UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Mejora de las prestaciones de un modelo para la generación de informes de radiología.

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

AUTOR/A: Aas Alas, Mohamed

Tutor/a: Paredes Palacios, Roberto

Cotutor/a: Albiol Colomer, Alberto

CURSO ACADÉMICO: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

Mejora de las prestaciones de un modelo para la generación de informes de radiología

TRABAJO FIN DE MÁSTER

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Autor: Mohamed Aas Alas

Tutor: Roberto Paredes Palacios
Alberto Albiol Colomer

Curso 2023-2024

Resum

La motivació d'aquest treball radica en la importància de millorar la precisió i eficiència en la generació automàtica d'informes de radiologia, un procés que pot agilitzar el diagnòstic i tractament dels pacients, així com reduir la càrrega de treball dels metges. Utilitzarem el *dataset* MIMIC-CXR, el qual conté radiografies de pacients i els seus respectius informes mèdics per a entrenar el model. La tasca específica consistix a generar informes detallats i precisos a partir d'imatges radiològiques utilitzant un model VED (*Vision Encoder Decoder*). L'objectiu principal d'aquest treball és millorar el rendiment del model de generació d'informes de radiologia esmentat, aplicant tècniques com el "*text augmentation*" sobre el *dataset* mitjançant l'ús de la API de OpenAI. L'estructura del treball s'enfocarà primer a intentar millorar les prestacions del model mitjançant diverses tècniques i, posteriorment, comparar les mètriques obtingudes amb l'estat de l'art actual, amb la finalitat d'avaluar l'efectivitat dels mètodes implementats i determinar possibles avanços en el camp.

Paraules clau: Radiografia, Informes mèdics, Generació automàtica

Resumen

La motivación de este trabajo radica en la importancia de mejorar la precisión y eficiencia en la generación automática de informes de radiología, un proceso que puede agilizar el diagnóstico y tratamiento de los pacientes, así como reducir la carga de trabajo de los médicos. Utilizaremos el *dataset* MIMIC-CXR, el cual contiene radiografías de pacientes y sus respectivos informes médicos para entrenar el modelo. La tarea específica consiste en generar informes detallados y precisos a partir de imágenes radiológicas utilizando un modelo VED (*Vision Encoder Decoder*). El objetivo principal de este trabajo es mejorar el rendimiento del modelo de generación de informes de radiología mencionado, aplicando técnicas como el "*text augmentation*" sobre el *dataset* mediante el uso de la API de OpenAI. La estructura del trabajo se enfocará primero en intentar mejorar las prestaciones del modelo mediante diversas técnicas y, posteriormente, comparar las métricas obtenidas con el estado del arte actual, con el fin de evaluar la efectividad de los métodos implementados y determinar posibles avances en el campo.

Palabras clave: Radiografía, Informes médicos, Generación automática

Abstract

The motivation of this work lies in the importance of improving the accuracy and efficiency in the automatic generation of radiology reports, a process that can speed up the diagnosis and treatment of patients, as well as reduce the workload of physicians. We will use the MIMIC-CXR dataset, which contains patient x-rays and their respective medical reports, to train the model. The specific task is to generate detailed and accurate reports from radiological images using a VED (Vision Encoder Decoder) model. The main objective of this work is to improve the performance of the mentioned radiology report generation model, applying techniques such as "text augmentation" on the dataset by using the OpenAI API. The structure of the work will first focus on trying to improve the performance of the model through various techniques and, subsequently, compare the obtained metrics with the current state of the art, in order to evaluate the effectiveness of the implemented methods and determine possible advances in the field.

Key words: Radiography, Medical Reports, Automatic Generation

Índice general

Índice general v
Índice de figuras vii
Índice de tablas vii

Agradecimientos ix
1 Introducción 1
1.1 Motivación . . . 1
1.2 Radiology Report Generation . . . 2
1.3 Fuentes de datos . . . 3
1.3.1 MIMIC-CXR . . . 3
1.4 Objetivos del trabajo . . . 4
1.5 Estructura de la memoria . . . 4
2 Estado del Arte 7
2.1 Principales propuestas . . . 7
2.1.1 Redes Neuronales Convolucionales (CNNs) y Recurrentes (RNNs) 7
2.1.2 Redes Preentrenadas . . . 11
2.1.3 Modelos GPT . . . 11
2.1.4 Técnicas Multimodales y Cross-Modal . . . 12
2.1.5 Modelos Vision Encoder-Decoder (VED) . . . 15
2.1.6 Grafos de Conocimiento . . . 18
2.1.7 Aprendizaje por Refuerzo (RL) . . . 20
2.1.8 Métricas de Evaluación y Calidad de informes . . . 20
2.2 Resumen general . . . 21
3 Metodología 23
3.1 Vision Encoder Decoder . . . 23
3.1.1 Vision Encoder . . . 23
3.1.2 Decoder . . . 24
3.2 Entrenamiento . . . 24
3.2.1 NLL . . . 25
3.2.2 RL . . . 25
3.3 Inferencia . . . 27
3.4 Preprocesamiento de datos . . . 27
3.4.1 Reordenamiento de las frase . . . 27
3.4.2 Text Augmentation con la API de OpenAI . . . 28
4 Experimentos y Resultados 31
5 Conclusiones Y Trabajos Futuros 35
5.1 Trabajos Futuros . . . 35

Apéndices
A Código en Python para la automatización de Reescritura de Informes de Radiología utilizando la API de OpenAI 41

B Relación del proyecto con los Objetivos del Desarrollo Sostenible **43**
B.0.1 Salud y bienestar (ODS 3) 43
B.0.2 Trabajo decente y crecimiento económico (ODS 8) 43
B.0.3 Industria, innovación e infraestructuras (ODS 9) 43
B.0.4 Ciudades y comunidades sostenibles (ODS 11) 44
B.0.5 Alianzas para lograr los objetivos (ODS 17) 44

Índice de figuras

1.1 Causas de consulta más frecuentes[9]. 1
1.2 imágenes del dataset MIMIC-CXR 3
1.3 Informe radiológico de las imágenes de la Figura 1.2. 4

2.1 CNN propuesta por Schlegl y cols.[29] 8
2.2 Modelo propuesto por Shin y cols.[30] 9
2.3 Modelo propuesto por Jing y cols.[12] 12
2.4 Modelo propuesto por Chen y cols.[3] 14
2.5 Arquitectura transformer.[32] 16
2.6 Esquema seq-seq.[14] 17
2.7 Modelo propuesto por Liu y cols.[18] 18

3.1 Diferencia Swin y ViT[21]. 24
3.2 Flujo de entrenamiento del modelo propuesto.[27] 25
3.3 Inferencia del modelo propuesto.[27] 27
3.4 Ejemplo de input tokenizado. 28
3.5 Ejemplo de Output tokenizado. 28
3.6 Informe antes del DA con gpt-3.5 30
3.7 Informe después del DA con gpt-3.5 30

Índice de tablas

3.1 Cálculo de costos por modelo para 152,173 informes 29

4.1 Resultados de los diferentes experimentos sobre la incorporación de informes generados en el entrenamiento del modelo. 31
4.2 Resultados de los experimentos realizados. 32
4.3 Comparación de los modelos del estado del arte para el conjunto de datos MIMIC-CXR 32

B.1 Evaluación de los Objetivos de Desarrollo Sostenible (ODS) según el nivel de importancia. 44

Agradecimientos

Quiero expresar mi más profundo agradecimiento a todas aquellas personas e instituciones que han sido fundamentales en el desarrollo de este proyecto y en mi formación durante este Máster en Inteligencia Artificial.

En primer lugar, a mis padres, por su apoyo incondicional y por ser una fuente constante de motivación. Gracias por siempre alentarme a seguir formándome y por estar a mi lado en cada paso del camino.

A mi tutor, Roberto Paredes Palacios, y a mi cotutor, Alberto Albiol Colomer, por ofrecerme la oportunidad de trabajar en un proyecto tan relevante y de vanguardia. Ha sido un tema que me ha fascinado desde el principio, y estoy profundamente agradecido por su ayuda constante, consejos y dedicación.

A Daniel Parres, investigador en inteligencia artificial en el PHRLT, quien me ayudó a arrancar con este proyecto y a quien debo muchas gracias por resolver cada una de mis dudas y por su apoyo técnico y personal.

Finalmente, a ValgrAI, la fundación que me brindó su apoyo para estudiar este máster. Agradezco profundamente su respaldo y compromiso con el futuro, este paso tan importante en mi vida académica ha sido más memorable gracias a su compromiso.

Gracias a todos por formar parte de este camino.

CAPÍTULO 1
Introducción

En este capítulo se presenta una introducción al Trabajo realizado, donde se abordará en primer lugar la motivación que ha impulsado este proyecto, seguido de una descripción del problema a tratar, la generación automática de informes de radiología, destacando su importancia en el ámbito médico. A continuación, se discutirán las fuentes de datos empleadas para entrenar y evaluar el modelo, y finalmente, se enunciarán los objetivos principales del trabajo, concluyendo con una visión general de la estructura de la memoria que guiará el desarrollo del documento.

1.1 Motivación

La radiología desempeña un papel fundamental en la medicina moderna, ya que es una herramienta crucial para el diagnóstico y tratamiento de diversas patologías. Según un estudio realizado en la Universidad de La Sabana, Colombia, un 7,36 % de las consultas médicas están relacionadas con lesiones osteomusculares[9], muchas de las cuales requieren radiografías para un diagnóstico preciso.

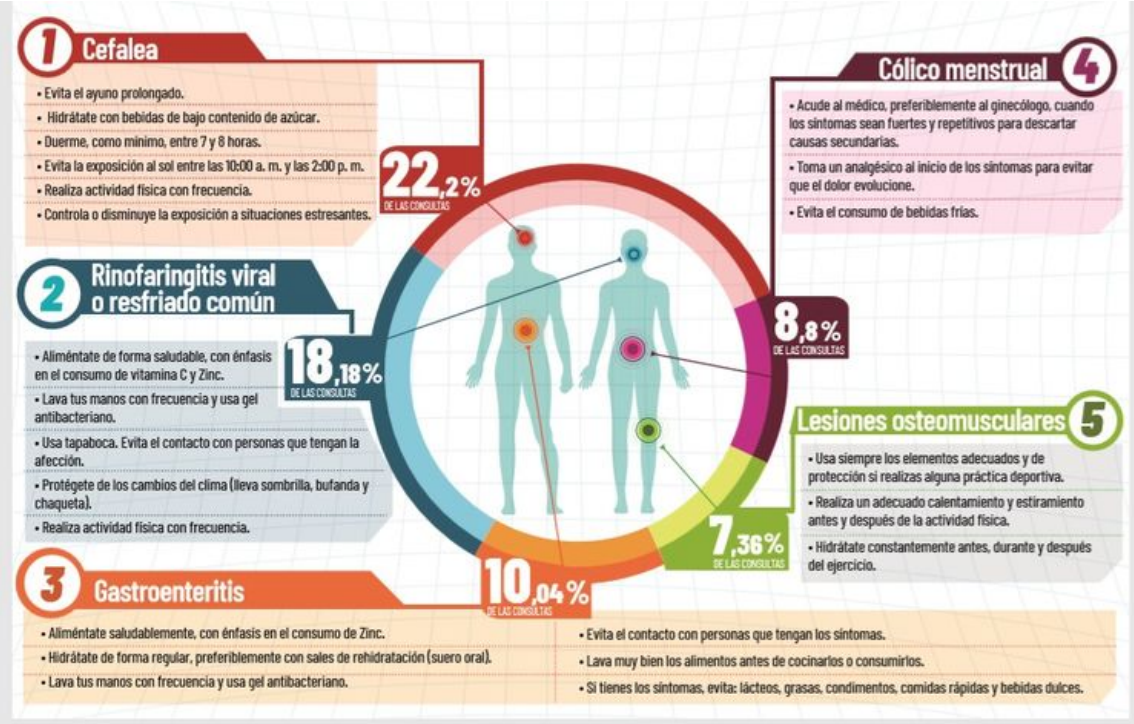


Figura 1.1: Causas de consulta más frecuentes[9].

Estas lesiones pueden involucrar fracturas o alteraciones óseas que, de no ser diagnosticadas correctamente, podrían llevar a complicaciones graves en la salud del paciente. En este contexto, la interpretación de imágenes radiológicas se convierte en un proceso esencial para la toma de decisiones clínicas.

Sin embargo, la interpretación de estas imágenes no está exenta de desafíos. Es un proceso complejo que depende en gran medida de la experiencia y el conocimiento del radiólogo, lo que puede introducir variaciones en los diagnósticos.

Además, la creciente lista de espera en los hospitales y el aumento en la cantidad de imágenes que se generan diariamente añaden una presión significativa sobre los profesionales de la salud. Estos deben producir informes de alta calidad en menos tiempo, lo que puede conducir a errores o retrasos en los diagnósticos, afectando negativamente la atención al paciente.

Bajo estas circunstancias, la necesidad de soluciones tecnológicas se vuelve evidente. La inteligencia artificial ha comenzado a ser explorada como una herramienta prometedora para automatizar la generación de informes radiológicos. Estas tecnologías no solo tienen el potencial de mejorar la eficiencia al reducir los tiempos de espera, sino que también pueden aumentar la precisión en el diagnóstico al minimizar las variaciones y errores humanos. En definitiva, la integración de la inteligencia artificial en la radiología representa un avance significativo hacia una atención médica más rápida, precisa y eficiente, alineada con las crecientes demandas del sistema de salud moderno.

1.2 Radiology Report Generation

La generación de informes médicos (MRG, por sus siglas en inglés) hace referencia a un problema en el que se usan técnicas de inteligencia artificial para producir automáticamente informes médicos a partir de datos obtenidos de imágenes médicas. Esta tecnología puede facilitar a los médicos la toma de decisiones más rápidas y precisas, ya que la tarea de redactar informes no solo es laboriosa, sino también propensa a errores, incluso para especialistas experimentados.[26]

Por otro lado, dentro de la generación de informes médicos tenemos la generación de informes radiológicos (RRG, por sus siglas en inglés), esta es una tarea compleja que se enfoca en interpretar imágenes de radiografías y elaborar informes detallados sobre posibles patologías en los pacientes. A diferencia de las tareas convencionales de visión por computador, que se centran en identificar objetos dentro de imágenes, la RRG se orienta hacia el diagnóstico de patologías, evaluando su presencia, ausencia o incertidumbre. Además, la limitada disponibilidad de datos y la diversidad inherente de los informes médicos representan desafíos importantes. Esta tarea es esencial para optimizar y mejorar la atención al paciente, ya que permite un análisis exhaustivo de su estado de salud y una detección temprana de enfermedades. Asimismo, sirve como una herramienta de apoyo crucial para los profesionales de la salud. La tarea de RRG comparte similitudes con otras áreas que manejan datos complejos, donde los modelos entrenados deben cumplir con estrictos criterios de seguridad debido a la importancia de diagnosticar con precisión, ya que la salud del paciente está en juego. Los enfoques actuales se basan principalmente en el *deep learning*, específicamente en la arquitectura *vision encoder–decoder*, incorporando componentes como memorias o *reinforcement learning* para mejorar el rendimiento.

1.3 Fuentes de datos

1.3.1. MIMIC-CXR

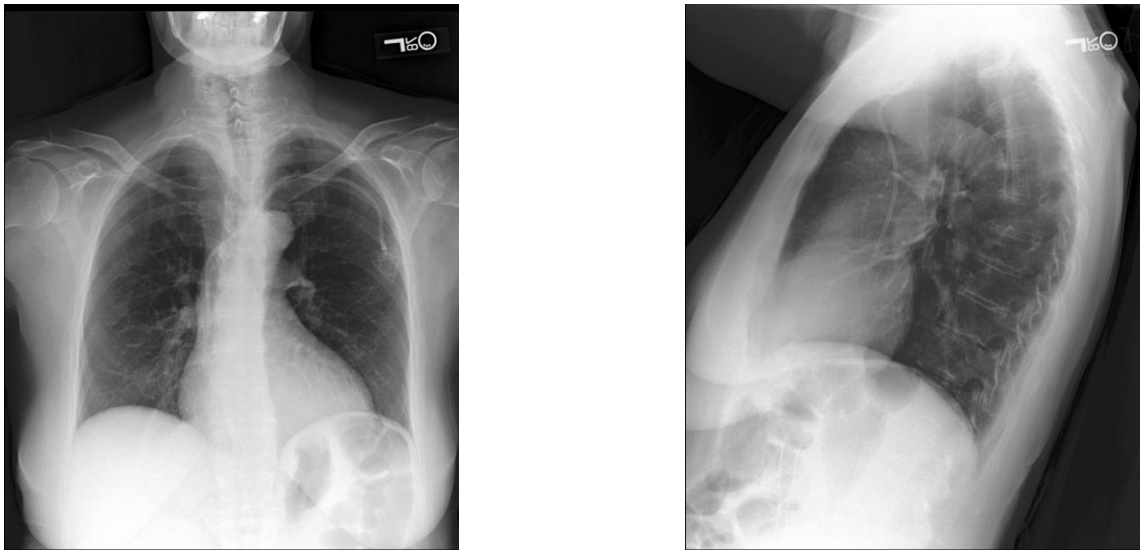
La radiografía de tórax es una herramienta de diagnóstico por imagen sumamente efectiva, permitiendo una inspección detallada del pecho de un paciente, aunque su correcta interpretación requiere formación especializada. Con el avance de los algoritmos de visión por computador, el análisis automatizado y preciso de radiografías de tórax ha despertado un creciente interés entre los investigadores.

En este contexto, presentamos *MIMIC-CXR*, uno de los conjuntos de datos más grandes y detallados disponibles para la investigación en radiología. Este *dataset* incluye 227,835 estudios de imágenes de 65,379 pacientes que acudieron al Departamento de Emergencias del Centro Médico Beth Israel Deaconess entre 2011 y 2016. Cada estudio puede contener una o más imágenes, generalmente una vista frontal y una vista lateral, con un total de 377,110 imágenes disponibles.[13]

Los estudios están acompañados por informes radiológicos en texto libre semi-estructurado, redactados por radiólogos durante la atención clínica rutinaria, los cuales describen los hallazgos radiológicos. En total, el conjunto de datos incluye 152,173 informes radiológicos para el entrenamiento de modelos, 2,300 para el conjunto de validación y 2,300 informes adicionales para *testing*. Cada informe está vinculado a una o más imágenes radiográficas de tórax, y los modelos entrenados con este *dataset* buscan generar un informe que detalle todas las patologías presentes en la imagen médica. Cabe destacar que en este proyecto se excluyeron las muestras cuyos informes no contenían hallazgos patológicos, ya que no aportan información relevante para el entrenamiento de los modelos.

Además, en cada sesión de Rayos X hay hasta 3 imágenes, lo que significa que cada informe puede hacer referencia a 1, 2 o 3 imágenes radiográficas. El conjunto de datos es público y se pone a disposición de manera gratuita para facilitar y fomentar una amplia gama de investigaciones en visión por computador, procesamiento del lenguaje natural y en este caso lo usaremos para resolver el problema 1.2 Radiology Report Generation.

A continuación, tenemos unas imágenes de muestra del *dataset* junto a su informe.



(a) Frontal view **(b)** Lateral view

Figura 1.2: imágenes del dataset MIMIC-CXR

"Left-sided Port-A-Cath tip terminates in the low SVC. Heart size is mildly enlarged, but decreased in size compared to the previous exam. The mediastinal and hilar contours are unchanged with tortuosity of thoracic aorta again noted. Also again noted is indentation upon the right aspect of the trachea at the thoracic inlet due to the presence of a large thyroid goiter, as seen on prior CT. The pulmonary vasculature is normal. The lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities. A common bile duct stent is incompletely assessed."

Figura 1.3: Informe radiológico de las imágenes de la Figura 1.2.

1.4 Objetivos del trabajo

El objetivo principal de este trabajo es implementar un modelo de generación de informes de radiología, implementando un enfoque robusto y eficaz que no solo se base en las capacidades inherentes del modelo, sino que también utilice técnicas avanzadas para optimizar los resultados. Entre estas técnicas, destaca el "*text augmentation*", una estrategia que se aplicará sobre el *dataset* con el fin de enriquecer y diversificar la información disponible para el entrenamiento del modelo. El "*text augmentation*" permite generar variantes de los textos existentes, incorporando sinónimos, modificando la estructura gramatical, o incluso introduciendo ligeras variaciones en el contenido que no comprometan la precisión de la información médica, pero que proporcionen al modelo una mayor exposición a diferentes formas de expresar el mismo diagnóstico.

Para llevar a cabo estas modificaciones y enriquecer los datos, se pueden utilizar diversas bibliotecas de *Machine Learning* (ML) que están diseñadas específicamente para realizar *data augmentation* en texto. Además, es posible emplear modelos de lenguaje (LLMs) para generar automáticamente estas variantes textuales, asegurando que el modelo se entrene con una amplia gama de ejemplos que representen la diversidad de expresiones y formulaciones posibles en los informes de radiología.

Este enfoque no solo tiene el potencial de mejorar la capacidad del modelo para generar informes coherentes y precisos, sino que también podría acercarse o incluso superar el estado del arte actual en la generación automática de informes médicos. La implementación de un modelo que integra estas técnicas proporciona mejoras significativas en términos de rendimiento y generalización, al dotar al sistema de una mayor flexibilidad lingüística y una comprensión más profunda de las posibles variaciones en los informes radiológicos.

En última instancia, este trabajo busca contribuir al avance de la inteligencia artificial en el campo de la salud, ofreciendo herramientas que apoyen a los profesionales médicos en la interpretación de las imágenes radiológicas de manera más eficiente y precisa.

1.5 Estructura de la memoria

Esta memoria se divide en los siguientes capítulos:

- El primer capítulo introduce el trabajo, comenzando con la motivación que impulsa el proyecto, seguido de una explicación sobre la generación de informes de radiología. También se describen las fuentes de datos utilizadas, se establecen los objetivos del trabajo y se presenta la estructura del documento.

- En el segundo capítulo, se presenta una revisión de las principales propuestas en el área de estudio. Se ofrece un resumen general de las tecnologías y enfoques relevantes, proporcionando el contexto necesario para entender la evolución del campo.
- El tercer capítulo detalla la metodología empleada en el desarrollo del proyecto. Se introduce el *Vision Encoder Decoder*, desglosando su funcionamiento en *Vision Encoder* y *Decoder*. Además, se explican los procesos de entrenamiento del modelo, diferenciando entre NLL y RL, y se describe el proceso de inferencia. Finalmente, se aborda el preprocesamiento de datos, incluyendo el uso de herramientas como OpenAI.
- El cuarto capítulo se enfoca en los experimentos realizados y los resultados obtenidos. Se documenta cómo se implementaron las pruebas y se analizan los resultados en detalle, destacando las implicaciones de los hallazgos.
- El quinto y último capítulo presenta las conclusiones derivadas del trabajo realizado. Además, se discuten posibles trabajos futuros que podrían llevarse a cabo para expandir y mejorar los resultados del proyecto.

CAPÍTULO 2
Estado del Arte

En este capítulo se comentará el estado del arte de las técnicas y modelos más relevantes en la clasificación y detección de patologías en imágenes médicas, particularmente en rayos X. A lo largo de este capítulo, se describirán los avances más significativos en la aplicación de redes neuronales profundas y otros enfoques de aprendizaje automático en el campo de la radiología, destacando tanto los modelos que han establecido nuevas bases teóricas como aquellos que han demostrado mejoras prácticas en la generación automática de informes radiológicos. Además, se discutirán los desafíos actuales en el uso de estas tecnologías, así como las limitaciones de las métricas de evaluación utilizadas para medir su efectividad.

2.1 Principales propuestas

En esta sección se presentan las principales propuestas desarrolladas en la literatura reciente para abordar el problema de la detección y clasificación de patologías en imágenes médicas, así como el de la generación de los respectivos informes médicos. Se organiza en función de las tecnologías empleadas, como redes neuronales convolucionales (CNNs), redes neuronales recurrentes (RNNs), modelos basados en transformers, y técnicas multimodales, entre otras. Cada subsección describe los trabajos más influyentes dentro de cada enfoque, resaltando sus contribuciones clave y su impacto en el avance del estado del arte.

2.1.1. Redes Neuronales Convolucionales (CNNs) y Recurrentes (RNNs)

Las Redes Neuronales Convolucionales (CNNs) son arquitecturas de *deep learning* que han demostrado ser sumamente eficaces en el problema de *Radiology Report Generation*. Las CNNs son especialmente útiles en el procesamiento y análisis de imágenes médicas debido a su capacidad para identificar y aprender patrones en dichas imágenes. Esto se logra a través de múltiples capas convolucionales que extraen características visuales de diferentes niveles de abstracción, lo cual es fundamental para capturar detalles específicos de las imágenes de radiología, como lesiones, anomalías o estructuras anatómicas.

Schlegl y cols. utilizaron una CNN para clasificar patrones de tejido en tomografías, con descripciones semánticas en los informes como etiquetas. Este método ayudó a abrir el camino para el uso del *deep learning* en la imagenología médica.[29]

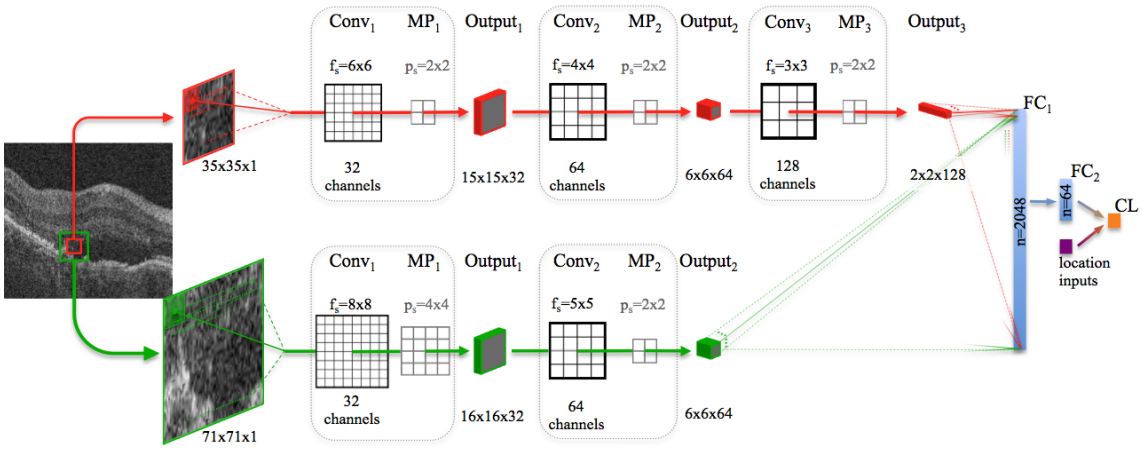


Figura 2.1: CNN propuesta por Schlegl y cols.[29]

A continuación, explicamos las diferentes partes del modelo:

- La arquitectura toma **dos entradas de parches de imagen** de diferentes tamaños:
  - **35x35** píxeles (representada por la ruta roja en el diagrama).
  - **71x71** píxeles (representada por la ruta verde en el diagrama).

  Estos parches de entrada se extraen de la misma posición de la imagen grande, pero a diferentes escalas, lo que permite que el modelo procese diferentes niveles de detalle de la imagen.
- Cada ruta de entrada (parches de 35x35 y 71x71) pasa por varias capas de **Convolución** y **Max-Pooling**. Estos son los bloques básicos de una CNN:
  - Las capas de **Convolución (Conv)** se utilizan para extraer características de la imagen aplicando filtros (denotados como f_s en el diagrama). El tamaño de estos filtros es diferente para cada capa de convolución.
  - Las capas de **Max-Pooling (MP)** se aplican después de cada convolución para reducir la dimensionalidad (denotadas como p_s en el diagrama), lo que ayuda al modelo a enfocarse en las características más importantes y a ser computacionalmente eficiente.
- Ruta 1 (Roja, Entrada 35x35):
  - **Conv1**: filtros de 6x6, seguida por **MP1**: agrupamiento de 2x2 → Salida: 15x15x32.
  - **Conv2**: filtros de 4x4, seguida por **MP2**: agrupamiento de 2x2 → Salida: 6x6x64.
  - **Conv3**: filtros de 3x3, seguida por **MP3**: agrupamiento de 2x2 → Salida: 2x2x128.
- Ruta 2 (Verde, Entrada 71x71):
  - **Conv1**: filtros de 8x8, seguida por **MP1**: agrupamiento de 4x4 → Salida: 16x16x32.
  - **Conv2**: filtros de 5x5, seguida por **MP2**: agrupamiento de 2x2 → Salida: 6x6x64.

  Es importante notar que esta ruta solo tiene dos pares de capas de convolución y max-pooling, mientras que la ruta roja tiene tres.
- Después de las operaciones de convolución y max-pooling, ambas **rutas roja y verde** envían sus salidas a una **capa *fully conected* (FC1)** con 2048 neuronas.

- Las salidas de ambas rutas están **densamente conectadas**, lo que significa que cada neurona en la primera capa *fully conected* recibe entrada de todas las neuronas de ambas rutas.

Luego sigue una segunda capa *fully conected* **(FC2)** con 64 neuronas.

- Después de la segunda capa totalmente conectada, las activaciones se pasan a la **capa de clasificación (CL)** final, que es la responsable de producir las predicciones de clase.

  - Además, se proporcionan ***location inputs*** en esta etapa. Estas entradas representan información adicional sobre la imagen, como características espaciales o relacionadas con la posición, y se combinan con las activaciones de la segunda capa totalmente conectada antes de la clasificación.

Por otro lado, las Redes Neuronales Recurrentes (RNNs), y en particular las LSTM (*Long Short-Term Memory*) o GRU (*Gated Recurrent Unit*), son especialmente adecuadas para procesar secuencias de datos, como las descripciones textuales en los informes de radiología. Las RNNs pueden modelar dependencias temporales y contextuales en las secuencias, lo que les permite generar texto coherente y clínicamente relevante al describir los hallazgos en las imágenes médicas. En el contexto del problema RRG, las RNNs pueden recibir las características extraídas por una CNN y generar descripciones detalladas que incluyen no solo la identificación de patologías, sino también su localización, gravedad y otras observaciones clínicas.

Shin y cols. desarrollaron una CNN para imágenes de rayos X de tórax, combinada con una RNN para anotar enfermedades, anatomía y severidad, marcando un paso importante en la automatización de la anotación de imágenes médicas.[30]

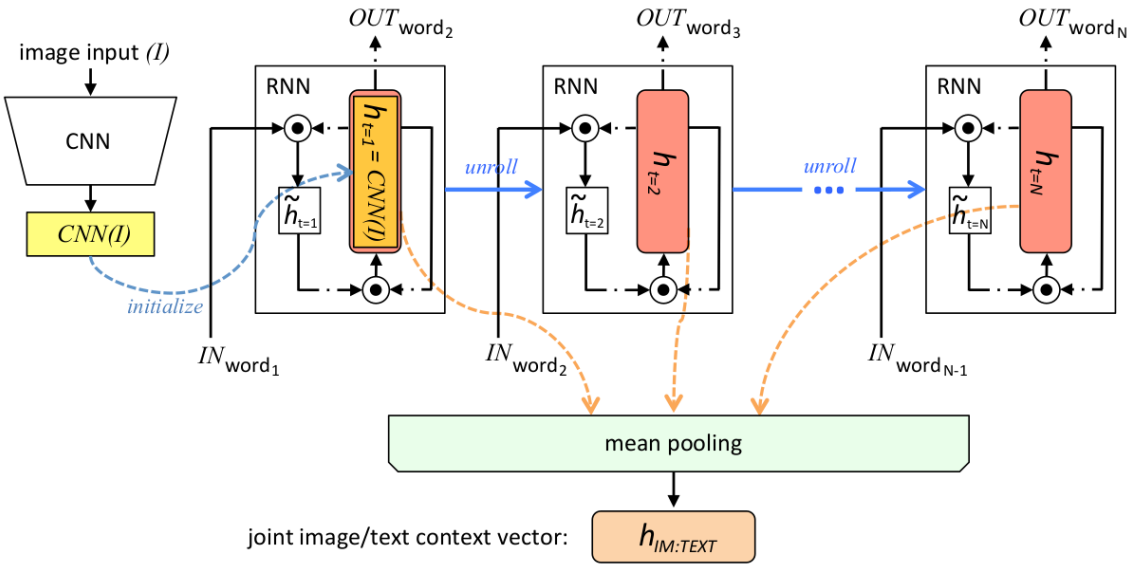


Figura 2.2: Modelo propuesto por Shin y cols.[30]

A continuación, explicamos las diferentes partes del modelo:

1. **Entrada de Imagen a la CNN**:

   - La imagen *I* se introduce primero en una **Red Neuronal Convolucional (CNN)**.

- La CNN procesa la imagen y genera una **representación o embedding de la imagen**, denotado como $CNN(I)$. Este embedding captura la información visual de la imagen.

2. **El Embedding de la CNN como Estado Inicial de la RNN**:
   - La salida de la CNN, $CNN(I)$, se utiliza para **inicializar el estado oculto** de la **Red Neuronal Recurrente (RNN)** en el paso de tiempo $t = 1$. Esto se muestra en el cuadro amarillo etiquetado como $h_{t=1} = CNN(I)$.
   - Este paso garantiza que la RNN comience su proceso de generación (o comprensión) de texto con el contexto proporcionado por la imagen.
3. **Procesamiento de Secuencias de Palabras por la RNN**:
   - A continuación, la RNN se **desenrolla** sobre una secuencia de palabras. Cada palabra de la secuencia se proporciona como entrada a la RNN **en cada paso de tiempo**.
   - La RNN actualiza su estado oculto h_t mientras procesa cada palabra secuencialmente: $IN_{word_1}, IN_{word_2}, \ldots, IN_{word_{N-1}}$. Estas entradas representan las palabras en el texto que se están procesando junto con la imagen.
   - Los estados ocultos $h_{t=2}, h_{t=3}, \ldots, h_{t=N}$ evolucionan conforme la RNN procesa cada palabra en la secuencia.
4. **Mean Pooling sobre los Estados Ocultos de la RNN**:
   - Después de que la RNN procesa todas las palabras en la secuencia, se aplica **mean-pooling** (promedio) a los estados ocultos a lo largo de todos los pasos de tiempo.
   - Esto significa que los estados ocultos de cada palabra en la secuencia se promedian, lo que resulta en un **vector de contexto conjunto de imagen y texto** $h_{IM:TEXT}$, que combina la información tanto de la imagen (a través de la CNN) como del texto (a través de la RNN).
5. **Parámetros Compartidos de la RNN**:
   - La misma RNN se utiliza en cada paso de la secuencia de palabras, es decir, la RNN que procesa la primera palabra, la segunda palabra, y así sucesivamente, es la misma, tanto en estructura como en sus parámetros. En otras palabras, **no se entrena una RNN distinta para cada paso de la secuencia**, sino que **los pesos y parámetros de la RNN se reutilizan en cada paso**.
   - Esto ayuda a reducir la complejidad del modelo y asegura un aprendizaje coherente a lo largo de la secuencia.

Otros proyectos relacionados con los previamente mencionados son los siguientes:

Moradi y cols. propusieron una combinación de bloques CNN y RNN para identificar regiones de interés en rayos X, lo que mejoró la precisión de las anotaciones de imágenes médicas.[23]

Rubin y cols. utilizaron CNNs entrenadas en paralelo para analizar vistas frontales y laterales de rayos X de tórax, con el objetivo de mejorar la estimación de patologías al considerar múltiples perspectivas.[28]

Liu y cols. introdujeron una CNN para extraer características visuales de las imágenes, seguida de un decodificador de oraciones y palabras para generar informes de

radiología. Su modelo se refinó utilizando aprendizaje por refuerzo para un mejor rendimiento.[19]

La RNN en el modelo de Shin y cols. trabajó junto con una CNN para anotar conjuntamente rayos X de tórax, ayudando a capturar patrones secuenciales en los datos médicos.[30]

En el trabajo de Moradi y cols., el bloque RNN se utilizó para procesar las salidas de la CNN, ayudando de manera efectiva en la identificación de regiones de interés al considerar la información temporal.[23]

2.1.2. Redes Preentrenadas

Las redes preentrenadas han demostrado ser extremadamente útiles en tareas complejas como la generación de informes de radiología (RRG). Estas redes, que han sido entrenadas previamente en grandes conjuntos de datos, poseen una capacidad inherente para captar y generalizar características de alto nivel, lo que les permite ser adaptadas eficientemente a nuevas tareas con una cantidad limitada de datos específicos del dominio. Al utilizar redes preentrenadas, los modelos pueden beneficiarse de este conocimiento previo, reduciendo el tiempo de entrenamiento y mejorando la precisión en la generación de informes. A continuación, describimos dos proyectos en los que se han usado CNN preentrenadas:

Modelos como TieNet aprovecharon redes preentrenadas y técnicas neuronales avanzadas como mecanismos de co-atención para mejorar significativamente el rendimiento de los modelos de generación de informes de radiología (RRG).[33]

Li y cols. avanzaron en la generación de informes de radiología utilizando redes preentrenadas, contribuyendo a una clasificación de enfermedades y generación de informes más precisa y eficiente.[16]

2.1.3. Modelos GPT

Los modelos de lenguaje como GPT (*Generative Pre-trained Transformer*) han demostrado ser altamente efectivos en tareas de procesamiento del lenguaje natural (NLP), incluida la generación de informes de radiología. Una de las principales ventajas de utilizar modelos GPT es su capacidad para comprender y generar texto coherente, lo que es esencial para la creación de informes médicos precisos y detallados. Los modelos GPT, al estar preentrenados en grandes volúmenes de datos textuales, poseen un conocimiento amplio y contextual que les permite producir informes con un alto grado de naturalidad y precisión.

Otra ventaja es la consistencia en la calidad de los informes generados. Los modelos GPT pueden ser entrenados para seguir un formato específico y utilizar una terminología médica estandarizada, lo que minimiza el riesgo de errores humanos y garantiza que los informes cumplan con los estándares requeridos. Además, la capacidad de los modelos GPT para aprender y adaptarse a nuevos datos significa que pueden mejorar continuamente a medida que se les proporciona más información, lo que es crucial para mantener la precisión en un campo tan dinámico como la radiología. A continuación, describimos dos proyectos en los que se han usado modelos GPT:

Alfarghaly y cols. utilizaron DistilGPT2 para la generación de informes de radiología, enfatizando tiempos de entrenamiento más rápidos y mejores métricas, aunque señalaron la necesidad de conjuntos de datos más grandes para mejorar la generalización.[1]

El modelo CvT2DistilGPT2 de Nicolson y cols. destacó la superioridad de GPT2 en el arranque en caliente del decodificador, mejorando el rendimiento de la generación de informes de radiología.[24]

2.1.4. Técnicas Multimodales y Cross-Modal

Dado que trabajaremos con dos tipos de datos distintos, informes médicos en formato de texto e imágenes de radiografías, resulta particularmente interesante explorar técnicas multimodales. Estas técnicas permiten combinar de manera efectiva la información obtenida de ambos dominios, potenciando la capacidad de análisis y la extracción de características complementarias a través de enfoques cross-modal.

El uso de técnicas multimodales y cross-modal en la generación de informes de radiología (RRG) presenta varias ventajas significativas. Estas técnicas permiten la integración y fusión de diferentes tipos de datos, como imágenes médicas y texto, lo que mejora la capacidad del modelo para captar y representar la información compleja presente en los estudios radiológicos.

En primer lugar, las técnicas multimodales permiten combinar la información visual de las imágenes de radiología con descripciones textuales, lo que proporciona un contexto más rico y detallado. Esto es crucial para la generación de informes precisos, ya que los modelos pueden considerar tanto las características visuales como las semánticas en su proceso de generación. Además, la capacidad de integrar múltiples fuentes de datos mejora la sensibilidad del modelo para detectar y caracterizar condiciones médicas sutiles que podrían pasarse por alto si se consideraran de manera aislada.[11]

Jing y cols. se centraron en la integración de datos multimodales con mecanismos de co-atención, mejorando la comprensión contextual necesaria para una efectiva generación de informes de radiología.[12].

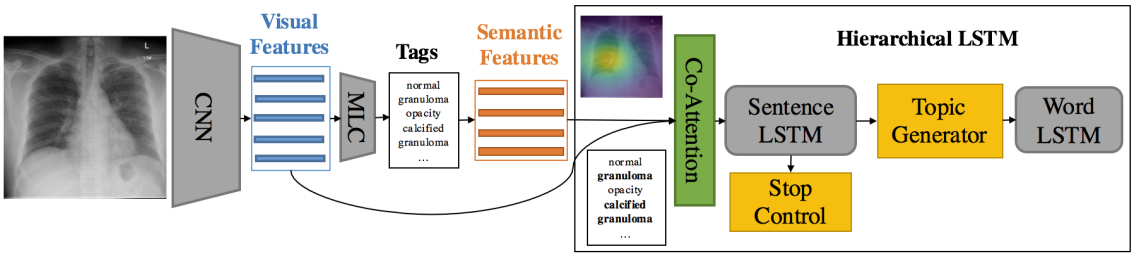


Figura 2.3: Modelo propuesto por Jing y cols.[12]

A continuación, explicamos las diferentes partes del modelo:

1. **Imagen de Entrada (Radiografía de Tórax)**: El proceso comienza con una imagen de rayos X de tórax como entrada principal del modelo. Esta imagen será procesada para extraer características visuales y semánticas que se utilizarán en las siguientes etapas.

2. **CNN (Red Neuronal Convolucional)**: Se utiliza una CNN para extraer las **características visuales** de la radiografía de tórax, como estructuras anormales, lesiones u otras áreas de interés.

3. **Clasificación Multi-etiqueta (MLC)**: La salida de la CNN se pasa a través de una red de **clasificación multi-etiqueta (MLC)**. Esta red predice múltiples etiquetas o

"tags" relacionados con el contenido visual de la radiografía, tales como "granuloma", "opacidad", "granuloma calcificado", etc. Estas etiquetas representan términos médicos que describen observaciones en la imagen.

4. **Etiquetas (Tags)**: Las etiquetas predichas por la red MLC (por ejemplo, "granuloma", "granuloma calcificado") se utilizan para generar las **características semánticas**.

5. **Características Semánticas**: Estas **características semánticas** representan embeddings de las etiquetas predichas. Se utilizan para proporcionar al modelo una comprensión a nivel superior de los conceptos médicos identificados en la radiografía. Los embeddings de palabras son probablemente pre-entrenados en un corpus médico, lo que permite al modelo mapear términos médicos a vectores significativos.

6. **Mecanismo de Co-Atención**: El modelo emplea un mecanismo de **co-atención**, que atiende tanto a las **características visuales** obtenidas de la CNN como a las **características semánticas** (los embeddings de las etiquetas predichas). En este caso, las etiquetas en negrita "granuloma" y "granuloma calcificado" indican que el mecanismo de co-atención se centra en estas características en relación con la imagen.

7. **LSTM**: Esta consta de varias etapas para la generación de descripciones basadas en radiografías. Primero, las características visuales y semánticas atendidas por el mecanismo de co-atención se pasan a un LSTM de oraciones, que forma oraciones coherentes a partir de los hallazgos médicos identificados. Un mecanismo de control de parada decide cuándo detener la generación de oraciones, evitando que el texto sea demasiado corto o excesivo. Además, un generador de temas asegura que las oraciones mantengan relevancia con los hallazgos clínicos, mientras que un LSTM de palabras genera las palabras individuales dentro de cada oración basándose en los temas y las características previamente obtenidas.

Por otro lado, las técnicas cross-modal facilitan la alineación y el intercambio de información entre diferentes modalidades, como la imagen y el texto. Esto permite que los modelos generen informes que no solo describen lo que se ve en una imagen, sino que también lo contextualizan en términos de la historia clínica del paciente y los conocimientos médicos actualizados. La fusión de características cross-modal también mejora la localización precisa de anomalías y la interpretación de patrones complejos, lo que es esencial en un campo tan crítico como la radiología.[11]

Chen y cols. proponen un enfoque basado en redes de memoria cross-modal (CMN) que mejora la interacción entre modalidades, permitiendo una mayor precisión en la generación de informes radiológicos. Su modelo utiliza una matriz de memoria para almacenar la información cross-modal, facilitando el proceso de consulta y respuesta de memoria, en el que se extraen los vectores de memoria más relevantes y se ponderan de acuerdo con las características visuales y textuales de entrada. Estas respuestas se introducen en el codificador y decodificador para generar informes más detallados y precisos, gracias a la información explícitamente aprendida de la interacción entre imágenes y textos.[3]

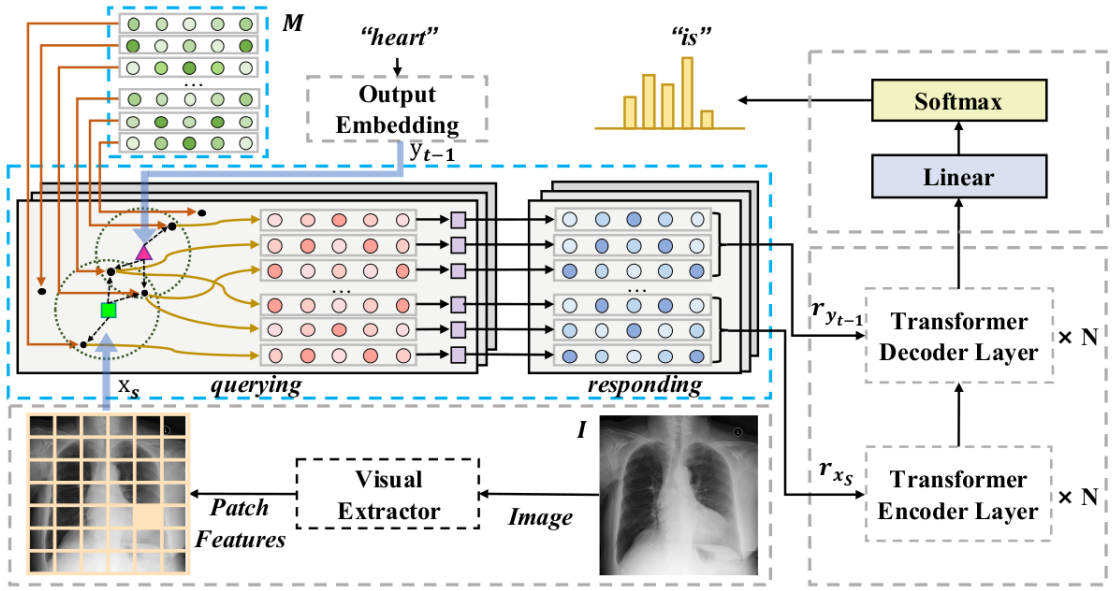


Figura 2.4: Modelo propuesto por Chen y cols.[3]

A continuación, explicamos las diferentes partes del modelo:

1. ***Visual Extractor* (Extractor Visual)**:
   - En la parte inferior izquierda de la imagen, se observa una radiografía de tórax etiquetada como *Image*. Esta imagen es procesada a través de un extractor visual que divide la imagen en *patch features* (características de parches), dividiendo la imagen en pequeñas regiones o *patches* para extraer información visual.
   - Este bloque convierte la información visual en un formato más manejable que puede ser usado por otras capas del modelo, como la capa de transformadores.
2. ***Transformer Encoder Layer* (Capa de Codificador del Transformer)**:
   - La información visual procesada (representada por x_s) se envía a la capa de codificación del *transformer*. El codificador es responsable de procesar las características de entrada de los parches visuales y crear una representación que luego será utilizada por las siguientes capas.
   - Esta capa incluye múltiples capas de *transformers* (indicadas como $\times N$).
3. ***Memory Network* (Red de Memoria Cross-Modal)**:
   - Esta red se encarga de manejar las interacciones entre diferentes modalidades (visual y textual). Se observa que la red de memoria está dividida en dos partes principales:
     - **Querying (Consultando)**: A la izquierda de la imagen, se muestran las consultas visuales (indicadas como *querying*) donde las representaciones visuales extraídas por el *encoder* interactúan con la memoria almacenada.
     - **Responding (Respondiendo)**: En la parte central, las respuestas desde la memoria (en azul) interactúan con el modelo a través de una serie de operaciones, proporcionando contexto visual y textual relevante.
4. ***Output Embedding* (Embeddings de Salida)**:

- En el proceso secuencial de generar texto a partir de las imágenes, el modelo también utiliza un *embedding* de salida de una palabra previa (y_{t-1}, que en el ejemplo es la palabra *"heart"*).

- Este *embedding* se utiliza junto con las representaciones visuales procesadas para crear una respuesta textual apropiada.

5. ***Transformer Decoder Layer*** **(Capa de Decodificación del Transformer)**:

- El decodificador procesa tanto la información obtenida del codificador (r_{x_s}) como el *embedding* de salida de la palabra anterior ($r_{y_{t-1}}$).

- Este proceso permite generar la secuencia de salida, en este caso, una descripción textual basada en la imagen. Varias capas de decodificación son aplicadas de manera similar al proceso de codificación.

6. ***Softmax Layer***:

- Finalmente, el vector resultante pasa por una capa linear y luego por una capa *Softmax* para producir una distribución de probabilidad sobre el vocabulario. Esto permite que el modelo seleccione la palabra más probable como salida en cada paso de la secuencia de generación de texto. En este caso, la palabra generada sería *"is"* (mostrada en la parte superior derecha).

Otros proyectos relacionados con los previamente mencionados son los siguientes:

Pan y cols. introdujeron la fusión de características cross-modal multiescala para mejorar la sensibilidad de localización y la caracterización de enfermedades en la generación de informes.[25]

Yang y cols. presentaron un marco para la actualización automática del conocimiento médico utilizando la alineación multimodal, mejorando la precisión de los informes de radiología.[35]

2.1.5. Modelos Vision Encoder-Decoder (VED)

Los modelos Transformer[32] y los mecanismos de atención han revolucionado no solo el campo de la inteligencia artificial sino que la generación automática de informes médicos también, especialmente en campos complejos como la generación de informes de radiología (RRG). Una de las principales ventajas de los Transformers es su capacidad para procesar y generar secuencias de texto de manera eficiente y contextualizada, permitiendo que el modelo tenga en cuenta no solo la información inmediata, sino también el contexto global de la secuencia. Esto es particularmente útil en RRG, donde la coherencia y la precisión del informe son críticas, ya que los errores pueden tener consecuencias graves para el diagnóstico y el tratamiento del paciente.

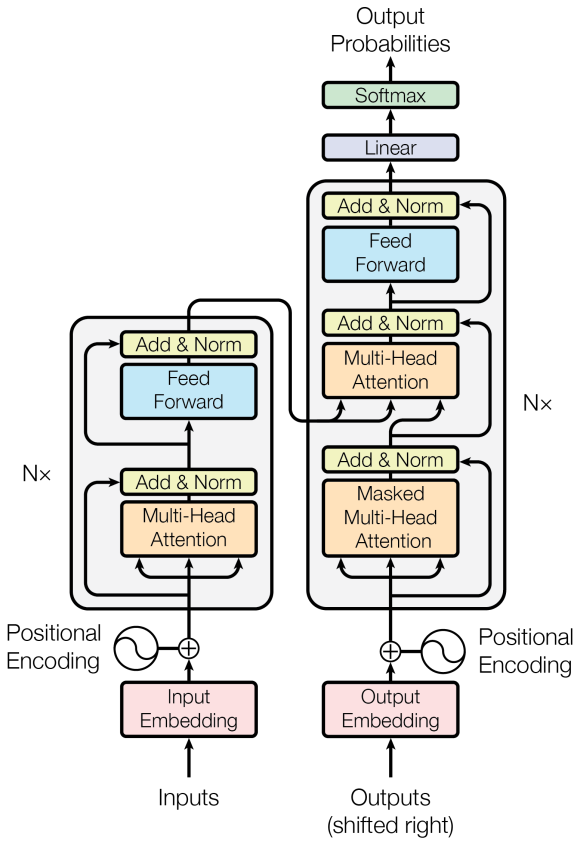


Figura 2.5: Arquitectura transformer.[32]

El mecanismo de atención, especialmente *Multi-Head Attention*, permite que el modelo enfoque su atención en diferentes partes de la imagen de rayos X, destacando las áreas más relevantes para el diagnóstico. Esto ayuda a que el modelo genere descripciones más precisas y detalladas de las anomalías observadas. Además, la capacidad de estos modelos para integrar conocimientos médicos específicos y generales permite una personalización más detallada del informe, adaptándolo mejor a las necesidades clínicas específicas.

Los Transformers también son especialmente útiles en tareas secuencia a secuencia (seq-to-seq)[8], un tipo de arquitectura que se emplea cuando se desea transformar una secuencia de entrada, como imágenes, en una secuencia de salida, como texto. Son ideales cuando los datos de entrada y salida del modelo son diferentes y son aplicables a una amplia gama de aplicaciones, como la traducción automática, el resumen de textos y la generación de texto a partir de imágenes entre otras. En el caso de la generación de informes radiológicos, el objetivo es procesar imágenes médicas y generar texto coherente que describa los hallazgos y el diagnóstico. Esta capacidad de los modelos seq-to-seq es crucial en el RRG porque nos permite traducir los detalles visuales de una imagen médica en descripciones textuales precisas, cubriendo cada aspecto importante del diagnóstico.

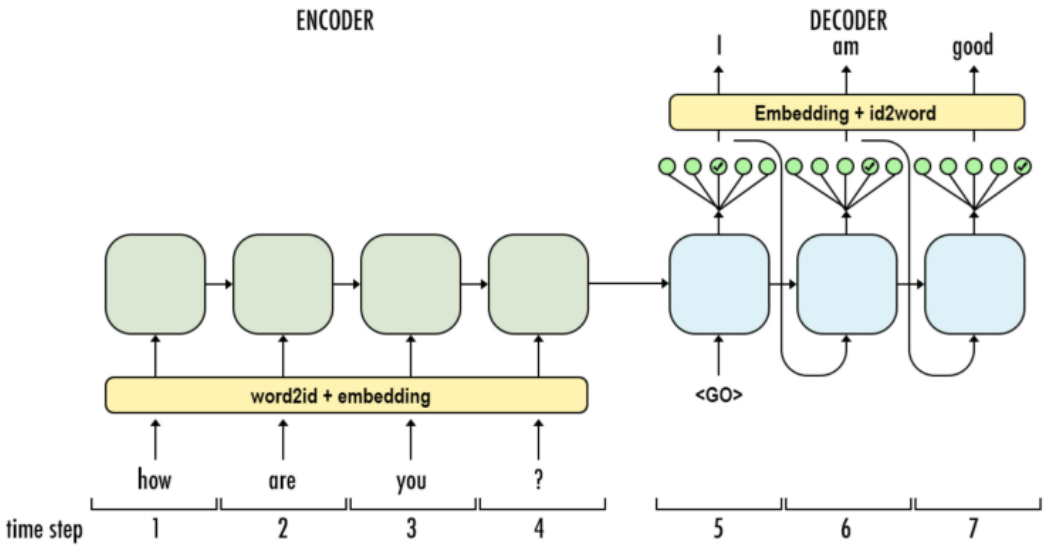


Figura 2.6: Esquema seq-seq.[14]

Además, los Vision Encoder Decoder (VED), que son una extensión de los Transformers, aprovechan esta arquitectura seq-to-seq para tareas multimodales como la generación de informes médicos. Los modelos VED utilizan un Transformer como codificador para procesar la información visual de las imágenes (como rayos X), y otro Transformer como decodificador para generar texto en lenguaje natural. Este enfoque ha demostrado ser efectivo para alinear y capturar mejor las relaciones entre las modalidades visuales y textuales, mejorando la coherencia y exactitud de los informes médicos generados automáticamente.

La utilización de Transformers en RRG también facilita la incorporación de información multimodal, combinando datos visuales y textuales de manera más efectiva. Los mecanismos de atención multinivel permiten una alineación más precisa entre las modalidades de imagen y texto, mejorando la coherencia del informe final. Asimismo, los modelos con una memoria mejorada, como los impulsados por Transformers, permiten una retención más robusta de la información relevante durante la generación del informe, lo que contribuye a producir informes más completos y detallados, capturando mejor las complejidades del diagnóstico radiológico.

A continuación se describen algunos de los proyectos que utilizan esta tecnología:

Yang y cols. emplearon un mecanismo de atención multi-cabeza en su modelo, que combinaba conocimiento médico general y específico para mejorar la precisión en la generación de informes de rayos X de tórax.[34]

Zhao y cols. utilizaron un método de alineación multinivel que incluía mecanismos de atención para alinear mejor las modalidades de texto e imagen, mejorando la coherencia de los informes.[37]

Chen y cols. introdujeron un transformer impulsado por memoria que mejoró la retención de información e incorporación en el decodificador, mejorando la profundidad de los informes generados.[4]

Chen y cols. propusieron un modelo VED basado en redes de memoria cross-modal, que capturaba alineaciones visuales-textuales para mejorar la coherencia y precisión de los informes.[3]

El modelo VED de Delbrouck y cols. utilizó DenseNet-121 como el encoder óptico y BERT como el decodificador, optimizado con RL, logrando informes de radiología de mayor calidad.[5]

2.1.6. Grafos de Conocimiento

El uso de grafos de conocimiento en la generación de informes de radiología (RRG) presenta diversas ventajas que pueden ser clave para superar los desafíos asociados a este proceso. En primer lugar, los grafos de conocimiento permiten estructurar de manera eficiente la información médica relevante, lo que facilita la integración de grandes volúmenes de datos provenientes de diferentes fuentes, como imágenes médicas y informes textuales previos. Esto mejora la precisión en la generación de informes al asegurar que la información relevante se relaciona correctamente con las observaciones clínicas pertinentes.

Además, los grafos de conocimiento permiten una mejor interpretación de las imágenes radiológicas, ya que pueden modelar relaciones complejas entre distintas entidades médicas, como diagnósticos, hallazgos y tratamientos. Al capturar estas relaciones de manera explícita, los sistemas de RRG pueden generar informes más coherentes y clínicamente relevantes. Asimismo, los grafos ayudan a mejorar la consistencia en la terminología empleada, algo crucial en el campo médico para evitar ambigüedades.

Otra ventaja importante es que los grafos de conocimiento pueden aprovechar el conocimiento experto ya existente en la literatura médica, codificando reglas y patrones comúnmente utilizados por radiólogos. Esto no solo reduce la variabilidad entre informes generados automáticamente, sino que también puede ayudar a identificar anomalías o posibles errores en el diagnóstico. Finalmente, los grafos de conocimiento facilitan la personalización de los informes según el contexto clínico específico del paciente, lo que puede conducir a recomendaciones más precisas.

Liu y cols. exploraron el uso de grafos de conocimiento no supervisados para replicar los patrones de los radiólogos en la generación de informes, con el objetivo de mejorar la automatización de la generación de informes.[18]

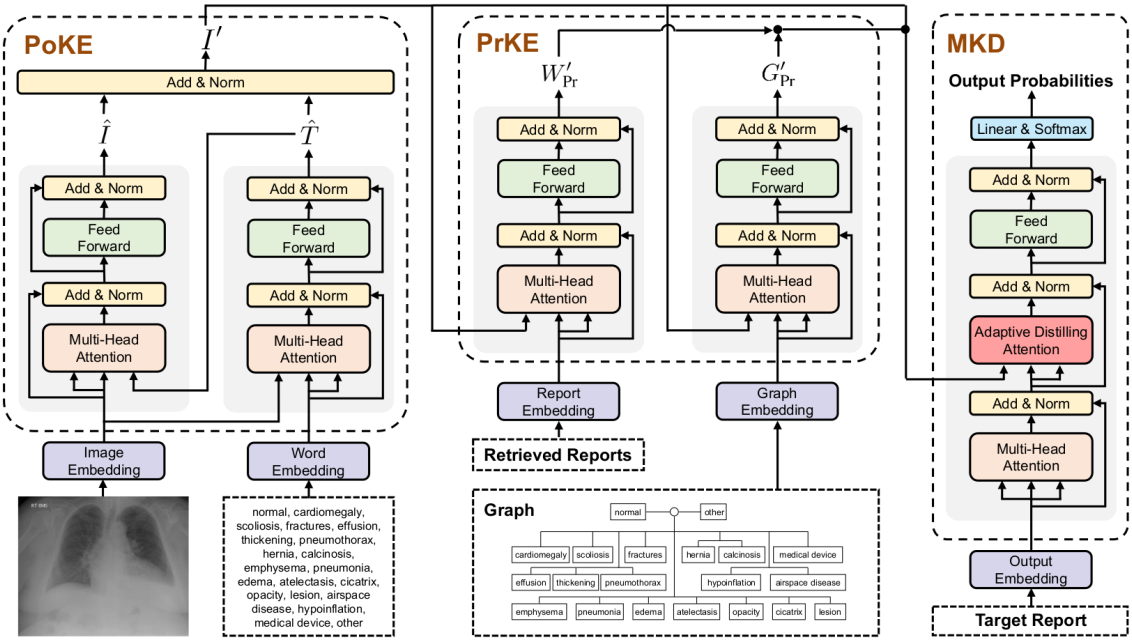


Figura 2.7: Modelo propuesto por Liu y cols.[18]

El modelo PPKED (Posterior-and-Prior Knowledge Exploring-and-Distilling) tiene como objetivo generar informes precisos mediante la exploración y destilación de conocimiento posterior y conocimiento previo. A continuación, se describen sus principales componentes:

- **Posterior Knowledge Explorer (PoKE)**:
  Este módulo explora el conocimiento posterior relacionado con la imagen de entrada, por ejemplo, una radiografía. Para esto, se utilizan representaciones de la imagen (*Image Embedding*) y de las palabras clave (*Word Embedding*) que representan términos médicos (como "cardiomegalia", "neumonía", etc.). Se emplea un mecanismo de atención (*Multi-Head Attention*) para identificar las áreas anormales explícitas en la imagen, mejorando así el conocimiento posterior acerca de posibles patologías.

- **Prior Knowledge Explorer (PrKE)**:
  Este componente explora el conocimiento previo relevante a la imagen de entrada, utilizando informes médicos previamente generados (*Report Embedding*). Además, se basa en un grafo (*Graph Embedding*) que organiza el conocimiento previo en términos de la relación entre las distintas patologías y condiciones médicas, como se muestra en la ilustración. El objetivo es encontrar patrones o información similar en informes previos que pueda ser útil para interpretar la imagen actual.

- **Multi-domain Knowledge Distiller (MKD)**:
  La tarea principal de este módulo es fusionar de manera efectiva el conocimiento posterior (explorado por PoKE) y el conocimiento previo (explorado por PrKE). Para ello, utiliza mecanismos de atención para destilar el conocimiento de manera adaptativa, y finalmente genera las probabilidades de salida que representan el informe médico final (*Target Report*).

Resumen del flujo del modelo

- Primero, PoKE extrae el conocimiento posterior basado en la imagen actual y sugiere posibles patologías o condiciones.

- Luego, PrKE analiza el conocimiento previo relevante, tomando en cuenta informes médicos existentes y sus relaciones.

- Finalmente, MKD destila ambos tipos de conocimiento y los combina de manera óptima para generar un informe médico preciso y contextualizado, utilizando técnicas avanzadas de atención como la ya mencionada *Multi-Head Attention* y una capa final de destilación adaptativa (*Adaptive Distilling Attention*).

Otros proyectos relacionados con los previamente mencionados son los siguientes:

Continuando su investigación, Liu y cols. se enfocaron en mejorar la robustez y precisión de la generación de informes mediante el desarrollo de técnicas más sofisticadas de grafos de conocimiento.[17]

Zhao y cols. incorporaron grafos de conocimiento en su método de alineación multinivel, utilizando el conocimiento médico en forma de diccionario para mejorar la alineación entre los datos textuales y visuales.[37]

2.1.7. Aprendizaje por Refuerzo (RL)

El uso de aprendizaje por refuerzo (RL, por sus siglas en inglés) en la generación de informes de radiología (RRG) ofrece varias ventajas clave. En primer lugar, permite la optimización de métricas específicas del dominio, lo que ayuda a generar informes más precisos y relevantes desde un punto de vista clínico. RL puede aprender a mejorar la coherencia y la exactitud en los informes mediante retroalimentación continua, lo que es especialmente importante en este contexto donde los errores o la falta de información pueden tener implicaciones graves para el diagnóstico. Además, RL permite la adaptación y personalización de los informes en función de criterios previamente definidos por los profesionales médicos, ajustando los textos generados a los estándares y necesidades de la práctica médica. En conjunto, estas características hacen que RL sea una herramienta poderosa para superar las limitaciones de los enfoques tradicionales basados en redes neuronales, mejorando no solo la calidad de los informes generados, sino también la interpretabilidad y utilidad para los médicos.

A continuación, describimos algunos proyectos en los que se han usado RL:

El modelo de Liu y cols. incorporó aprendizaje por refuerzo para el ajuste fino, lo cual fue crucial para mejorar la precisión y calidad de los informes de radiología generados.[19]

Miura y cols. sugirieron el uso de dos métricas optimizadas por RL para garantizar la generación de entidades específicas del dominio y descripciones coherentes, lo que ayudó a mejorar la calidad de los informes.[22]

Delbrouck y cols. aplicaron RL en su modelo VED, optimizándolo con múltiples métricas como RadGraph y BERTScore para mejorar la riqueza semántica de los informes.[5]

2.1.8. Métricas de Evaluación y Calidad de informes

Elegir una buena métrica es crucial para asegurar que los informes generados no solo sean precisos, sino también útiles desde un punto de vista clínico. Las métricas de evaluación juegan un papel fundamental en la mejora continua de los modelos, ya que permiten cuantificar tanto la calidad del lenguaje como la precisión de los hallazgos clínicos presentes en los informes.

En el problema de la generación de informes radiológicos (RRG), los trabajos más recientes buscan que el informe de referencia y el generado transmitan la misma información, incluso si lo hacen en un orden diferente o con distintas expresiones. Por esta razón, métricas clásicas como BLEU por ejemplo se vuelven obsoletas para este tipo de problemas, ya que BLEU se basa en la coincidencia exacta de n-gramas y no captura adecuadamente la equivalencia semántica cuando las frases están reordenadas o reformuladas. Además, no existe una única forma correcta de redactar un informe; las combinaciones lingüísticas son prácticamente infinitas, lo que hace necesario el uso de métricas más avanzadas que evalúen la calidad semántica y factual de los textos generados.

Algunos de los trabajos que tratan este tema son:

Delbrouck y cols. optimizaron su modelo VED utilizando las métricas RadGraph, BERTScore y NLL para mejorar la calidad y precisión de los informes de radiología, aunque reconocieron problemas con la repetitividad de los informes.[5]

Miura y cols. introdujeron métricas específicas del dominio como factENT y factENTNLI junto con BERTScore para refinar la evaluación de la calidad de los informes, enfocándose en la equivalencia semántica y la coherencia de las entidades.[22]

Nicolson y cols. criticaron las métricas de evaluación actuales como BLEU y ROUGE-L por su inadecuación para capturar las complejidades de los informes de patologías, abogando por métodos más sofisticados.[24]

2.2 Resumen general

En este capítulo, se ha presentado un análisis exhaustivo de las técnicas más relevantes en la clasificación y detección de patologías en imágenes médicas, enfocadas en la generación de informes automáticos de radiología (RRG). A lo largo de las diferentes secciones, se han discutido los enfoques basados en redes neuronales profundas, modelos preentrenados y arquitecturas transformers, así como el uso de técnicas multimodales, aprendizaje por refuerzo y grafos de conocimiento.

Las **Redes Neuronales Convolucionales (CNNs)** y las **Redes Neuronales Recurrentes (RNNs)** se han consolidado como pilares en el procesamiento de imágenes médicas y la generación de texto asociado, logrando avances significativos en la precisión y contextualización de los informes radiológicos. Los modelos combinados, como los que integran CNNs con RNNs, permiten un análisis más detallado y coherente de las imágenes, mejorando la anotación y descripción de las patologías.

El uso de **modelos preentrenados**, como las redes basadas en **Transformers** y los modelos GPT, ha demostrado una gran capacidad para la generación de textos coherentes y precisos. Estos modelos, al ser entrenados en grandes volúmenes de datos, logran una comprensión profunda del lenguaje y del contexto médico, facilitando la creación de informes detallados y clínicamente útiles.

Las técnicas **multimodales** y **cross-modal** han permitido integrar información visual y textual de manera más eficiente, mejorando la generación de informes mediante la alineación de características provenientes de diferentes modalidades. Además, el uso de **modelos Vision Encoder-Decoder (VED)** ha demostrado ser efectivo para generar informes que capturan tanto la información visual como la semántica.

El **aprendizaje por refuerzo (RL)** ha sido clave para optimizar la generación de informes ajustando las métricas utilizadas y asegurando la coherencia factual de los textos generados. De igual manera, los **grafos de conocimiento** han permitido estructurar mejor la información médica, integrando conocimiento previo y mejorando la precisión de los diagnósticos.

Finalmente, las **métricas de evaluación** han sido fundamentales para medir la calidad y precisión de los informes generados, destacando la necesidad de desarrollar nuevas métricas que capten mejor la equivalencia semántica y factual de los informes en lugar de depender únicamente de la coincidencia exacta de palabras.

En conjunto, estos avances tecnológicos han permitido una mejora sustancial en la generación automática de informes de radiología, aunque aún existen desafíos por superar, como la adaptación a nuevos conjuntos de datos y la generalización a diferentes dominios médicos.

CAPÍTULO 3
Metodología

En este capítulo se introducirá la metodología utilizada para desarrollar la solución al problema del RRG que hemos comentado. Para ello, se detallará la arquitectura que se empleará en nuestro modelo, abarcando tanto el proceso de entrenamiento como el de inferencia. Asimismo, se describirá el preprocesamiento de los datos necesario antes del entrenamiento para asegurar la calidad y mayor eficacia del modelo.

3.1 Vision Encoder Decoder

Como hemos comentado en el apartado 2.1.5 Modelos Vision Encoder-Decoder (VED), la tecnología transformer ha demostrado un gran rendimiento en múltiples tareas, por lo que hemos decidido implementar un modelo de Vision Encoder Decoder (VED), similar al utilizado por Parres y cols. [27]. Nuestro enfoque utiliza el modelo Swin como encoder de visión y BERT como decoder de lenguaje. Este tipo de arquitectura ha sido diseñada específicamente para abordar el problema de generación de informes de rayos X de tórax.

3.1.1. Vision Encoder

El Swin Transformer y los Vision Transformers tradicionales (ViT) se comparan en términos de cómo extraen los vectores de características de una imagen y su eficiencia computacional. El Swin Transformer, como se muestra en la Figura 3.1 (a), divide una imagen en parches y aplica autoatención solo dentro de ventanas locales en las capas iniciales. A medida que el modelo avanza en capas más profundas, fusiona estos parches, construyendo una representación jerárquica de la imagen. Esta estrategia permite capturar tanto detalles locales como globales con una complejidad computacional lineal en función del tamaño de la imagen. Al limitar la autoatención a ventanas locales, el Swin Transformer resulta más eficiente y escalable para tareas de clasificación y reconocimiento.

En contraste, los Vision Transformers tradicionales (Figura 3.1 (b)) dividen la imagen en parches pero aplican autoatención a nivel global, calculando las relaciones entre todos los parches. Esto genera una complejidad cuadrática en función del tamaño de la imagen, lo que los hace más costosos computacionalmente. Aunque los ViT pueden capturar dependencias globales de manera precisa, su enfoque no es tan eficiente, especialmente para imágenes grandes o tareas más complejas. La capacidad del Swin Transformer para mantener una estructura jerárquica y reducir la carga computacional lo hace más adecuado para una variedad de aplicaciones visuales.

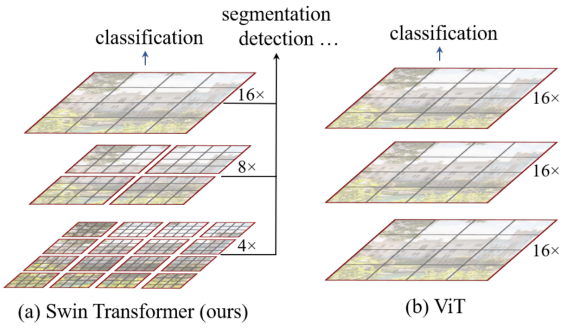


Figura 3.1: Diferencia Swin y ViT[21].

Dicho esto, el encoder de visión que hemos seleccionado es Swin, el cual ha mostrado un rendimiento sobresaliente en diversas tareas de visión por computador como la clasificación, la detección de objetos y la segmentación [21]. A diferencia de ViT [7], Swin emplea un mecanismo de atención más eficiente basado en ventanas desplazadas, lo que lo hace más adecuado para el análisis de imágenes de rayos X. Hemos optado por la variante base de esta arquitectura, conocida como SwinBase, preentrenada con el conjunto de datos ImageNet [6].

SwinBase trabaja con imágenes de entrada de 3 canales y un tamaño de 224x224 píxeles. La profundidad de las capas arquitectónicas es de 2, 2, 18 y 2, con un tamaño de parche de 4, y utiliza 4, 8, 16 y 32 cabezas de atención.

3.1.2. Decoder

Para el decoder de lenguaje, hemos elegido BERT debido a su eficacia comprobada en diversas tareas de procesamiento de lenguaje natural (NLP) [15, 20].

Chen y cols. utilizaron modelos de lenguaje preentrenados(BERT) para mejorar la generación de texto, y los experimentos muestran que su modelo supera a otros modelos robustos [2].

Nos hemos desviado de la configuración estándar, utilizando solo tres capas con un tamaño oculto de 1024, en lugar de las 12 capas originales de BERT. Para nuestro modelo, hemos optado por un decodificador basado en palabras con un vocabulario de aproximadamente 9.8k palabras.

La integración entre el modelo de visión y el decoder se realiza a través de capas de *cross-attention*, lo que permite la introducción de características visuales extraídas de las radiografías en el decoder. Este mecanismo posibilita la generación autoregresiva de informes médicos.

Siguiendo este enfoque, el modelo VED propuesto para el análisis de rayos X de tórax es SwinBase+BERT9k.

3.2 Entrenamiento

En el contexto radiológico, es común generar múltiples imágenes por estudio de paciente (como vistas anteroposterior y lateral). Empleamos un enfoque de imágenes múltiples, restringido a tres imágenes por estudio. Esto involucra concatenar características extraídas por el codificador Swin para cada imagen y procesarlas en el decodificador. Además, aplicamos transformaciones aleatorias, como traducción, escalado, rotación y

ajustes de brillo y contraste, para aplicar un *Image Augmentation* las imágenes de entrenamiento.

Nuestro proceso de entrenamiento consta de dos etapas, como se muestra en la Figura 3.2. En la etapa inicial, se emplea la función de pérdida NLL (Negative Log-Likelihood), utilizando "teacher forcing" para entrenar el modelo basándose en el informe generado y el informe de referencia. Esta etapa típicamente abarca aproximadamente veinte *epochs* de entrenamiento. Posteriormente, en la segunda etapa, se introduce RL (Reinforcement Learning) usando el algoritmo SCST (Self-Critical Sequence Training). Esta segunda fase conlleva aproximadamente quince *epochs* de entrenamiento.

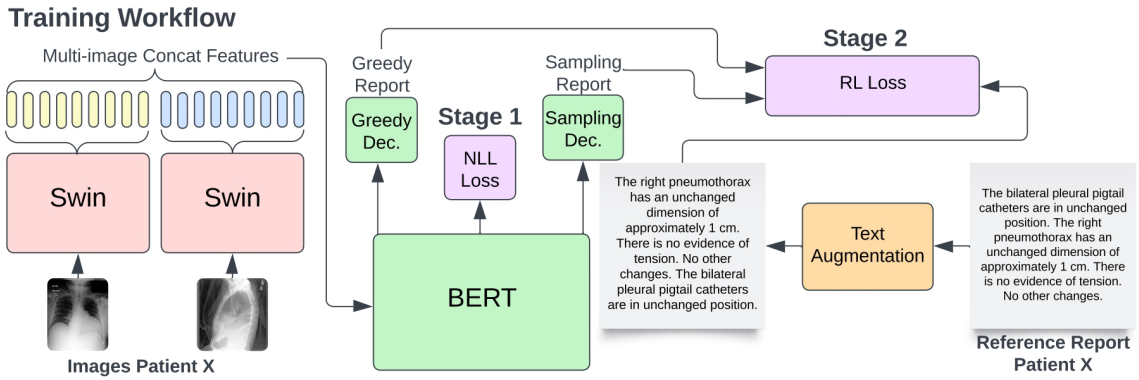


Figura 3.2: Flujo de entrenamiento del modelo propuesto.[27]

3.2.1. NLL

En la primera etapa de entrenamiento, se emplea la pérdida NLL (*Negative Log-Likelihood*), como se muestra en el flujo de trabajo en la Figura 3.2. Utilizamos "*teacher forcing*"para guiar el modelo en la generación de informes, comparándolos con los informes originales. Esto ayuda a que el modelo aprenda a predecir informes más precisos desde el principio del entrenamiento.

El *teacher forcing* es una técnica utilizada en el entrenamiento de modelos secuenciales, donde el valor verdadero de la secuencia esperada se introduce como entrada en lugar de utilizar la predicción generada por el modelo en la iteración anterior. Este enfoque permite acelerar el proceso de convergencia al evitar errores acumulativos durante las primeras etapas de entrenamiento y obliga al modelo a aprender a imitar correctamente las secuencias objetivo.

Sin embargo, con este enfoque no conseguimos resultados lo suficientemente buenos, ya que el modelo aún tiende a generar informes menos precisos, el modelo aprende a generar las palabras que más frecuentemente aparecen en los informes. Por lo tanto, planteamos una segunda fase de entrenamiento utilizando aprendizaje por refuerzo (RL) para mejorar aún más la calidad de los informes.

3.2.2. RL

El algoritmo Self-Critical Sequence Training (SCST) es un enfoque de aprendizaje por refuerzo para secuencias, como la generación de texto, que se basa en utilizar la propia salida del modelo como una referencia para calcular las recompensas. En el contexto de generación de informes, SCST permite comparar dos secuencias: una generada de forma "*greedy*"(Elige siempre la opción más probable) y otra generada por "*sampling*"(Elige con algo de aleatoriedad). La diferencia del *reward* entre estas dos secuencias impulsa

la optimización de la política del modelo. SCST es útil porque reduce la varianza de las estimaciones de gradientes, lo que resulta en una convergencia más estable y efectiva.

Las métricas que empleamos en nuestro enfoque son:

- BERTScore: Esta métrica compara la similitud entre representaciones semánticas de las frases, utilizando un modelo BERT preentrenado [36]. Es excelente para evaluar tanto la gramática como la coherencia semántica en textos generados, ya que captura relaciones contextuales profundas que las métricas tradicionales basadas en coincidencia de palabras no pueden.
  - Ventajas: Evalúa la calidad de la generación desde una perspectiva contextual y semántica, lo que permite una evaluación más matizada y realista de la calidad del informe en comparación con métricas basadas en la coincidencia superficial de palabras, como BLEU o ROUGE.
- RadGraph F1 (RGER): Esta métrica se centra en la precisión y exhaustividad en la detección de entidades y relaciones específicas de informes radiológicos, especialmente aquellas relacionadas con patologías [10]. Utiliza el gráfico de relaciones RadGraph para capturar la estructura subyacente de las entidades clínicas y sus relaciones, midiendo qué tan bien el modelo detecta y organiza esta información crucial.
  - Ventajas: Prioriza el reconocimiento correcto de entidades médicas relevantes y sus relaciones en el texto generado, lo cual es fundamental para la interpretación clínica correcta de los informes radiológicos. Esto asegura que el modelo no solo genere texto coherente, sino que también capture la información médica más importante.

Estas métricas son particularmente útiles para la generación de informes médicos porque no solo evalúan la calidad del lenguaje, sino también la precisión médica, lo que es fundamental para garantizar que el informe sea tanto claro como clínicamente útil.

La fase 2 introduce el aprendizaje por refuerzo (RL), que utiliza el algoritmo SCST para mejorar los informes generados. Esto es necesario porque, en el primer paso de RL, la decodificación "*greedy*" genera un informe (Yg) sin calcular gradientes. En el segundo paso, se utiliza la decodificación con "*sampling*" para generar el informe (Ys), sobre el cual se calculan los gradientes. Luego se optimizan las métricas propuestas, lo que permite mejorar la calidad de los informes generados por el modelo.

Para calcular la *loss* de la métrica (Lossmetric), se comparan las recompensas obtenidas para los informes de "*sampling*" y "*greedy*" (Ys y Yg) con respecto al informe de referencia (Yref), como se muestra en la ecuación 3.1:

$$Loss_{metric}(Y_s, Y_g, Y_{ref}) = -(r_{metric}(Y_s, Y_{ref}) - r_{metric}(Y_g, Y_{ref}))\log(Pr(Y_s)) \tag{3.1}$$

Cada Lossmetric contribuye como un término ponderado distinto a la pérdida final de RL, que se presenta en la ecuación 3.2. La pérdida final de RL se calcula utilizando las métricas BERTScore y F1 RGER, además de la pérdida NLL.

$$Loss_{RL} = \alpha Loss_{BERTScore} + \beta Loss_{F1RGER} + \gamma Loss_{NLL} \tag{3.2}$$

En nuestros experimentos, los factores de ponderación se establecieron en $\alpha = \beta =$ 0,495 y $\gamma = 0{,}010$.

3.3 Inferencia

El proceso de inferencia, como se muestra en la Figura 3.3, implica concatenar las características extraídas de múltiples imágenes del paciente usando el codificador Swin. Estas características se pasan luego al modelo BERT entrenado previamente, que genera el informe estimado utilizando decodificación por "*beam search*".

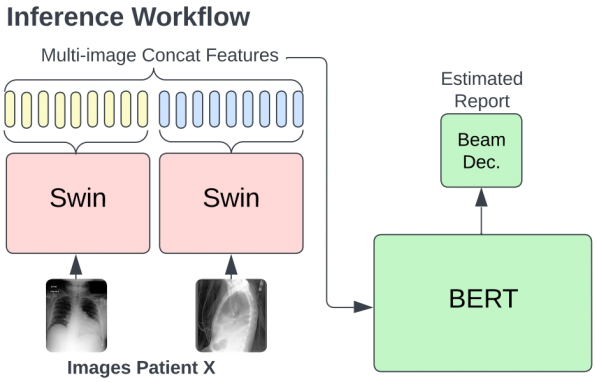


Figura 3.3: Inferencia del modelo propuesto.[27]

Beam search es un método de búsqueda que genera varias secuencias posibles y selecciona las más probables en cada paso, evaluando múltiples opciones simultáneamente. Esto mejora la calidad del informe generado al comparar y elegir las mejores secuencias, en lugar de optar por una única opción de forma determinista.

Este enfoque permite generar informes precisos a partir de múltiples imágenes, mejorando la exactitud del diagnóstico basado en imágenes radiológicas.

3.4 Preprocesamiento de datos

El objetivo del preprocesamiento de datos es preparar la información de entrada para los modelos de aprendizaje automático, mejorando su calidad, relevancia y formato. Este paso es crucial para asegurar que los modelos puedan generalizar mejor y producir resultados más precisos, evitando problemas como el sobreajuste y la falta de diversidad en los datos.

En el contexto de la generación de informes médicos, el preprocesamiento incluye técnicas de aumento de datos para aumentar la variabilidad de los textos disponibles, generando informes más diversos sin perder la coherencia diagnóstica. Una de estas técnicas es la reorganización de frases, que analizamos a continuación.

3.4.1. Reordenamiento de las frase

El reordenamiento de frases es una técnica de aumento de datos (*text augmentation*) utilizada para mejorar la diversidad de textos disponibles en un conjunto de datos. Esta técnica consiste en dividir un informe de referencia en frases y reorganizarlas aleatoriamente, manteniendo el diagnóstico original intacto. El *data augmentation* se ha utilizado en el campo del procesamiento de imágenes, pero su aplicación a informes médicos es más reciente.

Como se menciona en el estudio de [27], al aplicar esta técnica en informes radiológicos de tórax, se logró reducir la monotonía en los informes generados por los algoritmos de generación de informes radiológicos, además de mejorar la exactitud y calidad de los mismos. Este enfoque no solo previene el sobreajuste, sino que crea un modelo más generalista que predice informes de mayor calidad, garantizando que los resultados sean aplicables a una variedad más amplia de estudios clínicos y situaciones radiológicas diversas.

3.4.2. Text Augmentation con la API de OpenAI

En este trabajo, proponemos una nueva forma de aumento de datos textual que consiste en solicitar a la API de OpenAI el parafraseo de los informes.

Para seleccionar el modelo, calculamos cuántos *tokens* de entrada íbamos a tener y obtenemos una media de 130 *tokens*.

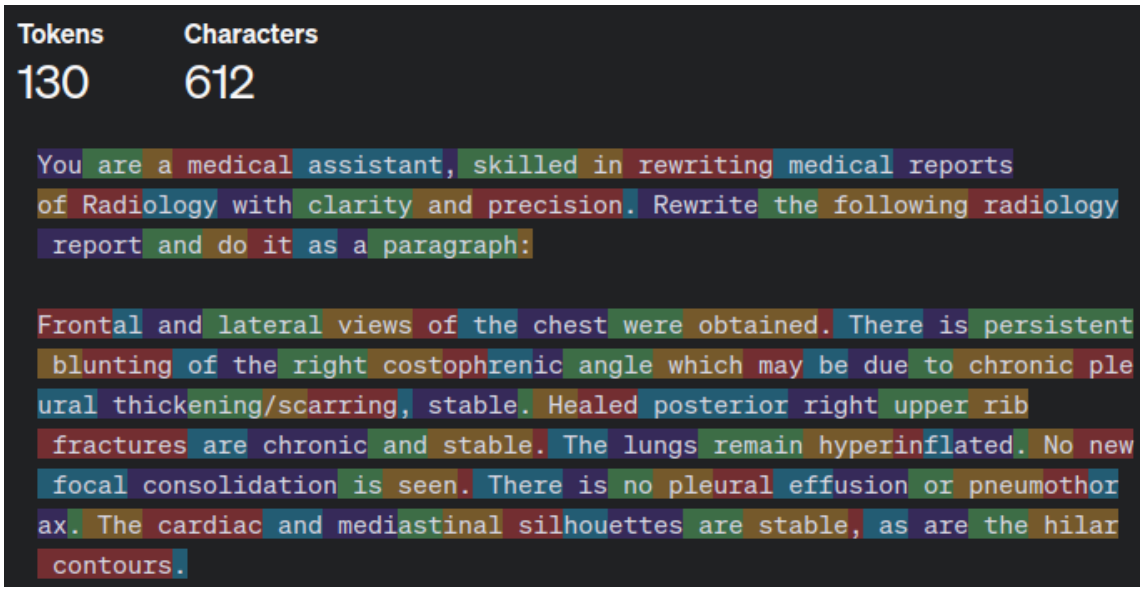


Figura 3.4: Ejemplo de input tokenizado.

Del mismo modo, calculamos cuántos *tokens* de salida íbamos a generar, con una media de 92 *tokens* por informe.

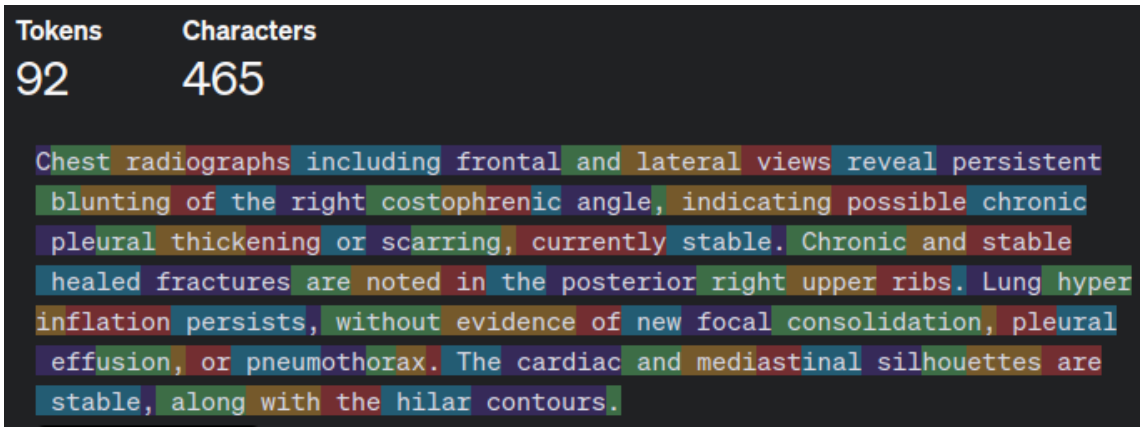


Figura 3.5: Ejemplo de Output tokenizado.

Sabiendo que los precios de los modelos por millón de *tokens* son los siguientes:

- **GPT-4o**: 5.00$ / 1M tokens (entrada), 15.00$ / 1M tokens (salida)
- **GPT-4**: 10.00$ / 1M tokens (entrada), 30.00$ / 1M tokens (salida)
- **GPT-3.5 Turbo**: 0.5$ / 1M tokens (entrada), 1.5$ / 1M tokens (salida)

Y que tenemos un total de 152,173 informes, calculamos los costos estimados para cada modelo. Los cálculos se realizan de la siguiente manera:

$$\text{Costo Total} = \frac{\text{tokens de entrada} \times 152,173}{1,000,000} \times \text{costo entrada} + \frac{\text{tokens de salida} \times 152,173}{1,000,000} \times \text{costo salida}$$

Los costos estimados por modelo son los siguientes:

Modelo	**Costo por Token**		**Costo Total**
	Costo Entrada	**Costo Salida**	**Total ($)**
GPT-4o	5.00$ ×19,78	15.00$ ×13,99	308.75
GPT-4	10.00$ ×19,78	30.00$ ×13,99	617.50
GPT-3.5 Turbo	0.5$ ×19,78	1.5$ ×13,99	30.88

Tabla 3.1: Cálculo de costos por modelo para 152,173 informes

Dado que el GPT-3.5 Turbo era considerablemente más económico que los otros modelos y ofrecía resultados suficientemente buenos para nuestro proyecto y similares al resto de modelos, elegimos este modelo. Su capacidad para manejar grandes secuencias de texto y generar salidas coherentes con variabilidad lo hacen ideal para nuestro proyecto. Además, controlamos la longitud del informe reescrito mediante el prompt, asegurándonos de que los textos generados no sean excesivamente largos ni cortos, pero mantengan el diagnóstico adecuado.

Los prompts utilizados fueron los siguientes:

System prompt: "Eres un asistente médico, especializado en reescribir informes de radiología con claridad y precisión."

User prompt: "Reescribe el siguiente informe de radiología y hazlo en un solo párrafo: *informe médico.*"

Este prompt fue utilizado dentro de la opción *chat* del modelo, el *script* de python usado se puede encontrar en el apéndice A. Este script se utilizó para recorrer todos los informes del conjunto de datos original, presentes en un archivo CSV, y generar un nuevo CSV con los informes reescritos por el modelo GPT-3.5 Turbo.

A continuación se muestran un ejemplo de un informe antes y después de aplicar la técnica de *text augmentation*:

Texto original:

"Moderately well inflated lungs with improvement in patchy opacities noted in the left lower lobe with new linear and patchy opacities in the right lower lobe likely subsegmental atelectasis. Improvement in bilateral layering pleural effusions. Cardiomediastinal silhouette appears normal. Right PICC terminates slightly beyond the cavoatrial junction, as before. ET tube tip terminates 5 cm above the carina in the midtrachea. EKG leads overlie the chest wall. Enteric tube traverses below the diaphragm, distal tip not visualized. Visualized bones are unremarkable."

Figura 3.6: Informe antes del DA con gpt-3.5

Texto tras el text augmentation:

"The lungs are moderately well inflated with a noticeable improvement in patchy opacities in the left lower lobe, while new linear and patchy opacities in the right lower lobe are suggestive of subsegmental atelectasis. Additionally, there has been improvement in bilateral layering pleural effusions. The cardiomediastinal silhouette appears normal. The right PICC terminates slightly beyond the cavoatrial junction, as previously noted. The ET tube tip ends 5 cm above the carina in the midtrachea, and EKG leads are observed overlying the chest wall. The enteric tube passes below the diaphragm, with the distal tip not visualized. Finally, the visualized bones show no abnormalities."

Figura 3.7: Informe después del DA con gpt-3.5

Este tipo de parafraseo introduce variabilidad en los datos sin modificar la información crítica del diagnóstico. Esta técnica de *text augmentation* es más completa que la anterior, en la que simplemente reordenábamos frases. En ese enfoque previo, el informe podría tener un orden incorrecto, lo que afectaría la coherencia. Sin embargo, en esta nueva técnica, al reescribir el texto, se mantiene un orden lógico y se utilizan sinónimos, lo cual no solo preserva la coherencia, sino que también permite que el modelo generalice mejor al enfrentarse a distintas redacciones del mismo contenido diagnóstico.

CAPÍTULO 4
Experimentos y Resultados

Dado que estamos entrenando un modelo de gran envergadura, compuesto por 147 millones de parámetros, con un *dataset* de tamaño considerable (1.3.1 MIMIC-CXR), el tiempo de entrenamiento es prolongado. La primera fase de entrenamiento (NLL) requiere aproximadamente dos días y medio, mientras que la segunda fase (RL) toma alrededor de ocho días.

Se llevaron a cabo varios experimentos para evaluar las diferentes técnicas de *text augmentation* mencionadas en el apartado 3.4 Preprocesamiento de datos. En estos experimentos, variamos los porcentajes de *text augmentation*, es decir, si incluimos un 20 % de texto modificado a través de la API de OpenAI, el otro 80 % corresponde a textos originales. Se realizaron diversas pruebas con el fin de encontrar un balance óptimo.

Para incluir los informes generados por OpenAI, primero se intentó modificar el código para que, durante cada *batch*, según el porcentaje deseado, se cargara el informe ya sea desde el archivo CSV del *dataset* original o desde el archivo CSV que contiene los informes reescritos por el modelo de OpenAI. Sin embargo, los resultados fueron inferiores al modelo anterior con informes de frases reordenadas, probablemente debido a que en cada fase de entrenamiento se seleccionaban frases diferentes. Es decir, un mismo informe seleccionado del CSV original en la primera fase podía ser seleccionado del CSV reescrito por OpenAI en la segunda fase.

Para evitar esta inconsistencia y garantizar que siempre se traten los mismos informes, sin importar en qué fase de entrenamiento nos encontremos, se decidió crear un CSV previo al entrenamiento que combine los dos CSV existentes según el porcentaje de *text augmentation* deseado.

Experimento	**% OpenAI**	**Método de Incorporación**	F_1cXb **Fase 1 (NLL)**
Exp 1	40 %	Cargar por batch	0.53
Exp 2	40 %	CSV combinado	0.59

Tabla 4.1: Resultados de los diferentes experimentos sobre la incorporación de informes generados en el entrenamiento del modelo.

Debido al extenso tiempo de entrenamiento requerido para el modelo completo, inicialmente se lanzaron varios experimentos y se entrenaron distintas configuraciones solo en la fase 1, para identificar cuáles eran las más prometedoras antes de continuar con la fase 2, que exige un mayor tiempo de entrenamiento. La métrica utilizada es CheXbert[31] (F1cXb), que compara los informes generados y los originales mediante un sistema de etiquetado para evaluar si transmiten la misma información. A continuación, se presentan las pruebas realizadas:

Experimento	% OpenAI	Reordenar Frases	F_1cXb	
			Fase 1 (NLL)	**Fase 2 (RL)**
Baseline	***0 %***	***No***	***0.448***	***0.622***
Exp 3	10 %	No	0.584	—
Exp 4	20 %	No	0.592	—
Exp 5	30 %	No	0.591	0.651
Exp 6	40 %	No	0.593	0.651
Exp 7	40 %	Sí	0.544	—

Tabla 4.2: Resultados de los experimentos realizados.

Debido al tiempo requerido para el entrenamiento y al hecho de que el rendimiento del modelo en la fase 1 no mostraba mejoras significativas después del cuarto experimento, estos son los resultados obtenidos hasta el momento. Aunque la diferencia entre los experimentos 4, 5 y 6 era mínima, como el experimento 6 mostraba el mejor rendimiento en la fase 1, se lanzó la fase 2 de entrenamiento para el mismo modelo.

Posteriormente, se lanzó la fase 2 para el experimento 5, y en ambos casos se obtenía el mismo resultado (0.651). No obstante, continuamos entrenando más modelos y realizando pruebas adicionales para mejorar los resultados.

En el Experimento 7 se probó aplicar ambas técnicas de *text augmentation* mencionadas. La técnica de reordenar las frases se aplicó a todos los informes en cada *batch*, pero esto resultó en un modelo de peor rendimiento.

Hasta el momento, el experimento 6 es el que ha proporcionado los mejores resultados en ambas fases de entrenamiento, por lo que se procederá a compararlo con los modelos del estado del arte actual.

Estado del arte	**Orientado a Chest-RRG**		**Orientado a NLG**	
	F_1cXb	F_1RGER	BLEU4	ROUGE-L
MIMIC-CXR: Modelos NLL				
Yang y cols. [34]	-	-	11.5	28.4
Pan y cols. [25]	-	-	11.2	**28.8**
Yang y cols. [35]	-	-	11.1	27.4
Zhao y cols. [37]	-	-	10.9	27.5
Nicolson y cols. [24]	-	-	**12.7**	28.6
Liu y cols. [19]	29.2	-	-	-
Chen y cols. [4]	34.6	-	8.6	27.7
Miura y cols. [22]	44.7	-	10.5	27.7
Chen y cols. [3]	40.5	-	10.6	27.8
Delbrouck y cols. [5]	44.8	20.2	10.5	25.3
Parres y cols. [27]	54.8	18.1	11.5	27.2
Nuestro modelo	**59.3**	**25.4**	10.5	24.8
MIMIC-CXR: Modelos RL				
Miura y cols. [22] (BERTScr+fact$_{ENT}$)	56.7	-	11.1	27.1
Miura y cols. [22] (BERTScr+fact$_{ENTNLI}$)	56.7	-	11.4	**27.1**
Delbrouck y cols. [5] (BERTScr+F_1RGER)	62.2	34.7	11.4	26.5
Parres y cols. [27] (BERTScr+F_1RGER)	62.8	36.1	8.4	25.6
Nuestro modelo (BERTScr+F_1RGER)	**65.1**	**36.9**	**12.6**	26.3

Tabla 4.3: Comparación de los modelos del estado del arte para el conjunto de datos MIMIC-CXR

Nuestro modelo demuestra ser altamente competitivo en comparación con el estado del arte actual en la generación de informes de radiología en la primera fase de entrenamiento (NLL). En particular, destaca como el mejor en las métricas F_1cXb y F_1RGER, superando a otros modelos con puntajes de 59.3 y 25.4, respectivamente. Esto evidencia su capacidad para identificar correctamente anomalías en las imágenes de tórax y generar informes coherentes. Estos resultados son muy satisfactorios, ya que somos el mejor modelo en ambas métricas clave, siendo estas especialmente significativas para nosotros, pues comparan el texto por etiquetas y asegura que el informe generado contenga semánticamente toda la información original.

En cuanto a la optimización por refuerzo (RL), nuestro modelo se sitúa en la primera posición en varias métricas, incluyendo F_1cXb con un puntaje de 65.1, F_1RGER con 36.9 y BLEU4 con 12.6. Aunque BLEU4 es una métrica más tradicional orientada a la generación de lenguaje natural, nuestro rendimiento superior en esta métrica reafirma la capacidad del modelo para generar informes con alta calidad textual, lo que refuerza la solidez de nuestra solución tanto en el reconocimiento clínico como en la generación de lenguaje natural.

CAPÍTULO 5
Conclusiones Y Trabajos Futuros

En este trabajo, hemos logrado desarrollar un modelo competitivo para la generación de informes radiológicos de tórax utilizando una arquitectura basada en Vision Encoder Decoder, con SwinBase como encoder de visión y BERT como decoder de lenguaje. Los resultados obtenidos hasta ahora son muy satisfactorios, destacando especialmente en las métricas F_1cXb y F_1RGER, donde superamos al estado del arte con puntajes de 59.3 y 25.4, respectivamente. Este hecho evidencia la capacidad del modelo para identificar correctamente anomalías en las imágenes de tórax y generar informes médicos coherentes y precisos.

En cuanto a la fase de aprendizaje por refuerzo (RL), nuestro modelo ha demostrado ser altamente competitivo, logrando la primera posición en las métricas F_1cXb, F_1RGER, y BLEU4, con puntuaciones de 65.1, 36.9 y 12.6, respectivamente. Estos resultados confirman la solidez de nuestro enfoque tanto en la identificación clínica como en la generación de lenguaje natural de alta calidad, destacándonos como el mejor modelo en ambas fases.

Uno de los hallazgos más significativos fue en relación con las técnicas de aumento de datos (*text augmentation*). Inicialmente probamos el reordenamiento de frases, pero descubrimos que el parafraseo utilizando la API de OpenAI con el modelo GPT-3.5-Turbo resultó ser más efectivo. Esto se debe a que el parafraseo introduce variaciones sutiles en el texto sin comprometer el contenido esencial, mejorando la diversidad de los datos y la capacidad del modelo para generalizar. Aunque al principio temíamos que el parafraseo pudiera alterar demasiado el texto, los resultados muestran que mantiene la coherencia y contribuye a mejorar la precisión del diagnóstico. Ajustar los *prompts* para asegurar que las variaciones sean controladas ha sido clave para maximizar su efectividad.

5.1 Trabajos Futuros

Existen diversas líneas de trabajo que todavía deben ser exploradas para seguir mejorando el rendimiento de nuestro modelo:

- **Finalizar la Fase 2 de los experimentos pendientes:** Algunos experimentos aún están en fase de NLL, y completar la fase de RL podría ofrecer mejoras adicionales. También se deben lanzar nuevos experimentos para evaluar diferentes configuraciones y técnicas de *text augmentation*.
- **Explorar el uso de otros modelos LLMs para *Text Augmentation*:** Además de GPT-3.5-Turbo, sería interesante probar otros modelos de lenguaje como LLaMA3 o modelos *open-source*. Estos modelos podrían ofrecer un mejor balance entre calidad de parafraseo y coherencia del informe.

- **Ajustar las ponderaciones de las pérdidas en la fase de RL:** Realizar pruebas variando los factores de ponderación en la función de pérdida (NLL, BERTScore, F1 RGER) podría mejorar la calidad de los informes generados en la fase de aprendizaje por refuerzo.
- **Evaluación en otros conjuntos de datos:** Aunque MIMIC-CXR es el conjunto de datos principal para este problema, sería importante evaluar cómo generaliza nuestro modelo en otros conjuntos de datos similares. Existen otros *datasets* en el ámbito de la radiología que podrían proporcionar una mejor visión del rendimiento de nuestro modelo en diferentes contextos y tipos de imágenes.
- **Optimización de la inferencia:** El método de búsqueda por *beam search* utilizado durante la inferencia ha mostrado buenos resultados, pero exploraremos técnicas alternativas, como *nucleus sampling*, que podrían generar informes más variados y precisos.
- **Optimización del tiempo de entrenamiento:** Dado que el tiempo de entrenamiento es considerablemente largo (especialmente en la fase de RL), explorar estrategias de optimización o preentrenamiento con otros *datasets* podría reducir el tiempo necesario para experimentar y ajustar el modelo.

En resumen, aunque hemos conseguido resultados competitivos y satisfactorios, estamos convencidos de que todavía podemos mejorar el modelo. Los futuros experimentos y optimizaciones tienen el potencial de consolidar aún más nuestra solución y hacerla más eficiente, generalista y precisa. Esto no solo será beneficioso para la generación automática de informes radiológicos, sino que también contribuirá al campo de la inteligencia artificial en la salud, proporcionando herramientas que faciliten el diagnóstico médico y mejoren la atención clínica.

Bibliografía

[1] Omar Alfarghaly et al. «Automated Radiology Report Generation Using Conditioned Transformers». En: *Informatics in Medicine Unlocked* 24 (2021), pág. 100557. DOI: 10.1016/j.imu.2021.100557.

[2] Yen-Chun Chen et al. *Distilling Knowledge Learned in BERT for Text Generation*. 2020. arXiv: 1911.03829 [cs.CL]. URL: https://arxiv.org/abs/1911.03829.

[3] Zhihong Chen et al. *Cross-modal Memory Networks for Radiology Report Generation*. 2022. arXiv: 2204.13258 [cs.CL]. URL: https://arxiv.org/abs/2204.13258.

[4] Zhihong Chen et al. *Generating Radiology Reports via Memory-driven Transformer*. 2022. arXiv: 2010.16056 [cs.CL]. URL: https://arxiv.org/abs/2010.16056.

[5] Jean-Benoit Delbrouck et al. *Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards*. 2022. arXiv: 2210.12186 [cs.CL]. URL: https://arxiv.org/abs/2210.12186.

[6] Jia Deng et al. «ImageNet: A large-scale hierarchical image database». En: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, págs. 248-255. DOI: 10.1109/CVPR.2009.5206848.

[7] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[8] Everton Gomede. *Understanding Sequence to Sequence (Seq2Seq) Models and Their Significance*. Accessed: 2024-09-04. 2023. URL: https://medium.com/@evertongomede/understanding-sequence-to-sequence-seq2seq-models-and-their-significance-d2f0fd5f6f7f.

[9] Noticias Institucionales. *Universidad de La Sabana*. 2018. URL: https://www.unisabana.edu.co/nosotros/noticias-institucionales/detalle-noticias-institucionales/noticia/estas-son-las-cinco-causas-de-consulta-mas-frecuentes-en-el-centro-medico-sabes-como-prevenirl/.

[10] Saahil Jain et al. *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports*. 2021. arXiv: 2106.14463 [cs.CL]. URL: https://arxiv.org/abs/2106.14463.

[11] Jina. *What is Cross-Modal Multi-Modal?* URL: https://docs.jina.ai/v3.8.4/get-started/what-is-cross-modal-multi-modal/.

[12] Baoyu Jing, Pengtao Xie y Eric Xing. «On the Automatic Generation of Medical Imaging Reports». En: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018. DOI: 10.18653/v1/p18-1240. URL: http://dx.doi.org/10.18653/v1/P18-1240.

[13] Alistair E. W. Johnson, Tom J. Pollard, Samuel J. Berkowitz et al. «MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports». En: *Scientific Data* 6 (2019), pág. 317. DOI: 10.1038/s41597-019-0322-0. URL: https://doi.org/10.1038/s41597-019-0322-0.

[14] Daniel Johnson. *Seq2seq (Sequence to Sequence) Model with PyTorch*. Accessed: 2024-09-08. 2024. URL: https://www.guru99.com/seq2seq-model.html.

[15] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL]. URL: https://arxiv.org/abs/1910.13461.

[16] Christy Y. Li et al. «Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation». En: 2018. arXiv: 1805.08298 [cs.CV]. URL: https://arxiv.org/abs/1805.08298.

[17] Fenglin Liu et al. «Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation». En: 2021. arXiv: 2111.04318 [cs.LG]. URL: https://arxiv.org/abs/2111.04318.

[18] Fenglin Liu et al. «Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation». En: 2021. arXiv: 2106.06963 [cs.CV]. URL: https://arxiv.org/abs/2106.06963.

[19] Guanxiong Liu et al. «Clinically Accurate Chest X-Ray Report Generation». En: 2019. arXiv: 1904.02633 [cs.CV]. URL: https://arxiv.org/abs/1904.02633.

[20] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

[21] Ze Liu et al. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. arXiv: 2103.14030 [cs.CV]. URL: https://arxiv.org/abs/2103.14030.

[22] Yasuhide Miura et al. *Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation*. 2021. arXiv: 2010.10042 [cs.CL]. URL: https://arxiv.org/abs/2010.10042.

[23] Mehdi Moradi et al. «Bimodal network architectures for automatic generation of image annotation from text». En: 2018. arXiv: 1809.01610 [cs.CV]. URL: https://arxiv.org/abs/1809.01610.

[24] Aaron Nicolson, Jason Dowling y Bevan Koopman. «Improving chest X-ray report generation by leveraging warm starting». En: *Artificial Intelligence in Medicine* 144 (oct. de 2023), pág. 102633. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2023.102633. URL: http://dx.doi.org/10.1016/j.artmed.2023.102633.

[25] Yue Pan et al. «Chest Radiology Report Generation Based on Cross-Modal Multi-Scale Feature Fusion». En: *Journal of Radiation Research and Applied Sciences* 17 (2024), pág. 100823. DOI: 10.1016/j.jrras.2024.100823. URL: https://www.sciencedirect.com/science/article/pii/S1687850724000074.

[26] *Papers with Code*. URL: https://paperswithcode.com/task/medical-report-generation.

[27] Daniel Parres, Alberto Albiol y Roberto Paredes. «Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation». En: *Bioengineering* 11 (2024). DOI: 10.3390/bioengineering11040351.

[28] Jonathan Rubin et al. *Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks*. 2018. arXiv: 1804.07839 [cs.CV]. URL: https://arxiv.org/abs/1804.07839.

[29] Thomas Schlegl et al. «Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks». En: *Information Processing in Medical Imaging (IPMI)*. Vol. 24. Springer, 2015, págs. 437-448. DOI: 10.1007/978-3-319-19992-4_34.

[30] Hoo-Chang Shin et al. «Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation». En: 2016. arXiv: 1603.08486 [cs.CV]. URL: https://arxiv.org/abs/1603.08486.

[31] Akshay Smit et al. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. 2020. arXiv: 2004.09167 [cs.CL]. URL: https://arxiv.org/abs/2004.09167.

[32] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[33] Xiaosong Wang et al. «TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays». En: 2018. arXiv: 1801.04334 [cs.CV]. URL: https://arxiv.org/abs/1801.04334.

[34] Song Yang et al. «Knowledge Matters: Chest Radiology Report Generation with General and Specific Knowledge». En: *Medical Image Analysis* 80 (2022), pág. 102510. DOI: https://doi.org/10.1016/j.media.2022.102510. URL: https://www.sciencedirect.com/science/article/pii/S1361841522001578.

[35] Song Yang et al. «Radiology Report Generation with a Learned Knowledge Base and Multi-Modal Alignment». En: *Medical Image Analysis* 86 (2023), pág. 102798. DOI: https://doi.org/10.1016/j.media.2023.102798. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000592.

[36] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. arXiv: 1904.09675 [cs.CL]. URL: https://arxiv.org/abs/1904.09675.

[37] Guo Zhao et al. «Radiology Report Generation with Medical Knowledge and Multilevel Image-Report Alignment: A New Method and its Verification». En: *Artificial Intelligence in Medicine* 146 (2023), pág. 102714. DOI: https://doi.org/10.1016/j.artmed.2023.102714. URL: https://www.sciencedirect.com/science/article/pii/S0933365723002282.

APÉNDICE A

Código en Python para la automatización de Reescritura de Informes de Radiología utilizando la API de OpenAI

```python
from openai import OpenAI
import pandas as pd

# Set your OpenAI API key
OPENAI_API_KEY = 'YOUR_API_KEY_HERE'

# Initialize the OpenAI client
client = OpenAI(api_key=OPENAI_API_KEY)

# Function to rewrite a radiology report using OpenAI GPT API
def rewrite_radiology_report(text):
    completion = client.chat.completions.create(
        model="gpt-3.5-turbo",
        messages=[
            {"role": "system", "content": "You are a medical assistant, skilled in rewriting medical reports of Radiology with clarity and precision."},
            {"role": "user", "content": f"Rewrite the following radiology report and do it as a paragraph:\n\n{text}"}
        ]
    )
    rewritten_text = completion.choices[0].message.content
    return rewritten_text

# Read the CSV file
df = pd.read_csv('prueba.csv')

# Get the first column 'text'
texts = df['text'].tolist()

```

```
# Rewrite each radiology report (first 10 entries for this example)
rewritten_texts = []
for text in texts:
    rewritten = rewrite_radiology_report(text)
    print(f"Original: {text}\nRewritten: {rewritten}\n")
    rewritten_texts.append(rewritten)

# Save the rewritten texts into a new CSV file
rewritten_df = pd.DataFrame({'rewritten_text': rewritten_texts})
rewritten_df.to_csv('rewritten_radiology_reports.csv', index=False)

print("The rewritten radiology reports have been saved to 'rewritten_radiology_reports.csv'.")
```

APÉNDICE B
Relación del proyecto con los Objetivos del Desarrollo Sostenible

La relación de nuestro proyecto con los Objetivos de Desarrollo Sostenible (ODS) es clave para comprender su impacto en diferentes ámbitos sociales, económicos y tecnológicos. Los ODS propuestos por la ONU buscan soluciones globales a los desafíos del mundo actual, y nuestro proyecto de generación automática de informes de radiología puede contribuir de manera significativa a varios de ellos. A continuación, se detallan los ODS más relevantes para nuestro proyecto.

B.0.1. Salud y bienestar (ODS 3)

Este objetivo tiene un alto nivel de relevancia para nuestro proyecto, ya que se centra en garantizar una vida sana y promover el bienestar para todas las personas. Nuestra solución es una aplicación directa para hospitales y centros de salud, mejorando la eficiencia en la elaboración de informes médicos. Al automatizar este proceso, los profesionales médicos podrán dedicar más tiempo a la atención directa de los pacientes, mejorando así la calidad del cuidado y reduciendo tiempos de diagnóstico.

B.0.2. Trabajo decente y crecimiento económico (ODS 8)

Este objetivo tiene una relación media con nuestro proyecto. La automatización de la generación de informes médicos reducirá la carga de trabajo de los profesionales de la salud, permitiéndoles centrarse en actividades de mayor valor. Además, al optimizar los recursos humanos y económicos en los hospitales, se puede redirigir parte del presupuesto hacia otras áreas críticas, fomentando así el crecimiento económico del sector salud.

B.0.3. Industria, innovación e infraestructuras (ODS 9)

Nuestro proyecto tiene una alta vinculación con este objetivo, ya que propone una solución innovadora que potencia tanto a la industria tecnológica como a las infraestructuras hospitalarias. La implementación de sistemas automatizados para la creación de informes médicos representa un avance significativo en la innovación tecnológica aplicada a la medicina, mejorando también las infraestructuras digitales de los hospitales al facilitar su modernización.

B.0.4. Ciudades y comunidades sostenibles (ODS 11)

Aunque la relación con este objetivo es baja, nuestro proyecto contribuye a crear comunidades más sostenibles al reducir las largas listas de espera en los hospitales. La mejora en la eficiencia del diagnóstico y la elaboración de informes disminuye los tiempos de atención, lo que repercute positivamente en el sistema de salud local y, en consecuencia, en el bienestar de las comunidades.

B.0.5. Alianzas para lograr los objetivos (ODS 17)

Este objetivo tiene una relación alta con nuestro proyecto, ya que su desarrollo ofrece una excelente oportunidad para fomentar la colaboración entre investigadores, equipos médicos y tecnológicos. Invitar a otros a participar en la mejora continua de esta solución permitirá alcanzar un modelo que sea viable para su implementación generalizada en hospitales y centros de salud, promoviendo así el avance hacia el cumplimiento de los ODS.

Tabla B.1: Evaluación de los Objetivos de Desarrollo Sostenible (ODS) según el nivel de importancia.

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**				**X**
ODS 2. **Hambre cero.**				**X**
ODS 3. **Salud y bienestar.**	**X**			
ODS 4. **Educación de calidad.**				**X**
ODS 5. **Igualdad de género.**				**X**
ODS 6. **Agua limpia y saneamiento.**				**X**
ODS 7. **Energía asequible y no contaminante.**				**X**
ODS 8. **Trabajo decente y crecimiento económico.**		**X**		
ODS 9. **Industria, innovación e infraestructuras.**	**X**			
ODS 10. **Reducción de las desigualdades.**				**X**
ODS 11. **Ciudades y comunidades sostenibles.**			**X**	
ODS 12. **Producción y consumo responsables.**				**X**
ODS 13. **Acción por el clima.**				**X**
ODS 14. **Vida submarina.**				**X**
ODS 15. **Vida de ecosistemas terrestres.**				**X**
ODS 16. **Paz, justicia e instituciones sólidas.**				**X**
ODS 17. **Alianzas para lograr objetivos.**	**X**			