Modelos Fundacionales en Imágenes Médicas: MI-Zero y CheXZero en la detección de Cáncer en histopatologías y radiografías

Laura Nicole Bermúdez Santa
Department of Mathematics
National University of Colombia
Bogotá D.C., Colombia
labermudezs@unal.edu.co

Juan Sebastián Castro Pardo
Department of Mathematics
National University of Colombia
Bogotá D.C., Colombia
jcastropa@unal.edu.co

Nicolás Álvarez Triana
Department of Mathematics
National University of Colombia
Bogotá D.C., Colombia
nialvarezt@unal.edu.co

***Resumen*—We explore the potential of foundation models (FMs) in cancer detection in histopathologies and radiographs by implementing the MI-Zero and CheXZero models. Github**

***Index Terms*—FM's, TPM, VPM, Zero-Shot, MI-Zero, foundational models, Medical imaging, WSI, IA, ChexZero**

I. Introducción

Exploramos el potencial de los FM's en la detección de cáncer en histopatologías y radiografías haciendo uso de los modelos MI-Zero y CheXZero.

La detección a temprana edad del cáncer mediante imágenes médicas es un campo relevante en la medicina moderna, identificar tumores en estados tempranos permite la aplicación de tratamiento inmediato, dando la posibilidad que este sea más eficaz y menos invasivo. Las técnicas actuales de imagen como la tomografía y la resonancia magnética son herramientas esenciales para los procesos de detección, diagnóstico, planificación de tratamiento y respuesta de terapia.

En este contexto, introducimos MI-zero y ChexZero como modelos fundacionales (FM's) diseñados para ser altamente adaptables y escalables, capaces de procesar grandes cantidades de datos no estructurados e ideales para aplicaciones médicas donde la precisión es fundamental.

II. Modelos fundacionales en imágenes médicas

Los modelos fundacionales (FM's) son un modelo de inteligencia artificial el cual en los últimos años han mostrado un avance significativo en términos de su desarrollo, se caracteriza por su entrenamiento el cual se basa en conjuntos de datos extensos y diversos utilizando técnicas de autosupervisión a gran escala dada la cantidad tan amplia y masiva de datos requeridos para su entrenamiento. A diferencia del paradigma convencional usado en el aprendizaje profundo (DL) el cual requiere de datos específicos y etiquetados para entrenar redes neuronales profundas (DNN's) los FM's pueden adaptarse fácilmente a nuevas tareas específicas con una cantidad mínima de datos. Un gran acercamiento a estos son los modelos de lenguaje a gran escala (LLM's) y los modelos de visión-lenguaje (VLM's) ambos tipos de modelo se entrenan con base en grandes conjuntos de datos y pueden aplicarse a múltiples tareas en sus respectivas especialidades.

II-A. Relevancia clínica

Por otro lado, los modelos fundacionales están teniendo un impacto significativo en la investigación y el desarrollo en el ámbito de la imagen médica, transformando la manera en la que se diseñan los métodos de investigación y se abordan los paradigmas. Considerarlos en el ámbito clínico implica:

- Precisión y consistencia en los diagnósticos al considerar detección temprana.
- Análisis rápido y automatizado considerando los grandes volúmenes de datos que puede ser procesados.
- Explicabilidad y generalización, La falta de explicabilidad en los modelos de aprendizaje profundo puede disminuir la confianza entre los clínicos y la capacidad de los modelos para generalizar adecuadamente es crucial para la aplicabilidad a distintos contextos.
- Preservación de la privacidad, los modelos fundacionales pueden transferir conocimiento sin necesidad de acceder directamente a datos sensibles, además pueden generar datos sintéticos similares a imágenes reales para entrenamiento.

II-B. Modelos fundacionales para imágenes médicas

1. **Modelos guiados por texto (TPM):**
   a) **Modelo contrastivo guiado por texto:** Actualmente, los modelos contrastivos guiados por texto están ganando fuerza en el ámbito de imágenes médicas por su capacidad de aprender representaciones semánticas a partir de imágenes médicas y sus descripciones textuales, utilizan aprendizaje contrastivo para acercar pares de imágenes y textos similares en el espacio de características y alejar pares disímiles, usados en la clasificación, segmentación y recuperación de imágenes. Estos modelos son eficaces para conectar imágenes médicas con descripciones textuales extrayendo representaciones significativas, reduce la dependencia de datos etiquetados y son adecuados para tareas predictivas Zero shot [1]. Sin embargo, el modelo presenta

[1]El objetivo principal de Zero-shot prediction o Zero-shot Learning es ganar la habilidad de predecir resultados sin ninguna muestra de entrenamiento.

limitaciones, ya que puede enfrentar imágenes médicas altamente complejas que requieres una caracterización más profunda.

b) **Modelo generativo guiado por texto:** Estos modelos están diseñados para generar imágenes médicas realistas a partir de descripciones textuales utilizan técnicas como autoencoders variacionales (VAE's) y redes adversariales generativas (GAN's) para el entendimiento de la distribución de imágenes médicas y crear nuevas muestras relacionadas. El siguiente modelo es aplicado en la generación y reproducción de imágenes específicas de enfermedades, lo cual resulta útil para la investigación y entrenamiento de modelos, al igual que para detectar anomalías. Se destaca de este modelo su capacidad para proporcionar explicaciones detalladas y apoyar decisiones clínicas, sin embargo, en términos de demanda, computación y generación de texto conciso presenta falencias.

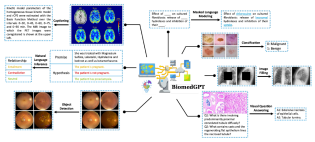
c) **Modelo hibrido guiado por texto:** Los modelos híbridos combinan enfoque generativo y contrastivo para integrar datos textuales e imágenes médicas. Adicional pueden realizar predicciones Zero-shot al alinear atributos visuales con características lingüísticas, mejorando así la precisión en cuestión de preguntas y respuestas visuales, por lo tanto, resultan valiosas para diagnósticos rápidos basados en imágenes y descripciones textuales.

d) **Modelo conversacional guiado por texto:** Modelos diseñados para facilitar diálogos interactivos entre profesionales médicos y IA permitiendo hacer preguntas, proporcionar instrucciones o buscar explicaciones. Mejoran la transferencia de conocimiento y el proceso de toma de decisiones en contextos clínicos; sin embargo, presentan grandes desafíos como la comprensión del contexto en consultas y precisión en escenarios más complejos.

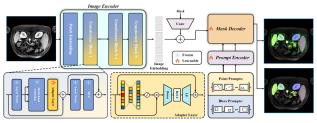
2. **Modelos guiados por imágenes (VPM):**

a) **Modelo adaptativo guiado por imágenes:** Modelos centrados en mejorar la especifidad y el rendimiento en la realización de tareas de segmentación de imágenes médicas, adaptan modelos fundacionales como Segment Anything Model (SAM) para abordar desafíos específicos en el ámbito médico. Destacan al mejorar el rendimiento del modelo en contextos donde hay escasez de datos etiquetados y abordan la segmentación de estructuras complejas en imágenes médicas. Estos modelos pueden requerir cantidades sustanciales de datos etiquetados y pueden no ser una opción viable para tareas que requieren generalización.

b) **Modelo generacional guiado por imágenes:** Modelos diseñados para manejar una amplia gama de tareas y datos médicos, como clasificación de imágenes, generación de texto y respuesta a preguntas. Estos modelos pueden adaptarse a diversas tareas sin necesidad de reentrenamiento, contribuyendo al soporte de decisiones clínicas.



(a) Modelo conversacional guiado por texto (TPM-Conversational):BioMedGPT



(b) Modelo adaptativo guiado por imágenes (VPM-Adaptations):SAM-Med2D

Figura 2: Taxonomía propuesta en [1] de los modelos fundacionales en imágenes médicas clasificados por tipos de VPM y TPM.

III. Costos y requerimientos computacionales

Al momento de implementar modelos fundaciones en el campo de imágenes médicas, los requerimientos computacionales se vuelven significativos, ya que requieren GPU's de alta gamma o aceleradores de hardware esto representa un desafío considerable. En la Figura 3 se puede observar una tabla que contrasta la información en cuanto de requerimientos computacionales de cada tipo.

IV. CheXZero

Modelo fundacional contrastivo guiado por texto e imagén de alto rendimiento con enfoque en aprendizaje Zero-shot autosupervisado, CheXZero no requiere de etiquetas manuales para la interpretación de imágenes de rayos X de tórax. Usa pares de imagen-texto para aprender representaciones que permiten clasificaciones en múltiples etiquetas.

ChexZero no requiere etiquetas excepto para el testeo y es capaz de identificar con precisión patologías sin anotaciones explícitas.

ID	Category	Sub-category	Short name	GPU Model	Number of GPUs	GPU Memory (GB)	Total GPU Memory (GB)	Training Time (GPU Hour)	Input Size	Total batch size	Epochs
1	TPM	Contrastive	MedCLIP	Nvidia RTX 3090	1	24	24	8	224x224	100	10
2	TPM	Contrastive	BioViL-T	Nvidia Tesla V100	8	32	256	-	448x448	240	50, 100
3	TPM	Contrastive	CLIPDM-OTS	NVIDIA RTX A5000	8	24	192	-	96x96x96	42	50
4	TPM	Contrastive	PTUnifier	NVIDIA A100	4	80	320	-	288x288-384x384	16-128	11-60
5	TPM	Contrastive	BiomedCLIP	NVIDIA A100	16	40	640	-	224x224-336x336	4k-64k (context)	40
6	TPM	Contrastive	KoBo	Nvidia RTX 3090	2	24	48	-	-	100	50
7	TPM	Contrastive	MI-Zero	NVIDIA A100	8	80	640	-	448x448	512	50
8	TPM	Contrastive	CITE	GeForce GTX 2080 Ti	2	11	22	0.37	224x224	128	(1000 iteration)
9	TPM	Generative	Clinical-BERT	Nvidia RTX 3090	2	24	48	96	224x224	256	50
10	TPM	Generative	Med-Flamingo	Nvidia A100	8	80	640	1296	-	400	-
11	TPM	Hybrid	MedBLIP	Nvidia RTX 3090	1	24	24	-	224x224x224	7	100
12	TPM	Hybrid	VLM for VQA in MI	GeForce GTX 1080 Ti	1	11	11	-	224x224	50	50
13	TPM	Conversational	DoID-GPT	Nvidia RTX 3090	>1	24					
14	TPM	Conversational	ChatDoctor	Nvidia A100	6	80	480	18	max-sq-len: 2048	192	3
15	TPM	Conversational	PMC-LLaMA	Nvidia A100	32	80	2560	-	max-sq-len: 2048	img:256, text:3200	8
16	TPM	Conversational	LLaVA-Med	Nvidia A100	8	40	320	120	-	128	100
17	TPM	Conversational	Radiology-Llama2	Nvidia A100	4	80	320	-	-	128	-
18	VPM	Adaptations	SAMed	Nvidia RTX 3090	2	24	48	-	512x512	12	200
19	VPM	Adaptations	MedSAM	Nvidia A100	20	80	1600	-	1024x2014	160	100
20	VPM	Adaptations	AutoSAM	NVIDIA Tesla V100	1	16	16	-	1024x1024	4	120
21	VPM	Adaptations	LVM-Med	Nvidia A100	16	80	1280	2688	224x224 - 1024x1024	16 - 64	20-200
22	VPM	Adaptations	SAM-Med2D	Nvidia A100	8	80	640	-	256x256	-	12
23	VPM	Generalist	SAM-B-ZSS	Nvidia RTX 3080	1	10	10	-	1024x1024	1	20
24	VPM	Generalist	RadFM	Nvidia A100	32	80	2560	-	256(3D), 512(2D)	1(3D), 4(2D)	8
25	VPM	Generalist	RETFound	Nvidia A100	8	40	320	2688	16x16	16, 1792	50, 800

Figura 3: Tabla resumen extraída de [1] de la demanda computación de cada tipo de modelo fundación según su categoria.

> *"We show that the performance of the self-supervised method is comparable to the performance of both expert radiologists and fully supervised methods on unseen pathologies in two independent test datasets collected from two different countries."* [2]

Se ha mostrado que no se requieren etiquetas explícitas para desempeñar bien tareas de interpretación de radiografías de tórax cuando los informes asociados a las imágenes médicas están disponibles.

IV-A. Arquitectura CheXZero

En la figura 4 se puede observar la arquitectura de CheXZero en un breve esquema de entrenamiento y funcionamiento para la clasificación Zero-shot. La arquitectura CheXZero

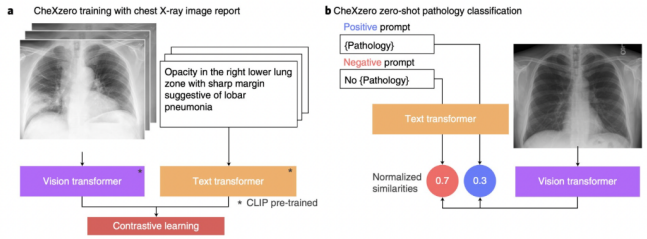


Figura 4: **a. Entrenamiento:** el modelo aprende características a partir de informes de radiologias **b. Predicción de patologías en radiografías de tórax:** Para cada patología se genera un aviso positivo y negativo. El método autosupervisado calcula un puntaje de probabilidad para la patología y con base en esta se realiza la clasificación.

utiliza una Vision transformer ViT-B/32

[2]para analizar imágenes de rayos X y un transformador para procesar el texto, este transformador procesa descripciones textuales asociadas con las imágenes, como informes radiológicos o etiquetas de patologías. La representación de texto y la longitud máxima de token permiten que el modelo maneje una variedad de descripciones y etiquetas [3] asociado a la imagen médica, el modelo detecta patologías y clasificarlas al correlacionar las características visuales extraídas por el transformador con las descripciones textuales procesadas por el transformador de texto el uso de CLIP (Contrastive language image pre training) como esquema de inicialización ayuda a garantizar una base sólida para el aprendizaje imagen-texto. El conjunto de datos obtenidos tras la inicialización del modelo se almacena en un archivo HDF5 y se realiza una búsqueda de hiperparámetros ajustando la tasa de aprendizaje y el número de etapas, con base en esto se realiza la evaluación del modelo, obteniendo así con descenso de gradiente estocástico una mayor ganancia con una tasa de aprendizaje de 0.0001 y un momentum de 0.9, entrenamiento de cuatro etapas y el batch size de 64

V. MI-Zero

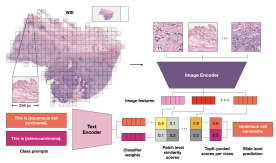
MI-Zero es un framework innovador que explota las capacidades de zero-shot en modelos contrastivos de imagen y texto aplicados a histopatología. Permite realizar múltiples tareas de diagnóstico usando codificadores preentrenados sin necesidad de etiquetas adicionales. Para alcanzar un nivel clínico en estos modelos, es esencial seguir un ciclo de vida de desarrollo bien estructurado.

V-A. Ciclo de Vida del Modelo

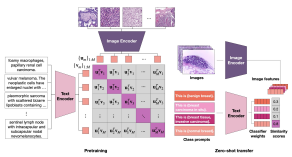
Para obtener modelos de grado clínico, se sigue el siguiente ciclo de vida:

1. **Recopilación de Datos**: Obtener más de 1000 imágenes diagnósticas completas con etiquetas clínicas.
2. **Tokenización de Imágenes**: Desglosar las imágenes de lámina completa en secuencias de parches detallados.
3. **Entrenamiento de Clasificadores**: Usar las etiquetas para entrenar un clasificador de láminas que agregue los parches detallados y realice predicciones.
4. **Despliegue Clínico**: Transferir el clasificador de láminas para su uso en clínica.

La transferencia zero-shot permite que un modelo generalice y aplique conocimientos aprendidos de otras tareas para manejar situaciones nuevas sin entrenamiento adicional.



(a) Arquitectura MI-Zero [5] i



(b) Arquitectura MI-Zero [5] ii

[2]El Vision Transformer es un tipo de red neuronal que ha demostrado ser eficaz en tareas de visión por computadora para la extracción de características visuales. ViT-B/32 acepta imágenes con una resolución de 224x224 píxeles.

[3]Procesa descripciones textuales, la representación de texto y la longitud máxima de Token permiten que el modelo maneje una variedad de descripciones y etiquetas. En [2] se usan 12 capas con un tamaño base de 63 millones de parámetros y una longitud máxima de 77 tokens

V-B. Pre entrenamiento sin Supervisión de los Codificadores Unimodales

V-C. Codificadores de Imagen

1. Inicializar con pesos preentrenados de ImageNet.
2. Usar una SOTA que codifica parches de imágenes de histopatologías no etiquetadas.

V-D. Codificadores de Texto

Recopilar un glosario específico en el campo de patologías, incluyendo diagnósticos reportados por hospitales y hepatologías relevantes.

Dada una secuencia de T palabras tokenizadas $w_1, w_2, \cdots, w_T$, se maximiza la log-verosimilitud bajo un modelo autorregresivo generativo parametrizado por θ:

$$L_{clm}(\theta) = -\sum_{t=1}^{T+1} \log p(w_t | w_{0:t-1}; \theta)$$

V-E. Alineación de Visión y Lenguaje

Para alinear los espacios latentes de visión y lenguaje, se utiliza una pérdida contrastiva cruzada con una constante M (tamaño de lotes de imágenes-texto) y un parámetro de temperatura τ.

Dado un lote de M pares de muestras, las representaciones vectoriales $l2$ normalizadas se computan vía dos codificadores $f(\cdot;\theta)$ y $g(\cdot;\phi)$ respectivamente:

$$u_m = \frac{f(x_m;\theta)}{||f(x_m;\theta)||}, \quad v_m = \frac{g(t_m;\phi)}{||g(t_m;\phi)||}$$

Las direcciones de imagen a texto y texto a imagen se consideran simétricas con igual peso durante el entrenamiento del modelo:

$$L_{i2t}(\theta,\phi) = -\sum_{i=1}^{M} \log \frac{\exp(\tau u_i^T v_i)}{\sum_{j=1}^{M} \exp(\tau u_i^T v_j)}$$

$$L_{t2i}(\theta,\phi) = -\sum_{j=1}^{M} \log \frac{\exp(\tau v_j^T u_j)}{\sum_{i=1}^{M} \exp(\tau v_j^T u_i)}$$

V-F. Transferencia Zero-Shot para Clasificación de Imágenes

Para cada clase de interés, se crea un prompt con el nombre de la clase (ejemplo, "adenocarcinoma") y una plantilla (ejemplo, "una imagen mostrando ").

Usando este prompt, se crea una representación vectorial de pesos y, tomando la representación vectorial de una imagen, la decisión de clasificación del modelo se calcula como:

$$\hat{y} = \arg\max_m u^T w_m$$

V-G. Transferencia Zero-Shot para WSIs Megapíxeles

Se divide cada región del tejido en cada WSI en N parches, y se calcula una representación vectorial normalizada para cada parche usando el codificador de imagen para obtener $\{u_i\}_{i=1,\ldots,N}$. Siguiendo el prompt de clasificación mencionado anteriormente, se computan los resultados $\{s_i\}_{i=1\ldots,N}$ donde $s_i = u_i^T[w_1, w_2, \ldots, w_C]$ (siendo C el número de clases).

V-H. Enfoques para Calcular Puntuaciones de Predicción a Nivel de Lámina

1. **Enfoque Basado en Conjuntos**: Utiliza operadores invariantes a la permutación como la media o el topK max-pooling para combinar las puntuaciones de los parches.
2. **Enfoque Basado en Grafos**: Considera las relaciones espaciales entre los parches, suaviza las puntuaciones utilizando un filtro de media y luego aplica un operador invariante a la permutación a las puntuaciones suavizadas.

Ambos enfoques buscan combinar las puntuaciones de los parches de manera robusta a la permutación y sin necesidad de actualizar parámetros, esencial para la transferencia zero-shot.

VI. Conclusiones y consideraciones

1. Estos modelos (MI-Zero y CheXZero) requieren de herramientas que optimicen el manejo de imágenes. Difieren en el procesamiento de estas, ya que MI-zero realiza parches de la imagen dado que estas pueden llegar a ser demasiado pesadas, mientras que el modelo CheXZero se hace un reajuste al tamaño de la imagen para aplicar el transformador, esto resulta en un mayor requerimiento computacional para el procesamiento de imágenes en MI-Zero con base en esto se considera realizar un análisis profundo en la implementación de este modelo. Por otro lado, a lo largo de la implementación de estos encontramos requerimiento en cuanto a sistema operativo y GPU que no fueron posibles satisfacer para su implementación.
2. MI-Zero y cheXZero, su capacidad para igualar e incluso superar el rendimiento humano en ciertas tareas, sugiere que estos modelos podrían revolucionar la patología computacional y otros campos relacionados con imágenes de alta resolución. Futuros esfuerzos deben centrarse en la recolección de datos de mayor calidad y en la mejora de las técnicas de pre entrenamiento para maximizar el impacto de estos modelos en la práctica clínica y más allá.
3. ChexZero evita la necesidad de esfuerzos de etiquetado manuales y que consumen mucho tiempo, dado que permite la detección automática de patologías en radiografías de tórax sin anotaciones explícitas.
4. En [2] se demostró un alto rendimiento (AUC ¿0.9) en 14 hallazgos y al menos 0.700 en 53 hallazgos de 107 hallazgos radiográficos que el método no había visto

durante el entrenamiento, respaldando su precisión en el proceso de detección.

REFERENCIAS

[1] B. Azad et al., "Foundational models in medical imaging: A comprehensive survey and future vision", Arxiv, 2310.18689v1, octubre de 2023. [En línea]. Disponible: https://arxiv.org/pdf/2310.18689

[2] Tiu, E., Talius, E., Patel, P. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat. Biomed. Eng (2022). https://doi.org/10.1038/s41551-022-00936-9

[3] "CheXzero: Detect Pathologies From Unannotated X-ray Images". Analytics Vidhya. Accedido el 12 de julio de 2024. [En línea]. Disponible: https://www.analyticsvidhya.com/blog/2022/10/chexzero-detect-pathologies-from-unannotated-x-ray-images/

[4] Pai, S., Bontempi, D., Hadzic, I. et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell 6, 354–367 (2024). https://doi.org/10.1038/s42256-024-00807-9

[5] Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, Faisal Mahmood; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19764-19775