Medical Report Generation through Radiology Images: An Overview

Olanda Prieto-Ordaz, Graciela Ramírez-Alonso, Manuel Montes-y-Gómez, Roberto López-Santillán

***Abstract*—The interpretation of medical images is a fundamental process for the diagnosis and treatment of patients. This process contributes determining the causes of symptoms as well as monitoring the effects of any treatment. Although the generation of medical reports from images is a complex task, deep learning strategies have been integrated with models that allow this arduous task to be tackled, achieving promising results. This work aims to present a compilation of the most outstanding deep learning strategies focused on the automatic generation of medical radiology reports from X-Ray images. Papers based on DenseNet, ResNet and VGG architectures, in combination with Long Short-Term Memories (LSTMs) and attention models, are analyzed in terms of the pre-processing strategies, databases used, model adaptations, and metric results. All these important findings are summarized in this survey, highlighting those models that reported the highest performance.**

***Index Terms*—Medical Reports, Deep Learning, Medical Images.**

I. Introducción

La interpretación de imágenes médicas es un proceso fundamental para el diagnóstico y tratamiento de enfermedades, ya que permite determinar las causas de alguna afección que pudiera presentar el paciente [1]–[4]. Además, permite monitorear y evaluar la utilidad de algún tratamiento, facilitando llevar un seguimiento de la evolución de la enfermedad en el paciente [5].

Para la elaboración de un diagnóstico y posterior reporte médico, el especialista analiza diferentes tipos de imágenes obtenidas con distintas técnicas de proyección, examinando minuciosamente cada región de las mismas. El objetivo de esta tarea es identificar las regiones normales, anormales o potencialmente anormales. Una vez que termina este proceso, se realiza una narrativa de lo encontrado generando así un reporte médico [6]. La interpretación de las imágenes médicas para la elaboración de un reporte médico es una tarea compleja que debe ser elaborada por radiólogos con estudios en programas especializados por varios años.

Un aspecto a considerar es la gran diversidad de los tipos de imágenes médicas, por ejemplo, dentro de las imágenes de radiología se encuentran: los rayos X (RX), las Tomografías Computarizadas (CT), las Imágenes de Resonancia Magnética (MRI), las Tomografías por Emisión de Positrones (PET) y los Ultrasonidos (US) [7]. Por consecuencia, elaborar un reporte médico utilizando imágenes de radiología demanda una mayor especialización en el área [5]. Aunado a esto, suelen existir variaciones en la interpretación de las imágenes médicas entre un especialista y otro [8]–[10]. De igual manera, los tiempos de ejecución para la realización de un reporte médico varían dependiendo del especialista que lo realiza. Por todo lo anterior, diferentes investigadores han enfocado sus esfuerzos al desarrollo de modelos computacionales que apoyen o aligeren la carga de trabajo de especialistas médicos en este importante proceso [11].

Modelos computacionales de aprendizaje profundo (DL) basados en redes convolucionales (CNN) en conjunto con redes recurrentes (RNN) han reportado resultados sobresalientes en tareas relacionadas con la generación automática de reportes médicos a partir de imágenes [5], [6], [11]–[15], [15]–[22]. De manera general, la estrategia de entrenamiento que siguen estos modelos es la siguiente: considerando como entradas imágenes médicas y sus reportes asociados, un modelo de CNN se encarga de extraer las características relevantes de la imagen, posteriormente, un modelo RNN (comúnmente una red *Long Short Term Memory* o LSTM) asocia las características extraídas por la CNN con cada palabra del reporte de entrada. Este proceso se realiza iterativamente generando palabra por palabra, cada sentencia del reporte de salida. Una vez que termina la etapa de entrenamiento, el modelo será capaz de generar reportes para nuevas imágenes que le sean alimentadas como entrada. Algunas propuestas solamente utilizan la salida de la CNN para generar predicciones que describen brevemente el tipo de lesión que se presenta [23]–[26]. Por su parte, otros modelos generan el texto del reporte médico utilizando una estrategia de entrenamiento de principio a fin (*end to end*) [15], [17], [19], [22], [27]–[31]. El diagrama de la Fig.1, representa de manera general el proceso de entrenamiento que siguen los modelos descritos anteriormente.

Aunque existen diferencias en los enfoques de DL propuestos en la literatura, el proceso general tiene en común los siguientes pasos:

1. Seleccionar una base de datos que contenga imágenes médicas y reportes médicos asociados.
2. Realizar tareas de pre-procesamiento con los datos de entrada, es decir con las imágenes y el texto.
3. Extraer características de la imagen médica con un modelo de CNN.
4. Asociar las características encontradas por la CNN con las descripciones textuales de los reportes médicos, utilizando un modelo de RNN.

Olanda Prieto Ordaz, Graciela Ramirez-Alonso and Roberto-López Santillán are with Facultad de Ingeniería, Universidad Autónoma de Chihuahua, Circuito Universitario Campus II Chihuahua,Chih, C.P. 31125, México.

Manuel Montes-y-Gómez is with Instituto Nacional de Astrofísica, Óptica y Electrónica INAOE, Luis Enrique Erro No.1, Sta María Tonanzintla, 72840 San Andrés Cholula, Puebla.

E-mail of corresponding author: galonso@uach.mx

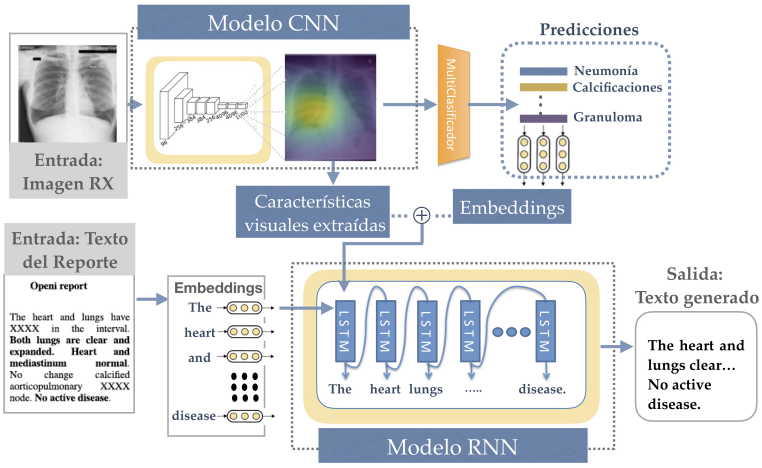


Fig 1. Representación general de un modelo de DL para la generación de reportes médicos durante el entrenamiento.

5. Generar el reporte médico para una imagen de entrada y evaluar los resultados del modelo.

El presente trabajo muestra una perspectiva general y los modelos propuestos más relevantes enfocados a la generación de reportes médicos a partir de imágenes de rayos X. Se decidió considerar solamente este tipo imágenes por ser las más utilizadas en los trabajos analizados. A diferencia de otras revisiones de literatura orientadas a la generación de reportes médicos [5], [13], [32], en este análisis se incluyen las diferentes estrategias de pre-procesamiento y pre-entrenamiento adoptadas en los enfoques analizados. Además, se identifican las arquitecturas de CNNs más utilizadas para el procesamiento de imágenes de RX, y los modelos de RNNs para el procesamiento de texto, enfatizando los componentes más relevantes en cada arquitectura. En este análisis se muestra una comparativa de los resultados con las distintas métricas utilizadas en esta tarea, identificando los modelos más sobresalientes. De igual manera, se muestran las bases de datos más utilizadas que consideran imágenes de RX para la generación de reportes médicos. Todo esto mediante el análisis de artículos que se publicaron del 2016 a la fecha.

II. Descripción General de Estrategias de Aprendizaje Profundo

A. *Redes Neuronales Convolucionales*

Dentro del área de DL, los modelos más utilizados para tareas de procesamiento de imágenes están basados en CNNs [33]. Partiendo de las investigaciones realizadas en [34]–[36], se propuso el modelo *LetNet*, una red neuronal artificial que estableció un marco de referencia para la CNN. Desde el año 2006, se han presentado e implementado diferentes modelos basados en una arquitectura básica de CNN, los cuales se caracterizan por incluir mayor cantidad de capas que los modelos iniciales [37], [38]. *AlexNet* [39], *VGGNet* [40], *GoogleNet* [37], *DenseNet* [41] y *ResNet* [42] son algunos de los modelos más representativos de CNNs utilizados en diversas tareas del área de visión por computadora. Dentro de las estrategias propuestas para la tarea de generación de reportes médicos mediante imágenes de RX, las CNNs más destacadas son: *DenseNet* [16], [28], [31], [43]–[46]; *ResNet* [17], [19], [20], [22], [27], [30], [47]–[50]; y *VGG* [6], [15], [51].

B. *Redes Neuronales Recurrentes*

Las Redes Neuronales Recurrentes *(RNN)* han demostrado buen desempeño en tareas que involucran el análisis de entradas secuenciales, como es el caso del Procesamiento de Lenguaje Natural (PLN). Una RNN procesa elemento por elemento la secuencia de entrada, mientras conserva en sus capas ocultas un vector de estado que contiene información de la secuencia anterior; por lo tanto, se define que el estado oculto actúa como memoria [36]. Las RNN son consideradas celdas de memoria a corto plazo dado que al recibir secuencias muy largas, las últimas celdas tienen dificultad de recibir la información de las capas iniciales [36]. Las Celdas de Memoria a Largo Plazo *(LSTM)* y las Unidades de Compuertas Recurrentes *(GRU)*, a diferencia de las RNN, poseen un mecanismo interno conformado por diferentes compuertas. Estas compuertas sirven para identificar cuál información dentro de una secuencia de entrada debe almacenar y cuál debe desechar [52], [53]. Lo anterior contribuye a que las LSTM y las GRU tengan la capacidad de recordar secuencias largas; es decir, que cuenten con un mecanismo que actúa como memoria.

Para la tarea de generación de reportes médicos se necesita del procesamiento de una secuencia de palabras. Comúnmente, este proceso suele realizarse utilizando redes LSTMs [6], [15], [17]–[20], [22], [27], [28], [31], [43], [46]–[48], [50], [54], [55]; o LSTM Bidireccionales (Bi-LSTMs) [29], [30], [43], [49], [49], [51]. Además, algunos de estos modelos incluyen mecanismos de atención que permiten ponderar con mayor relevancia algunos elementos de la entrada, contribuyendo así a mejorar su efectividad y eficiencia [6], [15], [21], [56]–[59].

C. *Representación de Palabras y Modelos de Lenguaje*

La exploración de estrategias que permitan el uso de modelos de DL para el PLN, es fundamental para tareas relacionadas con la generación automática de reportes médicos. Procesar un texto médico automáticamente a partir

de la imagen requiere de representaciones matemáticas que engloben las propiedades semánticas y sintácticas de cada palabra [60]. Considerando los estudios realizados en tareas de PLN, se ha destacado el uso de los *word embeddings* (WE) [61]. Dentro de las estrategias de WE se encuentran Word2vec [62], GloVe [63] y fastText [61], los cuales se consideran vectores de palabras no contextualizados. Sin embargo, la mayoría de los estudios realizados consideran una perspectiva del dominio general del texto y los resultados no necesariamente aplican para el procesamiento de texto en otros dominios como el biomédico [5]. Por otro lado, los *embeddings* contextualizados permiten generar diferentes vectores que representen los diversos significados de una palabra de acuerdo al contexto de la oración. De acuerdo a [60], dentro del estado del arte se encuentran modelos que permiten generar esta contextualización en los *embeddings* como: ELMo [64] y BERT [65]. Para la tarea de generación de reportes médicos algunos trabajos como [17], [22], [55], utilizan WE de GloVe y Word2vec pre-entrenados con el objetivo de robustecer la representación de los datos y obtener un mejor desempeño.

III. Bases de Datos: Imagen/Texto en Área Médica

Al implementar modelos que utilicen estrategias de DL para la tarea de generación de reportes médicos, se requiere de bases de datos que contengan imágenes médicas y reportes médicos asociados a cada imagen [19], [66], [67].

Dentro de las bases de datos más utilizadas para el entrenamiento y evaluación de modelos se encuentra *The Indiana University chest X-ray, IU X-Ray* [68]; la cual cuenta con 7,470 imágenes de RX de toráx que incluyen la vista frontal y lateral, así como 3,955 reportes médicos de dos hospitales.

Otra base de datos pública muy utilizada es *ChestX-ray14* [69], del *National Institute of Health (NIH) Clinical Center*. *ChestX-ray14* está conformada por 112,120 imágenes de RX e incluye 14 etiquetas de enfermedades de pecho.

CheXpert [70] es una base de datos pública de imágenes de RX de tórax que contiene 224,316 imágenes de 65,240 pacientes. La base de datos considera la clasificación de imágenes identificando la presencia o ausencia de 14 tipos de patología.

La base de datos *CX-CHR* [16] es una colección privada de una institución médica de salud conformada por 35,500 pacientes. Cada expediente contiene una o más imágenes de RX de tórax y los reportes médicos correspondientes están redactados en lenguaje mandarín.

Las bases de datos de *ImageCLEF Caption 2017* [71] e *ImageCLEF Caption 2018* [72], son grandes colecciones de imágenes biomédicas que incluyen imágenes de radiología de tórax y de diferentes partes del cuerpo. Las imágenes fueron extraídas del *PubMed Central (PMC)* y tienen asociada una oración, así como un conjunto de etiquetas UMLS *(Unified Medical Language System)*. La base de *ImageCLEF Caption* 2017 contiene 184,614 imágenes mientras que *ImageCLEF Caption* 2018 cuenta con 232,305 imágenes. *ImageCLEF Caption 2017* se ha utilizado en [15] para la tarea de generación de reportes médicos, mientras que *ImageCLEF Caption* 2018 solo ha reportado su uso para tareas de generación de subtítulos organizadas por *ImageCLEFcaption*.

MIMIC-CXR [73], [74] es una colección de datos reciente que cuenta con 371,920 imágenes de RX de tórax y 227,943 reportes asociados del *Beth Israel Deaconess Medical Center*. Aunque es una de las más grandes, su uso es limitado por ser relativamente reciente. Se ha utilizado en tareas orientadas a clasificación de imágenes médicas [25], [75], [76] y en la tarea de generación de reportes médicos [28].

La base de datos de *Eye-Gaze Dataset* es un subconjunto de la base de datos de *MIMIC-CXR*. La diferencia con *MIMIC-CXR* es que a un subconjunto de 1,083 imágenes de RX le agregaron los registros de movimientos oculares de un especialista durante el proceso del análisis [77]. Además, incluye un audio con la narrativa del diagnóstico. Aunque su uso no ha sido reportado para la generación de reportes médicos, se considera puede contribuir significativamente en esta área.

Las PACs son bases de datos privadas y utilizadas en centros médicos. Dentro de este conjunto se encuentra el sistema PACs del Hospital de Shaanxi, China, que contiene 16,569 imágenes RX de tórax y 16,569 reportes médicos [26]. El PACs-2 de *The National Library Medicine, National Institutes of Health*, contiene 8,121 imágenes de RX y 3,996 reportes [68]. El conjunto de PACs-3 del Hospital de Shanghai del 2015 [47], está constituido por 19,985 imágenes de RX de tórax, que incluyen los 19,985 reportes asociados.

Otra base de datos es *PadChest* [78], la cual contiene 160,868 imágenes de RX de tórax con 6 vistas diferentes y 109,931 reportes médicos. Además incluye 174 descripciones o *findings*, 19 diagnósticos y 104 anotaciones de la ubicación. Debido a que es una base de datos relativamente nueva, su uso es limitado.

PEIR Digital Library [6] contiene 4,732 imágenes de RX clasificadas en 20 categorías. Fue creada por *The University of Alabama* y se caracteriza por incluir descripciones a nivel de oración de 20 partes diferentes del cuerpo.

La Tabla I muestra las bases de datos públicas y privadas que incluyen imágenes de RX utilizadas para tareas de generación de reportes médicos a partir de imágenes.

IV. Pre-procesamiento y uso de Modelos Pre-entrenados

Para la generación de reportes médicos a partir de imágenes médicas, el pre-procesamiento, limpieza de los datos y uso de modelos pre-entrenados, son las consideraciones más comunes que han dado buenos resultados en este tipo de tareas [5], [11].

Las técnicas de aumento de datos permiten reducir problemas relacionados con el sobre-entrenamiento o el desbalance de las clases [11]. Esto se debe a que al crear variaciones de los datos de entrada se robustece el aprendizaje del modelo [82]. Dentro de las técnicas más utilizadas se encuentran: escalamiento, rotación, recorte y las variaciones en la intensidad de color [14], [26], [45], [79].

Para cumplir con criterios de calidad en la información que entra al modelo, varios autores reportan eliminar determinadas

TABLA I
BASES DE DATOS DE IMÁGENES MÉDICAS DE RADIOGRAFÍA CON TEXTO MÉDICO ASOCIADO.

Bases de datos	Imágenes	Descripción	Utilizado por
IU X-Ray [68]	7,470	3,955 reportes	[6], [14], [16], [17], [31], [54] [19], [22], [27], [28], [43], [44], [49], [50], [55] [20], [29], [45], [46], [48], [51]
ChestX-ray14 [69]	112,120	14 etiquetas	[17], [20], [79]
CheXpert [70]	224,316	14 tipos de patología	[50]
CX-CHR	35,500	35,500 reportes	[16], [20], [44]
ImageCLEF-Caption-2017 [71]	184,614	184,614 reportes	[15]
ImageCLEF-Caption-2018 [72]	232,305	232,305 reportes	-
MIMIC-CXR [73]	371,920	227,943 reportes	[25], [28]
Eye-Gaze Dataset [77]	1,083	Reportes transcritos Audio de reporte Datos de monitoreo de ojo	-
PACs from Hospital of Shaanxi, China(PACs-1)	16,569	16,569 reportes	[26]
PACs of the National Library Medicine, National Institutes of Health (PACs-2) [81]	8,121	3,996 reportes	[80]
PACs from Hospital of Shanghai(PACs-3) [47]	19,985	19,985 reportes	[47]
PadChest [78]	160,868	109,931 reportes	-
PEIR Digital Library [6]	4,732	20 multi-etiquetas	[6]

secciones del conjunto de datos. Por ejemplo, Biswal et al. [43] eliminaron las imágenes duplicadas de MIMIC-CXR; Yuan et al. [50] eliminaron las imágenes con menos de 2 vistas de IU X-Ray; Gu et al. [47] removieron las radiografías laterales de PACs-3; Gasimova et al. [27] eliminaron los reportes vacíos de IU X-Ray; Xue et al. [49] y Singh et al. [55] removieron los reportes que no cuentan con las secciones de *impressions* y *findings* de IU X-Ray. Además, Xue et al. [49] eliminaron los reportes que no disponían de dos imágenes completas de IU X-Ray. Por su parte, Li et al. [48] removieron las imágenes y los reportes no relevantes a 8 enfermedades de pecho de IU X-Ray.

Por otro lado, varios autores han optado por utilizar un modelo de CNN pre-entreando con la base de datos ImageNet o alguna base de datos de imágenes médicas con el fin de inicializar los pesos del modelo de manera más certera [15]–[17], [19], [22]–[24], [27], [30], [31], [44]–[46], [48], [49], [55], [79], [83].

En el área de radiología, el PLN se ha utilizado para extraer información de reportes médicos, librerías digitales o incluso de redes sociales. Lo anterior con la finalidad de proveer datos estructurados que puedan utilizarse por los modelos de DL [84]. Para el pre-procesamiento en el modelo de texto, uno de los procedimientos más comunes es remover todos los caracteres no alfabéticos y convertirlos a minúscula [6], [27], [29]–[31], [48]. Además del procedimiento anterior, otra estrategia consiste en filtrar todas las palabras incluidas en los reportes que no cumplan con determinado número de ocurrencias (frecuencia) [29], [47] [17], [31], [46], [50] [43]. Estas palabras se definen como *UNK* y finalmente se establece un estado de *START* y *END* para definir el inicio y fin de cada oración o párrafo [17], [29], [46], [48]–[50].

Algunas propuestas incluyen un pre-procesamiento más específico, por ejemplo, Harzig et al. [22] clasificaron manualmente como normal o anormal cada sentencia del reporte médico asociado de IU X-Ray. Además, para lograr una mejor representación del texto, utilizaron el algoritmo de Word2vec previamente pre-entrenado en Pubmed y Wikipedia [85].

Por otra parte, Li et al. [16] identificaron las sentencias con mayor número de ocurrencias en los reportes médicos de IU-XRay. Mediante este procedimiento, se establecieron 97 sentencias con ocurrencia mayor a 500. A su vez, Li et al. [44] definieron sentencias con descripciones anormales que cumplieran con determinado número de ocurrencias. Posteriormente, agruparon manualmente las sentencias con el mismo significado y seleccionaron las más frecuentes de cada grupo para establecerlas como 87 plantillas de IU X-Ray y 362 plantillas de CX-CHR.

En algunas propuestas como [19] las secciones de *indications*, *impressions* y *findings* fueron concatenadas para establecerlas como etiquetas (*groundtruth*) de IU X-Ray. En otros casos solo se consideraron las secciones de *impressions* y *findings* como en [22], [31], [46], [48], [50], [55] de IU X-Ray.

Wang et al. [17] utilizaron el algoritmo de Word2Vec pre-entrenado en artículos de PubMed para obtener una mejor representación del texto. Por otra parte, Singh et al. [55] utilizaron el algoritmo de *GloVe* pre-entrenado con texto genérico de *Common Crawl*. También utilizaron un LSTM entrenado con *RadGlove*, el cual tiene 4.5 millones de reportes de radiología de *Stanford University*. Dong et al. [26] dividieron la sección de diagnosis en pequeñas cláusulas de PACs-2 y posteriormente agruparon las que son similares. Además, aplicaron el algoritmo de *K-medoids* con el objetivo de identificar los 10 agrupamientos más grandes y asignar una etiqueta a cada uno de ellos. Por su parte, Gasimova et al. [27] ajustaron los reportes recortando o rellenando cada uno de ellos para establecerlo a 32 palabras.

La Tabla II muestra todas estas diferentes estrategias de pre-procesamiento que se han implementado, así como el uso de modelos pre-entrenados, indicando los artículos que los emplearon.

V. MÉTRICAS DE EVALUACIÓN

Las métricas que comúnmente se utilizan para evaluar el desempeño de los modelos orientados a la generación de

TABLA II
PRE-PROCESAMIENTO Y USO DE MODELOS PRE-ENTRENADOS

Descripción	Particularidad	Usado en
Pre-procesamiento en la imagen		
Aumento de datos	Escalamiento	[19], [24], [26], [29] [30], [45], [49], [79]
	Rotación	[45], [46]
	Recorte	[14], [45]
	Intensidad	[30]
Eliminación de datos	Vista Lateral	[47]
	Duplicidad	[28], [43]
	<2 vistas	[50]
	Datos no comunes	[48]
Modelos de CNN pre-entrenados		
Pre-entrenamiento	Imagenet	[17], [19], [24], [31] [27], [30], [49], [83] [15], [48], [55], [79]
	ChestX-ray8	[16], [44]–[46] [48]
	CheXpert	[48]
	CX-CHR	[22]
Fine-tuning	CX-CHR	[16], [44]
	PubMed Central Biomedical Image corpus	[15]
	PACs [23]	[24]
Pre-procesamiento en texto		
Palabra	Convertir a minúsculas	[6], [27], [29]–[31] [48]
	Eliminar caracteres	[29]–[31], [48] [6], [27]
	Considera aquellas ≥ 2	[29], [47]
	Considera aquellas ≥ 3	[17], [31], [46], [50]
	Considera aquellas ≥ 5	[43]
Sentencia	Clasifica Norma/Anormal	[22]
	Ocurrencia ≥ 500	[16], [44]
Sección	Concatena *findings-Impressions-indications*	[19]
	Concatena *findings-Impressions*	[22], [31], [50]
	findings-Impressions	[46], [48], [55]
	Divide *Diagnosis*	[26]
Reporte	Elimina Incompleto	[27], [28], [49], [55]
Modelos de WE pre-entrenados		
Embbedings	Word2Vec	[17], [22]
	GloVe	[55]
	RadGlove	[55]

reportes médicos a partir de imágenes son:

BLEU *Bilingual Evaluation Understudy* [86]. Esta métrica contabiliza el número de *n-gramas* de palabras de la sentencia generada por el modelo que coinciden con los de la sentencia objetivo. El número de palabras a considerar dentro del *n-grama* puede ser de 1 (BLEU1) hasta 4 (BLEU4), y sus valores corresponden al intervalo 0 a 1.

ROUGE-L *Recall-Oriented Understudy for Gisting Evaluation* [87]. La métrica ROUGE-L permite calcular la media armónica (*F-measure*) entre la Precisión y recuerdo a nivel de las sub-secuencias comunes más largas (*Longest Common Subsequence* o LCS) entre la sentencia generada y la sentencia objetivo. Sus resultados abarcan de un valor mínimo de 0 a un valor máximo de 1.

METEOR (M) [88]. Esta métrica evalúa la precisión de la sentencia generada por el modelo con respecto a la sentencia objetivo considerando un orden explícito de los *n-gramas*. Además, toma en cuenta el uso de sinónimos entre los *n-gramas*. Sus resultados van de un valor mínimo de 0 a un valor máximo de 1.

CIDEr *Consensus-based Image Description Evaluation* [89]. CIDEr determina la similitud entre sentencias basado en el cálculo del coseno entre el vector de pesos de los *n-gramas* de la sentencia generada y el vector de pesos de los *n-gramas* de la sentencia objetivo. Estos pesos se calculan de acuerdo al valor de *Term Frequency Inverse Document Frequency* (TF-IDF) que pondera los *n-gramas* más importantes según su frecuencia de aparición en el conjunto de sentencias objetivo. Además, esta métrica agrupa las palabras que tienen una misma raíz o *stemming* al realizar la ponderación de los *n-gramas*. Sus valores corresponden al intervalo 0 a 10.

Estas métricas se utilizan para calcular el desempeño del modelo propuesto observando la similitud o la diferencia entre los párrafos generados y las descripciones escritas por los radiólogos. Un buen desempeño se refleja en puntuaciones altas en BLEU, ROUGE, METEOR y CIDEr. Algunos autores como [14], [23], [27], [28], [43], [54], [80], además de utilizar las métricas para evaluar el reporte generado, incluyen métricas para evaluar la correcta clasificación de la imagen como:

Accuracy. Esta métrica se utiliza para evaluar la relación de las predicciones correctas e incorrectas en la clasificación de la imagen [90]. Sus valores corresponden al intervalo 0 a 1.

Area under the ROC curve **(AUC)**. El ROC *(Receiver Operating characteristic curve)* es una representación gráfica que proyecta la proporción de las predicciones correctas o verdaderos positivos (reflejadas en el eje y) con respecto a la proporción de falsos positivos (eje x). La AUC mide el área bajo la curva de la gráfica ROC y sus valores corresponden al intervalo 0 a 1 [90].

VI. ARQUITECTURAS ORIENTADAS A LA GENERACIÓN DE REPORTES MÉDICOS

La tarea de generación automática de reportes médicos mediante estrategias de DL, tiene como objetivo formar oraciones y/o párrafos bien estructurados, partiendo de una representación visual y un texto fluido [91]. Para cumplir con esta tarea, se utilizan modelos que asocian las regiones visualmente más relevantes de la imagen con un texto con la finalidad de generar las descripciones del reporte médico. Dentro de las propuestas de DL orientadas a esta tarea se encuentran los siguientes enfoques: (a) modelos que solamente clasifican las imágenes en categorías pre-definidas utilizando una CNN, (b) modelos basados en arquitecturas de CNN y RNN simples, (c) modelos jerárquicos, (d) modelos híbridos, y (g) modelos mixtos, todos estos últimos para la generación del reporte médico completo. Estas diferentes estrategias se detallan en esta sección.

A. *Modelos Enfocados a la Clasificación de Imágenes*

Estos modelos se caracterizan por no generar un reporte médico, sin embargo, utilizan estrategias de PLN para extraer información relevante de los reportes y realizar la clasificación de las imágenes utilizando una CNN. La Fig. 2a muestra un diagrama general de esta estrategia. Por ejemplo, Rajpurkar et al. [79] propusieron la red *ChextNet* que permite la detección de neumonía mediante imágenes de RX. La arquitectura consta de una red DenseNet que incluye 121 capas densas de convolución. El modelo fue entrenado con los datos de *ChestX-ray 14* [69] considerando solamente las imágenes que presentaban características de neumonía para establecer una nueva clasificación como positiva, mientras que al resto se definió como negativa. Para interpretar las predicciones del modelo utilizaron mapas de calor, lo que permitió localizar las áreas afectadas en la imagen.

Por otra parte, Dong et al. [26] utilizaron los modelos VGG y ResNet con el objetivo de clasificar imágenes de RX automáticamente. Para esta tarea se establecieron 10 etiquetas pre-definidas de normal, crecimiento de pulmón, aortosclerosis, crecimiento de corazón y otras más. Las etiquetas fueron extraídas de reportes médicos pertenecientes al PACs-2, mediante estrategias de PLN para formar agrupaciones. Posteriormente, se identificaron las agrupaciones más relevantes para establecer las etiquetas y asociarlas a una enfermedad.

A su vez, Rubin et al. [25] implementaron una arquitectura basada en DenseNet-121. Este modelo utilizó una estrategia para etiquetar las imágenes considerando las secciones de *impressions* y *findings* del reporte médico por medio de la herramienta NeoBio [69], [92].

B. *Modelos Basados en Arquitecturas Simples de CNN y RNN*

Este tipo de estrategias divide el problema en dos partes. La primera parte se basa en un modelo CNN encargado de extraer las características de la imagen. Luego, se tiene una red RNN (LSTM, Bi-LSTM o GRU) que genera los enunciados o sentencias palabra por palabra, Fig. 2b. Un ejemplo de este tipo de modelos lo presenta Hasan et al. [15] en donde una red VGG-19, junto con un LSTM con un mecanismo de atención, recibe secuencialmente las características de la imagen y la descripción asociada a ella.

Los modelos propuestos por Shin et al. [14] se basan en una red GoogleNet en conjunto con una red LSTM [93] o GRU [94]. Aunque el enfoque con el modelo LSTM se caracterizó por un entrenamiento más simple, el modelo de GRU obtuvo un mejor desempeño en la generación de etiquetas. En ambas propuestas se integró un vector de contexto que procesa las características de la imagen y su descripción, formando la entrada a la LSTM junto con el reporte asociado. Este modelo permite detectar y describir la ubicación de la enfermedad, severidad y los órganos afectados.

El modelo denominado MDNET propuesto por Zhang et al. [18] establece un mapeo multimodal directo entre las imágenes médicas y los reportes médicos. MDNet incluye un modelo basado en ResNet con variaciones en las conexiones residuales que permiten mejorar las características multi-escala. Para el modelo de lenguaje se adoptó un LSTM que integra un mecanismo de atención. Lo anterior, contribuye a la alineación de las características visuales con las palabras de las oraciones obteniendo mapas de atención más nítidos.

Utilizando una arquitectura de ResNet y un LSTM, Wang et al. [17] propusieron un modelo denominado TieNet. Su enfoque se caracteriza por integrar una atención multinivel para la clasificación y generación del reporte. La arquitectura abarca todo el proceso de principio a fin para incorporar representaciones del texto y la imagen. Aunado a ello, se generan mapas de calor para cada palabra lo que facilita la evaluación de sus predicciones.

Gasimova et al. [27] propusieron dos modelos, uno basado en la red ResNet y otro en la red VGG, para realizar una clasificación de la imagen en múltiples etiquetas generadas con la herramienta de *Manual Medical Subject Heading*. Para la generación de un texto estructurado implementaron un modelo LSTM.

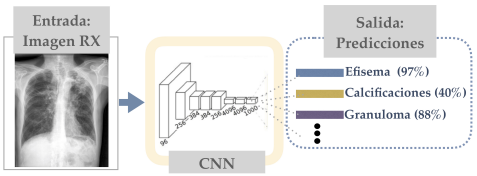
Gu et al. [47] utilizaron el modelo ResNet para extraer las características más relevantes de la imagen y clasificarla en múltiples etiquetas. Posteriormente, los mapas de características son utilizados para relacionar la información espacial y semántica. Para la generación de texto se adopta una red LSTM, la cual recibe como entrada la predicción del vector con la información espacial y semántica de la imagen.

Una arquitectura basada en la red Inception-V3 y LSTM fue propuesta por Singh et al. en [55]. En su estrategia se optó por pre-entrenar el modelo de lenguaje con *RadGlove embedding* [95] para lograr un dominio del contexto de radiología.

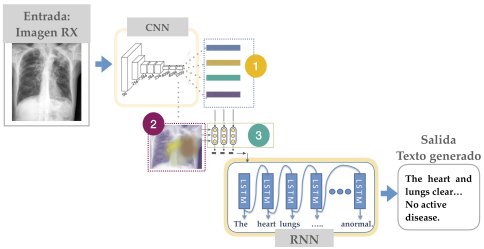
C. *Modelos Jerárquicos*

Los modelos jerárquicos se distinguen por la forma organizada de generar el texto. En la etapa inicial, se obtiene la información de la imagen por medio de una CNN. Posteriormente, una RNN recibe la información de la imagen para establecer la representación de la sentencia o tema que debe ser generado. A continuación, otra RNN recibe la representación de la sentencia para generar el texto palabra por palabra. Esta estructura permite dar pauta a la secuencia del texto que debe ser generado, ver Fig. 2c. Considerando este enfoque, Jing et al. [6] propusieron un modelo que permite desempeñar las tareas de predicción de etiquetas y la generación de las sentencias o párrafos de acuerdo a una etiqueta principal. La primera parte del modelo realiza la predicción de las etiquetas a través de una red VGG. Además, las características visuales aprendidas alimentan a una red multi-clasificador (MLC), que genera etiquetas a las características semánticas de la imagen.

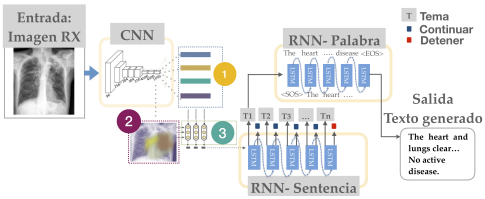
Otro enfoque es el propuesto por Xue et al. [49], el cual consiste en una arquitectura que permite generar un reporte de radiología automáticamente de principio a fin. El modelo procesa simultáneamente la vista frontal y lateral de la imagen mediante un modelo ResNet. El modelo generador de texto, basado en un Bi-LSTM, predice la primera palabra de la sentencia y la sentencia se produce palabra por palabra. Una vez que la sentencia es completada, el modelo toma como entrada la sentencia generada y las características globales de la imagen durante varias iteraciones hasta generar un párrafo de los hallazgos encontrados.



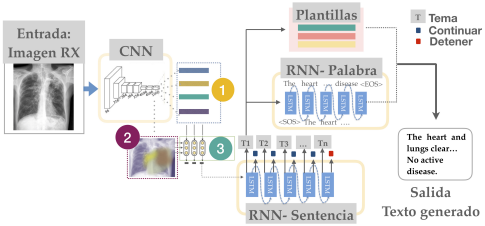
(a) Modelos enfocados a la clasificación de imágenes



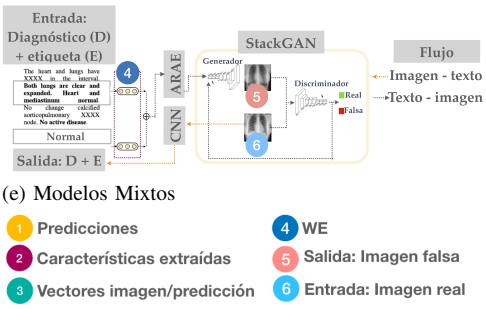
(b) Modelos basados en arquitecturas simples de CNN y RNN



(c) Modelos Jerárquicos



(d) Modelos Híbridos



(e) Modelos Mixtos

Fig 2. Diagramas representativos de las diferentes estrategias que se siguen para la generación de reportes médicos.

Por su parte, Harzig et al. [22] proponen un modelo que clasifica la imagen como normal o anormal y posteriormente genera su descripción. La estrategia de aprendizaje multi-tarea facilita un entrenamiento de principio a fin, donde las representaciones de la imagen y la generación de texto se llevan paralelamente. Su arquitectura está compuesta por un modelo ResNet, un mecanismo de atención, un modelo LSTM para generar el tema y dos LSTM para generar la sentencia asociada al tema. Esta sentencia es generada palabra por palabra.

Considerando una estrategia similar a [22], Xie, et al. [54] propusieron el modelo *Attention-based Abnormal-Aware Fusion Network (A3FN)*. La estrategia incluye una CNN y un modelo jerárquico de LSTMs con un mecanismo de atención. La diferencia principal con [22], es que utilizan un módulo de compuerta que permite clasificar a la imagen como normal o anormal y posteriormente generar el reporte médico.

Mediante una estrategia jerárquica recurrente con un mecanismo de atención, Yin et al. [31] propusieron una arquitectura compuesta por una DenseNet y LSTMs. A diferencia de los otros modelos, esta propuesta cuenta con un mecanismo de coincidencia que permite mapear los vectores de tema y las sentencias en un mismo espacio semántico.

La estrategia propuesta por Yuan et al. [50] es similar a [31], pero ellos utilizan una red pre-entrenada ResNet con imágenes de RX que permite reconocer hasta 14 enfermedades diferentes. Además, para robustecer la descripción semántica extrajeron conceptos médicos basados en reportes de radiología y los conceptos médicos más frecuentes de las imágenes de RX. Esta estrategia permitió generar múltiples sentencias estructuradas con una descripción acorde a las características de la imagen.

Liu, et al. [28] propusieron un modelo basado en DenseNet y LSTM jerárquico con atención que permite generar una secuencia de vectores. A diferencia de otros modelos jerárquicos con atención, utilizan *Reinforcement learning* (RL) [96] para mejorar sus resultados. Además, el incluir un mecanismo de atención en la LSTM, permite identificar las características relevantes mediante mapas de calor en la imagen.

La arquitectura propuesta por Xue [30] consiste en un modelo de tipo codificador - decodificador. La parte del codificador es una red ResNet-152 pre-entrenada que recibe como entrada dos imágenes de RX y genera una primera sentencia, la cual se conecta al decodificador (Bi-LSTM) para obtener una representación semántica. Luego, ambas representaciones se combinan y se representan en un solo vector. El vector obtenido será la entrada a la RNN multimodal que genera la próxima sentencia. Este proceso se repite hasta generar todo el texto del reporte.

Un mecanismo multi-atención que potencia el mapeo de las oraciones a la representación de las características de la imagen en las tareas de generación de párrafos es el propuesto por Huang et al. [19]. Este modelo utiliza una red ResNet y una estructura jerárquica LSTM de tres sub-módulos. El primer módulo se encarga de generar oraciones. El segundo módulo fusiona la información de los antecedentes del paciente. El último módulo es utilizado para la generación de palabras de acuerdo a cada vector de tema.

El enfoque propuesto por Jing [20], denominado *Co-operative Multi-Agent System*, basa su arquitectura en una red ResNet pre-entrenada, un módulo de *findings* y otro módulo de *impressions*. La red ResNet analiza y extrae las características de la imagen; posteriormente, el módulo de *findings* examina diferentes áreas de la imagen y genera una descripción de lo encontrado. Cuando el módulo de *findings* termina, el módulo de *impressions* genera las conclusiones de acuerdo a los hallazgos encontrados. Para generar el texto adecuadamente,

los autores utilizan 3 agentes jerárquicos compuestos por LSTMs denominados: *Planner* (PL), *Normality Writer* (NW) y *Abnormality Writer*(AW). El primer agente PL, identifica si el área contiene alguna anormalidad y envía la información al agente correspondiente. Los agentes NW y AW reciben la información y realizan la descripción asociada a la imagen.

El modelo propuesto por Tian et al. [29] utiliza una arquitectura jerárquica de BiLSTMs, Auto-econders (AE) y un mecanismo de atención que permite la generación de sentencias estructuradas. El mecanismo de atención relaciona las características visuales y los *embeddings*, adquiriendo predicciones precisas de las secciones de *findings* e *impressions*. La extracción de las características visuales se obtiene mediante una CNN que contiene 10 capas de convolución apiladas de manera secuencial. Además, la CNN comparte las características relevantes de la imagen con el mecanismo de atención que asocia al texto con la imagen. Para la generación del reporte médico, primero se forma una representación del tema a través de un BiLSTM y posteriormente se decodifica por un AE que produce la sentencia del tema, palabra por palabra.

D. Modelos Híbridos

Los modelos híbridos se caracterizan por generar el texto utilizando auxiliares de plantillas o prefijos para autocompletar el texto generado, Fig 2d. Li, et al. [16], por ejemplo, propusieron un modelo capaz de generar un reporte médico automáticamente mediante una plantilla pre-definida y el texto con las descripciones de la imagen. La estrategia incluye un modelo DenseNet y una RNN con un mecanismo de atención que permite mejorar la generación de texto.

El enfoque denominado *Knowledge-driven Encode, Retrieve, Paraphrase* (KERP), propuesto por Li, et al. [44], se caracteriza por generar el texto basándose en una plantilla pre-definida. La arquitectura es una red DenseNet y un *Graph Transformer* (GTR) el cual genera una representación gráfica de la información visual en términos de conceptos médicos y sus relaciones. Mediante un módulo de recuperación se decodifica la representación gráfica, la cual se genera como una plantilla de secuencia de palabras, las cuales posteriormente serán predicciones para la generación del reporte.

CLinicAl Report Auto-completion (CLARA) es un método propuesto por Biswal, et al. [43], el cual genera el reporte médico sentencia por sentencia utilizando términos clínicos o sentencias parciales definidas en el área médica. Su arquitectura se conforma por 4 módulos. El primer módulo recibe una entrada de una imagen de RX por medio de una arquitectura DenseNet pre-entrenada con ChestX-ray8 y obtiene la representación de sus características. Paralelamente el segundo módulo crea un repositorio que recibe como entrada los reportes médicos y selecciona diversas sentencias, su representación y frecuencia. El tercer módulo permite controlar si el informe es generado automáticamente o de manera interactiva, donde el especialista puede manipular la información del reporte utilizando palabras clave definidas o editando el texto.

Finalmente, el cuarto módulo extrae las sentencias más relevantes de los prototipos del repositorio, y envía la información al modelo secuencial que permite modificar la sentencia de acuerdo a la representación de la entrada, las palabras clave y el texto predefinido. Para la generación de texto se utilizó un modelo secuencial constituido por dos capas BiLSTM y 3 capas apiladas de LSTM.

E. Modelos Mixtos

En esta clasificación se distinguen arquitecturas muy particulares y diferentes a las anteriores orientadas a la generación de reportes médicos mediante imágenes de RX. Debido a la gran diversidad de las arquitecturas incluidas en esta clasificación, en la Fig. 2e solo se muestra el modelo propuesto por Spinks [80], en donde se propone una metodología compuesta por dos fases, la fase de entrenamiento y la fase de inferencia. En la fase de entrenamiento, el modelo de *Adversarially Regularized Autoencoders* (ARAE) [97] aprende los WE conformados por la concatenación del diagnóstico y etiqueta correspondientes a la imagen de RX asociada. Una vez que se obtienen estas representaciones, se entrena el modelo StackGAN [98] basado en una *Generative Adversarial Networks* (GAN) [99]. El propósito de esta GAN es recibir los WE obtenidos en el paso anterior y la imagen de RX correspondiente al diagnóstico, para aprender a generar una imagen RX con las distribuciones correctas de acuerdo a la entrada de texto. Finalmente, se entrena el modelo de CNN para aprender a realizar el proceso inverso de imagen a texto. En la fase de inferencia se trabaja con imágenes de RX reales y se obtiene su diagnóstico; además, de acuerdo al diagnóstico producido, se tiene la opción de recrear la imagen de RX para validar el resultado del modelo.

Por otra parte, Xiong [45] propusieron un enfoque utilizando la red DenseNet previamente entrenada, que asocia un *bounding box* con una clase predefinida utilizando el algoritmo de Grad-CAM [100]. Luego, un modelo basado en *Transformers* [101], constituido por 3 sub-capas apiladas de módulos con atención propia y un módulo *feed forward*, genera el reporte médico.

Li [48] propone una arquitectura que clasifica y localiza diferentes enfermedades en una imagen de RX de tórax; posteriormente, genera las sentencias correspondientes a la clasificación determinada. Para la tarea de clasificación se utilizó el modelo DenseNet pre-entrenado con ChestX-ray8. Para la tarea de detección se utilizó un modelo ResNet pre-entrenado con Imagenet y se aplicó la estrategia de Grand-GAMs [100]. La generación del reporte se realiza con diferentes LSTMs, donde para cada clase de enfermedad se especifica un par de LSTMs con la finalidad de obtener una mayor congruencia en el texto generado.

A su vez Gajbhiye [51] introduce un enfoque denominado *Multilevel Multi-Attention based encoder-decoder*, el cual combina la atención visual y la atención textual para la generación de reportes médicos. El modelo se conforma por una red VGG y un LSTM. Primero se codifica el contexto visual y textual, posteriormente son fusionados para alimentar un LSTM que cuenta con un mecanismo de atención. Paralelamente, el WE con la información del reporte alimenta un BiLSTM con atención. Después la salida del Bi-LSTM y la

salida del LSTM son concatenadas para realizar una predicción y generar el reporte médico palabra por palabra.

Zhang, et al. [46] propusieron una arquitectura basada en una Red Convolucional Gráfica (GCN). Mediante una DenseNet pre-entrenada con CheXpert se extraen las características de las imágenes. Posteriormente, las características alimentan a la GCN mediante un mecanismo de atención. Del GCN se especifican dos rutas, una para la clasificación de las características encontradas en la imagen y otra para la generación del reporte. Para la generación de reporte se utilizaron dos niveles de LSTM, uno para la generación del tema y otro para la generación de palabras. Los autores propusieron una nueva métrica denominada *Medical Image Report Quality Index* (MIRQI), que permite medir la exactitud de las enfermedades positiva o negativamente así como los atributos mencionados en el reporte médico.

La arquitectura propuesta por Alfarghaly et al. [102] está conformada por una red de tipo codificador - decodificador, donde la parte del codificador es la red *ChexNet* [79] y el decodificador es el modelo pre-entrenado de *DistilBERT*, propuesto por [103].

La Tabla III muestra los trabajos antes descritos indicando la base de datos que emplearon, las arquitecturas de los modelos implementados, la tarea principal que se realiza, los resultados que reportan resaltando los modelos que obtuvieron el mejor puntaje, y a manera de resumen, las principales observaciones que se identificaron en su análisis.

VII. Discusiones

Analizando los resultados de la Tabla III, una de las limitaciones que se identificó relacionada con el desempeño de los modelos, es la escasa cantidad de datos con los que se entrenan. Si consideramos tareas más sencillas como clasificación de imágenes naturales, se tiene la base de datos ImageNet, la cual cuenta con millones de imágenes que deben ser clasificadas de manera automática [39]. La base de datos más utilizada para generar reportes médicos cuenta solamente con 3,955 reportes, lo cual puede considerarse como una cantidad baja, aún y cuando se implementen estrategias de aumento de datos. Aunado a esto, existe una variabilidad entre los textos de los reportes para una misma enfermedad, lo que ocasiona una amplia dispersión entre los datos.

Otra limitación a considerar es la carencia de métricas orientadas a medir la calidad del reporte generado que sean validadas por expertos médicos. Es decir, queda una interrogante de si es mejor aquel modelo que logra buenos resultados con las métricas de BLEU, que aquel que alcanza buen desempeño con METEOR ya que este último considera sinónimos. No es claro si para un experto médico todas estas métricas deben ser ponderadas con igual peso o hay alguna que de manera particular refleje mejor la calidad del reporte generado.

De acuerdo a las Tablas I y III, el 73 % de los modelos analizados en este trabajo utilizan la base de datos IU X-Ray. Considerando estos modelos, se infiere que los modelos jerárquicos son los que obtienen un desempeño superior comparado con el promedio de los valores reportados por los otros modelos. El 50 % los modelos jerárquicos obtienen un desempeño superior a la media con BLEU1, un 63 % con BLEU2, un 75 % con BLEU3, un 71 % con BLEU4, 75 % con METEOR, un 80 % con ROUGE y un 50 % con CIDEr. El modelo con el mejor desempeño es el propuesto por Tian et al. [29]. La estrategia de utilizar una arquitectura jerárquica de BiLSTMs y AEs, en conjunto con un mecanismo de atención para relacionar las características visuales y los *embeddings*, logró muy buenos resultados. La arquitectura de CNN que utilizaron incluye varias capas de convolución apiladas y aplica estrategias de escalamiento para el aumento de datos. En relación al pre-procesamiento en texto, los caracteres fueron convertidos a minúscula y se eliminaron los caracteres no alfabéticos aplicando un filtrado de palabras basado en su ocurrencia. Al parecer, los modelos que procesan y analizan el texto asociado a la imagen tienen una mayor complejidad en comparación con los que extraen características visuales.

Por otro lado, los modelos de Yuan et al. [50] y Jing et al. [6] mantienen buen desempeño en todos los resultados de BLEU. Ambos modelos emplean una estructura jerárquica en conjunto con un mecanismo de atención. Además, el mecanismo de co-atención que implementa [6], permite asociar las características visuales extraídas de la red VGG con la representación del contexto. A esto se atribuye la generación de sentencias y/o párrafos bien estructurados. Para el pre-procesamiento de texto, las palabras se convirtieron a minúsculas y los caracteres no alfanuméricos se eliminaron. De igual manera, la estructura jerárquica de LSTMs con atención de Yuan et al. [50], en conjunto con la red pre-entrenada ResNet, permitió generar secuencias bien estructuradas con los datos de IU X-Ray.

En cuanto a los resultados con METEOR (M) y ROUGE, se observó una relación directa con BLEU. Con la métrica de CIDEr no se observó una tendencia clara ya que algunos autores reportan resultados muy altos en CIDEr pero bajas en BLEU [28], mientras que otros trabajos con buenos resultados en BLEU reportan bajos desempeños con CIDEr [6]. CIDEr agrupa palabras que tengan la misma raíz o *stemming* en una sola, posteriormente calculan su peso de acuerdo a su frecuencia y relevancia por medio del valor TF-IDF. Tal vez se tendría que revisar a detalle la manera en cómo se forman esas agrupaciones de palabras, lo cual pudiera explicar las variaciones en los resultados.

Es complejo identificar las causas de porqué los modelos propuestos no han logrado obtener mejores puntajes en las métricas de evaluación. La diversidad de arquitecturas y estrategias utilizadas en cada una de las propuestas puede arrojar factores muy particulares. Sin embargo, se ha identificado que el uso de WE no contextualizados como Word2vec, GloVe o fastText para alimentar al modelo de lenguaje contribuye a generar representaciones más pobres de las palabras sin identificar homónimos. Finalmente, no contar con etiquetas establecidas y estandarizadas para validar el significado del texto contribuye a aumentar la brecha para alcanzar puntuaciones óptimas en las métricas de evaluación.

VIII. Conclusiones

Aún y cuando los modelos de DL para la generación de reportes médicos es un área que se ha trabajado desde hace

TABLA III
MODELOS DE DL PROPUESTOS PARA LA GENERACIÓN DE REPORTES MÉDICOS MEDIANTE IMÁGENES DE RADIOLOGÍA

Referencia	Bases de datos	Arquitectura	Tarea	Accuracy	AUC	BLEU-1	BLEU-2	BLEU-3	BLEU-4	M	ROUGE	CIDEr	Observaciones
Rajpurkar et. al [79]	ChestX-ray14	CheXNet	Clasifica imagen	0.435	-	-	-	-	-	-	-	-	Produce mapas de calor No genera reporte
Dong et. al. [26]	PACs	K-medoids	Extrae etiqueta										BD privada
		VGG-16	Imagen normal o anormal	0.820	-	-	-	-	-	-	-	-	No genera reporte
		ResNet	Clas. de enfermedades	0.829	-	-	-	-	-	-	-	-	No genera reporte
Rubin et. al [25]	MIMIC-CXR	NeoBio	Genera etiqueta										No genera reporte
		DenseNet-121	Relaciona Imágenes/texto Clasifica imagen	-	0.721/0.668	-	-	-	-	-	-	-	Vista frontal / lateral
Hasan et. al. [15]	ImageCLEF CAPTION 2017	VGG-19-LSTM	Extraer características imagen Relaciona texto/imagen Genera una sentencia	0.1208	-	0.3211	-	-	-	-	-	-	Usa mecanismos de atención
Shin et.al. [14]	IU X-Ray	GoogleNet-GRU	Genera etiqueta Relaciona imágenes/ etiqueta Describe contexto	0.698	-	0.785	0.144	0.047	0.0	-	-	-	Detecta enfermedad Genera sentencia Usa sección *findings*
Zhang et. al. [18]	BCDIR	ResNet-LSTM	Extrae características imagen Relaciona imagen/texto Genera sentencia	-	-	**0.912**	0.829	0.75	0.677	0.396	0.701	0.204	BD privada y pequeña Posible sobre-entrenamiento
Wang et al. [17]	ChestX-ray14	ResNet-LSTM	Clasifica imagen	-	0.748	0.286	0.159	0.103	0.073	-	0.107	0.226	Texto no estructurado
	IU-X-Ray	ResNet-LSTM	Relaciona texto/imagen Genera múltiples sentencias	-	0.798	-	-	-	-	-	-	-	Usa mecanismo de atención Texto no estructurado Usa sección *findings*
Gasimova et al. [27]	IU X-Ray	ResNet-LSTM	Clasifica imagen	**0.994**	-	0.667	0.471	0.268	0.159	-	-	-	Evalúa reporte
		VGG-LSTM	Relaciona texto/imagen Genera una sentencia	-	-	0.069	0.023	0.07	0.01	-	-	-	
Gu et al. [47]	PACs-3	ResNet-LSTM	Clasifica imagen Relaciona texto/imagen Genera múltiples sentencias	-	-	0.762	0.716	0.681	0.651	**0.436**	0.490	0.809	Texto no estructurado Evalúa reporte
Singh et al. [55]	IU X-Ray	InceptionV3-LSTM	Extraer características imagen Relaciona texto/imagen Genera múltiples sentencias	-	-	0.374	0.224	0.152	0.109	0.163	0.307	0.359	Texto no estructurado Usa sección *findings+impressions*
Jing et. al. [6]	IU X-Ray	VGG-LSTM-	Extraer características imagen	-	-	0.517	0.386	0.306	0.247	0.217	0.447	0.327	Modelo jerárquico/atención
	PEIR Digital Library	MLC	Predicción de etiquetas Genera múltiples sentencias	-	-	0.300	0.218	0.165	0.113	0.149	0.279	0.329	Usa sección *findings + impression* Texto no estructurado
Xue et al. [49]	IU X-Ray	ResNet-BiLSTM	Extraer características imagen Relaciona texto/imagen Genera múltiples sentencias	-	-	0.464	0.358	0.270	0.195	0.274	0.366	0.343	Modelo recurrente con atención. Texto no estructurado Usa sección *findings*
Harzig et al. [22]	IU X-Ray	ResNet-LSTM	Clasifica imagen Relaciona texto/imagen Genera múltiples sentencias	-	-	0.373	0.246	0.175	0.126	0.163	0.315	0.359	Genera texto jerárquicamente Texto no estructurado Usa sección *findings*
Xie et al. [54]	IU X-Ray	CNN-LSTM	Clasifica imagen Relaciona texto/imagen Genera múltiples sentencias	0.132	-	0.443	0.337	0.236	0.181	-	0.347	0.374	Genera texto jerárquicamente Texto no estructurado Usa sección *findings+impressions*
Yin et al. [31]	IU X-Ray	DensNet-LSTM	Clasifica imagen Relaciona texto/imagen Genera múltiples sentencias	-	-	0.445	0.292	0.201	0.154	0.175	0.344	0.342	Genera texto jerárquicamente Usa mecanismo de atención Usa sección *findings+impressions*
Yuan et al. [50]	IU X-Ray	ResNet-LSTM	Clasifica imagen Relaciona texto/imagen Genera múltiples sentencias	-	-	0.529	0.372	0.315	0.255	0.343	0.453	-	Genera texto jerárquicamente Genera texto no estructurado Usa sección *findings+impressions* Usa mecanismo de atención
Liu et al. [28]	IU X-Ray	DensNet-LSTM	Clasifica imagen	0.916	-	0.369	0.246	0.171	0.115	-	0.359	1.490	Genera texto Jerárquicamente
	MIMIC-CXR	DensNet-LSTM	Relaciona texto/imagen Genera múltiples sentencias	0.834 -	- -	0.352	0.223	0.153	0.104	-	0.307	1.153 -	Texto estructurado Usa sección *findings*
Xue et al. [30]	IU X-Ray	ResNet 152-BiLSTM	Extraer características imagen Relaciona con texto Mapas de calor en imagen	-	-	0.489	0.340	0.252	0.195	0.230	0.478	0.565	Texto no estructurado Usa mecanismo de atención Usa sección *findings*
Huang et al. [19]	IU X-Ray	ResNet-LSTM	Clasifica imagen Clasifica imagen Genera múltiples sentencias	-	-	0.476	0.340	0.238	0.169	-	0.347	0.297	Texto no estructurado Usa sección *findings + impressions* Usa mecanismo de atención.
Jing et al. [20]	IU X-Ray	ResNet-LSTM	Clasifica imagen	-	-	0.464	0.301	0.210	0.154	-	0.362	0.275	Texto no estructurado.
	CX-CHR		Relaciona imagen/texto Genera múltiples sentencias	-	-	0.693	0.626	0.580	0.545	-	0.661	2.900	Usa sección *findings* Usa mecanismo de atención.
Tian et al. [29]	IU X-Ray	CNN-BiLSTM-AE	Clasifica imagen Genera múltiples sentencias	-	-	0.882	**0.874**	**0.867**	**0.860**	-	**0.929**	-	Genera texto jerárquicamente. Texto estructurado Usa sección *findings + impressions*. Usa mecanismo de atención.
Li et. al. [16]	IU X-Ray	DenseNet-RNN	Extraer características imagen	-	-	0.438	0.298	0.208	0.151	-	0.322	0.343	Híbrido plantilla/texto.
	CX-CHR	DenseNet-RNN	Relaciona texto/imagen Genera texto	-	-	0.673	0.587	0.530	0.486	-	0.612	2.895	Usa sección *findings*
Li et al. [44]	IU X-Ray	DenseNet-GTR	Clasifica imagen	-	-	0.482	0.325	0.226	0.162	-	0.339	0.280	Modelo híbrido
	CX-CHR	DenseNet-GTR	Genera texto	-	-	0.673	0.588	0.532	0.473	-	0.618	2.850	Usa plantilla. Usa sección *findings*. Usa mecanismo atención
Biswal et al. [43]	IU X-Ray	DenseNet-BiLSTM-LSTM	Clasifica imagen Genera texto	0.871	0.796	0.471	0.324	0.214	0.199	-	-	0.359	Utiliza prefijos. Usa sección *findings* Usa plantilla
Spinks et al. [80]	PACs-3	GAN-ARAE-CNN	Texto/ imagen	0.906	**0.948**	0.490	0.350	0.250	0.180	0.270	0.400	0.600	Usa mecanismo de atención
Xiong et al. [45]	IU X-Ray	DenseNet-Transformer	Clasifica imagen Relaciona imagen/texto Genera múltiples sentencias	-	-	0.350	0.234	0.143	0.096	-	-	0.323	Texto no estructurado. Usa sección *findings* Usa mecanismo de atención
Li et al. [48]	IU X-Ray	DenseNet-LSTM	Clasifica imagen Genera múltiples sentencias	-	-	0.419	0.280	0.201	0.150	-	0.371	0.553	Texto no estructurado. Usa sección *findings + impressions* Usa mecanismo de atención.
Gajbhiye et al. [51]	IU X-Ray	VGG-BiLSTM	Clasifica imagen Genera múltiples sentencias	-	-	0.500	0.380	0.317	0.278	0.281	0.440	1.067	Usa no estructurado. Usa sección *findings + impressions* Usa mecanismo de atención.
Zhang et al. [46]	IU X-Ray	DenseNet-GCN-LSTM	Clasifica imagen Genera múltiples sentencias	-	-	0.441	0.291	0.203	0.147	-	0.367	0.304	Texto no estructurado. Usa sección *findings+impressions* Usa mecanismo de atención
Alfarghaly et al. [102]	IU X-Ray	ChexNet-Transformer	Extraer características imagen Relaciona Semántica Genera múltiples sentencias	-	-	0.387	0.245	0.166	0.111	0.164	0.289	0.257	Texto no estructurado Usa sección *findings+impressions* Usa mecanismo de atención

algunos años, hay mucho por hacer. Analizando a detalle los modelos que lograron mejor desempeño, el componente de mayor complejidad en su definición es el encargado de analizar el texto. Sin embargo, algunos modelos que reportaron un pre-entrenamiento en sus modelos de CNN, reportaron mejor calidad en las características visuales extraídas de acuerdo a los mapas de calor generados. Por otra parte, utilizar modelos jerárquicos en la parte de LSTM (o BiLSTMs), así como modelos de atención y/o co-atención es un factor común en los modelos con mejores resultados. De igual manera, la parte de pre-procesamiento en el texto relacionada con la generación de WE es fundamental para el buen desempeño del modelo.

En cuanto a las bases de datos, la mayoría de los trabajos que se revisaron utilizan la base de datos de IU X-Ray para validar sus propuestas. A la fecha, se han propuesto nuevas y más complejas bases de datos, tales como MIMIC-CRX que contiene una gran cantidad de datos o Eye-Gaze que incluye información adicional del diagnóstico. Sin embargo, existe una carencia en colecciones de datos en español.

Al realizar este trabajo se observó que es necesario analizar y tal vez replantear las métricas que miden la calidad en el reporte generado, pues existe una gran redundancia entre ellas. Por ejemplo, se identificó que si el modelo reporta buenos resultados con BLEU1, BLEU2, BLEU3 y BLEU4, lo hará también con M y ROUGE. En el caso de CIDEr, su comportamiento es diferente. Además, se debe tener cuidado en que todos los modelos utilicen la misma definición de variables en sus métricas. Aunque en este trabajo no se profundizó en la complejidad computacional de los modelos y/o tiempo de entrenamiento, contar con equipo altamente especializado seguramente dará una gran ventaja ante aquellos que no lo usen por tratarse de modelos de DL.

REFERENCIAS

[1] A. D. Orjuela-Cañon and O. Perdomo, "Clustering proposal support for the covid-19 making decision process in a data demanding scenario," *IEEE LAT AM T*, vol. 19, p. 1041–1049, jun. 2021.

[2] A. Radiology, "Acr recommendations for the use of chest radiography and computed tomography (ct) for suspected covid-19. infection," *ACR website*, 2020.

[3] S. Yang, J. Niu, J. Wu, Y. Wang, X. Liu, and Q. Li, "Automatic ultrasound image report generation with adaptive multimodal attention mechanism," *Neurocomputing*, vol. 427, pp. 40–49, 2021.

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

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

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

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

[8] X. Liu, Y. Zhou, and Z. Wang, "Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 1–15, 4 2019.

[9] M. I. Neuman, E. Y. Lee, S. Bixby, S. Diperna, J. Hellinger, R. Markowitz, S. Servaes, M. C. Monuteaux, and S. S. Shah, "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children," *J HOSP MED*, vol. 7, pp. 294–298, 4 2012.

[10] R. M. Hopstaken, T. Witbraad, J. M. van Engelshoven, and G. J. Dinant, "Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections," *Clinical Radiology*, vol. 59, pp. 743–752, 8 2004.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 12 2017.

[12] I. Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz, "Automatic caption generation for medical images," in *Proceedings of the 3rd International Conference on Smart City Applications*, SCA '18, (New York, NY, USA), Association for Computing Machinery, 2018.

[13] B. Pandey, D. Kumar Pandey, B. Pratap Mishra, and W. Rhmann, "A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions," *Journal of King Saud University - Computer and Information Sciences*, 2021.

[14] H. C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2497–2506, IEEE Computer Society, 12 2016.

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

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

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

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

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

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

[21] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, pp. 2048–2057, International Machine Learning Society (IMLS), 2 2015.

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

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

[24] W. Xiaosong, L. Le, S. Hoo-chang, K. Lauren, N. Isabella, Y. Jianhua, and S. Ronald, "Unsupervised category discovery via looped deep pseudo-task optimization using a large scale radiology image database," *arXiv*, 2016.

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

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

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

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

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

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

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

[32] X. Liu, K. Gao, B. Liu, C. Pan, K. Liang, L. Yan, J. Ma, F. He, S. Zhang, S. Pan, and Y. Yu, "Advances in Deep Learning-Based Medical Image Analysis," *Health Data Science*, vol. 2021, pp. 1–14, 6 2021.

[33] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, pp. 5455–5516, 12 2020.

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

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

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

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

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

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

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2015.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[56] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 9 2015.

[57] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Association for Computational Linguistics (ACL), 8 2015.

[58] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, 9 2017.

[59] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

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

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

[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[63] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[64] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

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

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

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

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

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

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

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

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

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

[74] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, pp. 1–8, 12 2019.

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

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

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

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

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

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

[81] D. Demner-Fushman, M. Kohli, M. Rosenman, S. E. Shooshan, L. M. Rodriguez, S. Antani, G. Thoma, and C. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association : JAMIA*, vol. 23 2, pp. 304–10, 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

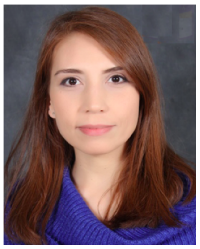
[99] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, p. 139–144, Oct. 2020.

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

[101] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 5998–6008, Curran Associates, Inc., 2017.

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

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



Olanda Prieto-Ordaz Obtuvo su grado de Ingeniero en Sistemas computacionales y Maestría en Administración en la Universidad Autónoma de Chihuahua (UACH). Trabajó en la industria privada como de Ingeniero de Soporte y Aplicaciones en el área de sistemas durante 8 años. Actualmente es catedrático en la Facultad de Ingeniería (UACH) y cursa el programa de Doctorado. Sus áreas de interés abarcan Ingeniería de Software, Aprendizaje Máquina y Visión por computadora.

Graciela Ramírez-Alonso Obtuvo el grado de Doctor en Ciencias en Ingeniería Electrónica, 2015 en el Instituto Tecnológico de Chihuahua, Chih., México. Actualmente trabaja como Profesor de Tiempo Completo en la Facultad de Ingeniería de la UACH. Su investigación actual incluye modelos de procesamiento de imagen y procesamiento digital de señales aplicando algoritmos de aprendizaje profundo. Actualmente es miembro del SNI en México.

Manuel Montes-Y-Gómez Manuel Montes-y-Gómez es investigador del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), de México. Su investigación se centra en las tecnologías del lenguaje, área en la que ha publicado más de 250 artículos en revistas y conferencias internacionales. Ha sido profesor invitado en la Universidad Politécnica de Valencia, la Universidad de Génova, y Universidad de Alabama en Birmingham. Es miembro de la Academia Mexicana de Ciencias, de la Sociedad Mexicana de Inteligencia Artificial, de la Academia Mexicana de Computación, y miembro fundador de la Asociación Mexicana de Procesamiento del Lenguaje Natural.

Roberto López-Santillán Es Doctor en Ingeniería por la UACH. Trabajó como desarrollador de software senior y administrador de bases de datos en el sector privado por 10 años. Actualmente es miembro del SNI en México. Sus áreas de interés incluyen análisis y diseño de algoritmos, aprendizaje máquina, aprendizaje profundo, procesamiento de lenguaje natural y programación genética.