UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

FACULTAD DE INGENIERÍA

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

GENERACIÓN AUTOMÁTICA DE REPORTES MÉDICOS A PARTIR DE IMÁGENES DE RAYOS X BASADO EN ESTRATEGIAS DE APRENDIZAJE PROFUNDO

POR:

OLANDA PRIETO ORDAZ

DISERTACIÓN PRESENTADA COMO REQUISITO PARA OBTENER EL GRADO DE

DOCTOR EN INGENIERÍA

CHIHUAHUA, CHIH., MÉXICO. JUNIO 2024

GENERACIÓN AUTOMÁTICA DE REPORTES MÉDICOS A PARTIR DE IMÁGENES DE RAYOS X BASADO EN ESTRATEGIAS DE APRENDIZAJE PROFUNDO.
Disertación presentada por OLANDA PRIETO ORDAZ como requisito parcial para obtener el grado de Doctor en Ingeniería, ha sido aprobada y aceptada por:

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

Dr. Fernando Martínez Reyes
Secretario de Investigación y Posgrado

Dra. María Isabel Flores Zamora
Coordinador Académico

Dra. Graciela María de Jesús Ramírez Alonso
Director de Tesis

Dr. Manuel Montes y Gómez
Co Director de Tesis
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

JUNIO 2024

Comité:

Dra. Graciela María de Jesús Ramírez Alonso
Dr. Manuel Montes y Gómez
Dr. Luis Carlos González Gurrola
Dr. Raymundo Cornejo García
Dr. Alain Manzo Martínez

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

05 de junio de 2024.

M.A. OLANDA PRIETO ORDAZ
Presente. -

En atención a su solicitud relativa al trabajo de tesis para obtener el grado de Doctor en Ingeniería, 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 **"Generación automática de reportes médicos a partir de imágenes de rayos x basado en estrategias de aprendizaje profundo"**.

Índice de Contenido

FACULTAD DE INGENIERÍA
Circuito No. 1, Campus Universitario II
Tel. (614) 442-9500
Chihuahua, Chih., México
www.uach.mx/fing

UNIVERSIDAD AUTÓNOMA DE
CHIHUAHUA

7. **Conclusiones**
   7.1. Trabajo futuro

Referencias

Anexos

A: Clasificación imágenes médicas

B: Publicaciones y Congresos

A T E N T A M E N T E

"naturam subiecit aliis"

EL DIRECTOR

M.I. FABIÁN VINICIO HERNÁNDEZ

FACULTAD DE INGENIERÍA
U.A.CH.

SECRETARIO DE INVESTIGACIÓN Y POSGRADO

DR. FERNANDO MARTÍNEZ REYES MARTÍNEZ

DIRECCIÓN

FACULTAD DE INGENIERÍA
Circuito No. 1, Campus Universitario II
Tel. (614) 442-9500
Chihuahua, Chih., México
www.uach.mx/fing

Resumen

La interpretación de imágenes médicas es un proceso fundamental para el diag nóstico y tratamiento clínico. Para la realización de un diagnóstico y reporte médico considerando diversas técnicas de proyección de imagen, es necesario examinar minuciosamente cada área de la imagen. El objetivo de esta tarea es identificar cuáles regiones se consideran normales, anormales o potencialmente anormales. Una vez que son identificadas estas regiones, se realiza una narrativa de los resultados obtenidos que conforman al reporte médico. Por lo anterior, realizar un reporte médico mediante imágenes médicas demanda una mayor especialización en el área. Aunado a esto, existen variaciones en la interpretación de las imágenes médicas entre un especialista y otro. Además los tiempos de ejecución para realizar un diagnóstico pueden variar dependiendo del especialista que lo realiza. Considerando el aumento de la carga de trabajo en los centros de salud, la integración de la Inteligencia Artificial (IA) en las herramientas tecnológicas para el diagnóstico médico es de gran utilidad. Lo anterior se puede destacar mediante aplicaciones inteligentes que sean capaces de generar un reporte médico inicial del cuál puedan partir los especialistas para realizar el diagnóstico médico en un menor tiempo.

El presente trabajo de investigación propone una arquitectura novedosa denominada *Enhanced Transformer Based - Medical Image Interpretation* (ETB-MII) para generar reportes médicos. La arquitectura propuesta es capaz de identificar en una imagen de rayos X características visuales tanto locales como globales y generar una descripción textual de los hallazgos encontrados. Esta propuesta pretende demostrar que una implementación basada principalmente en mecanismos de atención y aplicando una estrategia de aumento de datos eficaz puede lograr resultados competitivos en el estado del arte. Para evaluar adecuadamente el desempeño del modelo ETB-MII, se compararon las predicciones generadas con respecto a los resultados reportados por los modelos más relevantes del estado del arte utilizando los conjuntos de datos de IU X-ray y MIMIC-CXR asi como las métricas de BLEU, METEOR, ROUGE-L y CIDEr. Es importante mencionar que las métricas mencionadas son oficialmente utilizadas para evaluar la tarea de generación de reportes médicos mediante imágenes médicas. Los resultados obtenidos por ETB-MII demuestran un desempeño competitivo de acuerdo a las métricas de BLEU y ROUGE-L. Sin embargo, para las métricas de CIDEr y METEOR los resultados obtenidos sobrepasan el estado del arte.

Como parte de esta investigación se examinaron diversas predicciones generadas por el modelo ETB-MII y se realizó un análisis comparativo con predicciones generadas del modelo base R2GEN. Lo anterior permitió comprobar que el modelo ETB-MII genera descripciones más precisas de las imágenes de rayos X empleando una mayor diversidad en el vocabulario médico. Además, se realizó una evaluación de los resultados basada en la similitud de coseno para verificar que el modelo ETB-MII genera predicciones precisas y evita generar frases repetitivas. Finalmente, se identifica que la métrica de CIDEr es la más adecuada para evaluar este tipo de tareas. CIDEr permite evaluar al texto generado desde un punto de vista semántico, por lo que produce una valoración más adecuada de la interpretación textual producida a diferencia de métricas de BLEU y ROUGE las cuales favorecen a los modelos que memorizan las palabras más comunes.

Para Roberto, Natalia y Ana Lucia, los protagonistas de mi vida.

Agradecimientos

Quisiera expresar mi más profundo agradecimiento a la Universidad Autónoma de Chihuahua, por su invaluable apoyo durante todos mis estudios. En particular, agradezco a la Facultad de Ingeniería por abrirme las puertas para formar parte de esta gran comunidad académica.

Agradezco a mi directora de tesis, la Dra. Graciela María de Jesús Ramírez Alonso por toda su guía, compromiso, paciencia, comprensión y amistad. Sin su apoyo no hubiera sido posible alcanzar los objetivos propuestos y culminar con este gran proyecto.

Agradezco a mi Co-director, el Dr. Manuel Montes y Gómez, por enriquecer mi investigación mediante su orientación, enseñanzas y experiencia.

Agradezco a mi comité de evaluación, el Dr. Luis Carlos González Gurrola, el Dr. Raymundo Cornejo García, y al Dr. Alaín Manzo Martínez por sus enseñanzas y asesorías.

Agradezco al departamento de Investigación y Posgrado de la Facultad de Ingeniería, el Dr. Fernando Martínez Reyes y a la Dra. María Isabel Flores Zamora por su seguimiento y apoyo.

Agradezco al Dr. Alejandro Villalobos Aragón, a la Maestra Luly Flores, a la Maestra Karina Requena por el apoyo recibido y su gran amabilidad.

Gracias a mis maravillosos padres por todo su amor, por su ejemplo y sobre todo por creer siempre en mi.

Gracias a mis hermanos por estar en mi vida y por brindarme su apoyo.

Gracias a mi amado esposo por impulsarme a ser mejor día con día. Por tu amor incondicional, tus consejos, por ser mi soporte de vida; no hay palabras suficientes para agradecer todo lo que haces por mi.

Gracias a mis queridas hijas por todos sus abrazos, besos y palabras de apoyo en los momentos más complicados; por ser mi público durante mis largos ensayos de las presentaciones que tuve que impartir. Ustedes son mi razón de vivir.

Gracias a Dios por permitirme tanto.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

La interpretación de imágenes médicas es un proceso fundamental para el diagnóstico y tratamiento clínico. Este proceso contribuye a determinar las causas de los síntomas que presenta un paciente, para monitorear efectos y evaluar la utilidad de algún tratamiento. Además, la interpretación de imágenes médicas es indispensable para el seguimiento de la evolución de una enfermedad en un paciente durante un periodo de tiempo [7],[8].

Para la realización de un diagnóstico y reporte médico considerando diversas técnicas de proyección de imagen, es necesario examinar minuciosamente cada área de la imagen. El objetivo de esta tarea es identificar cuales regiones se consideran normales, anormales o potencialmente anormales. Una vez que son identificadas estas regiones, se realiza una narrativa de los resultados obtenidos que conforman al reporte médico [9].

Un aspecto a considerar en este proceso es la diversidad de los tipos de imágenes médicas [10–12], por ejemplo dentro de las imágenes de radiología se encuentran:

- Los rayos X.

- Las Tomografías Computarizadas (CT).
- Las Imágenes de Resonancia Magnética (MRI).
- Las Tomografías por Emisión de Positrones (PET).
- Los Ultrasonidos (US).

Por consecuencia, realizar un diagnóstico y reporte médico utilizando imágenes de radiología demanda una mayor especialización en el área [7]. Aunado a esto, existen variaciones en la interpretación de las imágenes médicas entre un especialista y otro. Al igual, los tiempos de ejecución para la realización de un reporte médico varían dependiendo del especialista que lo realiza. Según Alfarghaly et al. en [13], el proceso general para la realización de un reporte médico dura en promedio 10 minutos. Sin embargo, algunos reportes radiológicos pueden tardar más de 30 días debido a la creciente demanda de diagnósticos por imagen [14]. Estos aspectos contribuyen a que el proceso sea complejo y tardado, además se ha identificado que los errores humanos en la interpretación oscilan entre el 20 % y el 30 %, lo que puede dar lugar a un innecesario estrés en los pacientes [15]. En consecuencia, la incorporación de avances tecnológicos se ha convertido en un imperativo para mejorar los procesos del diagnóstico por imagen médica [8].

La automatización del proceso de diagnóstico y la generación de reportes médicos no es una tarea sencilla. En las últimas décadas, el aprendizaje automático (Machine Learning, ML por sus siglas en inglés), y en concreto los algoritmos de aprendizaje profundo (Deep Learning, DL), se han utilizado cada vez más para explorar diversas estrategias en la generación automatizada de reportes médicos con resultados sobresalientes [16–18]. Estas soluciones se basan en construir modelos de aprendizaje mediante el uso de arquitecturas de tipo codificador-decodificador que incluyen redes neuronales convolucionales (CNN), redes neuronales recurrentes (RNN)

[7],[8] y/o *Transformers* [19]. Sin embargo, la complejidad de los modelos propuestos va en aumento, ya que incluyen estructuras computacionalmente más costosas como modelos de *Tranformers* que incluyen dentro de su estructura memorias auxiliares basadas en recurrencias, modelos basados en grafos, CNNs, por nombrar algunos [6],[20],[21].

A su vez, es importante mencionar que este tipo de modelos computacionales requieren de una gran cantidad de imágenes médicas y reportes médicos para su entrenamiento. Por lo cuál antes de realizar el entrenamiento del modelo es imprescindible considerar técnicas de pre-procesamiento en los datos. Estas técnicas contribuyen a mejorar el aprendizaje y el desempeño del modelo. Por ejemplo, en las imágenes médicas, diferentes regiones dentro de la imagen pueden representar información relevante para el diagnóstico, mientras que otras regiones son inútiles. Si la imagen se utiliza sin aplicar ninguna técnica de pre-procesamiento, el modelo dedica parte de su entrenamiento en aprender información irrelevante. Por otro lado, si se aplica un pre-procesamiento en la imagen como recortar áreas innecesarias antes de entrenar al modelo, se logra mejorar el aprendizaje [22]. Otro ejemplo de su utilidad, es la aplicación de técnicas de aumento de datos para generar variaciones de la imagen o en el texto. Esto permite contribuir a balancear las clases y reducir la probabilidad de un sobre-entrenamiento [23]. Otro problema frecuente que se reduce con las técnicas de pre-procesamiento son los diferentes contrastes en las imágenes médicas que provienen de equipos de diagnóstico diferentes [8].

Por otra parte, considerando que los reportes médicos son elaborados manualmente por radiólogos de forma estructurada o semi-estructurada, el vocabulario varía de un reporte a otro. Uno de los primeros pasos a seguir es detectar palabras clave en los reportes para identificar el diagnóstico asociado. Posteriormente, etiquetar ese

diagnóstico con un vocabulario normalizado. Por consiguiente, se han propuesto varias herramientas basadas en estrategias de DL, para extraer texto de los reportes médicos e identificar las palabras clave que permitan homogeneizar los diagnósticos. Algunas de estas herramientas son NegBio labeler [10] y CheXpert labeler [24].

La mayoría de los enfoques propuestos para la automatización de reportes médicos analizados en este documento proponen una estrategia basada en un ensamble de redes neuronales artificiales (RNAs). En estos enfoques se presenta un primer modelo de visión, el cual aborda la tarea de procesamiento de la imagen y posteriormente la tarea de generación de texto se realiza en un modelo lenguaje. Algunas de estas propuestas señaladas anteriormente se centran en generar las etiquetas que describen los hallazgos encontrados en la imagen [25–27]. Por otra parte, otros de los enfoques analizados se centran en generar el texto del reporte médico utilizando una estrategia de entrenamiento de principio a fin (*end-to-end*)[28][29]. El diagrama de la figura 1.1 muestra de manera general el proceso de entrenamiento de estos modelos.

Aunque existen diferencias en los enfoques propuestos, el proceso general es definido mediante los siguientes pasos:

1. Seleccionar una base de datos que contenga imágenes médicas y reportes médicos asociados.

2. El modelo propuesto recibe como entrada la imagen médica y aprende las características más relevantes de la imagen.

3. Se asocian las descripciones médicas de los reportes a las características aprendidas de la imagen.

4. Se alimentan y entrenan los modelos de lenguaje para la generación de la des-

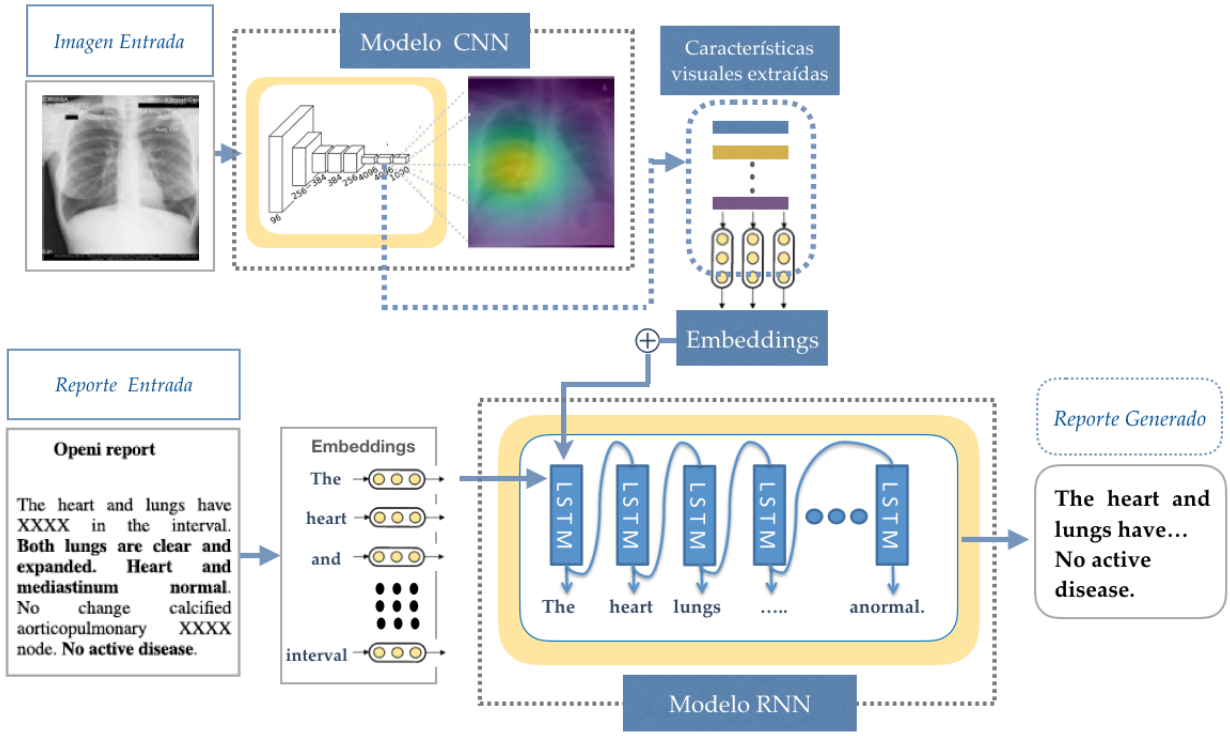


Figura 1.1: Representación gráfica del proceso de entrenamiento para un Modelo de DL basado en imagen/texto para generar reportes médicos.

cripción de la imagen.

Por consiguiente, el proponer un modelo competitivo para la generación automática de reportes médicos basado en estrategias de DL puede contribuir significativamente a esta tarea. Sin embargo, además de los costosos recursos computacionales requeridos para el entrenamiento de estos modelos, surgen retos adicionales como:

- La necesidad de homogeneizar la información semántica de la imagen y del texto del reporte médico para lograr una generación automática de párrafos largos, coherentes y comprensibles por los especialistas.
- La reducida cantidad de imágenes con reportes médicos asociados para entrenar los modelos de DL propuestos.

- La representación inadecuada de los textos que describen hallazgos anómalos.

Una desventaja que se identifica en algunos de los modelos propuestos es la falta de creatividad al predecir las descripciones de los hallazgos. Lo anterior se refleja en las frases repetitivas generadas para diferentes imágenes de rayos X. Esto motiva la incertidumbre del aprendizaje del modelo, ya que no queda claro si aprende a identificar características relevantes en la imagen o si solamente memoriza las palabras más frecuentes del texto logrando un desempeño aceptable de acuerdo a las métricas.

Considerando que la generación de lenguaje natural a través de la automatización es un proceso complejo, actualmente algunos modelos *Large Language Models* (LLMs) como: *Bidirectional Encoder Representations* (BERT)[30], *Large Language Model Meta AI* (LLaMa) [31], y *Generative Pre-trained Transformer GPT* [32], mejoran las capacidades de producción lingüística. Estos modelos se han entrenado utilizando enormes cantidades de datos y son capaces de mostrar una generalización significativamente mayor en tareas nuevas sin entrenamiento específico de la tarea mediante la "visualización"de unos pocos ejemplos dentro del dominio específico (tarea conocida como few/zero-shot learning) [33–36].

Basado en lo anterior, el presente trabajo de investigación tiene como objetivo contribuir a la tarea de generación automática de reportes médicos mediante imágenes de radiología de pecho (tórax). En este trabajo se propone una arquitectura denominada *Enhanced Transformer Based - Medical Image Interpretation* (ETB-MII) de tipo codificador-decodificador (*encoder-decoder*); la cuál aborda la tarea como un modelo de traducción, donde la imagen es la entrada del modelo y el reporte médico es la traducción obtenida de la imagen. La arquitectura incorpora un modelo basado en *Visual Transformer*, el cuál mediante mecanismos de atención y operadores

ponderados identifica las características visuales locales y globales para determinar su relevancia. Mientras que la parte del *decoder* está constituido por un modelo *Generative pre-trained Transformer* el cuál recibe las características visuales del *encoder* y asocia el texto médico correspondiente. Adicionalmente, inspirados en el trabajo realizado en [37], se definió una estrategia para el aumento de datos contextuales que permite incrementar la calidad en los textos médicos mejorando la diversidad de las descripciones. Para evaluar la eficacia del modelo ETB-MII se utilizaron los conjuntos de datos públicos IU X-ray y MIMIC-CXR; además se comparó el modelo ETB-MII contra los modelos del estado del arte mediante las métricas de *Natural Language Generation* (NLG) como *Bilingual Evaluation Understudy* (BLEU) [38], *Metric for Evaluation of Translation with Explicit Ordering* (METEOR) [39], *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) [40] y *Consensus-based Image Description Evaluation* (CIDEr) [41].

Como parte de esta investigación se examinaron diversas predicciones generadas por el modelo ETB-MII y se realizó un análisis comparativo con predicciones generadas del modelo base R2GEN. Lo anterior permitió comprobar que el modelo ETB-MII genera descripciones más precisas de las imágenes de rayos X empleando una mayor diversidad en el vocabulario médico. Además, se realizó una evaluación de los resultados basada en la similitud de coseno para verificar que el modelo ETB-MII genera predicciones precisas y evita generar frases repetitivas. Finalmente, se identifica que la métrica de CIDEr es la más adecuada para evaluar este tipo de tareas, donde el modelo ETB-MII obtuvo los resultados más sobresalientes en el estado del arte. CIDEr se basa en el significado semántico del texto y en el uso de sinónimos, a diferencia de métricas como BLEU y ROUGE las cuales favorecen a los modelos que sólo memorizan las palabras más comunes. Aunque las arquitecturas propuestas ofrecen un buen desempeño, aún existen áreas de oportunidad

considerables que motivan a explorar nuevas soluciones. Por ejemplo:

- El aumentar el volumen de las bases de datos con imágenes médicas para entrenar los modelos DL.
- El proponer modelos eficientes con arquitecturas simples que demanden menos costo computacional.
- El lograr que los textos generados por el modelo sean comprensibles y coherentes.

Además, es necesario desarrollar mejores enfoques que contemplen una estrategia *end-to-end* para la generación de reportes médicos, lo cuál es de gran utilidad para esta problemática.

En conclusión, la implementación de estrategias de DL aplicadas a la clasificación de imágenes médicas y a la generación de reportes médicos contribuyen a un mejor desempeño. Sin embargo, existen estrategias de DL que no se han explorado en su totalidad para estas tareas, como los modelos basados exclusivamente en mecanismos de atención. Por lo tanto, realizar una exploración a profundidad de este tipo de arquitecturas para la generación de reportes médicos basado en imágenes de rayos X es de gran utilidad.

1.1. Preguntas de Investigación

El presente trabajo de investigación tiene como objetivo contribuir a la tarea de generación automática de reportes médicos en inglés a partir de imágenes médicas. Para la realización de este objetivo se pretende analizar, definir e implementar un modelo basado en mecanismos de atención , donde el problema se plantea considerando a la imagen como la información inicial y el reporte médico como la tra-

ducción de la imagen. Las siguientes preguntas se abordarán durante la presente investigación.

1. ¿Qué particularidades de un modelo DL enfocado a tareas de traducción de texto pueden adaptarse a la generación de un reporte médico a partir de imágenes médicas?

2. ¿Qué técnicas para el pre-procesamiento de imágenes y texto contribuyen a mejorar el nivel de precisión de la representación visual y textual de un modelo para la generación de reportes médicos?

3. ¿Qué tipos de mecanismos de atención pueden generar una representación visual y textual de una imagen médica de rayos X con alto nivel de precisión?

4. ¿Qué impacto tendrá el contextualizar una imagen asociada a un texto médico utilizando modelos basados en atención?

5. ¿Cómo una arquitectura *end-to-end* basada en mecanismos de atención puede contribuir a mejorar los resultados del estado del arte para la generación de reportes médicos mediante de imágenes de rayos X?

1.2. Objetivo General

Proponer un modelo de DL enfocado a la generación automática de reportes de médicos a partir de imágenes rayos X de tórax basado en mecanismos de atención que genere textos estructurados con un alto nivel de calidad; mediante una arquitectura de tipo *encoder-decoder* empleando una estrategia de aumento de datos que permita entrenar al modelo en equipos de computo con capacidad limitada. El modelo debe ser capaz de obtener resultados competitivos comparados con los modelos del estado del arte.

Objetivos específicos

1. Analizar los diferentes modelos propuesto para la generación automática de reportes médicos a partir de imágenes médicas.

2. Definir procedimientos e implementar técnicas para el pre-procesamiento de imágenes médicas de rayos X que contribuya a la extracción de características visuales.

3. Definir e implementar un modelo basado en mecanismos de atención para la extracción de características de imágenes de rayos X.

4. Definir e implementar un modelo basado en mecanismos de atención para la generación automática de oraciones y palabras asociados a las características visuales de las imágenes médicas de rayos X.

5. Desarrollar una arquitectura de principio a fin (*end-to-end*) para la generación automática de reportes médicos mediante imágenes de rayos X de tórax con un mayor nivel de calidad comparado con el estado del arte.

1.3. Justificación

Un diagnóstico médico es crucial y fundamental para la toma de decisiones clínicas. Su contribución a la detección y monitoreo de enfermedades, al igual que la planificación de terapias o tratamientos repercute significativamente en la calidad de vida del paciente [42]. Para ayudar a prevenir un daño mayor al paciente a causa de la progresión de la enfermedad o para vigilar la progresión de cierto padecimiento, un diagnóstico temprano es esencial [42]. Por ejemplo, algunas enfermedades respiratorias no tratadas adecuadamente pueden traer complicaciones graves como es el caso de la neumonía, la cuál es responsable del 15 % de todas las defunciones de menores de 5 años [43]. Otro ejemplo son los casos notificados de MERS-Cov, donde

de acuerdo a la Organización Mundial de Salud (OMS) aproximadamente el 35 % de los casos han desembocado en la muerte del paciente [44]. Por otra parte, la falta de diagnóstico es un problema frecuente para la detección de cáncer oportuna. Por lo anterior, se ha identificado que menos del 30 % de los países de bajos ingresos cuentan con servicios de patología para atender a la población en general [45]. El mismo escenario aplica con los casos de glaucoma no diagnosticados, los cuales representan más del 50 % de los casos en todo el mundo [42].

El uso de imágenes médicas para la elaboración de un diagnóstico médico es una práctica comúnmente utilizadas para la detección y seguimiento de ciertos padecimientos como: cáncer, neumonía, Covid-19, Alzheimer, enfermedades del riñón, por nombrar algunos [23], [46–48]. Durante su elaboración, las imágenes médicas son analizadas por especialistas con el propósito de identificar ciertas características visuales que revelen aspectos de una región anormal o posiblemente anormal. Sin embargo, debido a su complejidad y al nivel de experiencia requerida para su ejecución, esta actividad suele ser tardada, compleja, y propensa a variabilidad por parte del ejecutor [49]. Tal es el caso de la realización de un procedimiento manual para identificar y segmentar determinada región en un MRI, donde puede existir una variabilidad en la operación de un 28 % en promedio y un tiempo de ejecución de 10 minutos por corte [50].

Con la llegada de las diversas metodologías de DL al área de análisis de imágenes médicas, han sido propuestos múltiples modelos para abordar tareas relacionadas a reportes médicos. Algunos ejemplos son modelos para las tareas de clasificación, detección, segmentación así como los modelos secuenciales para la generación automática de diagnósticos médicos. Sin embargo, aún queda una brecha por cubrir debido a que faltan modelos que incorporen estructuras de principio a fin me-

nos complejos que generen automáticamente diagnósticos médicos. Estos modelos deben ser precisos, con un lenguaje claro y entendible para los especialistas, y con un menor costo computacional. En consecuencia, un modelo de DL que incluya las estrategias más precisas para la generación automática de reportes médicos puede contribuir significativamente al estado del arte en el área de diagnósticos médicos.

Capítulo 2

Trabajo Relacionado

Uno de los aspectos más importantes para automatizar la tarea de generación de reportes médicos mediante imágenes de rayos X, es obtener una representación visual que permita identificar patrones característicos de una patología [51]. Para ello, es fundamental que el modelo de DL realice la extracción de características de la imagen discriminatoriamente; lo que implica que se identifiquen las características más significativas y relevantes al desempeñar la tarea.

Mediante la introducción de las Redes Neuronales Convolucionales (CNN), la obtención automática de las características relevantes de una imagen fue posible y gracias a su óptimo desempeño un gran número de modelos de CNN fueron propuestos causando impactantes resultados en el área de visión por computadora. AlexNet, Google Net, VGG, Residual Net, y otras más, son algunos de estos modelos [8]. Esto motivó a que se explorarán diversas estrategias de CNN's en el área de análisis de imágenes médicas logrando resultados relevantes [8],[9],[52].
Por ejemplo :

- La clasificación de un cancer de pecho y micro-calcificaciones en una imagen de mamografía [46].

- La detección y segmentación de tumores cerebrales a través de imágenes de resonancia magnética (MRI) [23],[22],[47], [48].
- La generación de reportes médicos mediante imágenes médicas. Esta tarea se enfoca en generar texto que describe las características relevantes encontradas en una imagen médica [7], [53].

Para llevar acabo la generación de reportes médicos mediante imágenes médicas, se han probado diversos métodos que involucran modelos para la clasificación de imágenes hasta modelos secuenciales (mejor conocidos en la literatura como *sequence to secuence* o "S2S") [54]. Dentro de los modelos S2S se han identificado varias arquitecturas que incluyen un ensamble de CNN y LSTMs, CNN y RNN, modelos híbridos que incluyen CNN, LSTMs y plantillas predefinidas, modelos jerárquicos compuestos por CNN y LSTMs que requieren ejecuciones por niveles y modelos mixtos que incluyen métodos auxiliares para la generación de reportes médicos [55], [56].

En la siguientes secciones se identifican las bases de datos que incluyen imágenes y reportes médicos asociados; además se analizan de manera general algunas de las arquitecturas más relevantes enfocadas a la tarea de clasificación de imágenes médicas de rayos X que emplearon estrategias de Procesamiento de Lenguaje Natural para el etiquetado de la imagen y arquitecturas enfocadas a la tarea de generación de reportes médicos mediante imágenes médicas de rayos X.

2.1. Bases de datos: imagen/texto área médica

La implementación de modelos de DL para la tarea de generación de reportes médicos, requiere de bases de datos que contengan grandes cantidades de imágenes médicas y textos médicos asociados [28],[46], [57].

Aunque a la fecha se cuenta con un limitado número de bases de datos públicas con reportes médicos, existen conjuntos de datos en inglés que incluyen diagnósticos de pacientes con una o más imágenes de rayos X junto con su reporte médico.

En algunas investigaciones se han utilizado PACS privados; éstos son repositorios de centros médicos o hospitales donde se almacenan diagnósticos e imágenes médicas. Dentro de este conjunto se identifica el sistema PACS-1 del Hospital de Shaanxi, China, que contiene 16,569 imágenes rayos X de tórax y 16,569 reportes médicos en lenguaje chino fue utilizado por [58]. Para la recolección de información se consideraron los reportes registrados en los años 2014 y 2015. Del reporte médico, solo las secciones de *findings* y *diagnosis* fueron consideradas. Para la definición de etiquetas se preprocesó la sección de *diagnosis* donde se corto el texto en sentencias y se identificaron las sentencias similares mediante técnicas de agrupamiento de NLP. De esta técnica se logró identificar 10 patologías con una frecuencia de 100 casos o más, las cuales fueron establecidas como etiquetas cubriendo la clasificación del 97 % de imágenes. Una de las desventajas de utilizar PAC's privados es que no existe disponibilidad de los datos utilizados en la investigación, por lo cual la propuesta no puede ser reproducida de acuerdo a lo reportado. De igual manera se reduce la posibilidad de ser comparada justamente con otros modelos del estado del arte.

La colección de datos de PACs-2 del Hospital de Shanghai del 2015 [59], incluye 19,985 imágenes de rayos X de tórax con 19,985 reportes asociados. Para esta colección se descartaron todas las imágenes laterales así como imágenes irregulares. Con respecto a los reportes médicos, primero fueron seleccionados los textos que incluyen la palabra pulmón y mediante técnicas de NLP, el texto fue separado en

tokens para identificar su frecuencia y considerar únicamente aquellos que tienen frecuencias mayores a 2. Posteriormente, se identificó un vocabulario y los 40 *tokens* más frecuentes se establecieron como etiquetas. Sin embargo, al igual que PACs-1, al ser un conjunto de datos privados su uso es muy limitado.

PEIR Digital Library[9] es una colección de datos pública utilizada por la educación médica. Esta colección contiene 7442 imágenes de rayos X y su descripción clasificadas en 20 categorías. Fue creada por *The University of Alabama* y se caracteriza por incluir descripciones a nivel de oración de 20 partes diferentes del cuerpo. Sin embargo, aunque es una base de datos pública solo es utilizada para tareas de clasificación ya que no incluye las diferentes secciones de un reporte médico.
Otra base de datos es *PadChest* [60], la cual contiene 160,868 imágenes de rayos X de tórax con 6 vistas diferentes y 109,931 reportes médicos. Las imágenes fueron interpretadas por 18 radiólogos del Hospital Universitario de San Juan, Alicante (España), durante los años 2009 a Diciembre 2017. Para la elaboración del conjunto de datos se realizó la extracción de imágenes de archivos DICOM, de los reportes médico se eliminó información sensible del paciente. Las imágenes que no contienen un reporte médico asociado fueron eliminadas de la colección así como aquellas que no incluyen la modalidad, la interpretación fotométrica o que incluyen una proyección horizontal. Se conservaron las imágenes con vistas laterales (L), anterior-posterior (AP) y posterior-anterior(PA). Adicionalmente fueron identificados diferentes protocolos de acuerdo a diferentes escenarios clínicos como *X-ray móviles* (utilizados en pacientes que no pueden estar de pie). Para el pre-procesamiento de los reportes médicos y obtención de etiquetas, se aplicaron técnicas de NLP de Doc2Vec y k-Means que permiten identificar agrupaciones en el vocabulario. Esta colección de datos incluye 174 descripciones o *findings*, 19 diagnósticos y 104 anotaciones de la ubicación. Debido a que es una base de datos relativamente nueva y en español, su

uso es muy limitado.

La base de datos de *Eye-Gaze Dataset* es un subconjunto de datos de *MIMIC-CXR* [61] y de *MIMIC-IV-ED* [62]. Esta base de datos además de incluir las observaciones (*findings*) y resultados (*impressions*) incluye resultados relacionados con algunos exámenes realizados al paciente. Para crear esta nueva colección de datos se seleccionaron 1,083 casos de la base de datos de MIMIC-CXR que incluyeran casos normales, estudios con casos con de neumonía y casos con insuficiencia cardiaca congestiva (CHF). Posteriormente un radiólogo con más de 5 años de experiencia realizó un análisis de cada uno de los 1,083 casos. Durante el análisis el radiologo utilizó un monitor de ojo *the Gazepoint GP3 Eye Tracke* para grabar todos los movimientos oculares [63]; un micrófono para grabar la voz mientras narraba los hallazgos y características relevantes encontradas en la imagen. Por cada sesión realizada con una duración de 30 minutos, el especialista logró analizar 30 casos. Aunque su uso no ha sido reportado para la generación de reportes médicos, se considera puede contribuir significativamente en esta área. La desventaja de este conjunto de datos es que demanda una alta capacidad de almacenamiento y se manejan diversas modalidades de datos lo cuál agrega mayor complejidad a los modelos.

Las bases de datos de *ImageCLEF Caption 2017* [64] e *ImageCLEF Caption 2018* [65], son grandes colecciones de imágenes biomédicas que incluyen imágenes de radiología de tórax y de diferentes partes del cuerpo. Las imágenes fueron extraídas del *PubMed Central (PMC)* y tienen asociada una oración, así como un conjunto de etiquetas UMLS *(Unified Medical Language System)*. La base de *ImageCLEF Caption* 2017 contiene 184,614 imágenes mientras que *ImageCLEF Caption* 2018 cuenta con 232,305 imágenes. *ImageCLEF Caption 2017* se ha utilizado para la tarea de generación de reportes médicos [66], mientras que *ImageCLEF Caption* 2018 solo ha

reportado su uso para tareas de generación de subtítulos organizadas por *ImageCLEFcaption*. Una de las desventajas que presenta el conjunto de datos es que texto médico asociado no incluye las secciones establecidas de un reporte médico como *findings* e *impressions*.

La base de datos *CX-CHR* [67] es una colección privada de imágenes de rayos X con textos médicos en lenguaje mandarín. La colección de datos fue recopilada por una institución médica de salud, la cuál incluye diagnósticos de 35,500 pacientes. Cada diagnóstico contiene una o más imágenes de rayos X de tórax de vista lateral y frontal, así como reportes médicos correspondientes. Al igual que los PACs anteriormente mencionados, al ser un conjunto de datos privado su uso es muy limitado.

CheXpert [24] es una base de datos pública de imágenes de rayos X de tórax que contiene 224,316 imágenes de 65,240 pacientes. La base de datos considera la clasificación de imágenes identificando la presencia o ausencia de 14 tipos de patología. Para su elaboración se diseño una herramienta para etiquetar la presencia o ausencia de las 14 patologías. Las imágenes de rayos X, fueron recolectadas del Hospital de Stanford entre el mes de Octubre del 2002 y Julio del 2017, junto con su respectivo reporte. De esta colección de datos se tomó una muestra de 1000 reportes para ser revisados manualmente por radiólogos experimentados y determinar la factibilidad de extraer la información de la sección de observaciones; esto con la finalidad de identificar las patologías para etiquetar las imágenes. La herramienta diseñada con la finalidad de etiquetar automáticamente la colección de datos, se basó en extraer información de los reportes médicos en la sección de *impressions* y *findings* las palabras clave asociadas a las patologias. Posteriormente, se clasificaron las observaciones extraídas de los reportes como negativo, incierto y presente. Finalmente,

la colección de datos se asignó de acuerdo a su clasificación. Aunque es una base de datos muy completa, no es muy utilizada para la tarea de generación de reportes médicos .

Otra base de datos pública que incluye imágenes y reportes médicos es *ChestX-ray8* [68]. Esta base de datos contiene 108,948 imágenes de rayos X con vista frontal de 32,717 pacientes, en la cuál se identifica la ausencia o presencia de 8 patologías. Para determinar las etiquetas se identificaron las patologías mas comunes encontradas en las imágenes de rayos X así como la retroalimentación de radiólogos experimentados. Posteriormente se realizó una búsqueda en el sistema PACS para identificar cuales reportes incluyen las 8 patologías. Diferentes técnicas de Procesamiento de Lenguaje Natural (NLP) fueron utilizadas para extraer información clave de los reportes médicos así como para remover la información innecesaria para la colección de datos. Cada reporte médico fue ligado a una palabra clave o marcado como ”normal”de acuerdo a su categoría. Las secciones de *findings* e *impressions* fueron las únicas que se consideraron para la colección de datos. Debido a que solo se utilizan 8 patologías, en la literatura se ha dado preferencia a otros conjuntos de datos que incluyen una mayor clasificación de patologías.

La colección del *National Institute of Health (NIH) Clinical Center* denominada *ChestX-ray14* está conformada por 112,120 imágenes de rayos X e incluye 14 etiquetas de enfermedades de pecho. Esta base de datos es una actualización de [68], la cual incluye 6 etiquetas más de patologías y 27 veces más imágenes de rayos X que en [4]. Además el conjunto de datos incluye una distribución más representativa de acuerdo a la población real de los pacientes y a los desafíos diagnósticos que se presentan al día.

MIMIC-CXR-JPG [61],[69] es una colección de datos reciente que cuenta con 377,110 imágenes de rayos X de tórax y 227,827 reportes asociados del centro médico *Beth Israel Deaconess Medical Center*. Los diagnósticos comprenden del año 2011 al 2016. Las imágenes fueron extraídas del PACS en formato DICOM, mismas que fueron relacionadas con sus respectivos reportes médicos. Cada registro e información sensible del paciente fue removida del archivo DICOM, así como información no relevante al diagnóstico. Posteriormente, las imágenes fueron exportadas a formato JPEG en formato estándar, utilizando la librería de Pydicom. Los píxeles fueron normalizados en un rango de [0,255]. Se valido que cada imagen pueda cambiar de intensidad en el contraste al subir los valores en los píxeles. Para la colección de reportes médicos, se utilizaron las sección de *findings* y *impressions*. Las 14 patologías más frecuentes en la sección *findings* fueron extraídas para establecerlas como etiquetas con la herramienta NegBio [70] y la herramienta diseñada en [24] y en la tarea de generación de reportes médicos [71],[27],[72],[73] . Además de se una colección de datos reciente es bastante utilizada en la literatura debido a la accesibilidad para obtener los datos y la presentación de los mismos.

Una de las bases de datos más utilizadas durante el entrenamiento y evaluación de modelos propuestos se encuentra *The Indiana University chest X-ray, IU X-ray* [4]. Esta base de datos esta constituida por 3,955 reportes médicos y 7,470 imágenes de rayos X de tórax que incluyen la vista frontal y lateral. La selección de los datos se realizó considerando un solo reporte médico por paciente, el cuál se incluyó en la base de datos junto con la o las imágenes de rayos X identificadas en el reporte médico. Estas imágenes fueron extraídas del Sistema de Archivos de Imagen (*Picture Archiving and Comunication System*, PACS). Los reportes fueron modificados, removiendo información sensible del paciente como nombre, número de seguro social, etc. Inicialmente se contemplaban 4000 reportes médicos, pero solo

fueron seleccionados aquellos que incluyeran las secciones de razones de estudio, recomendaciones (*findings*) y dignóstico (*impression*). Se analizaron y validaron cada una de las imágenes en formato DICOM así como sus respectivos reportes por 8 especialistas. Aquellos reportes que no fueran aprobados por los especialistas eran desechados de la colección al igual que las imágenes asociadas al reporte.

La Tabla 2.1 muestra las bases de datos públicas y PACS privados que incluyen imágenes de rayos X y texto médico, que se han utilizado para tareas de clasificación y/o generación de reportes médicos a partir de imágenes.

Tabla 2.1: Bases de datos de imágenes médicas de radiografía con reportes médicos asociados.

Bases de datos	Imágenes	Descripción	Utilizado por
PACs Hospital of Shaanxi China(PACs-1)	16,569	16,569 reportes	[58]
National Institutes of Health (PACs-2)[59]	19,985	19,985 reportes	[59]
PEIR Digital Library[9]	4,732	20 multi-etiquetas	[9]
PadChest[60]	160,868	109,931 reportes	-
Eye-Gaze Dataset[63]	1,083	Reportes transcritos Audio de reporte Movimiento Ocular	-
ImageCLEF-Caption-2017[64]	184,614	184,614 reportes	[66]
ImageCLEF-Caption-2018[65]	232,305	232,305 reportes	-
CX-CHR[67]	35,500	35,500 reportes	[67, 74, 75]
CheXpert[24]	224,316	65,240 reportes	[76]
ChestX-ray8 [68]	108,948	8 etiquetas	[53, 75, 77]
ChestX-ray14 [68]	112,120	14 etiquetas	[53, 75, 77]
MIMIC-CXR [61]	371,920	227,943 reportes	[27, 71]
IU X-ray [4]	7,470	3,955 reportes	[9, 53, 67, 74, 78, 79] [28, 71, 80–82] [75, 83–87] [76, 88–91]

2.2. Arquitecturas orientadas a la clasificación de imágenes médicas.

Algunas de las arquitecturas orientadas a la clasificación de imágenes médicas se caracterizan por utilizar estrategias de Procesamiento de Lenguaje Natural (NLP) para extraer información relevante de los reportes que corresponda a ciertas patologías. Posteriormente, definir la o las patologías encontradas como etiquetas y entrenar al modelo para realizar la clasificación de la imagen. La figura 2.1 muestra de manera general la arquitectura de un modelo DL basado en una CNN para la clasificación de una imagen de rayos X, considerando que ya se ha establecido su etiqueta.

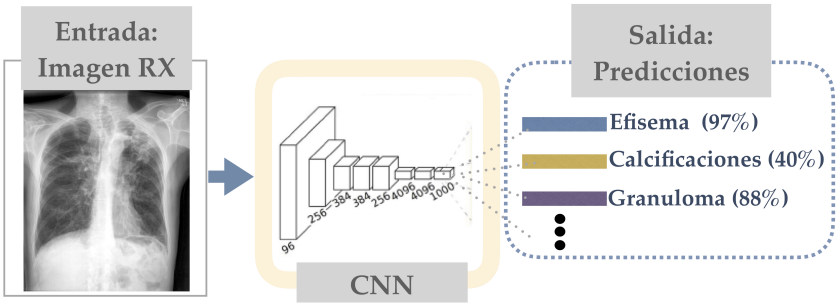


Figura 2.1: Modelos enfocados a la clasificación de imágenes.

En el año del 2015, Shin et al. [25], propusieron una estructura capaz de asociar automáticamente una imagen con su descripción utilizando modelos orientados a NLP. El modelo utilizó alrededor de 780,000 reportes de radiología y 216,000 imágenes de 2D para asociar la información semántica entre el reporte y la imagen. Por otra parte, Wang et al. [26], propusieron un modelo recursivo que forma agrupamientos a partir de los reportes de texto y de las características de la imagen. Lo anterior contribuyó a obtener una coherencia visual y un método balanceado al momento de agrupar. Mediante una CNN se identificaron las características visuales y se fueron asociando mediante un *fine-tuning* a las etiquetas principales. Durante

cada iteración del entrenamiento, se reafinaron los agrupamientos de las etiquetas los cuales de forma recursiva fueron alimentando nuevamente al modelo. En la etapa final, se aplicaron estrategias de NLP al reporte de radiología para contabilizar y medir la frecuencia de cada palabra. Este último proceso permitió asignar una posición a las palabras más frecuentes, las cuales fueron identificadas como palabras claves en cada agrupamiento. La evaluación de este enfoque fue realizada por un grupo de radiólogos certificados, los cuales revisaron las palabras claves resultantes y las imágenes asociadas. El modelo obtuvo un 0.8109 de *accuracy*, mostrando resultados coherentes con respecto a la imagen y las etiquetas asociadas.

Otros modelos se caracterizan por utilizan estrategias de NLP para extraer información relevante de los reportes y realizar la clasificación de las imágenes utilizando una CNN. Por ejemplo, Rajpurkar et al. [77] propusieron la red *ChextNet* que permite clasificar características de neumonía en imágenes de rayos X. La arquitectura consta de una red DenseNet que incluye 121 capas densas de convolución. El modelo fue entrenado con los datos de *ChestX-ray 14* [68] considerando solamente las imágenes que presentaban características de neumonía para establecer una nueva clasificación como positiva, mientras que al resto se definió como negativa. Para interpretar las predicciones del modelo utilizaron mapas de calor, lo que permitió visualizar las áreas afectadas en la imagen.

Por otra parte, Dong et al. [58] utilizaron los modelos VGG y ResNet con el objetivo de clasificar imágenes de rayos X automáticamente. Para esta tarea se establecieron 10 etiquetas pre-definidas de normal, crecimiento de pulmón, aortosclerosis, crecimiento de corazón y otras más. Las etiquetas fueron extraídas de reportes médicos pertenecientes al PACs-2, mediante estrategias de NLP para formar agrupaciones. Posteriormente, se identificaron las agrupaciones más relevantes para es-

tablecer las etiquetas y asociarlas a una enfermedad. Este modelo obtuvo un 97 % de *accuracy*.

A su vez, Rubin et al. [27] implementaron una arquitectura basada en DenseNet-121 y la base de datos de MIMI-CXR. Este modelo utilizó una estrategia para etiquetar las imágenes (laterales y frontales de tórax) considerando las secciones de *impressions* y *findings* del reporte médico por medio de la herramienta NeoBio[68],[92]. El modelo logró alcanzar un 72 % de *accuracy* al predecir enfermedades de la vista frontal y un 66 % de *accuracy* en la vista lateral.

2.2.1. Arquitecturas orientadas a la generación de reportes médicos

La tarea de generación automática de reportes médicos mediante imágenes de rayos X utilizando estrategias de DL, tiene como objetivo formar oraciones y/o párrafos bien estructurados partiendo de una representación visual [93]. Para cumplir con esta tarea, se han utilizado modelos que durante la etapa de entrenamiento identifican asociaciones entre las regiones visualmente más relevantes de la imagen con el texto médico asociado; esto con el propósito de generar un reporte médico. Principalmente los modelos más utilizados dentro de las arquitecturas propuestas involucran el uso de Redes Neuronales Convolucionales (CNNs), Redes Neuronales Recurrentes (RNN), particularmente modelos basados en *Long Short Term Memory* (LSTM) [7],[8] y recientemente los modelos basados en *Transformers* [6],[20],[21].

Para la evaluación de los modelos orientados a la generación de reportes médicos se utilizan las métricas orientadas a la generación de lenguage natural (NLG), que son las más utilizadas para tareas como subtítulos en imágenes, traducción, *Question-Answering*, por nombrar algunas [7]. Las métricas NLG incluyen:

- *Bilingual Evaluation Understudy* (BLEU) [94].
- *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) [95].
- *Metric for Evaluation of Translation with Explicit ORdering* (METEOR) [96].
- *Consensus-based Image Description Evaluation* (CIDEr) [97].

La finalidad de ponderar el desempeño mediante estas métricas es calcular la precisión del modelo propuesto, observando la similitud o la diferencia entre los párrafos generados y las descripciones escritas por los radiólogos. Además, son utilizadas para comparar los modelos propuestos contra otros modelos de referencia, donde una alta *precision* y *recall* se refleja en puntuaciones más altas de BLEU, ROUGE y METEOR. Mientras que métricas como CIDEr tiene en cuenta la semántica del texto generado lo que permite ponderar a favor textos con diferentes palabras pero con el mismo significado.

Partiendo de lo anterior, se clasificaron las propuestas más relevantes en la literatura de acuerdo a su arquitectura en los siguientes enfoques: modelos basados en arquitecturas simples de CNN y RNN, modelos jerárquicos, modelos híbridos, y modelos mixtos, todos estos utilizados para la generación de reportes médicos.

Modelos basados en arquitecturas simples de CNN y RNN

Este tipo de estrategias divide el problema en dos partes. La primera parte se basa en un modelo CNN encargado de extraer las características de la imagen. Luego, se tiene una red RNN (LSTM, Bi-LSTM o GRU) que genera los enunciados o sentencias palabra por palabra. La figura 2.2 muestra una arquitectura general de este tipo de modelos. Un ejemplo de esta arquitectura lo presentó Hasan et al. [66] en donde una red VGG-19, junto con un red LSTM que incluyó un mecanismo de atención, recibió secuencialmente las características de la imagen y la descripción

asociada a ella.

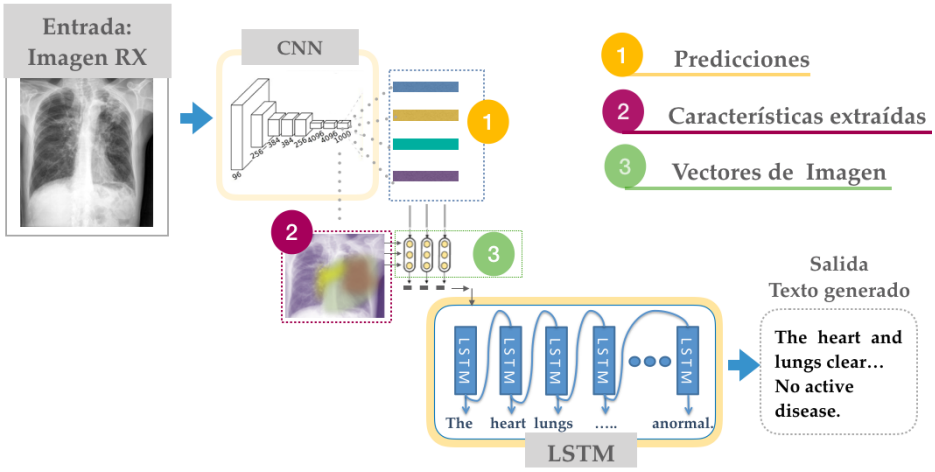


Figura 2.2: Modelos basados en arquitecturas simples de CNN y RNN.

Los modelos propuestos por Shin et al. [78] se basaron en una red GoogleNet en conjunto con una red LSTM [98] o GRU [99]. Aunque el enfoque con el modelo LSTM se caracterizó por ser un entrenamiento sencillo, el modelo de GRU obtuvo un mejor desempeño en la generación de etiquetas. En ambas propuestas se integró un vector de contexto que procesó las características de la imagen y su descripción, formando la entrada a la LSTM junto con el reporte asociado. Este modelo permitió describir la enfermedad, severidad y los órganos afectados.

El modelo denominado MDNET propuesto por Zhang et al. [100] estableció un mapeo multimodal directo entre las imágenes médicas y los reportes médicos. MDNet incluyó un modelo basado en ResNet con variaciones en las conexiones residuales que permitieron mejorar las características multi-escala. Para el modelo de lenguaje se adoptó una red LSTM que integró un mecanismo de atención. Lo anterior, contribuyó a la alineación de las características visuales con las palabras de las

oraciones obteniendo mapas de atención más nítidos.

Utilizando una arquitectura de ResNet y un LSTM, Wang et al. [53] propusieron un modelo denominado TieNet. Su enfoque se caracterizó por integrar una atención multinivel para la clasificación y generación del reporte. La arquitectura abarcó todo el proceso de principio a fin para incorporar representaciones del texto y la imagen. Aunado a ello, se generaron mapas de calor de las características relevantes lo que facilitó visualizar las predicciones.

Gasimova et al. [82] propusieron dos modelos, uno basado en la red ResNet y otro en la red VGG, para realizar una clasificación de la imagen en múltiples etiquetas (generadas con la herramienta de *Manual Medical Subject Heading*). Para la generación de un texto estructurado implementaron un modelo LSTM.

Gu et al. [59] utilizaron el modelo ResNet para extraer las características más relevantes de la imagen y clasificarlas en múltiples etiquetas. Posteriormente, los mapas de características se utilizaron para relacionar la información espacial y semántica. Para la generación de texto se adoptó una red LSTM, la cual recibió como entrada la predicción del vector con la información espacial y semántica de la imagen.

Una arquitectura basada en la red Inception-V3 y LSTM fue propuesta por Singh et al. en [88]. En su estrategia se optó por pre-entrenar el modelo de lenguaje con *RadGlove embedding* [101] para lograr un dominio del contexto de radiología.

La tabla 2.2 muestra los resultados obtenidos de estas arquitecturas en las métricas NLG orientadas a la generación de lenguaje, así como bases de datos utilizadas en cada una.

Modelos Jerárquicos

Los modelos jerárquicos se distinguen por la forma organizada de generar el texto. En la etapa inicial, se obtiene la información de la imagen por medio de una CNN. Posteriormente, una RNN recibe la información de la imagen para establecer la representación de la sentencia o tema que debe ser generado. A continuación, otra RNN recibe la representación de la sentencia para generar el texto palabra por palabra. Esta estructura permite dar pauta a la secuencia del texto que debe ser generado, la figura 2.3 muestra un esquema general de este tipo de arquitecturas.

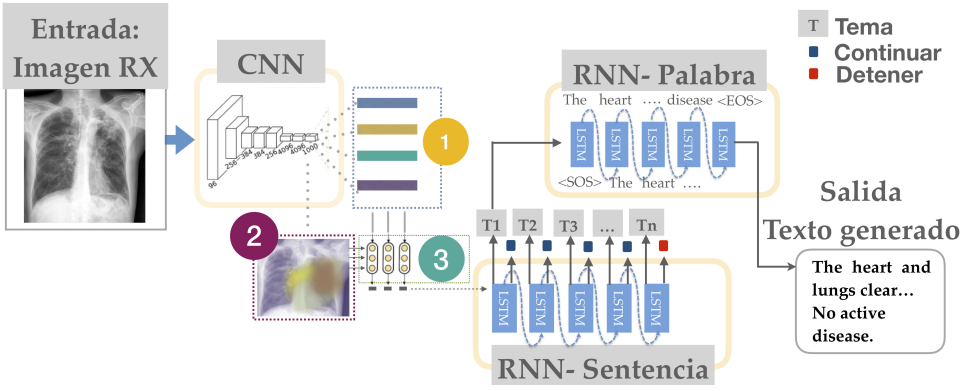


Figura 2.3: Arquitectura basada en modelos Jerárquicos. El (1) identifica las predicciones , (2) las características extraídas y (3) los vectores de imagen.

Considerando este enfoque, Jing et al. [9] propusieron un modelo que permitió desempeñar las tareas de predicción de etiquetas y la generación de las sentencias o párrafos de acuerdo a una etiqueta principal. La primera parte del modelo realizó la predicción de las etiquetas a través de una red VGG. Posteriormente, las características visuales aprendidas alimentaron a una red multi-clasificador (MLC), que generó etiquetas correspondientes a las características semánticas de la imagen.

Otro enfoque es el propuesto por Xue et al. [80], el cual consistió en una arquitectura que permitió generar un reporte de radiología automáticamente de principio a

fin. El modelo procesó simultáneamente la vista frontal y lateral de la imagen mediante un modelo ResNet. Posteriormente, el modelo generador de texto (basado en un Bi-LSTM), generó la primera palabra de la sentencia y la sentencia se fue produciendo palabra por palabra. Una vez que la sentencia es completada, el modelo tomó como entrada la sentencia generada y las características globales de la imagen durante varias iteraciones hasta obtener un párrafo de los hallazgos encontrados.

Por su parte, Harzig et al. [81] propusieron un modelo que clasificaba la imagen como normal o anormal y posteriormente generaba su descripción. La estrategia de aprendizaje multi-tarea facilitó un entrenamiento de principio a fin, donde las representaciones de la imagen y la generación de texto se llevaron paralelamente. Su arquitectura fue compuesta por un modelo ResNet, un mecanismo de atención, un modelo LSTM generador del tema principal y dos redes LSTM como generadores de la sentencia asociada al tema principal. Las sentencias fueron generadas palabra por palabra.

Considerando una estrategia similar a [81], Xie et al. [79] propusieron el modelo *Attention-based Abnormal-Aware Fusion Network (A3FN)*. La estrategia incluyó una CNN y un modelo jerárquico de LSTMs con un mecanismo de atención. La diferencia principal con [81], es que utilizaron un módulo de compuerta que permitió clasificar a la imagen como normal o anormal y posteriormente generar un reporte médico.

Mediante una estrategia jerárquica recurrente con un mecanismo de atención, Yin et al. [89] propusieron una arquitectura compuesta por una DenseNet y LSTMs. A diferencia de los otros modelos, esta propuesta incluyó un mecanismo de coincidencia que permitió mapear los vectores del tema principal y las sentencias en un

mismo espacio semántico.

La estrategia propuesta por Yuan et al. [76] fue similar a [89], pero ellos utilizaron una red pre-entrenada ResNet con imágenes de rayos X que permitió reconocer hasta 14 enfermedades diferentes. Además, para robustecer la descripción semántica extrajeron conceptos médicos basados en reportes de radiología y los conceptos médicos más frecuentes de las imágenes de rayos X. Esta estrategia permitió generar múltiples sentencias estructuradas con una descripción acorde a las características de la imagen.

Liu, et al. [71] propusieron un modelo basado en DenseNet y LSTM jerárquico con atención que permitió generar una secuencia de vectores. A diferencia de otros modelos jerárquicos con atención, utilizaron *Reinforcement learning* (RL) [102] para mejorar sus resultados. Además, el incluir un mecanismo de atención en la LSTM, lograron identificar las características relevantes mediante mapas de calor en la imagen.

La arquitectura propuesta por Xue [103] consistió en un modelo de tipo codificador - decodificador (*encoder-decoder*). La parte del codificador consitió en una red ResNet-152 pre-entrenada que era alimentada con dos imágenes de rayos X. La salida de la ResNet-152 pre-entrenada se conectaba a un decodificador (Bi-LSTM) para obtener una representación semántica. Posteriormente, ambas representaciones se combinaban y se representaban en un solo vector. El vector obtenido era enviado como entrada a una red RNN multimodal que generaba la próxima sentencia. Este proceso se repitió hasta generar todo el texto del reporte.

Un mecanismo multi-atención que potenciba el mapeo de las oraciones a la re-

presentación de las características de la imagen en las tareas de generación de párrafos es el propuesto por Huang et al. [28]. El modelo utilizó una red ResNet y una estructura jerárquica LSTM de tres sub-módulos. El primer módulo se encargó de generar oraciones. El segundo módulo fusionaba la información de los antecedentes del paciente. El último módulo fue utilizado para la generación de palabras de acuerdo a cada vector generado de los modulos anteriores.

El enfoque propuesto por Jing [75], denominado *Co-operative Multi-Agent System*, basó su arquitectura en una red ResNet pre-entrenada, un módulo de *findings* y otro módulo de *impressions*. La red ResNet analizaba y extraía las características de la imagen; posteriormente, el módulo de *findings* examinaba diferentes áreas de la imagen y generaba una descripción de lo encontrado. Cuando el módulo de *findings* terminaba, el módulo de *impressions* generaba las conclusiones de acuerdo a los hallazgos encontrados. Para generar el texto adecuadamente, los autores utilizaron 3 agentes jerárquicos compuestos por LSTMs denominados: *Planner* (PL), *Normality Writer* (NW) y *Abnormality Writer*(AW). El primer agente PL, identificaba si el área contiene alguna anormalidad y enviaba la información al agente correspondiente. Los agentes NW y AW recibían la información y realizaban la descripción asociada a la imagen.

El modelo propuesto por Tian et al. [85] utilizó una arquitectura jerárquica de BiLSTMs, Auto-econders (AE) y un mecanismo de atención que permitió la generación de sentencias estructuradas. El mecanismo de atención relacionaba las características visuales y el texto, adquiriendo predicciones precisas de las secciones de *findings* e *impressions*. La extracción de las características visuales se obtuvó mediante una CNN. Además, la CNN compartía las características relevantes de la imagen con el mecanismo de atención que asociaba al texto con la imagen. Para la genera-

ción del reporte médico, primero se formaba una representación del tema principal a través de un red BiLSTM y posteriormente se decodificaba por una red AE que producía la sentencia del tema, palabra por palabra.

Modelos Híbridos

Los modelos híbridos se caracterizan por generar el texto utilizando auxiliares de plantillas o prefijos para autocompletar el texto generado. La Figura 2.4 presenta un esquema general de este tipo de modelos.

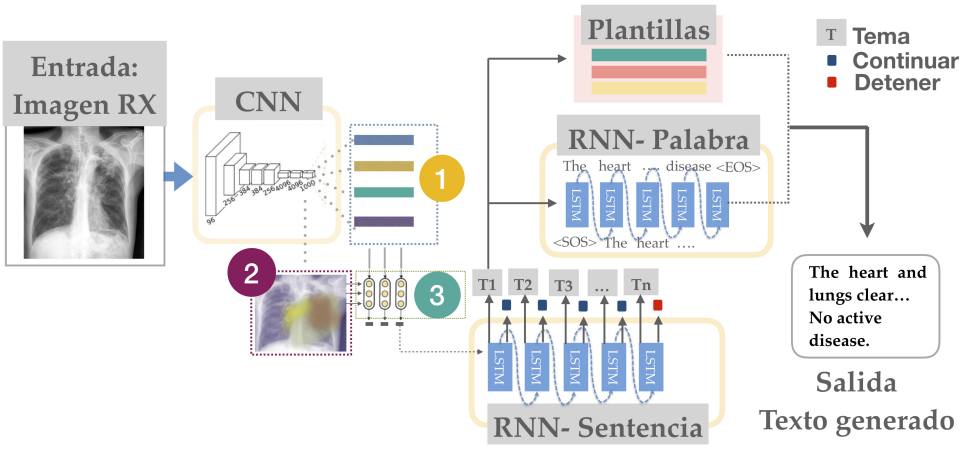


Figura 2.4: Arquitectura basada en modelos Híbridos. El número (1) identifica las predicciones , (2) las características extraídas y (3) los vectores de imagen.

Li, et al. [67] propusieron un modelo capaz de generar un reporte médico automáticamente mediante una plantilla pre-definida y el texto con las descripciones de la imagen. La estrategia incluyó un modelo DenseNet y una RNN con un mecanismo de atención que permitió mejorar la generación de texto.

El enfoque denominado *Knowledge-driven Encode, Retrieve, Paraphrase* (KERP), propuesto por Li, et al. [74], se caracterizó por generar el texto basándose en una

plantilla pre-definida. La arquitectura consistió en una red DenseNet y un *Graph Transformer* (GTR) el cual generaba una representación gráfica de la información visual en términos de conceptos médicos y sus relaciones. Mediante un módulo de recuperación se decodificaba la representación gráfica, la cual se generaba como una plantilla de secuencia de palabras, las cuales posteriormente eran las predicciones que conformaban al reporte.

CLinicAl Report Auto-completion (CLARA) propuesto por Biswal, et al. [83], generaba el reporte médico sentencia por sentencia utilizando términos clínicos o sentencias parciales definidas en el área médica. Su arquitectura se conformó por 4 módulos. El primer módulo recibía una imagen de rayos X y por medio de una arquitectura DenseNet pre-entrenada con ChestX-ray8 se obtenía la representación de las características. Paralelamente el segundo módulo creaba un repositorio que recibía como entrada los reportes médicos y seleccionaba diversas sentencias, su representación y frecuencia. El tercer módulo permitía controlar si el informe era generado automáticamente o de manera interactiva, donde un especialista podía manipular la información del reporte utilizando palabras clave definidas o editando directamente el texto. Finalmente, el cuarto módulo extraía las sentencias más relevantes de los prototipos del repositorio, y enviaba la información al modelo secuencial que permitía modificar la sentencia de acuerdo a la representación de la entrada, las palabras clave y el texto predefinido. Para la generación de texto se utilizó un modelo secuencial constituido por dos capas BiLSTM y 3 capas apiladas de LSTM.

Modelos Mixtos

En esta clasificación se distinguen arquitecturas muy particulares y diferentes a las anteriores orientadas a la generación de reportes médicos mediante imáge-

nes de rayos X. Debido a la gran diversidad de las arquitecturas incluidas en esta clasificación, en la Fig. 2.5 solo se muestra el modelo propuesto por Spinks , et al. [90], en donde se propone una metodología compuesta por dos fases, la fase de entrenamiento y la fase de inferencia. En la fase de entrenamiento, el modelo de *Adversarially Regularized Autoencoders* (ARAE) [104] aprendía el diagnóstico médico y etiqueta correspondientes a la imagen de rayos X asociada. Una vez que se obtenían estas representaciones, se entrenaba el modelo StackGAN [105] basado en una *Generative Adversarial Networks* (GAN) [106]. El propósito de esta GAN era recibir la salida generada del ARAE en el paso anterior y la imagen de rayos X correspondiente al diagnóstico; esto permitía que el modelo aprendiera a generar una imagen rayos X con las distribuciones correctas de acuerdo a la entrada de texto. Finalmente, se entrenaba el modelo de CNN, con el objetivo de que aprendiera a realizar el proceso inverso de imagen a texto. En la fase de inferencia se trabajó con imágenes de rayos X reales y se obtenía su diagnóstico; además, de acuerdo al diagnóstico producido, se tenía la opción de recrear la imagen de rayos X para validar el resultado del modelo.

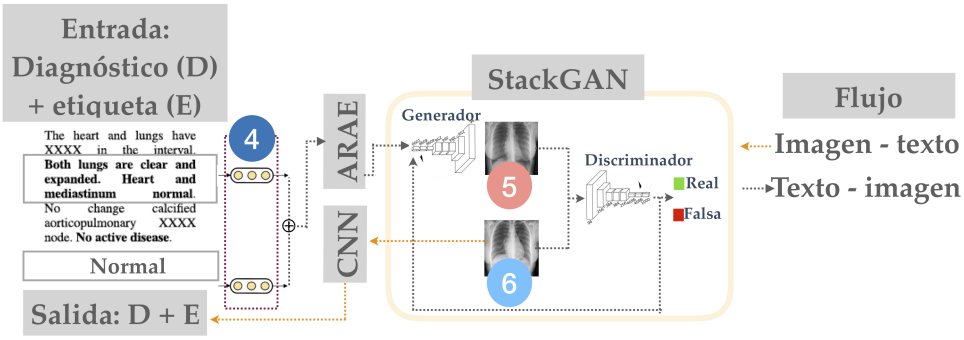


Figura 2.5: Arquitectura basada en modelos Mixtos. El (4) representan los vectores de texto, (5) la salida con la imagen generada por el modelo y (6) la imagen real de entrada.

Por otra parte, Xiong, et al. [84] propusieron un enfoque utilizando la red DenseNet previamente entrenada, que asociaba un área específica de la imagen (*bounding*

box) con una clase predefinida utilizando el algoritmo de Grad-CAM [107]. Luego, un modelo basado en *Transformers* [1], constituido por 3 sub-capas apiladas de módulos con atención y un módulo *feed forward*, generaba el reporte médico.

Li, et al. [86] propusieron una arquitectura que clasificaba y localizaba diferentes enfermedades en una imagen de rayos X; posteriormente, generaba las sentencias correspondientes a la clasificación determinada. Para la tarea de clasificación se utilizó un modelo DenseNet pre-entrenado con ChestX-ray8. Para la tarea de detección se utilizó un modelo ResNet pre-entrenado con Imagenet y se aplicó la estrategia de Grand-GAMs [107]. La generación del reporte se realizó con diferentes LSTMs, donde para cada clase de enfermedad se utilizaba un par de LSTMs con la finalidad de obtener una mayor congruencia en el texto generado.

A su vez Gajbhiye, et al. [87] presentaron un enfoque denominado *Multilevel Multi-Attention based encoder-decoder*, el cual combinaba la atención visual y la atención textual para la generación de reportes médicos. El modelo se conformó por una red VGG y un LSTM. Primero se codificaba el contexto visual y textual, posteriormente estas representaciones fueron fusionadas para alimentar una red LSTM que contaba con un mecanismo de atención. Paralelamente, la representación textual del reporte alimentaba una red BiLSTM con atención. Posteriormente la salida obtenida del Bi-LSTM y la salida del LSTM eran concatenadas para realizar una predicción y generar un reporte médico palabra por palabra.

Zhang, et al. [91] propusieron una arquitectura basada en una Red Convolucional Gráfica (GCN). Mediante una DenseNet pre-entrenada con CheXpert se extraían las características de las imágenes. Posteriormente, las características alimentaban a la GCN mediante un mecanismo de atención. Del GCN se especificaban dos rutas,

una para la clasificación de las características encontradas en la imagen y otra para la generación del reporte. Para la generación de reporte se utilizaron dos niveles de LSTM, uno para la generación del tema principal y otro para la generación de palabras. Los autores propusieron una nueva métrica denominada *Medical Image Report Quality Index* (MIRQI), que permitió medir la exactitud de las enfermedades positiva o negativamente así como los atributos mencionados en el reporte médico.

La arquitectura propuesta por Alfarghaly et al. [108] se conformó por una red de tipo codificador - decodificador (*encoder-decoder*), donde la parte del codificador consistió en la red *ChexNet* [77] y el decodificador en un modelo pre-entrenado de *DistilBERT*, propuesto por [109].

Tomando en consideración los enfoques analizados anteriormente se puede inferir que la automatización del diagnóstico médico a través de imágenes médicas, es una tarea compleja de gran relevancia. Aunque en la actualidad los modelos propuestos arrojan un buen desempeño, aún existen aspectos relevantes que faltan por cubrir. Por ejemplo, la obtención de una mejor representación de la información visual así como la generación de textos precisos y entendibles para los expertos en el área. Ambos aspectos contribuyen directamente a una mayor precisión en los resultados. Además, aunque los modelos LSTM apoyados de mecanismos de atención han demostrado obtener resultados competitivos, la falta de flexibilidad para adecuar cada unidad de memoria ocasiona una gran limitación en las arquitecturas.

Recientemente novedosos enfoques que incluyen modelos de *Transformers* han demostrado un mayor desempeño en las tareas de NLP. Esta estructura, propuesta por Vaswani et al. [1], es basada en una arquitectura de tipo codificador-decodificador (*encoder-decoder*). El modelo se caracteriza por contribuir a una mayor paralelización

y eficiencia en su ejecución gracias a los mecanismos de atención (*self attention*). Al no utilizar unidades recurrentes, el modelo asigna una posición relativa a la representación de cada palabra para identificar su secuencia e importancia en la oración a un menor costo computacional.

En el área de análisis de imágenes médicas, la aplicación de los *Transformers* sigue siendo explorada para tareas relacionadas al diagnóstico médico, como la propuesta por Li et al. [110]. En donde se presentó un enfoque para extraer características relevantes asociadas a imágenes de tórax con anomalías mediante reportes médicos. Este enfoque utilizó un modelo de *Transformer* denominado ERNIE, y por medio de un *fine-tuning* se transfirió información al modelo acerca de los signos de anomalías en el Tórax. Aunque el modelo no incluyó la extracción de información directamente de las imágenes en su arquitectura, el uso de ERNIE mediante un *fine-tuning* contribuyó a reducir el problema de la insuficiencia de los datos.

Otros enfoques como [6], [20],[111], propusieron estrategias basadas en mecanismos de atención, como *Transformers* que integran en su estructura el uso de GRU, LSTM o memoria auxiliares basadas en recurrencias, y redes basadas en grafos las cuales influyeron directamente en la ponderación a determinadas frecuencias de texto. Al emplear este tipo de estrategias, el texto producido tiende a ser limitado, dando la impresión de que el modelo genera secuencias de texto idénticas para varias imágenes. Además, estas estructuras son sumamente complejas demandando un alto costo computacional. Las tablas 2.2 y 2.3 muestran los trabajos mencionados indicando la base de datos que emplearon, las arquitecturas de los modelos implementados, la tarea principal que se realizó, los resultados que reportaron resaltando los modelos que obtuvieron el mejor puntaje, y a manera de resumen, las principales observaciones que se identificaron en su análisis.

Tabla 2.2: Modelos de DL propuestos para la generación de reportes médicos mediante imágenes de radiología

Referencia	Bases de datos	Arquitectura	Tarea	Accuracy	AUC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	M	ROUGE	CIDEr	Observaciones
Rajpurkar et. al [77]	ChestX-ray14	CheXNet	Clasifica imagen	0.435	-	-	-	-	-	-	-	-	Produce mapas de calor
													No genera reporte
Dong et. al.[58]	PACs	K-medoids	Extrae etiqueta										BD privada
		VGG-16	Imagen normal o anormal	0.820	-	-	-	-	-	-	-	-	No genera reporte
		ResNet	Clas. de enfermedades	0.829	-	-	-	-	-	-	-	-	No genera reporte
Rubin et. al[27]	MIMIC-CXR	NeoBio	Genera etiqueta										No genera reporte
		DenseNet-121	Relaciona Imágenes/texto	-	0.721/0.668	-	-	-	-	-	-	-	Vista frontal / lateral
			Clasifica imagen										
Hasan et. al.[66]	ImageCLEF	VGG-19-LSTM	Extraer características imagen	0.1208	-	0.3211	-	-	-	-	-	-	Usa mecanismos de atención
	CAPTION 2017		Relaciona texto/imagen										
			Genera una sentencia										
Shin et.al. [78]	IU X-ray	GoogleNet-GRU	Genera etiqueta	0.698	-	0.785	0.144	0.047	0.0	-	-	-	Detecta enfermedad
			Relaciona imágenes/ etiqueta										Genera sentencia
			Describe contexto										Usa sección *findings*
Zhang et. al. [100]	BCDIR	ResNet-LSTM	Extrae características imagen	-	-	**0.912**	0.829	0.75	0.677	0.396	0.701	0.204	BD privada y pequeña
			Relaciona imagen/texto										Posible sobre-entrenamiento
			Genera sentencia										
Wang et al. [53]	ChestX-ray14	ResNet-LSTM	Clasifica imagen	-	0.748	0.286	0.159	0.103	0.073	-	0.107	0.226	Texto no estructurado
	IU X-ray	ResNet-LSTM	Relaciona texto/imagen	-	0.798	-	-	-	-	-	-	-	Usa mecanismo de atención
			Genera múltiples sentencias										Texto no estructurado
													Usa sección *findings*
Gasimova et al. [82]	IU X-ray	ResNet-LSTM	Clasifica imagen	**0.994**	-	0.667	0.471	0.268	0.159	-	-	-	Evalúa reporte
		VGG-LSTM	Relaciona texto/imagen	-	-	0.069	0.023	0.07	0.01	-	-	-	
			Genera una sentencia										
Gu et al. [59]	PACs-3	ResNet-LSTM	Clasifica imagen	-	-	0.762	0.716	0.681	0.651	**0.436**	0.490	0.809	Texto no estructurado
			Relaciona texto/imagen										Evalúa reporte
			Genera múltiples sentencias										
Singh et al.[88]	IU X-ray	InceptionV3-LSTM	Extraer características imagen	-	-	0.374	0.224	0.152	0.109	0.163	0.307	0.359	Texto no estructurado
			Relaciona texto/imagen										
			Genera múltiples sentencias										Usa sección *findings+impressions*
Jing et. al.[9]	IU X-ray	VGG-LSTM-	Extraer características imagen	-	-	0.517	0.386	0.306	0.247	0.217	0.447	0.327	Modelo jerárquico/atención
	PEIR Digital	MLC	Predicción de etiquetas	-	-	0.300	0.218	0.165	0.113	0.149	0.279	0.329	Usa sección *findings + impression*
	Library		Genera múltiples sentencias										Texto no estructurado
Xue et al.[80]	IU X-ray	ResNet-BiLSTM	Extraer características imagen	-	-	0.464	0.358	0.270	0.195	0.274	0.366	0.343	Modelo recurrente con atención.
			Relaciona texto/imagen										
			Genera múltiples sentencias										Texto no estructurado
													Usa sección *findings*
Harzig et al. [81]	IU X-ray	ResNet-LSTM	Clasifica imagen	-	-	0.373	0.246	0.175	0.126	0.163	0.315	0.359	Genera texto jerárquicamente
			Relaciona texto/imagen										Texto no estructurado
			Genera múltiples sentencias										Usa sección *findings*
Xie et al. [79]	IU X-ray	CNN-LSTM	Clasifica imagen	0.132	-	0.443	0.337	0.236	0.181	-	0.347	0.374	Genera texto jerárquicamente
			Relaciona texto/imagen										Texto no estructurado
			Genera múltiples sentencias										Usa sección *findings+impressions*
Yin et al. [89]	IU X-ray	DensNet-LSTM	Clasifica imagen	-	-	0.445	0.292	0.201	0.154	0.175	0.344	0.342	Genera texto jerárquicamente
			Relaciona texto/imagen										Usa mecanismo de atención
			Genera múltiples sentencias										Usa sección *findings+impressions*
Yuan et al. [76]	IU X-ray	ResNet-LSTM	Clasifica imagen	-	-	0.529	0.372	0.315	0.255	0.343	0.453	-	Genera texto jerárquicamente
			Relaciona texto/imagen										Genera texto no estructurado
			Genera múltiples sentencias										Usa sección *findings+impressions*
													Usa mecanismo de atención
Liu et al. [71]	IU X-ray	DensNet-LSTM	Clasifica imagen	0.916	-	0.369	0.246	0.171	0.115	-	0.359	1.490	Genera texto Jerárquicamente
	MIMIC-CXR	DensNet-LSTM	Relaciona texto/imagen	0.834	-	0.352	0.223	0.153	0.104	-	0.307	1.153	Texto estructurado
			Genera múltiples sentencias	-	-							-	Usa sección *findings*
Xue et al. [103]	IU X-ray	ResNet 152-BiLSTM	Extraer características imagen	-	-	0.489	0.340	0.252	0.195	0.230	0.478	0.565	Texto no estructurado
			Relaciona con texto										Usa mecanismo de atención
			Mapas de calor en imagen										Usa sección *findings*

Tabla 2.3: Modelos de DL propuestos para la generación de reportes médicos mediante imágenes de radiología

Referencia	Bases de datos	Arquitectura	Tarea	Accuracy	AUC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	M	ROUGE	CIDEr	Observaciones
Huang et al. [28]	IU X-ray	ResNet-LSTM	Clasifica imagen Clasifica imagen Genera múltiples sentencias	-	-	0.476	0.340	0.238	0.169	-	0.347	0.297	Texto no estructurado Usa sección *findings + impressions* Usa mecanismo de atención.
Jing et al. [75]	IU X-ray CX-CHR	ResNet-LSTM	Clasifica imagen Relaciona imagen/texto Genera múltiples sentencias	- -	- -	0.464 0.693	0.301 0.626	0.210 0.580	0.154 0.545	- -	0.362 0.661	0.275 2.900	Texto no estructurado. Usa sección *findings* Usa mecanismo de atención.
Tian et al. [85]	IU X-ray	CNN-BiLSTM-AE	Clasifica imagen Genera múltiples sentencias	-	-	0.882	**0.874**	**0.867**	**0.860**	-	**0.929**	-	Genera texto jerárquicamente. Texto estructurado Usa sección *findings + impressions*. Usa mecanismo de atención.
Li et. al.[67]	IU-XRay CX-CHR	DenseNet-RNN DenseNet-RNN	Extraer características imagen Relaciona texto/imagen Genera texto	- -	- -	0.438 0.673	0.298 0.587	0.208 0.530	0.151 0.486	- -	0.322 0.612	0.343 2.895	Híbrido plantilla/texto. Usa sección *findings*
Li et al. [74]	IU X-ray CX-CHR	DenseNet-GTR DenseNet-GTR	Clasifica imagen Genera texto	- -	- -	0.482 0.673	0.325 0.588	0.226 0.532	0.162 0.473	- -	0.339 0.618	0.280 2.850	Modelo híbrido Usa plantilla. Usa sección *findings*. Usa mecanismo atención
Biswal et al. [83]	IU X-ray	DenseNet-BiLSTM-LSTM	Clasifica imagen Genera texto	0.871	0.796	0.471	0.324	0.214	0.199	-	-	0.359	Utiliza prefijos. Usa sección *findings* Usa plantilla
Spinks et al. [90]	PACs-3	GAN-ARAE-CNN	Texto/ imagen	0.906	**0.948**	0.490	0.350	0.250	0.180	0.270	0.400	0.600	Usa mecanismo de atención
Xiong et al. [84]	IU X-ray	DenseNet-Transformer	Clasifica imagen Relaciona imagen/texto Genera múltiples sentencias	-	-	0.350	0.234	0.143	0.096	-	-	0.323	Texto no estructurado. Usa sección *findings* Usa mecanismo de atención
Li et al. [86]	IU X-ray	DenseNet-LSTM	Clasifica imagen Genera múltiples sentencias	-	-	0.419	0.280	0.201	0.150	-	0.371	0.553	Texto no estructurado. Usa sección *findings + impressions* Usa mecanismo de atención.
Gajbhiye et al. [87]	IU X-ray	VGG-BiLSTM	Clasifica imagen Genera múltiples sentencias	-	-	0.500	0.380	0.317	0.278	0.281	0.440	1.067	Usa no estructurado. Usa sección *findings + impressions* Usa mecanismo de atención.
Zhang et al. [91]	IU X-ray	DenseNet-GCN-LSTM	Clasifica imagen Genera múltiples sentencias	-	-	0.441	0.291	0.203	0.147	-	0.367	0.304	Texto no estructurado. Usa sección *findings+impressions* Usa mecanismo de atención
Alfarghaly et al. [108]	IU X-ray	ChexNet-Transformer	Extraer características imagen Relaciona Semántica Genera múltiples sentencias	-	-	0.387	0.245	0.166	0.111	0.164	0.289	0.257	Texto no estructurado Usa sección *findings+impressions* Usa mecanismo de atención

2.2.2. Aumento de datos, pre-procesamiento y modelos pre-entrenados

Las técnicas de aumento de datos (DA) se aplican con la finalidad de cubrir varios propósitos como incrementar el tamaño del conjunto de datos, robustecer el entrenamiento del modelo aplicando variabilidad en los datos, reducir problemas relacionados con el sobre-entrenamiento o con el desbalance de las clases [8],[49],[112], [113]. Sin embargo, el escaso número de bases de datos públicas con imágenes médicas y reportes asociados es una de las principales razones por las cuales se adoptan sofisticados procedimientos que incluyan técnicas de DA. Debido a esta situación diferentes investigadores han buscado nuevas estrategias que contribuyan a robustecer las bases de datos médicas [63].

Una de las ventajas de utilizar las técnicas de DA para aumentar la diversidad de los datos de entrenamiento es obtener nuevos datos a partir de los existentes sin la necesidad de recopilar nuevos datos; esto permite reducir los problemas de sobre-entrenamiento en los modelos. La funciones de escalamiento, rotación, volteo, recorte y las variaciones en la intensidad de color son algunas de las estrategias DA más utilizadas en tareas que involucran el procesamiento de imágenes [77],[58],[78],[84],[114]. Otra alternativa son los modelos generativos, los cuales se han utilizado ampliamente como estrategia de DA cuando se cuenta con datos limitados o de poca calidad. Para aplicarla, es necesario entrenar al modelo para que aprenda a generar nuevos datos a partir de los originales [115],[116],[117],[118].

Para el análisis de conjuntos de datos que incluyen texto, diferentes enfoques de DA intentan capturar las mismas propiedades semánticas y sintácticas del texto original [119]. La inserción, sustitución, reemplazo y eliminación aleatoria de caracteres son algunas de las técnicas más utilizadas. Del mismo modo, las palabras pueden intercambiarse, eliminarse o sustituirse por sinónimos de forma aleatoria.

Otro enfoque implica el aumento contextual en el texto, donde se utilizan modelos lingüísticos enmascarados (MLM) como BERT para generar un nuevo texto basado en el contexto de una frase [120].

Conforme se han propuesto diversas arquitecturas para la tarea de generación de reportes médicos, como se ha mencionado anteriormente; el uso de grandes modelos de Lenguaje (LLMs) como BERT [2] o GPT [32], comienzan a ser explorados para generar datos sintéticos que permitan robustecer los entrenamientos de otros modelos de DL. Una de las ventajas de utilizar estos modelos para generar nuevos datos es la capacidad de aplicar contexto en las frases generadas, lo que permite incrementar el nivel de efectividad. Aunque esta estrategia a la fecha se ha utilizado para textos no especializados ni médicos, es necesario explorar su impacto en el área médica.

Para la generación de reportes médicos a partir de imágenes médicas, el preprocesamiento, limpieza de los datos y uso de modelos pre-entrenados, son las consideraciones más comunes que han dado buenos resultados en este tipo de tareas [7], [8]. Para cumplir con criterios de calidad en la información que entra al modelo, varios autores reportan eliminar determinadas secciones del conjunto de datos. Por ejemplo, Biswal et al. [83] eliminaron las imágenes duplicadas de MIMIC-CXR; Yuan et al. [76] eliminaron las imágenes con menos de 2 vistas de IU X-ray; Gu et al. [59] removieron las radiografías laterales de PACs-3; Gasimova et al. [82], eliminaron los reportes vacíos de IU X-ray; Xue et al. [80] y Singh et al. [88] removieron los reportes que no cuentan con las secciones de *impressions* y *findings* de IU X-ray. Además, Xue et al. [80] eliminaron los reportes que no disponían de dos imágenes completas de IU X-ray. Por su parte, Li et al. [86] removieron las imágenes y los reportes no relevantes a 8 enfermedades de pecho de IU X-ray.

Por otro lado, varios autores han optado por utilizar un modelo de CNN pre-entrenado con la base de datos ImageNet o alguna base de datos de imágenes médicas con el fin de inicializar los pesos del modelo de manera más certera [25],[26],[28], [53], [77],[66],[82],[88],[80],[81],[89],[103],[67], [74],[84],[86],[91], [121].

Para la extracción de características relevantes de la imagen, algunos trabajos como [122],[123],[124] utilizan una CNN así como la ubicación específica de la región. Otros trabajos como [5],[6],[125],[126] utilizan únicamente la CNN para la extracción de características.

En el área de radiología, el NLP se ha utilizado para extraer información de reportes médicos, librerías digitales o incluso de redes sociales. Lo anterior con la finalidad de proveer datos estructurados que puedan utilizarse por los modelos de DL [127]. Para el pre-procesamiento en el modelo de texto, uno de los procedimientos más comunes es remover todos los caracteres no alfabéticos y convertirlos a minúscula [9],[82],[89],[103],[85],[86]. Además del procedimiento anterior, otra estrategia consiste en filtrar todas las palabras incluidas en los reportes que no cumplan con determinado número de ocurrencias (frecuencia) [53],[59],[89],[76],[83],[91]. Estas palabras se definen como *UNK* y finalmente se establece un estado de *START* y *END* para definir el inicio y fin de cada oración o párrafo [53],[80],[76],[85],[86],[91].

Algunas propuestas incluyen un pre-procesamiento más específico, por ejemplo, Harzig et al. [81] clasificaron manualmente como normal o anormal cada sentencia del reporte médico asociado de IU X-ray. Además, para lograr una mejor representación del texto, utilizaron el algoritmo de Word2vec previamente pre-entrenado en Pubmed y Wikipedia [128].

Por otra parte, Li et al. [67] identificaron las sentencias con mayor número de ocurrencias en los reportes médicos de IU X-ray. Mediante este procedimiento, se establecieron 97 sentencias con ocurrencia mayor a 500. A su vez, Li et al. [74] definieron sentencias con descripciones anormales que cumplieran con determinado número de ocurrencias. Posteriormente, agruparon manualmente las sentencias con el mismo significado y seleccionaron las más frecuentes de cada grupo para establecerlas como 87 plantillas de IU X-ray y 362 plantillas de CX-CHR.

En algunas propuestas como [28] las secciones de *indications*, *impressions* y *findings* fueron concatenadas para establecerlas como etiquetas (*groundtruth* o GT) de IU X-Ray. En otros casos solo se consideraron las secciones de *impressions* y *findings* como en [88],[81],[89],[76],[86],[91].

Wang et al. [53] utilizaron el algoritmo de Word2Vec pre-entrenado en artículos de PubMed para obtener una mejor representación del texto. Por otra parte, Singh et al. [88] utilizaron el algoritmo de *GloVe* pre-entrenado con texto genérico de *Common Crawl*. También utilizaron un LSTM entrenado con *RadGlove*, el cual tiene 4.5 millones de reportes de radiología de *Stanford University*. Dong et al. [58] dividieron la sección de diagnosis en pequeñas cláusulas de PACs-2 y posteriormente agruparon las que son similares. Además, aplicaron el algoritmo de *K-medoids* con el objetivo de identificar los 10 agrupamientos más grandes y asignar una etiqueta a cada uno de ellos. Por su parte, Gasimova et al. [82] ajustaron los reportes recortando o rellenando cada uno de ellos para establecerlo a 32 palabras.

Las Tablas 2.4 y 2.5, muestran las diferentes estrategias de pre-procesamiento que se han implementado, así como el uso de modelos pre-entrenados, indicando

los autores que los emplearon. Como se puede observar el utilizar modelos pre-entrenados en *imagenet* para la parte visual es una de las estrategias más utilizada por diversos autores al momento de implementar el modelo extractor de características visuales. Aunque algunos autores han reportado el uso de modelos pre-entrenados con imágenes médicas, no se identifica una diferencia significativa en el desempeño. Sin embargo, en términos generales se ha comprobado que el utilizar modelos pre-entrenados tiene varias ventajas como: transferencia de conocimiento, Ahorro de recursos computacionales y tiempo, Menor cantidad de datos de entrenamiento requeridos, generalización mejorada y facilita la implementación del modelo.

Por otra parte, las estrategias de escalamiento, rotación y recorte siguen utilizándose por autores que buscan robustecer el entrenamiento del modelo ampliando la variabilidad de los datos. Esto permite que los modelos tengan más robustez al recibir datos de menor calidad. Otro de los aspectos importantes en los conjuntos de datos, es que la mayoría de los trabajos presentados utilizan la vista frontal y lateral de las imágenes de rayos X.

Para el pre-procesamiento en el texto, los autores reportan la limpieza de caracteres especiales, la conversión a minúsculas y existe cierta variabilidad en considerar la frecuencia de palabras en el texto. Al realizar la normalización del texto se ayuda a evitar la duplicación de *tokens* debido a diferencias de mayúsculas o minúsculas. La limpieza de los caracteres especiales pueden agregar ruido innecesario a los datos; la eliminación de estos caracteres contribuye a simplificar el texto y a reducir la complejidad del modelo. Por otra parte, al eliminar palabras que aparecen con poca frecuencia en el corpus de texto, se reduce el tamaño del vocabulario, lo que puede mejorar la eficiencia computacional y reducir la cantidad de datos necesarios para

Tabla 2.4: Pre-procesamiento y uso de modelos pre-entrenados en imagen

Pre-procesamiento en la imagen		
Descripción	**Particularidad**	**Usado en**
Aumento de datos	Escalamiento	[26, 28, 58, 85] [77, 80, 84, 103]
	Rotación	[84, 91]
	Recorte	[78, 84]
	Intensidad	[103]
Eliminación de datos	Vista Lateral	[59]
	Duplicidad	[71, 83]
	$<$ 2 vistas	[76]
	Datos no comunes	[86]
Modelos de CNN pre-entrenados		
Pre-entrenamiento	Imagenet	[26, 28, 53, 89] [80, 82, 103, 121] [66, 77, 86, 88]
	ChestX-ray8	[67, 74, 84, 91] [86]
	CheXpert	[86]
	CX-CHR	[81]
Fine-tuning	CX-CHR	[67, 74]
	PubMed Central Biomedical Image corpus	[66]
	PACs [25]	[26]

entrenar el modelo.

Finalmente otro aspecto a considerar es determinar la o las secciones del reporte médico. Generalmente la sección de *findings* o las secciones concatenadas de *findings-impressions* son las más utilizadas por los autores. Este aspecto es importante considerarlos al momento de comparar el desempeño de los modelos.

Tabla 2.5: Pre-procesamiento y uso de modelos pre-entrenados en texto

Pre-procesamiento en texto		
Descripción	**Particularidad**	**Usado en**
Palabra	Convertir a minúsculas	[9, 82, 85, 89, 103]
		[86]
	Eliminar caracteres	[85, 86, 89, 103]
		[9, 82]
	Considera aquellas ≥ 2	[59, 85]
	Considera aquellas ≥ 3	[53, 76, 89, 91]
	Considera aquellas ≥ 5	[83]
Sentencia	Clasifica	
	Norma/Anormal	[81]
	Ocurrencia ≥ 500	[67, 74]
Sección	Concatena *findings-*	
	Impressions-indications	[28]
	Concatena	
	findings-Impressions	[76, 81, 89]
	findings-Impressions	[86, 88, 91]
	Divide *Diagnosis*	[58]
Reporte	Elimina Incompleto	[71, 80, 82, 88]
Modelos de WE pre-entrenados		
Embbedings	Word2Vec	[53, 81]
	GloVe	[88]
	RadGlove	[88]

Capítulo 3

Marco Teórico

La generación de reportes médicos a partir de imágenes de rayos X ha sido una tarea de gran interés en el ámbito del diagnóstico médico. En los últimos años, los avances en modelos de DL han revolucionado esta área, proporcionando herramientas poderosas para mejorar el proceso de diagnóstico. Los modelos de DL aplicados a la generación de reportes médicos utilizan arquitecturas de tipo *encoder-decoder* que incluyen redes neuronales convolucionales (CNN) para extraer características relevantes de las imágenes de rayos X y redes neuronales recurrentes (RNN) como *Long Short-Term Memory* (LSTM), *Gated Recurrent Unit* (GRU) y *Transformers* para generar texto coherente y descriptivo. Recientemente las arquitecturas basadas en modelos de *Transformers* se han ido adaptando para realizar la extracción de características visuales obteniendo resultados competitivos.

La siguiente sección presenta de manera general los modelos DL esenciales para concebir nuevas tácticas capaces de generar texto de forma automática a partir de imágenes de rayos X.

3.1. Aprendizaje Profundo

El área de Aprendizaje Profundo (DL) es un subcampo del área de Aprendizaje Máquina (ML). Este subcampo ha cobrado mayor relevancia debido a los aportes realizados en diversas áreas como: visión por computadora, Procesamiento de Lenguaje Natural (NLP), reconocimiento de voz, videojuegos, por mencionar algunos [7]. Al igual que ML, los modelos de DL pueden abordarse como supervisados, no supervisados, semisupervisados o como aprendizaje con refuerzo (RL). El enfoque supervisado consiste en una función de mapeo que permita asociar al modelo una entrada X a una salida Y. En este tipo de modelos se encuentran el Perceptrón Multicapa (MLP), las Redes Neuronales Recurrentes (RNN), Redes Neuronales Convolucionales (CNN), entre otros. Por el contrario, el aprendizaje no supervisado no cuenta con una etiqueta Y que permita asociar la entrada X. Por lo tanto, debe identificar características relevantes así como su relación en la distribución en X. Un ejemplo de estos modelos son los *autoencoders* (AE) [129]. En el aprendizaje semi-supervisado, se cuenta con un conjunto de datos etiquetados y no etiquetados. Los datos etiquetados permiten al modelo aprender a identificar características relevantes de los datos que son asociados a determinada etiqueta; y partiendo del aprendizaje obtenido se empiezan a agrupar los datos no etiquetados de acuerdo a sus características. Por último el aprendizaje por refuerzo, basa su estrategia en incentivos o penalizaciones de acuerdo a las predicciones del modelo [129].

3.1.1. Redes Neuronales Convolucionales

Dentro del área de DL, una de las arquitecturas más utilizadas es la CNN; la cual fue inspirada del mecanismo que hace posible la percepción visual en los seres vivos [130]. En 1959, Hubel and Wiesel, identificaron que las células más superficiales localizadas en la corteza visual de un gato, son responsables de detectar la luz

en los campos receptivos [130]. Inspirado por este descubrimiento, [131], propuso el *neocognitron*, lo que puede ser identificado como el predecesor de la CNN. En el año de 1990, Lecun publicó un artículo en donde estableció un marco de referencia de la CNN. En este artículo trata de la implementación de una CNN denominada **LetNet-5**, la cual clasifica dígitos escritos a mano [132].

Como otras redes neuronales, **LetNet-5** contiene múltiples capas y puede ser entrenada con el algoritmo de retropropagación *(backpropagation)* [132]. Una de las principales características de **LetNet-5**, es que permite obtener representaciones efectivas de la imagen original. Lo anterior hace posible reconocer patrones visuales directamente de los píxeles con un poco o nada de pre-procesamiento en los datos.

Desde el año 2006 se han desarrollado diferentes métodos para superar las arquitecturas propuestas de las CNN, [133–135]. Una de las arquitecturas más notables fue **AlexNet**, propuesta por Krizhevsky et. al. [133]. Básicamente es un modelo clásico de CNN, pero con mayor profundidad comparada con los modelos CNN antecesores, y presenta mejoras significativas en cuanto a las tareas de clasificación de imágenes. En general, la arquitectura utilizada en **AlexNet** es similar a la **LetNet-5**, una de sus diferencias es que utiliza 5 capas de convolución, de las cuales algunas son seguidas de una capa *max-pooling*. Posteriormente utiliza 3 capas totalmente conectadas *(fully connected)* y finaliza con una capa de salida con *1000-way softmax*. La figura 3.1 muestra una representación general de la arquitectura de **AlexNet**.

Con el éxito de **AlexNet**, se han propuesto más arquitecturas con el objetivo de obtener un mayor desempeño en tareas de procesamiento de imagen; entre los modelos más representativos se encuentran **VGGNet** propuesta por Jadenberg et. al [136], **GoogleNet** propuesta por Szegedy et. al [134] y **ResNet** propuesta por He et.

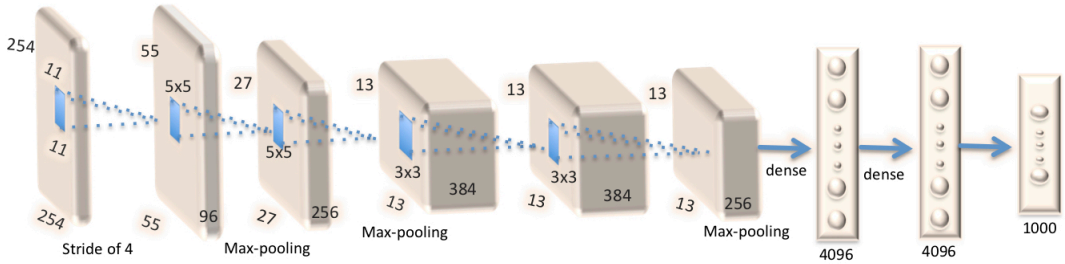


Figura 3.1: Arquitectura de AlexNet

al. [137].

Desde los orígenes de las CNN, una de las principales tendencias es la profundidad en las redes. Por ejemplo **ResNet** que ganó el *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) 2015, se caracteriza por tener 152 capas, lo cual la hace 20 veces más profunda que **AlexNet** y 8 veces más profunda que **VGGNet**. La figura 3.2 muestra la arquitectura general del modelo **VGGNet**, donde se apreciar las capas de convolución y sus tamaños.

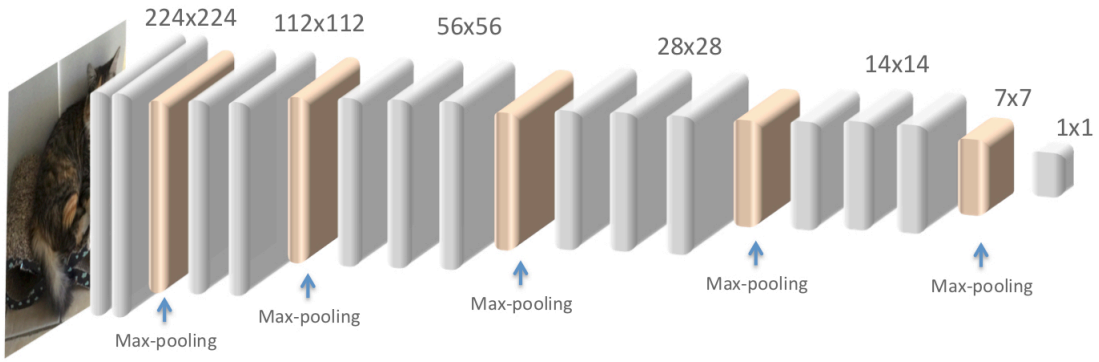


Figura 3.2: Arquitectura de VGG-16

Algunos de los modelos más representativos de CNN para las diversas tareas del área de visión por computadora como **AlexNet, VGGNet y ResNet**, basaron su arquitectura en los componentes básicos de **LetNet-5** [135]. Con el objetivo de com-

prender una arquitectura básica de un modelo de CNN, a continuación se abordarán brevemente sus principales componentes.

Capas de Convolución: Una capa de Convolución tienen como objetivo el aprender una representación de características del valor entrada. Esta compuesta por un número determinado de filtros *kernels*, los cuales permiten calcular un mapa de características de dos dimensiones (ancho y altura) por filtro. Cada filtro posee un determinado número de neuronas, las cuales están conectadas a una región de neuronas del filtro anterior (campo receptivo). Estas conexiones son con el objetivo de calcular un nuevo valor entre sus pesos y una pequeña región a la que están conectados en el volumen de entrada. Este nuevo valor es almacenado en el mapa de características, el cual será activado cuando se detecta algún tipo de característica visual. De acuerdo a [72], el mapa completo de todas las características se obtiene de diversos filtros. Matemáticamente el valor de la característica en la locación (*i,j*) en el mapa de características *k* de la *l*capa, $z^l_{i,j\ ,k}$ es calculado por:

$$z^l_{i,j\ ,k} = W^{lT}_k X^l_{i,j} + b^l_k \tag{3.1}$$

donde W^l_k y b^l_k son el peso del vector y del bias del filtro *k*, de la *l-ésima* capa, respectivamente; y donde $X^l_{i,j}$ es la entrada a la locación (*i,j*) de la *l-ésima* capa.

Existen tres hiper-parámetros que controlan el tamaño del volumen de la salida:

1. La profundidad del volumen de salida, corresponde al número de filtros que se utilizarán.

2. *Stride*, el cual determina el número de píxeles que será deslizado el filtro. Por

ejemplo, si el valor del *stride*= 1, esto indica que el filtro se moverá un píxel, antes de identificar el campo receptivo con el cuál estará trabajando.

3. *Zero-Padding*, permite rellenar con un valor de cero los bordes de la entrada, esto permite controlar el tamaño del espacio del volumen de salida.

Es posible calcular el tamaño del volumen de salida mediante la siguiente la función:

$$\frac{W - F + 2P}{S} + 1 \tag{3.2}$$

Donde *W* es el tamaño del volumen de entrada, *F* es el tamaño del filtro, *S* es *stride* que será aplicado y *P* es el número de *zero padding* que será aplicado en el borde.

Función de activación: Es una función que identifica cambios no lineales en la CNN, lo que permite en una red multicapa detectar características no lineales [132].

El valor de activación de una característica transformada (*convolutional feature*) $z^l_{i,j\ k}$ puede ser computarizada como:

$$a^l_{i,j\ ,k} = a(z^l_{i,j\ ,k}) \tag{3.3}$$

Pooling: El objetivo de esta capa es reducir el volumen de la representación de salida. De igual manera, permite reducir la cantidad de parámetros, el costo computacional y por lo tanto el sobre-entrenamiento (*overfitting*). Generalmente esta capa se coloca entre dos capas de convolución. Cada mapa de características de una capa de *pooling* es conectado a su predecesor de la capa de convolución. Denotando la función de *pool* para cada mapa de características $a^l\ :,:,_k$ se tiene:

$$Y^l_{i\ ,j\ ,k} = pool(a^l_{m,n\ ,k}), \forall (m,n) \in R_{I,J} \tag{3.4}$$

Donde $R_{I,J}$ es una región local de (*i,j*).

De manera general, la capa de *pooling:*

1. Acepta un tamaño de volumen $W_1 \times H_1 \times D_1$
2. Requiere dos hiper-parámetros: tamaño de filtro f y *stride*
3. Produce un volumen de tamaño $W_2 \times H_2 \times D_2$ donde:
   - $W_2 = \frac{W_1 - F}{S} + 1$
   - $H_2 = \frac{H_1 - F}{S} + 1$
   - $D_2 = D_1$
4. Produce cero parámetros.

Capas FC: Las capas totalmente conectadas (*fully connected*), son posicionadas después de las capas de convolución y capas de *pooling*, con el objetivo de lograr un mejor desempeño [136]. En esta capa cada una de las neuronas son conectada con todas las neuronas de la capa anterior, esto permite obtener información semántica global.

Al final, la última capa de la CNN, permite transformar la imagen original a una clase determinada.

3.1.2. Redes Neuronales Recurrentes

Las Redes Neuronales Recurrentes *(RNN)*, han sido utilizadas para tareas que involucran entradas secuenciales como reconocimiento de lenguaje o de texto. La RNN procesa uno por uno cada elemento de la secuencia de entrada. Mientras cada elemento es procesado, la red conserva en sus capas ocultas un vector de estado que contiene información de los elementos pasados de la secuencia. Es decir, el vector contiene el estado oculto de la secuencia anterior y el cuál será pasado a la secuencia siguiente. El estado oculto actúa como la memoria, ya que mantiene la información

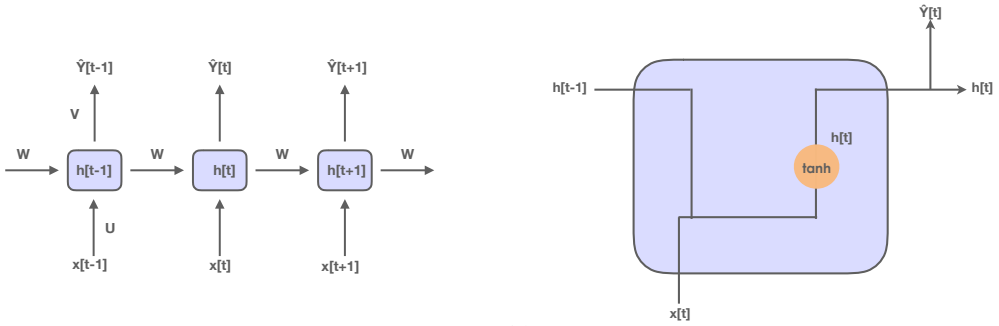
de la entrada anterior [132].

Durante la retropropagación (*backpropagation*), las RNN sufren del problema del desvanecimiento del gradiente *(vanishing gradient*). Como los valores del gradiente son utilizados para actualizar los pesos de las neuronas, es común que se presente el problema del *vanishing gradient*. Esto es porque a medida que retrocede y se va propagando el valor se vuelve demasiado pequeño. Por lo anterior, las capas que obtienen este valor dejan de aprender lo que ocasiona que las RNN olviden lo que adquirieron durante las secuencias anteriores [132].

Por esta razón las RNN, se consideran celdas de memoria a corto plazo. Esto indica que al momento de utilizar una secuencia suficientemente larga, se tendrá dificultad para pasar la información de las celdas anteriores a las últimas. Por lo cuál, al utilizar una RNN para realizar predicciones de un texto con secuencias largas, se pierde información relevante de las primeras etapas.

El estado oculto en una RNN, se calcula de la combinación de la entrada actual *x[t]* y el estado oculto de la celda anterior *h[t-1]*. Este cálculo forma un vector que contiene información de la entrada actual y de la entrada anterior. El vector pasa por una función de activación *tanh*, y la salida $\hat{Y}$[t-1] forma el nuevo estado oculto, o la memoria de la red. La función de activación *tanh* es utilizada para regular los valores entre -1 y 1. En la figura 3.3, se muestra una secuencia de RNN y una unidad de la RNN.

Las celdas de memoria a largo plazo *(LSTM)* y las Unidades de Compuertas Recurrentes *(GRU)*, fueron creadas para solucionar el problema de la memoria a corto plazo de las RNN. Cada celda de un LSTM, posee un mecanismo interno llamado compuertas que permiten regular el flujo de información. Además este mecanismo



a) b)

Figura 3.3: a) RNN representando varias secuencias , b) Unidad de una RNN.

aprende a identificar cual información dentro de una secuencia es importante almacenar y cual se debe desechar. Haciendo esto, las celdas LSTM pueden transmitir información relevante a lo largo de la secuencia que permita realizar predicciones. La mayoría de los modelos más utilizados en el estado del arte para tareas de NLP son basados en RNNs. Estos modelos incluyen las unidades de LSTM y GRU, y son utilizados para el reconocimiento de lenguaje, generación de texto, además para generar subtítulos para vídeo. En esencia los modelos LSTM y GRU pueden aprender a almacenar solo información relevante para realizar predicciones y desechar aquellos datos no significativos [138][139].

3.1.3. LSTM

Los modelos LSTM son una categoría de RNN capaces de almacenar un valor por largos tiempos de ejecución. Fue introducido por [98], a partir del cuál fueron utilizados para diferentes áreas de investigación obteniendo mejores resultados.

Un LSTM tiene un flujo similar que una *RNN*, la cual procesa los datos que transmiten información a medida que se propaga. La diferencia radica en las operaciones que se realizan dentro de las celdas. Estas operaciones son utilizadas para

permitir que el LSTM almacene o deseche información considerando su relevancia [98].

El concepto clave de este modelo es la celda de estado, la cuál esta compuesta por varias compuertas. La celda de estado es una banda transportadora de información a lo largo de la secuencia, a la que se puede añadir o remover datos. Esto es lo que representa la memoria de la red [140].

La celda de estado, puede cargar información relevante a lo largo del procesamiento de la secuencia. Incluso la información de etapas tempranas puede llegar a etapas posteriores, reduciendo el efecto de la memoria a corto plazo. Mientras que la celda de estado va realizando su recorrido a través de la secuencia, las compuertas se encargan de evaluar las entradas. Es decir definir que debe almacenarse y que debe desecharse durante el entrenamiento.

Algunas de estas compuertas son funciones de activación como *tanh* y *sigmoid*, las cuales tienen la función de conservar los valores entre -1 y 1 si corresponde a *tanh*, y valores entre 0 y 1 si corresponde a *sigmoid*. De esta manera la red puede aprender cual dato es importante y es necesario conservar así como cuál deberá ser desechado. La LSTM, utiliza tres tipos de compuertas para regular la información que recibe de entrada. Una compuerta para eliminar datos de la memoria *(forget gate)*, compuerta de entrada *(input gate)* y la compuerta de salida *(output gate)* [140]. En la figura 3.4 se muestra la arquitectura de un LSTM.

- ***Forget Gate:*** Esta compuerta permite seleccionar la información que debe ser almacenada y cuál debe ser removida. Durante este proceso la información del estado oculto anterior e información de la entrada actual se procesa a través de la función de activación *sigmoid*. Esta función no considera los valores de

salida cercanos o iguales a 0, mientras que los valores iguales o cercanos a 1 son almacenados. La ecuación (3.5) muestra como se calcula la salida de la compuerta, donde σ representa la función de activación *sigmoid*; *Wf* son los pesos asociados a $h_{(t-1)}$ y x_t; y el b_f es el *bias* asociados a los pesos del vector.

$$f_t = \sigma(Wf.[h_{(t-1)}, x_t] + b_f) \tag{3.5}$$

- ***Input Gate:*** Esta compuerta se utiliza para actualizar la celda de estado. Para ello combina la entrada actual, la salida de la última unidad de LSTM y el valor de la celda de estado en su última iteración. El primer paso es pasar el estado oculto anterior y la entrada actual a una función de activación *sigmoid*. Esta función identificar que valores se van a actualizar transformando los valores entre 0 y 1. Una vez transformados los datos que sean iguales a 0 son irrelevantes y son desechados, mientras que aquellos igual a 1, son relevantes para el proceso de actualización. Además de esto, es necesario pasar los datos anteriores del estado oculto y la entrada actual a una función de activación *tanh*. Lo anterior con la finalidad de reducir los valores en un rango de -1 y 1 para regular la red. Posteriormente se multiplican la salida de la función de activación *tanh* y la salida obtenida de la función de activación de *sigmoid*. La función de activación *sigmoid* va a ser quien defina que información es relevante de la salida de la función *tanh* y que debe conservarse. En la ecuación (3.6) se muestra como se calcula la compuerta de entrada *(input gate)*. El símbolo de σ representa la función de activación *sigmoid*; *Wi* son los pesos asociados a $h_{(t-1)}$ y x_t; y el b_i es el *bias* asociados a los pesos del vector.

$$i_t = \sigma(Wi.[h_{(t-1)}, x_t] + b_i) \tag{3.6}$$

$$\hat{C}_t = tanh(W_C.[h_{(t-1)}, x_t] + b_C) \tag{3.7}$$

La ecuación (3.7) representa la función de activación *tanh*. Esta función evalúan los posibles candidatos que se guardaran en la celda de estado. $W_C$representa los pesos asociados a $h_{(t-1)}$ y x_t, y b_c es el *bias* asociados a los pesos del vector.

- ***Cell State:*** Primeramente la Celda de estado obtiene el elemento multiplicador por el vector obtenido de la compuerta para eliminar *(Forget Gated)*. Esto tiene la posibilidad de remover los valores en celda de estado si se multiplica por valores cercanos al "0". Posteriormente se obtiene la salida de la compuerta de entrada *(Input Gate)*. Además se realiza una adición al elemento multiplicador que actualiza la celda de estado con un nuevo valor que la red neuronal encuentra relevante. El procedimiento anterior proporciona el nuevo valor. La ecuación (3.8) muestra el como se obtiene la nueva información que contiene la celda de estado *(cell state)*. En la ecuación f_t representa *Forget Gate*, $C_{(t-1)}$ representa la celda de estado, i_t representa la compuerta de entrada *(Input Gate)* y $\hat{C}_t$ representa los nuevos candidatos para la celda de estado.

$$C_t = f_t * C_{(t-1)} + i_t * \hat{C}_t \tag{3.8}$$

- ***Output Gate:*** Por último tenemos la compuerta de salida *(Output Gate)*. Esta compuerta decide cuál será el siguiente estado oculto. El estado oculto contiene información de la entrada anterior, además se utiliza para la predicción. Primero se requiere enviar el estado oculto anterior y la entrada actual en una función de activación *sigmoid*. Posteriormente la nueva celda de estado *(Cell State)* pasa por la función de activación *tanh*. En esta función se multiplica la salida de la función de activación *tanh* y la salida de la función de activación *sigmoid*, con la finalidad de identificar la información que va almacenanarse. La ecuación (3.9), muestra como se obtiene la compuerta de salida *(Output Gate)*, donde σ representa la función de activación *sigmoid*. *Wo* son los pesos

asociados a $h_{(t-1)}$ y x_t así como el b_o es el *bias* asociados a los pesos del vector. La ecuación (3.10) muestra como se obtiene el nuevo estado oculto, donde O_t representa la compuerta de salida *(Output Gate)* y C_t representa la celda de estado *(Cell State)*.

$$O_t = \sigma(W_O.[h_{(t-1)}, x_t] + b_O) \tag{3.9}$$

$$h_t = O_t * tanh(C_t) \tag{3.10}$$

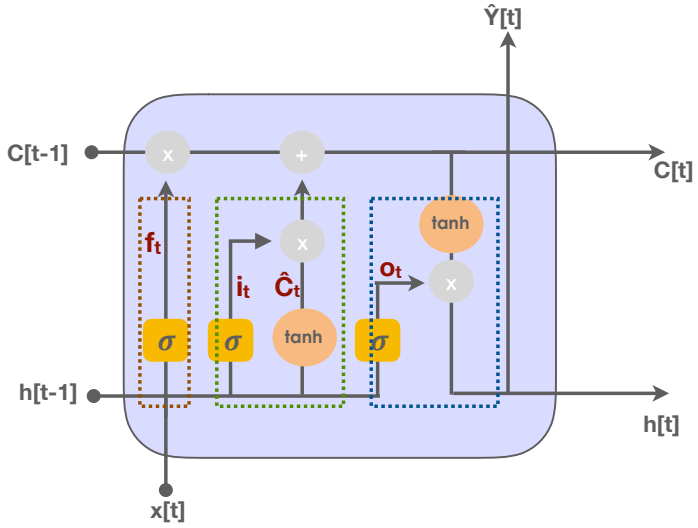


Figura 3.4: LSTM

3.2. Procesamiento de Lenguaje Natural (NLP)

La exploración de estrategias que permitan el uso de máquinas para el Procesamiento del Lenguaje Natural (NLP), es fundamental para tareas como la generación automática de reportes médicos. El procesar un texto médico por medio de una máquina requiere de representaciones matemáticas que engloben las propie-

dades semánticas y sintácticas de cada palabra [141]. Considerando los estudios realizados para las tareas de NLP se ha destacado el uso de los *Word Embedding* (WE). El uso de estas representaciones de palabras como entradas para los modelos lenguajes han obtenido resultados competitivos [142]. Sin embargo, la mayoría de los estudios realizados consideran una perspectiva del dominio general del texto y los resultados no necesariamente aplican para el texto de otros dominios como el biomédico. Para la tarea de generación de reportes médicos se han propuesto diversos enfoques basados en modelos de DL y estrategias de WE. Estos enfoques fueron inspirados por soluciones implementadas para tareas de NLP pero con el objetivo extraer conceptos a partir de textos médicos [7].

3.2.1. *Word Embeddings* (WE)

El fundamento de los algoritmos utilizados para la representación de palabras sugiere que los elementos léxicos con distribuciones similares comparten significados similares [141]. Estos algoritmos utilizan vectores para representar las palabras y determinan su similitud calculando la distancia entre ellos y la similitud del coseno. Para la realización de tareas de NLP, se requieren técnicas que permitan representar el texto que se servirá como entrada para alimentar los algoritmos de DL. Las técnicas de WE realizan diversos procedimientos para convertir cada palabra en vectores. Lo anterior con el objetivo de obtener una representación del texto más enriquecida que contribuya al aprendizaje del modelo. Dentro de estas estrategias de WE como Word2vec, GloVe y FastText [142].

3.2.2. *Embeddings* Contextualizados

Los algoritmos tradicionales de WE como GloVe, Word2vec se caracterizan por representar cada palabra por un solo vector y calcular su distribución con respecto a otras palabras. Sin embargo, al no considerar los diferentes significados que puede

tener una palabra dependiendo del contexto en que se utiliza, su funcionalidad queda limitada. Los embeddings contextualizados permiten generar diferentes vectores que representen los diversos significados de una palabra de acuerdo al contexto de la oración. De acuerdo a Chiu et. al. [141], en el estado del arte se encuentran algunos modelos que permiten generar la contextualización en los *embeddings* como:

1. *Embeddings* con modelos de lenguaje: Modelos como ELMo permiten enriquecer la representación de cada palabra considerando el contexto sintáctico y la estructura morfológica de cada palabra en el texto. La arquitectura de este modelo consiste en un conjunto de LSTM bidireccionales multinivel. El objetivo es tomar una representación de los modelos de lenguaje de ambas direcciones [141].

2. *Bidirectional Encoder Representation from Transformers*: Similar al modelo de ELMo, BERT es un modelo neuronal contextualizado, el cuál aprende embeddings de acuerdo de la relación de las palabras en un texto. Utilizando una técnica *Bi-Transformer*, que de manera efectiva explora la información semántica de cada sentencia. La ventaja de esta caracterización es que puede fusionar el contexto en ambas direcciones [141].

3.3. Transformers

Inicialmente la arquitectura de *Transformers* fue propuesta con el objetivo de contribuir a las tareas de traducción en el área de NLP. Sin embargo, debido a su desempeño y novedosa arquitectura, su uso se expandió en otras tareas diferentes a NLP. Recientemente esta arquitectura es explorada en el área de visión por computadora [3] para tareas de reconocimiento de imágenes, ofreciendo resultados competitivos comparados con en el estado del arte. La figura 3.5 muestra de manera general la estructura del modelo.

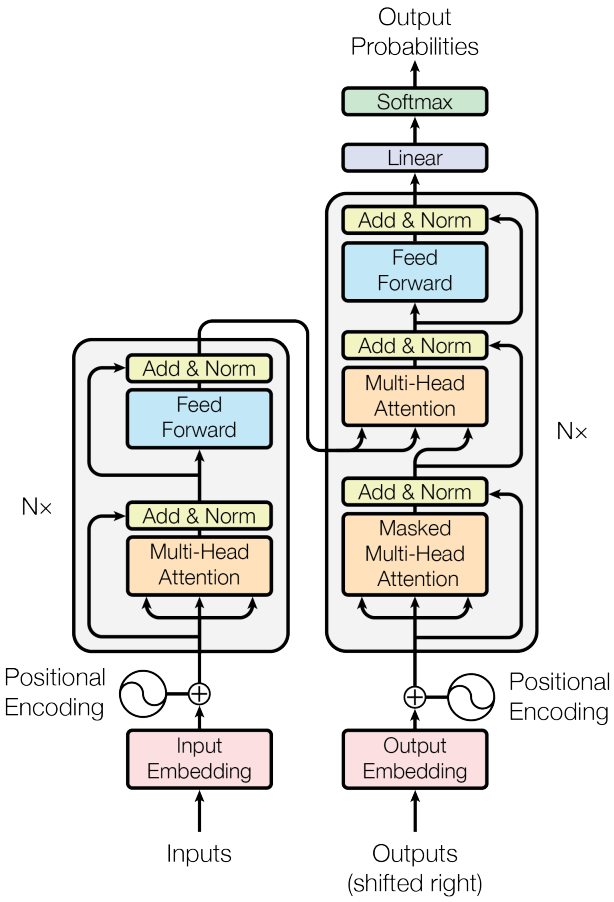


Figura 3.5: Arquitectura del *Transformer*, tomada de [1]

La arquitectura propuesta por Vaswani et.al. en el 2017 [1], consiste en 6 *encoders* apilados y sus correspondientes *decoders*. Su estructura varia de los modelos secuenciales basados en arquitecturas de RNN y CNN, debido a que remplaza la recurrencias y convoluciones con modelos de atención. Esto permite realizar una ejecución más rápida y eficiente. La figura 3.5 muestra un diagrama de su estructura.

- *Encoder*: Cada *encoder* esta constituido con dos capas, la primera capa es un

mecanismo de atención mientras que la segunda capa es una red neuronal *Feed Forward* (FFNN). El mecanismo de atención analiza los vectores de entrada simultáneamente para determinar la relación de una palabra con respecto al resto del texto. Este mecanismo de atención posee múltiples cabezales que permiten identificar diferentes representaciones de la información. Utilizando 3 diferentes vectores por entrada: *Query vector (Q), Value vector (V), Key vector (K)*, se obtiene un vector de salida que almacena la relación e importancia de cada vector con respecto a los otros. El vector resultante alimenta a la siguiente capa (FFNN) junto con el vector de entrada (conexión residual), lo que permite que la información que recibe como entrada sea agregada con el vector de salida.

- *Decoder*: A diferencia del *encoder*, en cada *decoder* se incluyen dos capas de atención, lo que permite identificar a que partes del vector de entrada debe prestar más atención. El primer mecanismo de atención es alimentado con el vector de salida del *encoder* y el segundo trabaja como un mecanismo de atención con múltiples cabezales.

- Codificación Posicional *(Positional Encoding)*: Debido a que los *Transformers* reciben los *Word Embeddings* (WE) simultáneamente, el *Positional enconding* permite al modelo identificar la secuencia de cada uno de ellos. En otras palabras el *Positional encoding* representa la posición de la palabra en el tiempo. Mediante la suma de los WE con el *Positional encoding*, el algoritmo tiene la capacidad de identificar la secuencia del texto [143].

3.3.1. BERT: *Pre-training of Deep Bidirectional Transformers for Language Understanding*

El modelo de BERT propuesto por Devlin et.al. [2], es caracterizado por incluir una arquitectura basada en el modelo de *Transformer*. Inicialmente, el modelo de BERT fue diseñado para tareas orientadas a NLP, sin embargo algunos autores como [5] han expandido su uso adecuando al modelo para realizar tareas relacionadas a visión por computadora.

La estructura es un incluye un *encoder* bidireccional multi-capa que recibe como entrada la concatenación de dos segmentos (A,B) de secuencias de *tokens*. Esto permite que el modelo sea ideal para tareas relacionadas a preguntas-respuestas (QA) y tareas de traducción de lenguaje.

La definición de la secuencia de entrada, esta conformada por un *token* especial [CLS]. Posteriormente la entrada se conforma por la secuencia de *tokens* del segmento A, un *token* especial [SEP] y la secuencia de *tokens* del segmento B.

Dentro de la literatura se reportan dos tamaños de modelo, BERT-Base que incluye 12 *encoders* apilados y BERT-Large que incluye 24 *encoders* apilados. El preentrenamiento de BERT se realiza con dos tareas no supervisadas. La primera tarea denominada *Masked LM*, consiste en enmascarar aleatoriamente un porcentaje de *tokens* que conforman la entrada y posteriormente el modelo debe predecir el *token* enmascarado. La segunda tarea consiste en la predicción de la siguiente sentencia (NSP, por sus siglas en inglés). Con el objetivo de que el modelo identifique una relación entre sentencias (por ejemplo para la tarea de QA), el modelo se alimenta de pares de sentencias y se le pide que determine si la segunda oración sigue lógicamente a la primera o no. Esto permite que el modelo pueda capturar infor-

mación sobre la relación y coherencia entre diferentes partes del texto. La figura 3.6 muestra un esquema general del modelo así como las tareas requeridas para su entrenamiento.

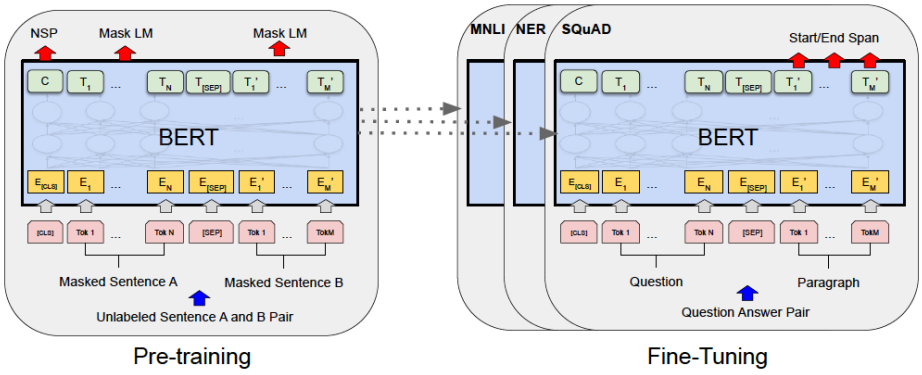


Figura 3.6: Arquitectura general del modelo *BERT*, tomada de [2]

3.3.2. *Visual Transformer* (ViT)

Una de las arquitecturas más novedosas en el campo de procesamiento de imágenes es el *Visual Transformer* ViT. Esta arquitectura adapta los principios fundamentales del *Transformer* [1], el cuál originalmente fue desarrollado para tareas de NLP. A diferencia de los modelos tradicionales de CNNs utilizados en el procesamiento de imágenes, el ViT elimina por completo las capas convolucionales y se basa únicamente en mecanismos de atención para capturar las relaciones espaciales y semánticas en las imágenes. Este enfoque permite una mayor capacidad de generalización y una mejor captura de contextos globales y locales en las imágenes, lo que lo hace especialmente efectivo para tareas de visión por computadora. La figura 3.7 muestra un esquema general de la arquitectura ViT.

La arquitectura del *Visual Transformer* (ViT) se caracteriza por incluir una estructura basada en el modelo *Transformer* [1] que incluye un *encoder-decoder*. Sin embargo, el modelo ViT esta orientado a la tarea de clasificación de imágenes y solo

incluye en su estructura el *encoder*. La figura 3.7 muestra un esquema general de la arquitectura propuesta.

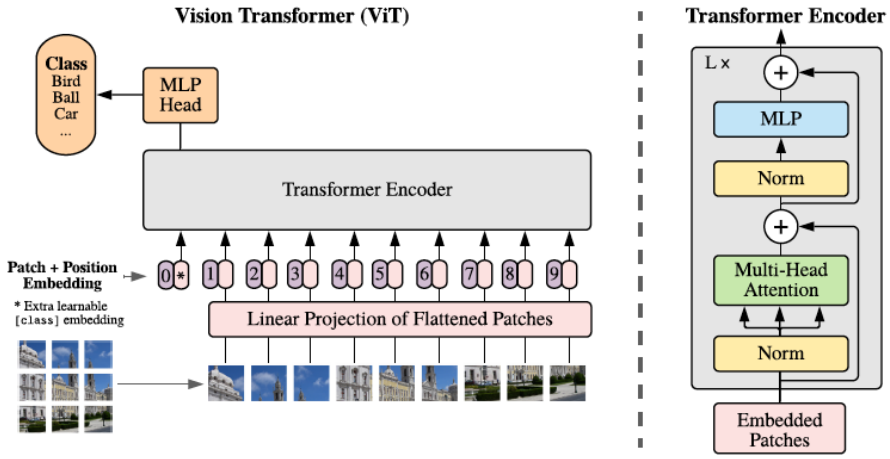


Figura 3.7: Arquitectura del *Visual Transformer (ViT)*, tomada de [3]

El modelo procesa una imagen I, donde $I \in R^{H \times W \times C}$, es dividida en una secuencia de N segmentos $\{X_1, X_2, ..., X_N\}$. Posteriormente cada segmento $X_i \in R^{C \times P}$, donde C representa el número de canales y [P,P] el número de resolución del *patch*. Posteriormente se obtiene la representación de *tokens* mediante la proyección lineal que incluye la vectorización de cada segmento.

$$\check{x} = \{X_1 E, X_2 E, .., X_N E\}, E \in R^{C \times P^2 \times D} \tag{3.11}$$

Donde D representa la dimensión del vector. Posteriormente, se identifica la posición mediante un *positional embedding* identificado como E_{pos}. El vector de entrada que alimenta al *encoder* del modelo se identifica como:

$$x = \check{x} + E_{pos}, E_{pos} \in R^{N \times D} \tag{3.12}$$

Cada bloque del *encoder* del modelo ViT incluye un módulo con múltiples cabezales de atención (*multihead attention*, MSA), un perceptrón multicapa (*multi-layer perceptron*, MLP) incluyendo una capa de normalización (LN).

$$Z' = MSA(LN(Z_{l-1})) + Z_{l-1} \tag{3.13}$$

$$Z_l = MLP(LN(Z'_L)) + Z'_L \tag{3.14}$$

La arquitectura ViT incluye las variantes de acuerdo a su tamaño. El ViT-Base incluye 12 *encoders* apilados, ViT-Large incluye 24 *encoders* apilados y ViT-Huge incluye 32 *encoders* apilados.

3.4. Métricas de Evaluación

Las métricas generales de Generación de Lenguage Natural (NLG) que comúnmente se utilizan para evaluar el desempeño de los modelos orientados a la generación de reportes médicos a partir de imágenes son las siguientes:

BLEU *Bilingual Evaluation Understudy* [94]. Esta métrica contabiliza el número de *n-gramas* de palabras (1-unigrama, 2-bigramas, 3-trigramas y 4-cuatrigramas) de la sentencia generada por el modelo que coinciden con los de la sentencia objetivo (*target*). El número de palabras a considerar dentro del *n-grama* puede ser de 1 (BLEU1) hasta 4 (BLEU4), y sus valores corresponden al intervalo 0 a 1.

ROUGE-L *Recall-Oriented Understudy for Gisting Evaluation* [95]. La métrica ROUGE-L permite calcular la media armónica (*F-measure*) entre la *precision* y *recall* a nivel de las sub-secuencias comunes más largas (*Longest Common Subsequence* o LCS) entre la sentencia generada y la sentencia objetivo. Sus resultados abarcan de un valor

mínimo de 0 a un valor máximo de 1. Sin embargo no considera ningún aspecto que pueda evaluar la calidad del texto generado.

METEOR (M) [96]. Esta métrica evalúa la precisión de la sentencia generada por el modelo con respecto a la sentencia objetivo considerando un orden explícito de los *n-gramas*. Además, toma en cuenta el uso de sinónimos entre los *n-gramas* así como la raíz de las palabras basandose en el conjunto de datos denominado WordNet [144] . Sus resultados van de un valor mínimo de 0 a un valor máximo de 1. Aunque esta métrica incluye más estrategias en su evaluación, está limitada por el uso de un vocabulario no especializado para detectar sinónimos y paráfrasis en texto médico.

CIDEr *Consensus-based Image Description Evaluation* [97]. CIDEr determina la similitud entre sentencias basado en el cálculo del coseno entre el vector de pesos de los *n-gramas* de la sentencia generada y el vector de pesos de los *n-gramas* de la sentencia objetivo. Estos pesos se calculan de acuerdo al valor de *Term Frequency Inverse Document Frequency* (TF-IDF) que pondera los *n-gramas* más importantes según su frecuencia de aparición en el conjunto de sentencias objetivo. Además, esta métrica agrupa las palabras que tienen una misma raíz o *stemming* al realizar la ponderación de los *n-gramas*. Esto garantiza una evaluación justa de los n-gramas. CIDEr se considera la métrica más completa para evaluar la tarea de subtitulado de imágenes. Otro aspecto de CIDEr es la medición de las similitudes semánticas entre las frases generadas y las verdaderas. Para ello se calcula la media de todas las frases, lo que facilita la correlación de la calidad del texto con los juicios humanos. Además de la calidad del texto, CIDEr tiene en cuenta la precisión y la recuperación (*recall*) así como la calidad del texto generado. Por ello es ideal para evaluar informes médicos.

Estas métricas se utilizan para calcular el desempeño del modelo propuesto ob-

servando la similitud o la diferencia entre los párrafos generados y las descripciones escritas por los radiólogos. Un buen desempeño se refleja en puntuaciones altas en BLEU, ROUGE, METEOR y CIDEr.

Algunos autores como [25],[78],[82],[79],[71],[83], [90], además de utilizar las métricas para evaluar el reporte generado, incluyen métricas para evaluar la correcta clasificación de la imagen como:

Accuracy. Esta métrica se utiliza para evaluar la relación de las predicciones correctas e incorrectas en la clasificación de la imagen [145]. Sus valores corresponden al intervalo 0 a 1.

***Area under the ROC curve* (AUC)**. El ROC *(Receiver Operating characteristic curve)* es una representación gráfica que proyecta la proporción de las predicciones correctas o verdaderos positivos (reflejadas en el eje y) con respecto a la proporción de falsos positivos (eje x). La AUC mide el área bajo la curva de la gráfica ROC y sus valores corresponden al intervalo 0 a 1 [145].

Capítulo 4

Modelo ETB-MII

El modelo ETB-MII consiste en una arquitectura constituida por un codificador y un decodificador (*encoder-decoder*) con mecanismos de atención. Para alimentar al modelo, se requirió aplicar un pre-procesamiento en las imágenes así como en el texto asociado. Las imágenes fueron normalizadas y recortadas mientras que al texto asociado se aplicó un cambio a letras minúsculas, se removieron los caracteres especiales y se filtraron las palabras con menor frecuencia. Además una técnica de aumento de datos fue aplicada tanto a la imagen como al texto.

El módulo del *Visual Encoder* (codificador visual) está basado en un *Transformer* Visual y conformado por 12 *encoders* apilados. Para recibir la imagen como entrada, primero es necesario dividirla en pequeños segmentos. Posteriormente, el *Visual Encoder* recibe como entrada la secuencia de los segmentos obtenidos con el objetivo de identificar las características relevantes de cada uno y su importancia en relación a los demás. El *Semantic Decoder* (Decodificador semántico), que incluye 12 *decoders* apilados, recibe y transmite la salida de cada uno de los *encoders* a todos los bloques de *decoders* que constituyen al modelo. Una vez realizado lo anterior, se genera la descripción correspondiente para cada región asociada. El resultado final es el texto

médico, el cual es generado palabra por palabra. Las interacciones entre las características de la imagen y el texto se realizan a través de un mecanismo de *Atención Cruzada Multicapa* (MCA), en el que se implementa una atención producto-punto escalada. La figura 1 muestra una visión general del modelo propuesto.

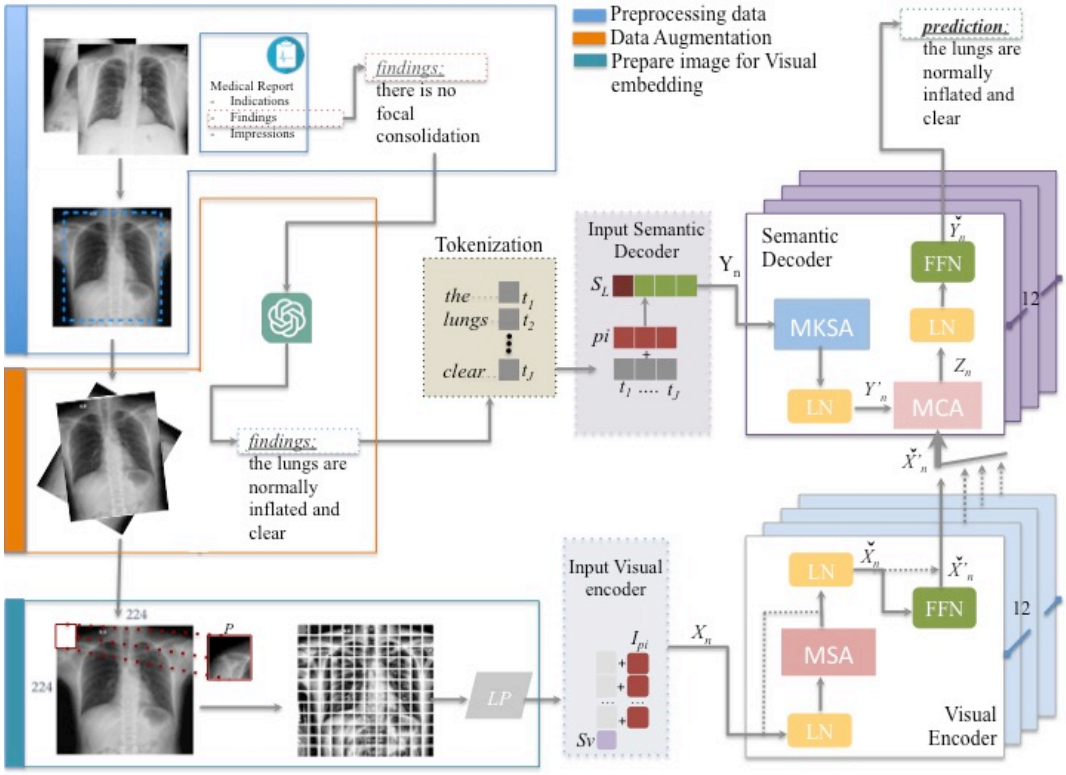


Figura 4.1: Representación general del modelo propuesto, Enhanced Transformer Based - Medical Image Interpretation (ETB-MII) durante la etapa de entrenamiento. Las imágenes y su texto correspondiente son procesados, junto con la estrategia de aumento de datos (DAS), antes de convertirlos en *embeddings* para la entrada al modelo. El *Visual Encoder* recibe el *embedding* visual X_n, mientras que el *Semantic Decoder* recibe el *embedding* contextual Y_n. Ambos procesos ocurren simultáneamente y comparten la información relevante a través del bloque MCA para producir el texto médico final.

4.1. Preprocesamiento de datos

Durante el preprocesamiento de datos se realiza una normalización a cada una de las imágenes y al texto de *findings* asociado que conforman el conjunto de entre-

namiento D_{train}. La figura 4.2 incluye un ejemplo de una imagen de rayos X y el reporte médico asociado.

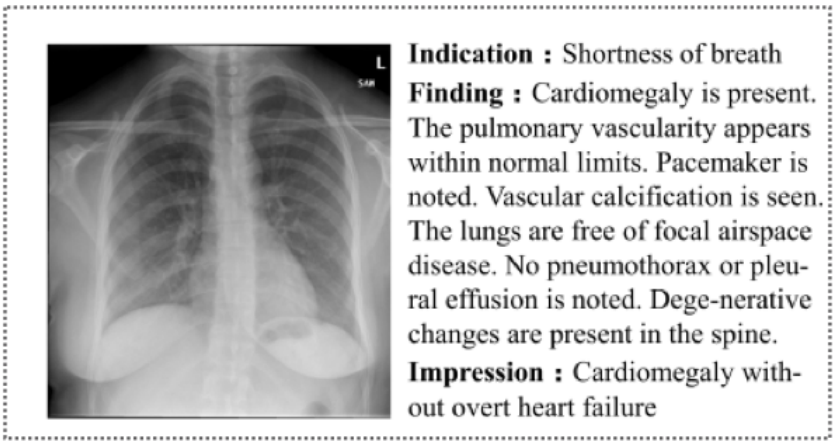


Figura 4.2: Imagen tomada del conjunto de datos de IU-Xray [4]

.

Durante el preprocesamiento se aplica una normalización a cada $Imagen$ que implica ajustar los valores de los píxeles para que su distribución tenga una media de cero y una desviación estándar de uno.

$$I = \frac{Imagen - \eta}{\sigma} \tag{4.1}$$

Donde $I_k \in R^{H \times W \times C}$ representa la imagen de entrada, mientras que la altura, anchura y número de canales de la imagen son representados por H, W y C. El símbolo η es la media de los valores de los píxeles de la imagen original y σ es la desviación estándar de los píxeles de la imagen original. Posteriormente se aplica un recorte en la imagen partiendo desde el centro hacia afuera estableciendo un tamaño determinado de 224×224.

Para aplicar el preprocesamiento en el texto se identifica la sección de *findings* del conjunto de entrenamiento D_{train} como $fk = \{t_1, t_2, .., t_j\}$; $t_{1:J} \in U^{\omega}$, donde J define el número total de *tokens* o palabras y U^{ω} es un conjunto finito de representaciones de palabras. Para cada fk, se aplica una transformación a minúsculas, se remueven los caracteres especiales, puntuaciones y espacios en blanco. Aunado a lo anterior, se realiza un conteo general de la frecuencia de cada *token* mediante un diccionario y se eliminan aquellos que tienen una frecuencia menor a 3.

4.2. Estrategia de Aumento de datos (DAS)

Para aplicar la estrategia de aumento de datos (DAS) se selecciona aleatoriamente un subconjunto s del conjunto de datos D_{train}, el cual está constituido por imágenes y el texto de la sección de *findings* del reporte médico asociado. Este subconjunto es determinado por $D_M = \{M_1, M_2, ...M_k\}$, $k = \{1, 2, ..., s\}$. Cada elemento M_k está constituido por $M_k = (I_k, f_k)$, $I_k \in R^{H \times W \times C}$ donde I_k representa la imagen de entrada y H, W, y C corresponde a altura (*height*), anchura (*width*) y número de canales (*number of channels*). El texto de *findings* es representado por $f_k = \{t_1, t_2, ..., t_J\}$; $t_{1:J} \in U^{\omega}$, donde *J* define el número total de *tokens* y U^{ω} es un conjunto finito de representaciones de palabras.

La representación sintética del conjunto de datos es generada a partir de D_M aplicando la estrategia DAS, la cuál consiste en dos fases. En la primera fase se aplica una rotación aleatoria a cada I_k de D_M. La segunda fase incluye un aumento de datos contextual, donde el texto f_k es parafraseado utilizando un modelo autogenerativo de OpenIA[1] definido como GPT3.5-turbo. Este modelo ha sido entrenado en una gran cantidad de texto, lo que le permite entender y generar texto basado en el contexto y significado. Para comprender el contexto en un texto determina-

[1]https://openai.com/

do, el modelo GPT3.5 tiene en cuenta las palabras anteriores y posteriores a una palabra concreta. Lo anterior permite que el modelo entienda el significado de una palabra dentro de una frase o párrafo. Por último, el modelo genera el texto de forma probabilística, considerando una serie de posibles palabras o frases siguientes y asignando probabilidades a cada una de ellas en función de las asociaciones aprendidas.

La tabla 4.1 incluye un ejemplo de texto *findings* original y el texto *findings* aplicando el parafraseo generado por GPT-3.5.

Tabla 4.1: Ejemplo de *findings* y el parafraseo *findings* generado por ChatGPT.

Original *findigs*
1)the heart is normal in size 2)the mediastinum is unremarkable 3)the lungs are hyperinflated there is biapical scarring 4)no acute infiltrate or pleural effusion seen
Phrasing *findigs* generated with GPT-3.5
1)the heart appears within normal dimensions 2)the mediastinal region shows no significant abnormalities 3)the lungs exhibit excessive inflation biapical scarring is present 4)there are no signs of recent infiltration or accumulation of fluid in the pleural cavity

Por lo tanto, el conjunto de datos sintéticos producido se define como:

$$\begin{aligned} D_{synthetic} &= \{M_{b1}, M_{b2}, ..., M_{bk}\}; k = \{1, 2, ..., s\} \\ M_{bk} &= (f_{Rotation}(I_k), f_{prompting}(f_k)) \end{aligned} \tag{4.2}$$

Donde M_{bk} es una muestra que incluye la imagen rotada I_k y el parafraseo generado a partir de la sección de *findings* f_k.

Una vez creado el conjunto de datos sintético, se incluye en nuestros datos de entrenamiento originales que utilizará el modelo ETB-MII. Los datos de entrena-

miento finales se definen como: $D_T = \{D_{train}, D_{synthetic}\}$.

4.3. *Visual Encoder*

A diferencia de los modelos descritos en el trabajo relacionado, que incorporan una CNN para la extracción de características visuales, el modelo ETB-MII se inspira en la arquitectura de *Vision Transformer* (ViT) propuesta en [3]. Dado que el *Visual Encoder* requiere una secuencia de entrada 1D de *embeddings* conformada por *tokens*, es necesario manipular las imágenes para encontrar una mejor representación. Dado lo anterior, una imagen I, se divide en una serie de segmentos 2D no translapados $\{I_1, I_2, \ldots, I_N\}; I_i \in R^{(P^2 \times C)}$, donde P^2 es la resolución de cada segmento de imagen, y $N = HW/P^2$ es el número total de segmentos por imagen. Posteriormente, todos los segmentos se proyectan a una dimensión D a través de una capa de proyección lineal (*learnable linear projection layer*), cuyo tamaño es definido por P^2C. Una vez proyectados, los segmentos N se representan en forma de matriz $X \in R^{N \times D}$. Se añade un token especial S_v a la secuencia resultante y un *embedding* posicional I_{pi} para conservar la información de posición. Finalmente, el *embedding* visual de entrada X_n que alimenta al *Visual Encoder*, se define como:

$$X_n = [X, S_v] + I_{pi}, \in R^{(N+1) \times D} \tag{4.3}$$

En esta implementación, X_n pasa a través de 12 bloques de *encoders*, donde cada bloque incluye un multicabezal de atención propia mejor conocido como *multi-head self-attention* (MSA) y una Red Neuronal Feed Forward (FFN) con una capa de normalización (LN). Además, en cada bloque se incluyen conexiones residuales. El proceso que realiza el *Visual Encoder* entre bloques se representa mediante:

$$\begin{aligned} \check{X}_n &= MSA(LN(X_n)) + X_n, \\ \check{X}'_n &= FFN(LN(\check{X}_n)) + \check{X}_n \end{aligned} \tag{4.4}$$

En la ecuación 4.4, el *embedding* visual X_n pasa a través de un *encoder* donde una capa de normalización (LN) recibe la entrada. El módulo MSA determina la relevancia de cada segmento de la imagen, y la salida del módulo MSA se añade a la conexión residual que contiene el *embedding* visual X_n. El resultado se envía a la capa FFN y, por último, la salida FFN se combina con la entrada $\check{X}_n$ a través de las conexiones residuales, produciendo $\check{X}'_n$. El proceso del módulo MSA se describe en detalle en la siguiente sección.

4.4. *Multi-headed Self Attention* (MSA)

Después de que el *embedding* visual X_n pase por una capa LN, el módulo MSA la recibe y pasa por una capa de proyección (*learnable linear projection layer*). A continuación, X_n se mapea a tres *matrices* denominadas *query* Q, key K y value V mediante tres parámetros entrenables $W_q, \in R^{D \times D_q}, W_k, \in R^{D \times D_k}$, y $W_v, \in R^{D \times D_v}$ como se representa a continuación:

$$Q = X_n \times W_q \tag{4.5}$$

$$K = X_n \times W_k \tag{4.6}$$

$$V = X_n \times W_v \tag{4.7}$$

Donde D_q, D_k, y D_v representan el número de características que son entrenables en las matrices W^q, W^k, W^v, que permiten mapear la información de entrada en diferentes dimensiones.

Dado que el objetivo del MSA es calcular los pesos de asociación entre diferentes características de la imagen, una atención de producto punto escalado (mejor conocido como *scaled dot-product attention*) es requerido en esta etapa. Esto se define como se muestra en la ecuación 4.8.

$$A(Q,K) = softmax\Big(\frac{QK^T}{\sqrt{Dq}}\Big) \tag{4.8}$$

La salida de la capa *self-attention* $H \in R^{N\times D_v}$, es definida como:

$$H = \textit{self-attention}(X_n) = A(Q,K) \times V \tag{4.9}$$

Después de *self-attention*, H se desempeña en paralelo en todos los cabezales de atención, la salida final es producida por la concatenación de todos los cabezales H, como se define a continuación:

$$MHA(Q,K,V) = Concat(H_1, H_2, .., H_i)W^O \tag{4.10}$$

Donde i, representa el número de cabezal y W^O es un peso de proyección entrenable.

4.5. *Semantic Decoder*

Considerando las características visuales más relevantes $\check{X}'_n$ que son extraídas del *Visual Encoder*, el *Semantic Decoder* va a predecir el texto médico del *findings*, el cuál representa la descripción de la imagen. Para realizar esta tarea, el texto médico de *findings* el cuál es el texto objetivo o *target*; se divide en una secuencia de *tokens* como $\{t_1, t_2, ..., t_J\}$; $t_{1:J} \in U^\omega$, donde *J* representa el total de *tokens* y U^ω es un conjunto

finito de representación de palabras. Aquí $\omega : \{1, 2, .., S\} \longmapsto U$, S corresponde al índice total que asigna la representación a las palabras [146]. Posteriormente, para producir el *embedding* semático, Y_n, se agrega un *token* especial el cuál es agregado al inicio del *embedding* semático así como un *token posicional* p_i es agregado como se muestra a continuación:

$$Y_n = [S_L, t_1, t_2, ..., t_J] + p_i, \in R^{J \times D_M} \tag{4.11}$$

Donde D_M es la dimensión del *decoder*. El *decoder* del modelo propuesto está basado en el modelo propuesto por Radford et al. in [32] con una dimensión del 768.
De acuerdo con la implementación presentada por Cornia et al. in [147], el *Semantic Decoder* propuesto incluye una atención enmascarada (*Masked Self Attention*, MKSA), una multi-atención cruzada (*Multi-headed Cross Attention*, MCA), una capa FFN y una capa LN en cada bloque. La salida de cada bloque del *decoder* $\check{Y}_n$ es formulada como:

$$\begin{aligned} Y'_n &= LN(MKSA(Y_n)) \\ Z_n &= MCA(\check{X}'_n, Y'_n) \\ \check{Y}_n &= FFN(Z_n). \end{aligned} \tag{4.12}$$

Donde $\check{X}'_n$ es la salida del *Visual Encoder* definida en la Eq.(3).

4.6. *Mask-Self Attention* (MKSA)

La capa MKSA incluye una propiedad auto-regresiva, que permite que la atención sea calculada por cada *token* $\{t_1, t_2, .., t_J\}$ de Y_n basada en su posición. Como MKSA es un proceso de atención, el *embedding* semántico Y_n es mapeado con las matrices de *query* Q_Y, key K_Y, y value V_Y utilizando los parámetros entrenables

de $W_{qy} \in R^{D_M \times D_{qy}}, W_{ky} \in R^{D_M \times D_{ky}}$, and $W_{vy} \in R^{D_M \times D_{vy}}$ como se muestra a continuación:

$$Q_Y = Y_n \times W_{qy} \tag{4.13}$$

$$K_Y = Y_n \times W_{ky} \tag{4.14}$$

$$V_Y = Y_n \times W_{vy} \tag{4.15}$$

La operación del módulo MKSA es implementada así como una atención de producto punto escalado (*scale dot product attention*), como se muestra en la figura 4.3. La correlación y similitud de cada *token* es calculada mediante la operación de producto punto escalado, pero sólo se tienen en cuenta los *tokens* anteriores al actual.

La ecuación 4.16 define el proceso de atención propia (*self-attention*).

$$A(Q_Y, K_Y) = softmax\Big(\frac{Q_Y (K_Y)^T}{\sqrt{Dqy}} + M\Big) \tag{4.16}$$

Donde $M \in R^{J \times J}$, $M_{i,j} \in \{0, -\infty\}$. Aqui, M representa la atención enmascarada que permite un aprendizaje contextual restringido en el que sólo se admiten los *tokens* anteriores, como se muestra en la figura 4.4.

La salida de la capa de atención propia $H_{Yi} \in R^{D_M \times D_{qy}}$ es calculado como se muestra en la ecuación 4.18.

$$H_{Yi} = \textit{self-attention}(Y_n) = A(Q_Y, K_Y) \times V_Y \tag{4.17}$$

$$MHA(Q_Y, K_Y, V_Y) = Concat(H_{Y1}, H_{Y2}, .., H_{Yi})W^{\theta} \tag{4.18}$$

Donde Yi, representa el numero de cabezales de atención y W^{θ} el peso de proyec-

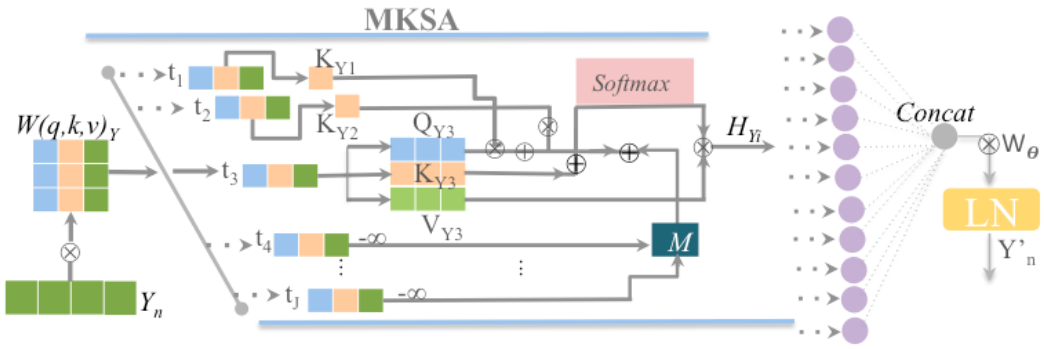


Figura 4.3: El módulo de MKSA recibe el *token* semántico actual y obtiene Q_Y, K_Y, y V_Y basado en la relevancia de los *tokens* subsecuentes $< t_3$. El valor de los *tokens* $> t_3$ se establece a -∞, valor de acuerdo a la matriz M.

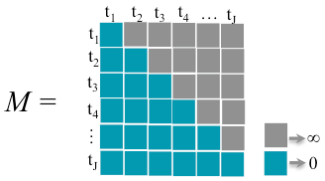


Figura 4.4: *M self-attention* Atención enmascarada y los valores asignados de acuerdo a la posición de los *tokens*.

ción entrenable.

4.7. Multi-headed Cross Attention (MCA)

La relevancia de las características visuales detectadas en cada bloque del *Visual Encoder* se asigna a una secuencia de *embedding* semántico Y'_n obtenido del módulo de MKSA a través del módulo de MCA. Se realiza una operación de suma para integrar todas las características visuales, definida como:

$$MCA(\check{X}'_n, Y'_n) = \sum_{i=1}^{n} \alpha_i \odot C(\check{X}'_i, Y'_n)$$
$$\alpha_i = \sigma(W_i[Y'_n, C(\check{X}'_i, Y'_n)] + b_i) \tag{4.19}$$

Donde α_i representa una matriz de pesos del mismo tamaño que la matriz resultante de la atención cruzada (*Cross-Attention*). Los pesos en σ modifican tanto el impacto individual de cada capa del *encoder* como la importancia relativa entre las distintas capas. La atención cruzada resultante del *encoder-decoder* es computarizada basada en los *queries* del *decoder* y los *keys* y *values* del *encoder* como se expresa a continuación:

$$C(\check{X}'_n, Y'_n) = A(W_{qy}Y'_n, W_k\check{X}'_n, W_v\check{X}'_n) \tag{4.20}$$

Como se muestra en la figura 4.5, el módulo MCA fusiona la información visual X'_n

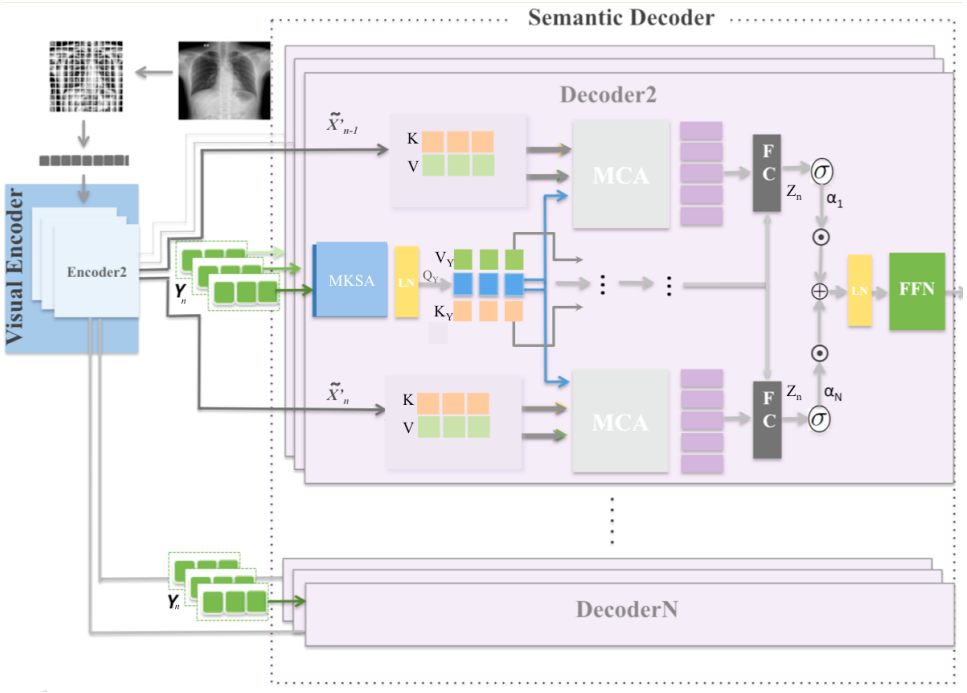


Figura 4.5: El módulo de MCA recibe el *token* Y_n el cual es procesado por el módulo de MKSA, y el Q_Y del *token* semántico es asociado con el k y v de las características visuales $\check{X}'_n$. El proceso implica alimentar cada bloque del *Semantic Decoder* con la salida de todos los bloques del *Visual Encoder*.

de todos los bloques del *visual Encoder* con cada Y'_n de los bloques del *Semantic Decoder*. El objetivo de esta operación es identificar la información semántica vinculada a las características visuales.

Capítulo 5

Diseño de Experimentos

5.1. Datasets

Después de realizar una exhaustiva revisión de las bases de datos, se destacaron las bases de datos IU-Xray [4] y MIMIC-CXR [69] como las más completas y utilizadas en la literatura. Ambos conjuntos de datos contienen imágenes de rayos X de tórax, así como los informes médicos correspondientes. Además, la disponibilidad, contenido y estructura permite identificarlas como la primera opción para esta investigación. Al ser los conjuntos de datos que más utilizan en el estado del arte, permite realizar comparaciones más justas en los modelos propuestos.

Para la base de datos de MIMIC-CXR, se utilizó la partición oficial y sólo se considero la sección de *findings* del informe médico.

Dado que no existe una partición oficial definida para el conjunto de datos de IU-Xray, se empleó el enfoque descrito en [6]. En concreto, el 70 % de los datos se destinó al entrenamiento, el 10 % a la validación y el 20 % a las pruebas. Además, se validó de que un mismo paciente no apareciera en varias particiones. Como cada paciente pueden tener una o más imágenes, sólo se restringió a dos imágenes por

paciente (lateral y frontal). Las imágenes se recortaron a 224 × 224 píxeles y se normalizaron considerando la media y la desviación estándar.

Para el pre-procesamiento del texto, se empleo la estrategia propuesta en [67], que consiste en cambiar a minúsculas los *tokens*, eliminar los caracteres especiales y filtrar los *tokens* con una frecuencia inferior a 3.

5.2. Modelos Base

5.2.1. Modelo MedViLL

El modelo de MedViLL [5], se caracteriza por incluir una arquitectura de tipo *Transformer-encoder* denominado *Bert*, el cuál se ha utilizado recientemente para tareas que involucran el área de visión y texto [122], [124],[148].

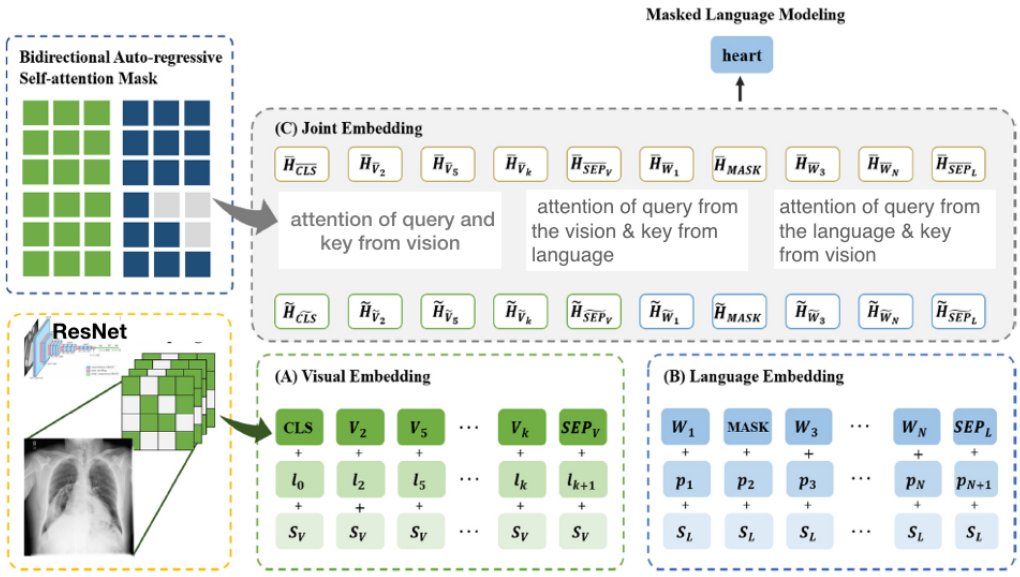


Figura 5.1: Arquitectura del modelo MedViLL. Referencia tomada de [5]

El modelo requiere emplear dos sub-tareas durante su entrenamiento [149]. Es-

tas tareas tienen la finalidad de alinear las características visuales y el texto mediante un pre-entrenamiento que consiste en recibir una imagen y una secuencia de texto como entrada. Donde antes de alimentar al modelo, se enmascara aleatoriamente una parte de texto (generalmente palabras). Posteriormente, se realiza el entrenamiento para que el modelo aprenda a predecir el texto que se encuentra enmascarado.

Representación de características visuales

Para obtener una representación de características visuales, el modelo utiliza una arquitectura RESNET-50 previamente entrenada en la tarea de clasificación de 14 patologías con la base de datos de *ChestXray-14*. Posteriormente las caracteristicas visuales son extraídas de la última capa. Al obtener las características visuales de la imagen denominadas V, se agrega un valor posicional denominado l el cual permite identificar el orden de la imagen.

$$\begin{aligned} v &= \{v_1, v_2, ..., v_K\}, v_i \in R^c \\ l &= \{l_1, l_2, ..., l_K\}, l_i \in R^c \end{aligned} \tag{5.1}$$

Donde K indica el número de características visuales y c el tamaño de canales. Entonces la representación de las características visuales que alimentan al modelo se identificaron como $\widetilde{v} = \{\widetilde{v}_1, \widetilde{v}_2, ..., \widetilde{v}_K\}$.

$$\widetilde{v}_i = v_i + l_i + s_v \tag{5.2}$$

Donde s_V es el identificador del segmento visual. Este identificador es agregado con la finalidad de identificar la información visual del la textual y se agrega des-

pués de que las características visuales son identificadas con su posición.

Representación textual

Para representar las secuencias de texto del reporte asociado a cada imagen, se utiliza la estrategia de [2]. Donde w representa el reporte médico compuesto por la sección *findings*. Cada reporte w, es dividido en N palabras *tokens* utilizando el *tokenizador WordPiece* [150]. Posteriormente, cada *token* es convertido a una representación vectorial como $W = \{W_1, W_2, ..., W_N\}$,$W_i \in R^d$, donde d es la dimensión del vector. Posteriormente, se concatena un el valor posicional p, donde $p = \{p_1, p_2, ..., p_N\}$, $p_i \in R^d$ y el identificador del segmento de lenguaje como S_L. Finalmente la representación textual que alimenta al modelo es representada por:

$$\check{W}_i = W_i + p_i + S_L$$

Representación visual/textual

Una vez que se obtiene la representación visual (*embedding* visual) $\check{V} \in R^d$ y la representación textual (*embedding* de texto) $\check{W} \in R^d$, se concatenan ambos segmentos para formar una sola secuencia de entrada al modelo. Para identificar la secuencia de entrada y los segmentos visual y textual se agregan los *tokens* especiales de CLS y SEP como se muestra a continuación:

$$\check{H} = \{\overline{CLS}, \check{V}_1, ..., \check{V}_K, \overline{SEP}_V, \check{W}_1, ..., \check{W}_N, \overline{SEP}_L \in R^{d \times S}\} \tag{5.3}$$

Donde $S = N + K + 3$. El final *embedding* contextualizado es:

$$\overline{H} = \{\overline{CLS}, \overline{V}_1, ..., \overline{V}_K, \overline{SEP}_V, \overline{W}_1, ..., \overline{W}_N, \overline{SEP}_L\} \tag{5.4}$$

5.2.2. Modelo R2GEN

El modelo de R2GEN esta basado en la estructura base del *Transformer* propuesto por [6]. Dentro de su arquitectura se utiliza un modelo CNN para obtener una representación visual de la imagen, además se incluye una memoria auxiliar que permite influenciar la ponderaciones del módulo de atención del *Transformer*. La figura 5.2 muestra una representación general de esta arquitectura.

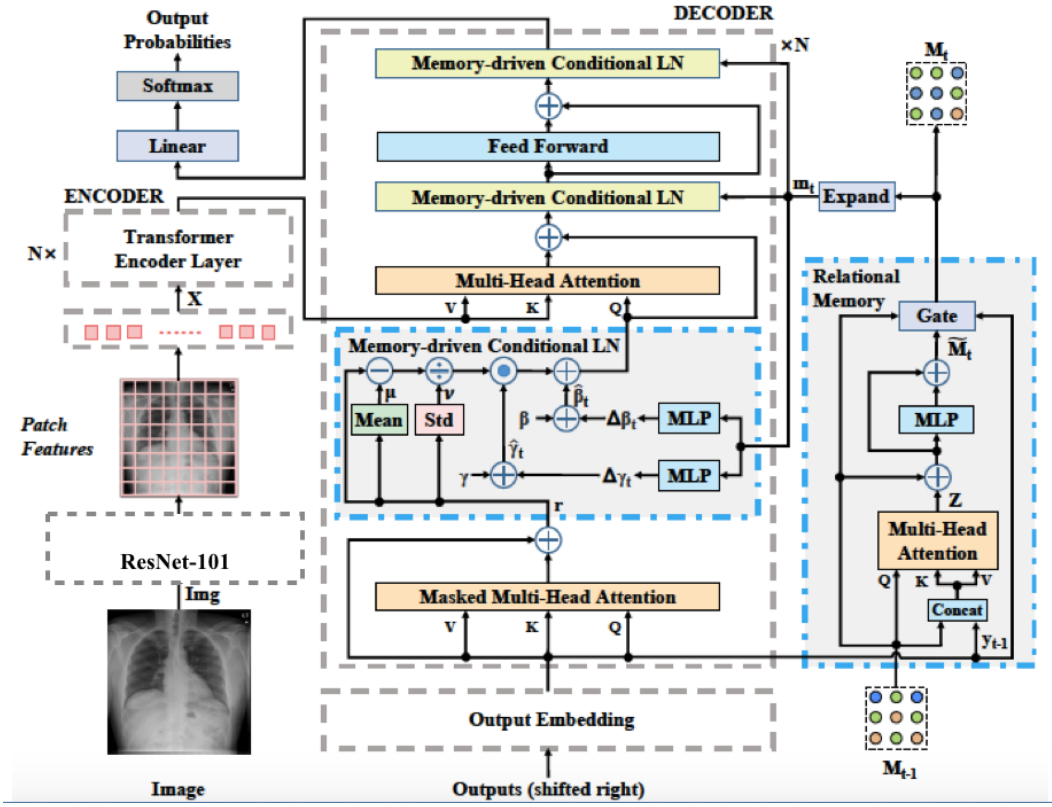


Figura 5.2: Arquitectura del modelo R2GEN. Referencia tomada de [6]

Extractor de características Visuales

La arquitectura de R2GEN utiliza una CNN pre-entrenada denominada ResNet para obtener la representación visual (*embedding* visual) de la imagen Img. El pro-

ceso se representa mediante la siguiente ecuación:

$$f_v(Img) = \{x_1, x_2, ..., x_S\} \tag{5.5}$$

Donde $f_v(\cdot)$ representa el extractor de características visuales.

***Encoder* (Codificador)**

R2GEN utiliza la estructura base del *Transformer encoder* propuesto en [1], donde los estados ocultos (*hidden states*) h_i contienen las características visuales obtenidas del extractor visual definidas en la ecuación 5.5. La representación de la información de h_i se refleja como:

$$\{h_1, h_2, ..., h_S\} = f_e\{x_1, x_2, ..., x_S\} \tag{5.6}$$

Donde $f_e(\cdot)$ se refiere al *encoder* del *Transformer*

***Decoder* (Decodificador)**

El *decoder* utilizado en R2GEN esta basado en la estructura original del *decoder* propuesta en [1]; sin embargo se modificó la capa de normalización (LN) por una nueva capa identificada como (*Memory-driven Conditional Layer Normalization*, MCLN). Esta modificación se realizó a cada bloque del *decoder*. El proceso se establece como:

$$y_t = f_d(h_1, ..., h_S, MCLN(RM(y_1, ..., y_{t-1}))) \tag{5.7}$$

Donde $f_d(\cdot)$ se refiere al *decoder* y al proceso MCLN.

MCLN (*Memory-driven Conditional Layer Normalization*)

El proceso MCLN se incorpora a la arquitectura del modelo con el objetivo de incluir la memoria auxiliar en el proceso del *decoder* y mejorar sus predicciones. Al incluir esta memoria se requiere incluir los parámetros $gamma$ y $beta$ para escalar las representaciones aprendidas. Por cada *decoder* del modelo se utilizan 3 MCLNs, donde la primer salida de MCLN se fusiona con el parámetro $query$ para alimentar los múltiples cabezales de atención junto con los estados ocultos del codificador de key y $value$. El MCLN recibe la salida m_t obtenida de la memoria auxiliar M_t. Posteriormente un MLP es utilizado para predecir un cambio entre $\Delta\gamma_t$ de γ_t. Este proceso se define como:

$$\Delta\gamma_t = f_{mlp}(m_t) \tag{5.8}$$

$$\check{\gamma}_t = \gamma + \Delta\gamma_t \tag{5.9}$$

Al igual $\Delta\beta_t$ de $\check{\beta}_t$ se calcula mediante:

$$\Delta\beta_t = f_{mlp}(m_t) \tag{5.10}$$

$$\check{\beta}_t = \beta + \Delta\beta_t \tag{5.11}$$

Una vez que se obtiene $\check{\beta}_t$ y $\check{\gamma}_t$ estos son aplicados a los resultados obtenidos en los cabezales de atención de los resultados obtenidos previamente como se muestra a continuación:

$$f_{mcln(r)} = \check{\gamma}_t \odot \frac{r - \mu}{v} + \check{\beta}_t \tag{5.12}$$

Donde r representa la salida del bloque anterior; μ y v, son la media y la desviación estandar de r, respectivamente. El resultado obtenido en $f_{mcln}(r)$ del modulo MCLN alimenta al siguiente modulo MCLN.

5.3. Configuración de Modelos

5.3.1. Modelo Propuesto: ETB-MII

Se implementó el *Visual Encoder* pre-entrenado (ViT) [3] en *Imagenet* y el *Semantic Decoder* pre-entrenado (GPT-2) [32] mediante una arquitectura *encoder-decoder*. Posteriormente, al modelo ETB-MII se le realizó un entrenamiento fino (fine-tuned) con la base de datos de IU X-ray y MIMIC-CXR. Se estableció el optimizador Adamw con una tasa de aprendizaje (*learning rate)* de $1e{-}4$ y con un decaimiento (*decay*) de $5e^{-5}$. Se definió un tamaño de lote de 6 y número de épocas a 10. El parámetro de *beam* se estableció a 4. Se seleccionó el modelo que obtuvo los mejores resultados con el conjunto de validación el cuál fue utilizado para probar al modelo con la partición de datos de prueba. El modelo ETB-MII se implementó en Pytorch y para su entrenamiento se utilizó un GPU 11GB GeForce RTX 2080Ti.

5.3.2. Modelo Base: MedViLL

Con la finalidad de implementar una estructura robusta que permita generar reportes médicos mediante imágenes de rayos X y comparar el modelo propuesto, se implementó localmente el modelo MEDVILL [5].

Debido a que el modelo de MEDVILL es multimodal y fue entrenado en tareas de clasificación, preguntas-respuestas (QA) y generación de texto el modelo requiere ser pre-entrenado en determinadas tareas para mejorar su desempeño. En lo que compete a esta investigación el modelo únicamente fue utilizado para la generación de texto a partir de una imagen. Por lo anterior, la implementación no fue replicada en su totalidad solo la recomendada por el autor [5] para la generación de reportes médicos.

Siguiendo las indicaciones publicadas en el sitio oficial del autor [1]; el modelo fue replicado utilizando un el base de MedViLL. Para alinear las características visuales con las características textuales el autor sugiere aplicar un modelado de lenguaje enmascarado (MLM) basado en los trabajos de [123],[148]. La tarea de MLM consiste en remplazar aleatoriamente el 15 % de los *tokens* textuales de entrada $\{w_1, ...w_N\}$ o el *token* original con una probabilidad de 80 %, el cual es sustituido con un token especial denominado MASK. El objetivo de este pre-entrenamiento es que el modelo sea entrenado para identificar el *token* enmascarado aprendiendo el contexto de las secuencias de entrada textuales minimizando las diferencias.

$$L_{MLM(\theta} = -E_{v,w} \sim D[logP_\theta(w_m|v, m/)] \tag{5.13}$$

Donde θ son parámetros entrenables del modelo MedViLL.

Para la representación visual de la imagen se utilizó una red CNN ResNet-50 pre-entrenada en *Imagenet*, la cual se identifica como extractor de características[5]. La imagen de entrada con un tamaño de $(512 \times 512 \times 3)$ alimenta al la red ResNet-50 y se extrae el última capa obteniendo una representación de tamaño $(16 \times 16 \times 2048)$ la cual pasa por una capa *flattened*. Para la representación del texto se establece un máximo tamaño de la secuencia de entrada a 253 *tokens*. Para unir la representación visual y textual se utiliza el modelo de BERT el cuál contiene 12 cabezales de atención, un tamaño de 768 en la representación oculta de cada token (*embedding hidden size*) y un *drop-out* de 0.1. Se utiliza el optimizador AdamW con una tasa de aprendizaje *learning rate* de $1e^{-5}$.

La implementación de acuerdo a la publicación del autor fue realizado utilizan-

[1]https://github.com/SuperSupermoon/MedViLL

do 8 GPUs RTX-3090, sin embargo para esta implementación se utilizó una tarjeta GPU 2080-Ti y se trabajó directamente con el modelo publicado en el sitio oficial. [2]

5.3.3. Modelo Base: R2GEN

El modelo utilizó una red CNN ResNet-101 [137], pre entrenada en *Imagenet* como extractor visual. El modelo *Transformer* es inicializado con pesos aleatorios. La memoria relacional del modelo incluye 8 cabezales de atención. Para su entrenamiento se estableció el optimizador de Adam, una tasa de aprendizaje de $5e^{-5} - 1e^{4}$ con una tasa de decremento (*Decay*) de 0.8. Y un tamaño de *beam* igual a 3.

En el trabajo publicado por el autor [6] no se especifica el equipo utilizado para entrenar al modelo, sin embargo para realizar esta implementación se siguieron las instrucciones publicadas en el sitio oficial del autor[3] y se entrenó con una tarjeta GPU 2080Ti.

5.4. Métricas de Evaluación

Para evaluar el desempeños del modelo propuesto ETB-MII y los modelos base, se implementó el conjunto de herramientas de evaluación proporcionado por COCO en [151]. Este conjunto de herramientas esta diseñada en lenguaje Python e incluye las métricas de Generación de Lenguaje Natural (NLG) como BLEU [38], METEOR [96], ROUGE-L [95] y CIDEr [97].

[2]https://github.com/SuperSupermoon/MedViLL
[3]https://github.com/cuhksz-nlp/R2Gen

Capítulo 6

Resultados

6.1. ETB-MII aplicando la estrategia para el aumento de datos (DAS)

Para evaluar la utilidad de la estrategia de aumento de datos (DAS), se realizaron diversos experimentos que implicaron aumentar los datos del conjunto de entrenamiento en un 15 %, 20 %, o 30 % del conjunto de datos de IU X-ray, y un 5 %, 10 %, o 15 % del conjunto de datos de MIMIC-CXR. La tabla 6.1 presenta los resultados obtenidos al aplicar la estrategia DAS con cada porcentaje de aumento de datos. La columna *"strategy"* indica un *"None"* cuando los datos no fueron pre-procesados y se incluye un "PPS", cuando los datos fueron pre-procesados. Además, en esta columna se indica el porcentaje utilizado para el aumento de datos.

Como se puede observar para ambos conjuntos de datos, al aplicar un pre-procesamiento en la imagen y el texto se obtienen mejores resultados para la mayoría de las métricas NLG. Esto se debe a que al aplicar el pre-procesamiento en los datos, se contribuye a eliminar ruido, datos atípicos, irrelevantes o inconsistentes que podrían afectar negativamente el rendimiento del modelo. Por ejemplo, las pa-

labras con errores ortográficos identificadas en la sección de *findings* del reporte se identifican y remueven al establecer una frecuencia menor a 3.

Por otra parte, se observa que al aplicar la estrategia DAS se obtiene un incremento general en los resultados obtenidos para ambos conjuntos de datos. Esto se debe a que al generar los datos sintéticos por DAS, el modelo se somete a una amplia variedad de datos lo que permite capturar una representación más completa y variada del vocabulario. Lo anterior contribuye a una mayor capacidad de generalización del modelo. Dentro de los porcentajes establecidos para el aumento de datos los mejores resultados se obtuvieron con un 15 % para IU X-ray y un 10 % para MIMIC-CXR.

Analizando el comportamiento de las métricas aplicando la estrategia DAS, la métrica CIDEr es la que más destaca en su desempeño para ambos conjuntos de datos. Una de las posibles causas es que al agregar datos adicionales que representan diferentes estilos de escritura, jergas, y estructuras sintácticas, se enriquece la diversidad semántica y sintáctica del conjunto de entrenamiento.

Finalmente, es importante destacar que aunque en ambos conjuntos se obtuvieron mejores resultados con la estrategia DAS, se identifica un menor incremento en el conjunto de datos de MIMIC-CXR. Sin embargo, esto no es solamente con el modelo propuesto, esta tendencia se identifica en la mayoría de los trabajos presentados en el estado de arte. El conjunto de datos de MIMIC-CXR, se caracteriza por ser una colección grande de datos pero a la vez por incluir datos de menor calidad y consistencia.

Tabla 6.1: Resultados con el conjunto de datos de prueba de IU X-ray y MIMIC-CXR utilizando las métricas de NLG, donde se muestra el desempeño del modelo ETB-MII utilizando la estrategia de aumento de datos (DAS). La estrategia definida como *"None"* muestra el rendimiento del modelo sin eliminar las palabras con una frecuencia <3, sin remover los caracteres especiales y sin rotar la imagen; mientas que la siglas *"PPS"* indican que se aplicó el pre-procesamiento en texto e imagen. La métrica BLEU se representa con B, y M corresponde a la métrica METEOR

.

Database	Strategy	B1	B2	B3	B4	M	ROUGE-L	CIDEr
IU X-ray	None	0.235	0.138	0.092	0.066	0.192	0.209	0.204
	PPS & 0 %	0.420	0.281	0.206	0.158	0.186	0.316	0.703
	PPS & 15 %	**0.437**	**0.300**	**0.222**	**0.178**	**0.194**	**0.333**	**0.864**
	PPS & 20 %	0.434	0.295	0.222	0.176	0.193	0.328	0.858
	PPS & 30 %	0.425	0.286	0.211	0.165	0.189	0.327	0.745
MIMIC	None	0.281	0.134	0.654	0.032	0.102	0.161	0.095
	PPS & 0 %	0.317	0.181	0.109	0.067	0.129	0.212	0.101
	PPS & 5 %	0.326	0.183	0.108	0.070	0.128	0.213	0.288
	PPS & 10 %	**0.340**	**0.192**	**0.120**	**0.080**	**0.136**	**0.214**	**0.336**
	PPS & 15 %	0.326	0.184	0.108	0.069	0.134	0.212	0.289

6.2. ETB-MII vs Modelos Base

Los resultados de esta comparación se presentan en la tabla 6.2. Cabe señalar que la evaluación se realizó utilizando el mismo GPU y las mismas particiones de datos para los modelos de MedViLL, R2GEN y ETB-MII, lo que garantiza una evaluación equitativa y sólida. Sin embargo, existe una ligera discrepancia entre los resultados de la implementación realizada de MedViLL y R2GEN con respecto a resultados publicados oficialmente. De acuerdo a las especificaciones de entrenamiento del modelo MedViLL [5], las diferencias presentadas en los resultados publicados y los obtenidos en la implementación podrían deberse a que el modelo de MedViLL fue entrenado utilizando 8 RTX-3090 GPUs con un tamaño de lote de 128. Para el modelo R2GEN, los recursos de hardware computacional no se publicaron oficialmente.

Al examinar los resultados obtenidos de cada modelo se evaluaron los reportes

generados por los modelos base y el modelo propuesto. En los resultados obtenidos de la implementación del modelo R2GEN se obtuvo un mejor rendimiento que el modelo MedViLL. Un hallazgo digno de mención en relación al modelo R2GEN es su tendencia a generar textos repetitivos. Analizando la arquitectura del modelo, un factor que contribuye a ello puede ser la memoria auxiliar ya que conserva la información de textos anteriores pero también incorpora una gran cantidad de información redundante que puede resultar perjudicial [6]. Otra consideración importante es que la utilización de un sistema de memoria basado en conexiones recurrentes aumenta la complejidad y los gastos computacionales asociados al proceso. Además, el uso de este tipo de memoria merma la capacidad de aprender dependencias de largo alcance entre los datos, como se ha demostrado en varios estudios [3],[19],[20].

Tabla 6.2: Resultados del conjunto de datos de prueba de IU X-ray y MIMIC-CXR con las métricas de NLG que muestran el desempeño de los modelos base y ETB-MII. La métrica BLEU se representa con B, y M corresponde a la métrica METEOR.

Dataset	Model	B1	B2	B3	B4	M	ROUGE-L	CIDEr
IU X-ray	R2GEN	0.427	0.257	0.172	0.124	0.166	**0.335**	0.354
	MedViLL	0.221	0.160	0.120	0.080	0.113	0.187	0.176
	ETB-MII	**0.437**	**0.300**	**0.222**	**0.178**	**0.194**	0.333	**0.864**
MIMIC	R2GEN	**0.343**	**0.194**	**0.121**	**0.082**	0.131	**0.246**	0.238
	MedViLL	0.289	0.168	0.102	0.060	0.123	0.223	0.198
	ETB-MII	0.340	0.192	0.120	0.080	**0.136**	0.214	**0.336**

6.3. ETB-MII vs Modelos del estado del arte

La tabla 6.3 muestra los resultados obtenidos del modelo propuesto ETB-MII y los resultados publicados de los modelos más relevantes en el estado del arte utilizando las métricas de NLG con las bases de datos de IU X-ray y MIMIC-CXR.

Tabla 6.3: Comparación con los modelos del estado del arte con el conjunto de datos de prueba de IU X-ray y MIMIC-CXR. La métrica BLEU se representa con B, y M corresponde a la métrica METEOR. El signo "-representa que no hay resultados de evaluación en esa métrica. Un valor más alto es mejor, y los mejores resultados se resaltan en negro.

IU X-ray							
MODEL	B1	B2	B3	B4	M	ROUGE-L	CIDEr
S&T [152][153]	0.216	0.124	0.087	0.066	-	0.306	0.294
SA&T[80][154]	0.304	0.177	0.112	0.077	-	0.249	0.083
AdaAtt[155]	0.220	0.127	0.089	0.068	-	0.308	0.295
COAtt [9][153]	0.455	0.288	0.205	0.154	-	0.369	0.277
HRGR[67]	0.438	0.298	0.208	0.151	-	0.322	0.343
CMAS[153]	0.464	0.301	0.210	0.154	-	0.362	0.275
KERP[156]	0.482	0.325	0.226	0.162	-	0.339	0.28
CDGPT2[13]	0.387	0.245	0.166	0.111	0.164	0.289	0.257
MedViLL[5]	-	-	-	0.049	-	-	-
MedviLL**	0.221	0.160	0.120	0.080	0.113	0.187	0.176
R2Gen[6]	0.47	0.304	0.219	0.165	0.187	0.371	-
R2Gen**	0.427	0.257	0.172	0.124	0.166	0.335	0.354
PPKED[157]	0.483	0.315	0.224	0.168	-	0.376	0.351
KEMHAM[154]	**0.496**	**0.327**	**0.238**	**0.178**	-	**0.381**	0.382
ASGK[158]	-	-	-	0.125	-	0.279	0.306
ETB-MII	0.437	0.300	0.222	**0.178**	**0.194**	0.333	**0.864**
MIMIC-CXR							
S&T [152][154]	0.256	0.157	0.102	0.070	-	0.249	0.063
SA&T[80] [154]	0.304	0.177	0.112	0.077	-	0.249	0.083
AdaAtt [155]	0.311	0.178	0.111	0.075	-	0.250	0.073
TopDown [159] [154]	0.280	0.169	0.108	0.074	-	0.250	0.073
R2Gen [6]	0.353	0.218	0.145	0.103	**0.142**	0.277	-
R2Gen**	0.343	0.194	0.121	0.082	0.131	0.246	0.238
PPKED[157]	0.36	0.224	0.149	0.106	-	**0.284**	-
MedViLL**	0.289	0.168	0.102	0.060	0.123	0.223	0.198
CA[160]	0.35	0.219	0.152	0.109	-	0.283	-
CMCL [161]	0.334	0.217	0.14	0.097	-	0.281	-
AlignTransformer [20]	**0.378**	**0.235**	**0.156**	0.112	-	0.283	-
M2TR [162]	**0.378**	0.232	0.154	0.107	-	0.272	-
KEMHAM[154]	0.363	0.228	**0.156**	**0.115**	-	**0.284**	0.203
ETB-MII	0.340	0.192	0.120	0.080	0.136	0.214	**0.336**

Los resultados de evaluación que reporta el modelo KEMHAM [154] muestran un mejor desempeño en las métricas de BLEU y ROUGE-L en el conjunto de datos de IU X-ray. El modelo de KEMHAM consiste en cuatro módulos principales que integran las características visuales de la imagen de rayos X con conocimientos generales y específicos. De acuerdo con la arquitectura del modelo, en el proceso de extracción de características visuales intervienen tres CNNs, un codificador BERT pre-entrenado en texto clínico y un modelo de red neuronal de grafos (GNN). Estos modelos son utilizados para obtener las características visuales que servirán como entrada al decodificador. Como se puede inferir, el modelo KEMHAM se caracteriza por utilizar una arquitectura compleja. Considerando lo anterior y de acuerdo a la Tabla 6.3, se identifica que el modelo ETB-MII refleja una discrepancia marginal en la métrica BLEU-3 con un 6.72 %, mientras que en BLEU-4, ETB-MII obtiene el mismo resultado que KEMHAM. Por el contrario, el modelo ETB-MII demuestra un desempeño superior al resto de los modelos en lo que respecta a las métricas de CIDEr y METEOR, con un aumento de hasta el 226 % y el 103 % en estas métricas respectivamente como se muestra en la figuras 6.1 y 6.2.

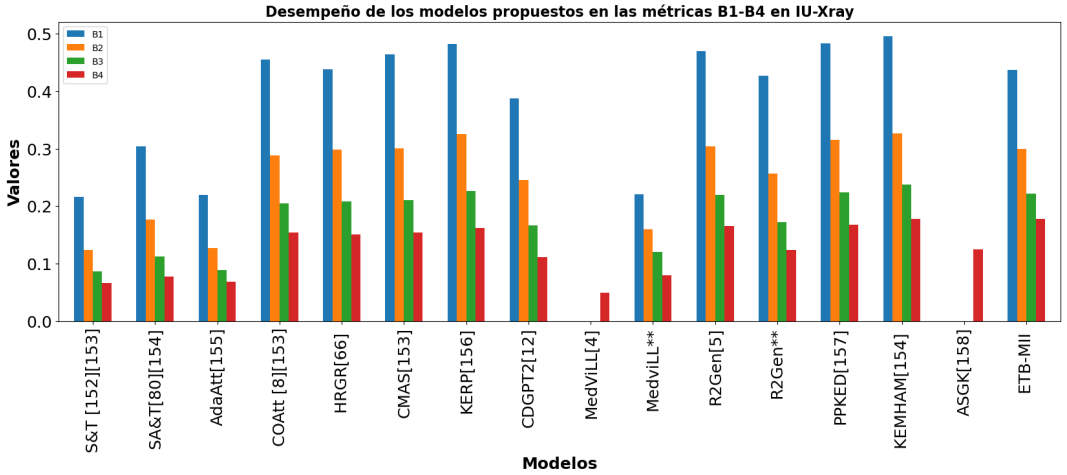


Figura 6.1: Representación gráfica del desempeño de los modelos propuestos en las métricas NLG en IU-Xray.

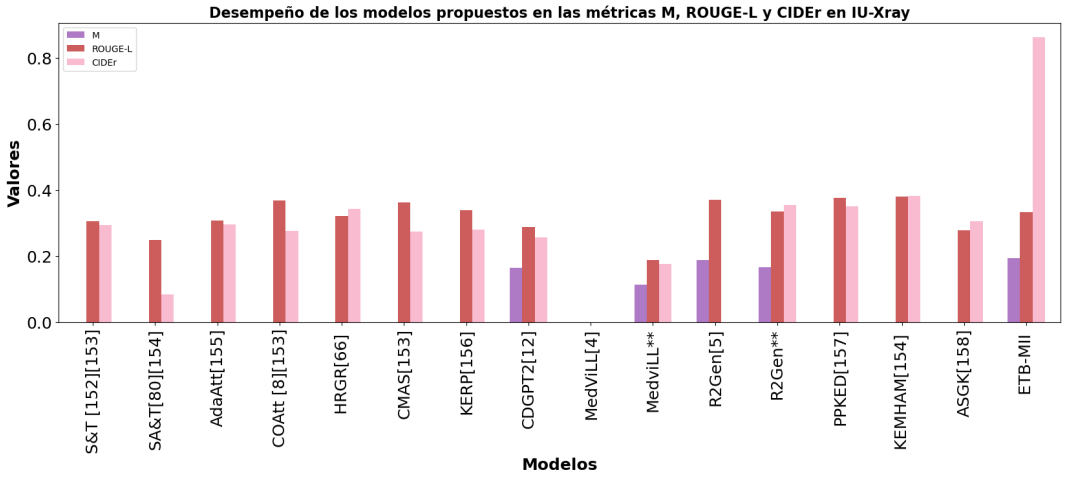


Figura 6.2: Representación gráfica del desempeño de los modelos propuestos en las métricas M, ROUGE-L y CIDEr en IU-Xray.

En el conjunto de datos MIMIC-CXR, ETB-MII muestra la misma tendencia para la métrica CIDEr, con una mejora del 41 % como se puede observar visualmente en la figura 6.4. Según los resultados publicados del modelo R2GEN [6], se obtuvo el mejor desempeño en la métrica de METEOR. No obstante, ETB-MII supera a la implementación local de R2GEN en un 3.8 %. Por otra parte, el modelo AlignTransformer [20] alcanza los mejores valores para la métrica BLEU como se muestra en la figura 6.3, y el modelo PPKED [157] para la métrica ROUGE-L.

Sin embargo, al igual que KEMHAM, AlignTransformer incluye una arquitectura compleja. Esta arquitectura está conformada por una red ResNet-50 pre-entrenada en *Imagenet* y con un *fine-tuning* utilizando el conjunto de datos CheXpert; además incorpora una CNN, un módulo de atención (*multi-head attention*) que permite alinear las características visuales con las etiquetas de la imagen y un modelo de *Transformer* que incluye sólo las capas del *decoder*. Por otro lado, el modelo PPKED sigue la misma configuración que el modelo KEMHAM, ya que incluye una red ResNet-152, un *Transformer* BERT, una GNN y un *Transformer decoder*.

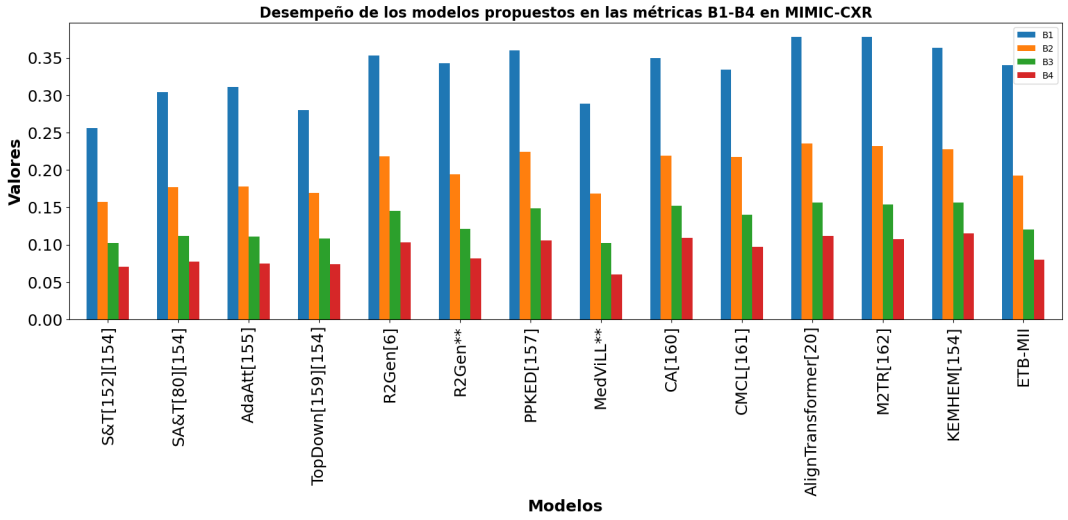


Figura 6.3: Representación gráfica del desempeño de los modelos propuestos en la métrica de BLEU en MIMIC-CXR.

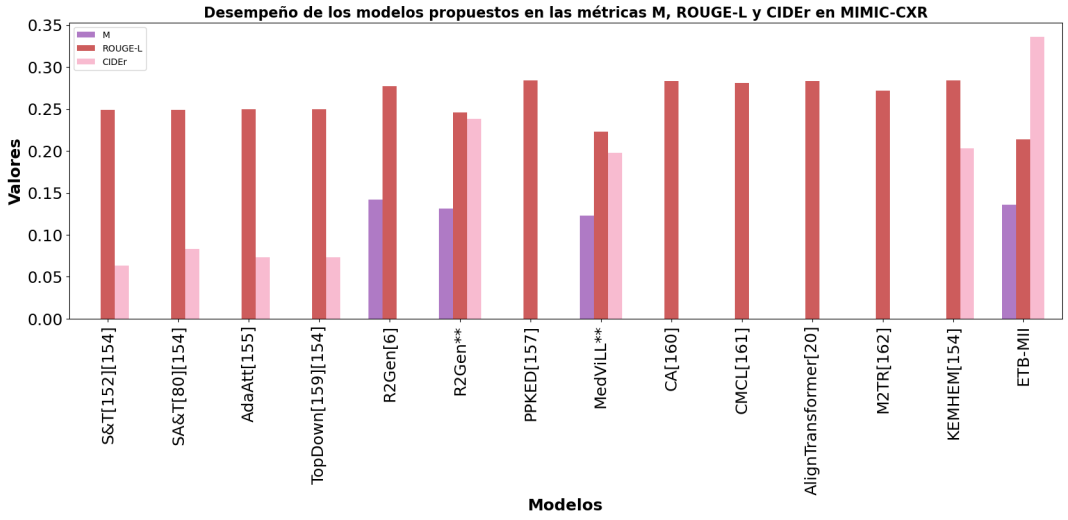


Figura 6.4: Representación gráfica del desempeño de los modelos propuestos en las métricas M, ROUGE-L y CIDEr en MIMIC-CXR.

Es evidente que emplear una estructura compleja para la tarea de generación de reportes médicos no garantiza una mejora significativa en el desempeño, ni altas ponderaciones en las métricas NLG. Por otro lado, ETB-MII utiliza una arquitectura simple al integrar modelos pre-entrenados que permiten reducir el procesamiento y que producen un rendimiento competitivo en las métricas más críticas.

6.4. Análisis de resultados

Para realizar un análisis de la eficacia del modelo ETB-MII, se compararon algunos de los textos generados (reportes generados) comparados con el texto objetivo (*Ground Truth*, GT). También se incluyó un análisis de los reportes generados por el mejor modelo base: R2GEN.

La figura 6.5 incluye tres ejemplos de la imágenes de rayos X, las etiquetas (GT), los reportes generados del modelo R2GEN[3], y los reportes generados de ETB-MII. Los textos que conforman los reportes generados por los modelos que coinciden con la etiqueta (GT), se identifican con colores idénticos. Algunas de las sentencias con colores iguales representan terminologías médicas diferentes pero con el mismo significado. Lo anterior fue verificado por un experto médico. [4]

Por ejemplo, en el caso 1: *"there is no focal consolidation"* está resaltado en color azul en el GT, y en el reporte generado por el modelo ETB-MII, el texto identificado con azul es: *"there is no focal airspace consolidation"*. Esto significa que ambas sentencias proporcionan el mismo diagnóstico pero tienen diferente terminología. El reporte generado por ETB-MII que incluye las palabras *"airspace"* y *"the lungs are clear"*, que están marcadas en color negro no están incluidas en el GT. No obstante, son un complemento útil que mejora la claridad y comprensibilidad del texto. De acuerdo al experto médico que participó en la verificación de los reportes generados, un dato relevante es que el modelo propuesto ETB-MII predice la presencia de cambios degenerativos mínimos en la columna toráxica. Sin embargo, esta observación no se incluye en el GT, pero es evidente en la radiografía.

[3]Implementación realizada localmente

[4]Especialista Médico Familiar con 10 años de experiencia en el IMSS Unidad Medicina Familiar 18, ubicado en Av. México 98, La Cruz, La Magdalena Contreras, 10800 CDMX. México.

En el segundo caso hay variaciones en ambos reportes generados; el GT es: *"no focal consolidation pneumothorax or large pleural effussion"*, y el modelo R2GEN genera lo siguiente: *"the lungs are clear of focal air space disease pneumothorax or large pleural effussion"*, mientras que el reporte generado por el modelo ETB-MII es: *"the lungs are normally inflated and clear"*, lo que indica que el paciente está libre de la enfermedad. Aunque ambos modelos predicen la ausencia de anomalías utilizando una terminología diferente al GT, ambos proporcionan un diagnóstico que indica la ausencia de enfermedad. En concreto, ETB-MII genera un texto conciso y comprensible, lo que demuestra su eficacia para predecir los textos médicos de la sección de *findings*.

El tercer ejemplo muestra que ETB-MII genera el siguiente texto: *"a mild spondylosis"*. Esta palabra está subrayada en negro porque no está en el GT. Este término se refiere a una afección en la que se produce un desgaste anormal del cartílago y el hueso, y está relacionada con los cambios degenerativos de la columna vertebral [163]. Por el contrario, R2GEN genera un reporte especificando una condición normal sin anomalías.

Como parte de este análisis también se investigó por qué el modelo R2GEN obtiene puntuaciones más altas en las métricas BLEU y ROUGE-L, a pesar de las similitudes entre los reportes generados del modelo R2GEN y los generados por el modelo ETB-MII. La hipótesis es que el modelo tiende a memorizar frases esenciales y, por tanto, produce informes repetitivos. Para verificar esta hipótesis, se utilizaron representaciones visuales del GT y de los reportes generados por los modelos R2GEN y ETB-MII. En primer lugar, se seleccionaron aleatoriamente 50 imágenes de rayos X para generar los reportes utilizando ambos modelos, R2GEN y ETB-MII. A continuación, los reportes generados y los GT se representaron en *embeddings* y se proyectaron en un espacio bidimensional para comparar similitudes. Para este

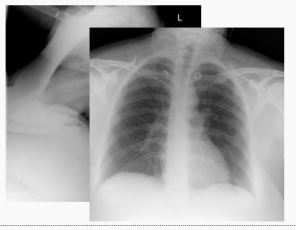
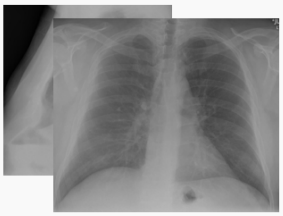
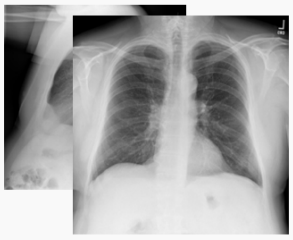
	Ground-truth	Prediction (R2GEN Model)	Prediction (ETB-MID)
	normal cardiomediastinal silhouette there is no focal consolidation there are no XXXX of a large pleural effusion. there is no pneumothorax there is no acute bony abnormality seen	normal cardiomediastinal silhouette there is no focal consolidation there are no xxxx of a large pleural effusion there is no pneumothorax there is no acute bony abnormality seen	the lungs are clear there is no focal airspace consolidation no pleural effusion or pneumothorax normal cardiomediastinal silhouette there are minimal degenerative changes of the thoracic spine no acute bony abnormality
	the cardiomediastinal silhouette is normal in size and contour no focal consolidation pneumothorax or large pleural effusion negative for acute displaced rib fracture	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size the lungs are clear of focal airspace disease pneumothorax or pleural effusion there are no acute bony findings	the cardiomediastinal silhouette and vasculature are within normal limits for size and contour the lungs are normally inflated and clear no acute bone abnormality
	the heart pulmonary xxxx and mediastinum are within normal limits . there is no pleural effusion or pneumothorax . there is no focal air space opacity to suggest a pneumonia . there are mild degenerative changes of the spine .	the lungs are clear . there is no pleural effusion or pneumothorax . the heart and mediastinum are normal . the skeletal structures are normal .	the heart is normal in size. the mediastinum is unremarkable. the lungs are clear. mild spondylosis is noted. there is mild degenerative changes of the spine. no pleural effusions or pneumothorax

Figura 6.5: Ejemplos de los reportes generados del modelo R2GEN y ETB-MII. Los textos generados por los modelos que son iguales a la etiqueta (GT) son identificados con el mismo color. El texto subrayado son palabras no incluidas en GT. Como puede ver el lector, los reportes generados por el modelo ETB-MII mantienen un diagnóstico preciso, pero el orden de las frases y la diversidad del texto varían.

análisis se empleo un *framework* de Python que permite generar la transformación de frases a nivel de sentencia [164], y el conocido método de Análisis de Componentes Principales (PCA) [165] para obtener una representación vectorial que permita visualizar cada reporte. La figura 6.6 muestra dos gráficos; en el gráfico de la izquierda se incluyen la representación de los reportes generados de R2GEN. Como se puede observar esta gráfica presenta menos de 50 puntos porque muchas de las representaciones de cada reporte generado son las mismas, lo que resulta en un solo punto por múltiples reportes. Por otro lado, el modelo ETB-MII de la figura 6.6 (diagrama de dispersión derecho) muestra una mayor correlación con los textos del GT en términos de similitud de frases. Dado que esta representación incluye 50 radio-

grafías de rayos X, es evidente que el modelo R2GEN repite el mismo texto en más de un reporte generado, mientras que el modelo ETB-MII refleja los 50 puntos, donde cada punto representa el reporte generado de las imágenes evaluadas. Además, los reportes generados por el modelo ETB-MII están positivamente correlacionados con el GT, como se aprecia en la forma de su representación.

Para culminar con este análisis se obtuvieron dos representaciones de mapas de calor con la finalidad de identificar patrones recurrentes en la generación de texto por parte de los modelos. Esto puede ayudar a comprender e identificar la similitud entre los reportes generados de cada modelo. Para obtener los mapas de calor se utilizaron los mismos 50 reportes generados por cada modelo. Posteriormente se obtiene la representación de cada reporte mediante el *framework Sentence-BERT* [164]. El mapa de calor que se presenta en la figura 6.7 proporciona información adicional sobre la visualización de los reportes generados por R2GEN. Como se muestra, R2GEN genera frases idénticas para distintas imágenes de rayos X. La frase más frecuente es *"heart size normal, lungs are clear, xxxx are normal, no pneumonia effusions edema pneumothorax adenopathy nodules or masses"* que aparecen en el 32 % en los reportes generados de los ejemplos. Además, la frase *"heart size normal, lungs are clear, xxxx are normal, no pneumonia effusions edema pneumothorax adenopathy nodules or masses"* aparece en el 18 % de los reportes generados de los ejemplos, mientras que *"the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size, the lungs are clear of focal airspace disease pneumothorax or pleural effusion, there are no acute bony findings"* aparece en el 10 % de los reportes generados como *"the cardiomediastinal silhouette is normal in size and contour, no focal consolidation pneumothorax or large pleural effusion, negative for acute bone abnormality"*. Entre las 50 muestras seleccionadas, 44 reportes generados presentan frases repetitivas.

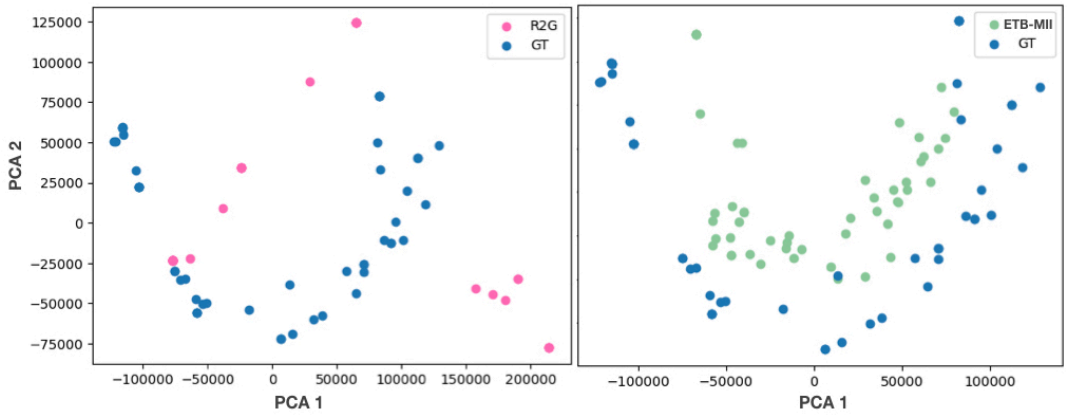
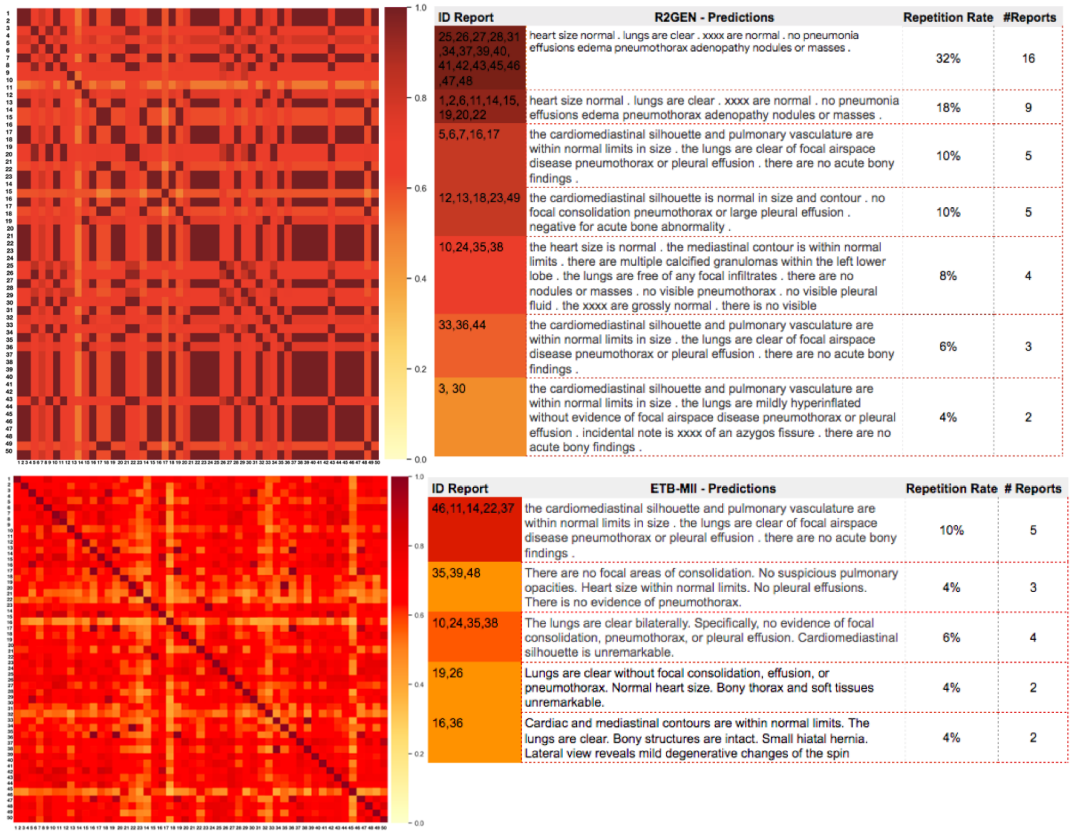


Figura 6.6: Visualización de los reportes generados por R2GEN vs GT, y ETB-MII vs GT en términos de similitud. Los puntos rosas representan los reportes generados de R2GEN, los verdes los de ETB-MII y los azules representan el GT.

Por otro lado, la figura 6.7 muestra el mapa de calor de los reportes generados por el modelo ETB-MII. Como se puede apreciar, la frase *"the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size, the lungs are clear of focal airspace disease pneumothorax or pleural effusion, there are no acute bony findings"* sólo se repite cinco veces en los reportes generados.

Aunque el modelo R2GEN obtiene buenos resultados según las métricas BLEU y ROUGE-L, tiene tendencia a generar textos repetitivos. Esto puede deberse al uso de una memoria auxiliar que almacena las frases más comunes de los textos del GT, y que influye en los mecanismos de atención para generar textos que incluyan las secuencias más frecuentes del conjunto de GT.

Analizando los resultados que se muestran en la tabla 6.3, los resultados más elevados que corresponden a las métricas de BLEU y ROUGE-L pueden no representar necesariamente una mayor calidad en los reportes generados. El caso anterior puede deberse al uso repetido de palabras específicas que contiene el GT. Por lo cuál, es más útil utilizar métricas como METEOR y CIDEr para evaluar el desempeño los



ID Report	R2GEN - Predictions	Repetition Rate	#Reports
25,26,27,28,31 ,34,37,39,40, 41,42,43,45,46 ,47,48	heart size normal . lungs are clear . xxxx are normal . no pneumonia effusions edema pneumothorax adenopathy nodules or masses .	32%	16
1,2,6,11,14,15, 19,20,22	heart size normal . lungs are clear . xxxx are normal . no pneumonia effusions edema pneumothorax adenopathy nodules or masses .	18%	9
5,6,7,16,17	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .	10%	5
12,13,18,23,49	the cardiomediastinal silhouette is normal in size and contour . no focal consolidation pneumothorax or large pleural effusion . negative for acute bone abnormality .	10%	5
10,24,35,38	the heart size is normal . the mediastinal contour is within normal limits . there are multiple calcified granulomas within the left lower lobe . the lungs are free of any focal infiltrates . there are no nodules or masses . no visible pneumothorax . no visible pleural fluid . the xxxx are grossly normal . there is no visible	8%	4
33,36,44	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .	6%	3
3, 30	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are mildly hyperinflated without evidence of focal airspace disease pneumothorax or pleural effusion . incidental note is xxxx of an azygos fissure . there are no acute bony findings .	4%	2

ID Report	ETB-MII - Predictions	Repetition Rate	# Reports
46,11,14,22,37	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .	10%	5
35,39,48	There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pneumothorax.	4%	3
10,24,35,38	The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. Cardiomediastinal silhouette is unremarkable.	6%	4
19,26	Lungs are clear without focal consolidation, effusion, or pneumothorax. Normal heart size. Bony thorax and soft tissues unremarkable.	4%	2
16,36	Cardiac and mediastinal contours are within normal limits. The lungs are clear. Bony structures are intact. Small hiatal hernia. Lateral view reveals mild degenerative changes of the spin	4%	2

Figura 6.7: Mapas de calor generados por el modelo R2GEN y por el ETB-MII, con un tamaño de muestra aleatorio de 50 reportes. El cuadro de texto de la derecha ofrece un análisis del porcentaje de secuencias de texto repetidas en el reporte generado.

modelos propuestos para la tarea de generación automática de reportes médicos.

Análisis de textos generados por ETB-MII con bajo desempeño en CIDEr

Para obtener una mejor comprensión de los resultados, se consideraron los textos generados por los modelos ETB-MII y R2GEN de la partición de pruebas del conjunto de datos IU-Xray. Posteriormente, se obtuvo la ponderación individual de cada texto con la métrica CIDEr. Debido a la amplia gamma de valores obtenidos en ambos modelos, se utilizó una representación logarítmica (figura 6.8) con el objetivo de visualizar de manera más efectiva los datos [166]. Como se puede observar en la figura 6.8, existe un amplio número de texto generado con ponderaciones en CIDEr

de 0.0; aunque R2GEN presenta un mayor número de textos generados ponderados con 0.0, ETB-MII también incluye una cantidad considerable. La hipótesis de este comportamiento en CIDEr es que aunque sea una métrica flexible al momento de evaluar los n-gramas diferentes al GT, si no encuentra una similitud de coseno positiva se otorga una ponderación baja. Esto se puede identificar en el ejemplo 4 incluido en la figura 6.10, donde el texto generado especifica la ausencia de *"focal consolidation"*, mientras que en el GT aparece el texto *"there is focal consolidation"*.

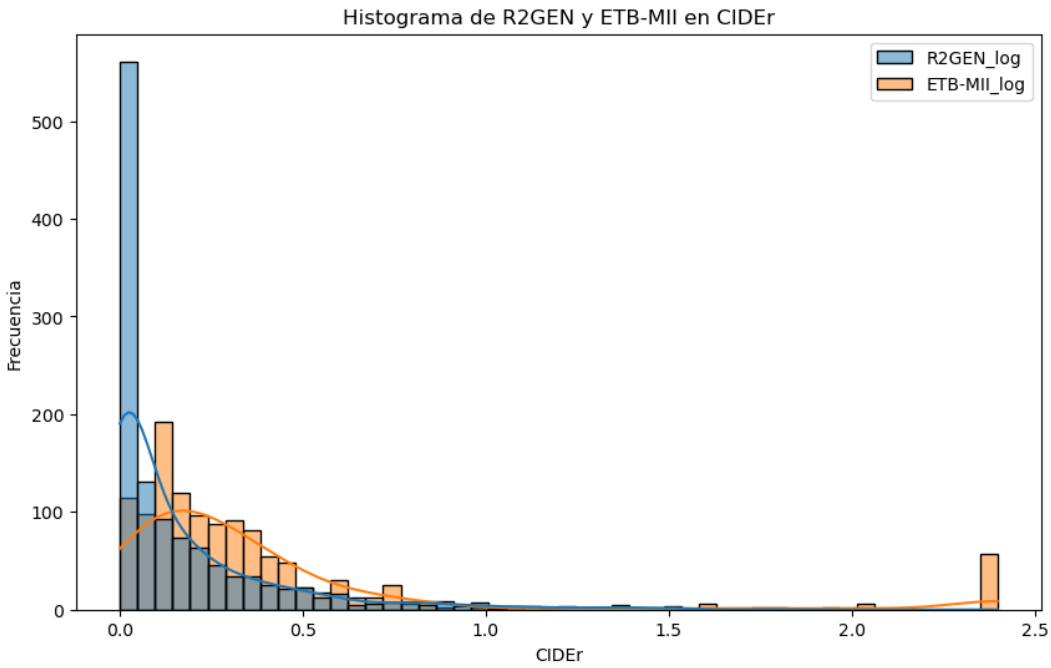


Figura 6.8: Histograma de los ETB-MII y R2GEN de la partición de pruebas de IU-Xray con respecto a la métrica de CIDEr.

Para realizar el análisis de los reportes generados por el modelo ETB-MII, se seleccionaron aleatoriamente 11 reportes con una ponderación menor a 0.4 en la métrica de CIDEr. Posteriormente, cada uno de los textos generado por el modelo con su correspondiente GT se representaron por medio de *embeddings*, los cuales fueron proyectados en un espacio bidimensional con la finalidad de visualizar si-

militudes. Para realizar esta representación se empleo el *framework: Sentencebert*, el cuál permite generar la transformación de frases a nivel de sentencia [164] así como el método PCA [165] para obtener una representación bidimensional. La figura 6.9 muestra dos gráficos, la representación de la izquierda incluye los reportes generados por el modelo ETB-MII con un valor menor a 0.4 en CIDEr en relación a su GT, mientras que la representación de la derecha muestra los reportes generados del modelo ETB-MII con un valor mayor o igual a 0.4 en CIDEr en relación a su GT.

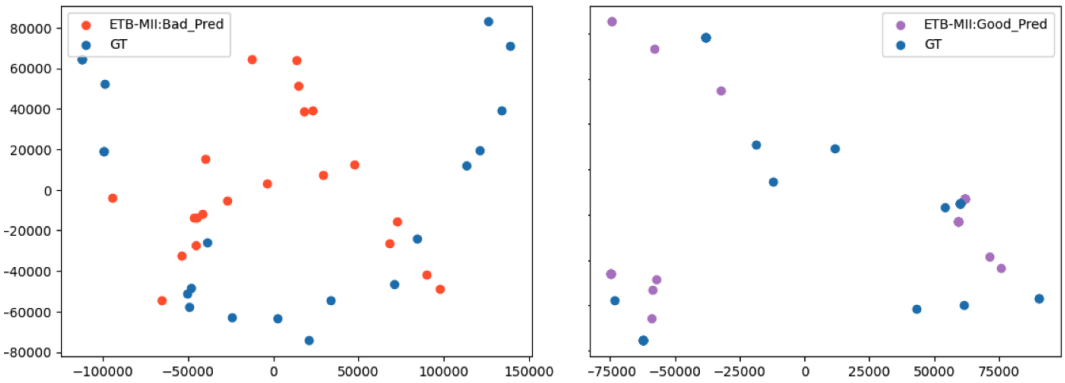


Figura 6.9: Visualización de los reportes generados por ETB-MII con un bajo valor en CIDEr (puntos rojos) vs GT (puntos azules) , y los reportes generados por ETB-MII con un alto valor en CIDEr (puntos morados) vs GT en términos de similitud.

Como se puede observar, los texto que obtienen una menor ponderación en CIDEr reflejan una correlación negativa con respecto al GT, mientras que los reportes generados que obtienen mayor valor en CIDEr reflejan una correlación positiva con respecto al GT.

La figura 6.10 incluye 4 ejemplos de los reportes generados por el modelo con sus respectivos GTs y la ponderación obtenida en CIDEr. Para mayor comprensión del lector los textos que incluyen las mismas palabras o términos son remarcados con el mismo color. En el primer ejemplo, el texto generado por ETB-MII incluye el texto completo del GT pero complementa la información describiendo la presencia de grapas quirúrgicas. A pesar de que en la imagen se identifica el hallazgo, la

descripción es imprecisa ya que solo menciona que están presentes en el cuadrante superior derecho. En el segundo ejemplo, la mayor parte del texto generada por el modelo es diferente al GT, incluso se identifica una calcificación incorrecta en la imagen. En el siguiente ejemplo, el modelo ETB-MII no identifica el hallazgo de *"soft tissues"* incluido en el GT, aunque en términos generales el diagnóstico es "normal"tanto en el GT como en el texto generado. Por último, el ejemplo 4 presenta un texto generado por el modelo con el hallazgo de *"no evidence of focal consolidation"*, mientras que el GT incluye el texto *"there is focal consolidation"*. Este diagnostico es totalmente contrario, ya que el modelo ETB-MII especifica la ausencia de este hallazgo mientras que el GT la presencia del mismo.

Aunque este análisis solo muestra unos cuantos ejemplos mal ponderados, podemos identificar que la métrica CIDEr realiza ponderaciones de acuerdo a la generación de la calidad del texto, es decir tiene congruencia al representar el desempeño del modelo. También se identifica que aunque el modelo ETB-MII sigue demostrando una gran capacidad por generar textos estructurados, también refleja diagnósticos incorrectos. Lo anterior puede darse en gran medida a que la cantidad de diagnósticos anormales o descripción de ciertos hallazgos estén en menor cantidad que los diagnósticos normales, como normalmente se presenta en las bases de datos médicas. Es de gran relevancia mencionar que el 73 % de los diagnósticos del conjunto de IU-Xray son normales es decir no presentan alguna condición o hallazgo anormal, mientras que el 27 % de los diagnósticos corresponden a diferentes hallazgos anormales en los cuales puede incluirse más de una anomalía en el mismo diagnóstico. De los hallazgos anormales más frecuentes en IU-Xray se encuentra el texto *"lung/hypoinflation"* con un 2.3 %, mientras que el texto *"Thoracic Vertebrae/degenerative/mild"* se presenta un 1.5 % y el texto *"Spine/degenerative"* un 1 %. Para el 4.8 % de los diagnósticos se identifica el texto *"No Indexing"*, lo que in-

dica que no se marco ningún hallazgo. Lo anterior demuestra que el modelo esta un poco sesgado hacia los casos normales, lo cuál da pauta a probar nuevas estrategias que permitan eficientar el desempeño del modelo.

	Ground-truth	Prediction (ETB-MII)	Cider
	cardiac and mediastinal contours are within normal limits the lungs are clear bony structures are intact	cardiac and mediastinal contours are within normal limits the lungs are clear bony structures are intact left axillary surgical clips are present in the right upper quadrant	0.3704
	bibasilar airspace opacities right greater than left the heart size and mediastinal silhouette are within normal limits for contour no pneumothorax or pleural effusions the XXXX are intact	the heart is top normal in size the mediastinum is unremarkable the lungs are hypoinflated with elevated right hemidiaphragm XXXX XXXX opacities in the left lung base	0.1043
	the trachea is midline the cardiomediastinal silhouette is normal lung XXXX are clear without evidence of effusion infiltrate or pneumothorax visualized bony structures are intact visualized soft tissues appear normal	the heart and lungs have XXXX XXXX in the interval both lungs are clear and expanded no change in the small calcified granuloma in the right upper lobe heart and mediastinum normal	0.1071
	the heart size is within normal limits trachea is midline no pleural effusions or pneumothorax cardiomediastinal contours are normal there is focal consolidation in the posterior segment of the right lower lobe no bony or soft tissue abnormalities	the lungs are clear bilaterally specifically no evidence of focal consolidation pneumothorax or pleural effusion cardiomediastinal silhouette is unremarkable visualized osseous structures of the thorax are without acute abnormality	0.0708

Figura 6.10: Ejemplos de los reportes generados por el modelo ETB-MII con un bajo nivel de desempeño en CIDEr. Se identifican con el mismo color los textos generados por el modelo y el GT que son iguales. Además se agrega la ponderación obtenida en CIDEr de cada texto generado.

6.4.1. Eficiencia del modelo

Para analizar la eficiencia de los modelos ETB-MII y R2GEN se calculó el número de parámetros, las operaciones en punto flotante por segundo en gigas (FLOPs),

las operaciones de multiplicación y suma por segundo en gigas (MACs) y el tiempo de inferencia. El número de parámetros del modelo es utilizado para tener una referencia del tamaño del modelo, donde un valor más alto representa un mayor tamaño. De la misma manera, el número de FLOPs es una métrica utilizada como referencia para calcular la complejidad computacional del modelo, donde un valor más alto refleja una complejidad computacional mayor [167, 168].

Para obtener las métricas anteriormente mencionadas para ambos modelos R2GEN y ETB-MII se implementó la librería de *flops-profile*[5]. Mediante *Flops-profile* se realiza un diagnóstico del modelo el cuál calcula el número de parámetros, FLOPS, MACs así como el tiempo de inferencia. La tabla 6.4, las figuras 6.11 y 6.12 muestran los resultados obtenidos de cada modelo con el número de parámetros, FLOPs, MACs y el tiempo de inferencia. El modelo ETB-MII obtiene un menor número de parámetros y de FLOPs, lo que hace referencia a una menor complejidad computacional comparado con el modelo R2GEN. Aunado a lo anterior, el modelo ETB-MII obtiene un menor tiempo de inferencia, el cual refleja la sobrecarga de tiempo del modelo en la etapa de inferencia y, por lo general, un menor tiempo de inferencia puede mejorar el rendimiento en tiempo real del modelo [168].

De acuerdo a los resultados obtenidos, el modelo ETB-MII muestra una mayor eficiencia computacional, lo que resulta en un modelo más práctico y eficaz para generar texto médico estructurado de calidad.

6.4.2. Discusión

Desde la implementación de modelos de DL para generar reportes médicos, se han utilizado las métricas de Generación de Lenguaje Natural (NLG, por sus siglas

[5]https://github.com/cli99/flops-profiler/blob/main/LICENSE

Tabla 6.4: Número de parámetros, FLOPs, MACs y tiempo de inferencia de los modelos ETB-MII y R2GEN.

Model	Input	Params	FLOPs	MACs	Inference Time (ms)
R2GEN	224×224	83.54G	417.23G	208.55G	4320 ms
ETB-MII	224×224	**211.01M**	**110.74G**	**55.35G**	**50.99ms**

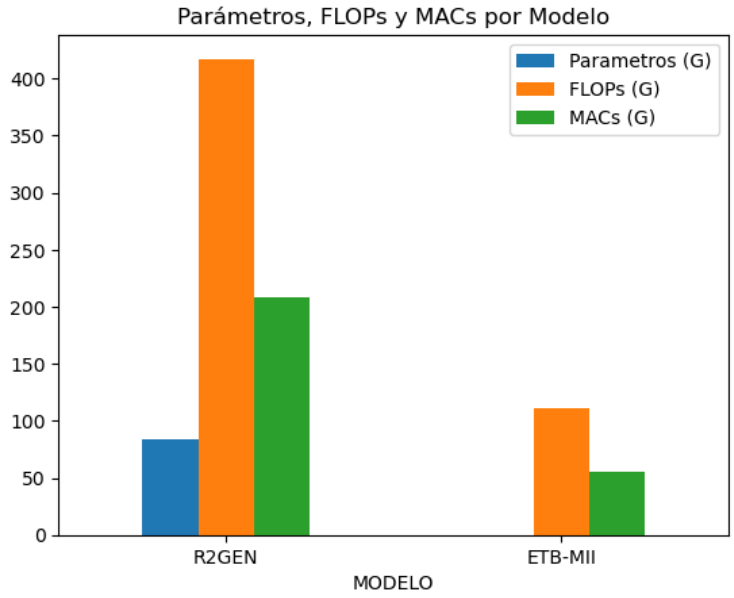


Figura 6.11: Número de parámetros, FLOPs y MACs de los modelos ETB-MII y R2GEN

en inglés) para evaluar objetivamente el rendimiento de los modelos. Sin embargo, los modelos que muestran una mayor diversidad en la generación de texto son penalizados por métricas como BLEU y ROUGE-L. Estas penalizaciones son considerables y se imponen a los reportes generados por el modelo que no incorporan los mismos n-gramas así como su orden. Lo anterior, desalienta el potencial de proponer nuevos modelos que tengan la capacidad de formular descripciones más creativas de los hallazgos identificados en las imágenes médicas. En consecuencia, los expertos prefieren proponer arquitecturas complejas que incluyan memorias auxiliares, modelos basados en grafos creados con vocabularios predeterminados y limitados, o incluso adaptar plantillas predefinidas para modificar los textos gene-

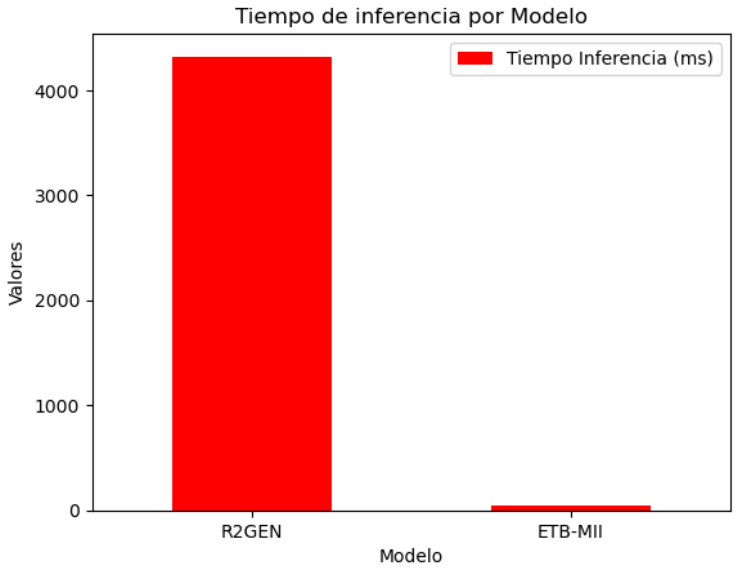


Figura 6.12: Tiempo de inferencia de los modelos ETB-MII y R2GEN en milisegundos (ms).

rados por los modelos propuesto. Todo esto con la finalidad de obtener mejores ponderaciones en las métricas de evaluación.

Capítulo 7

Conclusiones

En este trabajo se presenta una arquitectura DL novedosa, eficiente y precisa basada en una arquitectura *encoder-decoder* para la generación automática de reportes médicos a partir de imágenes de rayos X. Esta arquitectura se denomina como *Enhanced Transformer Based - Medical Image Interpretation* (ETB-MII). Mediante la aplicación de una estrategia de aumento de datos (DAS) basada en la rotación de imágenes y el aumento contextual, el modelo propuesto mejora los resultados del estado del arte hasta un 226 % en la métrica de CIDEr y en un 103 % en la métrica de METEOR utilizando el conjunto de datos de IU X-ray; mientras que para las métricas de BLEU-3 se muestra una diferencia marginal del 6.72 %, y para Rouge-L un 12.5 %. Para el conjunto de datos de MIMIC-CXR el desempeño del modelo ETB-MII mejora su resultado en un 141 % en la métrica de CIDEr. Es importante considerar que las métricas de BLEU-1 a BLEU-4 sólo se basan en medir la *precision* y ROUGE-L mide la *precision* así como *recall*. Sin embargo, para obtener la máxima puntuación en estas métricas (BLEU y ROUGE-L), el texto generado debe coincidir exactamente con el GT y el orden de palabras. Por el contrario, las métricas CIDEr y METEOR tienen mayor flexibilidad en su evaluación, considerando aspectos como similitud semántica, raíz de palabras, sinónimos; por ello es necesario considerar

que son más adecuadas para la tarea de generación de reportes médicos mediante imágenes de rayos X.

Mediante el análisis de errores realizado durante esta investigación se identificaron dos aspectos relevantes. El primero es que la métrica CIDEr realiza una evaluación congruente en el desempeño del modelo, ya que logra penalizar los textos que no igualan el significado del GT pero permite que el modelo se exprese en *n-gramas* diferentes al especificado en el GT. El Otro aspecto, es el sesgo que se induce al modelo por el conjunto de datos cargado con textos normales y en una menor cantidad con textos anormales. Esto nos muestra la necesidad por orientarnos más a buscar generar colecciones de datos médicos mas enriquecidos.

Durante el curso de esta investigación, se ha observado otro aspecto relevante que merece atención: el aumento en la complejidad de los modelos propuestos. Cada vez es más frecuente encontrar arquitecturas que incorporan una variedad de componentes que van desde memorias auxiliares basadas en recurrencias, modelos de grafos, hasta plantillas predefinidas utilizadas para adaptar los textos generados por los modelos. En el análisis de eficiencia realizado, el modelo ETB-MII obtiene un menor número de parámetros y FLOPs con respecto al modelo R2GEN. Además, los tiempos de inferencia de ambos modelos varían notoriamente siendo ETB-MII el que obtiene un tiempo de inferencia menor. Lo anterior nos lleva a cuestionar si lo que se pretende es proponer modelos capaces de describir patrones visuales de forma automática y precisa o si solamente se están desarrollando nuevos modelos con el objetivo de satisfacer las métricas de evaluación de NLG dando más prioridad a BLEU y ROUGE-L.

7.1. Trabajo futuro

El área de diagnóstico médico mediante imágenes de rayos X ha cobrado mayor relevancia en los últimos años. Durante esta investigación se han analizado diferentes modelos enfocados a la tarea de generación de reportes médicos mediante imágenes de rayos X, de los cuales se destaca el uso de estructuras secuenciales que incluyen modelos visuales y modelos de lenguaje.

Considerando las propuestas más recientes en el estado del arte para esta tarea, cada vez es más frecuente encontrar arquitecturas más complejas que demandan un alto nivel computacional para su entrenamiento e implementación. Además, el utilizar las métricas orientadas a la generación de lenguaje como BLEU, ROUGE-L y METEOR no garantiza que los modelos estén aprendiendo a identificar características relevantes en una imagen.

Partiendo de lo anterior, es necesario generar nuevas métricas orientadas a la generación de texto médico que evalúen correctamente las predicciones de los modelos considerando diversos aspectos como la calidad del texto generado, la similitud semántica empleando diferentes representaciones textuales y la coherencia del texto.

Otra área de oportunidad a considerar para trabajos futuros, es la posibilidad de adaptar los modelos propuestos a otro tipo de diagnósticos médicos que involucren diversas imágenes médicas como las resonancias magnéticas, los ultrasonidos, las tomografías computarizadas, etc. Esto con la finalidad de robustecer la calidad del diagnóstico médico.

Además de lo anterior, es necesario explorar diversas metodologías que permi-

tan desarrollar modelos adaptativos que puedan ser adecuados a diferentes conjuntos de datos y condiciones clínicas, facilitando la implementación en entornos médicos diversos.

Las metodologías para el aumento de datos es otra de las áreas importantes que faltan de explorar así como el impacto de las mismas en los entrenamientos de los modelos centrados en mecanismos de atención.

Las colecciones de datos médicos es otro reto importante, ya que el no contar con los suficientes datos de hallazgos anormales se limita el potencial de aprendizaje en los modelos. Por lo anterior, es necesario buscar como enriquecer más esta área.

Como podemos inferir, existe una gran diversidad de alternativas que faltan de explorar para esta tarea en específico. Esto permite abrir el horizonte a nuevas posibilidades que contribuyan eficazmente a crear sistemas que actúen como asistentes para los profesionales de la salud, proporcionando recomendaciones y análisis adicionales basados en las imágenes de rayos X.

Referencias

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 10 2020.

[4] D. Demner-Fushman, M. Kohli, M. Rosenman, S. Shooshan, L. Rodriguez, S. Antani, G. Thoma, and C. Mcdonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association : JAMIA*, vol. 23, 07 2015.

[5] J. H. Moon, H. Lee, W. Shin, Y. H. Kim, and E. Choi, "Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, 2022.

[6] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating Radiology Reports via Memory-driven Transformer," *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1439–1449, 10 2020.

[7] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *ARTIF INTELL MED*, vol. 106, p. 101878, 2020.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep

learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 12 2017.

[9] B. Jing, P. Xie, and E. Xing, "On the Automatic Generation of Medical Imaging Reports," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 2577–2586, 11 2017.

[10] R. M. Thanki and A. Kothari, "Data Compression and Its Application in Medical Imaging," in *Hybrid and Advanced Compression Techniques for Medical Images*, pp. 1–15, Springer International Publishing, 2019.

[11] Y. Liao, H. Liu, and I. Spasić, "Deep learning approaches to automatic radiology report generation: A systematic review," *Informatics in Medicine Unlocked*, vol. 39, p. 101273, 2023.

[12] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning," *BioMedical Engineering Online*, vol. 22, p. 48, 12 2023.

[13] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 1 2021.

[14] S. Mayor, "Waiting times for x ray results in england are increasing, figures show," in *BMJ (Clinical research ed.*, vol. 350, p. h1598, 3 2015.

[15] G. Maskell, "Error in radiology – where are we now?," *The British Journal of Radiology*, vol. 92, p. 20180845, 11 2018.

[16] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in Medical Image Analysis: A Review," *Intelligent Medicine*, 8 2022.

[17] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Medical Image Analysis*, vol. 85, p. 102762, 2023.

[18] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. andía, C. Tejos, C. Prieto, and D. Capurro, "A survey on deep learning and explainability for automatic report generation from medical images," 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, (Red Hook, NY, USA), Curran Associates, Inc., 2017.

[20] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12903 LNCS, pp. 72–82, Springer Science and Business Media Deutschland GmbH, 2021.

[21] G. van Tulder, Y. Tong, and E. Marchiori, "Multi-view analysis of unregistered medical images using cross-view transformers," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 104–113, Strasbourg, France: Springer International Publishing, 2021.

[22] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic Brain Tumor Segmentation using Cascaded Anisotropic Convolutional Neural Networks," 9 2017.

[23] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No New-Net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, eds.), (Cham), pp. 234–244, Springer International Publishing, 2019.

[24] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, pp. 590–597, 1 2019.

[25] H. C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, "Interleaved text/image Deep Mining on a large-scale radiology database," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1090–1099, IEEE Computer Society, 10 2015.

[26] W. Xiaosong, L. Le, S. Hoo-chang, K. Lauren, N. Isabella, Y. Jianhua, and S. Ro-

nald, "Unsupervised category discovery via looped deep pseudo-task optimization using a large scale radiology image database," *arXiv*, 2016.

[27] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, "Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks," *arXiv*, 4 2018.

[28] X. Huang, F. Yan, W. Xu, and M. Li, "Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation," *IEEE Access*, vol. 7, pp. 154808–154817, 2019.

[29] X. Zeng, L. Wen, Y. Xu, and C. Ji, "Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105700, 12 2020.

[30] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference NAACL-HLT* (J. Burstein, C. Doran, and T. Solorio, eds.), pp. 4171–4186, Association for Computational Linguistics, 2019.

[31] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[33] Q. Xie, E. J. Schenck, H. S. Yang, Y. Chen, Y. Peng, and F. Wang, "Faithful ai in medicine: A systematic review with large language models and beyond," 2023.

[34] J. Deng and Y. Lin, "The benefits and challenges of chatgpt: An overview," *Frontiers in Computing and Intelligent Systems*, vol. 2, no. 2, pp. 81–83, 2022.

[35] Z. Liu, A. Zhong, Y. Li, L. Yang, C. Ju, Z. Wu, C. Ma, P. Shu, C. Chen, S. Kim, *et al.*, "Tailoring large language models to radiology: A preliminary approach to llm adaptation for a highly specialized domain," in *International Workshop on Machine Learning in Medical Imaging*, pp. 464–473, Springer, 2023.

[36] N. H. Shah, D. Entwistle, and M. A. Pfeffer, "Creation and adoption of large language models in medicine," *Jama*, vol. 330, no. 9, pp. 866–869, 2023.

[37] E. Villa-Cueva, J. M. Valles-Silva, A. P. López-Monroy, F. Sanchez-Vega, and R. Lopez-Santillan, "Few shot profiling of cryptocurrency influencers using natural language inference & large language models," 2023.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," *ACL*, pp. 311–318, 2001.

[39] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

[40] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Annual Meeting of the Association for Computational Linguistics*, 2004.

[41] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 4566–4575, 11 2014.

[42] A. H. Shahid and M. P. Singh, "Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments," 7 2019.

[43] O. de la Salud (OMS), "Neumonía," Dic 2019. último acceso: 2020-12-01. [online], Disponible:https://www.who.int/es/news-room/fact-sheets/detail/pneumonia.

[44] O. de la Salud (OMS), "Coronavirus causante del síndrome respiratorio de Oriente Medio (MERS-CoV)," Dic 2019. último acceso: 2020-12-01. [online], Disponible: https://www.who.int/es/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov).

[45] O. de la Salud (OMS), "Cáncer," Dic 2019. último acceso: 2020-12-01. [online], Disponible: https://www.who.int/es/news-room/fact-sheets/detail/cancer.

[46] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S. B. Ko, "Breast Cancer Classification in Automated Breast Ultrasound Using Multiview Convolutional Neural Network with Transfer Learning," *Ultrasound in Medicine and Biology*, vol. 46, pp. 1119–1132, 5 2020.

[47] S. Hassantabar, M. Ahmadi, and A. Sharifi, "Diagnosis and detection of infected tissue of COVID-19 patients based on lung x-ray image using convolutional neural network approaches," *Chaos, Solitons and Fractals*, vol. 140, p. 110170, 11 2020.

[48] S. Zahia, B. Garcia-Zapirain, I. Saralegui, and B. Fernandez-Ruanova, "Dyslexia detection using 3D convolutional neural networks and functional magnetic resonance imaging," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105726, 12 2020.

[49] W. T. Le, F. Maleki, F. P. Romero, R. Forghani, and S. Kadoury, *Overview of Machine Learning: Part 2: Deep Learning for Medical Image Analysis*, vol. 30, pp. 417–431. W.B. Saunders, 11 2020.

[50] G. Mohan and M. M. Subashini, "MRI based medical image analysis: Survey on brain tumor grade classification," 2018.

[51] M. Usman, T. Zia, and A. Tariq, "Analyzing Transfer Learning of Vision Transformers for Interpreting Chest Radiography," *Journal of Digital Imaging*, 2022.

[52] M. Ghaffari, A. Sowmya, and R. Oliver, "Automated Brain Tumor Segmentation Using Multimodal Brain Scans: A Survey Based on Models Submitted to the BraTS 2012-2018 Challenges," *IEEE Reviews in Biomedical Engineering*, vol. 13, pp. 156–168, 2020.

[53] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 9049–9058, IEEE Computer Society, 12 2018.

[54] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. andía, C. Tejos, C. Prieto, and D. Capurro, "A survey on deep learning and explainability for automatic image-based medical report generation," 2020.

[55] R.-S. G. Ramirez-Alonso, O. Prieto-Ordaz and M. Montes-Y-Gómez, "Medical Report Generation through Radiology Images: An Overview.," *IEEE Latin America Transactions*, vol. 20, p. 986–999, 6 2022.

[56] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. Chapman, T. Amrhein, D. Mong, D. L. Rubin, O. Farri, and M. P. Lungren, "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," *Artificial Intelligence in Medicine*, vol. 97, pp. 79–88, 6 2019.

[57] R. Jain, P. Nagrath, G. Kataria, V. Sirish Kaushik, and D. Jude Hemanth, "Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning," *Measurement: Journal of the International Measurement Confederation*, vol. 165, p. 108046, 12 2020.

[58] Y. Dong, Y. Pan, J. Zhang, and W. Xu, "Learning to Read Chest X-Ray Images from 16000+ Examples Using CNN," in *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017*, pp. 51–57, Institute of Electrical and Electronics Engineers Inc., 8 2017.

[59] M. Gu, X. Huang, and Y. Fang, "Automatic generation of pulmonary radiology reports with semantic tags," *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*, pp. 162–167, 2019.

[60] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101797, Dec 2020.

[61] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv*, 2019.

[62] "MIMIC-IV (Versión 0.4," 2020.

[63] A. Karargyris, S. Kashyap, I. Lourentzou, J. T. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E. A. Krupinski, and M. Moradi, "Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development," *Scientific Data*, vol. 8, pp. 1–18, 12 2021.

[64] B. Ionescu, H. Müller, M. Villegas, H. Arenas, G. Boato, D. T. Dang Nguyen, Y. Dicente Cid, C. Eickhoff, A. García Seco de Herrera, C. Gurrin, M. B. Islam, V. Kovalev, V. Liauchuk, J. Mothe, L. Piras, M. Riegler, and I. Schwall, "Overview of imageclef 2017: Information extraction from images," in *Lecture Notes in Computer Science*, vol. 10456, pp. 11–14, 08 2017.

[65] B. Ionescu, H. Müller, M. Villegas, A. G. S. de Herrera, C. Eickhoff, V. Andrearczyk, Y. D. Cid, V. Liauchuk, V. Kovalev, S. A. Hasan, Y. Ling, O. Farri, J. Liu, M. Lungren, D.-T. Dang-Nguyen, L. Piras, M. Riegler, L. Zhou, M. Lux, and C. Gurrin, "Overview of ImageCLEF 2018: Challenges, datasets and evaluation," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), (Avignon, France), LNCS Lecture Notes in Computer Science, Springer, September 10-14 2018.

[66] S. A. Hasan, Y. Ling, J. Liu, R. Sreenivasan, S. Anand, T. R. Arora, V. Datla, K. Lee, A. Qadir, C. Swisher, and O. Farri, "Attention-based medical caption generation with image modality classification and clinical concept mapping," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11018 LNCS, pp. 224–230, Springer Verlag, 9 2018.

[67] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation," *Advances in Neural Information Processing Systems*, vol. 2018-December, pp. 1530–1540, 5 2018.

[68] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[69] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly

available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, pp. 1–8, 12 2019.

[70] "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports - PubMed."

[71] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, "Clinically accurate chest x-ray report generation," in *Proceedings of the 4th Machine Learning for Healthcare Conference* (F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, eds.), vol. 106 of *Proceedings of Machine Learning Research*, pp. 249–269, PMLR, 09–10 Aug 2019.

[72] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 5 2018.

[73] M. M. A. Monshi, J. Poon, and Y. Y. Chung, "Convolutional neural network to detect thorax diseases from multi-view chest x-rays," in *Neural Information Processing,26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part IV*, 2019.

[74] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," *arXiv*, 2019.

[75] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest x-ray reports," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2577–2586, 2019.

[76] J. Yuan, H. Liao, R. Luo, and J. Luo, *Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment*, vol. 11769 LNCS. Springer Science and Business Media Deutschland GmbH, 10 2019.

[77] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv*, 2017.

[78] H. C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation," in *Proceedings of the IEEE Computer Society Confe-*

rence on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 2497–2506, IEEE Computer Society, 12 2016.

[79] X. Xie, Y. Xiong, P. S. Yu, K. Li, S. Zhang, and Y. Zhu, "Attention-Based Abnormal-Aware Fusion Network for Radiology Report Generation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11448 LNCS, pp. 448–452, Springer Verlag, 4 2019.

[80] Y. Xue, T. Xu, L. Rodney Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, "Multimodal recurrent model with attention for automated radiology report generation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11070 LNCS, pp. 457–466, Springer Verlag, 2018.

[81] P. Harzig, Y.-Y. Chen, F. Chen, and R. Lienhart, "Addressing data bias problems for chest x-ray image report generation," *arXiv*, 2019.

[82] A. Gasimova, "Automated enriched medical concept generation for chest x-ray images," *Lecture Notes in Computer Science*, p. 83–92, 2019.

[83] S. Biswal, C. Xiao, L. M. Glass, B. Westover, and J. Sun, *CLARA: Clinical Report Auto-Completion*. WWW '20, New York, NY, USA: Association for Computing Machinery, 2020.

[84] Y. Xiong, B. Du, and P. Yan, "Reinforced Transformer for Medical Image Captioning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11861 LNCS, pp. 673–680, Springer, 10 2019.

[85] J. Tian, C. Zhong, Z. Shi, and F. Xu, "Towards automatic diagnosis from multimodal medical data," in *Lecture Notes in Computer Science*, vol. 11797 LNCS, pp. 67–74, Springer, 2019.

[86] X. Li, R. Cao, and D. Zhu, "Vispi: Automatic visual perception and interpretation of chest x-rays," *arXiv*, 2020.

[87] G. O. Gajbhiye, A. V. Nandedkar, and I. Faye, *Automatic report generation for chest X-Ray images: A multilevel multi-attention approach*, vol. 1147 CCIS, pp. 174–182. Springer, 9 2020.

[88] S. Singh, S. Karimi, K. Ho-Shon, and L. Hamey, "From Chest X-Rays to Radiology Reports: A Multimodal Machine Learning Approach," in *2019 Digital Image Computing: Techniques and Applications, DICTA 2019*, Institute of Electrical and Electronics Engineers Inc., 12 2019.

[89] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 728–737, 2019.

[90] G. Spinks and M.-F. Moens, "Justifying diagnosis decisions by deep neural networks," *Journal of Biomedical Informatics*, vol. 96, p. 103248, 07 2019.

[91] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12910–12917, Apr. 2020.

[92] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, and Z. Lu, "Negbio: a high-performance tool for negation and uncertainty detection in radiology reports," *arXiv*, 2017.

[93] X. He and L. Deng, "Deep learning in natural language generation from images," in *Deep Learning in Natural Language Processing*, pp. 289–307, Springer International Publishing, 1 2018.

[94] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (USA), p. 311–318, Association for Computational Linguistics, 2002.

[95] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[96] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, (Prague, Czech Republic), pp. 228–231, Association for Computational Linguistics, June 2007.

[97] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *arXiv*, 2015.

[98] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[99] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (A. Moschitti, B. Pang, and W. Daelemans, eds.), pp. 1724–1734, ACL, 2014.

[100] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 3549–3557, 7 2017.

[101] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," in *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis.*, p. 204–213, 2018.

[102] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

[103] Y. Xue and X. Huang, "Improved disease classification in chest x-rays with transferred features from report generation," in *Information Processing in Medical Imaging, IPMI 2019, Proceedings* (A. Chung, S. Bao, J. Gee, and P. Yushkevich, eds.), Lecture Notes in Computer Science, (Germany), pp. 125–138, Springer Verlag, 2019.

[104] J. Zhao, Y. Kim, K. Zhang, A. Rush, and Y. LeCun, "Adversarially regularized autoencoders," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 5902–5911, PMLR, 10–15 Jul 2018.

[105] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5908–5916, IEEE Computer Society, oct 2017.

[106] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, p. 139–144, Oct. 2020.

[107] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, p. 336–359, Oct 2019.

[108] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics in Medicine Unlocked*, vol. 24, p. 100557, 2021.

[109] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv*, 2020.

[110] Z. Li and J. Ren, "Fine-tuning ERNIE for chest abnormal imaging signs extraction," *Journal of Biomedical Informatics*, vol. 108, p. 103492, 8 2020.

[111] F. Nooralahzadeh, N. Perez Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive transformer-based generation of radiology reports," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, (Punta Cana, Dominican Republic), pp. 2824–2832, Association for Computational Linguistics, Nov. 2021.

[112] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial Intelligence Review*, vol. 56, pp. 12561–12605, 11 2023.

[113] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023.

[114] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[115] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "Gans for medical image analysis," *Artificial Intelligence in Medicine*, vol. 109, p. 101938, 2020.

[116] S. N. Kasthurirathne, G. Dexter, and S. J. Grannis, "Generative Adversarial Networks for Creating Synthetic Free-Text Medical Data: A Proposal for Co-

llaborative Research and Re-use of Machine Learning Models," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2021, pp. 335–344, 2021.

[117] P. Celard, E. L. Iglesias, J. M. Sorribes-Fdez, R. Romero, A. S. Vieira, and L. Borrajo, "A survey on deep learning applied to medical images: from simple artificial neural networks to generative models," *Neural Computing and Applications*, vol. 35, no. 3, pp. 2291–2323, 2023.

[118] "The use of generative adversarial networks in medical image augmentation," *Neural Computing and Applications*, vol. 35, pp. 24055–24068, 12 2023.

[119] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Applied Soft Computing*, vol. 132, p. 109803, 2023.

[120] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, L. Sun, Q. Li, D. Shen, T. Liu, and X. Li, "Auggpt: Leveraging chatgpt for text data augmentation," 2023.

[121] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-net based fully convolutional networks," in *Communications in Computer and Information Science*, vol. 723, pp. 506–517, Springer Verlag, 2017.

[122] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," 8 2019.

[123] Y. C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: UNiversal Image-TExt Representation Learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12375 LNCS, pp. 104–120, Springer Science and Business Media Deutschland GmbH, 9 2020.

[124] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," tech. rep., 2019.

[125] B. Yan and M. Pei, "Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 2982–2990, 6 2022.

[126] C. I. Orozco, E. Xamena, C. A. Martínez, and D. A. Rodríguez, "Covid-xr: A web management platform for coronavirus detection on x-ray chest images," *IEEE LAT AM T*, vol. 19, p. 1033–1040, jun. 2021.

[127] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki, and D. Mitsouras, "Natural language processing technologies in radiology research and clinical applications," *Radiographics*, vol. 36, pp. 176–191, 1 2016.

[128] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing," *Proceedings of Languages in Biology and Medicine*, 01 2013.

[129] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," *arXiv*, 3 2018.

[130] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of Physiology*, vol. 195, pp. 215–243, 3 1968.

[131] K. Fukushima and S. Miyake, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition," in *Biological Cybernetics*, pp. 267–285, Springer, Berlin, Heidelberg, 1982.

[132] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[133] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 5 2017.

[134] C. Szegedy, W. Wei Liu, Y. Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, IEEE, 6 2015.

[135] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing Journal*, vol. 70, pp. 41–65, 2018.

[136] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks," *International Journal of Computer Vision*, vol. 116, pp. 1–20, 1 2016.

[137] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 6 2016.

[138] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Springer International Publishing, 2019.

[139] S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian, and X. Zhang, "A real-time object detection algorithm for video," *Computers & Electrical Engineering*, vol. 77, pp. 398–408, 2019.

[140] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, pp. 1–27, 5 2020.

[141] B. Chiu and S. Baker, "Word embeddings for biomedical natural language processing: A survey," *Language and Linguistics Compass*, vol. 14, 12 2020.

[142] S. Wang, W. Zhou, and C. Jiang, "A survey of word embeddings based on deep learning," *Computing*, vol. 102, pp. 717–740, 3 2020.

[143] "The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time.."

[144] C. Fellbaum, "Wordnet publications — wordnet," 2005.

[145] B. J. Erickson and F. Kitamura, *Magician's corner: 9. performance metrics for machine learning models*, vol. 3. Radiological Society of North America Inc., 5 2021.

[146] M. von Davier, "Training optimus prime, m.d.: Generating medical certification items by fine-tuning openai's gpt2 transformer model," 2019.

[147] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," tech. rep.

[148] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," 2020.

[149] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13041–13049, Apr. 2020.

[150] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016.

[151] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," 2015.

[152] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[153] B. Jing, Z. Wang, and E. Xing, "Show, describe and conclude: On exploiting the structure information of chest X-ray reports," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 6570–6580, Association for Computational Linguistics, July 2019.

[154] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Medical Image Analysis*, vol. 80, p. 102510, 2022.

[155] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3242–3250, 2017.

[156] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6666–6673, Jul. 2019.

[157] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13748–13757, 2021.

[158] M. Li, R. Liu, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," *World Wide Web 2022 26:1*, vol. 26, pp. 253–270, 8 2022.

[159] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018*, IEEE Conference on Computer Vision and Pattern Recognition, (United States), pp. 6077–6086, Institute of Electrical and Electronics Engineers (IEEE), dec 2018.

[160] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive Attention for Automatic Chest X-ray Report Generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (Stroudsburg, PA, USA), pp. 269–280, Association for Computational Linguistics, 2021.

[161] F. Liu, S. Ge, and X. Wu, "Competence-based Multimodal Curriculum Learning for Medical Report Generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Stroudsburg, PA, USA), pp. 3001–3012, Association for Computational Linguistics, 2021.

[162] F. Nooralahzadeh, N. A. Perez Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive transformer-based generation of radiology reports," 01 2021.

[163] J. R. Mccormick, A. J. Sama, N. C. Schiller, A. J. Butler, and C. J. Donnally, "Cervical spondylotic myelopathy: A guide to diagnosis and management,"

[164] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[165] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *The Journal of Finance and Data Science*, vol. 2, no. 4, pp. 265–278, 2016.

[166] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. .ºReilly Media, Inc.", 2018.

[167] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12021–12031, 2023.

[168] W. Lu, J. Jiang, Y. Shi, X. Zhong, J. Gu, L. Huangfu, and M. Gong, "Application of entity-bert model based on neuroscience and brain-like cognition in electronic medical record entity recognition," *Frontiers in Neuroscience*, vol. 17, 2023.

[169] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[170] "How many images do you need to train a neural network? « Pete Warden's blog."

[171] B. Mandal, A. Okeukwu, and Y. Theis, "Masked Face Recognition using ResNet-50," 4 2021.

[172] M. Elgendi, M. U. Nasir, Q. Tang, D. Smith, J.-P. Grenier, C. Batte, B. Spieler, W. D. Leslie, C. Menon, R. R. Fletcher, N. Howard, R. Ward, W. Parker, and S. Nicolaou, "The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective," *Frontiers in Medicine*, vol. 8, p. 153, 3 2021.

[173] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, 11 2018.

Anexos

A: Clasificación imágenes médicas

Implementación de modelos para Clasificación imágenes

Con la finalidad de identificar cuál de las arquitecturas CNN o ViT puede contribuir con mayor eficiencia a la extracción de características de la imagen se realizaron dos implementaciones de modelos visuales. La primer implementación consistió en reproducir la red CNN propuesta por [77] (*Chexnet*), para identificar las 14 patologías de la base de datos de *Chest-xray14*; además identificar si un pre-procesamiento en las imágenes contribuye al aprendizaje del modelo. La segunda implementación consistió en reproducir el modelo ViT propuesto por [1], con la finalidad de evaluar su desempeño en la tarea de clasificar imágenes médicas así como la aportación del pre-procesamiento de los datos para el aprendizaje del modelo.

Base de Datos

Para la implementación del modelo al igual [77] se utilizó la base de datos denominada *ChestX-ray14* [68] del *National Institute of Health Clinical Center* (NIH). Además de ser una base de datos pública, incluye 112,120 imágenes clasificadas con la ausencia o presencia de 14 patologías en idioma inglés. Con la finalidad de llevar una mayor claridad y transparencia en los datos del experimento, las tablas de resultados e información de datos hacen referencia a las patologías en idioma inglés, que es el incluido originalmente en la base de datos. La tabla 2.1 incluye una breve descripción de la base de datos, así como algunas características relevantes.

Al igual que el entrenamiento llevado por [77], se utilizaron las particiones oficiales de los datos; el 70 % de las imágenes se destinó al entrenamiento del modelo, el 10 % para validación y el 20 % para prueba. Además, se validó que no existieran imágenes del mismo paciente en más de 1 partición (entrenamiento, validación y

pruebas).

Todas las imágenes de rayos X fueron normalizadas de acuerdo a la media y desviación estándar de las imágenes de la base de datos *Imagenet*. Para el entrenamiento del modelo, se modificó el tamaño de las imágenes a 224×224.

Modelo CNN y configuración de entrenamiento.

Para la clasificación de 14 patologías en imágenes médicas de rayos X se utilizó un modelo base pre-entrenado en *Imagenet*. La arquitectura del modelo corresponde a una red *ChexNet*[77] que incorpora una estructura del modelo DenseNet [169] y sustituye la capa final totalmente conectada (FC) por una capa FC con una salida de 14 dimensiones, seguida de una capa de activación *sigmoid*. Para el entrenamiento del modelo se utilizó la estrategia de entrenamiento propuesta por [77] en la cuál define el optimizador *Adam* con parámetros estables de (β_1 = 0.9 y β_2 = 0.999) así como una tasa de aprendizaje inicial de *lr* = 0.001 con una disminución de un factor de 10 por época en caso de no reducir el valor de pérdida en sus resultados. Se estableció un número de épocas de 100 y un tamaño de lote de 32. Además, se configuró una transformación aleatoria de volteo horizontal (*horizontal flipping*) durante el entrenamiento con la finalidad de contribuir al aprendizaje del modelo de acuerdo a [77].

Modelo ViT y configuración de entrenamiento.

El modelo propuesto por [3], se caracteriza por manipular la imagen para obtener un *embedding* visual que pueda ser reconocido como entrada para alimentar al *encoder* del *Transformer*. Durante este proceso, es necesario dividir la imagen I en pequeños segmentos de 2D (dimensiones), los cuales son concatenados y transformados por una capa de proyección lineal de 1D. Una vez que se obtiene esta representación, un *token* especial es agregado a la representación visual así como la información de posición de cada segmento; esto permitirá identificar el orden cada elemento del *embedding* de entrada. El *encoder* del modelo utiliza un MLP(*Multi Layer Perceptron*) que permite arrojar la clasificación de la imagen. Para el entrenamiento del modelo se homogeneizaron los parámetros de la función de pérdida de *Cross-Entropy* y el optimizador Adam definidos anteriormente.

Clasificación binaria de Patologías

Debido al desbalance en las clases que incluye la base de datos *Chexnet-xray14*, y como se aprecia en la tabla 1, primero se realizó un entrenamiento binario por cada clase y se aplicaron dos estrategias para contribuir al aprendizaje del modelo.

- La primera estrategia, fue aplicar una serie transformaciones en las imágenes que pertenecen a clases con menos de 1000 unidades como se sugiere en [170],[171]. La clase de *Hernia* es la única que contiene menos de 1000 imágenes de casos positivos, por lo cuál es la única clase donde se aplicaron una serie de transformaciones geométricas basadas en [172]. Estas transformaciones incluyen una rotación por 10^o (en ambos lados), una translación en la imagen (tanto en eje x, como en el eje y)) y cambios en contraste de la imagen . La serie de transformaciones fue aplicada individualmente a las 227 imágenes de clases positivas de Hernia. En la figura 1 se incluye un ejemplo de las transformaciones realizadas. Las nuevas imágenes fueron almacenadas y registradas en la clase de *Hernia* obteniendo un nuevo total de clases positivas de 1135, como se muestra en la tabla 1 en la columna Pos + transformadas.

- La segunda estrategia utilizada fue reducir el tamaño de la clase mayoritaria que cuentan con más cantidad de imágenes. Esta técnica se identifica en la literatura como *subsampling* y consiste en reducir el tamaño de datos. Para este caso, se seleccionaron aleatoriamente una parte de las clases negativas con la finalidad de igualar la contribución de los casos positivos con los casos negativos durante el aprendizaje del modelo. Por ejemplo, la clase *Atelectasis* que incluye 11,559 clases positivas y 100,561 clases negativas, al aplicar esta estrategia solo se consideran 11,559 clases negativas.

Figura 1: Aplicación de técnicas para aumentar número de imágenes en la clase Hernia.

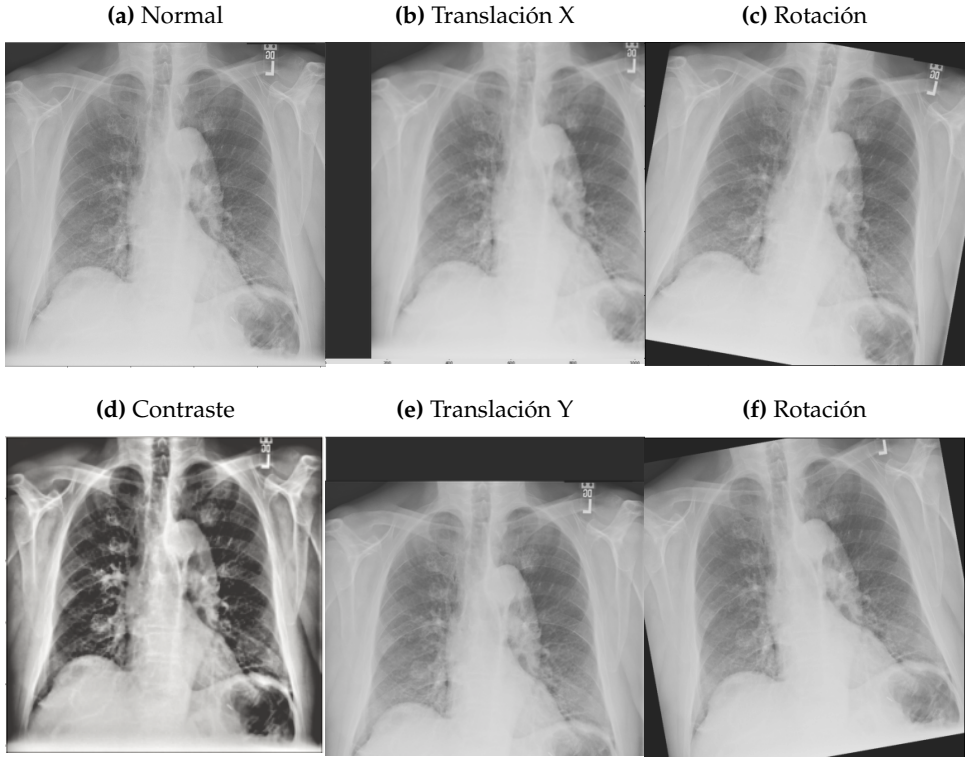


Tabla 1: Distribución de Clases en Base de datos Chest-xray14.

Patología	Positivos	Negativos	Pos+transformadas
Atelectasis	11559	100561	–
Cardiomegaly	2776	109344	–
Consolidation	4667	107453	–
Edema	2303	109817	–
Effusion	13317	98803	–
Emphysema	2516	109604	–
Fibrosis	1686	110434	–
Hernia	227	111893	1135
Infiltration	19894	92226	–
Mass	5782	106338	–
Nodule	6331	105789	–
Pleural Thickening	3385	108735	–
Pneumonia	1431	110689	–
Pneumothorax	5302	106818	–

Tabla 2: Partición de datos de imágenes de Chest-xray14 durante entrenamiento y prueba.

Experimento-1			
Patología	Entrenamiento	Validación	Prueba
Atelectasis	7319(+), 37817(-)	1878(+), 16245(-)	2362(+), 20363(-)
Cardiomegaly	1671(+), 43465(-)	507(+), 17616(-)	598(+), 22127(-)
Consolidation	2901(+), 42235(-)	789(+), 17334(-)	977(+), 21748(-)
Edema	1475(+), 43661(-)	408(+), 17715(-)	420(+), 22305(-)
Effusion	8169(+), 36967(-)	2190(+),15933(-)	2959(+), 19767(-)
Emphysema	1659(+), 43477(-)	336(+), 17787(-)	521(+),22204(-)
Fibrosis	1088(+), 44048(-)	253(+), 17870(-)	345(+), 22380(-)
Hernia	132(+), 45004(-)	38(+), 18005(-)	57(+), 22668(-)
Infiltration	12426(+), 32710(-)	3365(+), 14758(-)	4103(+), 18622(-)
Mass	3742(+), 41394(-)	803(+), 17320(-)	1237(+), 21488(-)
Nodule	4070(+), 41066(-)	949(+), 17174(-)	1312(+), 21413(-)
Pleural Thickening	2060(+), 43076(-)	544(+), 17579(-)	781(+), 21944(-)
Pneumonia	929(+), 44207(-)	235(+),17888(-)	267(+), 22458(-)
Pneumothorax	3429(+), 41707(-)	738(+), 17385(-)	1135(+), 21590(-)
Experimento-2			
Atelectasis	7319(+), 7319(-)	1878(+), 1878(-)	2362(+), 2362(-)
Cardiomegaly	1671(+), 1671(-)	507(+), 507(-)	598(+), 598(-)
Consolidation	2901(+), 2901(-)	789(+), 789(-)	977(+), 977(-)
Edema	1475(+), 1475(-)	408(+), 408(-)	420(+), 420(-)
Effusion	8169(+), 8169(-)	2190(+),2190(-)	2959(+), 2959(-)
Emphysema	1659(+), 1659(-)	336(+), 336(-)	521(+),521(-)
Fibrosis	1088(+), 1088(-)	253(+), 253(-)	345(+), 345(-)
Hernia	817(+), 817(-)	91(+), 91(-)	227(+), 227(-)
Infiltration	12426(+), 12426(-)	3365(+), 3365(-)	4103(+), 4103(-)
Mass	3742(+), 3742(-)	803(+), 803(-)	1237(+), 1237(-)
Nodule	4070(+), 4070(-)	949(+), 949(-)	1312(+), 1312(-)
Pleural Thickening	2060(+), 2060(-)	544(+), 544(-)	781(+), 781(-)
Pneumonia	929(+), 929(-)	235(+),235(-)	267(+), 267(-)
Pneumothorax	3429(+), 3429(-)	738(+), 738(-)	1135(+), 1135(-)

Los modelos fueron entrenados con cada una de las clases como clasificadores binario durante 100 épocas con un lote de 32 y utilizando la función de pérdida *Binary Cross Entropy* (BCELoss). En el Entrenamiento-1 no se incluyó el aumento de datos ni la transformación en la imagen como en [77]. En la tabla 2, se muestra la cantidad de imágenes utilizadas por patología durante el entrenamiento, validación y prueba.

Por otra parte, para el Entrenamiento-2 se aplicó el aumento de datos y la estra-

tegia de *subsampling* a las diferentes clases. El modelo se entrenó como clasificador binario durante l00 épocas, con un lote de 32 y con la función de perdida *Binary Cross Entropy* (BCELoss). La tabla 2 en la sección de Entrenamiento-2 muestra la cantidad de imágenes utilizadas para el entrenamiento, validación y prueba.

Entrenamiento Multiclase para la clasificación de Patologías

Considerando que una imagen de rayos X puede incluir varias patologías, se realizó un entrenamiento para identificar más de una patología en una imagen. Para realizar esta clasificación, los modelos de CNN y ViT fueron modificados para obtener una salida de 1×14 dimensiones; donde cada dimensión corresponde a una patología. Posteriormente, se estableció la definición de una etiqueta por imagen por medio de un vector y con 14 dimensiones:

$$y = \{y1, y2, ..., yn\} \quad \text{donde} \quad y \in \{0, 1\} \tag{1}$$

donde y sólo puede incluir valores binarios de 0 y 1. El valor de 1 identifica cuando una patología está presente, de lo contrario el valor es 0.

Ambos modelos fueron entrenado por 100 épocas, con un lote de 32 y utilizando la función de pérdida *Binary Cross Entropy* (BCELoss). Para esta clasificación que incluye las 14 clases, la estrategia *undersampling* no fue implementada debido a que las clases con una mayor cantidad de imágenes se reducen hasta en un 94 %. Por esta razón, se optó por incluir el peso por cada clase. Esto permite que una clasificación incorrecta de una clase minoritaria sea penalizada con un mayor peso que una clasificación incorrecta de una clase mayoritaria.

$$wc_i = 1 - \frac{fc_i}{N} \tag{2}$$

Donde wc_i representa el peso de la clase, fc_i el número de ocurrencias que se presentan en la muestra y N el total muestras incluidas en el lote.

Resultados orientados a la clasificación de imágenes de rayos X

La evaluación del modelo se realizó utilizando la métrica de área bajo la curva (AUC). Esta métrica permite realizar una representación de la efectividad del modelo para identificar las predicciones correctas. En otras palabras, si el modelo puede

distinguir entre un caso positivo y un caso negativo de las 14 patologías.

De acuerdo a los resultados obtenidos en la clasificación binaria del modelo, se identifica un mejor desempeño en el modelo propuesto por [3]. Estos resultados se pueden observar en la tabla 3, donde los valores más altos están remarcados con **negritas**. Sin embargo, también se observó que el aumento en los datos y la estrategia de *subsampling* contribuyeron a mejorar el desempeño de ambos modelo. Para algunas clases como *Hernia*, los resultados de AUC reflejan una gran contribución del balance de clases positivas y negativas así como por el aumento de datos; ya que se obtiene un incremento de 0.6015 a un 0.9545 en el modelo de CNN y de 0.6267 a un 0.9638. Por otra parte, es necesario considerar que los índices de detección por un especialista para la clase positiva de *Hernia* son mayores que una clase positiva de *Infiltration* [173]. Los resultados de los entrenamientos se muestran en la tabla 3 y la tabla 4

Tabla 3: Resultados de la clasificación Binaria en 14 patologías mediante Chexnet, CNN utilizando la métrica de AUC.

		CNN		**ViT[3]**	
Patología	**ChexNet[68]**	**SDA**	**DA**	**SDA**	**DA**
Atelectasis	0.8094	0.5078	0.8009	0.6034	**0.8534**
Cardiomegaly	**0.9248**	0.5815	0.8431	0.6832	0.9232
Consolidation	**0.7901**	0.6265	0.7520	0.6576	**0.8421**
Edema	**0.8878**	0.6626	0.8560	0.7626	**0.8960**
Effusion	0.8638	0.5781	0.8752	0.6781	**0.9052**
Emphysema	**0.9371**	0.5207	0.8904	0.6207	**0.9104**
Fibrosis	0.8047	0.5779	**0.8260**	0.5979	**0.8560**
Hernia	0.9164	0.6015	0.9545	0.6237	**0.9638**
Infiltration	0.7345	0.5726	0.7093	0.5926	**0.7493**
Mass	**0.8676**	0.5644	0.8317	0.5944	0.8517
Nodule	**0.7802**	0.5597	0.7488	0.5971	**0.7818**
Pleural Thickening	**0.8062**	0.5470	0.7650	0.6470	0.8050
Pneumonia	**0.7680**	0.5493	0.7478	0.6093	0.7578
Pneumothorax	**0.8887**	0.7425	0.8553	0.7825	0.8653
Promedio	0.8413	0.568	0.8183	0.6448	**0.8544**

Para el entrenamiento multi-clase, los resultados se muestran en la tabla 4 aplicando el aumento de datos. Como se mencionó anteriormente, los resultados con mejor desempeño se encuentran resaltados en negritas de los cuales se identifica una diferencia entre los valores obtenidos por [77] y el entrenamiento ViT. Esto in-

dica que las estrategias implementadas para combatir el desbalance en los datos permiten contribuir al aprendizaje del modelo.

Tabla 4: Resultados de la Multi-Clasificación en 14 patologías mediante Chexnet, CNNs y ViT utilizando la métrica de *AUC*

Patología	**ChexNet[68]**	**CNN**	**ViT[3]**
Atelectasis	0.8094	0.7937	**0.8367**
Cardiomegaly	**0.9248**	0.8954	0.9138
Consolidation	0.7901	0.7984	**0.8178**
Edema	0.8878	0.8854	**0.9102**
Effusion	0.8638	0.8762	**0.8823**
Emphysema	**0.9371**	0.8731	0.9268
Fibrosis	0.8047	0.8015	**0.8611**
Hernia	0.9164	0.8871	**0.9403**
Infiltration	**0.7345**	0.6923	0.7336
Mass	**0.8676**	0.8074	0.8586
Nodule	0.7802	0.7318	**0.7981**
Pleural Thickening	**0.8062**	0.7507	0.8010
Pneumonia	0.768	0.7432	**0.7877**
Pneumothorax	**0.8887**	0.8454	0.8827
Promedio	0.8414	0.8129	**0.8536**

De acuerdo a los resultados obtenidos en la tabla 3 y la tabla4, el modelo ViT obtiene resultados competitivos que permite extraer las características relevantes de las imágenes de rayos X de toráx. La estructura del modelo basada principalmente en mecanismos de atención demuestra su efectividad para obtener una representación de las imágenes médicas. Además aplicando estrategias de aumento de datos mejora sus resultados para la tarea de clasificación.

B: Publicaciones y Congresos

Publicaciones

Revistas arbitradas

- G. Ramirez-Alonso, O. Prieto-Ordaz, R. López-Santillan and M. Montes-Y-Gómez, "Medical Report Generation through Radiology Images: An Overview,ïn IEEE Latin America Transactions, vol. 20, no. 6, pp. 986-999, June 2022, doi: 10.1109/TLA.2022.9757742.

Revistas de Divulgación

- Ordaz, O. P.,& Flores, D. M. (2019). Segmentación semántica para reconocimiento de escenas. FINGUACH. Revista de Investigación Científica de la Facultad de Ingeniería de la Universidad Autónoma de Chihuahua, 6(19), 6-7.

Congresos

- O. Prieto-Ordaz, G. Ramírez-Alonso, L. C. González, R. López-Santillán and M. Montes-y-Gómez, "Brain Tumor Segmentation using an Encoder-Decoder Network with a Multiscale Feature Module,"2020 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2020, pp. 1-6, doi: 10.1109/ROPEC50909.2020.9258718.
- O. Prieto-Ordaz, 2022, 19-21 de Septiembre, Generación Automática de Reportes Médicos a partir de imágenes de rayos X basado en estrategias de aprendizaje profundo, 35º Congreso Nacional de Posgrado, 3MT.