

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO - CHILE



**“DETECCIÓN ASISTIDA DE ANOMALÍAS EN
TOMOGRAFÍAS TORÁCICAS 3D UTILIZANDO
APRENDIZAJE DE MÁQUINAS CON
MECANISMOS DE ATENCIÓN INCORPORADOS”**

DIEGO IVÁN GALLEGOS PULQUILLANCA

**MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO
CIVIL ELECTRÓNICO MENCIÓN COMPUTADORES**

PROFESOR GUIA:

MAURICIO ARAYA

PROFESOR CORREFERENTE:

NICOLÁS TORRES

Diciembre de 2023

Agradecimientos

Agradezco a mi familia, en especial a madre y a mi abuela, quienes han sido un pilar fundamental en mi vida. A mis amigos Lucas, Ankatu y Misael, quienes han sido un apoyo incondicional en estos años de universidad. A mi profesor referente, por su paciencia y guía en este trabajo. Agradezco también a la universidad y en especial al departamento de electrónica, por las herramientas y conocimientos brindados.

A mi madre, por todo su esfuerzo y amor que me han hecho la persona que soy, y a mis tías Julia y Sandra, que llevo siempre dentro mío.

Resumen

Los transformers son grandes modelos de Deep Learning (DL) ampliamente utilizados en la actualidad, la aplicación de estos, ya sea para tareas de procesamiento de lenguaje natural (NLP), visión por computadora (CV) u otras, es una tarea sustancial, que requiere grandes esfuerzos de investigación, planeación, ejecución y validación, tanto como requisitos concretos para su buen desempeño. La disposición de maquinaria adecuada y cantidades considerables de data son fundamentales, pues está demostrada la relación entre el entrenamiento por tiempos prolongados y con grandes cantidades de data a disposición, y la capacidad de generalización adquirida por estos modelos. Pero a pesar de estas dificultades, los transformers parecen ser una solución más que destacada en muchos ámbitos, dominando en la actualidad indiscutiblemente el campo de NLP y compitiendo codo a codo con modelos de redes convolucionales (CNN) en tareas de CV.

En el marco de este trabajo, se plantea la creación de un sistema de detección de anomalías en tomografías torácicas en 3D, haciendo uso de un clasificador de aprendizaje automático que integra mecanismos de atención como elemento fundamental en su estructura. El enfoque inicial implica explorar el modelo Vision Transformer (ViT) y evaluar la transferencia de conocimiento desde el entrenamiento con imágenes naturales hacia el dominio de imágenes médicas. Este enfoque no solo se centra en abordar los desafíos técnicos relacionados con la implementación de modelos transformers en el ámbito médico, sino que también destaca la importancia de otros módulos esenciales, como la segmentación y la explicabilidad. La inclusión de estos componentes se orienta hacia la mejora integral de la detección de anomalías en tomografías torácicas 3D.

Abstract

The transformers are large Deep Learning (DL) models widely used today. The application of these, whether for NLP, CV, or other tasks, is a substantial endeavor that demands significant efforts in research, planning, execution, and validation. The availability of suitable machinery and considerable amounts of data are crucial, as the relationship between prolonged training and large amounts of data availability has been demonstrated to enhance the generalization capacity of these models. Despite these challenges, transformers seem to be a gold-standard solution in various domains, currently indisputably dominating the NLP field and competing neck-and-neck with CNN models in CV tasks.

Within the scope of this work, the creation of a 3D thoracic tomography anomaly detection system is proposed. This involves employing a machine learning classifier that integrates attention mechanisms as a fundamental element in its structure. The initial focus includes exploring the Vision Transformer (ViT) model and assessing knowledge transfer from training on natural images to the domain of medical images. This approach not only addresses technical challenges related to implementing transformer models in the medical domain but also emphasizes the significance of other essential modules, such as segmentation and explainability. The inclusion of these components aims at comprehensive enhancement in 3D thoracic tomography anomaly detection.

Glosario

CNN: Convolutional Neural Network, Red Neuronal Convolucional..

RNN: Recurrent Neural Network, Red Neuronal Recurrente..

SA: Self-Attention, Auto-Atención.

MHSA: Multi-Head Self-Attention, Auto-Atención con Múltiples Cabezas.

ViT: Vision-Transformer, Transformador de Visión.

Kernel: Núcleo.

Encoder: Codificador.

Decoder: Decodificador.

MLP: Multi Layer Perceptrón, Red de Perceptrones Multicapa.

NLP: Natural Language Processing, Procesamiento del Lenguaje Natural.

CV: Computer Vision, Visión por Computadora.

CT: Computerized Tomography, Tomografía Computarizada.

3D: Tres Dimensiones.

LSTM: Long Short-Term Memory, Memoria a Corto y Largo Alcance.

GPT: Generative Pre-Trained Transformer, Transformador Generativo Preentrenado.

ResNet: Residual Network, Red Neuronal Residual.

HU: Hounsfield Units.

SVM: Support Vector Machine, Máquina de vectores de soporte.

TP: True Positive, Verdadero Positivo.

TN: True Negative, Verdadero Negativo.

FP: False Positive, Falso Positivo.

FN: False Negative, Falso Negativo.

Índice general

<i>1.. Introducción</i>	5
1.1. Motivación y contexto	5
1.2. Planteamiento del problema y búsqueda de solución	6
<i>2.. Marco teórico y estado del arte</i>	8
2.1. Tomografías torácicas computarizadas y DICOM	8
2.2. Redes neuronales artificiales para clasificación	9
2.2.1. Redes neuronales convolucionales	10
2.2.2. Mecanismos de atención	12
2.2.3. Transformadores	18
2.2.4. Transformadores de visión	19
2.2.5. Transformadores de visión en clasificación de imágenes médicas	23
<i>3.. Sistema solución propuesto</i>	25
3.1. Discriminación de data y preprocesamiento	26
3.2. Modelo de clasificación multietiqueta	29
3.2.1. Estrategias de adaptación para Vision Transformers	29
3.3. Módulo de segmentación	32
3.4. Módulo de explicabilidad	33
<i>4.. Experimentos de validación</i>	34
4.1. Exploración de extracción de características por ViT preentrenado en imágenes médicas 2D	37
4.2. Fine-tuning de ViT preentrenado para clasificación de muestras médicas	40

4.3. Adaptación de modelo preentrenado para imágenes médicas 3D . . .	42
4.4. Análisis de resultados y comparación con benchmark	44
5.. Conclusiones y trabajo futuro	48

Índice de figuras

2.1. Diagrama del perceptrón de Rosenblatt.	10
2.2. Operación de convolución. Fuente: [7].	11
2.3. Generación de la salida y_t , dada una secuencia de entrada $(x_1, x_2, \dots, x_T)$. Fuente [9]	14
2.4. Estructura general de un modelo de tarea. Fuente [10]	15
2.5. Mecanismo interno de un modelo de atención general. Fuente [10] . .	18
2.6. Atención de producto punto escalado (izquierda). Atención multi-cabecal (derecha). Fuente [11]	20
3.1. Prototipo de sistema de asistencia en la detección de anomalías en tomografías torácicas 3D.	26
3.2. Clasificador de COVID-19 por umbral de votación de cortes. Fuente [30].	30
3.3. Clasificador basado en Swin con una capa de pooling para agregación de características [31].	31
3.4. Localización de regiones detalladas para la predicción de una sola cla- se de ViT-Base por diferentes métodos. Fuente [43].	33
4.1. Mosaico de muestras presentes en OrganCMNIST 20x20.	38
4.2. Sistemas usados en el primer experimento.	39
4.3. Mosaico de cortes en las muestras 3D presentes en NoduleMNIST3D 20x20.	42
4.4. Sistemas usados en el tercer experimento.	43

Índice de cuadros

3.1. Atributos de importancia para procesamiento de volúmenes CT. . . .	28
4.1. Características de los datasets utilizados en la parte experimental de la memoria.	37
4.2. Distribución de clases a través de los subconjuntos del dataset OrganCMNIST.	38
4.3. Detalles de las variantes de ViT [16].	40
4.4. Métricas obtenidas para los clasificadores entrenados el primer experimento.	40
4.5. Métricas obtenidas para los clasificadores del segundo experimento. .	41
4.6. Distribución de clases en el dataset NoduleMNIST3D.	43
4.7. Métricas obtenidas para los clasificadores del tercer experimento. . . .	45
4.8. Benchmark de rendimiento en OrganCMNIST con los mejores clasificadores obtenidos en el primer y segundo experimento.	46
4.9. Benchmark de rendimiento en NoduleMNIST3D con los mejores clasificadores obtenidos en el tercer experimento.	47

1. INTRODUCCIÓN

1.1. Motivación y contexto

La investigación y trabajo propuesto en esta memoria surge del desarrollo un sistema de asistencia en el diagnóstico médico de pacientes, mediante de la detección temprana de anomalías en tomografías torácicas computarizadas. Este enfoque se fundamenta en un sistema automático de clasificación multi-etiqueta. La motivación principal radica en la posibilidad de aprovechar la extensa base de datos de tomografías torácicas computarizadas recopiladas a lo largo de años de exámenes realizados por el Hospital Clínico de la Universidad de Chile.

La detección temprana y precisa de anomalías en este u otros tipos de exámenes no puede subestimarse. Históricamente, la responsabilidad de interpretar y descubrir estas señales recae en el profesional de la salud a cargo del examen, quien proporciona sus impresiones y registra los hallazgos. Esto es crucial para la correcta gestión de los pacientes en términos de pruebas diagnósticas adicionales, planificación del tratamiento y propósitos pronósticos. Sin embargo, los errores en la detección de estas anomalías son comunes, y esto puede tener consecuencias graves en la toma oportuna de decisiones frente a patologías o afecciones que el paciente pueda desarrollar. Distintas publicaciones, [1] [2], cifran un estimado de los errores en la interpretación de los radiólogos entre un 3 % a 5 %, y con un estimado de mil millones de exámenes de radiología al año, se producirían alrededor de 40 millones de errores anualmente.

Según [1], los errores de percepción, es decir, los errores producidos por la omi-

sión de anomalías en la fase de detección, pueden llegar a corresponder al 80 % de los errores cometidos por los radiólogos. Una herramienta de asistencia en la detección de anomalías podría contribuir a disminuir especialmente este tipo de errores.

Para ayudar a dar solución a esta problemática, nos dirigiremos al desarrollo y validación de una herramienta basada en aprendizaje de máquina, que incorpore los ya probados mecanismos de atención en tareas de visión por computadora. Esta herramienta sería un clasificador multi-etiqueta, que pueda reconocer patrones y clasificarlos como anomalías, para asistir así al radiólogo encargado de documentar los hallazgos del examen.

1.2. Planteamiento del problema y búsqueda de solución

Se cuenta entonces con un conjunto de datos considerable de tomografías torácicas computarizadas 3D, cuyas etiquetas de clasificación son generadas por mecanismos de NLP, a partir de los informes médicos asociados a cada muestra. El conjunto de datos se compone de CT's tanto con contraste aplicado como sin él, factor importante en el proceso de discriminación y selección de las muestras a incluir en el conjunto final a utilizar. Cada muestra es además anonimizada, para la protección de la privacidad de las personas naturales de quienes provienen los estudios médicos.

Los datos se presentan en formato de fichero DICOM, que sigue un estándar para el almacenamiento y transmisión de imágenes médicas, y asegura la integración de la imagen con la información del paciente, formando un conjunto no separable. La información de la imagen en sí debe extraerse tomando en cuenta ciertos parámetros clave, contenidos en los campos de cabecera de cada fichero, como por ejemplo; la orientación o la escala de píxeles para obtener la información de tamaño real, y luego ser procesada y manejada en un formato más acorde con las prácticas habituales en

tareas de visión por computadora.

En cuanto al modelo buscado en sí, se menciona la necesidad de un clasificador multi-etiqueta, es decir, que pueda asignar múltiples etiquetas no excluyentes entre sí a cada muestra. Para ello se consideran dos tipos de redes neuronales dominantes en el campo de clasificación, las redes neuronales convolucionales, o CNN, y los transformadores de visión, o ViT. Se explorarán las diferencias entre estos tipos de modelos más adelante, basados en estudios empíricos presentes en [3]. Pero se decanta por la utilización de modelos basados en ViT, o al menos que incorporen mecanismos de atención, pues una característica de éstos, demostrada empíricamente, es su capacidad superior en capturar el contexto global en las tareas asignadas, lo que se teoriza sería beneficioso en el caso de estudio de esta memoria, pues las muestras de exámenes CT contemplan volúmenes considerables que podrían no ser bien interpretados por el enfoque local de las CNN.

Gran parte de la bibliografía referenciada en esta memoria aborda estudios llevados a cabo en el marco de la reciente pandemia de COVID-19. La notable capacidad de contagio de este virus resultó en un colapso de los sistemas de salud a nivel global, desencadenando una intensa crisis, tanto desde el punto de vista sanitario como humanitario. Y es entonces que surgen distintos estudios, muchos utilizando datasets de CT torácicas, en un esfuerzo por emplear herramientas basadas en aprendizaje de máquinas para el diagnóstico de esta patología. Aunque predominantemente se desarrollaron clasificadores binarios, las técnicas y modelos empleados pueden extrapolarse al caso de uso que nos ocupa.

Los modelos escogidos se utilizarán en primera instancia en datasets de prueba, para ver su comportamiento respecto de otros modelos basados en convoluciones en el estado del arte.

2. MARCO TEÓRICO Y ESTADO DEL ARTE

En este capítulo, se introducen y exploran conceptos fundamentales y estudios relevantes que desempeñan un papel fundamental en el desarrollo y búsqueda de una solución al problema abordado. Se examinan de manera detallada los principios teóricos y prácticos que sirven como cimientos esenciales para la investigación.

2.1. *Tomografías torácicas computarizadas y DICOM*

Las tomografías computarizadas (CT), también conocidas como tomografías axiales computarizadas (TAC), representan una tecnología médica establecida y confiable que ha transformado profundamente el campo de la imagenología diagnóstica. Durante décadas, las CT han sido una herramienta indispensable en la práctica médica, proporcionando imágenes detalladas y precisas del interior del cuerpo humano y desempeñando un papel crucial en la detección, diagnóstico y tratamiento de diversas anomalías pulmonares y orgánicas.

En el ámbito de las anomalías pulmonares, las tomografías computarizadas torácicas se han consolidado como el estándar de referencia para la detección temprana y la caracterización de patologías como nódulos pulmonares, neumonías, tumores y enfermedades pulmonares intersticiales. La capacidad de las CT para generar imágenes tridimensionales de alta resolución del tejido pulmonar ofrece a los radiólogos y otros profesionales de la salud una base fundamental para el diagnóstico y tratamiento de estas condiciones.

El formato DICOM (Digital Imaging and Communications in Medicine) es el estándar de facto en el almacenamiento y transmisión de imágenes médicas entre dispositivos de uso médico. DICOM posee un alcance a nivel mundial y se encuentra implementado en casi todo dispositivo de radiología, cardiología y radioterapia, por lo que es natural contemplar el tratamiento de la data en este formato para cualquier aplicación referente a la imagenología médica. DICOM garantiza la interoperabilidad y la coherencia en la interpretación y el intercambio de datos de imagen, proporcionando a los profesionales de la salud acceso a información precisa y contextual que es esencial para el manejo de este y otros tipos de imágenes médicas.

2.2. Redes neuronales artificiales para clasificación

El aprendizaje de máquinas es una rama en el campo de la inteligencia artificial, basada en el uso de data y algoritmos, que se enfoca en imitar el proceso de aprendizaje de los humanos. Utilizando modelos de redes neuronales artificiales u otro tipo de modelos de aprendizaje automático, éstas generan modelos matemáticos automáticamente, sin la necesidad de programarlos explícitamente para el fin requerido, en cambio utilizan la data disponible para mejorar gradualmente sus predicciones.

En el transcurso de las décadas, las aplicaciones del aprendizaje de máquinas han avanzado de tareas rutinarias a desafíos sustanciales, como sistemas de recomendaciones complejos, vehículos auto-conducidos o detección de objetos en tiempo real. El nacimiento de las redes neuronales artificiales puede ser rastreado desde la introducción del perceptrón en 1957 por Frank Rosenblatt, Figura 2.1, que toma entradas no booleanas, calcula su suma ponderada y genera una salida por medio de una función de activación escalón, pero la capacidad de reconocer patrones era limitada y provocaron su abandono. Sin embargo, el panorama cambió con el tiempo gracias a avances en la investigación y la tecnología. La adopción de estructuras más complejas, como las

redes neuronales multicapa (MLP), permitió la formación de modelos más profundos capaces de aprender representaciones jerárquicas, mejorando sustancialmente la capacidad de clasificación. En la actualidad, las redes neuronales profundas, respaldadas por técnicas como el aprendizaje profundo y arquitecturas como las redes neuronales convolucionales (CNN) y recurrentes (RNN), han revolucionado la clasificación, demostrando un rendimiento excepcional en tareas complejas, como el reconocimiento de imágenes, el procesamiento del lenguaje natural y la clasificación de datos multidimensionales.

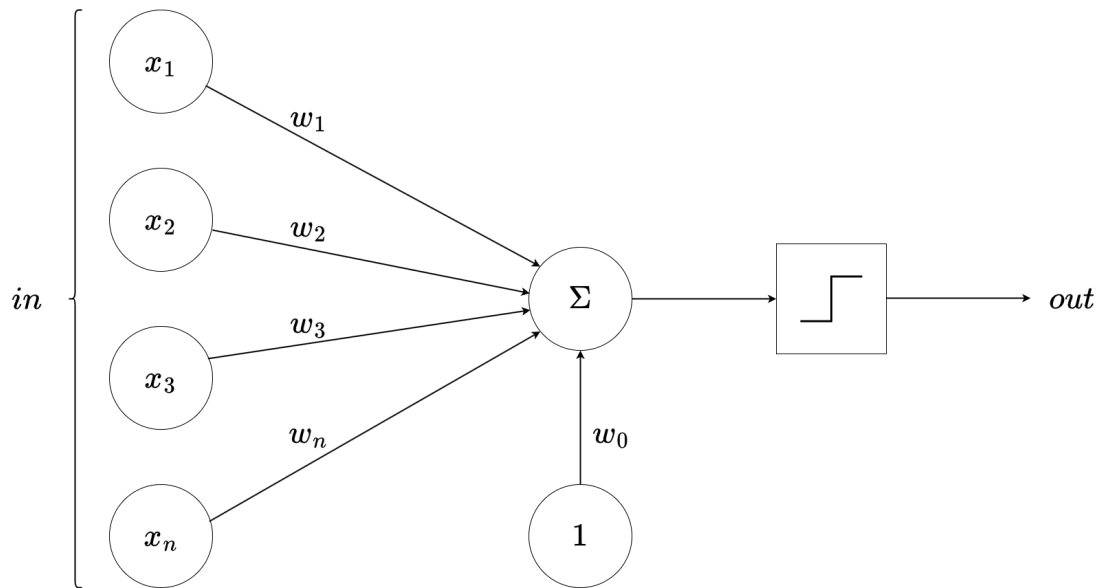


Fig. 2.1: Diagrama del perceptrón de Rosenblatt.

2.2.1. Redes neuronales convolucionales

Una red neuronal convolucional es una variación de una red neuronal tradicional, donde existe al menos una capa convolucional. En estas capas, cada neurona aplica operaciones de convolución a la entrada, por medio de un filtro, o kernel, aplicado a una ventana deslizante en la entrada para producir un mapa de activación, una ilustración de esto puede verse en la Figura 2.2.

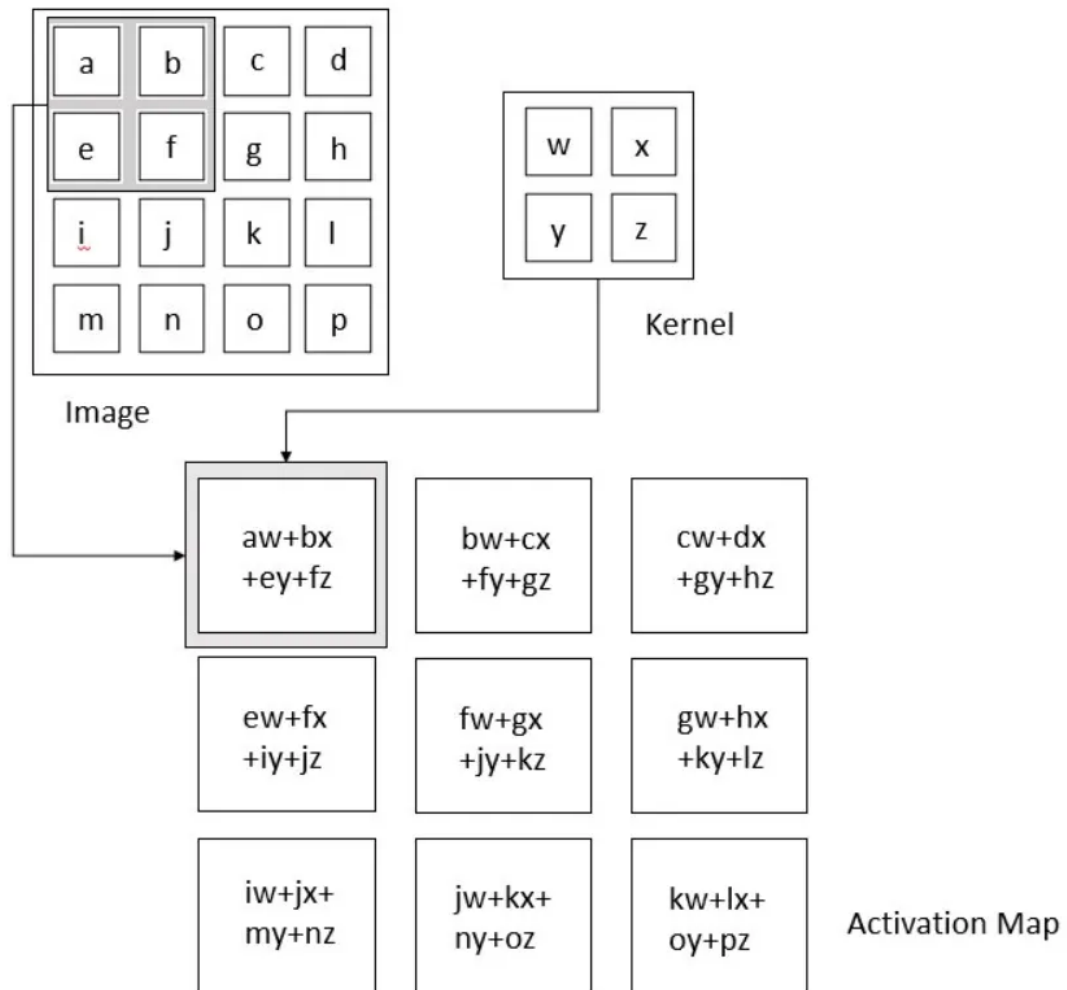


Fig. 2.2: Operación de convolución. Fuente: [7].

Una capa de pooling es aplicada luego sobre cada mapa de activación; estas capas tienen como objetivo reducir la dimensionalidad de las representaciones intermedias. El tipo de pooling más comúnmente utilizado es el max-pooling, el cual aplica la función “MAX” sobre una ventana del mapa de activación, reduciendo así su dimensionalidad. La utilización de estas capas puede generalmente reducir el tamaño original a un 25 %. Las redes neuronales convolucionales utilizadas para clasificación suelen contar con capas totalmente conectadas (fully-connected) como capas finales. Estas capas finales poseen una función de activación sigmoide, lo cual es la norma en la

clasificación multi-clase.

La aplicación de estos filtros ha demostrado ser exitoso en la obtención características locales, las que luego son combinadas y ensambladas en patrones más complejos. A medida que se procede en las capas convolucionales se obtienen representaciones cada vez más abstractas, las que luego son categorizadas por las capas fully-connected.

El concepto de redes neuronales convolucionales ha estado presente en el campo del aprendizaje de máquina desde la década de los noventa, Neocognitron de Fukushima [4] es considerado uno de los precursores de las redes convolucionales modernas, mientras que el término por el que conocemos este tipo de redes hoy en día es acuñado por Yan LeCun *et al.* en [5], ambos enfocados en el reconocimiento de escritura. Desde entonces, múltiples publicaciones han aportado al desarrollo y la evolución de redes neuronales convolucionales, pero es el establecimiento de un reto de reconocimiento de imágenes, ILSVRC (ImageNet large scale visual recognition challenge), basado en el dataset ImageNet, el que impulsa realmente el explosivo avance que han experimentado este tipo de estructuras de aprendizaje profundo.

Desde 2012, con la introducción de AlexNet [6], todos los modelos ganadores de ILSVRC han sido variaciones de redes neuronales convolucionales. Naturalmente, el dominio mostrado por las CNN en tareas de reconocimiento y clasificación de imágenes, motivan la introducción de éstas a aplicaciones médicas. Sin embargo, si bien la fortaleza en la localidad de las CNN las hace poderosas en el área, es conocido que su alcance para obtener un contexto global de la entrada no es su fuerte.

2.2.2. Mecanismos de atención

En el campo del procesamiento del lenguaje natural hemos visto grandes avances en el último tiempo, con potentes modelos, profundos y complejos, de parte de

los gigantes en el área, como Google y OpenAI Ft. Microsoft. Sin embargo, previamente, las tareas de NLP eran dominadas por otro tipo de redes, las redes neuronales recurrentes o RNN, las cuales tuvieron un impacto muy grande, en especial en el reconocimiento de lenguaje hablado (speech recognition), como puede verse en el reporte hecho por Google sobre sus avances utilizando LSTM [8].

Las RNN se caracterizan por el procesamiento secuencial, almacenando estados ocultos que son realimentados a la red para procesar las siguientes salidas y estados. Pero en tareas de secuencia-secuencia (entrada y salida de secuencia de tokens), o seq2seq, la predicción de RNNs tradicionales en realidad se basa en una sola representación codificada (salida del encoder) de toda la secuencia de entrada. Sin embargo, ciertas partes de la secuencia de entrada podrían ser más importantes para la generación de alguna parte de la secuencia de salida que otras, es aquí que nace el concepto de atención.

Los mecanismos de atención se incorporan para codificar la relevancia entre pares de tokens decoder-encoder, donde cada token del decoder posee una representación de la importancia, para sí, de todos los tokens del encoder, compuesta por una distribución de probabilidades. Éste mecanismo ataca directamente el problema de la pérdida de información producida por la compresión de la información secuencial a un único vector de estado, aunque incorpora la complejidad en el manejo de aún más parámetros en la red.

El concepto de atención en una RNN surge por primera vez por Badhanau *et al.* [9], dónde, utilizando una RNN bidireccional como encoder y una RNN como decoder, implementa un mecanismo de atención entre éstos. La arquitectura propuesta por Badhanau *et al.* puede ser vista en la Figura 2.3, donde la palabra de la secuencia de salida en el instante t depende del estado oculto s_{t-1} en el decoder en el instante $t - 1$, del vector de contexto en el instante t , c_t , de las anotaciones h_j (que reúne informa-

ción de cada palabra de la secuencia de entrada x_j al concatenar los estados ocultos obtenidos de la RNN bidireccional), los pesos $\alpha_{t,i}$ y un parámetro de atención $e_{t,i}$ que describe que tan bien s_i y h_j se combinan. Las ecuaciones 2.1, 2.2 y 2.3 describen el mecanismo de atención propuesto en [9].

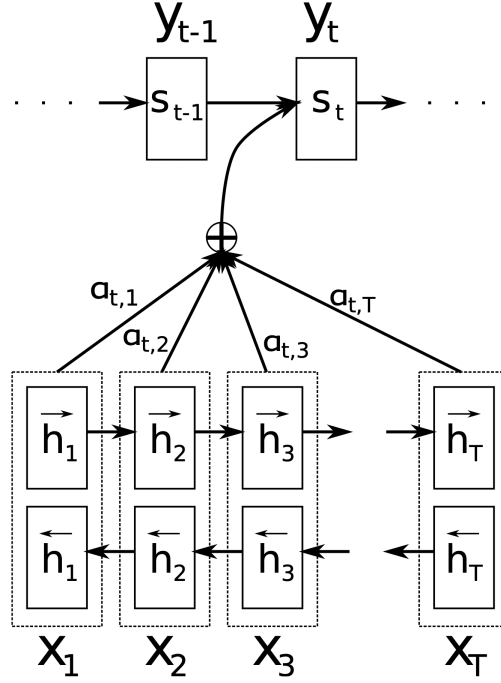


Fig. 2.3: Generación de la salida y_t , dada una secuencia de entrada $(x_1, x_2, \dots, x_T)$. Fuente [9]

$$P(y|y_1, \dots, y_{i-1}, X) = g(y_{i-1}s_i, c_i) \quad (2.1)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (2.2)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (2.3)$$

Con

$$e_{ij} = a(s_{i-1}, h_j)$$

Desde su introducción por Badhanau *et al.* [9], distintas publicaciones han refinado el concepto de atención, y rápidamente fue adoptado en distintos campos. En [10] se explican varios de los mecanismos de atención presentes en los distintos campos del aprendizaje de máquinas, contando con un modelo de atención general y notación uniforme.

Un modelo que emplea atención puede describirse como un conjunto de cuatro submódulos, el modelo en sí se denota como un modelo de tarea (*task model*), y su estructura puede verse en la Figura 2.4. Un modelo de tarea es aquel que toma una entrada, ejecuta una tarea específica y genera una salida deseada, la tarea puede referirse a traducción de texto o análisis de sentimientos, para modelos de procesamiento de language natural, generación de subtítulos o segmentación de imágenes, para modelos en visión por computadora, entre muchas otras tareas.

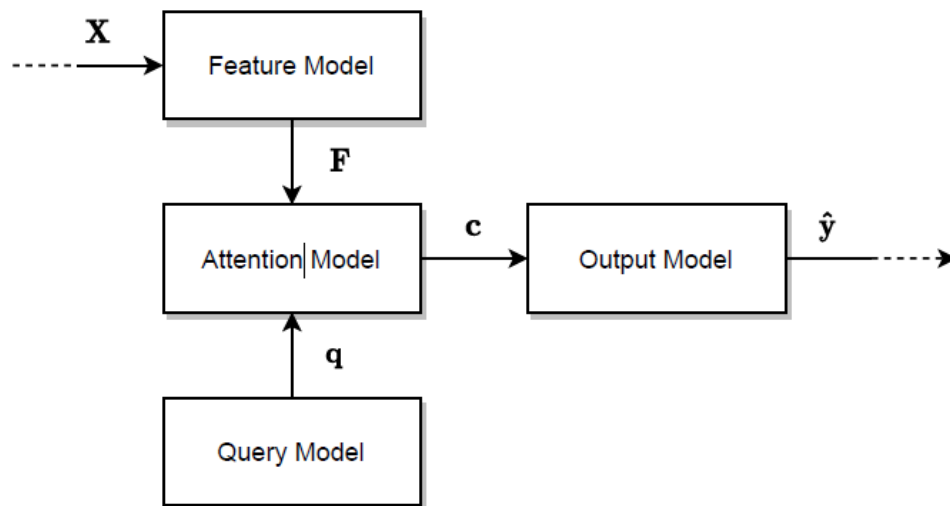


Fig. 2.4: Estructura general de un modelo de tarea. Fuente [10]

Los submódulos que componen el modelo de tarea se describen a continuación:

- **Modelo de características (*Feature model*):** El modelo de características es

utilizado para extraer n_f vectores de características $f_1, \dots, f_n \in \mathbb{R}^{d_f}$ de la entrada $X \in \mathbb{R}^{d_x \times n_x}$, donde d_x representa la dimensión de los vectores de entrada y n_x representa el número de estos vectores. El modelo de características puede ser una RNN, una CNN, una capa de embeddings, una transformación lineal de la data de entrada, o ninguna transformación en lo absoluto.

- **Modelo de consulta (*Quer y model*):** Este modelo se encarga de decirle al modelo de atención en qué vectores de características enfocarse. Para ello es necesario el vector query $q \in \mathbb{R}^{d_q}$, donde d_q es la dimensión del vector de query. Este modelo se estructura para una tarea en específico, por ejemplo, se pueden utilizar los estados ocultos de un decoder RNN como query, para conocer en cada instante t a qué vectores de características (generados por algún modelo de características) prestarle mayor atención para la generación de y_t en la secuencia de salida, que puede ser, por ejemplo, la generación de subtítulos de una imagen.
- **Modelo de atención (*Attention model*):** La Figura 2.5 presenta una vista de la estructura interna de un modelo de atención general. Este modelo toma como entradas los vectores de características $F = [f_1, \dots, f_n] \in \mathbb{R}^{d_f \times n_f}$ y el query $q \in \mathbb{R}^{d_q}$, dos matrices son generadas a partir de F , la matriz de valores $V = [v_1, \dots, v_n] \in \mathbb{R}^{d_v \times n_f}$ y la matriz de claves $K = [k_1, \dots, k_n] \in \mathbb{R}^{d_k \times n_f}$, por medio de las matrices de pesos entrenables $W_V \in \mathbb{R}^{d_v \times d_f}$ y $W_K \in \mathbb{R}^{d_k \times d_f}$. K y V deben poseer el mismo número de columnas.
La matriz de claves K y el query q son usados para obtener los puntajes de atención (*attentions scores*) por medio de una función $score()$, como por ejemplo, un producto punto, como puede verse en la ecuación 2.4, el puntaje de atención e_l representa entonces cuán importante es la información contenida en el vector de claves k_l .

$$e_l = \text{score}(q, k_l) \quad (2.4)$$

Luego, los puntajes de atención e_l son procesados para obtener un peso de atención a_l , ecuación 2.5, por medio de una capa de alineación. Como los pesos de atención pueden recorrer un rango amplio, la función de alineación se encarga de normalizarlos, una función de alineación podría ser entonces la función softmax. El vector de pesos de atención $a = [a_1, \dots, a_l]$ acarrea consigo la información de la importancia de cada vector de valor de la matriz V .

$$a_l = \text{align}(e_l; e) \quad (2.5)$$

Luego un vector de contexto c es producido realizando una suma ponderada sobre la matriz de valores V , utilizando los pesos de atención a_l para cada vector de valores v_l , según 2.6.

$$c = \sum_{l=1}^{n_f} a_l \times v_l \quad (2.6)$$

- **Modelo de salida (*Output model*):** El modelo de salida genera, como su nombre lo indica, la salida del modelo de tarea, para ello hace uso del vector de contexto c , y puede ser una gran variedad de modelos, capas MLP con una capa softmax al final es una práctica común.

Y aunque los mecanismos de atención son introducidos en arquitecturas RNN, el gran salto y éxito de éstos en realidad se puede reconocer con la introducción de los *Transformers* (Vaswani *et al.* [11]).

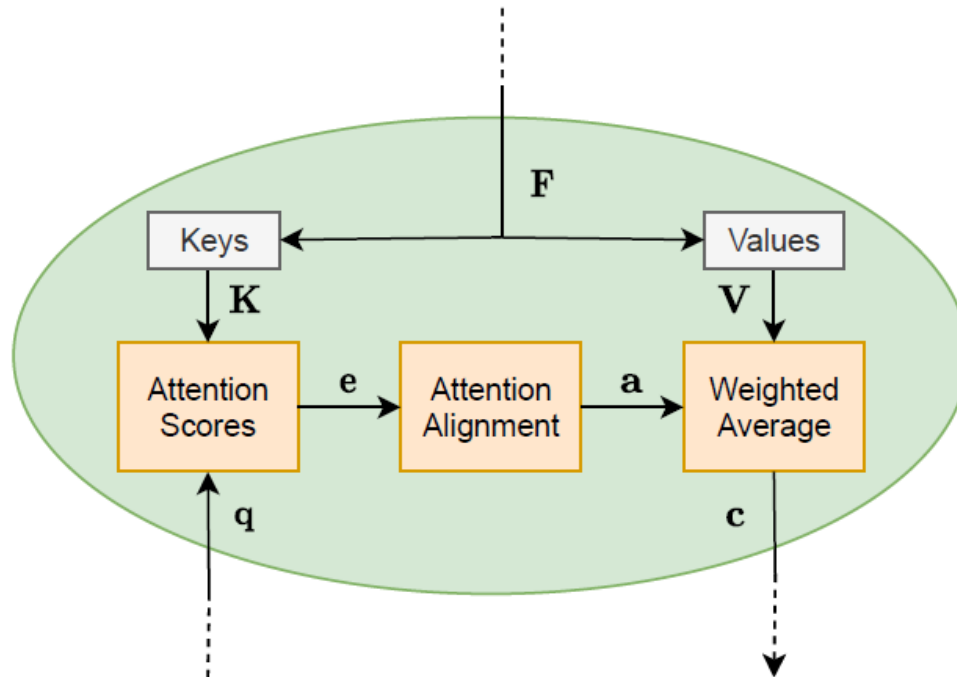


Fig. 2.5: Mecanismo interno de un modelo de atención general. Fuente [10]

2.2.3. Transformadores

Los transformadores nacen de la falta de paralización de las redes neuronales recurrentes [11], siendo estas últimas subyugadas a un procesamiento secuencial de la data. Se basan puramente en mecanismos de atención para obtener un contexto global de dependencias entre la entrada y salida secuenciales, y han tomado las riendas en la actualidad en diversas tareas, obteniendo estado del arte en muchas de ellas.

A grandes rasgos un transformer es construido usando N encoders, apilados uno sobre otro, enlazados a N decoders, también apilados uno sobre otro. Donde no hay presencia de convoluciones o recurrencia, todo lo que necesitas es la atención, para aprender representaciones e interacciones complejas en entradas secuenciales.

Los transformadores son introducidos, como ya se mencionó, por Vaswani *et al.*

[11]. BERT (Bidirectional Encoder Representations from Transformers), Devlin *et al.* [12], es un gran modelo pre-entrenado de Google que puede ser ajustado (*fine-tuned*) para diversas tareas de NLP, logrando estado del arte en 11 tareas distintas. Por otra parte, y probablemente lo que ha atraído en mayor medida la opinión pública a modelos de NLP, es la serie de modelos GPT de OpenAI ([13], [14], [15] y el más reciente GPT-4), y su producto abierto al público ChatGPT, el que ha estado en la palestra mediática desde su lanzamiento.

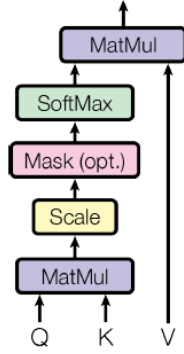
En resumen, se ha mencionado previamente el impacto significativo que los mecanismos de atención han tenido en el procesamiento seq2seq en tareas de NLP, pero la tarea que nos atañe en cuestión es la de clasificación (enfocada a muestras médicas 3D), entonces ¿Es posible aprovechar estos mecanismos de atención y posiblemente la estructura de los transformadores para la tarea de clasificación?, pues la respuesta probablemente la tengan los *Vision Transformer*.

2.2.4. Transformadores de visión

En una época donde tareas de procesamiento de imágenes en computer vision con deep learning se mantenían dominadas casi en su totalidad por redes convolucionales profundas, el éxito alcanzado por los transformers en tareas de NLP motiva su incursión en computer vision, y el concepto de *Vision Transformer* o ViT es introducido por Dosovitskiy *et al.* [16]. Los transformadores de visión han demostrado grandes habilidades, obteniendo desempeños comparables e incluso superiores a redes convolucionales en el estado del arte en diversas tareas de computer vision. Aunque modelos de CNN demuestran mejores resultados en entrenamientos desde cero [17], los transformadores de visión logran superar éstos, tanto en desempeño como en robustez, cuando se utilizan modelos pre-entrenados con grandes cantidades de data [18] [19].

En simples términos ViT es un encoder de transformador que alimenta directamen-

Scaled Dot-Product Attention



Multi-Head Attention

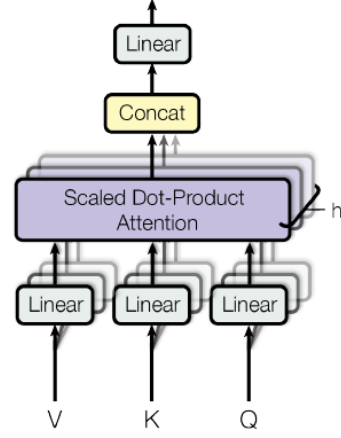


Fig. 2.6: Atención de producto punto escalado (izquierda). Atención multi-cabezal (derecha).
Fuente [11]

te un bloque MLP (Multi Layer Perceptron), del cuál se obtienen las predicciones de clase. ViT hace uso de *Multi-Head Self-Attention*, Figura 2.6, misma utilizada en [11], que es implementada utilizando producto punto, también llamada *scaled dot-product attention*, múltiples operaciones de SA son llevadas a cabo en paralelo, llamados cabezales, MHSA es descrito en [16] por las siguientes ecuaciones.

$$[q, k, v] = zU_{qkv} \quad U_{qkv} \in \mathbb{R}^{D \times 3D_h} \quad (2.7)$$

$$A = \text{softmax} \left(qk^T / \sqrt{D_h} \right) \quad A \in \mathbb{R}^{N \times N} \quad (2.8)$$

$$SA(z) = Av \quad (2.9)$$

$$MHSA = [SA_1(z), \dots, SA_k(z)]U_{msa} \quad U_{msa} \in \mathbb{R}^{kD_h \times D} \quad (2.10)$$

Donde $z \in \mathbb{R}^{N \times D}$ es el vector de entrada al módulo de Self-Attention, el cual es una proyección de una parche de la imagen (típicamente por una capa de embedding entrenable). A través de la matriz de proyección lineal U_{qkv} se obtienen los vectores de consulta q , clave k y valor v . El vector q representa la importancia relativa del

parche de características respecto de los demás, k representa la similitud del parche de características y el resto, y v contiene la información del parche de características para ser empleado en la ponderación con la matriz de atención. En las arquitecturas ViT se añade además un token [cls], que condensa toda la información de la imagen, y generalmente se utiliza el último estado oculto de este para clasificación.

A_{ij} es entonces el peso de atención calculado para cada par de elementos en la imagen, con sus respectivos q^i y k^j , con los cuales se realiza una suma ponderada sobre todos los vectores de valores. U_{mhsa} es una matriz de proyección lineal para MHSA.

ViT es capaz de capturar dependencias de largo alcance (a diferencia de las CNN que son conocidas por extraer características locales) e información contextual de la data de entrada, lo cual le permite a este modelo representar la relevancia de distintas regiones de la data de entrada.

ViT ha inspirado múltiples variantes, entre las cuales podemos destacar DeiT (Data-efficient Image Transformer) [20], que presenta además una estrategia de destilación basada en tokens en una arquitectura profesor-estudiante, que asegura que el estudiante aprenda del profesor a través de la atención.

Por otro lado Swin Transformer (Shifted Windows Transformer) [21] propone representaciones jerárquicas de las características de la entrada con respecto de las capas de encoder, además de permitir conexión (superposición) entre particiones por ventanas desplazadas, a diferencia de ViT que trabaja únicamente con particiones que no se superponen. Ambos modelos señalados han demostrado resultados destacables en clasificación con ImageNet.

Raghu *et al.* presentan un estudio interesante en [3], donde se plantean una explo-

ración de cómo resuelven los transformadores de visión las tareas relacionadas con imágenes, haciendo a su vez una comparación con las predominantes redes neuronales convolucionales en dichas tareas, y apuntando a obtener una mejor comprensión de la representación interna generada por los transformadores de visión. Las 6 ideas principales con las que contribuyen son las siguientes:

- ViT tiene mayor similitud entre las representaciones obtenidas de capas superficiales y capas profundas que CNN.
- ViT incorpora mayor información espacial que CNN en capas superficiales.
- Incorporar información espacial en capas superficiales es vital, y el preentrenamiento con grandes cantidades de data ayuda a las capas superficiales de atención a aprender esto.
- Las conexiones de salto (skip connections) son incluso más beneficiosas en ViT que en ResNet (CNN con conexiones de salto), teniendo fuerte influencia en el desempeño y la similitud de las representaciones.
- ViT mantiene mejores representaciones espaciales que ResNet.
- ViT puede aprender representaciones de gran calidad contando con grandes cantidades de data.

En cuanto al campo médico, los mecanismos de atención se han incorporado en gran medida en la segmentación de imágenes, el cual ha estado marcado desde 2015 por U-Net [22]. Hemos visto el surgimiento de modelos híbridos basados en U-Net, que contemplan tanto la clásica convolución como mecanismos de atención, como por ejemplo TransUNet [23] y UNETR [24], como también esfuerzos para aprovechar la topología en U usando puramente mecanismos de atención, como con Swin-Unet [25]. Pero por supuesto, el campo donde en realidad está el foco de este proyecto es la clasificación de imágenes médicas.

2.2.5. Transformadores de visión en clasificación de imágenes médicas

Producto de la pandemia del virus COVID-19, y dados los resultados notables obtenidos por los transformadores de visión en otros campos, surgieron varias investigaciones centradas en el diagnóstico de esta patología, He *et al.* realizan una gran recopilación de éstos y otros casos en su review [26]. De las múltiples investigaciones referenciadas por He *et al.* es preciso destacar, en el contexto de este proyecto, unas cuantas formuladas para el uso de CT scans, y en primera instancia se identifican las cuales utilizan la arquitectura ViT sin alteraciones considerables. Than *et al.* [27] estudian el efecto del tamaño de los patches en su uso de ViT para la clasificación de COVID-19 y pulmones con presencia de enfermedades, encontrando que el desempeño decae a mayor tamaño de patches, revelando un compromiso entre información local y global. Costa *et al.* [28] obtienen resultados competitivos al utilizar modelos pre-entrenados, incluyendo DeiT [20], para distinguir COVID-19 y otras neumonías de casos normales. Li *et al.* [29] desarrollaron una plataforma para el diagnóstico de COVID-19 basada en ViT, ellos convierten las muestras de imágenes CT en una serie de patches planos para ajustarse a la entrada de ViT, también adoptan un modelo de destilación profesor-estudiante para destilar el conocimiento de una CNN pre-entrenada con imágenes naturales. Gao *et al.* [30] aplicaron ViT a CT scans en 2 y 3 dimensiones para diagnosticar COVID-19, ellos construyen un sub-volumen de imágenes, extrayendo un número fijo de rebanadas para formarlo. Zhang *et al.* [31] entrenaron el popular Swin-Transformer en imágenes CT, su entorno de trabajo primero segmenta el pulmón por medio de UNet para luego entregar la región del pulmón al extractor de características, esta estrategia ayuda a reducir la carga computacional del transformador.

Y aunque los ViTs puros han demostrado muy buenos resultados en el área, también ha habido una exploración extensa de posibilidades de combinar ViT con otros componentes de aprendizaje para obtener una representación de distribuciones de da-

ta complejas, siendo las más populares las capas de convolución, en un intento de aprovechar la extracción de características locales de las CNN con el contexto global proporcionado por ViT. He *et al.* destacan interesantes casos en el contexto descrito anteriormente. Liang [32] usa una CNN para minar características y un transformer para realizar la agregación de éstas, enfocados en imágenes CT 3D. Barhoumi *et al.* [33] combinan varias CNN con un ViT, al alimentar al ViT con las características extraídas, además pre-entrenan la CNN utilizando una arquitectura GAN (Generative Adversarial Network) y concluyen que éste último proceso ayuda a disminuir las diferencias entre los dos dominios.

Manzari *et al.*, en una publicación reciente [34], proponen un modelo híbrido, teniendo como propósitos el mitigar el peso computacional de los *transformers*, que crece cuadráticamente con la dimensión espacial o de *embeddings*. Su objetivo es proponer un modelo robusto y general para la clasificación de imágenes médicas, que generalice para un amplio rango de dominios médicos, como CT, x-ray y ultrasonidos. También tratan con gran atención el tema de *adversarial attacks*, apuntando a un modelo propuesto robusto ante este tipo de riesgos. También hacen uso del dataset MedMNISTv2, propuesto recientemente [39], el cual se propone como un *benchmark* para modelos de aprendizaje de máquinas en el contexto de análisis de imágenes médicas y visión por computadora, este dataset será utilizado en el presente trabajo para probar los modelos que puedan ser aplicados al objetivo final y poder hacer comparaciones con modelos en el estado del arte, cuyos resultados en MedMNISTv2 están documentados.

3. SISTEMA SOLUCIÓN PROPUESTO

Un sistema de asistencia en la detección de anomalías en tomografías torácicas basado en técnicas de deep learning debería contemplar todas las fases del tratamiento de las muestras médicas, tanto como el modelo a utilizar para las predicciones. Además, la incorporación de técnicas de explicabilidad para las predicciones permitirían soportar en parte el comportamiento de “caja negra” con el que se trabaja al utilizar redes neuronales profundas, ya que en aplicaciones como la presente se beneficiarían enormemente de lograr explicar los segmentos de las muestras en donde el modelo toma decisiones. Lo anterior, al ser combinado con la opinión y criterio de un radiólogo experto, permitiría aumentar la confiabilidad en el proceso de análisis.

La Figura 3.1 muestra un esquema básico de cómo se podría estructurar un sistema que de solución al problema planteado. Donde los cortes de cada muestra 3D son primero alimentados a un módulo de segmentación, el que es capaz de alterarlas para conservar solo la información relevante perteneciente al órgano en cuestión, o proporcionar una máscara binaria que puede ser entregada al procesador de muestras para ejecutar la segmentación. Luego el modelo de DL realiza la clasificación de la muestra. Mientras que el módulo de explicabilidad extrae información de las zonas más relevantes para la predicción realizada, y con ello proporciona una guía visual para el usuario.

La sección 3.1 detalla los principales aspectos a considerar durante el preprocesamiento de las muestras médicas para un dataset compuesto por imágenes médicas en formato DICOM, que es el caso de las muestras recopiladas por el Hospital Clínico de

la Universidad de Chile. Sin embargo, para el propósito de esta memoria, se trabajará con un dataset simplificado, MedMNIST v2 [39], con objeto de validar el uso de los modelos propuestos previo a la utilización del dataset original. Más información sobre MedMNIST v2 es proporcionada en el capítulo 4.

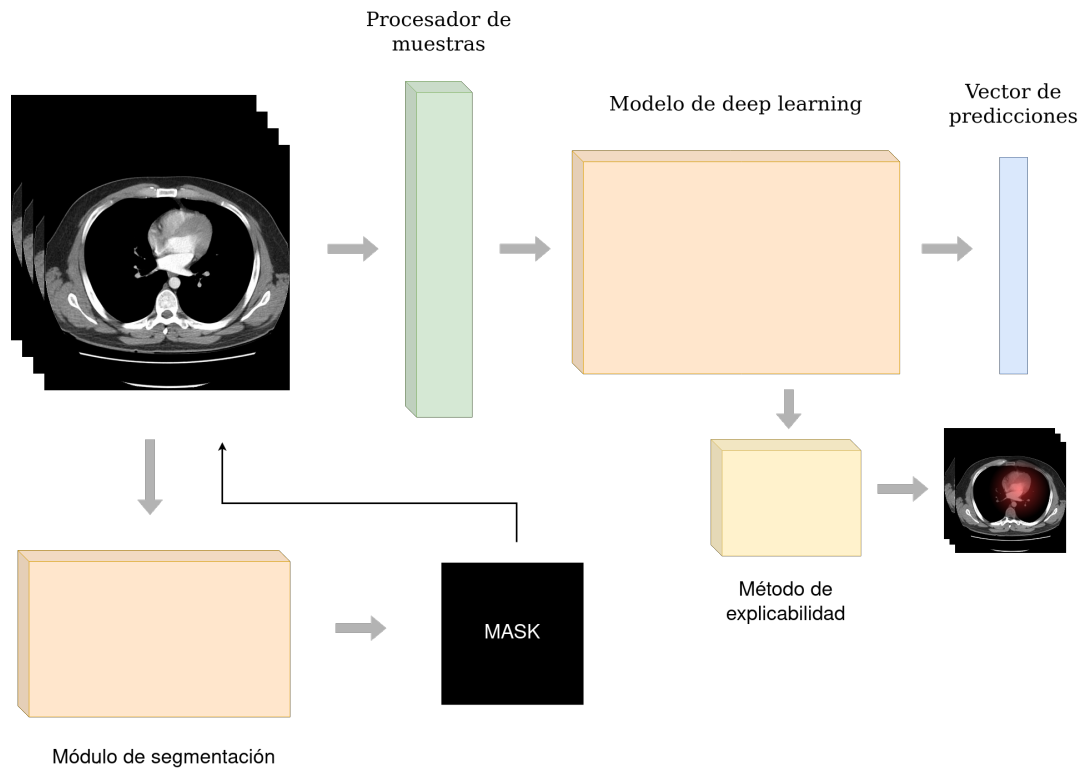


Fig. 3.1: Prototipo de sistema de asistencia en la detección de anomalías en tomografías torácicas 3D.

3.1. Discriminación de data y preprocesamiento

La elección del conjunto de muestras de imágenes CT a incluir en el dataset final que se propone usar para entrenamiento y validación es crítica, ya que puede existir una variabilidad considerable en ciertos aspectos entre conjuntos de muestras. Un aspecto clave es el uso de contraste al momento del examen, pues existe una diferencia considerable entre una muestra de un examen realizada con un medio de contraste

aplicado al paciente que sin él. Esta característica puede ser obtenida tanto de los reportes escritos como de la muestra en formato DICOM, pues esta posee un campo llamado *contrast* que indica la inclusión o no de contraste intravenoso, y en un principio se recomendaría utilizar solo muestras sin contraste aplicado, pues los efectos de este agente pueden ser distintos dependiendo de cada paciente y el tiempo transcurrido entre la aplicación de la sustancia y la toma del examen.

En el proceso de selección de las muestras que se consideraran en el dataset final, también se debe considerar que las etiquetas de clases deben ser extraídas de los reportes, escritos por profesionales, asociados a cada fichero, y entonces es preciso descartar las instancias duplicadas, reportes preliminares, aquellos que no reportan hallazgos e informes vacíos.

Las imágenes CT en formato DICOM poseen mucha más información de la necesaria para la tarea en cuestión, y entonces es preciso rescatar solo lo necesario de ellas para el uso correcto de la memoria de la que dispondrá durante el proceso de entrenamiento. Se debe fijar una ventana HU coherente con las ventanas que comúnmente fija un radiólogo al realizar el estudio de una CT, para ello se pueden usar los campos *RescaleSlope* y *RescaleIntercept* para fijar una ventana común para todo el dataset de [-1000 HU, +1000 HU], que corresponden a las radiodensidades del aire y hueso denso respectivamente. Es necesario también hacer un reescalado para tener consistencia de tamaño físico real representado por cada píxel entre el universo de imágenes CT, para ello el campo *PixelSpacing* entrega información sobre el tamaño físico que se encuentra codificado como la distancia entre el centro de dos píxeles adyacentes. El campo *SliceThickness* proporciona información sobre la tercera dimensión de la imagen, y aunque su utilización podría ser más compleja, sería recomendable realizar un estudio para evaluar la variabilidad de esta propiedad en el universo seleccionado. Otros campos que se deben tener en cuenta para asegurar la coherencia en el dataset son *ImageOrientationPatient* e *ImagePositionPatient*, para asegurarse de que el orden

y la orientación de los cortes de cada imagen CT 3D sea el correcto, y *GantryDetectorTilt*, para garantizar que todas las muestras tengan inclinación cero. La Tabla 3.1 muestra los atributos del formato DICOM mencionados y una descripción de ellos. Una buena referencia, y en que se basa esta propuesta de procesamiento, es la provista en el apéndice A de [36].

Tag	Atributo	Descripción
(0028,1052)	Rescale Intercept	Valor b en la relación entre la información de píxeles (SV) y las unidades de salida (Output units). $\text{Output units} = m \cdot \text{SV} + b$.
(0028,1053)	Rescale Slope	m en la ecuación especificada en Rescale Intercept.
(0020,0037)	Image Orientation (Patient)	Los cosenos de dirección de la primera fila y la primera columna con respecto al paciente.
(0020,0032)	Image Position (Patient)	Coordenadas x,y,z de la esquina superior izquierda (centro del primer voxel transmitido) de la imagen, en mm.
(0018,1120)	Gantry/Detector Tilt	Ángulo nominal de inclinación en grados del gantry de escaneo.

Tab. 3.1: Atributos de importancia para procesamiento de volúmenes CT.

Con el conjunto final consolidado se contempla la implementación de un módulo de procesamiento de imágenes, o volúmenes, para adaptar cada muestra a las necesidades del modelo a utilizar para clasificación. Esto dependerá de la estructura de cada modelo, y puede contemplar reescalado para adaptar a la entrada, normalización y técnicas de *data augmentation*. Un ejemplo de esto puede verse en la sección 4.2, donde se utiliza una herramienta de preprocesamiento para la adaptación del dataset de prueba a las dimensiones de entrada del modelo utilizado.

3.2. *Modelo de clasificación multietiqueta*

La siguiente etapa contempla la búsqueda y planteamiento del modelo de clasificación multietiqueta utilizado para la detección de anomalías en las muestras médicas, como es la premisa del proyecto, se propone trabajar con modelos que incorporen mecanismos de atención.

El primer modelo donde se centran las miradas naturalmente es ViT. La implementación clásica de ViT, presentada en [11], está ideada para el trabajo con imágenes en dos dimensiones, por lo que es importante desarrollar una estrategia de adaptación a imágenes 3D sólida. Otros modelos candidatos para este módulo son las variantes de ViT, las más populares siendo Swin Transformer [21] y DeiT [20], aunque como ViT y la mayoría de estos modelos, presentan el mismo desafío de adaptación a imágenes 3D.

Varios de los estudios citados en esta memoria presentan estrategias para la adaptación de estos modelos al trabajo con CT 3D, otros proponen modelos propios basados en el uso de mecanismos de atención, como también modelos híbridos con módulos de convolución. Se presentan a continuación algunos modelos y estrategias de adaptación candidatos para la integración en el sistema propuesto.

3.2.1. *Estrategias de adaptación para Vision Transformers*

En [30] se adopta una estrategia basada en un umbral de votación por predicción de rebanada, la estructura del clasificador puede verse en la Figura 3.2, considerando como caso positivo cuando se tiene $> 25\%$ de predicciones positivas para los cortes del volumen CT.

En [31] utilizan un Swin Transformer preentrenado como extractor de característi-

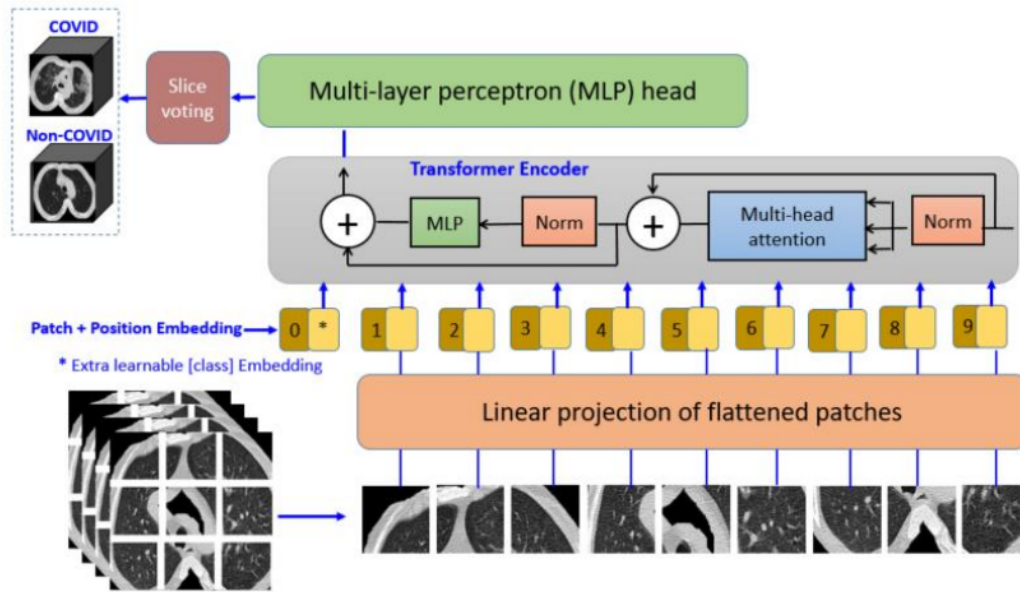


Fig. 3.2: Clasificador de COVID-19 por umbral de votación de cortes. Fuente [30].

cas y luego una capa de pooling seguida de un cabezal de clasificación. La Figura 3.3 muestra la estructura de dicho modelo, donde se puede apreciar además el módulo de segmentación basado en U-Net utilizado por los autores para el preprocesamiento de las muestras.

En [37] utilizan un promedio de las predicciones por cortes, y aunque en este estudio se trabaja con CNN como extractor de características, la técnica de agregación puede ser considerada para aplicarse al caso con un modelo basado en ViT. Además de lo anterior, esta publicación puede resultar de gran utilidad en el desarrollo del proyecto, tanto para esta memoria como para el trabajo futuro, ya que de igual forma persigue y presenta un clasificador multietiqueta de anomalías en CT 3D. La función de pérdida utilizada en esta técnica de agregación puede ser replicada para nuestro caso.

Otros trabajos presentan modelos híbridos, que también contemplan capas convolucionales, como en [38], que utiliza una CNN para extraer características de los cortes

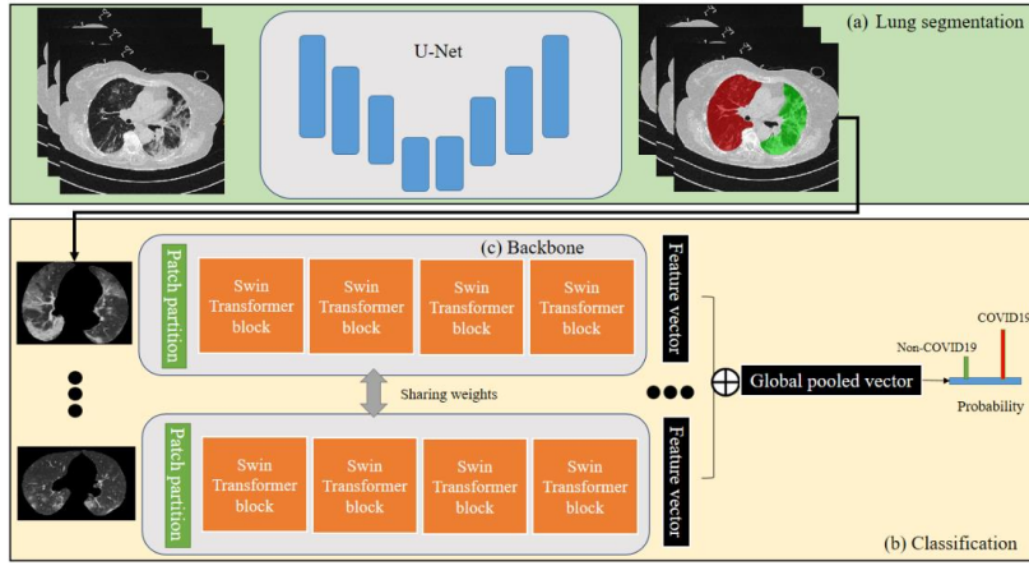


Fig. 3.3: Clasificador basado en Swin con una capa de pooling para agregación de características [31].

de las CT 3D y luego BERT para la agregación y clasificación. En [32] se utiliza un modelo un poco más complejo, con capas convolucionales, un bloque residual que implementa *SE attention*, y una capa MLP en paralelo a un bloque de transformador para la agregación y clasificación.

En primera instancia se propone la implementación del sistema con un ViT preentrenado en ImageNet21k, y realizar un proceso de *fine-tuning* para la tarea en cuestión, este enfoque está sumamente explorado y documentado en diversas fuentes, y serviría como un pie inicial para plantear un modelo mas complejo, que pueda contar incluso con una etapa de segmentación para guiar al clasificador a tomar decisiones en las zonas de interés. Se presupone que el conocimiento adquirido por este modelo durante la etapa de preentrenamiento es transferible al análisis de muestras médicas. Para corroborar esta hipótesis, se proponen tres experimentos en la siguiente sección, los cuales servirán como validadores y motivadores para la aplicación de este u otros modelos más complejos.

3.3. *Módulo de segmentación*

Este bloque cumple una función mas bien de asistencia para el módulo de clasificación, pues su propósito es reconocer y proveer una guía en las zonas del volumen, o cortes, en donde existe presencia del órgano objetivo en el que se pretende que el clasificador discrimine. Este módulo puede ser implementado igualmente mediante técnicas de aprendizaje automático. En este campo, se han realizado significativos avances y se dispone de una variados enfoques, desde aquellos basados en CNN hasta modelos que incorporan mecanismos de atención. También es factible ejecutarlo como parte de un proceso de preprocesamiento, utilizando técnicas clásicas de segmentación de visión por computadora. De igual forma, la ejecución de este módulo está subyugada a la elección del clasificador, pues este puede contemplar entradas de cortes 2D o los volúmenes completos de las tomografías.

Este módulo tomaría entonces los cortes 2D, o volúmenes en su completitud, y entregaría una máscara que denote las zonas permitidas, con la cual se puede retirar la información de las muestras que no sea relevante, antes de ser entregadas al clasificador, o bien alimentar al clasificador con la muestra íntegra y la máscara generada. Un ejemplo de esto último es lo realizado en [38], donde al extractor CNN se le alimenta con la máscara en un canal de la entrada.

La utilización de un modelo de DL con arquitectura de U, ya sea basado en convoluciones o que implemente mecanismos de atención, naturalmente es una opción que despierta interés, dado el éxito de U-Net en aplicaciones médicas. En este escenario es que se debería contar con un dataset similar al disponible para el proyecto, pero que cuente con anotaciones para entrenar al modelo de DL.

3.4. Módulo de explicabilidad

Además de proporcionar las etiquetas que el modelo detecte para cada caso, el hecho de ser un asistente en la detección de anomalías implica que el sistema debería entregar información de las zonas en que se centra la predicción, en pos de que el operador pueda tener una impresión de las características que son detectadas por el clasificador para la discriminación. Existen múltiples técnicas para generar mapas de calor de las zonas de importancia en una predicción, y en particular para este caso, se propone utilizar algún método que aproveche los mecanismos de atención presentes en el modelo a utilizar para la clasificación. En [44] se recopilan múltiples métodos que podrían ser base para implementar este módulo, tanto explícitamente para arquitecturas ViT como otros generales para modelos de DL, como por ejemplo GradCAM [41]. En la Figura 3.4 se muestra un ejemplo de distintos métodos de explicabilidad, este ejemplo es extraído de [43].

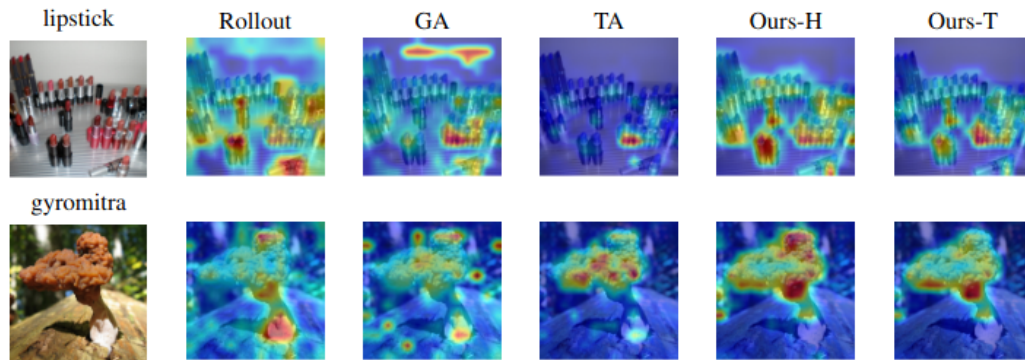


Fig. 3.4: Localización de regiones detalladas para la predicción de una sola clase de ViT-Base por diferentes métodos. Fuente [43].

4. EXPERIMENTOS DE VALIDACIÓN

Dado el sistema propuesto, expuesto en la Figura 3.1, el componente principal es evidentemente el módulo que contempla el modelo de ML para clasificación. Es por esta razón que, debido a las restricciones de maquinaria y tiempo, el alcance de la experimentación se limita al módulo de clasificación en este trabajo, dejando los módulos de segmentación y explicabilidad para trabajos futuros.

Se proponen tres experimentos para evaluar el comportamiento de un ViT preentrenado para la clasificación de muestras médicas. Como la discriminación de muestras, el procesado y entrenamiento en el dataset real requeriría grandes tiempos de ejecución y maquinaria dedicada y de poca disponibilidad en un estudio previo de factibilidad, se ejecutan estos en un dataset de prueba, MedMNIST v2 [39]. Este dataset de prueba cuenta con varios subconjuntos para la clasificación de imágenes biomédicas, tanto en 2D como 3D, contemplando CT, rayos X, ultrasonidos, entre otros, para tareas de clasificación binaria, multiclase y multietiqueta. Todos los subconjuntos son preprocesados a un tamaño común de 28x28, para datasets 2D, y 28x28x28, para datasets 3D, con el objetivo de proveer un conjunto de datos ligero y estandarizado, de uso directo y sencillo, para uso educacional y validación de modelos biomédicos.

Los objetivos específicos de los experimentos son:

- Verificar la habilidad de estos modelos, preentrenados en imágenes comunes, para extraer información de imágenes médicas similares a las que se tienen a disposición para el proyecto.

- Utilizar y comparar las representaciones codificadas de las muestras respecto del trabajo con su versión original en un método de ML tradicional.
- Adaptar un modelo preentrenado en imágenes comunes a la utilización de muestras médicas 2D y verificar su comportamiento con el *benchmark* provisto por los autores del dataset [39].
- Idear una o mas estrategias de adaptación de un modelo preentrenado en imágenes comunes 2D al trabajo con muestras médicas 3D, y comparar su comportamiento con el *benchmark* provisto por los autores del dataset [39].
- Con los resultados experimentales de prueba, concluir sobre la validez de la aplicación de estas metodologías en el dataset real a utilizar en el proyecto.

A continuación, se ofrece una síntesis concisa de los experimentos presentados.

1. **Exploración de extracción de características por ViT preentrenado en imágenes médicas 2D:** Para estas pruebas se utiliza el dataset OrganCMNIST del conjunto de MedMNIST v2, compuesto por imágenes CT abdominales 2D de plano coronal (también están presentes en el conjunto OrganAMNIST y OrganSMNIST, de planos axial y sagital respectivamente), para la tarea de clasificación multiclase de órganos. Como todos los demás conjuntos, las muestras cuentan con preprocesamiento para el rápido despliegue de metodologías de clasificación, el cual contempla la selección de una ventana abdominal y disposición en formato *grayscale*.

Utilizando esta data, se entrena un clasificador basado en un método de machine learning clásico SVM. Un SVM (Support Vector Machine) es un algoritmo de aprendizaje automático supervisado, cuyo objetivo es encontrar el hiperplano óptimo que separe las distintas clases. Es conocido por su eficacia en conjuntos

de datos de alta dimensionalidad y su capacidad para manejar datos no lineales al utilizar funciones kernel.

El SVM es entrenado a partir de la imagen original y luego utilizando un ViT preentrenado para extraer una representación codificada de ella y comparar los resultados de estos procedimientos.

2. Fine-tuning de ViT preentrenado para clasificación de muestras médicas:

Para esta prueba se utiliza el mismo dataset que en la prueba anterior. Pero en esta oportunidad se realiza un procedimiento de *fine-tuning*, esta vez con un cabezal de clasificación, para su comparación con los modelos basados en convoluciones que son provistos en el *benchmark* de referencia.

3. Adaptación de modelo preentrenado para imágenes médicas 3D: Este experimento contempla la adopción de una estrategia para utilizar un modelo ViT a un caso similar al objetivo del proyecto, aunque reducido a un dataset de prueba, en este caso NoduleMNIST3D. Este dataset está diseñado para clasificación binaria de volúmenes CT según la presencia de nódulos malignos, a partir de tomografías torácicas.

Se propone en esta instancia utilizar un promedio o una capa de pooling para realizar la agregación de las características, extraídas por el encoder de ViT.

La Tabla 4.1 muestra información de la distribución de los subconjuntos de entrenamiento, validación y testeo de los datasets utilizados en las pruebas de validación.

En tanto para las métricas, se consideran *accuracy* y *roc_auc* para los tres casos experimentales, pues estas son las reportadas por los autores para distintos modelos en

Dataset	Tipo de data	Tarea (clases)	# Entrenamiento / # Validación / # Testeo
OrganCMNIST	CT abdominal 2D	Multiclase (11)	13,000 / 2,392 / 8,268
NoduleMNIST3D	CT torácica 3D	Binaria	1,158 / 165 / 310

Tab. 4.1: Características de los datasets utilizados en la parte experimental de la memoria.

el estado del arte [39]. *Accuracy* expresa la proporción de aciertos en las predicciones sobre el universo de datos procesados, ver ecuación 4.1, mientras que *roc_auc* expresa el área bajo la curva (AOC) de la curva característica de operación del receptor (ROC), esta métrica muestra que tan bien es capaz de distinguir el clasificador entre clases negativas y positivas (para clasificador binario), como el primer dataset considerado es multiclase, se utiliza una adaptación de esta métrica para el caso, la cual se basa en la ponderación de cada clase tratado como clase binaria (con las demás siendo la clase negativa), en una estrategia denominada *One Versus Rest* u OVR.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

4.1. Exploración de extracción de características por ViT preentrenado en imágenes médicas 2D

Para este experimento entonces se utiliza el dataset OrganCMNIST, detallado en la Tabla 4.1, el cual esta disponible como paquete para Python, para realizar clasificación multiclase utilizando SVC con dos tipos distintos de kernel y a sobre dos espacios de datos, uno a partir de los píxeles en crudo de las imágenes del dataset, y el otro a partir de la representación codificada de estos por un encoder ViT (que no contempla cabezal de clasificación). La distribución de clases y subconjuntos puede verse en la Tabla 4.2, mientras que la Figura 4.1 presenta un mosaico con algunas de las imágenes presentes en el subconjunto de entrenamiento del dataset mencionado.

Índice	Etiqueta	Entrenamiento	Validación	Testeo
0	bladder	1148	191	828
1	femur-left	619	347	431
2	femur-right	595	96	421
3	heart	600	202	421
4	kidney-left	1088	132	727
5	kidney-right	1170	157	735
6	liver	2986	429	828
7	lung-left	1002	347	549
8	lung-right	1022	352	557
9	pancreas	1173	179	750
10	spleen	1572	205	962

Tab. 4.2: Distribución de clases a través de los subconjuntos del dataset OrganCMNIST.



Fig. 4.1: Mosaico de muestras presentes en OrganCMNIST 20x20.

Los sistemas empleados en este experimento pueden verse en la Figura 4.2.

Para los clasificadores sobre las imágenes crudas solo se aplica una etapa de estandarización, teniendo entonces entradas de 784 dimensiones (28x28 píxeles), la implementación es simple y directa, utilizando la biblioteca de Scikit-Learn, se utilizan

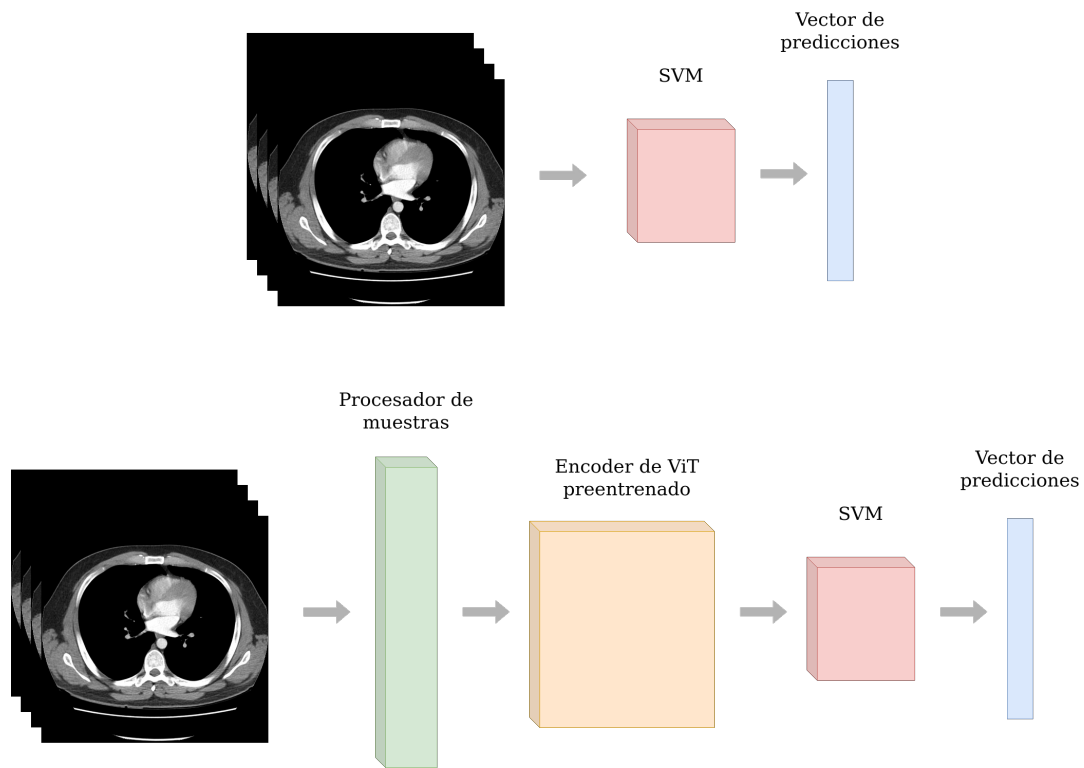


Fig. 4.2: Arriba: SVM sobre las imágenes crudas. Abajo: SVM sobre características extraídas por ViT preentrenado.

parámetros por defecto y no se considera búsqueda de parámetros, pues el objetivo es la comparación con un clasificador similar sobre el espacio transformado por el encoder de ViT.

Para el siguiente par de clasificadores se utiliza un ViT preentrenado en ImageNet21k, sin la capa MLP al final que se utiliza normalmente para clasificación, se escoge el modelo ViT-Base, con patches de tamaño 16x16, las características de este modelo, junto con las otras variantes presentadas en [16], pueden verse en la Tabla 4.3. Las muestras deben pasar por un preprocesado de reescalado y normalización para adaptarse a la entrada de 224x224 píxeles, de igual forma, se agrega una etapa de estandarización para ser consistentes con los clasificadores anteriores. Tanto para el preprocesamiento como para el modelo en si, se utiliza la librería transformers de Hugging Face, una librería muy potente y de fácil uso para el trabajo con transfor-

madores. La versión codificada de las imágenes se extrae del estado oculto del token [cls], pues en la arquitectura ViT completa es esta información la que se pasa a la etapa MLP para clasificación.

Modelo	Capas	Tamaño oculto	Tamaño MLP	Cabezas	Parametros
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Tab. 4.3: Detalles de las variantes de ViT [16].

Los resultados para los clasificadores entrenados, para el subconjunto de testeo del dataset empleado, pueden verse en la Tabla 4.4.

Tipo de data	Clasificador	Accuracy	ROC_AUC
Imágenes puras	SVC con kernel lineal	0.582	0.928
Imágenes puras	SVC con kernel rbf	0.787	0.975
Características extraídas por ViT	SVC con kernel lineal	0.821	0.986
Características extraídas por ViT	SVC con kernel rbf	0.831	0.987

Tab. 4.4: Métricas obtenidas para los clasificadores entrenados el primer experimento.

4.2. Fine-tuning de ViT preentrenado para clasificación de muestras médicas

Dados los resultados anteriores, el siguiente paso es usar un modelo preentrenado para la tarea de clasificación multiclase en el mismo dataset descrito anteriormente. Utilizando el mismo modelo ViT-Base, ahora contemplando una etapa MLP sobre la última capa del encoder, se realiza el proceso de fine-tuning haciendo uso de la clase Trainer de la misma biblioteca de transformers, que provee una API de fácil uso para

el entrenamiento de los modelos.

Los pesos preentrenados, tanto para este y los demás experimentos, son proporcionados por los autores de la misma publicación que presenta la arquitectura [16], excepto que los pesos del cabezal MLP son borrados previo a hecho público el modelo, aunque esto naturalmente no es gran problema, pues dicha capa es específica para cada tarea de clasificación. En este caso la capa MLP también se alimenta de los estados ocultos del token [cls] de la última capa del encoder.

Como se mencionó, el proceso de fine tuning es bastante directo utilizando Trainer de HuggingFace, usa optimizador AdamW y solamente se configuran unos pocos hiperparámetros, como lo son la tasa de aprendizaje o *learning rate* a 1×10^{-3} , se entrena por 2 épocas, en un principio, y por 4 épocas posteriormente para analizar si se consigue un aumento perceptible en el comportamiento del clasificador.

Los resultados de los modelos, evaluados en el subconjunto de testeo, se pueden ver en la Tabla 4.5. Estos pueden ser comparados con los reportados para distintos modelos en el estado del arte, por los autores del dataset, presentes en la Tabla 4.8 de la sección 4.3.

Modelo	Subconjunto	Epocas	Acc	ROC_AUC
ViT-Base	Entrenamiento	2	0.985	0.999
ViT-Base	Testeo	2	0.928	0.996
ViT-Base	Entrenamiento	4	0.985	0.999
ViT-Base	Testeo	4	0.938	0.996

Tab. 4.5: Métricas obtenidas para los clasificadores del segundo experimento.

4.3. Adaptación de modelo preentrenado para imágenes médicas 3D

Este último proceso experimental contempla la búsqueda y aplicación de una estrategia de adaptación de un modelo que implementa mecanismos de atención, ideado para el trabajo con imágenes 2D, a un dataset de imágenes médicas 3D, nuevamente se selecciona un ViT preentrenado, con el objetivo de evaluar tanto el aprendizaje transferido como el método de adaptación seleccionado.

El detalle de la división de los subconjuntos del dataset NoduleMNIST3D puede verse en la Tabla 4.1, mientras que la Figura 4.3 muestra un mosaico de cortes (en la posición numero 10) de las imágenes 3D contenidas en el subconjunto de entrenamiento del dataset. Este dataset está pensado para la tarea de clasificación binaria, siendo la clase negativa (0) “*benign*” y la clase positiva (1) “*malignant*”, la distribución de clases para los subconjuntos del dataset se expone en la tabla 4.6.

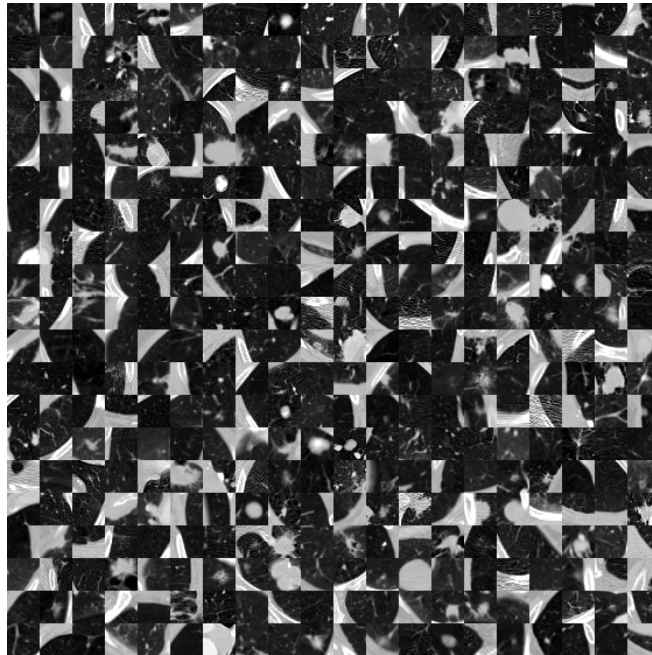


Fig. 4.3: Mosaico de cortes en las muestras 3D presentes en NoduleMNIST3D 20x20.

Se construyen entonces dos modelos que incorporan el ViT preentrenado como un

Índice	Etiqueta	Entrenamiento	Validación	Texteo
0	benign	863	123	246
1	malignant	295	42	64

Tab. 4.6: Distribución de clases en el dataset NoduleMNIST3D.

submódulo, y toman entradas de imágenes bidimensionales. El primer modelo sería un ViT convencional, con capa MLP directamente sobre los estados ocultos del token [cls], pero en el proceso de entrenamiento la pérdida es calculada sobre el promedio de las predicciones de todas las cortes de una muestra tridimensional. Mientras que el segundo caso toma los estados ocultos y los agrega por medio de una capa de max-pooling, luego la capa MLP realiza la predicción sobre este vector agregado. Ambos modelos son adaptados para clasificación binaria, por lo que la función de pérdida utilizada es *Binary Cross Entropy*.

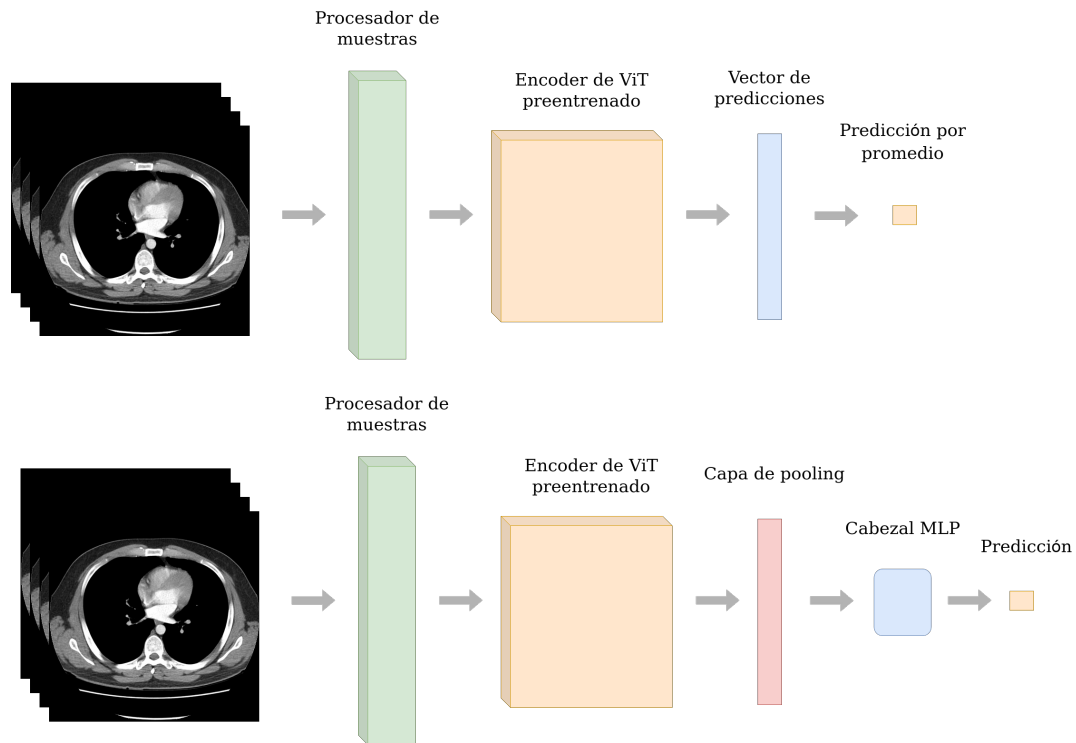


Fig. 4.4: Arriba: ViT con agregación por promedio de predicciones. Abajo: ViT con agregación por capa de pooling.

Los sistemas utilizados en esta sección pueden verse en la Figura 4.4. Nuevamente se hace uso de un preprocesador para adaptar los cortes de las imágenes 3D a la entrada de los modelos que contienen el ViT como submódulo, no se utilizan técnicas de *data augmentation* y se utiliza un bucle de entrenamiento en esencia simple, sin decaimiento de pesos o un programador de ratio de aprendizaje. Se hace una prueba previo al entrenamiento para verificar que existe cierto grado de aprendizaje de los modelos, luego se entrena cada uno por 4 épocas donde solo se ajustan los pesos de los cabezales MLP, para luego entrenar por 6 épocas considerando los pesos del submódulo ViT con un ratio de aprendizaje menor, finalmente se evalúan los modelos con el subconjunto de testeo, los resultados en este y los demás subconjuntos (considerando las métricas obtenida de la ultima época de entrenamiento) se pueden ver en la Tabla 4.7.

4.4. *Análisis de resultados y comparación con benchmark*

De los resultados obtenidos para el primer experimento, presentados en la Tabla 4.4, resulta fácil apreciar que los clasificadores SVM entrenados con un espacio de datos constituido por características extraídas de las imágenes por un ViT preentrenado con imágenes naturales, presentan una mejora sobre los clasificadores entrenados sobre las imágenes crudas, con el SVM con kernel lineal y extractor ViT superando por 41 % a su contraparte en precisión, y en 6 % en área bajo la curva. Además, el SVM con kernel RBF exhibe un aumento del 5 % en precisión y del 1 % en el área bajo la curva. Estos resultados sugieren que las representaciones obtenidas mediante la extracción de características con el modelo preentrenado son más fácilmente separables mediante técnicas de clasificación clásicas de aprendizaje automático, como el SVM.

El segundo experimento dio como resultado muy buenos clasificadores, como puede apreciarse en los resultados expuestos en la Tabla 4.5, con valores de precisión y

ViT	Método de agregación	Subconjunto	Épocas	Acc	ROC_AUC
ViT-Base	Promedio de predicción por cortes	Testeo	0	0.206	0.450
ViT-Base	Promedio de predicción por cortes	Entrenamiento	4	0.745	0.493
ViT-Base	Promedio de predicción por cortes	Validación	4	0.745	0.6959
ViT-Base	Promedio de predicción por cortes	Testeo	4	0.793	0.457
ViT-Base	Promedio de predicción por cortes	Entrenamiento	10	0.745	0.492
ViT-Base	Promedio de predicción por cortes	Validación	10	0.745	0.472
ViT-Base	Promedio de predicción por cortes	Testeo	10	0.793	0.515
ViT-Base	Max Pooling	Entrenamiento	4	0.742	0.510
ViT-Base	Max Pooling	Validación	4	0.745	0.733
ViT-Base	Max Pooling	Testeo	4	0.793	0.720
ViT-Base	Max Pooling	Entrenamiento	10	0.745	0.489
ViT-Base	Max Pooling	Validación	10	0.745	0.456
ViT-Base	Max Pooling	Testeo	10	0.793	0.512

Tab. 4.7: Métricas obtenidas para los clasificadores del tercer experimento.

área bajo la curva bastante altos. Con esto resulta preciso aseverar que el conocimiento obtenido por estos modelos de un entrenamiento, con grandes cantidades de datos de imágenes naturales, es transferible a imágenes de muestras médicas.

En tanto el tercer experimento muestra resultados que ofrecen perspectivas menos alentadoras, pero aún con oportunidades de mejora, la precisión varía muy poco y solo en el subconjunto de entrenamiento, mientras que el área bajo la curva presenta valores cercanos a 0.5, lo que se considera un clasificador aleatorio. Lo anterior es claro indicador de que las estrategias de agregación no son exitosas o el conocimiento adquirido del proceso de preentrenamiento no es compatible con las estrategias seleccionadas. También el hecho de que se trabaje sobre los cortes en su totalidad y no en

segmentos acotados de interés puede ser un gran factor a considerar.

El repositorio del proyecto MedMNIST v2 también provee una clase evaluadora, con la que se reportan las métricas obtenidas para los modelos provistos en el *benchmark* [39]. La Tabla 4.8 muestra los resultados obtenidos por los clasificadores en comparación con los resultados provistos en [39], para el dataset OrganCMNIST, donde se aprecia que el clasificador obtenido por *fine-tuning* supera a los otros modelos reportados por los autores.

Modelo	AUC	ACC
ResNet-18 (28)	0.992	0.900
ResNet-18 (224)	0.994	0.920
ResNet-50 (28)	0.992	0.905
ResNet-50 (224)	0.993	0.911
auto-sklearn	0.976	0.829
AutoKeras	0.990	0.879
Google AutoML Vision	0.988	0.877
ViT encoder + SVM rbf kernel	0.983	0.830
ViT-Base	0.996	0.938

Tab. 4.8: Benchmark de rendimiento en OrganCMNIST con los mejores clasificadores obtenidos en el primer y segundo experimento.

La tabla 4.9 muestra los resultados obtenidos por los clasificadores del tercer experimento, después de 10 épocas, en comparación con los resultados provistos para modelos en el estado del arte para el trabajo con imágenes 3D [39]. Los resultados obtenidos por los clasificadores presentados para este caso tienen resultados bastante menores que los reportados por los autores, por ende se debe explorar alguna solución alternativa.

Modelo	AUC	ACC
ResNet-18 + 2.5D	0.838	0.835
ResNet-18 + 3D	0.863	0.844
ResNet-18 + ACS	0.873	0.847
ResNet-50 + 2.5D	0.835	0.848
ResNet-50 + 3D	0.875	0.847
ResNet-50 + ACS	0.886	0.841
auto-sklearn	0.914	0.874
AutoKeras	0.844	0.834
ViT average pred	0.504	0.794
ViT pooler	0.512	0.794

Tab. 4.9: Benchmark de rendimiento en NoduleMNIST3D con los mejores clasificadores obtenidos en el tercer experimento.

5. CONCLUSIONES Y TRABAJO FUTURO

El sistema propuesto en esta memoria contempla la implementación de distintos módulos, basado en la investigación previa referida a la tarea en cuestión. De entre estos módulos, se dirige la atención de manera específica hacia el modelo de ML utilizado como clasificador, considerándolo como el elemento crucial del sistema.

En este contexto, los experimentos desarrollados para la validación del uso de modelos basados en mecanismos de atención arrojan resultados dispares. Por un lado se evidencia que el conocimiento adquirido en la fase de preentrenamiento en imágenes naturales puede ser aprovechado cuando se trabaja con muestras médicas 2D, esto en relación a los buenos resultados mostrados por los clasificadores obtenidos en el experimento numero dos, y en parte por la mejora del clasificador SVM al utilizar un espacio de datos con características extraídas por un modelo con mecanismos de atención, en comparación con el mismo entrenado en las imágenes originales, del primer experimento. Mientras que por otro lado los resultados obtenidos del tercer experimento no son satisfactorios, pues no representan un clasificador efectivo, capaz de discriminar con buen margen entre casos positivos y negativos.

Si bien los resultados obtenidos para el tercer experimento resultan menos alentadores que los anteriores, el bajo rendimiento puede atribuirse a algunos factores. En primera instancia, y producto de las restricciones de tiempo y maquinaria, el sistema y proceso propuesto no implementa un módulo o bloque que pueda guiar al clasificador a las zonas de interés, donde se espera que el modelo tome decisiones. Este hecho puede incidir enormemente en el comportamiento del clasificador, pues luego de extraer

las representaciones de las características extraídas por el encoder, estas contendrían información de muchas zonas de poco interés para la tarea en cuestión, como cortes sin la presencia de la anomalía o contornos ajenos al órgano en cuestión. En consecuencia, el cabezal MLP, combinado con la técnica de agregación correspondiente, contaría con una información poco prolija, donde la información de los segmentos relevantes puede verse diseminada por la presencia de características poco importantes para la diferenciación del caso.

Si bien el modelo puede generar una representación rica sobre las rebanadas 2D, la transferencia del conocimiento al trabajo con imágenes 3D es una tarea no trivial, debido a la diferencia en la dimensionalidad entre imágenes naturales en 2D e imágenes médicas en 3D. Las estrategias de agregación utilizadas cumplen la función de reducir la dimensionalidad del espacio latente, en el caso de la utilización de una capa de max-pooling, o del vector de predicciones (en este caso a un escalar, pues la tarea es clasificación binaria), en el caso de la predicción por promedio de los puntajes por rebanada. Sin embargo, la información posicional de profundidad se pierde al utilizar este acercamiento para abordar el problema.

En conforme con lo anterior, la necesidad de manejar un clasificador 3D efectivo puede bifurcarse en dos ramas. La primera de estas sería la búsqueda de una estrategia de adaptación que demuestre mejor desempeño, y considerando la implementación del módulo de segmentación para guiar al modelo. La segunda es la búsqueda de una arquitectura diferente, ideada explícitamente para el trabajo con imágenes 3D. Por ejemplo, en [40] desarrollan tres estrategias para adaptar las entradas al encoder de transformer, luego de que los voxels son proyectados por la capa de embeddings, de imágenes 2D a 3D, y así aprovechar el conocimiento adquirido por un ViT preentrenado sin modificar a grandes rasgos la estructura de éste.

El módulo de segmentación en tanto, persiste en considerarse un punto crítico pa-

ra el entrenamiento efectivo del modelo, varios de los sistemas mencionados en esta memoria implementan un modelo o etapa que cumple esta función [30] [31] [38] [37]. La implementación de un módulo como este podría contemplar técnicas de segmentación morfológica o basada en modelos de ML, donde nuevamente existen opciones de implementación que contemplan mecanismos de atención [24] [23], y es un paso y desafío a considerar en el trabajo futuro.

Por ultimo, como se ha mencionado anteriormente, un sistema de asistencia en la detección de anomalías debería poder entregar información sobre dónde el modelo toma decisiones, utilizando algún método de explicabilidad. Existen diferentes técnicas desarrolladas para atacar esta problemática, GradCAM [41] y sus variantes son ampliamente utilizadas, y también existen técnicas específicas para aprovechar los mecanismos de atención presentes en los transformers, como *Attention Rollout* [42] o el más reciente *Beyond Intuition* [43].

En resumen, aplicar modelos como los vision transformers, para abordar tareas de clasificación de imágenes médicas en tres dimensiones, representa un desafío significativo, especialmente al intentar capitalizar el conocimiento previamente adquirido por un modelo preentrenado en imágenes naturales 2D. Cada módulo debe ser meticulosamente planificado y validado en un trabajo extenso y bien estructurado, ya que cada componente contribuye de manera sustancial al sistema, ya sea en la tarea de clasificación misma o en la interacción con el usuario final. A pesar de los obstáculos y complejidades inherentes, los modelos que incorporan mecanismos de atención, particularmente los vision transformers, han demostrado su notable capacidad para abordar diversas tareas, y el ámbito de la imagenología médica no es una excepción. Por lo tanto, la aplicación de estos modelos en el problema abordado en esta memoria es tan prometedora como desafiante.

Bibliografía

- [1] BRUNO, Michael A.; WALKER, Eric A.; ABUJUDEH, Hani H. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 2015, vol. 35, no 6, p. 1668-1676.
- [2] ALOTAIBI, Norah A., et al. Diagnostic errors in clinical FDG-PET/CT. *European Journal of Radiology*, 2020, vol. 132, p. 109296.
- [3] RAGHU, Maithra, et al. Do vision transformers see like convolutional neural networks?. *Advances in Neural Information Processing Systems*, 2021, vol. 34, p. 12116-12128.
- [4] FUKUSHIMA, Kunihiko. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 1980, vol. 36, no 4, p. 193-202.
- [5] LECUN, Yann, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, vol. 86, no 11, p. 2278-2324.
- [6] KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012, vol. 25.
- [7] GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep learning*. MIT press, 2016.
- [8] SAK, Hasim; SENIOR, Andrew; RAO, Kanishka; BEAUFAYS, Françoise; SHCALKWYK, Johan. "Google voice search: faster and more accu-

rate.” Archived 9 March 2016 at the Wayback Machine, September 2015. <<https://ai.googleblog.com/2015/09/google-voice-search-faster-and-more.html>>. Accessed: June 6th, 2023.

- [9] BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
- [10] BRAUWERS, Gianni; FRASINCAR, Flavius. A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineering, 2021.
- [11] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 2017, vol. 30.
- [12] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [13] RADFORD, Alec, et al. Improving language understanding by generative pre-training. 2018.
- [14] RADFORD, Alec, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, vol. 1, no 8, p. 9.
- [15] BROWN, Tom, et al. Language models are few-shot learners. Advances in neural information processing systems, 2020, vol. 33, p. 1877-1901.
- [16] DOSOVITSKIY, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020. arXiv preprint arXiv:2010.11929, 2010.
- [17] YUAN, Li, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet. En Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 558-567.

- [18] FILIPIUK, Michal; SINGH, Vasu. Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems. En SafeAI@ AAAI. 2022.
- [19] UPARKAR, Om, et al. Vision Transformer Outperforms Deep Convolutional Neural Network-based Model in Classifying X-ray Images. *Procedia Computer Science*, 2023, vol. 218, p. 2338-2349.
- [20] TOUVRON, Hugo, et al. Training data-efficient image transformers & distillation through attention. En *International conference on machine learning*. PMLR, 2021. p. 10347-10357.
- [21] LIU, Ze, et al. Swin transformer: Hierarchical vision transformer using shifted windows. En *Proceedings of the IEEE/CVF international conference on computer vision*. 2021. p. 10012-10022.
- [22] RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. En *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18. Springer International Publishing, 2015. p. 234-241.
- [23] CHEN, Jieneng, et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [24] HATAMIZADEH, A., et al. Transformers for 3d medical image segmentation. *arXiv* 2021. *arXiv preprint arXiv:2103.10504*.
- [25] CAO, Hu, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. En *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022. p. 205-218.
- [26] HE, K., et al. Transformers in medical image analysis: A review. *arXiv* 2022. *arXiv preprint arXiv:2202.12165*.

- [27] THAN, Joel CM, et al. Preliminary study on patch sizes in vision transformers (vit) for covid-19 and diseased lungs classification. En 2021 IEEE National Bio-medical Engineering Conference (NBEC). IEEE, 2021. p. 146-150.
- [28] COSTA, Gabriel Sousa Silva, et al. COVID-19 automatic diagnosis with ct images using the novel transformer architecture. En Anais do XXI simpósio brasileiro de computação aplicada à saúde. SBC, 2021. p. 293-301.
- [29] LI, Jingxing; YANG, Zhanglei; YU, Yifan. A medical ai diagnosis platform based on vision transformer for coronavirus. En 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI). IEEE, 2021. p. 246-252.
- [30] GAO, Xiaohong; QIAN, Yu; GAO, Alice. COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models. arXiv preprint arXiv:2107.01682, 2021.
- [31] ZHANG, Lei; WEN, Yan. MIA-COV19D: a transformer-based framework for COVID19 classification in chest CTs. En Proceeding of the IEEE/CVF International Conference on Computer Vision Workshops. 2021. p. 513-8.
- [32] LIANG, Shuang. A hybrid deep learning framework for covid-19 detection via 3d chest ct images. arXiv preprint arXiv:2107.03904, 2021.
- [33] BARHOUMI, Yassine; GHULAM, Rasool. Scopeformer: n-CNN-ViT hybrid model for intracranial hemorrhage classification. arXiv preprint arXiv:2107.04575, 2021.
- [34] MANZARI, Omid Nejati, et al. MedViT: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine, 2023, vol. 157, p. 106791.

- [35] YANG, Jiancheng, et al. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 2023, vol. 10, no 1, p. 41.
- [36] DRAELOS, Rachel Lea, et al. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 2021, vol. 67, p. 101857.
- [37] DRAELOS, Rachel Lea; CARIN, Lawrence. Explainable multiple abnormality classification of chest CT volumes. *Artificial Intelligence in Medicine*, 2022, vol. 132, p. 102372.
- [38] TAN, Weijun; LIU, Jingfeng. A 3d cnn network with bert for automatic covid-19 diagnosis from ct-scan images. En *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. p. 439-445.
- [39] YANG, Jiancheng, et al. MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 2023, vol. 10, no 1, p. 41.
- [40] Wang, Yi, et al. Çan We Solve 3D Vision Tasks Starting from A 2D Vision Transformer?...^aarXiv preprint arXiv:2209.07026 (2022).
- [41] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [42] Chefer, Hila, Shir Gur, and Lior Wolf. "Transformer interpretability beyond attention visualization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
- [43] Chen, Jiamin, et al. "Beyond Intuition: Rethinking Token Attributions inside Transformers." *Transactions on Machine Learning Research* (2022).

- [44] Kashefi, Rojina, et al. .^Explainability of Vision Transformers: A Comprehensive Review and New Perspectives..^arXiv preprint arXiv:2311.06786 (2023).