertas (OPEN)

Desempeño en Preguntas Cerradas (CLOSED)

Desempeño General (OVERALL)

Resultados cualitativos

Conclusiones y Trabajo Futuro

Conclusiones

Trabajo Futuro

Referencias

Curriculum Vitae

A T E N T A M E N T E

"naturam subiecit aliis"

EL DIRECTOR

M.I. FABIÁN VINICIO HERNÁNDEZ MARTÍNEZ

FACULTAD DE INGENIERÍA U.A.CH.

SECRETARIO DE INVESTIGACIÓN Y POSGRADO

DR. FERNANDO MARTÍNEZ REYES

DIRECCIÓN

Dedicatoria

Dedico este logro a mis padres, el Ing. José Payán Soto y Blanca Estela Marta Portillo, a mi abuela Socorro Portillo Hidalgo†, quienes han sido mi apoyo constante, mi inspiración y mis más grandes motivadores. Gracias por brindarme vida, amor, paciencia y por impulsarme siempre a ser una mejor persona. Este esfuerzo también es suyo.

A mi prometida, la Lic. Alba Celeste Bejarano Ordóñez, gracias por acompañarme en cada paso de este viaje, por tu amor y tu inquebrantable paciencia. Has sido mi fortaleza en los momentos difíciles, y sin ti, este sueño no sería realidad.

Agradecimientos

Agradezco a mi directora de tesis por guiarme durante el transcurso de mi maestría, al Consejo Nacional de Humanidades, Ciencias y Tecnología (CONAHCYT) por brindarme el apoyo económico necesario para realizar una maestría, sin ese apoyo no habría podido obtener dicho grado en mi trayectoria profesional. La mejor manera de retribuir el apoyo que se me brindó es con el presente trabajo de investigación y con los conocimientos que deseo compartir.

Asimismo, agradezco a la Universidad Autónoma de Chihuahua especialmente a la Facultad de Ingeniería por haberme dado la oportunidad de ser parte del cuerpo estudiantil de la Maestría en Ingeniería en Computación. Gracias a esta oportunidad pude enriquecer mi formación académica, profesional y personal.

Por último, a mi familia quienes siempre estuvieron conmigo a lo largo de este proceso de mi vida. A mi prometida Alba quien siempre estuvo a mi lado con su amor y ayuda incondicional. Sin su gran apoyo esto no hubiera sido posible.

Resumen

Este trabajo de investigación explora el análisis e implementación de tres diferentes modelos de aprendizaje profundo diseñados para abordar tareas de *Visual Question Answering* (VQA) en el contexto de interpretación de imágenes médicas. El desempeño y funcionamiento de los modelos *Medical Vision Language Learner* (MedViLL), *Masked language modeling, Masked image modeling, Image text matching and Image text contrastive learning* (M2I2), y *Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses* (MUMC) se analiza a través de diferentes métricas usando la base de datos de VQA-RAD. Estos modelos combinan algoritmos de visión por computadora y procesamiento de lenguaje natural para responder preguntas relacionadas con imágenes radiológicas, con el fin de mejorar el diagnóstico médico y la eficiencia en el análisis de imágenes médicas. El entrenamiento de estos modelos incluye técnicas avanzadas que enmascaran partes de las imágenes y el texto para que se aprenda a predecir regiones o palabras en función del contexto, mejorando así su capacidad de generalización. Así mismo, con base en la arquitectura de estos modelos base, en este trabajo de tesis, se propone uno nuevo, que lleva por nombre BioMedViLL, al cual se le incorpora un *Vision Transformer* ViT-B32 como codificador visual y un *Bio ClinicalBERT* como codificador textual. Este trabajo presenta el resultado de varios experimentos que se llevaron a cabo para medir y comparar el desempeño de dichos modelos, evaluando su precisión en las respuestas a preguntas abiertas y cerradas. En este análisis, MUMC destacó por obtener respuestas más completas y de mejor calidad, mientras que las respuestas generadas por BioMedViLL aún y cuando superaron cuantitativamente a MUMC, no alcanzaron el mismo nivel de calidad, lo que sugiere que obtener buenas métricas no siempre garantiza resultados de alta calidad.

Palabras clave: VQA-RAD, Vision Transformer, BioMedViLL, MUMC, M2I2, MedViLL, ViT-B-32, Bio_ClinicalBERT.

Índice

Índice de cuadros

Índice de figuras

Introducción

Los avances en modelos de aprendizaje automático, en especial en el área de aprendizaje profundo (Deep Learning) han abierto nuevas posibilidades en el ámbito de la inteligencia artificial (IA). Por ejemplo, en el campo de la visión por computadora (Computer Vision o CV), se han logrado avances significativos en la detección de objetos a través del procesamiento de imágenes con el uso de redes neuronales convolucionales (Jahne y HauBecker, 2000). De manera similar, en el procesamiento de lenguaje natural (Natural Language Processing o NLP), se han alcanzado mejoras en el reconocimiento del habla mediante el procesamiento de texto, gracias a los modelos de lenguaje grandes que permiten capturar largas secuencias de texto, lo cual mejora las tareas de traducción de texto (Otter, Medina, y Kalita, 2020).

Visual Question Answering o "VQA" es un enfoque que combina visión por computadora y procesamiento de lenguaje natural con el objetivo de responder preguntas relacionadas con imágenes. En el ámbito de la medicina, donde las imágenes son cruciales para el diagnóstico y tratamiento, herramientas de modelos multimodales mediante VQA ofrecen oportunidades significativas en radiología. Estos modelos pueden ayudar a los profesionales médicos a gestionar eficientemente grandes volúmenes de imágenes al proporcionar respuestas a preguntas específicas sobre su contenido. Sin embargo, el progreso en el desarrollo de modelos de VQA se ha visto obstruido por la escasez de conjuntos de datos con pares de preguntas-respuestas diseñadas específicamente para imágenes médicas en radiología.

VQA-RAD (Lau, Gayen, Ben Abacha, y Demner-Fushman, 2018), es un conjunto de datos construido manualmente donde un grupo de médicos formularon preguntas y proporcionaron respuestas de referencia sobre imágenes de radiología. La base de datos VQA-RAD fue elaborada con imágenes de casos clínicos de MedPix, un archivo de radiología de acceso abierto. Se crearon preguntas y respuestas a través de la contribución de 15 clínicos, utilizando un enfoque de dos partes: preguntas libres y preguntas reformuladas. Este conjunto de datos está archivado y disponible en formatos JSON, XML y Excel, contiene 3,515 preguntas visuales, clasificadas y validadas manualmente, con 315 imágenes correspondientes.

Los autores de VQA-RAD implementaron dos modelos de VQA para la base de datos, los cuales llevan por nombre *Multimodal Compact Bilinear Pooling for Visual Question Answering (MCB_VQA)* (Fukui y cols., 2016) y *Stacked Attention Networks for Image Question Answering (SAN)* (Yang, He, Gao, Deng, y Smola, 2016), realizando un análisis de las preguntas y respuestas, así como de las relaciones entre tipos de preguntas, tipos de respuestas e imágenes de diferentes regiones anatómicas. Para medir el rendimiento de dichos modelos, los autores utilizaron las métricas de precisión simple, precisión promedio y "Bilingual Evaluation Understudy" o BLEU (Papineni, Roukos, Ward, y Zhu, 2002), tanto en preguntas abiertas y cerradas. En el caso de la precisión general, los autores lograron resultados de 54.2 % y 54.6 % para preguntas abiertas y cerradas, respectivamente.

Algunas de las limitaciones que tiene VQA-RAD es el tamaño reducido de datos, lo que puede dificultar la generalización de los modelos entrenados, ya que no cubre una variedad suficiente de escenarios clínicos y radiológicos. Además, la naturaleza de las preguntas y respuestas tiende a ser sencilla y específica, lo que limita la evaluación de la capacidad de los modelos para responder preguntas más complejas o abiertas. También existe un desequilibrio en la distribución de las preguntas, con predominancia de preguntas de tipo binario ("sí/no") o de categorías específicas, lo que puede sesgar el rendimiento de los modelos. Por último, la base de datos depende de anotaciones realizadas por humanos, lo que puede introducir subjetividad o errores en las respuestas de referencia, afectando la precisión de la evaluación.

En esta investigación, se plantea un análisis detallado y una experimentación rigurosa con diversos modelos de *Visual Question Answering* (VQA), utilizando la base de datos VQA-RAD como punto de referencia para su evaluación. El principal objetivo es profundizar en la comprensión de los mecanismos y el rendimiento de estos modelos en la interpretación de imágenes médicas, un área crítica para el desarrollo de herramientas de apoyo en el diagnóstico clínico. Este estudio ha permitido, además, el desarrollo de un nuevo modelo denominado BioMedVill, diseñado específicamente para abordar algunas de las limitaciones observadas en los modelos previos. La evaluación presentada en este trabajo abarca tanto aspectos cuantitativos, mediante métricas de precisión y rendimiento, como cualitativos, centrándose en la interpretación de las respuestas generadas por

los modelos. De esta forma, se busca identificar qué enfoques de aprendizaje profundo pueden ofrecer beneficios concretos para profesionales y estudiantes del ámbito de la salud, mejorando su capacidad para interpretar imágenes médicas de manera precisa y eficiente.

Justificación

En el área de medicina, la interpretación de imágenes es fundamental para el correcto diagnóstico de enfermedades, un diagnóstico incorrecto debido a una interpretación errónea de una imagen médica puede causar graves daños en el paciente. VQA puede mejorar la comprensión y el análisis de imágenes médicas al permitir que los médicos formulen preguntas específicas sobre estas imágenes y obtengan respuestas precisas. Esto lleva a diagnósticos más acertados y tratamientos más efectivos.

Los modelos de VQA pueden utilizarse en aplicaciones educativas para enseñar a estudiantes de medicina a analizar imágenes médicas y tomar decisiones basadas en información visual. Asimismo, ayudan a mejorar la calidad de la atención médica y, en última instancia, a mejorar la formación de profesionales de la salud.

También ayuda a reducir errores humanos, ya que las herramientas de VQA proporcionan información objetiva y precisa en respuesta a preguntas clínicas. La investigación en VQA médica es crucial para abordar desafíos específicos en el campo de la salud, como el diagnóstico, la toma de decisiones clínicas, la educación médica y la mejora de la atención al paciente. Además de contribuir a la eficiencia y la seguridad de los procesos médicos, así como a la investigación clínica y la colaboración interdisciplinaria.

Objetivos

Objetivo general:

Analizar y evaluar distintos modelos de *Visual Question Answering* (VQA) aplicados a la interpretación de imágenes médicas con el fin de desarrollar y proponer un nuevo modelo que logre resultados competitivos con el estado del arte.

Objetivos específicos:

1. Identificar y analizar los diferentes modelos que utilizan la base de datos de VQA-RAD para su evaluación.
2. Implementar y entender su funcionamiento en la generación de respuestas a preguntas de índole médica.
3. Proponer un nuevo modelo, el cual logre resultados competitivos con los encontrados en el estado del arte.
4. Realizar un análisis cualitativo y cuantitativo de dichos modelos, para la evaluación de la calidad de sus predicciones.

Antecedentes

VQA-RAD es una base de datos popularmente utilizada por diferentes investigadores para evaluar modelos VQA, aunque no es la única con la que se cuenta. A continuación se mencionarán trabajos importantes que la han utilizado en conjunto con otras bases de datos.

"Towards Visual Question Answering on Pathology Images"

X. He y cols. (2021) centraron su investigación en el desarrollo de un marco de trabajo de VQA para tareas médicas con el objetivo de analizar imágenes patológicas y responder preguntas médicas relacionadas con estas imágenes. Los autores presentaron el conjunto de datos PathVQA, que consta de 32,795 preguntas generadas a partir de 4,998 imágenes médicas. Además, propusieron un marco de trabajo de optimización de tres niveles que involucra pre entrenamiento autosupervisado y ajuste fino de VQA para aprender representaciones visuales y textuales.

Para superar los desafíos de crear conjuntos de datos médicos para VQA, los autores utilizaron libros de texto médicos y bibliotecas digitales en línea para extraer imágenes y subtítulos. Además, médicos verificaron los pares de preguntas y respuestas para garantizar la coherencia clínica y la

corrección.

El conjunto de datos resultante abarca diversos conceptos médicos. El marco de trabajo de optimización de tres niveles incluye:

1. Pre entrenamiento autosupervisado: se realizó un pre entrenamiento autosupervisado cruzado de imágenes médicas y preguntas, donde una tarea del procesamiento del texto implica determinar si una pregunta se refiere a una imagen dada.

2. Ajuste fino (Fine-Tuning) de VQA: los codificadores de la imagen y el texto pre entrenados se ajustan finamente en el conjunto de datos PathVQA. Se adecuaron los codificadores para estar cerca de los pesos óptimos obtenidos del pre entrenamiento autosupervisado, facilitando la transferencia de representaciones aprendidas.

3. Validación: el modelo entrenado se valida en un conjunto separado.

Los autores experimentan con dos métodos de VQA y demuestran que su marco de trabajo de optimización de tres niveles mejora el rendimiento de estos métodos. Evalúan los modelos utilizando métricas de exactitud (Accuracy), F1-Score y BLEU. Los resultados para los 2 métodos en la métrica de exactitud fueron de 63.4 % y 58.9 %, respectivamente.

El artículo contribuye creando el conjunto de datos PathVQA, proponiendo el marco de trabajo de optimización de tres niveles y demostrando su eficacia en VQA para tareas médicas.

Pmc-vqa: Visual instruction tuning for medical visual question answering

Zhang y cols. (2023), abordan el problema de VQA Médica (MedVQA), que es crucial para interpretar eficientemente imágenes médicas con información relevante para el diagnóstico clínico. Los autores proponen un modelo llamado MedVInT, que combina un codificador de imágenes pre entrenado con un modelo de lenguaje.

Se utiliza el conjunto de datos de MedVQA a gran escala llamado PMC-VQA, con 227,000 pares de preguntas y respuestas de 149,000 imágenes que abarcan diversas enfermedades. El modelo

propuesto es pre entrenado en PMC-VQA y se ajusta finamente en conjuntos de datos públicos, superando a los trabajos existentes.

En cuanto a la arquitectura, se presentan dos variantes del modelo, MedVInT-TE y MedVInT-TD, adaptados para modelos de lenguaje con codificador y decodificador, respectivamente.

1. MedVInT-TE, utiliza una ResNet-50 (K. He, Zhang, Ren, y Sun, 2016) pre entrenada como extractor de características visuales (Codificador Visual o Visual encoder), en el caso del texto, se utiliza GPT (Radford, Narasimhan, Salimans, Sutskever, y cols., 2018) como modelo de lenguaje (Codificador de Lenguaje o Language encoder) para codificar las preguntas y respuestas en embeddings; un embedding es una representación vectorial para elementos como palabras o frases, que permite capturar sus características semánticas y relaciones en un espacio numérico.

   Por último, las salidas del codificador visual y del codificador de lenguaje son concatenadas para ser procesadas por un decodificador multimodal (Transformer de 4 capas), que se encarga de generar las respuestas a las preguntas de las imágenes.

2. MedVInT-TD, el codificador visual y el codificador de lenguaje son iguales que los de MedVInT-TE, para el decodificador multimodal se pre entrenan los pesos para generar secuencias de texto, mientras que para el modelo MedVInT-TE el decodificador multimodal se entrena desde cero.

El trabajo evalúa el rendimiento del modelo propuesto en comparación con otros conjuntos de datos existentes como VQA-RAD y SLAKE (B. Liu y cols., 2021). MedVInT-TE y MedVInT-TD lograron un desempeño de 78.2% y 81.6%, respectivamente. Los resultados muestran un rendimiento superior en términos de exactitud.

Una de las limitaciones de los modelos es su capacidad de adaptación a la base de datos VQA-RAD. Dado que PMC-VQA es significativamente más grande, permite una mejor representación de las relaciones visuales y textuales, algo que no se logra completamente con el alcance limitado de VQA-RAD. Las contribuciones principales incluyen el modelo MedVInT, el conjunto de datos

PMC-VQA y una evaluación para MedVQA.

Does clip benefit visual question answering in the medical domain as much as it does in the general domain?

Eslami, de Melo, y Meinel (2021), presentan el trabajo sobre la evaluación de Contrastive Language–Image Pre-training o CLIP (Radford y cols., 2021) para la tarea de Visual Question Answering Médica (MedVQA). Este trabajo evaluó la eficacia de CLIP para el dominio médico, presentando PubMedCLIP, una versión de CLIP para tareas médicas basadas en artículos de PubMed.

Se utilizan dos conjuntos de datos de MedVQA (VQA-RAD y SLAKE) y dos métodos (MEVF y QCR) para evaluar el rendimiento de PubMedCLIP en comparación con el CLIP original y el Model-Agnostic Meta-Learning o MAML (Finn, Abbeel, y Levine, 2017).

Los autores proponen PubMedCLIP como un codificador visual pre entrenado para imágenes médicas, demostrando mejoras en las métricas para la tarea de MedVQA en comparación con enfoques anteriores. PubMedCLIP utiliza un ViT32 (Dosovitskiy y cols., 2020) como codificador visual pre entrenado en una ResNet-50.

Como codificador de lenguaje, se utiliza un Transformer (Vaswani y cols., 2017) con 8 Multi-Head Attention (Vaswani y cols., 2017) para crear los embeddings de las preguntas. Una vez que se tienen los embeddings visuales y de texto, se utiliza una Long Short Term Memory o LSTM (Hochreiter y Schmidhuber, 1997) para codificar las respuestas para finalmente concatenar los embeddings y comparar los resultados en un perceptrón multicapa (Multi-Layaer Perceptron o MLP) como clasificador.

Los resultados indican que el rendimiento varía según el tipo de pregunta y la región del cuerpo en las imágenes médicas, como huesos o tejidos fácilmente identificables, que tienden a obtener mejores resultados debido a patrones visuales más claros en las imágenes. Por otro lado, las preguntas que involucran regiones más complejas o con menor contraste visual, como tejidos blandos o áreas superpuestas, presentan un rendimiento inferior.

Los autores discutieron los desafíos y mejoras en la interpretación de preguntas médicas y respuestas en imágenes. También identificaron diferencias de distribución en los conjuntos de datos de MedVQA (VQA-RAD y SLAKE), lo que afectó el rendimiento del codificador visual de PubMedCLIP, el cual obtuvo un desempeño general de 66.5 % y 72.1 % con los métodos MEVF y QCR, respectivamente; aun así, supera al codificador visual de MAML con mejoras de hasta el 3 % en la precisión general de las respuestas.

Self-supervised vision-language pretraining for medial visual question answering

Li, Liu, Tan, Liao, y Zhong (2023), presentan un modelo llamado M2I2 para VQA médico. M2I2 utiliza un método de auto-entrenamiento que incluye: modelado de imágenes enmascaradas (Bao, Dong, Piao, y Wei, 2021) (Masked Image Modeling o MIM). Modelado de lenguaje enmascarado (Devlin, Chang, Lee, y Toutanova, 2018) (Masked Language Modeling o MLM), coincidencia de texto e imagen (Faghri, Fleet, Kiros, y Fidler, 2017) (Image Text Matching o ITM) y aprendizaje contrastivo de texto e imagen (Radford y cols., 2021) (Image Text Contrastive Learning o ITC).

Este enfoque evalúa su rendimiento en tres conjuntos de datos de VQA médica (VQA-RAD, PathVQA y SLAKE), destacando la eficacia del enfoque propuesto. La arquitectura de M2I2 consta de dos etapas principales: pre entrenamiento y ajuste fino.

1. Pre entrenamiento:

    a) Codificador Visual. Se utiliza un codificador de imágenes basado en un ViT de 12 capas para representar las imágenes radiográficas.

    b) Codificador Textual. Se emplea el modelo BERT (Devlin y cols., 2018) pre entrenado con las primeras 6 capas para extraer características del texto asociado a las imágenes.

    c) Codificador Multimodal. Se utilizan las últimas 6 capas de BERT como codificador multimodal para aprender patrones entre la información visual y del lenguaje.

    d) Decodificador Visual. Se utiliza un decodificador de imágenes de 8 capas inspirado en Masked Autoencoder (K. He y cols., 2022) (MAE) para reconstruir parches de imágenes enmascaradas aleatoriamente.

2. Ajuste Fino:

    a) Decodificador Textual. Se incorpora un decodificador BERT pre entrenado para generar respuestas.

    b) Pérdida de Modelado de Lenguaje Condicional (Brown y cols., 2020) (Conditional language-modeling loss). Se utiliza una función de pérdida para guiar el ajuste fino, considerando el contexto de la salida del codificador multimodal y las etiquetas (ground truth).

MIM utiliza un auto-codificador enmascarado para reconstruir parches de imágenes enmascaradas. MLM aplica un modelo de lenguaje enmascarado para predecir tokens de texto enmascarado, considerando tanto la información del texto como de la imagen. ITM entrena el modelo para realizar clasificación binaria de pares de texto e imagen, con las salidas obtenidas del codificador multimodal. ITC maximiza las similitudes entre imágenes y texto utilizando un enfoque de aprendizaje contrastivo.

M2I2 obtiene un desempeño general de 73.7 %, lo cual demuestra que la combinación del pre entrenamiento y el ajuste fino contribuye a aprender representaciones de características unimodales y multimodales que se transfieren a tareas específicas de VQA médica, logrando un mejor rendimiento en comparación con métodos existentes en conjuntos de datos VQA médica. Estos resultados resaltan la importancia del pre entrenamiento y se destaca la capacidad del modelo para abordar preguntas tanto abiertas como cerradas en VQA médica.

PeFoMed: Parameter Efficient Fine-tuning on Multimodal Large Language Models for Medical Visual Question Answering

En G. Liu, He, Li, Zhao, y Zhong (2024), presentan un enfoque para abordar las tareas de VQA médica, utilizando Modelos de Lenguaje Multimodales Grandes (Radford y cols., 2021) (Multimodal Large Language Models o MLLMs).

Los MLLMs combinan visión por computadora y el procesamiento del lenguaje natural, para generar respuestas en forma de texto libre para preguntas sobre imágenes médicas. El marco propuesto,

llamado PeFoMed, se centra en el ajuste fino de parámetros de MLLMs para aplicaciones de VQA médica.

El modelo se entrena en dos etapas, aprovechando un MLLMs pre entrenado en un conjunto de datos multimodal general, haciendo un ajuste fino en pares de imágenes y descripciones médicas y un conjunto de datos específico de VQA médica llamado VQA-Med (Ben Abacha, Hasan, Datla, Demner-Fushman, y Müller, 2019). El marco propuesto utiliza una estrategia de ajuste fino en dos etapas.

En la primera etapa, el modelo se ajusta en un conjunto de datos general de imágenes y descripciones, y en la segunda etapa, se somete a un ajuste fino en un conjunto de datos específico de VQA-Med.

La arquitectura de PeFoMed consta de tres partes principales:

1. Codificador Visual. Utiliza un esquema de codificación visual denominado Extended Vision Transformer Architecture (EVA), que pertenece a la categoría de modelos Transformers Visuales (ViT). Este codificador toma los tokens (unidades mínimas que representan información en texto o imágenes) de imágenes médicas y los transforma en representaciones visuales.

2. Modelo de Lenguaje Multimodal Grande. Utiliza un MLLM pre entrenado denominado Mini GPT-v2. El MLLM ha sido entrenado en un conjunto de datos multimodal en un dominio general. Actúa como un decodificador que procesa entradas multimodales y genera respuestas en forma de texto.

3. Capa de Proyección Lineal (Linear Projection Layer). Esta capa realiza una proyección lineal de las representaciones visuales codificadas por el ViT en el espacio del MiniGPT-v2. En otras palabras, transforma las características visuales de las imágenes en un formato que puede ser interpretado por el MiniGPT-v2.

Además, se utiliza una técnica llamada Low-Rank Adaptation (Hu y cols., 2021) (LoRA) en el MiniGPT-v2 durante la fase de ajuste fino. Esta técnica permite ajustar de manera eficiente solo

una pequeña parte del modelo, reduciendo así el número de parámetros que se actualizan durante el entrenamiento y, por lo tanto, minimizando el tiempo de entrenamiento.

La estructura general de la entrada multimodal se define mediante una serie de instrucciones que incluye información visual y de pregunta y se pasa al MiniGPT-v2 pre entrenado para generar respuestas.

PeFoMed logra una precisión general del 81.9 %, superando al modelo GPT-4v por un margen significativo del 26 % en precisión absoluta en preguntas cerradas. El modelo propuesto demuestra superioridad, especialmente en la resolución de preguntas abiertas.

PeFoMed se compara con métodos anteriores, demostrando su eficacia en la generación de respuestas precisas tanto para preguntas abiertas como cerradas en tareas de VQA-Med.

Contrastive training of a multimodal encoder for medical visual question answering

En (Silva, Martins, y Magalhães, 2023), proponen un modelo que combina un codificador de imágenes basado en una EfficientNetV2 (Tan y Le, 2021) con un codificador multimodal basado en la arquitectura RealFormer (R. He, Ravula, Kanagal, y Ainslie, 2020).

El modelo se pre entrena utilizando aprendizaje contrastivo (Chen, Kornblith, Norouzi, y Hinton, 2020) (Contrastive Learning) y se ajusta finamente a la tarea de VQA utilizando una función de pérdida que aborda específicamente el desbalance de clases.

Los experimentos realizados en el conjunto de datos VQA-Med de ImageCLEF2019, demuestran la eficacia del enfoque, el cual obtuvo un desempeño de 58.2 %, destacando los beneficios del pre entrenamiento multimodal. Los autores presentan el modelo Multimodal Medical BERT (MMBERT), que es un modelo base para VQA médica que utiliza un codificador de imágenes con una ResNet152 y un codificador textual basado en Transformers.

MMBERT se pre entrena en el conjunto de datos Radiology Objects in COntext (Pelka, Koitka, Rückert, Nensa, y Friedrich, 2018) (ROCO) utilizando Modelado de Lenguaje Enmascarado (MLM) y se ajusta finamente para VQA.

A continuación se describen los componentes clave de la arquitectura del modelo:

1. Codificador visual. Utiliza un EfficientNetV2 como un codificador de imágenes. Este componente se encarga de procesar y extraer características importantes de las imágenes médicas.

2. RealFormer. Se incorpora una capa de autoatención multimodal basada en RealFormer, que es una variante de los Transformers y que son conocidos por su capacidad para encontrar relaciones en datos secuenciales extensos.

   La autoatención multimodal permite al modelo considerar simultáneamente información visual y textual, mejorando la comprensión de las relaciones entre las imágenes y las preguntas.

3. SERF. Se utiliza una función de activación SERF para mejorar el rendimiento del modelo. La elección de esta función de activación específica se traduce en mejoras tanto en la precisión general como en la puntuación BLEU.

4. Aprendizaje contrastivo multimodal (Radford y cols., 2021) (Multimodal Contrastive Learning o MCL). En el pre entrenamiento, el modelo se entrena utilizando aprendizaje contrastivo multimodal. Este enfoque implica aprender representaciones semánticas al hacer que las instancias similares se acerquen en el espacio de representación, mientras que las instancias no relacionadas se alejen.

   Este pre entrenamiento mejora la capacidad del modelo para comprender la similitud entre las imágenes y las preguntas.

Los autores intentan reproducir los resultados originales de MMBERT, pero los valores predeterminados del código fuente no logran los mismos resultados, llevando a una exploración de diferentes hiperparámetros. A pesar de los ajustes en el tamaño del lote y el número de épocas, no se alcanzan los resultados originales de MMBERT.

Después, los autores presentan los resultados del nuevo modelo propuesto. Se realizan ajustes en el codificador visual, la arquitectura del Transformer multimodal, la función de activación, la función de pérdida para el ajuste fino y la tarea de pre entrenamiento. El rendimiento se mide en precisión general y con la métrica BLEU en varias categorías.

Los resultados indican mejoras al utilizar EfficientNetV2, una arquitectura multimodal específica, la función de activación SERF y la pérdida asimétrica, mejorando un 4.1 % global. La tarea de pre entrenamiento con aprendizaje contrastivo muestra beneficios, especialmente al emplear la similitud de Jaccard y la similitud de coseno. Finalmente, los resultados demostraron que el modelo propuesto, que combina una EfficientNetV2 con un Transformer multimodal y aprendizaje contrastivo, superan los resultados de MMBERT.

La Tabla 1 presenta un resumen de las métricas de los artículos previamente analizados, resaltando que los modelos M2I2 y PeFoMed lograron los porcentajes más altos en la exactitud para las respuestas abiertas, cerradas y en general, respectivamente, dentro de la base de datos de VQA-RAD.

Tabla 1: Resumen de trabajos del estado del arte que han utilizado VQA-RAD.

Modelos	**VQA-RAD**		
Exactitud (Accuracy)	Abiertas	Cerradas	En general
MedVInT-TE (Zhang y cols., 2023)	69.3 %	84.2 %	78.2 %
MedVInT-TD (Zhang y cols., 2023)	73.7 %	86.8 %	81.6 %
PubMedCLIP (MEVF) (Eslami y cols., 2021)	48.6 %	78.1 %	66.5 %
PubMedCLIP (QCR) (Eslami y cols., 2021)	60.1 %	80 %	72.1 %
M2I2 (Li, Liu, Tan, y cols., 2023)	**83.5 %**	66.5 %	76.8 %
PeFoMed (G. Liu y cols., 2024)	73.7 %	**87.4 %**	**81.9 %**

Considerando todo lo anterior, en el marco de esta investigación, se pretende construir un modelo de VQA para tareas médicas, basado en Transformers, destinado a mejorar el rendimiento en las respuestas a preguntas médicas.

Marco Teórico

Transformer

Un "Transformer" es una arquitectura de aprendizaje profundo que se basa en un mecanismo de atención (self-attention), que permite a la red neuronal capturar relaciones entre palabras en una oración de manera simultánea, independientemente de su posición en el texto.

La arquitectura del modelo de Transformer presentada en el trabajo de Vaswani y cols. (2017) se muestra en la Figura 1 la cual consta de dos partes principales, un codificador o "Encoder" (izquierda) y un decodificador o "Decoder" (derecha). Ambos están formados por una serie de capas como se describe a continuación:

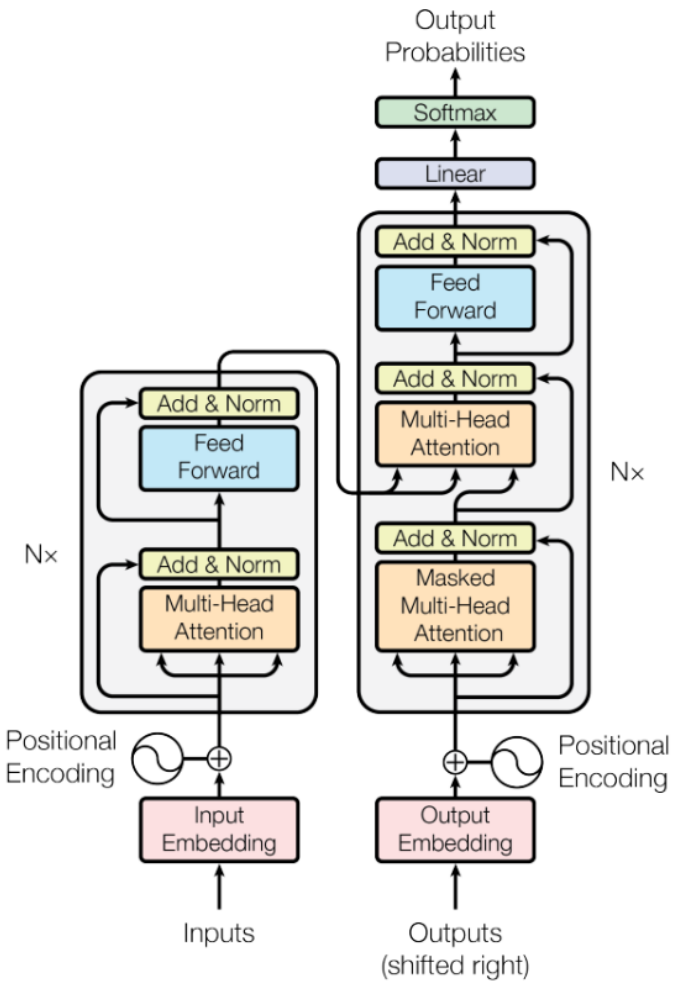


Figura 1: Arquitectura de un Transformer (Vaswani y cols., 2017, 3).

1. Atención multi-cabeza (Multi-Head Attention): esta capa se muestra en la Figura 2 y calcula la atención ponderada sobre un conjunto de vectores de entrada, donde los pesos de atención actúan como los pesos de la suma. Se divide en tres sub capas: la capa de consulta o "Q" (Query), la capa de llave o "K" (Key) y la capa de valor o "V" (Value). Dentro de la capa de atención ocurren varios procesos, los cuales se describen a continuación.

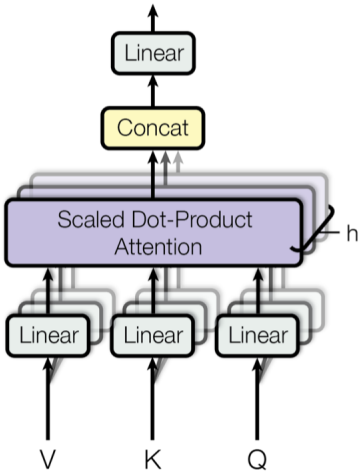


Figura 2: Atención multi-cabeza (Vaswani y cols., 2017, 4).

- Q, K y V: cada palabra en la secuencia de entrada se representa en estas tres formas diferentes, estas representaciones se utilizan para calcular la atención. Se calculan mediante transformaciones lineales de la entrada.
- Atención con producto punto escalado entre Q y K (Scaled Dot-Product Attention): la Figura 3 muestra el cálculo de la afinidad entre cada par de palabras en la secuencia de entrada. Esta afinidad se utiliza para calcular los pesos de atención. La atención con producto punto escalado se calcula con la siguiente ecuación:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

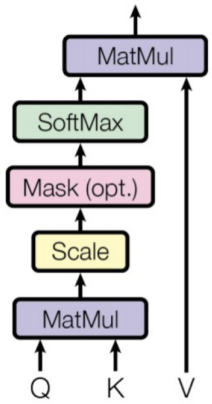


Figura 3: Atención escalonada de producto escalado (Vaswani y cols., 2017, 4).

Donde Q, K, V son las matrices de consulta, llave y valor, respectivamente, y d_k es la dimensión de K.

La ecuación para calcular la atención ponderada es la siguiente:

$$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O \tag{2}$$

Donde $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ y las proyecciones son matrices de parámetros $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ y $W_i^O \in \mathbb{R}^{hd_{\text{v}} \times d_{\text{model}}}$.

- Softmax y pesos de atención: los pesos de atención se calculan aplicando la función softmax sobre los productos punto obtenidos. Estos pesos indican cuánta importancia debe asignarse a cada palabra en función de su relación con las demás palabras en la secuencia.

2. Red neuronal pre-alimentada (feed forward network o FFN): después de la capa de atención, se aplica una red neuronal a cada posición de la secuencia por separado. La ecuación que describe ese proceso es la siguiente:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

Donde x es la secuencia, W_1, b_1, W_2 y b_2 son los parámetros de la red neuronal.

3. Capas de normalización: después de cada capa (tanto la capa de atención como la red neuronal), se aplica una capa de normalización para reducir la complejidad computacional y acelerar el entrenamiento de la red.

4. Conexiones residuales: se agregan conexiones residuales al final de la capa de normalización y de la red neuronal para evitar el desvanecimiento de gradiente (vanishing gradient). Se calculan como:

$$Output = LayerNorm\left(x + SubLayer\left(x\right)\right) \tag{4}$$

   Donde $SubLayer(x)$ es la conexión residual.

5. Capa de codificador-decodificador de atención (encoder-decoder attention layer): esta capa permite que el decodificador acceda a la información del codificador para ayudar en la generación de la salida del Transformer.

   - Salida del codificador-decodificador de atención: la salida es una combinación lineal ponderada de los valores del codificador, donde los pesos son determinados por la atención calculada.

En la entrada de un Transformer, las secuencias de datos se dividen en tokens. Cada token representa una unidad discreta en la secuencia (por ejemplo, una palabra). Cada token es una representación vectorial que servirá como la entrada inicial al Transformer.

El Transformer se entrena utilizando algoritmos de optimización y back-propagation, y se ha demostrado que es altamente efectivo en tareas de procesamiento de lenguaje natural, traducción automática, generación de texto y una amplia gama de aplicaciones de visión por computadora.

Vision Transformer

Un Vision Transformer o "ViT" (Dosovitskiy y cols., 2020) es un modelo basado en Transformers enfocado a tareas de visión por computadora. La arquitectura de ViT se muestra en la Figura 4, la

cual tiene como componentes principales un codificador de un Transformer y embedding de parches de imágenes. Ambos están formados por una serie de capas como se describe a continuación:

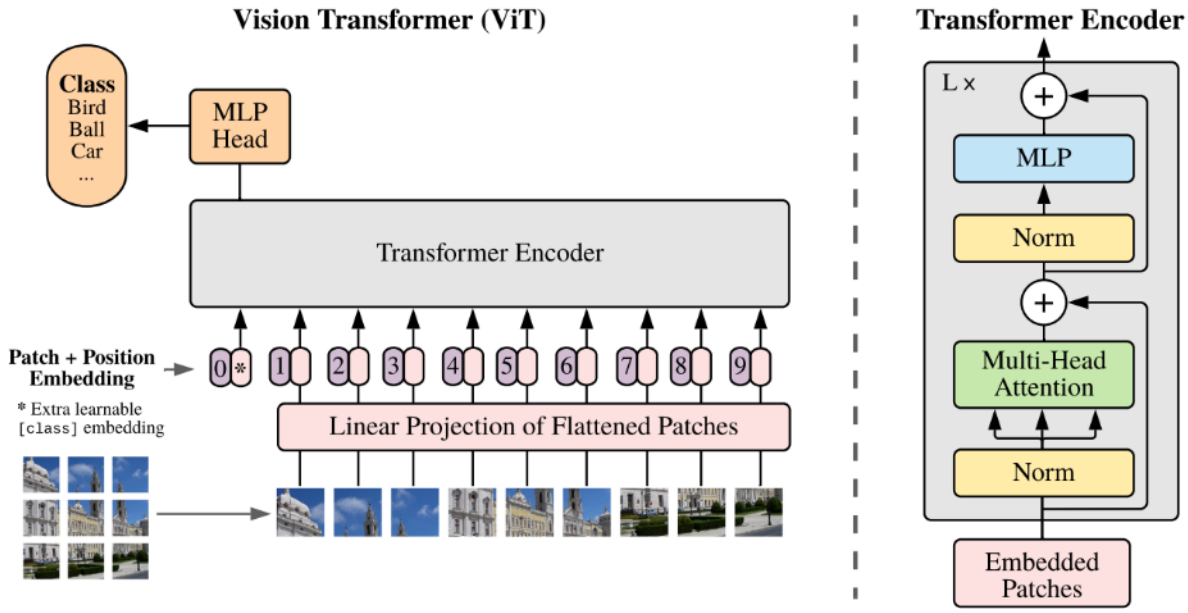


Figura 4: Arquitectura de un Vision Transformer "ViT" (Dosovitskiy y cols., 2020, 3).

1. Embedding de parches (Patch Embedding): la entrada de la imagen se divide en "parches" y cada parche se trata como un token, similar a las palabras en el Transformer original. Cada parche se convierte en un vector utilizando una capa de proyección lineal. Si la imagen de entrada es de tamaño $H \times W$ y se divide en parches de tamaño $P \times P$, entonces $N = \frac{H \times W}{P^2}$ siendo N el total de parches.

2. Embedding de posición (Position Embedding): al igual que en el Transformer original, se agrega codificación posicional a los vectores de los parches para que el modelo pueda entender la posición en la secuencia de estos parches. Esto se realiza mediante la suma de vectores de posición aprendidos a los vectores de parches.

3. Codificador del Transformer: para este modelo, las sub capas del codificador son las siguientes:

   a) Atención multi-cabeza: esta capa calcula las relaciones entre todos los pares de embedding de los parches. La atención se calcula igual que en el Transformer original con la

ecuación 1. Después, de igual manera, se aplica una función softmax para obtener los pesos de atención, y finalmente se calcula la suma ponderada de los valores, donde los pesos provienen de la atención.

b) Red neuronal: después de la capa de atención, se aplica una red neuronal a cada vector de parche de forma independiente.

c) Conexiones Residuales: al igual que en el Transformer original, se agregan conexiones residuales al final de la capa de atención y de la red neuronal para evitar el desvanecimiento de gradiente.

4. Capa de salida (Output Layer): la salida del último bloque del codificador se pasa a una capa densa (Fully Connected) seguida de una función de activación softmax para clasificación o regresión, dependiendo de la tarea.

La adaptación de los Transformers a tareas de visión por computadora ha demostrado ser efectiva en gran medida debido a la capacidad de los mecanismos de atención para capturar relaciones de largo alcance (capacidad del mecanismo de atención para identificar y analizar dependencias o conexiones entre elementos distantes dentro de una imagen.) y contextos en las imágenes, lo que puede ser especialmente útil en tareas como la detección de objetos en imágenes más complejas como lo son las imágenes médicas, debido a su alta resolución, las estructuras pequeñas y complejas, el ruido, y la necesidad de comprender un contexto médico especializado.

El ViT ha demostrado un buen rendimiento en procesamiento de imágenes y ha competido con arquitecturas más tradicionales como las redes neuronales convolucionales (Convolutional Neural Network o CNN), superando sus resultados en un 8 % global.

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers o “BERT” (Devlin y cols., 2018) es una arquitectura basada en Transformers y su principal innovación es que utiliza una representación bidireccional profunda, para capturar tanto el contexto izquierdo como el derecho en todas las capas, mejorando los resultados en tareas de procesamiento de lenguaje natural.

La arquitectura de BERT se muestra en la Figura 5. A diferencia del Transformer original, este utiliza solo el bloque del codificador. El codificador del Transformer utiliza varias capas de atención (Self-attention) y redes neuronales.

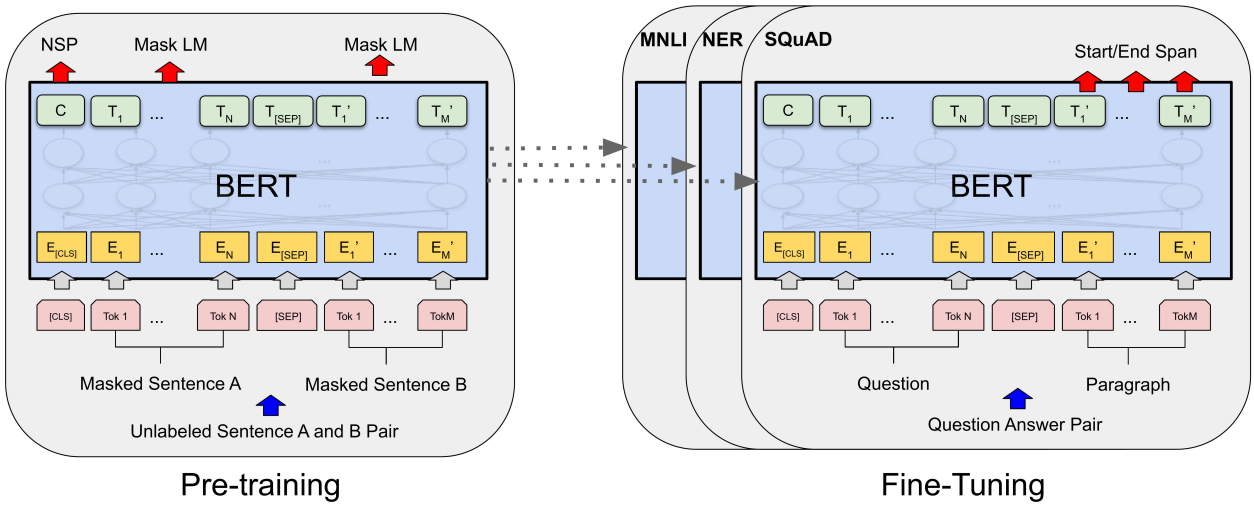


Figura 5: Arquitectura de "BERT" (Devlin, Chang, Lee, y Toutanova, 2018, 3).

Cada capa consiste en:

1. Atención multi-cabeza: mecanismo de atención que permite que cada palabra en una oración preste atención a todas las demás palabras de la misma oración. La atención se calcula con la ecuación 1 y la atención multi-cabeza con la ecuación 2.

2. Red neuronal: redes neuronales totalmente conectadas aplicadas de manera independiente a cada posición. Se calculan con la ecuación 3.

3. Conexiones residuales y normalización por capas (LayerNorm). se calcula con la ecuación 4

Lo que hace diferente a BERT es que está diseñado para ser bidireccional. Esto significa que, en lugar de procesar el texto de izquierda a derecha, BERT toma en cuenta el contexto completo de cada palabra, considerando tanto lo que está antes como lo que está después, en todas las capas.

1. Pre entrenamiento de BERT

   BERT se pre entrena en dos tareas:

a) Modelado de Lenguaje Enmascarado (MLM): en lugar de predecir la siguiente palabra en una secuencia, como hacen los modelos autorregresivos, se enmascaran algunas palabras aleatoriamente (15 %) en la entrada y trata de predecir esas palabras basándose en el contexto. Formalmente, para una secuencia $x_1, x_2, ..., x_n$, se seleccionan palabras al azar y se reemplazan por un token especial [MASK]. MLM se puede calcular con la siguiente ecuación:

$$L_{MLM} = \sum_{m \in M} \log P\left(x_m | x_{\setminus m}; \theta\right) \tag{5}$$

Donde:

- M es el conjunto de palabras enmascaradas.
- $P\left(x_m | x_{\setminus m}; \theta\right)$ es la probabilidad de predecir el token x_m en la posición enmascarada m dado el contexto de los tokens no enmascarados $x_{\setminus m}$ y los parámetros del modelo θ.
- θ representa los parámetros aprendidos por el modelo.
- log es el logaritmo natural, se utiliza porque maximizar la probabilidad logarítmica es numéricamente más estable.

b) Predicción de la siguiente oración (Next Sentence Prediction o NSP): para modelar relaciones entre oraciones, se entrena en una tarea de predicción de la siguiente oración. Se le dan al modelo pares de oraciones (A y B), y el modelo debe predecir si la segunda oración sigue a la primera en el texto original o si es una oración aleatoria.

La pérdida se calcula con una función de clasificación binaria, donde:

$$P\left(label = 1 | A, B\right) \tag{6}$$

Indica si la oración B sigue a la oración A.

BERT predice los tokens enmascarados maximizando la probabilidad logarítmica de las palabras enmascaradas basándose en el contexto de las demás palabras en la secuencia.

Su enfoque bidireccional de pre entrenamiento fue una innovación clave respecto a modelos anteriores, como GPT (Radford y cols., 2018) o ELMo (Embeddings from Language Models) (Peters y cols., 2018), que consideraban el contexto de manera unidireccional.

Medical Vision Language Learner

Medical Vision Language Learner o "MedViLL" (Moon, Lee, Shin, Kim, y Choi, 2022) es una arquitectura basada en Transformers. La arquitectura de MedViLL se muestra en la Figura 6 y se compone de dos módulos principales:

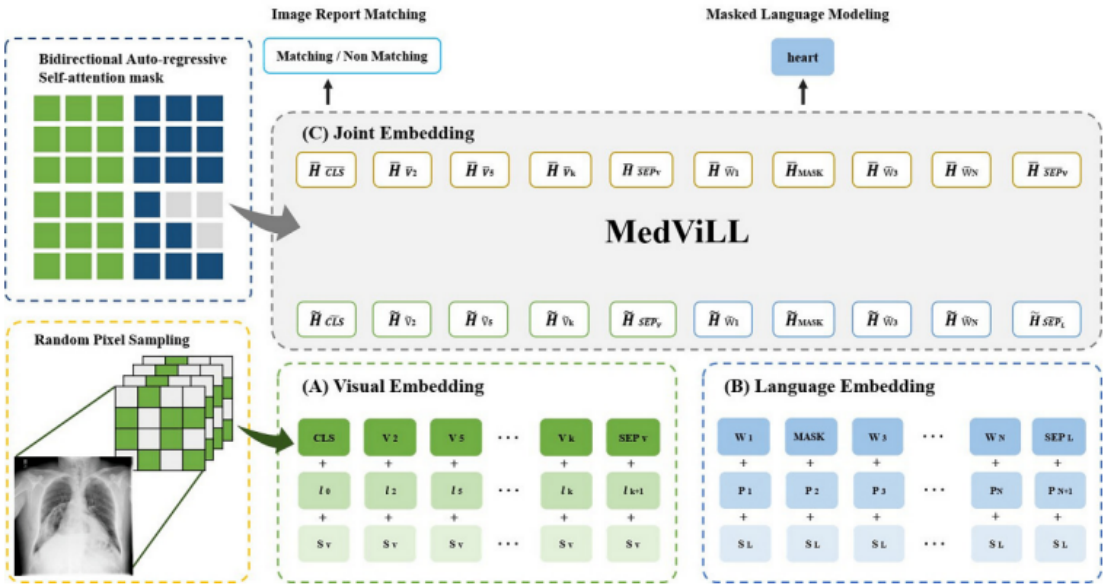


Figura 6: Arquitectura de "MedViLL" (Moon, Lee, Shin, Kim, y Choi, 2022, 5).

1. Codificador visual: utiliza una red convolucional ResNet-50 para extraer características de las imágenes médicas. Las características visuales (v) extraídas se representan como un conjunto de vectores de características $v = \{v_1, v_2, \ldots, v_K\}$, donde K es el número de regiones de la imagen.

2. Codificador textual: utiliza un BERT para procesar y codificar textos médicos. Cada palabra (w) del texto se representa como un vector de embeddings $w = \{w_1, w_2, \ldots, w_N\}$, donde N es el número de palabras en el texto.

Ambos codificadores están integrados en conjunto para capturar las características de las imágenes y el texto. Las representaciones de imágenes y texto se integran mediante un Transformer multimodal que aprende las interacciones entre ambas modalidades.

Las características de imagen y texto se concatenan y pasan a través de las capas del Transformer con la siguiente ecuación:

$$H = \{\overline{CLS}, \bar{v}_2, \ldots, \bar{v}_K, \overline{SEP}_V, \bar{w}_1, \ldots, \bar{w}_N, \overline{SEP}_L\} \tag{7}$$

Donde $\overline{CLS}$ y $\overline{SEP}$ son tokens especiales para la concatenación de las características visuales y de texto, mientras que H representa la salida del Transformer multimodal. El token $\overline{CLS}$ se coloca al inicio de la secuencia y se utiliza para agregar la representación global de toda la entrada, mientras que el token $\overline{SEP}$ se usa para separar diferentes segmentos o partes de la entrada.

1. Pre entrenamiento y ajuste fino. El proceso de pre entrenamiento se realiza con dos objetivos principales:

   a) Modelado de lenguaje enmascarado (MLM): se enmascaran palabras en el texto y se entrena el modelo para predecir las palabras enmascaradas basándose en el contexto restante y la imagen asociada.

   La pérdida del MLM se calcula como:

   $$L_{MLM}(\theta) = -\mathbb{E}_{(v,w)\approx \mathrm{D}}[\log P_\theta(w_m | v, w_{\setminus m})] \tag{8}$$

   donde θ son los parámetros de entrenamiento del modelo.

   Se toma una muestra de un par de imágenes y texto (v, w) del conjunto de entrenamiento D, donde w se puede dividir en los tokens enmascarados w_m y sus complementos $w_{\setminus m}$. $\mathbb{E}_{(v,w)\approx D}$ es el promedio del conjunto de entrenamiento D, y $P_\theta(w_m | v, w_{\setminus m})$ es la probabilidad de w_m dados v y $w_{\setminus m}$.

b) Coincidencia de imagen-texto (Image-Text Matching o ITM): evalúa la relación entre las imágenes y los textos asociados, entrenando al modelo para identificar pares correctos e incorrectos.

La pérdida del ITM se calcula como:

$$L_{ITM}(\theta) = -\mathbb{E}_{(v,w)\approx \mathrm{D}}[y \log P_\theta(v,w)] - \mathbb{E}_{(v,w')\approx \mathrm{D}}[(1-y)\log(1-P_\theta(v,w'))] \quad (9)$$

donde (v,w) representan un par de imagen-texto que sí coincide, mientras que (v,w') uno que no coincide. y es una variable binaria que indica si el par imagen-texto es correcto (1) o incorrecto (0). $\mathbb{E}_{(v,w)\approx D}$ es el promedio para el conjunto de entrenamiento D y $P_\theta(v,w)$ es la probabilidad de que (v,w) esté emparejado.

La función de pérdida total para el pre entrenamiento combina las pérdidas del ITM y MLM:

$$L = L_{MLM} + L_{ITM} \quad (10)$$

Después del pre entrenamiento, el modelo se ajusta finamente para tareas específicas, como la generación de reportes médicos y la respuesta a preguntas médicas.

El modelo MedViLL demuestra una mejora significativa en varias tareas médicas en comparación con los métodos existentes. Esto incluye una mejor precisión en la generación de reportes médicos y una mayor precisión en la respuesta a preguntas médicas, validando la efectividad del enfoque multimodal propuesto por los autores.

M2I2 (Masked language modeling, Masked image modeling, Image text matching and Image text contrastive learning)

Masked language modeling, Masked image modeling, Image text matching and Image text contrastive learning o "M2I2" (Li, Liu, Tan, y cols., 2023) es una arquitectura basada en Transformers

en donde se propone un enfoque que combina varias técnicas de pre entrenamiento para modelos multimodales.

La arquitectura de M2I2 se muestra en la Figura 7 en donde se puede observar que se basa en codificadores visuales y textuales, seguido de un mecanismo de combinación que une ambas modalidades para tareas conjuntas. La combinación de los enfoques de pre entrenamiento permite una representación robusta de ambos tipos de datos, como se describe a continuación:

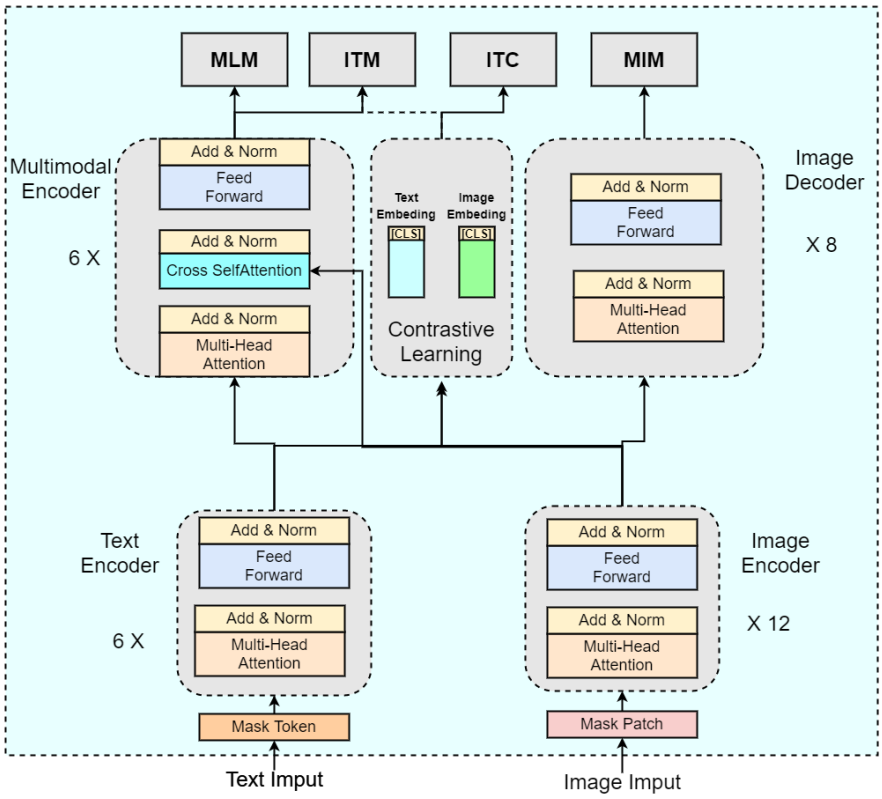


Figura 7: Arquitectura de "M2I2" Li, Liu, Tan, Liao, y Zhong, 2023, 2).

1. Codificador textual:

   - Utiliza un BERT para procesar el texto.
   - Durante el pre entrenamiento, ciertos tokens del texto se enmascaran y el modelo debe predecirlos. Esto fuerza al modelo a aprender una representación contextual de las palabras.

2. Codificador visual:

   - Utiliza un modelo ViT para procesar la imagen.
   - En el Modelado de imagen enmascarada (Masked Image Modeling o MIM), se enmascaran regiones de la imagen y el modelo debe reconstruir estas regiones.

3. Coincidencia de imagen-texto (Image Text Matching o ITM):

   - El modelo debe aprender a identificar si una imagen y un texto son compatibles o no.
   - Se utilizan pares de imágenes y texto, algunos de los cuales están emparejados correctamente (positivos), y otros están emparejados incorrectamente (negativos).

     ITM se calcula con la siguiente ecuación:

$$L_{ITM} = E_{(V,T)D} H\left(y_{itm}, p_{itm}\left(V,T\right)\right) \tag{11}$$

   Donde H es el cálculo del cross-entropy, y_{itm} es el ground truth de las etiquetas de la clase y p_{itm} es la clase predicha.

4. Aprendizaje contrastivo de imagen-texto (Image Text Contrastive Learning o ITC):

   - El aprendizaje contrastivo busca alinear las representaciones de imágenes y texto en el espacio de características, de modo que las imágenes y texto correspondientes estén cerca entre sí, mientras que los no correspondientes estén lejos.

Pre entrenamiento y ajuste fino.

1. Modelado de lenguaje enmascarado (MLM). Durante el pre entrenamiento, se ocultan ciertos token en la entrada de texto, y el objetivo del modelo es predecirlas correctamente basándose en el contexto restante. Esto permite que el modelo aprenda representaciones significativas del texto. Se calcula como:

$$L_{mlm} = E_{\left(V,\hat{T}\right)D} H\left(y_{mlm}, p_{mlm}\left(V,\hat{T}\right)\right) \tag{12}$$

Donde $\hat{T}$ es el token textual enmascarado, $p_{mlm}\left(V,\hat{T}\right)$ es la predicción del modelo y p_{mlm} es el ground truth del token textual enmascarado.

2. Modelado de imagen enmascarada (MIM). En lugar de enmascarar palabras, se enmascaran parches de una imagen. El modelo debe reconstruir las regiones faltantes de la imagen enmascarada, lo que obliga al modelo a aprender representaciones significativas de las imágenes. Se calcula con la siguiente ecuación:

$$L_{mim} = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - \hat{P}_l\right)^2 \tag{13}$$

Donde P_i es la imagen enmascarada, $\hat{P}_l$ es la imagen reconstruida por el modelo y n es el número de parches de imagen enmascarados.

3. Coincidencia de imagen-texto (ITM). Se utiliza para entrenar al modelo a identificar si una imagen y un texto están correctamente emparejados. Esto es importante para la generación de texto a partir de imágenes, donde el modelo debe comprender si el texto describe correctamente la imagen.

   La pérdida del ITM es una pérdida de clasificación binaria que puede formularse como:

$$L_\theta = -\sum_{j=1}^{|y|}\log P_\theta\left(y_j|y<j,x,c\right) \tag{14}$$

Donde θ son los parámetros del modelo, "y" es la salida del decodificador, x es el ground truth de las respuestas y c es la característica del contexto fusionado del codificador multimodal.

4. Aprendizaje contrastivo de imagen-texto (ITC). La idea es que las representaciones de una imagen y su descripción textual deberían estar muy cerca en el espacio de características, mientras que las representaciones de imágenes y textos no relacionados deberían estar lejos.

   La pérdida del ITC se calcula:

$$L = L_{mim} + L_{mlm} + L_{itm} + L_{itc} \tag{15}$$

La combinación de MLM, MIM, ITM e ITC durante el pre entrenamiento mejora significativamente el rendimiento en tareas de VQA médico.

Al utilizar múltiples tareas de pre entrenamiento, el modelo aprende representaciones más significativas y generalizables, lo que reduce la necesidad de grandes volúmenes de datos específicos para cada tarea.

Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses for Medical Visual Question Answering

MUMC (Li, Liu, He, Zhao, y Zhong, 2023) es una arquitectura basada en Transformers que busca mejorar el rendimiento de VQA en el área médica. Los autores proponen un enfoque de pre entrenamiento que aprovecha tanto pérdidas contrastivas unimodales como multimodales para mejorar las representaciones del modelo. El uso de estas pérdidas contrastivas permite al modelo aprender representaciones alineadas entre las modalidades de imagen y texto, lo que mejora su capacidad de respuesta a las preguntas. La arquitectura de MUMC se muestra en la Figura 8 en donde se puede observar que consta de tres componentes principales:

1. Codificadores: un codificador visual para imágenes y un codificador de lenguaje para texto.

   Codificador visual. Se utiliza un ViT pre entrenado para extraer características de la imagen. La salida de este codificador es un conjunto de características visuales $v \in \mathbb{R}^{n_v x d_v}$, donde n_v es el número de tokens visuales, y d_v es la dimensión de las características visuales.

   Codificador textual. Se emplea un BERT pre entrenado para procesar la pregunta en lenguaje natural. La salida es una representación de características textuales $t \in \mathbb{R}^{n_t x d_t}$, donde n_t es el número de tokens textuales, y d_t es la dimensión de las características textuales.

2. Combinación multimodal: combina las representaciones de ambos codificadores.

   Se implementa con un mecanismo de atención (ecuación 1), el cual permite que las características textuales se ajusten en función de las características visuales. Esta atención permite que la representación del texto se unifique con la representación visual, y viceversa, lo que resulta en una comprensión más profunda entre el texto y la imagen.

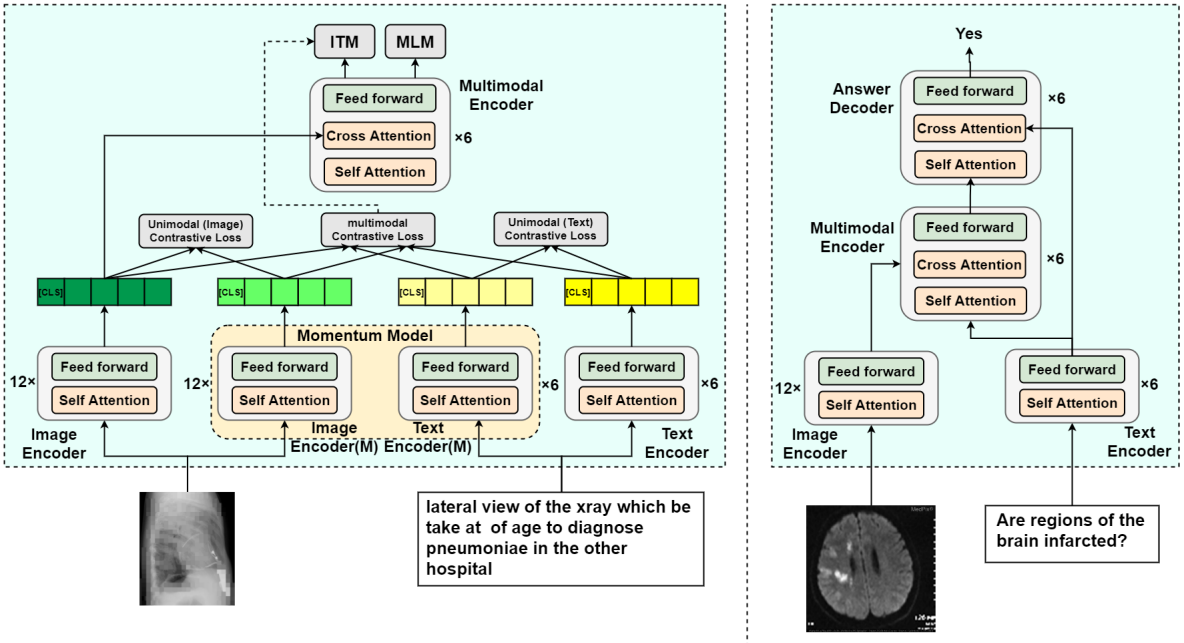


Figura 8: Arquitectura de "MUMC" (Li, Liu, He, Zhao, y Zhong, 2023, 3).

3. Decodificador: genera respuestas a partir de las representaciones combinadas.

   Una vez obtenida la representación visual y textual, se utiliza un decodificador (BERT) para generar la respuesta a la pregunta.

El pre entrenamiento del modelo utiliza dos tipos de pérdidas contrastivas:

1. Pérdida contrastiva unimodal: esta pérdida se aplica al texto e imagen de manera independiente para aprender representaciones contrastivas dentro de cada modalidad.

2. Pérdida visual: se seleccionan pares de imágenes (positivo y negativo). La pérdida está diseñada para minimizar la distancia (o maximizar la similitud) entre las representaciones de imágenes similares (positivas) y maximizar la distancia entre imágenes diferentes (negativas).

3. Pérdida textual: de manera similar, se seleccionan pares positivos y negativos de textos, y la pérdida provoca representaciones similares para textos similares y diferentes para textos diferentes. Las ecuaciones de las pérdidas unimodal y multimodal son las siguientes:

$$L_{ucl} = \frac{1}{2}\mathbb{E}_{(V,T)D}\left[H\left(y_{i2i}(V), \frac{\exp\left(s\left(V,V_i\right)/t\right)}{\Sigma_{n=1}^{N}\exp\left(s\left(V,V_i\right)/t\right)}\right)\right] + \left[H\left(y_{t2t}(T), \frac{\exp\left(s\left(T,T_i\right)/t\right)}{\Sigma_{n=1}^{N}\exp\left(s\left(T,T_i\right)/t\right)}\right)\right] \tag{16}$$

$$L_{mcl} = \frac{1}{2}\mathbb{E}_{(V,T)D}\left[H\left(y_{i2t}(V), \frac{\exp\left(s\left(V,T_i\right)/t\right)}{\Sigma_{n=1}^{N}\exp\left(s\left(V,T_i\right)/t\right)}\right)\right] + \left[H\left(y_{t2i}(T), \frac{\exp\left(s\left(T,V_i\right)/t\right)}{\Sigma_{n=1}^{N}\exp\left(s\left(T,V_i\right)/t\right)}\right)\right] \tag{17}$$

Donde:

a) s es la similitud de coseno.

b) $s(V,V_i) = g_v(v_{cls})^T g_v(v_{cls})_i$, es la representación de las características visuales.
$s(T,T_i) = g_t(t_{cls})^T g_t(t_{cls})_i$, es la representación de las características textuales.

c) $s(V,T_i) = g_v(v_{cls})^T g_t(t_{cls})_i$, es la representación combinada de las características visuales y textuales.

d) $s(T,V_i) = g_t(t_{cls})^T g_v(v_{cls})_i$, es un par positivo (texto e imagen correspondientes).

La pérdida total es una combinación ponderada de las pérdidas unimodal y multimodal y se calcula con la siguiente ecuación:

$$L_{itm} = \mathbb{E}_{(V,T)D} H\left(y_{itm}, p_{itm}\left(V,T\right)\right) \tag{18}$$

La función H representa el cálculo del cross-entropy, donde y_{itm} denota el ground truth de las etiquetas y $p_{itm}(V,T)$ es la función para predecir las clases.

MUMC presenta un enfoque para mejorar el VQA médico mediante el uso de técnicas de pre entrenamiento contrastivo. La combinación de pérdidas unimodales y multimodales permite al modelo aprender representaciones contrastivas para imágenes médicas y lenguaje, mejorando así la precisión y relevancia de las respuestas generadas.

Metodología

En este capítulo se presentan los detalles de la base de datos que se utiliza en el trabajo de investigación, así como las técnicas de preprocesamiento de los datos que se realizaron para la extracción de características. Además, se describe la implementación del modelo de aprendizaje automático propuesto llamado BioMedViLL y las métricas de evaluación. Posteriormente, se detalla la implementación que se realizó sobre la base de datos VQA-RAD para medir el desempeño de los diferentes modelos del estado del arte en la metodología que se definió. En última instancia, se define el experimento de clasificación binaria que se realizó sobre las respuestas en la base de datos VQA-RAD, esto con el propósito de realizar una comparación contra lo publicado por otros autores que reportan resultados de clasificación binaria y resultados de predicción de respuestas abiertas.

Adquisición, descripción y análisis de la base de datos

VQA-RAD (Lau y cols., 2018) es un conjunto de datos compuesto principalmente por imágenes de rayos X de la cabeza y el pecho, que incluye también imágenes de tomografías computarizadas y resonancias magnéticas. El conjunto de datos incluye 3,515 pares de preguntas y respuestas generados por expertos médicos, basados en 315 imágenes con una resolución de 1024 x 1151 píxeles, en formato JPG. Cabe mencionar que la base de datos hace una partición de los conjuntos de entrenamiento y prueba por medio de una columna que determina a cuál conjunto pertenece.

Esta base de datos se descargó de su página oficial, la cual incluye:

1. Imágenes radiológicas: proporcionadas en formato JPG, correspondientes a diferentes estudios como tomografías computarizadas (CT), radiografías del tórax y resonancias magnéticas (MRI). Algunos ejemplos de estas imágenes se muestran en las Figuras 9, 10, 11, respectivamente. Las imágenes se dividen en 3 tipos de órganos: tórax (CHEST), abdomen (ABD) y cabeza (HEAD).

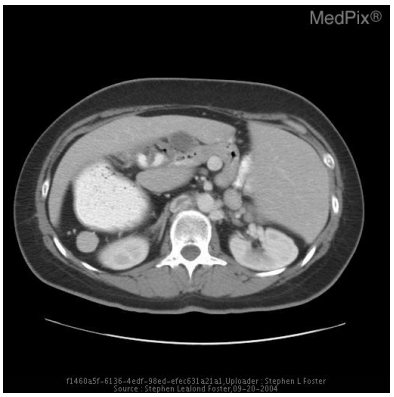


Figura 9: Tomografía del abdomen.

Figura 10: Radiografía del tórax.

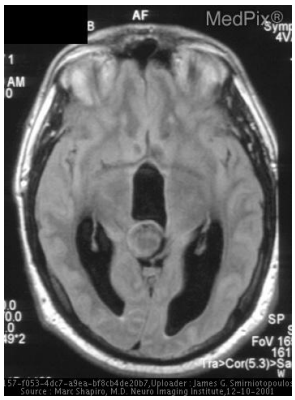


Figura 11: Resonancia de la cabeza.

2. Preguntas y respuestas: cada imagen tiene asociada una o más preguntas médicas, junto con sus respuestas generadas por expertos médicos.

   Las preguntas incluyen temas como diagnóstico de fracturas, presencia de anomalías, entre otros, como se muestra a continuación.

Tabla 2: Ejemplos de preguntas y respuestas abiertas sobre las imágenes de VQA-RAD.

Pregunta Abierta	**Respuesta Abierta**
What type of imaging does this not represent?	Ultrasound
What is not pictured in this image?	The extremities
Where is the abnormality?	Left temporal lobe
Where is the pathology in this image?	Vasculature

Tabla 3: Ejemplos de preguntas y respuestas cerradas sobre las imágenes de VQA-RAD.

Pregunta Cerrada	**Respuesta Cerrada**
Are regions of the brain infarcted?	Yes
Are the lungs normal appearing?	No
Is there evidence of a pneumothorax	No
Is the trachea midline?	Yes

Tipos de preguntas

Las preguntas se clasifican en 11 tipos diferentes, los cuales se describen en la Tabla 4.

Tabla 4: Tipos de preguntas sobre las imágenes de VQA-RAD.

Tipo	**Descripción**
Modalidad	Cómo se toma una imagen: tomografía computarizada, rayos X, resonancia magnética, etc.
Plano	Orientación de una imagen que atraviesa el cuerpo: axial, sagital, coronal.
Órgano	Categorización que conecta las estructuras anatómicas: sistema pulmonar, cardíaco y músculo-esquelético.
Anormalidad	Normalidad de una imagen u objeto. Por ejemplo: ¿Hay algún problema con la imagen?, o ¿Se ve el hígado normal?
Presencia	Los médicos pueden referirse a la presencia de condiciones en una imagen o en un paciente. Por ejemplo: fracturas, desplazamiento de la línea media, infarto.
Ubicación	Posición o ubicación de un objeto u órgano, incluido el lado de un paciente, con respecto a los bordes de la imagen o en relación con otros objetos de la imagen.
Color	Intensidad de la señal, incluyendo realce u opacidad.
Tamaño	Medición del tamaño de un objeto, por ejemplo, agrandamiento, atrofia.
Otros atributos	Otros tipos de preguntas de descripción.
Conteo	Se centra en una cantidad de objetos, por ejemplo, número de lesiones.
Otro	Categorización general para preguntas que no entran en las categorías anteriores.

Tipos de respuestas

Las respuestas se clasifican en 2 tipos diferentes, los cuales se describen en la Tabla 5 que se presenta a continuación:

Tabla 5: Tipos de respuestas sobre las imágenes de VQA-RAD.

Tipo	**Descripción**
Cerrada (Closed-ended)	"Si", "No" y otras opciones limitadas que se consideran binarias. Por ejemplo: ¿Está la masa a la izquierda o a la derecha?
Abierta (Open-ended)	No tiene una estructura de pregunta limitada y podría tener múltiples respuestas correctas.

Análisis de las imágenes, preguntas y respuestas

Para obtener una visión general del dataset, se realizaron los siguientes análisis.

1. Frecuencia de imágenes según el tipo de órgano: se analizó la distribución de las imágenes por tipo de órgano, la cual se muestra en la Figura 12, siendo el tórax el más frecuente, seguido por el abdomen, y finalmente la cabeza.

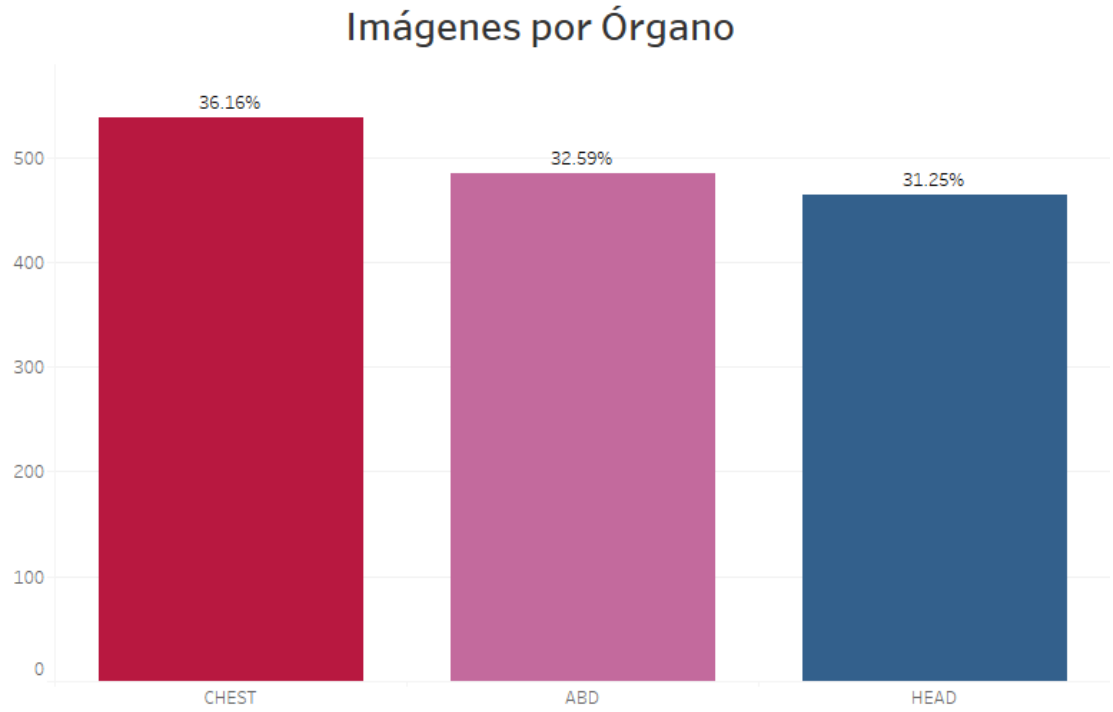


Figura 12: Distribución de las imágenes por órgano.

2. Frecuencia de preguntas: se analizó la distribución de las preguntas más comunes, la cual se muestra gráficamente en la Figura 13. Las preguntas relacionadas con presencia [PRES] es la categoría más dominante, con un 40.22 % del total de preguntas. Esto indica que gran parte de las preguntas están enfocadas en determinar la presencia o ausencia de alguna condición en las imágenes. Con un 18.03 %, la posición [POS] es la segunda categoría más frecuente. Esto significa que las preguntas sobre la ubicación o posición de ciertos bordes u objetos en las imágenes tienen un papel importante en la interpretación médica. Las preguntas sobre anormalidad [ABN], tamaño [SIZE], modalidad [MODALITY] y atributos [ATTRIB] también son relativamente comunes, lo que indica que estas características también son importantes para el diagnóstico.

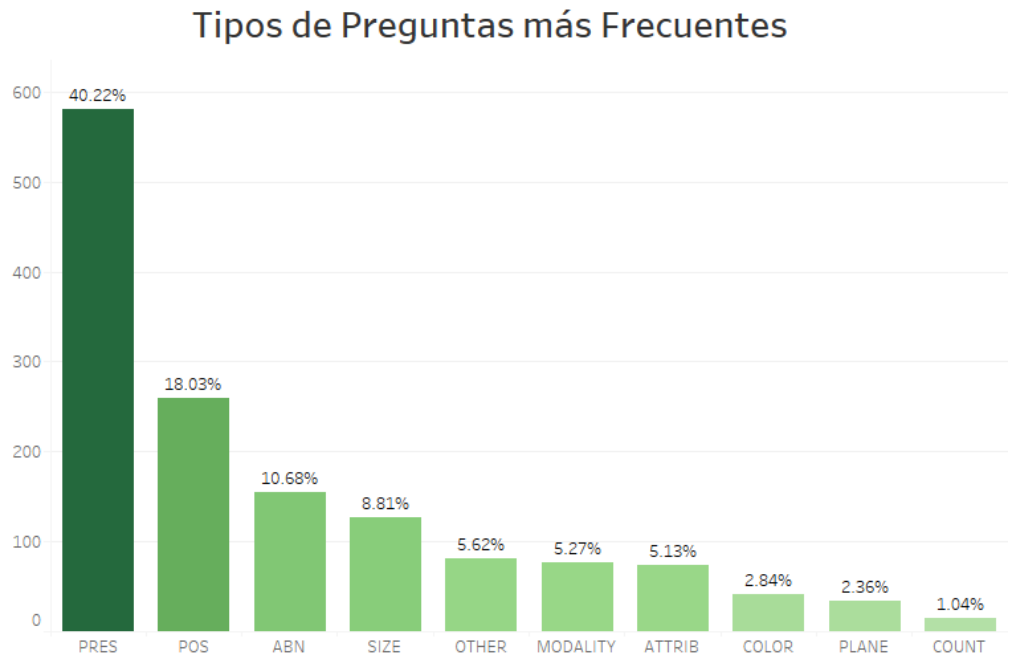


Figura 13: Preguntas más frecuentes.

3. Frecuencia de preguntas por órgano: se analizaron los tipos de preguntas más frecuentes y se clasificaron por órgano. La Figura 14 muestra el diagrama que se obtiene de este análisis. Las preguntas se clasifican en las nueve categorías mostradas en la Tabla 4: PRES (presencia), POS (posición), ABN (anormalidad), SIZE (tamaño), OTHER (otras), ATTRIB (atributo), MODALITY (modalidad), COLOR y PLANE (plano). Además, se categorizan por los tres órganos del conjunto de datos: ABD (abdomen), CHEST (tórax) y HEAD (cabeza).

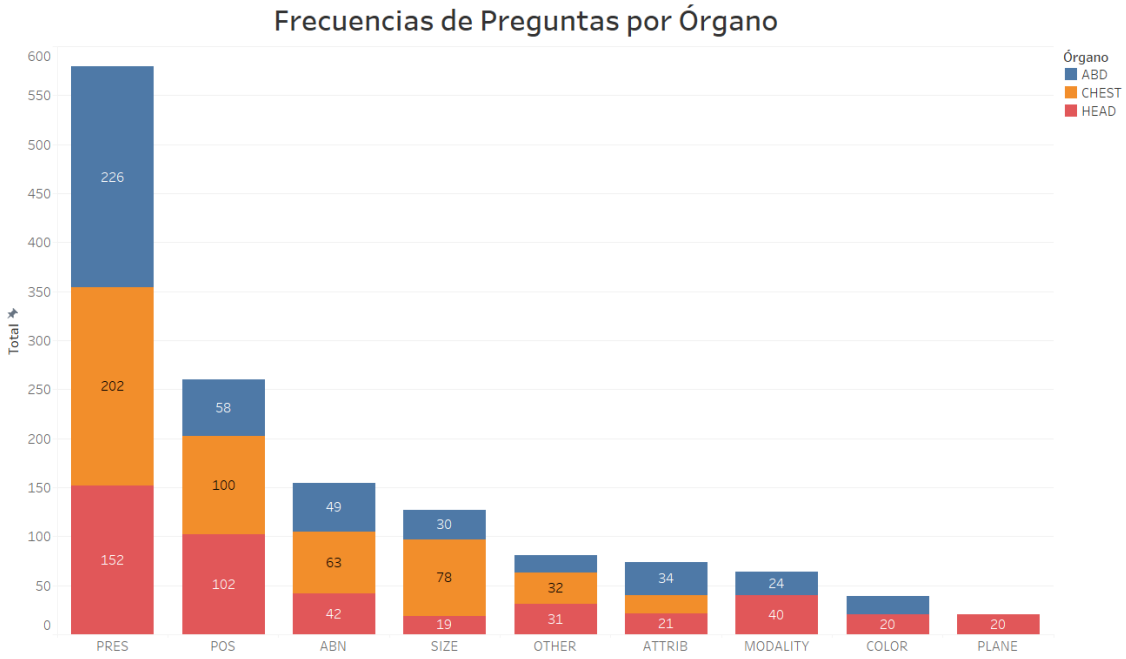


Figura 14: Frecuencia de preguntas por órgano.

El tipo de pregunta más común es el de presencia, con un total de 580 preguntas. Las preguntas sobre el abdomen (226) y el tórax (202) predominan en esta categoría, mientras que las relacionadas con la cabeza son 152.

En segundo lugar, están las preguntas relacionadas con la posición, con un total de 260 preguntas. Aquí las preguntas sobre la cabeza (102) supera levemente a las relacionadas con el tórax (100) seguida del abdomen (58).

En esta figura se observa que el mayor número de preguntas se centra en presencia [PRES] y POS (posición), particularmente relacionadas con el abdomen y el tórax. El tipo de pregunta sobre PLANE (plano) es exclusivamente para la cabeza, lo que indica que es más relevante para estudios de imágenes de esta región. Las preguntas relacionadas con el tórax [CHEST] son las más comunes en múltiples categorías (ABN, SIZE, OTHER), lo que indica que este órgano es un área de interés en la base de datos.

4. Distribución de respuestas: se calculó la distribución de respuestas abiertas y cerradas, la cual

se muestra en la Figura 15). Las respuestas cerradas resultaron ser las más frecuentes con un 56.97 %, mientras que las respuestas abiertas fueron las menos comunes con un 43.03 %.

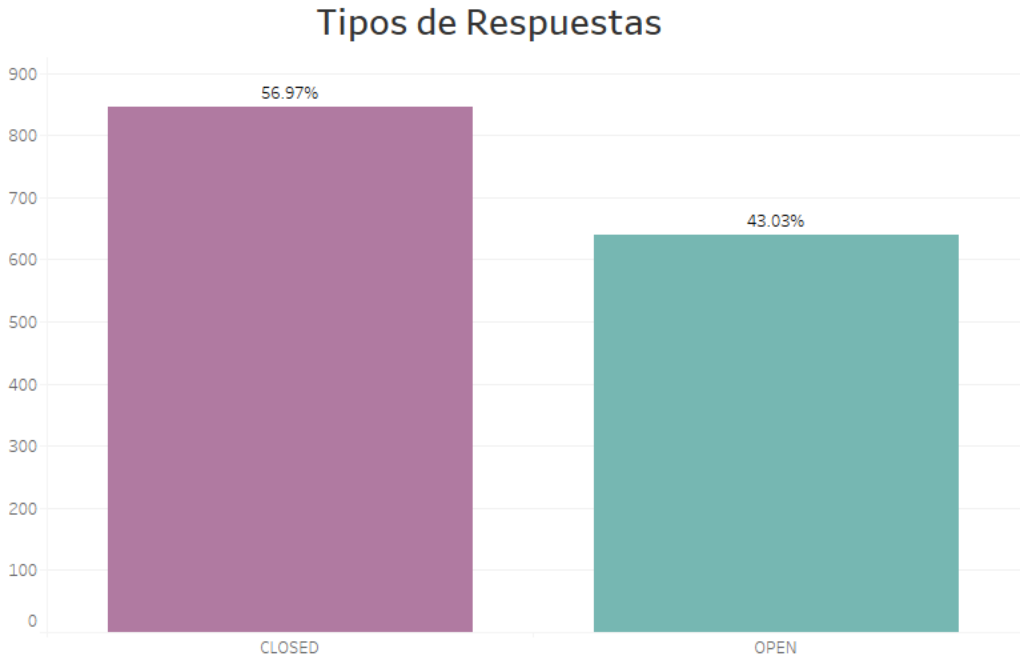


Figura 15: Distribución del tipo de respuestas.

5. Distribución de respuestas por órgano: se analizó la distribución de las respuestas y se clasificaron según el tipo de órgano. La Figura 16 muestra el resultado de este análisis. En el caso de las respuestas cerradas, el órgano con mayor frecuencia fue el tórax, mientras que la cabeza fue el de menor frecuencia. Por otra parte, para las respuestas abiertas, la cabeza presentó la mayor frecuencia, mientras que el tórax fue el menos frecuente.

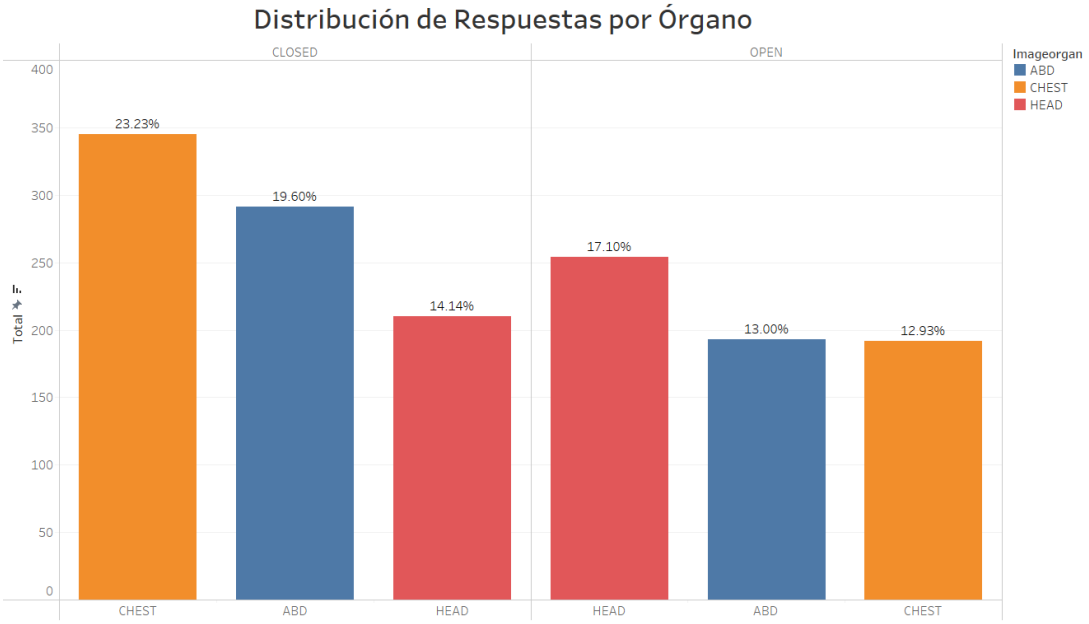


Figura 16: Distribución de respuestas por órgano.

Preprocesamiento de datos

Se cargaron las preguntas y respuestas del conjunto de datos VQA-RAD desde archivos JSON y PKL. Los archivos de datos están en formato JSON para las preguntas y PKL para las respuestas.

Las imágenes se identificaron mediante un diccionario *img_id2val* que vincula identificadores de imagen con los datos reales de las mismas.

Filtrado y clasificación de las entradas

Las preguntas y respuestas se clasifican por su identificador único *qid* para asegurar la alineación correcta entre las preguntas y las respuestas correspondientes. Además, se realizó un filtrado adicional basado en el tipo de órgano (CHEST, HEAD, ABD) si el modelo se está entrenando en una categoría específica.

Limpieza del texto

El texto de las preguntas pasó por un proceso de normalización antes de ser tokenizado. Esto incluye:

1. Convertir el texto a minúsculas.

2. Remover caracteres innecesarios como signos de puntuación (?, ´, .), el plural 's, y algunas combinaciones específicas como ? -yes/no o ? -open.

3. Reemplazar ciertas palabras para la estandarización, por ejemplo, cambiar "x ray"por "x-ray".

Tokenizacion, codificación, y enmascaramiento

La tokenizacion del texto de las preguntas fue realizada por un modelo *BERTTokenizer*, truncando la longitud de la secuencia a 253 tokens como máximo. Este proceso convierte las preguntas en secuencias de tokens que representan las palabras y sub-palabras en el vocabulario de BERT. Se agregaron tokens especiales, [CLS] al inicio de la secuencia y [SEP] al final.

Para las secuencias que son más cortas que la longitud máxima permitida (253 tokens), se utilizó padding con el token especial [PAD] para completar la longitud a 255 tokens.

Las máscaras de atención para BERT se generaron asignando 1 a los tokens válidos y 0 al padding. La segmentación no es relevante en este contexto (dado que solo se maneja una secuencia de texto), por lo que todos los tokens pertenecen al mismo segmento (marcado como 1).

Procesamiento de imágenes y etiquetas

Las imágenes fueron cargadas utilizando PIL y convertidas a escala de grises con 3 canales de colores, simulando el formato RGB. Después, fueron redimensionadas a 224×224 píxeles y se realizaron una serie de transformaciones como lo son la normalización de los canales de colores y la conversión a tensores.

Las respuestas correspondientes a las preguntas fueron convertidas en vectores de etiquetas. Las etiquetas de las respuestas se almacenan como tensores, y si la respuesta es una pregunta cerrada (por ejemplo, sí o no), se aplica una codificación binaria para representar la clase correcta.

Las respuestas abiertas o cerradas se identifican a partir del tipo de pregunta, y se asignan tipos específicos (0 para cerradas, 1 para abiertas).

Modelo BioMedViLL

BioMedViLL está diseñado para tareas multimodales que requieren procesar tanto texto como imágenes, como por ejemplo la interpretación de preguntas visuales. La arquitectura de BioMedViLL se compone de dos partes principales: un codificador textual basado en una variante de BERT pre entrenado en tareas médicas llamado *Bio_ClinicalBERT* y un codificador visual basado en un Vision Transformer de 32 capas llamado ViT-B-32, los cuales se combinan en una capa final de clasificación, como se muestra en la Figura 17.

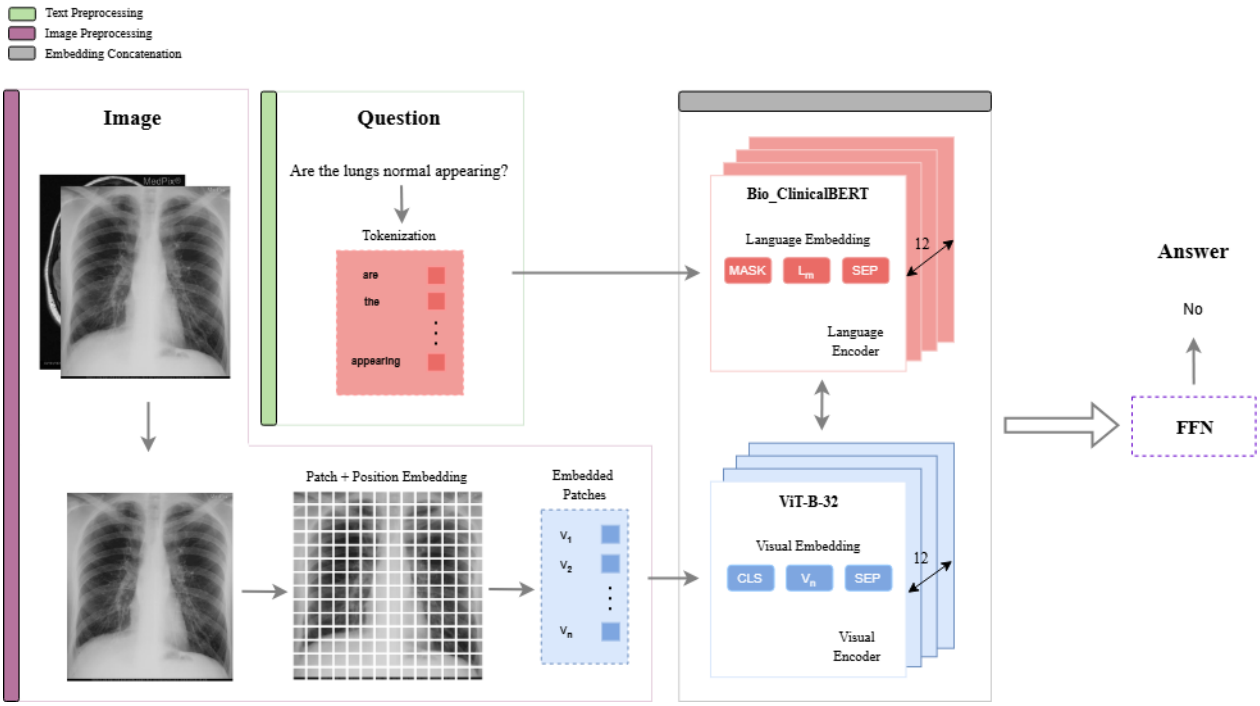


Figura 17: Arquitectura de BioMedViLL.

Las secuencias de texto se tokenizan utilizando un *BertTokenizer* correspondiente al modelo "Bio-ClinicalBERT" y se utiliza codificación de posición con el token $[p_i]$. Este proceso convierte el texto en una secuencia de tokens que el modelo *Bio_ClinicalBERT* puede procesar. En esta eta-

pa, se realizan operaciones adicionales como la truncación de secuencias largas y la creación de máscaras [MASK] de atención para marcar la relevancia de los tokens. El texto de entrada pasa por varias capas de autoatención y transformación dentro de *Bio_ClinicalBERT*, y se toma la salida correspondiente al token [CLS], que es un resumen global de la secuencia de texto. El token especial [SEP] indica el final de la secuencia de texto, que sirve para separar las diferentes partes de una entrada de texto. Estas características se usarán posteriormente para combinarse con las características visuales [V_i].

Las imágenes se transforman mediante operaciones de preprocesamiento como el redimensionado y la normalización, para adaptarlas al tamaño de entrada requerido por el ViT-B-32. Las imágenes pre procesadas se utilizan para extraer características visuales relevantes. El ViT-B-32 transforma la imagen a una secuencia de parches y aplica un mecanismo de autoatención para aprender representaciones visuales, el resultado es una representación vectorial de la imagen que posteriormente se combinará con las características textuales.

Después, las características textuales y visuales se combinan en una única representación multimodal. La concatenación se realiza en los vectores de embeddings obtenidos de los codificadores de texto e imagen y se representa de la siguiente manera:

1. Las salidas del ViT-B-32 y Bio_ClinicalBert se concatenan en un solo vector.
2. El vector combinado se pasa a una capa completamente conectada que transforma este vector multimodal en una predicción final. Esta predicción puede ser para clasificación, regresión u otra tarea según la aplicación.

Por último, una vez que las representaciones del texto e imagen se han combinado, se pasa el vector resultante por una capa completamente conectada (FFN) para obtener la salida final.

Para su entrenamiento se utilizó el optimizador *AdamW* con una tasa de aprendizaje de 0.00003 ($3e^{-5}$) para ajustar los pesos del modelo durante el entrenamiento. La función de pérdida utilizada fue *BCEWithLogitsLoss*. El entrenamiento se llevó a cabo en 50 épocas, con un tamaño de lote (batch) de 16. Este proceso se realizó en una NVIDIA RTX 4080. Durante el entrenamiento, las

predicciones del modelo fueron evaluadas tanto en preguntas de tipo cerradas (sí o no) como en preguntas abiertas.

Primero se cargaron los datos de entrada (secuencias de tokens y sus correspondientes imágenes) y las etiquetas. El modelo recibió las secuencias tokenizadas y las imágenes, pasando por ViT-B-32 para extraer las características de las imágenes, y por BERT para extraer las representaciones del texto. Ambos procesos de extracción de características se integraron para generar una predicción de las respuestas. Para el cálculo de la pérdida del modelo, se comparó la predicción del modelo con las etiquetas reales utilizando la función de pérdida *BCEWithLogitsLoss*, que toma en cuenta los logits (Valor de probabilidad) sin normalizar. Después, se calculó el gradiente de la pérdida con respecto a los pesos del modelo, y estos pesos se actualizaron con el optimizador *AdamW*. Se guardó la pérdida media del entrenamiento y la precisión de las predicciones tanto para preguntas cerradas como abiertas.

Métrica de evaluación

Para evaluar el desempeño del modelo durante el entrenamiento y prueba, se utilizó la métrica de exactitud (Accuracy). Esta métrica refleja la proporción de predicciones correctas hechas por el modelo respecto al total de ejemplos evaluados. En este caso, se calcularon tres tipos de exactitud:

- Exactitud total (Total Accuracy): evalúa la proporción de predicciones correctas sobre todo el conjunto de datos.
- Exactitud en preguntas cerradas (Closed Accuracy): mide el desempeño del modelo en preguntas cuya respuesta es binaria (sí/no).
- Exactitud en preguntas abiertas (Open Accuracy): mide el desempeño del modelo en preguntas abiertas.

1. Cálculo de la exactitud por lote (batch)

   El modelo genera un conjunto de predicciones denominadas “logits”, que representan la probabilidad de que cada clase sea la correcta para una determinada entrada. Para calcular la

exactitud, se selecciona la clase con la mayor probabilidad usando la operación de argmax (función que devuelve el número de la columna con el valor máximo de cada fila):

$$\hat{y} = argmax\left(logit\left(x\right)\right) \tag{19}$$

Donde $\hat{y}$ es la clase predicha y $logit\left(x\right)$ es la salida del modelo para una entrada x.

Después, se comparan las predicciones $\hat{y}$ con las etiquetas reales y (ground-truth), que están codificadas en formato one-hot (vectores donde la clase correcta es 1 y las demás son 0). Esto genera un valor de exactitud por lote, donde el cálculo del puntaje (Score) se realiza a través de la multiplicación de las predicciones con las etiquetas reales:

$$Score\left(y,\hat{y}\right) = \sum_{i=1}^{N}\left(\hat{y}_i = y_i\right) \tag{20}$$

Donde $\left(\hat{y}_i = y_i\right)$ es una función que toma el valor 1 si la predicción $\hat{y}$ coincide con la etiqueta y_i, y 0 en caso contrario. Esta operación se realiza para cada conjunto en el lote y se promedia para obtener la exactitud total del lote:

$$Batch\_accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i = y_i\right) \tag{21}$$

Donde N es el número total de conjuntos en el lote.

2. Precisión total

   La exactitud total se calcula como el promedio de las precisiones de todos los lotes en el conjunto de validación:

$$Total\_accuracy = \frac{1}{M}\sum_{i=1}^{M} Batch\_accuracy\left(i\right) \tag{22}$$

   Donde M es el número total de lotes en el conjunto de evaluación.

3. Exactitud para Preguntas Cerradas y Abiertas

   En estos modelos, las preguntas se dividen en dos tipos: cerradas y abiertas. Para cada tipo de pregunta, se calculan exactitudes separadas:

- Exactitud para preguntas cerradas:

$$Closed\_accuracy = \frac{\sum_{i \in Closed} (\hat{y}_i = y_i)}{\text{Total Closed Questions}} \tag{23}$$

- Exactitud para preguntas abiertas:

$$Open\_accuracy = \frac{\sum_{i \in Open} (\hat{y}_i = y_i)}{\text{Total Open Questions}} \tag{24}$$

Evaluación

Al final de cada época de entrenamiento, se realizó una evaluación sobre el conjunto de prueba. Durante este proceso, el modelo se congeló para el cálculo de gradientes, y las métricas de exactitud y pérdida se calcularon de manera similar al entrenamiento.

Por último, se calculó la exactitud para preguntas abiertas, cerradas y en general, dividiendo los aciertos entre el número total de preguntas de cada tipo.

Al finalizar cada época de entrenamiento y prueba, se guardó una versión del modelo en la carpeta de salida, lo que permitió recuperar la mejor versión del modelo basada en su rendimiento durante el entrenamiento y prueba.

Resultados cuantitativos

A continuación, se presentan los resultados de la evaluación de los modelos de VQA aplicados a la base de datos VQA-RAD.

Los modelos evaluados incluyen MedViLL, BioMedViLL (Propio), M2I2, y MUMC, cada uno con diferentes arquitecturas y enfoques de procesamiento de imagen y texto.

Se analizó el desempeño de estos modelos tanto en preguntas abiertas como cerradas, proporcionando una visión general sobre su capacidad para responder con precisión en distintos tipos de

preguntas.

A lo largo de esta sección, se detallan los resultados en la métrica de evaluación presentada anteriormente, permitiendo identificar las fortalezas y debilidades de cada modelo.

Estos resultados son importantes para determinar qué modelos son más efectivos en la resolución de preguntas y cómo sus predicciones se alinean con las respuestas correctas (ground truth) provistas en la base de datos.

En las siguientes subsecciones, se mostrarán gráficos comparativos y análisis cuantitativos que ilustran el rendimiento de cada modelo, permitiendo obtener conclusiones sobre su fiabilidad en el entorno médico.

En este apartado, se evaluó el desempeño de cuatro modelos de VQA: MedViLL, BioMedVill (ViT-B-32/BioClinicalBERT), M2I2, y MUMC, utilizando la base de datos de VQA-RAD.

El análisis se estructuró en tres categorías de pregunta-respuesta: OPEN, CLOSED, y OVERALL, donde se compararon los puntajes obtenidos por cada modelo considerando el resultado de 10 corridas por modelo.

Para visualizar la distribución de los resultados, se emplearon gráficos de caja (boxplots), que permiten observar las métricas de mediana, cuartiles, dispersión, y valores atípicos (outliers) de cada modelo en las diferentes categorías.

Desempeño en Preguntas Abiertas (OPEN)

En la categoría de preguntas abiertas (Figura 18 izquierda), se evaluaron los modelos respecto a su capacidad para responder preguntas con respuestas más detalladas, observando los siguientes resultados:

BioMedViLL fue el modelo con mejor desempeño, con una mediana cercana a 60, lo que indica

una alta consistencia en la calidad de las respuestas. Cabe resaltar que este modelo presentó un valor atípico, sugiriendo que en ciertos casos específicos, el desempeño puede variar ligeramente.

MedViLL alcanzó un segundo lugar con una mediana alrededor de 45. Este modelo mostró una menor dispersión de resultados, lo que indica un comportamiento más estable, aunque no tan preciso como BioMedViLL en este tipo de preguntas.

MUMC tuvo una mediana cercana a 41, pero con una dispersión ligeramente mayor, lo que indica variabilidad en el rendimiento para diferentes preguntas abiertas.

M2I2 obtuvo el peor rendimiento, con una mediana alrededor de 30. La baja dispersión sugiere que el modelo es consistentemente débil en la resolución de este tipo de preguntas.

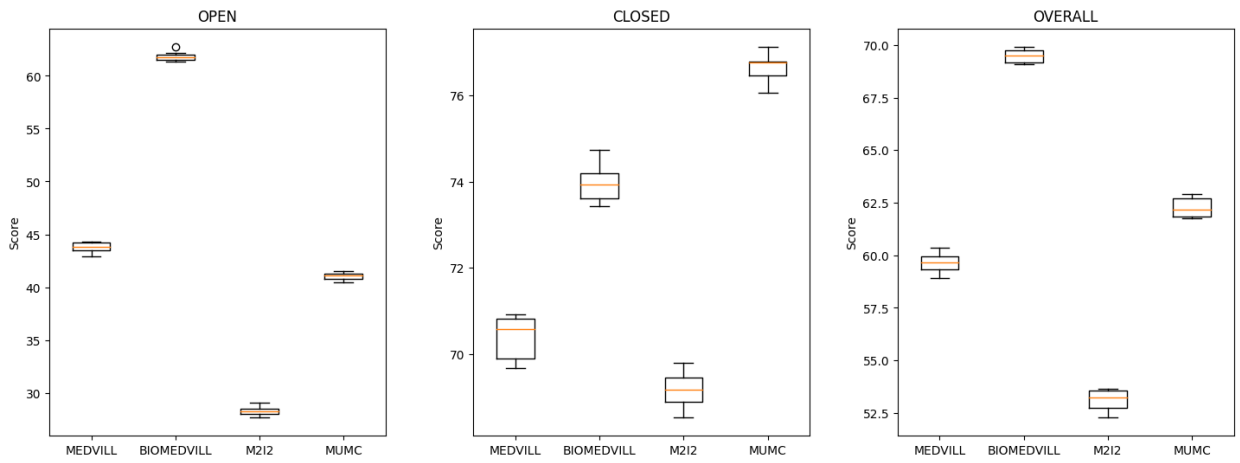


Figura 18: Análisis estadístico.

Desempeño en Preguntas Cerradas (CLOSED)

El análisis en la categoría de preguntas cerradas (Figura 18 central) evaluó la capacidad de los modelos para responder a preguntas binarias (“si” o “no”). Los resultados fueron los siguientes:

MUMC mostró el mayor rendimiento de todos los modelos, con una mediana alrededor de 77. Este resultado, junto con una baja dispersión, sugiere que el modelo es muy efectivo en la identificación

de respuestas cerradas con alta precisión.

BioMedViLL obtuvo una mediana de aproximadamente 74, lo que indica un buen desempeño, aunque ligeramente inferior al de MUMC.

MedViLL mostró un rendimiento inferior, con una mediana cercana a 70. A pesar de la menor precisión, la dispersión de los resultados sugiere un desempeño más consistente.

M2I2 nuevamente presentó el peor rendimiento, con una mediana alrededor de 68. Aunque su desempeño es bajo, la menor dispersión sugiere que el modelo tiene una precisión limitada pero estable en esta categoría.

Desempeño General (OVERALL)

El análisis global (Figura 18 derecha) integró los resultados obtenidos en las categorías anteriores para proporcionar una visión general del desempeño de los modelos en todas las preguntas:

BioMedViLL sigue mostrando el mejor rendimiento global, con una mediana cercana a 70. Esto confirma que es el mejor modelo, tanto en preguntas abiertas como cerradas.

MUMC obtuvo un buen rendimiento global, con una mediana alrededor de 62,5, lo que lo posiciona como una buena alternativa, especialmente en preguntas cerradas.

MedViLL presentó un rendimiento intermedio, con una mediana de 60, destacándose por un comportamiento equilibrado pero con un menor puntaje general.

M2I2 continuó siendo el modelo más débil, con una mediana alrededor de 52,5, lo que lo posiciona como el de menor rendimiento entre los evaluados.

Estos resultados sugieren que, dependiendo del tipo de preguntas (abiertas o cerradas), los modelos pueden variar en su capacidad de respuesta. Mientras que MUMC sobresale en preguntas

cerradas, BioMedViLL es el modelo más equilibrado, ofreciendo el mejor desempeño general, especialmente en preguntas abiertas. M2I2, por otro lado, muestra una debilidad significativa en todas las categorías, lo que indica la necesidad de ajustes o mejoras en su arquitectura para poder competir con los otros modelos en este tipo de tareas.

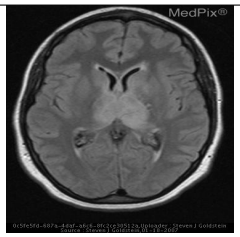
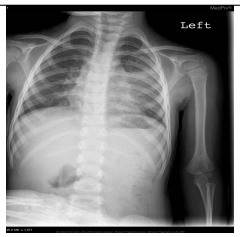
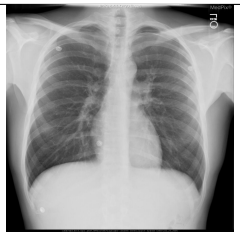
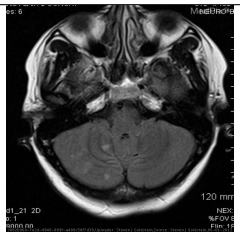
Resultados cualitativos

Al momento de comparar el desempeño de los cuatro modelos antes mencionados de manera cualitativa, se observó que el modelo MedViLL utiliza representaciones de embeddings para procesar el texto y las imágenes. Las salidas del modelo, tanto para la parte textual como visual, son vectores que no necesariamente corresponden a palabras directamente, sino a representaciones numéricas de las entradas. MedViLL no tiene una capa de decodificación diseñada para convertir las representaciones numéricas a texto. Esto significa que, si bien el modelo puede procesar la información de la imagen y el texto, no está diseñado para generar secuencias de palabras o frases en su salida. En otras palabras, MedViLL no cuenta con un decodificador que permita mapear las características aprendidas de regreso a texto para tareas de VQA.

La salida de MedViLL es una concatenación de vectores de características visuales y textuales. Esta concatenación es lo que se usa para la tarea de clasificación de las respuestas, pero no se traduce en palabras o secuencias de texto que permitan interpretar la salida de manera cualitativa. Debido a esta falta de mapeo de la salida a palabras, no es posible comparar cualitativamente estos modelos con otros que sí pueden generar texto de manera explícita, como M2I2 y MUMC. Como la propuesta de este trabajo de tesis se basó en el modelo MedViLL, este comportamiento se heredó a nuestra propuesta, por lo que no fue posible analizar de manera cualitativa el desempeño de MedViLL y BioMedViLL. Los modelos M2I2 y MUMC incluyen un decodificador, que les permite generar palabras en su salida, lo que facilita una comparación cualitativa y detallada de las respuestas.

En la Tabla 6 se presentan algunos de los resultados cualitativos de los modelos M2I2 y MUMC.

Tabla 6: Comparación de los resultados de los modelos M2I2 y MUMC.

Imagen	Pregunta	Respuesta	Predicción	
			M2I2	MUMC
	Is this a normal X-ray?	No	Yes	No
	Are the basal ganglia enlarged?	No	No	Yes
	Is there a fracture of the left humerus?	No	Yes	Yes
	What plane is demonstrated?	Coronal	Portal vein	A bullous lesion
	From what structure is the large hypodense mass emanating from?	The base of the cecum	Chest xray	Chest xray
	In which area of the brain are the infarcts found?	Right Cerebellum	Psoas Major Muscle	Fat

Conclusiones y Trabajo Futuro

Conclusiones

En este trabajo de tesis, se propone un nuevo modelo llamado BioMedViLL el cual está basado en Transformers, ofrece mejoras significativas en la precisión de las respuestas a preguntas sobre las imágenes de la base de datos VQA-RAD, en comparación con otros enfoques previos. Los resultados muestran un rendimiento competitivo en las tareas de VQA médica, especialmente en preguntas cerradas, donde se alcanza un alto grado de precisión.

Sin embargo, aunque BioMedViLL logra resultados cuantitativos sobresalientes en la interpretación de imágenes médicas, sus respuestas no pueden evaluarse de manera cualitativa en comparación con los modelos M2I2 y MUMC. Esto se debe a que BioMedViLL no emplea una arquitectura de tipo encoder-decoder, lo que limita su capacidad para generar respuestas en forma textual basado en el contexto de la imagen y la pregunta.

Por otra parte, los modelos M2I2 y MUMC sí utilizan arquitecturas encoder-decoder, lo que les permite predecir secuencias textuales con mayor precisión, basándose en una comprensión más profunda del contexto de la imagen y el texto, lo que resulta en respuestas más precisas y completas.

Los resultados refuerzan la idea de que el uso de arquitecturas avanzadas, como las empleadas en M2I2 y MUMC, junto con técnicas de pre entrenamiento como modelado de imagen enmascarada (MIM) y modelado de lenguaje enmascarado (MLM), mejoran significativamente el rendimiento en tareas de VQA médica.

En este trabajo de investigación, se destaca la importancia de seguir explorando arquitecturas de VQA más robustas y técnicas de pre entrenamiento, que permitan a BioMedViLL integrar de manera efectiva las modalidades visual y textual, de manera que se puedan obtener respuestas más precisas y completas. Esta investigación establece una base sólida para la implementación de modelos de VQA en el ámbito médico, abriendo oportunidades para mejoras y avances en este campo

tan relevante como lo es el diagnóstico médico.

Trabajo Futuro

Se pretende explorar nuevas arquitecturas de VQA médica, mejorar la calidad de los conjuntos de datos, empleando técnicas de preprocesamiento tanto en la imagen como en el texto, y aplicar técnicas avanzadas de aprendizaje y ajuste fino para obtener mejores resultados. También se pretende agregar un decodificador a la arquitectura BioMedViLL con el fin de evaluar las predicciones de manera cualitativa.

Además, es necesario validar estos modelos en entornos clínicos reales, con el fin de mejorar la precisión de los diagnósticos y la toma de decisiones en el ámbito de la salud.

Referencias

Bao, H., Dong, L., Piao, S., y Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Ben Abacha, A., Hasan, S. A., Datla, V. V., Demner-Fushman, D., y Müller, H. (2019). Vqa-med: overview of the medical visual question answering task at imageclef 2019. En *Proceedings of clef (conference and labs of the evaluation forum) 2019 working notes.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., y Dhariwal, P. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chen, T., Kornblith, S., Norouzi, M., y Hinton, G. (2020). A simple framework for contrastive learning of visual representations. En *International conference on machine learning* (pp. 1597–1607).

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eslami, S., de Melo, G., y Meinel, C. (2021). Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*.

Faghri, F., Fleet, D. J., Kiros, J. R., y Fidler, S. (2017). Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Finn, C., Abbeel, P., y Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. En *International conference on machine learning* (pp. 1126–1135).

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., y Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., y Girshick, R. (2022). Masked autoencoders are scalable vision learners. En *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16000–16009).

He, K., Zhang, X., Ren, S., y Sun, J. (2016). Deep residual learning for image recognition. En

Proceedings of the ieee conference on computer vision and pattern recognition (pp. 770–778).

He, R., Ravula, A., Kanagal, B., y Ainslie, J. (2020). Realformer: Transformer likes residual attention. *arXiv preprint arXiv:2012.11747*.

He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., y Xie, P. (2021). Towards visual question answering on pathology images. En *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (Vol. 2).

Hochreiter, S., y Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jahne, B., y HauBecker, H. (2000). *Computer vision and applications: a guide for students and practitioners*. Elsevier.

Lau, J. J., Gayen, S., Ben Abacha, A., y Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, *5*(1), 1–10.

Li, P., Liu, G., He, J., Zhao, Z., y Zhong, S. (2023). Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. En *International conference on medical image computing and computer-assisted intervention* (pp. 374–383).

Li, P., Liu, G., Tan, L., Liao, J., y Zhong, S. (2023). Self-supervised vision-language pretraining for medial visual question answering. En *2023 ieee 20th international symposium on biomedical imaging (isbi)* (pp. 1–5).

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., y Wu, X.-M. (2021). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. En *2021 ieee 18th international symposium on biomedical imaging (isbi)* (pp. 1650–1654).

Liu, G., He, J., Li, P., Zhao, Z., y Zhong, S. (2024). Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. *arXiv preprint arXiv:2401.02797*.

Moon, J. H., Lee, H., Shin, W., Kim, Y.-H., y Choi, E. (2022). Multi-modal understanding and

generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, *26*(12), 6070–6080.

Otter, D. W., Medina, J. R., y Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(2), 604–624.

Papineni, K., Roukos, S., Ward, T., y Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Pelka, O., Koitka, S., Rückert, J., Nensa, F., y Friedrich, C. M. (2018). Radiology objects in context (roco): a multimodal image dataset. En *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis: 7th joint international workshop, cvii-stent 2018 and third international workshop, labels 2018, held in conjunction with miccai 2018, granada, spain, september 16, 2018, proceedings 3* (pp. 180–189).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., y Zettlemoyer, L. (2018). Deep contextualized word representations. arxiv: 180205365. *arXiv*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. En *International conference on machine learning* (pp. 8748–8763).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., y cols. (2018). Improving language understanding by generative pre-training.

Silva, J. D., Martins, B., y Magalhães, J. (2023). Contrastive training of a multimodal encoder for medical visual question answering. *Intelligent Systems with Applications*, *18*, 200221.

Tan, M., y Le, Q. (2021). Efficientnetv2: Smaller models and faster training. En *International conference on machine learning* (pp. 10096–10106).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Yang, Z., He, X., Gao, J., Deng, L., y Smola, A. (2016). Stacked attention networks for image question answering. En *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 21–29).

Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., y Xie, W. (2023). Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Curriculum Vitae

Grados obtenidos:

- Licenciatura en Ingeniería en Ciencias de la Computación - Universidad Autónoma de Chihuahua (agosto 2017 - junio 2022).

Experiencia laboral:

- Ingeniero de Soporte de Aplicaciones - TATA Consultancy Services (junio 2022 - marzo 2023).

Estudios:

- Maestría en Ingeniería en Computación - Universidad Autónoma de Chihuahua (enero 2023 - diciembre 2024).
- Diplomado en Seguridad de la Información - Universidad Autónoma de Chihuahua (septiembre 2023 - febrero 2024).

Correo Electrónico: albertopayanm@gmail.com

Esta tesis fue mecanografiada por el Ing. José Alberto Payán Marta.