Jornadas de Automática

Aplicación de grandes modelos de lenguaje en diagnóstico de glaucoma

Hernández, Jorge [a,*], Barrios, Eduardo J. [a], Alayón, Silvia [a], Diaz-Aleman, Tinguaro [b]

[a] *Dpto. de Ingeniería Informática y de Sistemas, Escuela Superior de Ingeniería y Tecnología, Camino San Francisco de Paula, nº 19, 38200, San Cristóbal de la Laguna, Santa Cruz de Tenerife, España.*
[b] *Departamento de Oftalmología, Hospital Universitario de Canarias, 38320 Santa Cruz de Tenerife, España.*

To cite this article: Hernández, Jorge, Barrios, Eduardo J., Alayón, Silvia, Diaz-Aleman, Tinguaro. 2025. Application of Large Language Models in glaucoma diagnosis. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12085

Resumen

En esta investigación se explora el potencial de los Grandes Modelos de Lenguaje-Visión (Visual LLM) en el diagnóstico de glaucoma a partir de retinografías. En concreto, se investiga el uso de la Visual LLM conocida como Moondream. Empleando técnicas de transferencia de aprendizaje, el modelo se ha re-entrenado con imágenes de retinografías, con el objetivo de que aprenda a distinguir ojos sanos y ojos con signos glaucomatosos. La metodología diseñada combina la extracción de características visuales y el razonamiento textual, abriendo nuevas vías para la interpretación clínica automatizada. Este trabajo sitúa a los Visual LLMs como una opción atractiva para integrar la Inteligencia Artificial multimodal en Oftalmología y mejorar la detección del glaucoma.

Palabras clave: Soporte a la toma de decisiones, Imágenes médicas y procesamiento, Identificación y validación, Formulación de modelos y diseño de experimentos, Procesamiento y sistemas de imágenes biomédicas y médicas

Application of Large Language Models in glaucoma diagnosis

Abstract

This research explores the potential of Visual Large Language-Language Models (Visual LLM) in the diagnosis of glaucoma from retinographies. Specifically, the use of the Visual LLM known as Moondream is analysed. Using transfer learning techniques, the model has been re-trained with retinal images, with the aim of learning to distinguish between healthy eyes and eyes with glaucomatous signs. The designed methodology combines visual feature extraction and textual reasoning, opening new ways for automated clinical interpretation. This work positions Visual LLMs as an attractive option for integrating multimodal Artificial Intelligence in Ophthalmology and improving glaucoma detection.

Keywords: Decision support, Medical imaging and processing, Identification and validation, Model formulation and experiment design, Biomedical and medical image processing and systems.

1. Introducción

El glaucoma es una enfermedad ocular que daña el nervio óptico provocando una pérdida de visión progresiva que puede derivar en ceguera. Actualmente es la principal causa de ceguera irreversible en el mundo. La detección temprana permite frenar su progresión, sin embargo, esta detección es complicada debido a que las fases iniciales de la enfermedad son asintomáticas (Jonas et al., 2017).

Los especialistas médicos analizan visualmente las imágenes del fondo de ojo (retinografía) para buscar signos de glaucoma. En concreto, se centran en la inspección de la cabeza del nervio óptico (CNO), que es la región por la que emergen los axones de las células ganglionares de la retina y por la que discurren los vasos sanguíneos retinianos. Esta región anatómica es la que más se altera en caso de glaucoma. En las retinografías, la cabeza del nervio óptico se distingue con claridad, siendo sus principales partes el disco y la copa

*Autor para correspondencia: jhernanv@ull.edu.es

óptica (Wang et al., 2021), tal y como se muestra en la figura 1.

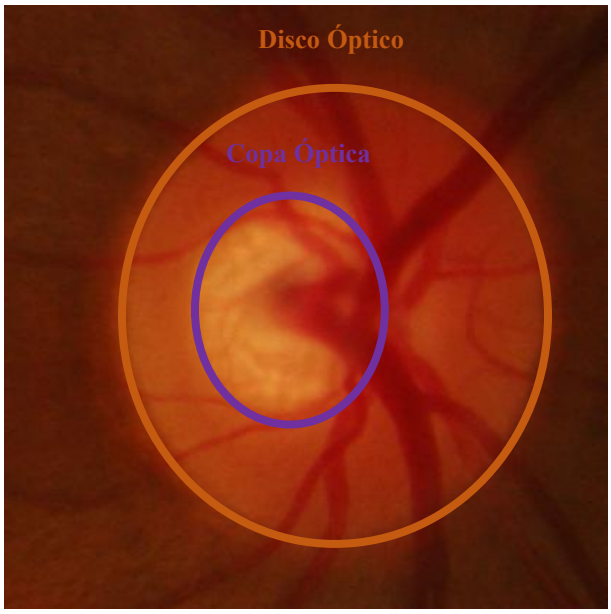

Figura 1 – Retinografía de CNO con disco y copa señalados

El análisis visual de estas imágenes es una tarea complicada y subjetiva, que depende mucho de la experiencia del médico. Por este motivo, la aplicación de herramientas basadas en Inteligencia Artificial para la ayuda al diagnóstico médico es cada vez más frecuente: para mejorar el diagnóstico por imagen (radiología, patología, oftalmología), apoyar decisiones clínicas, personalizar tratamientos con datos genómicos, acelerar el descubrimiento de fármacos, extraer y procesar información de historias clínicas, analizar señales fisiológicas (ECG, EEG), etc. (Mienye et al., 2024), (Rajpurkar et al., 2022).

En el caso concreto abordado, la detección del glaucoma a partir de imágenes de fondo de ojo, podemos encontrar varios trabajos en la literatura científica que proponen el uso de diversas arquitecturas de redes neuronales. Las Redes Convolucionales (Convolutional Neural Network, CNN) se han consolidado en este campo gracias a su capacidad para extraer características discriminativas de estructuras retinianas complejas. Por otro lado, los Transformadores de Visión (Vision Transformers, ViT), desarrollados más recientemente, han comenzado a aplicarse en este campo, logrando resultados similares o incluso superiores a los obtenidos por las CNNs (Chen et al., 2015), (Wassel et al., 2022), (Fan et al., 2023).

En los últimos años han emergido Grandes Modelos de Lenguaje (LLM) multimodales, capaces de procesar simultáneamente imágenes y texto, conocidos como Visual LLMs. Estas arquitecturas integran la potencia de extracción de características de los ViTs con el razonamiento contextual y la generación de lenguaje natural de los LLMs (Van et al., 2024). En este trabajo se estudiará la aplicación de estos modelos al diagnóstico del glaucoma.

La aplicación de los Visual LLMs multimodales en medicina es tan reciente, que son pocos los trabajos publicados hasta la fecha. Uno de ellos propone un sistema, denominado SkinGPT-4, para el diagnóstico de enfermedades cutáneas, con el uso de un modelo visual alineado con un LLM para describir características clínicas en lenguaje natural. SkinGPT-4 mostró una precisión diagnóstica comparable a la de los especialistas, al tiempo que redujo los tiempos de respuesta y generó informes interactivos comprensibles para el paciente (Zhou et al., 2024). Más relacionado con el problema abordado, es el trabajo presentado en (Tan et al., 2024), en el que emplean GPT-4 para evaluar de forma automática varios LLM (GPT3.5 y diferentes variantes de LLAMA2) entrenados con 368 pares de preguntas y respuestas en oftalmología. En el estudio demostraron la capacidad de GPT-4 para valorar otros modelos, dada la alta correlación de sus juicios con los de expertos humanos. Además, hallaron que, entre los modelos ajustados, GPT-3.5 obtuvo mejores resultados que las variantes de LLAMA2.

En las siguientes secciones se describirán las herramientas y las bases de imágenes utilizadas en el presente estudio, el diseño de los experimentos y los resultados encontrados. Finalmente, se ofrecerán las conclusiones y las líneas que deja abiertas el trabajo.

2. Materiales y Métodos

En esta sección se describirá la base de retinografías que se ha utilizado, el modelo Visual LLM concreto elegido, y el desarrollo experimental diseñado.

2.1. Descripción del base de retinografías

El conjunto de datos utilizado para entrenar y evaluar el modelo está compuesto por:

- La base de datos pública Rim-ONE DL (Batista et al., 2020), que incluye 172 imágenes de ojos con glaucoma y 313 de ojos normales.
- Imágenes de fondo de ojo recopiladas en el Hospital Universitario de Canarias. Este lote consta de 191 imágenes de ojos con glaucoma y 63 de ojos normales. Estas imágenes, que no son públicas, se adquirieron con el retinógrafo no midriático multifuncional Topcon TRC-NW8. El estudio se llevó a cabo conforme a la Declaración de Helsinki y fue aprobado por el Comité de Ética de Investigación del Hospital Universitario de Canarias (CHUC_2023_41, 27 de abril de 2023). Se garantizó la confidencialidad de los datos personales.

En total se emplearon 739 imágenes: 363 retinografías de ojos con glaucoma y 376 de ojos normales.

2.2. Moondream

Moondream (Moondream AI, 2024) es un Visual LLMs de código abierto que comprende imágenes mediante sencillas instrucciones de texto. Es rápido, muy capaz y solo ocupa 1 GB. Por estas características ha sido seleccionado en el presente trabajo.

Este modelo presenta una arquitectura encoder–decoder, esta arquitectura se ha desarrollado en los modelos de procesamiento de lenguaje natural conocidos como Transformadores (Transformers) (Vaswani et al., 2017). Combina un encoder convolucional multi-escala con un

decoder de texto, conectados mediante atención cruzada para fusionar información visual y lingüística en un espacio semántico común.

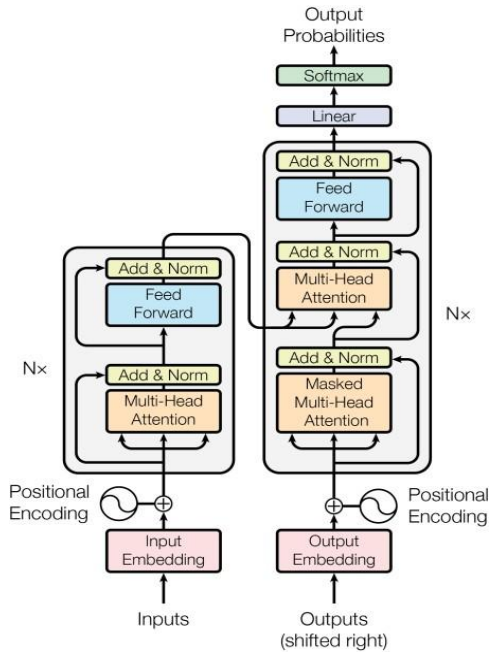

Figura 2. Encoder-Decoder (Vaswani et al., 2017)

Su encoder extrae características a distintas resoluciones y emplea saltos en las conexiones del estilo U-Net para preservar detalles finos, mientras que el decoder genera explicaciones en lenguaje natural centradas en regiones relevantes de la imagen.

Entrenado end-to-end con objetivos contrastivos imagentexto y de generación lingüística (por ejemplo, entropía cruzada más pérdida perceptual), y optimizado con AdamW, Moondream combina robustez al ruido, eficiencia computacional y elevada capacidad de generalización, lo que lo hace ideal para aplicaciones clínicas de diagnóstico y descripción automatizada (Moondream AI, 2024).

2.3. Inserción de tablas y márgenes de página

En este apartado se describe la metodología: la composición de los conjuntos de entrenamiento y de test, la estrategia de entrenamiento, y el valor seleccionado para los parámetros.

Se han realizado dos entrenamientos del modelo Moondream, la diferencia entre ellos es el número de épocas (10 y 20). Estos entrenamientos aprovechan la transferencia de aprendizaje partiendo del modelo pre-entrenado, y realizan un ajuste de precisión, o ajuste fino (fine tuning) con las imágenes de retinografías.

El conjunto disponible de retinografías, compuesto por un total de 739 imágenes, se dividió aleatoriamente en conjuntos de entrenamiento y test con una proporción del 80% y del 20%, respectivamente:

- Conjunto de entrenamiento: 290 retinografías de glaucoma y 301 de ojos sanos.
- Conjunto de test: 73 retinografías de glaucoma y 75 de ojos sanos.

Cada retinografía está etiquetada como “normal” o “glaucoma” por varios expertos médicos. A estas etiquetas iniciales, compuestas por una sola palabra, se le añade una breve descripción genérica de lo que se ve en la imagen cuando es normal o cuando tiene glaucoma, la misma para todas las imágenes. El objetivo de esta extensión es dotar de más contexto lingüístico al modelo. Las nuevas etiquetas son las siguientes:

- Para imágenes con glaucoma: “Yes. The optic disc shows signs of glaucomatous damage such as increased cup-to-disc ratio and thinning of the neuroretinal rim”.
- Para imágenes normales: “No. The optic disc appears healthy with a regular neuroretinal rim and no signs of cupping”.

Los parámetros utilizados en los entrenamientos se muestran de manera resumida en la tabla 1.

Tabla 1: Parámetros de los entrenamientos

Hiperparámetro	Valor utilizado
Input size	378 x 378
Learning Rate (LR)	5e-7
Epochs	10 / 20
Gradient Accumulation Steps	8
Learning Rate Scheduler	Warm-up lineal (10%) + Coseno (90%)
Optimizador	AdamW (β_1=0.9, β_2=0.95, ε=1e-6, weight decay=0.01)
Loss function	Cross-Entropy sobre logits generados
Tamaño original del batch	1
Tamaño efectivo del batch	8 (por acumulación de gradientes)

Para valorar la eficiencia del Visual LLM, se han calculado las siguientes métricas (Brzezinski et al., 2018):

- Precisión: proporción de todas las predicciones que el modelo clasifica correctamente, tanto positivas como negativas, respecto al total de muestras.
- Sensibilidad: porcentaje de imágenes con glaucoma que el modelo identifica correctamente, respecto al total de imágenes de glaucoma.
- Especificidad: porcentaje de imágenes normales que el modelo identifica correctamente, respecto al total de imágenes normales.
- Precisión balanceada: promedio entre sensibilidad y especificidad, que refleja la exactitud del modelo considerando ambas clases por igual.
- F_1-score: métrica que combina el valor predictivo positivo y sensibilidad en un solo valor, penalizando los desequilibrios entre falsos positivos y falsos negativos.

3. Resultados

Los resultados obtenidos tras realizar dos ajustes finos diferentes del modelo Moondream con las imágenes de retinografías se muestran en la tabla 2.

Tabla 2: Resultados experimentales. Experimento 1: 10 épocas, Experimento 2: 20 épocas, P: precisión, S: sensibilidad, E: especificidad, PB: precisión balanceada, F1:F1-score

Experimento	P(%)	S	E	PB(%)	F1
1	83,13	1	0,666	83,33	0,8538
2	85,63	0,9310	0,800	86,55	0,8710

La duración aproximada de los entrenamientos fue de 8 y 16 horas para los experimentos 1 y 2 respectivamente. Fueron realizados en un equipo dotado con una GPU RTX4070.

De los resultados obtenidos se pueden extraer dos conclusiones diferentes:

1) El modelo es capaz de distinguir imágenes con una precisión aceptable.
2) Esta precisión podría ser mayor si se entrenara el modelo durante más épocas.

4. Discusión

Los resultados obtenidos en dos experimentos de entrenamiento con ajuste fino (de 10 y 20 épocas) muestran que la arquitectura encoder–decoder de Moondream es capaz de aprender representaciones discriminativas del nervio óptico y del ratio copa/disco, alcanzando una precisión balanceada superior al 84 % y un F_1-score de hasta 0,86 tras 20 épocas.

Por limitaciones de hardware no se han podido realizar experimentos más largos (con más épocas). La tendencia observada en los resultados indica que alargar el entrenamiento podría mejorar los resultados del modelo.

El desempeño conseguido con 20 épocas (experimento 2) es comparable al de numerosas CNNs y ViTs reportadas en la literatura. Al evaluar nuestro rendimiento frente al de algunas CNNs, observamos que en el artículo (Sallam et al., 2021) se obtuvo una precisión del 86,9 % sobre la base de datos de Glaucoma basado en atención a gran escala(LAG) con su mejor red neuronal. En la investigación de (Li et al., 2020), se alcanzó una precisión del 96,2 % sobre LAG, aunque en el conjunto RIM-ONE DL esta disminuyó al 85,2 %, lo cual es comprensible considerando que dicha base de datos no formaba parte del conjunto de entrenamiento. Por su parte, en el artículo de (Haouli et al., 2023) se trabajó tanto con CNNs como con ViTs sobre diferentes bases de datos, logrando precisiones superiores al 92 % con ViTs y al 88 % con CNNs.

Si bien nuestras precisiones no alcanzan estos valores, es importante destacar el valor añadido que aportan los modelos VLLMs (Vision-Language Large Models). Estos modelos no solo permiten realizar tareas de clasificación, sino que también generan descripciones en lenguaje natural sobre el contenido de las imágenes. Esta capacidad facilita la interpretación de los resultados, aportando un nivel de explicabilidad que va más allá de la simple predicción numérica. En contextos clínicos, las descripciones generadas pueden ayudar a los profesionales a comprender por qué un modelo ha detectado ciertas anomalías facilitando la toma de decisiones médicas.

No ha sido posible realizar una comparativa con otros modelos LLM porque no hemos encontrado ningún VLLM ajustado para el problema de glaucoma. De hecho, el ajuste fino de VLLMs es algo muy reciente, cuyo potencial aún no ha sido explorado en muchos campos.

5. Conclusiones

En este trabajo se ha explorado el potencial de los modelos visual LLMs para la detección automática de glaucoma a partir de retinografías. Concretamente, se ha reentrenado un modelo de acceso abierto, y de tamaño pequeño (Moondream) con retinografías de ojos sanos y glaucomatosos.

El objetivo de este trabajo es comprobar si un Visual LLM se puede adaptar a un problema de diagnóstico médico. Los resultados encontrados indican que sí, aunque quedan muchas líneas abiertas que explorar:

- Ampliar y diversificar la base de datos, incorporando retinografías de múltiples centros, dispositivos y grupos étnicos, así como anotaciones detalladas de expertos sobre distintos grados de glaucoma.
- Comparar otras arquitecturas Visual LLM y variantes de Moondream, así como técnicas de fusión multimodal que incluyan datos clínicos (presión intraocular, OCT, historial del paciente) para enriquecer el diagnóstico.
- Investigar la influencia de las etiquetas en el comportamiento del modelo.

En conclusión, Moondream demuestra ser una herramienta viable y prometedora para el apoyo al diagnóstico precoz de glaucoma. El fortalecimiento de bases de datos, la validación externa y la incorporación de fuentes de información adicionales serán pasos clave para llevar esta tecnología desde el laboratorio hasta la práctica clínica, contribuyendo así a reducir la ceguera irreversible por glaucoma.

Como futura línea de investigación, se podría comparar el modelo Moondream que hemos entrenado con otros VLLMs diferentes, entrenados con las mismas imágenes, para poder realizar una comparativa de eficiencia fiable.

Agradecimientos

Trabajo cofinanciado por la Agencia Canaria de Investigación, Innovación y Sociedad de la Información de la Consejería de Universidades, Ciencia e Innovación y Cultura y por el Fondo Social Europeo Plus (FSE+) Programa Operativo Integrado de Canarias 2021-2027, Eje 3 Tema Prioritario 74 (85%). Código de subvención TESIS2022010056.

Proyecto (PROID2024010027) financiado por la Agencia Canaria de Investigación Innovación y Sociedad de la Información (ACIISI) y por el Fondo Europeo de Desarrollo Regional en el marco del programa FEDER Canarias 2021-2027

Referencias

Batista, F. J. F., Diaz-Aleman, T., Sigut, J., Alayon, S., Arnay, R. & AngelPereira, D., 2020. RIM-ONE DL: A unified retinal image database for assessing glaucoma using deep learning. Image Analysis & Stereology 39, 161–167. DOI: 10.5566/ias.2346

Brzezinski, D., Stefanowski, J., Susmaga, R. & Szczęch, I., 2018. Visualbased analysis of classification measures and their properties for class imbalanced problems. Information Sciences 462, 242–261. DOI: 10.1016/j.ins.2018.06.020

Chen, Y., Xu, D. W., Kee Wong, T. Y., Wong, J. & Liu, J., 2015. Glaucoma detection based on deep convolutional neural network. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 715–718. DOI: 10.1109/EMBC.2015.7318462

Fan, R. et al., 2023. Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization. Ophthalmology Science 3, 100233. DOI: 10.1016/j.xops.2022.100233

Haouli, I.-E., Hariri, W., Seridi-Bouchelaghem, H., 2023. Exploring Vision Transformers for Automated Glaucoma Disease Diagnosis in Fundus Images, in: 2023 International Conference on Decision Aid Sciences and Applications (DASA). pp. 520–524. DOI: 10.1109/DASA59624.2023.10286714

Jonas, J. B., Aung, T., Bourne, R. R., Bron, A. M., Ritch, R. & Panda-Jonas, S., 2017. Glaucoma. The Lancet 390, 2183–2193. DOI: 10.1016/S0140-6736(17)31469-1

Li, L., Xu, M., Liu, H., Li, Y., Wang, X., Jiang, L., Wang, Z., Fan, X., Wang, N., 2020. A Large-Scale Database and a CNN Model for Attention-Based Glaucoma Detection. IEEE Transactions on Medical Imaging 39, 413–424. DOI: 10.1109/TMI.2019.2927226

Mienye, I. D. et al., 2024. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. Informatics in Medicine Unlocked 51, 101587. DOI: 10.1016/j.imu.2024.101587

Moondream AI, 2024. Moondream.ai. [En línea]. Disponible en: https://moondream.ai/

Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J., 2022. AI in health and medicine. Nature Medicine 28, 31–38. DOI: 10.1038/s41591-021-01614-0

Sallam, A., Gaid, A.S.A., Saif, W.Q.A., Kaid, H.A.S., Abdulkareem, R.A., Ahmed, K.J.A., Saeed, A.Y.A., Radman, A., 2021. Early Detection of Glaucoma using Transfer Learning from Pre-trained CNN Models, in: 2021 International Conference of Technology, Science and Administration (ICTSA). pp. 1–5. DOI: 10.1109/ICTSA52017.2021.9406522

Tan, T., Elangovan, K. & Ting, D., 2024. Fine-tuning large language model (LLM) artificial intelligence chatbots in ophthalmology and LLM-based evaluation using GPT-4. arXiv preprint 2402.10083. DOI: 10.48550/arXiv.2402.10083

Van, M.-H., Verma, P. & Wu, X., 2024. On large visual language models for medical imaging analysis: An empirical study. DOI: arXiv:2402.14162

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. DOI: arXiv:1706.03762

Wang, Y. X., Panda-Jonas, S. & Jonas, J. B., 2021. Optic nerve head anatomy in myopia and glaucoma, including parapapillary zones alpha, beta, gamma and delta: Histology and clinical features. Progress in Retinal and Eye Research 83, 100933. DOI: 10.1016/j.preteyeres.2020.100933

Wassel, M., Hamdi, A. M., Adly, N. & Torki, M., 2022. Vision Transformers based classification for glaucomatous eye condition. In: 26th International Conference on Pattern Recognition (ICPR), pp. 5082– 5088. DOI: 10.1109/ICPR56361.2022.9956086

Zhou, J. et al., 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nature Communications 15, 5649. DOI: 10.1038/s41467-024-50043-3