Universidad Católica
San Pablo

FACULTAD DE INGENIERÍA Y COMPUTACIÓN

DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

Escuela Profesional de Ciencia de la Computación

Optimización Computacional de Vision Transformer (ViT) para el Cribado Eficiente del Glaucoma

Tesis

Presentada por el bachiller

Rodrigo Alonso Torres Sotomayor

Para Optar por el Título profesional de:

Licenciado en Ciencia de la Computación

Asesor: Dr. Juan Carlos Gutiérrez Cáceres

Arequipa, 2024

Agradezco a Dios por todas sus bendiciones, a mis profesores por su invaluable enseñanza, y a mis padres, hermano y algunos amigos.

Abreviaturas

ViT Vision Transformer

CNN Redes Convolucionales

CLAHE Ecualización de Histograma Adaptativo de Contraste Limitado

Agradecimientos

Quiero expresar mi gratitud a Dios por haberme guiado a lo largo de estos cinco años de estudio.

Agradezco profundamente a mis padres y a mi hermano por su constante apoyo y por haberme ayudado a formarme como profesional.

Agradezco a mi *alma mater*, la universidad, por haberme acogido y proporcionado la educación necesaria que me permitirá contribuir positivamente a la sociedad.

Quiero agradecer de manera especial a mi asesor, el Dr. Juan Carlos Gutiérrez Cáceres, por su orientación invaluable durante el desarrollo de esta tesis.

Resumen

En la era actual, las redes transformer son prometedoras para predicciones y clasificación, aprovechando técnicas avanzadas para identificar patrones en grandes conjuntos de datos. Originalmente desarrolladas para procesamiento del lenguaje natural, han demostrado un rendimiento excepcional en tareas como traducción automática,y su habilidad para capturar relaciones a larga distancia las hace igualmente efectivas para el procesamiento de imágenes. Esta tesis se centra en una variante específica para imágenes: Vision Transformer (ViT), esta transforma una imagen en una secuencia de tokens procesados mediante capas transformer, capturando información local y global. Este enfoque ha mostrado resultados prometedores, desafiando las Redes Convolucionales (CNN) en muchas tareas de visión por computadora. Sin embargo, ViT presenta un costo computacional significativo debido a su estructura basada en atención, limitando su aplicación en entornos con recursos limitados. Esta tesis explora técnicas para mitigar estos desafíos sin comprometer la precisión en la detección de enfermedades, como el glaucoma. El glaucoma, una enfermedad ocular prevalente, requiere detección temprana debido a su naturaleza asintomática inicial. Aprovechando la capacidad de ViT para capturar patrones globales y adaptarse a diferentes tamaños de entrada, esta investigación desarrolla un sistema eficiente y preciso para la detección temprana del glaucoma. Los resultados preliminares indican que es posible lograr un rendimiento comparable o superior con una arquitectura de ViT, siendo esta nueva arquitectura más eficiente y con menor costo computacional.

Abstract

Las redes Vision Transformer (ViT) son una innovación en el procesamiento de imágenes, pero su alto costo computacional limita su aplicabilidad en entornos con recursos limitados. Esta investigación busca optimizar la arquitectura ViT, enfocándose en reducir su complejidad sin comprometer la precisión, aplicándola al cribado del glaucoma, una enfermedad ocular que requiere detección temprana. Para ello, se utilizaron ajustes en parámetros clave del modelo y se aplicó la técnica CLAHE (Contrast Limited Adaptive Histogram Equalization) para mejorar la calidad de las imágenes. El modelo ViT optimizado logró una reducción del 65 % en el tiempo de entrenamiento, pasando de 3 horas a 30 minutos por época, y alcanzó un accuracy del 87.11 %. Los resultados sugieren que la arquitectura optimizada no solo es eficiente en términos computacionales, sino también lo suficientemente precisa para ser implementada en entornos clínicos para la detección temprana del glaucoma.

Índice general

1. Introducción **2**

1.1. Motivación y Contexto 3

1.2. Planteamiento del Problema 3

1.3. Objetivos 5

1.3.1. Objetivos Específicos 5

1.4. Organización de la tesis 5

1.5. Cronograma 6

2. Estado del arte **9**

2.1. Aprendizaje Automático 9

2.2. Redes Neuronales CNN 10

2.3. Vision Transformer 13

2.4. Consideraciones Finales 15

3. Marco Teórico **16**

3.1. Redes Convolucionales 16

3.2. Funciones Softmax y Maxpooling 17

3.3. Función de activación 18

3.3.1. ReLU 18

3.3.2. GeLU 19

3.4. Validación Cruzada 19

3.5. Vision Transformer 20

3.6. Swin Transformer 22
3.7. CLAHE 24

4. Propuesta **26**
4.1. Filtros a las imágenes 27
4.2. Arquitectura de modelo ViT 28
4.2.1. Normalización por Lote 29
4.2.2. Regularización L2 30
4.3. Métricas de Evaluación 31
4.3.1. Matriz de Confusión 31
4.3.2. Accuracy 31
4.3.3. Precision 31
4.3.4. Recall (Recuperación) 32
4.3.5. F1-score 32
4.4. Consideraciones Finales 32
4.5. Futuras Direcciones 33

5. Pruebas y Resultados **34**
5.1. Aplicación del CLAHE a las Retinografías 34
5.2. Descripción de los Experimentos 35
5.2.1. Experimentos Realizados 35
5.2.2. Experimento N° 1 35
5.2.3. Experimento N°6 36
5.2.4. Comparación de arquitecturas 38

6. Conclusiones Preliminares **40**
6.1. Problemas encontrados 41

Bibliografía **44**

Índice de tablas

1.1. Cronograma de Actividades 1 7

1.2. Cronograma de Actividades 2 8

2.1. Ref: referencia, AUC: Área Bajo la Curva ROC, ML: *Machine Learning*, DLS: Aprendizaje Profundo Supervisado, CDR: radio de la copa a disco, GMP: Patrón de momento generalizado, DBL: Detección de lesiones brillantes, SVM: *Support Vector Machine* 13

4.1. Conjuntos de datos que componen la información almacenada en Kaggle. . 27

5.1. Experimentos realizados, Dim Proy: Dimensión de Proyección, Head Att.: Cabezales de atención, T. Units: Unidades Transformers, T. Layers: Capas Transformer, MLP head: Unidades MLP. 35

5.2. Tabla comparativa de Arquitecturas. Exp: Experimento, Dim Proy: Dimensión de Proyección, Head Att.: Cabezales de atención, T. Units: Unidades Transformers, T. Layers: Capas Transformer. 38

Índice de figuras

1.1. Glaucoma [Miranza, 2021] 4
1.2. Retinografías de pacientes que sufren de glaucoma de *Kaggle* [Kiefer, 2023]. 4
1.3. Retinografías de pacientes que no padecen de glaucoma de *Kaggle* [Kiefer, 2023]. 4

3.1. Arquitectura y componentes CNN [Jesús, 2020] 17
3.2. Función de activación Softmax [BotPenguin, 2023] 17
3.3. Procesamiento Max-Pooling de 4x4 a 2x2 [Ahamed et al., 2020] 18
3.4. Codificador ViT propuesto por *Dosovitskiy et al.*[Dosovitskiy et al., 2020] . 20
3.5. Arquitectura ViT propuesta por *Dosovitskiy et al.* [Dosovitskiy et al., 2020] 21
3.6. Arquitectura Swin Transformer [Sanchez-Bocanegra et al., 2023] 22

4.1. Propuesta. 26
4.2. Aplicación de CLAHE a las retinografías. 28
4.3. Estructura del modelo ViT. 29

5.1. Ejemplo, Aplicación del CLAHE a una retinografía. 34
5.2. Entrenamiento del modelo base. 36
5.3. Matriz de Confusión de conjunto de prueba. 36
5.4. Matriz de Confusión de conjunto de prueba porcentaje. 37
5.5. Entrenamiento del Experimento N°6, modelo ViT. 37
5.6. Entrenamiento del Experimento N°6, modelo ViT. 38
5.7. Matriz de Confusión de conjunto de prueba del Experimento N°6. 39

5.8. Matriz de Confusión de conjunto de prueba en porcentajes del Experimento N°6. 39

Capítulo 1

Introducción

En la época actual, uno de los recursos más recientes y prometedores para realizar predicciones y clasificación son las redes transformer. Esta innovadora tecnología aprovecha técnicas avanzadas de análisis estadístico y matemático para identificar patrones y relaciones en conjuntos masivos de datos, lo que permite anticipar eventos futuros y clasificar datos con una gran precisión.

Las redes transformer surgieron originalmente para el procesamiento de lenguaje natural, donde *Alexey Dosovitskiy et al.* [Dosovitskiy et al., 2020] demostraron un rendimiento excepcional en tareas como traducción y la generación de textos. Sin embargo, su capacidad para capturar relaciones a largo plazo en secuencias de datos las hace aplicables a una amplia gama de problemas, incluido el procesamiento de imágenes.

Para esta tesis, nos enfocaremos en una variante particular de las redes transformer adaptada específicamente para el procesamiento de imágenes: ViT *(Vision Transformer)*. Este modelo, introducido en un artículo por *Alexey Dosovitskiy et al.* [Dosovitskiy et al., 2020], representa una desviación significativa de las CNN tradicionales para tareas de clasificación de imágenes.

La arquitectura de ViT se enfoca en transformar una imagen en una secuencia de tokens, que luego se procesan mediante capas transformer para capturar tanto la información local como global. Esto contrasta con el enfoque convolucional, que se centra en la extracción de características locales mediante operaciones de convolución y agrupación.

La adopción de ViT en el campo del reconocimiento de imágenes ha demostrado resultados prometedores en una variedad de conjuntos de datos, desafiando la supremacía de las redes convolucionales en muchas tareas de visión por computadora. Su capacidad para capturar relaciones a largo plazo en imágenes y su flexibilidad para adaptarse a diferentes dominios y escalas hacen de ViT un modelo digno de exploración y análisis en el contexto de este trabajo.

1.1. Motivación y Contexto

Esta tesis se enmarca en el campo de la Visión por Computador, un área de estudio en constante evolución que busca desarrollar sistemas capaces de comprender, interpretar y procesar información visual de manera automatizada. Dentro de este contexto, la clasificación de imágenes desempeña un papel crucial al permitir la identificación y categorización de objetos, patrones y características visuales en imágenes digitales. Este proceso es fundamental en una amplia gama de aplicaciones, desde el reconocimiento facial hasta el diagnóstico médico por imágenes.

Sin embargo, a pesar de los avances significativos logrados con las redes neuronales convolucionales (CNN), aún existen desafíos importantes en el campo de la clasificación de imágenes, especialmente en términos de eficiencia, escalabilidad y capacidad de generalización a diferentes conjuntos de datos. En este sentido, la arquitectura *Vision Transformer* (ViT) emerge como una alternativa prometedora que ha demostrado su eficacia en tareas de clasificación de imágenes al utilizar mecanismos de auto-atención para capturar relaciones de largo alcance entre los diferentes elementos de la imagen. El potencial de ViT para superar las limitaciones de las CNN y su capacidad para adaptarse a imágenes de tamaño variable hacen de esta arquitectura un candidato atractivo para la mejora de los sistemas de clasificación de imágenes en diversos dominios de aplicación.

Además, es importante considerar que las Vision Transformers tienen un costo computacional significativo debido a su estructura basada en atención, lo cual puede limitar su aplicación en entornos con recursos limitados o aplicaciones que requieren tiempos de respuesta rápidos, mencionado por *Dosovitskiy et al.* [Dosovitskiy et al., 2020]. En esta tesis, se explorarán técnicas y mejoras de rendimiento específicas para mitigar estos desafíos computacionales sin comprometer la precisión en la detección de enfermedades, como el glaucoma.

En relación con el diagnóstico del glaucoma, una enfermedad ocular prevalente que afecta a millones de personas en todo el mundo, la detección temprana es crucial debido a su naturaleza asintomática en las etapas iniciales. Aprovechando la capacidad de ViT para capturar patrones globales en imágenes y su escalabilidad para adaptarse a diferentes tamaños de entrada, esta investigación busca desarrollar un sistema de clasificación de imágenes eficiente y preciso que pueda contribuir significativamente a la detección temprana y al tratamiento efectivo de esta enfermedad visual.

1.2. Planteamiento del Problema

El glaucoma es una enfermedad ocular crónica que afecta el nervio óptico del ojo y puede llevar a la pérdida irreversible de la visión. Generalmente, se asocia con un aumento de la presión intraocular, que daña gradualmente las fibras del nervio óptico como se puede ver en la Figura 1.1. El tratamiento del glaucoma generalmente implica medicamentos para reducir la presión intraocular, cirugía láser o cirugía convencional, y su objetivo principal es controlar la progresión de la enfermedad y preservar la visión. Por lo tanto, el glaucoma requiere una atención médica constante y un seguimiento adecuado para evitar

complicaciones visuales graves.

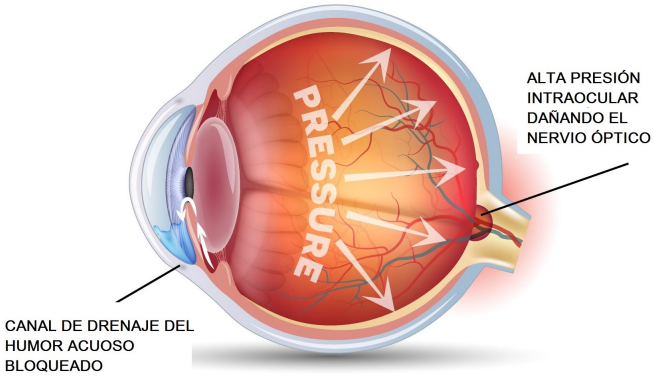

Figura 1.1: Glaucoma [Miranza, 2021]

Evelyn C O'Neill et al. en [O'Neill et al., 2014] indicaron en el 2014 que "Si se muestra una imagen del nervio óptico a diferentes expertos, es probable que no estén de acuerdo en si es glaucomatoso o no, especialmente en las primeras etapas de la enfermedad", continúa. "Los calificadores tienden a sobrestimar o subestimar el daño glaucomatoso y tienen baja reproducibilidad de calificación y poca concordancia. Por lo tanto, entrenar modelos de IA para predecir clasificaciones subjetivas es problemático", mencionado en la revista Oftalmologo al día [Review, 2023].

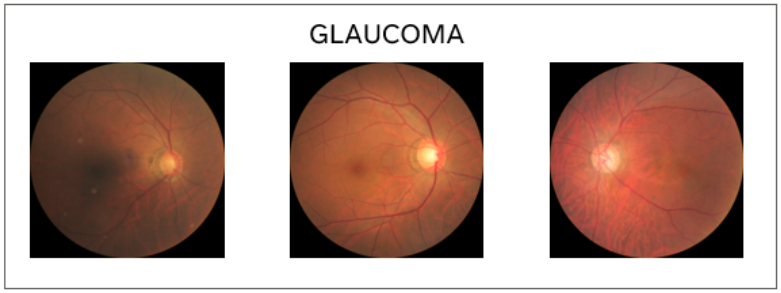

Figura 1.2: Retinografías de pacientes que sufren de glaucoma de *Kaggle* [Kiefer, 2023].

Figura 1.3: Retinografías de pacientes que no padecen de glaucoma de *Kaggle* [Kiefer, 2023].

Al analizar las retinografías, mostrando un ejemplo de la base de datos de Kaggle en la Figuras 1.2 y 1.3, se observa que la falta de brillo en algunas zonas, como por ejemplo

las zonas más lejanas a la copa óptica, ocasiona que sólo los nervios que se pueden apreciar son los más resaltantes o gruesos, sin embargo, los nervios más delgados a simple vista no son percibidos; esto podría generar errores o falsos positivos al entrenar un modelo, dando resultados no esperados.

1.3. Objetivos

El objetivo general de esta tesis es el cribado del glaucoma aplicando *Vision Transformer* con un costo computacional bajo.

1.3.1. Objetivos Específicos

- Se aplicarán filtros para modificar la base de datos para mejorar la imagen.
- Implementar el *Vision Transformer* ViT B-16 para la clasificación de retinografías.
- Crear una arquitectura *Vision Transformer* ajustando parámetros para reducir el costo computacional sin disminuir los resultados.
- Se aplicarán las métricas *accuracy*, *precision*, matriz de confusión y *F1 score*, para evaluar los modelos.

1.4. Organización de la tesis

El restante de este documento de tesis está organizado de la siguiente forma:

- El capítulo 2 describe el Estado del Arte de la tarea del cribado del glaucoma, donde se detallan enfoques utilizados y las técnicas basadas en *Deep Learning*.
- El capítulo 3 describe el Marco Teórico, donde primero se trata el concepto de las CNN. Después se presentas complementos de las arquitecturas Deep Learning, como funciones de activación y procesos de submuestreo de imágenes. Además la introducción a los modelos ViT y sus variantes. Finalmente, se muestra el proceso de aplicación de un filtro que se utilizará más adelante.
- El Capítulo 4 presenta la Propuesta de solución al problema presentado. Se exponen las modificaciones al data set, las características del modelo base a aplicar y los componentes a variar durante los experimentos.
- El capítulo 5 se abordarán los Experimentos y Resultados realizados. En este capítulo, se describen detalladamente los experimentos llevados a cabo. Además, se presentan los resultados obtenidos en cada uno de los experimentos realizados. Este capítulo proporciona una visión completa de las pruebas realizadas y los logros alcanzados.

- El capítulo 6 presentará la conclusiones preliminares de la propuesta presentada. Donde se analiza si el modelo creado es mejor a modelo base.

1.5. Cronograma

Semestre 2024-1		
Actividad	Detalles	Semana
Definición de Tema y Asesor	A partir del tema estudiado en el curso previo, se consulta para continuar con el mismo asesor, firmar el compromiso de asesoría.	23/03/2024
Definición de Conceptos Teóricos	Extraer conceptos teóricos de los papers a usar en la propuesta	30/03/2024
Redacción del Cronograma	Planificación de avances en el curso.	06/04/2024
Capitulo de Introducción	Objetivo General, Objetivos Específicos, Cronograma, Motivación y Contexto, Planteamiento del Problema.	13/04/2024
Redacción del Estado del arte y Marco Teórico	Recopilación de papers, clasificarlos y resumen de cada propuesta encontrada.	20/04/2024
Redacción Capítulo Propuesta	Descripción de la arquitectura ViT B-16 y CLAHE	27/04/2024
Revisión de Herramientas para la Implementación	Revisión Keras y Conjuntos de datos	04/05/2024
Exposición		
Implementación de la propuesta	Configuración de CLAHE a la base de datos	25/05/2024
Implementación de la propuesta	Implementación del ViT B-16 y ejecución	01/06/2024
Creación de modelo ViT desde cero	Implementación de un modelo ViT desde cero, hacer pruebas y redactar resultados	08/06/2024
Pruebas con parámetros distintos y Redacción de resultados	COntinuar con las pruebas y redacción de resultados	15/06/2024
Redacción de resultados	Aplicar las métricas para evaluar los resultados del modelo ViT B-16 y el mejor modelo encontrado	22/06/2024
Revisión del Capitulo Introducción, Estado del Arte y Marco Teórico	Revisión de los capitulos	25/06/2024
Revisión del Capitulo de Propuesta y Resultados	Revisar los Capitulos Propuesta y Resultados	06/07/2024
Exposición		

Tabla 1.1: Cronograma de Actividades 1

Cronograma 2 - Otro Campo de Datos		
Semestre 2024-2		
Actividad	Detalles	Semana
Revisión de Introducción		
Revisión Marco Teórico y estado del arte		
Completar el capítulo de Propuesta		
Documento de plan de tesis		
Exposición		
Implementación de la segunda parte de la propuesta		
Implementación de la segunda parte de la propuesta		
Implementación de la segunda parte de la propuesta		
Implementación de la segunda parte de la propuesta y Redacción de resultados		
Redacción de resultados		
Documento de plan de tesis		
Exposición		

Tabla 1.2: Cronograma de Actividades 2

Capítulo 2

Estado del arte

En esta sección se realizará una recopilación de diversos estudios que aportan al campo de las técnicas para la detección temprana del glaucoma. Se incluirá una tabla de síntesis para organizar y evaluar estos trabajos con el objetivo de identificar las contribuciones más destacadas.

2.1. Aprendizaje Automático

La introducción de la IA y el alto rendimiento computacional ha tenido un impacto significativo en el progreso de diversas disciplinas, incluyendo el diagnóstico automático de signos de glaucoma. Es por ello que se pensó que podría darse un enfoque a las imágenes con el glaucoma, es decir, enfocarse más en la retina y lo que pasaba con esta. Es así que, según *Deepak et al.* [Deepak et al., 2012], el glaucoma se caracteriza por la presencia de atrofia en la retina. En imágenes de la retina, estos cambios se presentan en forma de variaciones sutiles en las intensidades locales. Estas variaciones suelen describirse mediante estadísticas basadas en la forma local, que son propensas a errores. Es por ello que se propone un enfoque automatizado basado en características globales para detectar el glaucoma en imágenes. Se ha diseñado una representación de la imagen para resaltar los indicadores sutiles de la enfermedad, de manera que las características globales de la imagen puedan discriminar de manera efectiva entre casos normales y casos de glaucoma.

El método propuesto por *Deepak et al.* [Deepak et al., 2012] ha demostrado en un gran conjunto de imágenes anotadas por tres expertos médicos. Los resultados muestran que el método es eficaz en la detección de indicadores sutiles de glaucoma. El rendimiento de clasificación en 1186 retinografías a color, que contiene una mezcla de casos normales, casos sospechosos y casos confirmados de glaucoma, es del 97 % de sensibilidad con un 87 % de especificidad. Esto mejora aún más cuando se eliminan los casos sospechosos de los casos anormales.

En el 2019, *Mennato-Allah Talaa et al.* [Talaat et al., 2019] señala que la Diagnosis Asistida por Computadora (CAD) puede representar un avance significativo en el cribado del glaucoma. Además, CAD tiene la ventaja de ser no invasivo, simple y rentable. En

este trabajo, se presenta un algoritmo genérico automatizado de detección de glaucoma en el que se calculan características estadísticas y texturales a partir de la región del disco óptico en imágenes de la retina. Se realizaron varios análisis para comparar el rendimiento de la clasificación del glaucoma considerando diferentes técnicas de mejora de contraste (ecualización de histograma - ecualización de histograma adaptativa limitada de contraste) y modelos de color (RGB - HSV - CIELAB). A continuación, se aplica la extracción de características para identificar el conjunto óptimo de características en cada uno de los diferentes experimentos. El mejor rendimiento se logró cuando se calcularon características texturales a partir de los canales CIELAB con ecualización de histograma, lo que resultó en: precission al 92.5 %, *specify* al 90.0 % y sensibilidad al 95.0 % considerando datos públicos, en este caso se usó la base de datos REFUGE.

Cuando se aplican adecuadamente, los modelos basados en algoritmos de aprendizaje automático pueden ofrecer diagnósticos más precisos en comparación con los métodos previos. En el caso de *Tékouabou et al.* [Tékouabou et al., 2020] investigan una nueva estrategia basada en el algoritmo SVM para la clasificación automática de ojos con signos favorables a la patología del glaucoma. Durante la construcción de este modelo, se llevaron a cabo pruebas exhaustivas de varios parámetros de los algoritmos SVM, lo que resultó en un mejor rendimiento en las pruebas después de validar el modelo durante el entrenamiento. Este modelo, evaluado en términos de *accuracy*, AUC y *precision*, se ha demostrado tanto más confiable como eficiente en comparación con los enfoques anteriores para este problema, y también requiere menos tiempo de procesamiento.

En 2022, *Huarote Zegarra Raúl et al.* [Huarote Zegarra Raúl, 2022] llevaron a cabo una investigación que abordaba la necesidad de clasificar retinografías según la retinopatía diabética, el glaucoma o la salud ocular. Para lograr esto, se desarrolló una estrategia específica que incluía preparar las imágenes digitales mediante varios pasos: convertirlas a escala de grises, realizar una ecualización, aplicar el algoritmo de resalte de borde Canny y efectuar operaciones morfológicas. Estas imágenes se ingresaron en una red neuronal SOM para su clasificación. Las imágenes se etiquetaron como 0 para retinopatía diabética, 1 para glaucoma y 3 para ojos sanos. Para validar esta estrategia, se utilizó una base de datos pública de imágenes de fondo de ojo, con 45 imágenes para el entrenamiento y 15 imágenes adicionales para pruebas, todas escaladas a 256 x 256 píxeles en escala de grises. Esta metodología demostró una efectividad del 93.7 % en la identificación precisa de la enfermedad ocular.

2.2. Redes Neuronales CNN

A través del tiempo surgieron las CNN con un buen rendimiento en el área de clasificación de imágenes y abordaron la detección del glaucoma en la impresión del iris.Para abordar esto, los investigadores *Ting et al.* [Ting et al., 2017] desarrollaron y entrenaron un sistema de *deep learning* para la evaluación del glaucoma, utilizando datos del Programa Integrado de Retinopatía Diabética de Singapur (SiDRP). Utilizaron un conjunto de datos que incluía 125,189 imágenes de fotografía retinal diabética (DRP), validando el rendimiento del algoritmo en 71,896 imágenes. Los resultados revelaron una prevalencia del 0.1 %, con un área bajo la curva para la detección de posible glaucoma de 0.942, una

sensibilidad del 0.964 y una especificidad del 0.872.

En el 2018, *Liu et al.* [Liu et al., 2018] vió que mediante la detección de la relación copa-disco (la proporción del área de la copa óptica al área del disco óptico, CDR) de los pacientes, se puede realizar un cribado del glaucoma. Utilizando imágenes de retina de alta resolución en la base de datos HRF, se propone un conjunto completo de métodos de cribado. En primer lugar, el realiza la extracción de canales y la mejora de la imagen, luego utiliza un algoritmo de segmentación por umbral para separar el disco óptico y utiliza el método de crecimiento de regiones para separar la copa óptica. De esta manera, se puede calcular automáticamente el CDR. Se utiliza 30 imágenes para realizar pruebas, 15 de ellas son imágenes de ojos normales y las otras 15 son de ojos con glaucoma. Los resultados mostraron que los CDR de los pacientes con glaucoma son mayores de 0,6 en todas las 15 imágenes de glaucoma, y los CDR de los ojos normales están entre 0,2 y 0,6. Esto concuerda con el juicio de los expertos.

En 2019, *Andres Diaz-Pinto, et al.* [Diaz-Pinto et al., 2019] investigaron el desempeño de cinco arquitecturas de redes neuronales convolucionales (CNN) entrenadas en ImageNet (VGG16, VGG19, InceptionV3, ResNet50 y Xception) como clasificadores de glaucoma. La arquitectura Xception destacó por su rendimiento superior, evaluado por el equilibrio entre el área bajo la curva (AUC) y el número de parámetros de la CNN. Utilizando exclusivamente bases de datos públicas, el estudio utilizó 1707 imágenes con técnicas de aumento de datos, logrando un AUC promedio de 0.9605 con un intervalo de confianza del 95 % de 95.92-97.07 %, una especificidad promedio de 0.8580 y una sensibilidad promedio de 0.9346 después de ajustar la arquitectura Xception, superando significativamente investigaciones previas.

Además, se realizó un análisis adicional del rendimiento del modelo ajustado al probarlo en imágenes de una base de datos completamente diferente. Este enfoque difiere del método convencional donde un subconjunto de datos se usa para entrenamiento y otro para pruebas. Al usar exclusivamente la base de datos ACRIMA como conjunto de prueba, se obtuvo un AUC de 0.7678 con un intervalo de confianza del 95 % de 68.41-81.81 %. Se replicó este experimento en otras cuatro bases de datos públicas (HRF, Drishti-GS1, RIM-ONE, sjchoi86-HRF), obteniendo AUC de 0.8354, 0.8041, 0.8575 y 0.7739, respectivamente.

En el 2020, *Batista et al.* [Batista et al., 2020] donde se usó la base de datos RIM-ONE DL (RIM-ONE para Aprendizaje Profundo). En este artículo se describe este conjunto de imágenes, que consta de 313 retinografías de sujetos normales y 172 retinografías de pacientes con glaucoma. Todas estas imágenes han sido evaluadas por dos expertos e incluyen una segmentación manual del disco y la copa. También se describe un conjunto de pruebas de evaluación con diferentes modelos de redes neuronales convolucionales ampliamente conocidos. Donde se concluyó que el modelo de red VGG19 no solo proporcionó el AUC más alto, sino que su sensibilidad también igualó a 1, el valor más alto posible. El otro modelo de red con características similares, VGG16, también produjo buenos resultados. Aunque no es posible realizar una comparación directa con los resultados del desafío REFUGE del 2018, es interesante notar que el equipo ganador logró un AUC de 0.9885 con una sensibilidad del 0.9752 para una muestra de prueba que consistía en 360 retinografías de pacientes sanos y 40 que no lo estaban. El marco de evaluación propuesto en este trabajo contiene cuatro elementos principales: La descripción de los conjuntos de

entrenamiento y prueba, los tipos de redes neuronales empleadas, el enfoque utilizado para entrenar y evaluar los modelos, así como las medidas métricas consideradas durante la evaluación.

En lo que respecta al *train* y *test* se consideran dos variantes. En la primera variante, el conjunto de imágenes se dividió al azar en imágenes de entrenamiento y prueba utilizando una proporción de 70:30, respectivamente. En la segunda variante, las imágenes tomadas en el HUC se utilizaron para el entrenamiento (195 normales y 116 con glaucoma), y las imágenes tomadas en los otros dos hospitales se utilizaron para la prueba (118 normales y 56 con glaucoma). El único procesamiento realizado en las imágenes consistió en reescalarlas en intensidad en el rango de 0 a 1 y cambiar su tamaño a 224x224x3.

En 2022, *Mehedi Hasan Raju, et al.* [Raju et al., 2022] proponen que la autenticación biométrica basada en el iris es una modalidad biométrica ampliamente utilizada debido a su precisión, entre otros beneficios. Mejorar la resistencia de la biometría del iris a los ataques de suplantación es un tema de investigación importante. El seguimiento ocular y los dispositivos de reconocimiento del iris tienen hardware similar que consta de una fuente de luz infrarroja y un sensor de imagen. Esta similitud potencialmente permite que los algoritmos de seguimiento ocular se ejecuten en sistemas de biometría impulsados por el iris. El trabajo actual avanza en el estado del arte de la detección de ataques de impresión del iris, en los que un impostor presenta una impresión del iris auténtico de un usuario a un sistema de biometría. La detección de estos ataques se logra mediante el análisis de la señal de movimiento ocular capturada con un modelo de aprendizaje profundo. *Mehedi Hasan Raju, et al.* [Raju et al., 2022] explica que se implementó una versión personalizada de la arquitectura ResNet 18 y la llamaron C-ResNet 18. En C-ResNet 18, se realizó varios cambios en comparación con ResNet 18 básico. Las razones principales detrás de estos cambios son adaptarla a la declaración del problema, conjunto de datos y lograr un mejor rendimiento en términos de métricas de evaluación. Cambiaron convoluciones 2D a 1D, se modificó la forma en que se usaban las conexiones de salto y se cambió el número de canales de entrada y salida en cada bloque de convolución. En resumen, C-ResNet 18 consta de 17 bloques de convolución, seguidos de un promedio global de pooling, una capa de aplanamiento y una capa lineal (totalmente conectada). Se utilizó normalización por lotes (BN) y la función de activación utilizada fue ReLU después de cada operación de convolución.

Abordando el problema de la aplicación de distintas técnicas apartir de imágenes del ojo para la detección del glaucoma, se recopilaron distintos artículos donde se analizó los trabajos y estudios realizados.

Se pensó en realizar unas modificaciones en las imágenes, tales como la segmentación. Un proceso así se realizó por *Lianyi Wu, et al.* [Wu et al., 2019] donde se descubrió que la innovación clave fue la combinación SegNet y ADDA juntos. Se aplicó GAN para entrenar un codificador SegNet objetivo y combinandoló con un decodificador SegNet preentrenado. Este enfoque redujo el MSE (Error Cuadrático Medio) entre el CDR calculado con la segmentación y el CDR calculado con las anotaciones en el conjunto de datos objetivo. La contribución que trajo también incluye experimentos sobre funciones de pérdida y experimentos sobre el tamaño del discriminador. Al mismo tiempo, se señaló las debilidades de ADDA en el artículo. Como trabajo a futuro de este artículo fue que se probarían más enfoques de aprendizaje por transferencia y se utilizaría alguna técnica

para ampliar el conjunto de datos de entrenamiento.

Después de analizar los distintos trabajos presentados en está sección, se realizó una tabla para poder resaltar los datos más relevantes de estos trabajos como: la técnica o arquitectura que utilizaron, en lo que más se enfocaron para utilizar en sus pruebas, la base de datos y la confiablididad obtenida. Todos estos aspectos se detallan en la Tabla 2.1. Se destacó la relevancia de la base de datos ACRIMA, que incluye 396 imágenes de glaucoma y 309 imágenes normales, siendo idónea para ser utilizada como banco de pruebas en comparaciones y análisis adicionales. Los autores alientan a la comunidad científica a validar sus modelos utilizando esta nueva base de datos disponible públicamente y a comparar los resultados obtenidos con el método propuesto en este artículo. Además, recientemente ha surgido la base de datos de Kaggle, que consiste en una compilación de bases de datos públicas, la cual se discutirá más adelante.

Ref	IA	Enfoque	Técnica	Base de datos	AUC
[Tékouabou et al., 2020]	ML	-	SVM	Glaucoma Center of Semmelweis University Budapest	89.47 %,
[Raju et al., 2022]	CNN	Impresión de iris	C-ResNet 18	ETPAD v2	87.78 %
[Ting et al., 2017]	DLS	Retinografías	-	DRP	94.2 %
[Huarote Zegarra Raúl, 2022]	UML	-	SOM	Fundus-Images	93.7 %
[Talaat et al., 2019]	ML	Retinografías	SVM, RGB - HSV - CIELAB	REFUGE	92.5 %
[Batista et al., 2020]	CNN	Retinografías	VGG19	RIM-ONE	92.72 %
[Liu et al., 2018]	-	CDR	Segmentation	HRF	98.74 %
[Wu et al., 2019]	CNN	CDR	SegNet - ADDA	-	-
[Deepak et al., 2012]	ML	DBL	GMP	-	97 %
[Diaz-Pinto et al., 2019]	CNN	Retinografías	Xception	RIM-ONE, ACRIMA, HRF, Drishti-GS1, sjchoi86-HRF	96.05 %

Tabla 2.1: Ref: referencia, AUC: Área Bajo la Curva ROC, ML: *Machine Learning*, DLS: Aprendizaje Profundo Supervisado, CDR: radio de la copa a disco, GMP: Patrón de momento generalizado, DBL: Detección de lesiones brillantes, SVM: *Support Vector Machine*

2.3. Vision Transformer

Aunque los transformadores han sido ampliamente adoptados como estándar para el procesamiento del lenguaje natural, su aplicación en visión por computadora sigue siendo limitada. Tradicionalmente, se han combinado con redes neuronales convolucionales (CNN) o se han utilizado para reemplazar ciertos elementos de estas redes manteniendo su estructura general. Sin embargo, Dosovitskiy et al. [Dosovitskiy et al., 2020] demostraron que esta dependencia de las CNN no es necesaria, mostrando que un transformador puro aplicado directamente a secuencias de parches de imágenes puede tener un rendimiento excepcional en tareas de clasificación de imágenes. Al ser preentrenado con grandes volúmenes de datos y transferido a múltiples conjuntos de datos de reconocimiento de imágenes de tamaño mediano o pequeño (como ImageNet, CIFAR-100, VTAB, entre otros), ViT logra resultados sobresalientes en comparación con las redes convolucionales más avanzadas, a la vez que requiere considerablemente menos recursos computacionales para su entrenamiento.

En el contexto del concurso MIA-COV19 realizado en el 2021, Gao et al. [Gao et al., 2021] realizaron una comparativa entre las arquitecturas de transformadores de visión (ViT) y las redes neuronales convolucionales (CNN) en el análisis de imágenes médicas. Se destacó la eficiencia y escalabilidad de las arquitecturas basadas en transformadores, especialmente valiosas en entornos médicos donde los conjuntos de datos suelen ser más limitados en comparación con bases de datos de referencia como ImageNet. Los resultados revelan que

el modelo ViT supera al DenseNet en la detección de COVID-CT, logrando un *accuracy* del 76.6 % frente al 73.7 % respectivamente. Además, se aborda el desafío de procesar imágenes de tomografía computarizada (TC) de tórax en formato 3D, señalando posibles limitaciones en el rendimiento de los modelos 3D en comparación con los modelos 2D debido a la proporción reducida de regiones lesionadas.

En 2022, Aníbal Bregón Bregón el al. [Aníbal Bregón Bregón, 2022] exploró tres áreas fundamentales en su Trabajo Fin de Máster: estimación de profundidad a partir de imágenes monoculares, arquitecturas basadas en transformers y técnicas para mejorar el rendimiento de los modelos. Realizó modificaciones en una arquitectura avanzada para la estimación de profundidad utilizando imágenes monoculares, entrenando los modelos resultantes en el conjunto de datos KITTI, orientado a la conducción autónoma. Se evaluó el impacto de estas modificaciones en diversas métricas como la calidad de los resultados, la velocidad de inferencia y el tamaño del modelo. Se eligió una configuración que equilibraba el aumento en la velocidad de inferencia, la reducción del tamaño del modelo y la posible pérdida de calidad de los resultados. Se proporcionó el código necesario para entrenar y utilizar estos modelos, con el objetivo de promover la expansión de este trabajo y sus hallazgos.

En 2023, Sanchez-Bocanegra y colaboradores [Sanchez-Bocanegra et al., 2023] llevaron a cabo un estudio sobre la detección de cáncer de piel, comparando los modelos ViT B-16, Swin Transformer y varias CNN. Para los modelos Transformer, fue necesario redimensionar las imágenes a 224x224 píxeles, lo cual implicó procesar el archivo de metadatos completo y dividir las imágenes en grupos de 100 unidades para luego concatenarlas en un nuevo dataframe. Una vez obtenida una tabla con las características de las 10,015 imágenes en el formato 224x224x3, se seleccionaron 750 imágenes por categoría para el entrenamiento del modelo.

El modelo ViT B-16 fue entrenado con lotes de tamaño 8, 12 y 16, utilizando tasas de aprendizaje de 0.001, 0.0005 y 0.0001. El objetivo principal fue evitar exceder la capacidad de memoria del equipo durante el procesamiento de las imágenes y el entrenamiento del modelo. El modelo obtuvo un accuracy: 90.19 %, precision: 89.83 %, recall: 89.76 % y f1 score: 89.70 %. Para entrenar el modelo Swin Transformer se utiliza un tamaño de lote de 2. Las tasas de aprendizaje utilizadas son: 0.00001 y 0.000005. El modelo obtuvo un accuracy: 87.81 %, precision: 87.51 %, recall: 87.29 % y f1 score: 87.24 %.

A comienzos de 2024, García Molina [Garcia Molina, 2024] realizó un estudio de Trabajo de Fin de Grado centrado en la evaluación de una red neuronal transformer diseñada para procesar nubes de puntos voxelizadas, comparándola con la arquitectura establecida de PointNet en la tarea de clasificación de nubes de puntos 3D. Se introdujo una capa de atención multi-cabeza personalizada y se ajustó la entrada para incorporar las coordenadas tridimensionales de los puntos. Los resultados mostraron que la arquitectura PointNet alcanzó un 87 % de precisión, mientras que su transformer, denominado Optimus, logró un 85 %. Aunque las diferencias entre ambas arquitecturas no fueron significativas, la red transformer demostró un rendimiento competitivo. Se identificó un potencial considerable para mejorar la red propuesta en futuros trabajos.

2.4. Consideraciones Finales

Es importante aclarar que en la Tabla 2.1 no se incluyen los trabajos Transformers debido a que no hay un trabajo que aplique alguno de los modelos presentados al cribado del glaucoma. Sin embargo, fueron colocados trabajos que utilizan estas arquitecturas para mostrar la efectividad del *Vision Transformer* en casos de clasificación de imágenes en el área biomédico.

Capítulo 3

Marco Teórico

3.1. Redes Convolucionales

En esta tesis se aborda la aplicación de las redes convolucionales, conocidas como CNN, las cuales son una forma especializada de redes neuronales profundas diseñadas específicamente para procesar y analizar datos visuales como imágenes y videos. Estas redes son altamente efectivas en la extracción automática de características relevantes de las imágenes mediante capas de convolución y pooling, lo que les permite identificar patrones visuales complejos como bordes, texturas y formas de manera jerárquica. Las CNN se utilizan ampliamente en diversas aplicaciones de visión por computadora, incluyendo la clasificación de imágenes, detección de objetos, reconocimiento facial y segmentación de imágenes, demostrando un rendimiento excepcional en el procesamiento de imágenes y videos en la actualidad.

Para extraer estas características relevantes de los datos de entrada, como una imagen, se emplean operaciones convolucionales mediante la aplicación de filtros o núcleos de convolución sobre la imagen. Estos filtros son matrices pequeñas que se deslizan sobre la imagen de entrada, detectando patrones o características específicas como bordes, texturas o formas. El proceso de convolución implica seleccionar un filtro, colocarlo en diversas ubicaciones de la imagen, realizar multiplicaciones y sumas para obtener respuestas del filtro en cada ubicación, y crear así un ”mapa de características”que representa estas respuestas en toda la imagen.

Estos mapas de características resultantes pueden pasar por otras capas de la red neuronal, como capas de agrupación y capas completamente conectadas, para realizar tareas específicas, como clasificación de objetos en imágenes.

Las operaciones convolucionales son eficaces para detectar patrones locales en los datos de entrada, lo cual las hace especialmente útiles en aplicaciones de visión por computadora como detección de objetos, reconocimiento de imágenes y segmentación semántica. Además, estas operaciones contribuyen a reducir la cantidad de parámetros en la red en comparación con las redes neuronales completamente conectadas. Esta reducción ayuda a prevenir el sobreajuste y facilita el aprendizaje eficiente de características importantes en los datos.

Se puede ver la arquitectura básica de estas CNN en la Figura 3.1.

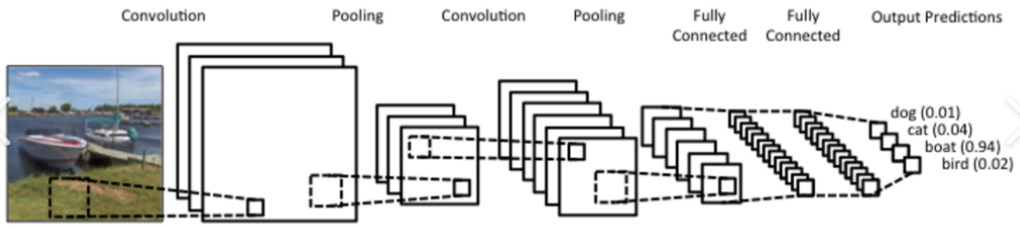

Figura 3.1: Arquitectura y componentes CNN [Jesús, 2020]

3.2. Funciones Softmax y Maxpooling

La función SoftMax se emplea como función de activación en la capa de salida de una red neuronal, especialmente en problemas de clasificación multiclase. Su propósito es transformar las salidas de la capa anterior en probabilidades normalizadas que suman uno. Estas probabilidades indican la confianza del modelo en la asignación a cada clase posible.

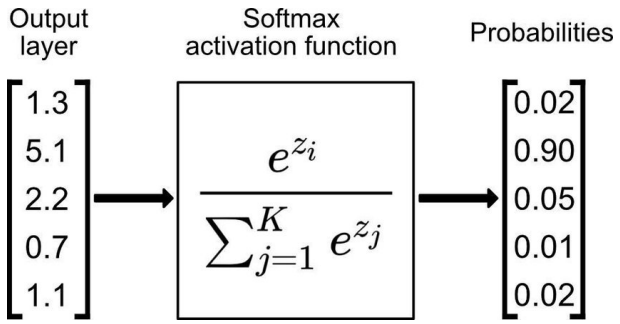

Figura 3.2: Función de activación Softmax [BotPenguin, 2023]

La función SoftMax opera sobre un vector de entrada y produce un vector de salida con la misma dimensión. Realiza dos pasos principales: primero, exponencia cada elemento del vector de entrada, y segundo, divide cada elemento exponenciado por la suma de todos los elementos exponenciados, como se ilustra en la Figura 3.2.

Esta función es crucial en la clasificación multiclase porque asigna probabilidades a cada clase. Esto permite determinar la probabilidad de que una instancia pertenezca a cada una de las clases posibles. Es ampliamente utilizada en aplicaciones como la clasificación de imágenes, texto y voz.

El Max-Pooling es un método de submuestreo que discretiza una representación de entrada (como una imagen o una matriz de salida de una capa oculta) reduciendo su tamaño. Este proceso busca reducir el costo computacional al disminuir el número de parámetros que la red neuronal debe aprender, además de proporcionar invarianza frente a pequeñas translaciones: si una translación no modifica el máximo de una región cubierta

por el filtro, el máximo de cada región seguirá siendo el mismo, asegurando que la nueva matriz resultante sea idéntica.

Para ilustrar el funcionamiento del Max-Pooling, consideremos el ejemplo de una matriz inicial de 4×4, como se muestra en la Figura 3.3. Aquí, un filtro de ventana 2×2 se desliza sobre la entrada. En cada región cubierta por el filtro, el Max-Pooling selecciona el valor máximo, creando así una nueva matriz de salida donde cada elemento representa el máximo de cada región detectada.

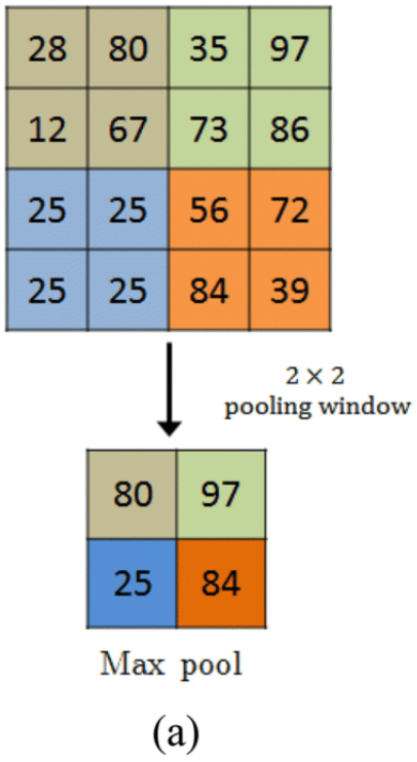

Figura 3.3: Procesamiento Max-Pooling de 4x4 a 2x2 [Ahamed et al., 2020]

3.3. Función de activación

3.3.1. ReLU

La función de activación ReLU (*Rectified Linear Unit*) es crucial en las redes neuronales convolucionales (CNN) y otros tipos de redes neuronales artificiales. Su propósito fundamental es agregar no linealidad a la red, lo cual facilita que pueda aprender y representar relaciones más complejas en los datos.

La función ReLU se define de la siguiente manera:

$$ReLU(x) = max(0, x)$$

Donde x es la entrada a la función. La función ReLU toma un valor de cero para cualquier entrada negativa y pasa directamente cualquier entrada positiva sin modificarla.

En otras palabras, si la entrada es positiva, la función ReLU la deja pasar tal como es; si es negativa, la función ReLU la transforma en cero. Esto crea una activación esparsa en la red, ya que solo las neuronas cuya entrada es positiva se activan.

3.3.2. GeLU

El gelu (*Gaussian Error Linear Unit*) es una función de activación empleada en redes neuronales, particularmente en modelos como BERT y otros basados en Transformers. Su propósito principal es agregar no linealidad a la red, lo que facilita que pueda aprender y representar relaciones más complejas en los datos.

La función gelu se define de la siguiente manera:

$$gelu(x) = 0.5x \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right)$$

Donde x es la entrada a la función. La función gelu combina una función lineal con una función no lineal (tangente hiperbólica) para producir una salida que se asemeja a la distribución gaussiana acumulativa. Esto permite una transición suave entre los valores positivos y negativos de la entrada, evitando el problema de la saturación de gradientes que se observa en otras funciones de activación como la sigmoidal.

3.4. Validación Cruzada

Finalmente, la validación cruzada de 5 pliegues, a menudo referida como "validación cruzada 5-fold", es una técnica ampliamente utilizada en el aprendizaje automático y la evaluación de modelos. Su propósito es evaluar la capacidad de generalización de un modelo y mitigar el sesgo en la evaluación del rendimiento, permitiendo una evaluación más robusta y fiable del modelo al ser probado en diferentes subconjuntos de datos. El procedimiento de la validación cruzada 5-fold incluye los siguientes pasos:

1. Partición en pliegues (*folds*): El conjunto de datos se divide en aproximadamente 5 subconjuntos o pliegues del mismo tamaño.

2. Ciclo de entrenamiento y evaluación: Se realiza un ciclo de entrenamiento y evaluación 5 veces, correspondiente al número de pliegues:

   Pliegue de prueba: Un pliegue se reserva como conjunto de prueba en cada ciclo.

   Pliegues de entrenamiento: Los otros 4 pliegues se utilizan como conjunto de entrenamiento en cada ciclo.

   Entrenamiento y evaluación: Se entrena el modelo en los pliegues de entrenamiento y se evalúa su rendimiento en el pliegue de prueba. Esto implica ajustar el modelo a los datos de entrenamiento y medir su precisión en los datos de prueba.

3. Promedio de resultados: Al completar los 5 ciclos, se obtienen 5 medidas de rendimiento (una para cada pliegue de prueba). Estas medidas suelen ser métricas como precisión, exactitud, sensibilidad, especificidad, entre otras, dependiendo de la tarea. Para evaluar el rendimiento general del modelo, se calcula el promedio de estas medidas.

3.5. Vision Transformer

Los modelos ViT se construyen sobre la arquitectura Transformer, inicialmente desarrollada para tareas de procesamiento del lenguaje natural (NLP). Esta arquitectura se compone de varias capas de auto-atención (*self-attention*), que emplean el mecanismo de atención (*attention*) para permitir al modelo enfocarse selectivamente en diferentes partes de la secuencia de entrada al hacer predicciones. Este enfoque fue introducido por *Dosovitskiy et al.* [Dosovitskiy et al., 2020]. En ViT, se utiliza el Transformer de manera estándar, destacándose por aplicar este enfoque directamente a imágenes sin depender de las convoluciones convencionales (CNN), como es detallado en el artículo mencionado.

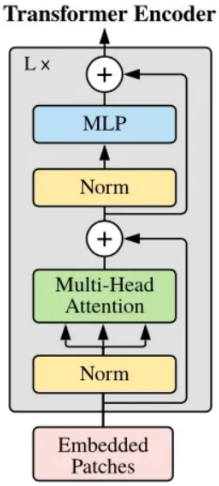

Figura 3.4: Codificador ViT propuesto por *Dosovitskiy et al.*[Dosovitskiy et al., 2020]

En ViT, el proceso comienza dividiendo la imagen en una secuencia de pequeños parches (*patches*) de tamaño fijo, explorando en el artículo tamaños como 16x16 y 14x14, como se ilustra en la Figura 3.5. Cada parche se transforma en un vector 1D aplanado. Estos parches son luego codificados (*embedding*) en vectores de longitud 796 mediante la multiplicación por una matriz de proyección lineal E. Posteriormente, los parches se concatenan con la clase de la imagen y su correspondiente codificación de posición (*position embedding*). Durante el entrenamiento, el *position embedding* se encarga de aprender la posición relativa de cada parche dentro del contexto global de la imagen.

Los parches son sometidos a varias capas de auto-atención, donde cada parche se compara con los demás para aprender a enfocarse selectivamente en diferentes regiones de la imagen, codificando así la información en una secuencia de representaciones, como se muestra en la Figura 3.4. Estas representaciones se combinan luego a través de una capa de agrupación global para obtener un vector de características de longitud fija, que

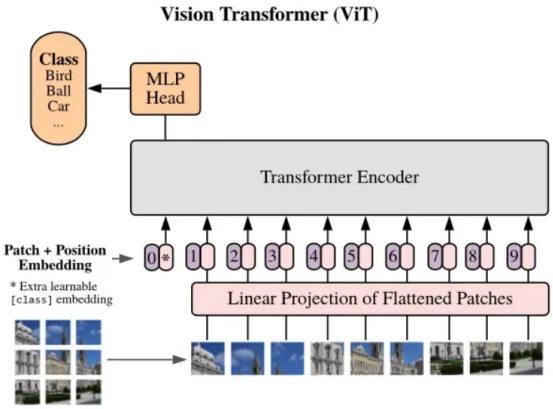

Figura 3.5: Arquitectura ViT propuesta por *Dosovitskiy et al.* [Dosovitskiy et al., 2020]

se procesa a través de un perceptrón multi-capa (MLP) para realizar la predicción final de clasificación.

La principal ventaja del modelo ViT reside en su capacidad para capturar patrones globales en una imagen desde las primeras capas de auto-atención, abarcando prácticamente toda la imagen en el caso de conjuntos de datos como ImageNet, en contraste con las CNN tradicionales que se centran en características locales. Este enfoque de atención ha demostrado ser muy efectivo para identificar patrones complejos en imágenes, especialmente en tareas como la detección y segmentación de objetos.

Otra ventaja clave del modelo ViT es su capacidad de escalar. Debido a que el mecanismo de auto-atención opera sobre una secuencia de parches de tamaño fijo, los modelos ViT pueden adaptarse fácilmente a imágenes de cualquier tamaño ajustando los *Positional Embeddings*. En el contexto del estudio mencionado, esta adaptación se logra mediante interpolación, lo que los hace más versátiles en comparación con las CNN tradicionales que requieren tamaños de entrada específicos. Además, la arquitectura ViT puede pre-entrenarse en grandes conjuntos de datos para aprender representaciones generales de imágenes, similar al enfoque utilizado en las CNN.

El proceso de clasificación de una imagen utilizando ViT B-16 se puede desglosar en los siguientes pasos:

- Preprocesamiento: La imagen de entrada se ajusta al tamaño y formato requerido por la arquitectura, lo que incluye dividirla en parches y aplicar la codificación posicional necesaria.
- Codificación de los parches: Cada parche de la imagen se convierte en un vector de características utilizando una capa de proyección específica.
- Incorporación de información de posición: Se agrega información de posición a los vectores de características para mantener la estructura espacial original de la imagen.
- Paso a través de los bloques Transformer: Los vectores de características, ahora con información posicional, se procesan a través de múltiples bloques Transformer, donde se calculan las interacciones entre los diferentes parches.

- Pooling: Se aplica una técnica de pooling, como el promedio o el máximo, para combinar la información de todos los parches en un único vector representativo.
- Clasificación: El vector resultante se introduce en una capa de clasificación que genera las probabilidades de pertenencia a cada clase.
- Selección de la clase: Se elige la clase con la probabilidad más alta como la etiqueta final de clasificación para la imagen.

3.6. Swin Transformer

La adaptación del Transformer del lenguaje al dominio de la visión presenta desafíos significativos debido a las diferencias fundamentales entre ambos, como las variaciones en la escala de las entidades visuales y la alta resolución de los píxeles en las imágenes en comparación con las palabras en el texto. En respuesta a estas diferencias, *Liu et al.* [Liu et al., 2021] proponen un enfoque innovador con el Swin Transformer, un modelo jerárquico que utiliza esquemas de ventanas desplazadas para calcular representaciones. Este método optimiza la eficiencia al restringir el cálculo de la autoatención a ventanas locales no superpuestas, mientras facilita la interconexión entre estas ventanas. La arquitectura jerárquica del Swin Transformer ofrece flexibilidad para modelar a diferentes escalas y mantiene una complejidad computacional lineal respecto al tamaño de la imagen. Estas características han demostrado ser altamente efectivas en diversas tareas de visión, como la clasificación de imágenes (con una precisión top-1 del 87.3 en ImageNet-1K) y tareas de predicción densa como la detección de objetos (con AP de caja del 58.7 y AP de máscara del 51.1 en COCO testdev) y segmentación semántica (con mIoU del 53.5 en ADE20K val). Estos resultados subrayan el potencial de los modelos basados en Transformer, como el Swin Transformer, como estructuras fundamentales para aplicaciones de visión avanzada.

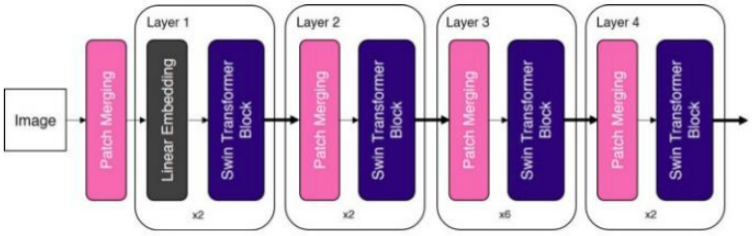

Figura 3.6: Arquitectura Swin Transformer [Sanchez-Bocanegra et al., 2023]

La arquitectura de Swin Transformer se comforma principalmente por:

- Pirámide de Parches (Patch Pyramid): En lugar de dividir la imagen en parches de tamaño fijo, Swin Transformer utiliza una pirámide de parches, donde los parches de diferentes escalas se agrupan en etapas sucesivas. Esto permite capturar tanto información local como global de la imagen.
- Bloques de Transición (Transformer Blocks): Cada etapa de la pirámide de parches incluye múltiples bloques Transformer. Estos bloques procesan la información

en cada escala de manera similar a los bloques Transformer estándar, permitiendo capturar relaciones a diferentes escalas.

- Conexiones de Atención a Escala (Hierarchical Shifted Windows): En lugar de una atención global, Swin Transformer utiliza conexiones de atención a escala, donde cada bloque atiende a un conjunto específico de parches y se desplaza gradualmente a lo largo de la pirámide de parches. Esto permite capturar relaciones a diferentes distancias y escalas.

- Reducción de Dimensionalidad (Downsampling Layers): Entre las etapas de la pirámide de parches, se incluyen capas de reducción de dimensionalidad para reducir la resolución espacial de la representación, lo que ayuda a manejar la complejidad computacional y capturar información a escalas más grandes.

- Capa de Clasificación (Classification Head): Al final de la arquitectura, se incluye una capa de clasificación que transforma la salida de la última etapa de la pirámide de parches en una distribución de probabilidades sobre las clases de salida.

El proceso de clasificación de imágenes utilizando Swin Transformer se puede descomponer en varios pasos clave:

- Preprocesamiento: La imagen de entrada se ajusta al tamaño y formato requerido por la arquitectura Swin Transformer, dividida en una pirámide de parches de diferentes escalas para capturar información detallada a diversas resoluciones.

- Codificación de Parches: Cada parche de la imagen se codifica y transforma en un vector de características mediante una capa de proyección, asegurando que cada región de la imagen contribuya adecuadamente a la representación global.

- Paso a través de los Bloques de Transición: Los vectores de características pasan a través de múltiples bloques Transformer en cada etapa de la pirámide de parches. Estos bloques capturan relaciones a diferentes escalas y distancias, permitiendo al modelo procesar características contextuales a lo largo de la jerarquía de la imagen.

- Reducción de Dimensionalidad: Entre las etapas de la pirámide de parches, se realiza una reducción de dimensionalidad para manejar la complejidad computacional y consolidar la información capturada a escalas más amplias, preservando al mismo tiempo la información relevante para la clasificación.

- Capa de Clasificación: La salida de la última etapa de la pirámide de parches se introduce en una capa de clasificación que genera las probabilidades de pertenencia a cada clase.

- Selección de Clase: La clase con la probabilidad más alta se selecciona como la etiqueta final de clasificación para la imagen, determinando así la categoría a la que pertenece la imagen en cuestión.

3.7. CLAHE

Técnica de procesamiento de imágenes para optimizar el contraste localmente al aplicar la ecualización del histograma en regiones pequeñas, evitando la amplificación del ruido al limitar el rango de ecualización en cada región. A diferencia de la ecualización de histograma tradicional aplicada por *Mennato-Allah Talaa et al.* [Talaat et al., 2019], que puede aumentar el contraste globalmente pero también amplificar el ruido, CLAHE descrita por Alwazzan et al. en 2021 para una investigación de mejora en fondos de ojo [Alwazzan et al., 2021] opera a nivel local, lo que significa que mejora el contraste en regiones específicas de la imagen sin afectar otras regiones.

Esta técnica opera de la siguiente manera:

1. División en Regiones: En lugar de aplicar la ecualización del histograma a toda la imagen de manera global, CLAHE divide la imagen en regiones más pequeñas. Esta división facilita una adaptación local del contraste en cada región específica de la imagen.
2. Ecualización Adaptativa: En cada región, se aplica la ecualización del histograma de manera independiente. Esto significa que el mapeo de intensidades se calcula en función del histograma local de cada región, lo que permite adaptarse a las características locales de contraste.
3. Contraste Limitado: Para evitar amplificar el ruido en áreas de baja varianza, la CLAHE incluye un límite para el rango de ecualización en cada región. Esto significa que la ecualización se realiza dentro de un rango limitado de intensidades, lo que ayuda a controlar el efecto de amplificación del ruido.

La ecuación de la ecualización del histograma:

$$v(i,j) = \frac{Fu(u(i,j) - Fu(a))}{1 - Fu(a)}(L-1) + 0.5$$

- $v(i,j)$: Es el valor del píxel ecualizado en la posición (i,j) de la imagen de salida. Es el valor que se asigna al píxel después de aplicar la ecualización del histograma.
- $Fu(x)$: Es una función acumulativa de distribución de probabilidad (CDF, por sus siglas en inglés) de los niveles de intensidad de la imagen de entrada. $Fu(x)$ representa la probabilidad acumulada de que un píxel tenga un valor de intensidad menor o igual a x.
- $u(i,j)$: Es el valor del píxel original en la posición (i,j) de la imagen de entrada antes de aplicar la ecualización del histograma.
- a: Es el valor de intensidad mínimo en la imagen de entrada. Representa el nivel de intensidad más oscuro en la imagen original.
- L: Es el número de niveles de intensidad posibles en la imagen de salida. Es decir, si la imagen de salida es de 8 bits, L sería 256 (desde 0 hasta 255).

- El término $(L1)$ normaliza el valor de intensidad al rango completo de la imagen de salida.
- 0.5: Es un término de redondeo. Se agrega 0.5 antes de realizar la conversión a entero para redondear correctamente el resultado al valor de intensidad más cercano.

Capítulo 4

Propuesta

En este capítulo se presenta la propuesta de solución diseñada para realizar la ecualización de retinografías del conjunto de datos de entrada y extraer todas las características necesarias para lograr una clasificación efectiva. Se emplea tanto la transformación de datos como la arquitectura transformer para abordar los objetivos específicos mencionados. El proceso general se ilustra en la Figura 4.1. A continuación se detalla cada etapa del proceso propuesto.

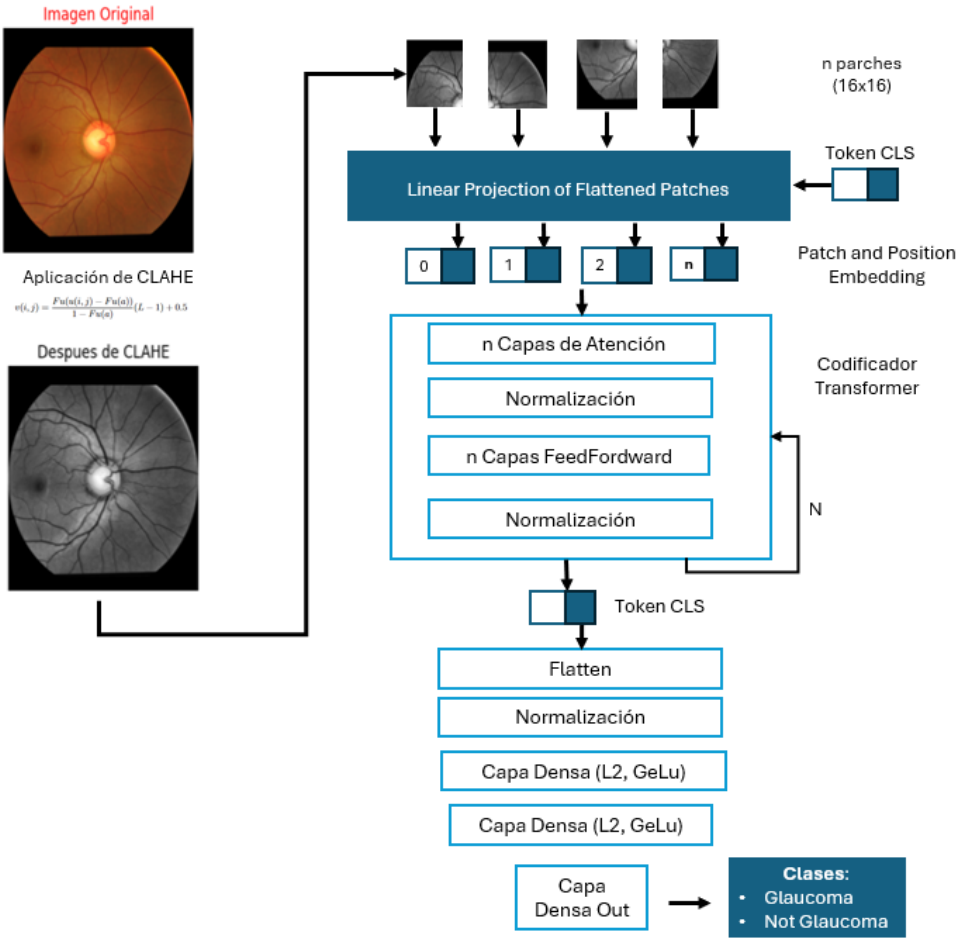

Figura 4.1: Propuesta.

Dataset	No Glaucoma	Glaucoma
BEH (Bangladesh Eye Hospital)	463	171
CRFO-v4	31	48
DR-HAGIS	0	10
DRISHTI-GS1-TRAIN	18	32
DRISHTI-GS1-TEST	13	38
EyePACS-AIROGS	0	3269
FIVES	200	200
G1020	724	296
HRF (High Resolution Fundus)	15	15
JSIEC-1000 (Joint Shantou International Eye Center)	38	0
LES-AV	11	11
OIA-ODIR-TRAIN	2932	197
OIA-ODIR-TEST-ONLINE	802	58
OIA-ODIR-TEST-OFFLINE	417	36
ORIGA-light	482	168
PAPILA	333	87
REFUGE1-TRAIN	360	40
REFUGE1-VALIDATION	360	40
sjchoi86-HRF	300	101

Tabla 4.1: Conjuntos de datos que componen la información almacenada en Kaggle.

4.1. Filtros a las imágenes

Se empleará la retinografías ofrecidas por Kaggle, la cual consiste en una recopilación de 19 *datasets* como se observa en la Tabla 4.1, gracias a Kiefer [Kiefer, 2023] que realizó esta recopilación para la minería de datos. Algunos de estas bases de datos ya han sido mencionados en el Capítulo 2 lo cual brinda mayor seguridad para utilizar estos datos para entrenar nuestro modelo. Esta base de datos consta de 8621 retinografías para el entrenamiento, 5747 para validación y 2874 para pruebas, con dimensiones de 512 x 512; estos datos están balanceados en un 60 % para casos sin glaucoma y un 40 % para casos con glaucoma.

Al conjunto de datos se le aplicará el método Ecualización de Histograma Adaptativo de Contraste Limitado (CLAHE). El CLAHE es una técnica de procesamiento de imágenes utilizada para mejorar el contraste local en una imagen, tal como se describe en el Capítulo 3 presentada para fondos de ojo por Alwazzan et al. [Alwazzan et al., 2021].

Mejorando el contraste local en las retinografías mediante el CLAHE, los detalles sutiles, como los nervios, pueden hacerse más distintos y, por ende, más fáciles de detectar para el modelo, como se ilustra en la Figura 4.2. Esto puede resultar en un entrenamiento más efectivo del modelo, ya que tendrá más información útil para aprender y generalizar patrones importantes en las imágenes de retinografía.

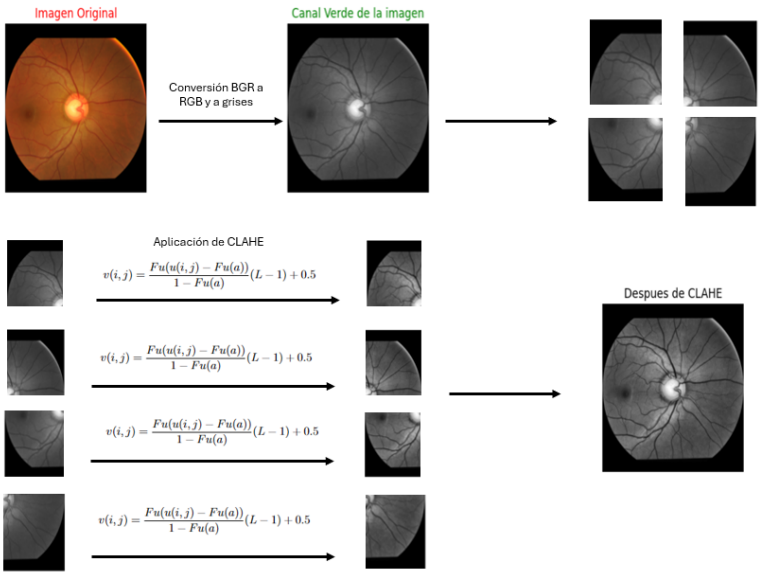

Figura 4.2: Aplicación de CLAHE a las retinografías.

4.2. Arquitectura de modelo ViT

En la actualidad, varios clasificadores de imágenes como las redes neuronales convolucionales CNN y los Transformers de Visión ViT se han utilizado para detectar enfermedades con éxito. Los ViT, a pesar de ser una tecnología relativamente nueva, han mostrado resultados prometedores en diversas tareas de visión por computadora, incluida la detección de enfermedades a partir de imágenes médicas.

En este estudio, nos enfocaremos en la detección del glaucoma utilizando como base la arquitectura Vision Transformer B-16 (ViT B-16). Esta elección se fundamenta en su capacidad para capturar relaciones de largo alcance en las imágenes y su flexibilidad para adaptarse a diferentes aplicaciones en el campo médico. Nos basaremos en los éxitos previos de esta arquitectura en la clasificación de imágenes para la detección de enfermedades, como se discutió detalladamente en el Capítulo 2 de este trabajo.

El proceso experimental que se lleva a cabo se resume en la Figura 4.3.

Comenzaremos aplicando el método de Ecualización de Histograma Adaptativo de Contraste Limitado (CLAHE) a las retinografías de la base de datos. Esta técnica mejora el contraste local en las imágenes, lo que puede hacer que los detalles sutiles, como los nervios, sean más distintos y, por ende, más fáciles de detectar para el modelo.

Luego, transformaremos las imágenes en vectores de características utilizando la Proyección Lineal de Parches Aplanados. Este proceso tiene como objetivo transformar las imágenes en parches, es decir, vectores unidimensionales de características, para prepararlos para ser procesados por el codificador Transformer; a este parche le agregaremos su posicionamiento incrustado. Se agrega un token especial llamado CLS al principio de la secuencia de vectores, como indica Devlin et al. [Devlin et al., 2018], que presenta el uso de este token para clasificación de imágenes a través de un Transformer. El estado final de este token se utiliza para la clasificación final.

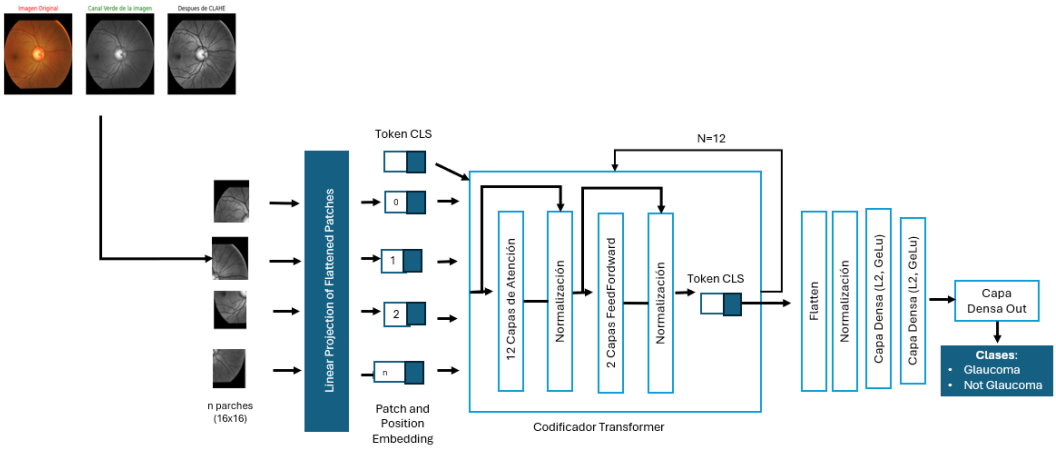

Figura 4.3: Estructura del modelo ViT.

Estos vectores de características se introducirán en el codificador del Transformer para comenzar el entrenamiento del modelo. La arquitectura de ViT B-16 que se toma como base para este modelo consta de 12 bloques de atención, donde hay 12 capas de atención y 2 capas feedforward. En estas capas se aplica la normalización después de cada capa y se aplica la técnica de una red residual, donde la entrada original se suma a la salida de la subcapa. Cada capa de atención calcula las interacciones entre los vectores de parches para capturar relaciones de largo alcance en la imagen, mientras que las capas feedforward aplican transformaciones no lineales para aprender características más abstractas de los datos de entrada. En estas capas, se aplica la función de activación GELU (*Gaussian Error Linear Unit*) a los resultados antes de pasarlos a la siguiente capa. Esto añade no linealidad al modelo, permitiendo que aprenda relaciones y patrones más complejos en los datos, lo cual es fundamental para la tarea característica de un Vision Transformer [Hendrycks and Gimpel, 2016]. Durante el procesamiento en el codificador Transformer, cada token (incluido el CLS) se actualiza con información contextual de los tokens vecinos. Esto permite que los vectores de representación capturen relaciones a largo plazo y características relevantes. A medida que el token CLS pasa por las capas del codificador, se combina con la información de todos los demás tokens. La atención y las conexiones residuales ayudan a agregar características relevantes.

El estado final del token CLS se conectará con una capa Flatten, aplicada para asegurarnos de que continúe la unidimensionalidad.

Finalmente, los resultados serán procesados por una capa de clasificación final, la cual está compuesta por una capa densa con dos neuronas, una correspondiente a cada clase de salida. Posteriormente, se aplica la función de activación Softmax para obtener las probabilidades de clasificación asociadas.

4.2.1. Normalización por Lote

Se utilizará el método de Normalización por lotes para ajustar y estandarizar las activaciones, asegurando que tengan una media cercana a cero y una desviación estándar cercana a uno. Este proceso, según Santurkar et al. [Santurkar et al., 2018], suaviza el

panorama de optimización durante el entrenamiento, lo que conduce a un comportamiento más predecible y estable de los gradientes. Esto, a su vez, facilita un entrenamiento más rápido.

La fórmula para la normalización por lotes se define como:

$$BN(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \times \gamma + \beta$$

Donde:

- x es el vector de activación de la capa.
- μ es la media de x en el lote.
- σ^2 es la varianza de x en el lote.
- ϵ es una pequeña constante para evitar la división por cero.
- γ es un parámetro de escala aprendido.
- β es un parámetro de sesgo aprendido.

Después de la etapa de Normalización, se implementarán tres capas densas completamente conectadas para llevar a cabo la clasificación final. Esto permitirá que la red neuronal aprenda representaciones más complejas y abstractas de las características presentes en la imagen. Estas capas serán activadas utilizando la función GELU, la misma empleada en la arquitectura ViT B-16. El propósito de utilizar GELU es introducir no linealidad en la red, facilitando así el aprendizaje de relaciones y patrones más sofisticados en los datos.

4.2.2. Regularización L2

Para prevenir el sobreajuste en estas capas densas, aplicaremos la técnica de regularización L2 a los pesos de estas capas, tecnica aplicada por David Ramírez [David, 2022] para eliminar el *Overfitting* en procesos de *Deep Learning*. Se agrega un término de penalización a la función de pérdida que es proporcional a la norma L2 de los pesos. En esta arquitectura, la regularización L2 se aplica a las capas densas finales del modelo para ayudar a controlar el sobreajuste y mejorar la generalización del modelo.

La fórmula para la regularización L2 se define como:

$$Loss_{L2} = \frac{\lambda}{2} \sum_{i} w_i^2$$

Donde:

- w_i son los pesos del modelo.
- λ es el coeficiente de regularización, que controla la fuerza de la penalización.

4.3. Métricas de Evaluación

La evaluación se realizará utilizando métricas estándar, *accuracy*, precisión, *recall*, *F1-score* y matriz de confusión. Estas métricas nos proporcionarán una comprensión detallada del rendimiento del modelo Transformer en la clasificación para el cribado del glaucoma en las retinografías.

4.3.1. Matriz de Confusión

Es un cuadro que resume el desempeño del modelo en la clasificación de clases, mostrando el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Proporciona una representación visual del rendimiento del modelo en términos de precisión y errores en la clasificación.

- TP (True Positive) – Son los valores que el algoritmo clasifica como positivos y que realmente son positivos.
- TN (True Negative) – Son valores que el algoritmo clasifica como negativos (0 en este caso) y que realmente son negativos.
- FP (False Positive) – Falsos positivos, es decir, valores que el algoritmo clasifica como positivo cuando realmente son negativos.
- FN (False Negative) – Falsos negativos, es decir, valores que el algoritmo clasifica como negativo cuando realmente son positivos.

4.3.2. Accuracy

La métrica de (*accuracy*) indica el porcentaje total de valores que el modelo ha clasificado correctamente, incluyendo tanto los positivos como los negativos.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

4.3.3. Precision

La métrica de precision se utiliza para determinar el porcentaje de valores clasificados como positivos que realmente son positivos.

$$Precision = \frac{TP}{(TP + FP)}$$

4.3.4. Recall (Recuperación)

El recall es una medida que indica la fracción de instancias positivas que el modelo identificó correctamente. Su fórmula es:

$$Recall = \frac{TP}{(TP + FN)}$$

4.3.5. F1-score

Es una métrica que combina precisión y recall en una medida única que proporciona un balance entre ambas. Su fórmula es:

$$F1{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.4. Consideraciones Finales

Es importante tener en cuenta algunas consideraciones y limitaciones potenciales de nuestro enfoque. Por ejemplo, la calidad de las imágenes puede variar entre conjuntos de datos, lo que podría afectar el rendimiento del modelo. Además, la aplicación de CLAHE puede modificar la apariencia de la imagen y potencialmente introducir artefactos o cambios no deseados, lo que requerirá una evaluación cuidadosa de sus efectos.

Además de los ajustes mencionados anteriormente, es importante considerar estudios adicionales que analicen cómo diferentes configuraciones de los parámetros del Transformer afectan la eficiencia y precisión del modelo en tareas de visión. Un estudio relevante en este contexto es el de Touvron et al. [Touvron et al., 2021], que aborda la eficiencia de datos y la destilación en transformers para imágenes. Es por ello que se realizarán diferentes experimentos para encontrar un modelo eficiente y de bajo costo computacional. Los parámetros del Transformer que se ajustarán son:

- **Dimensión de Proyección:** La dimensión de proyección en un Transformer determina la cantidad de características que se aprenden en cada capa. Ajustar esta dimensión puede mejorar la capacidad del modelo para capturar detalles específicos en los datos de entrada. Según Dosovitskiy et al. [Dosovitskiy et al., 2020], ajustar la dimensión de proyección es crucial para el rendimiento del modelo en diferentes tamaños de imágenes y conjuntos de datos.

- **Cabezas de Atención:** Las cabezas de atención en un Transformer permiten al modelo enfocarse en diferentes partes de la entrada simultáneamente. Aumentar el número de cabezas puede mejorar la capacidad del modelo para capturar dependencias complejas en los datos, aunque también aumenta el costo computacional. Vaswani et al. [Vaswani et al., 2017] demostraron que el uso de múltiples cabezas de atención mejora significativamente el rendimiento de los Transformers en tareas de procesamiento de lenguaje natural, lo cual también es aplicable a tareas de visión.
- **Unidades Transformer:** Las unidades o bloques de Transformer consisten en múltiples capas de atención y capas MLP (Perceptrón Multicapa). Aumentar el número de estas unidades puede mejorar la capacidad del modelo para aprender representaciones jerárquicas profundas, pero también aumenta el tiempo de entrenamiento y el uso de memoria. En investigaciones de ViT y BERT, Devlin et al. [Devlin et al., 2018] han demostrado que más capas pueden mejorar la precisión en tareas complejas, aunque con un incremento en los requisitos computacionales.
- **Capas MLP:** Las capas MLP en los bloques de Transformer son responsables de aprender combinaciones no lineales de las características. Ajustar el número de unidades en estas capas puede mejorar la capacidad del modelo para capturar patrones complejos. El estudio de Liu et al. [Liu et al., 2021] sobre los Transformers en visión mostró que la profundidad y la anchura de las capas MLP son críticas para el rendimiento del modelo.

4.5. Futuras Direcciones

En los siguientes pasos de esta investigación, se planea explorar la viabilidad de utilizar el Swin Transformer como una alternativa al Vision Transformer ViT para la detección de glaucoma en retinografías. El Swin Transformer es una arquitectura más moderna que ha mostrado un rendimiento prometedor en diversas aplicaciones de visión por computadora, destacándose por su capacidad para capturar relaciones a larga distancia de manera más eficiente que las arquitecturas tradicionales de Transformers.

Se llevará a cabo un estudio comparativo exhaustivo entre el Swin Transformer y el ViT B-16, evaluando su rendimiento en términos de precisión en la detección de glaucoma, eficiencia computacional y capacidad de generalización. Este análisis permitirá identificar las fortalezas y limitaciones de cada arquitectura, proporcionando así información crucial para futuras investigaciones en el ámbito de la detección de enfermedades mediante el uso de modelos Transformers en visión por computadora.

Capítulo 5

Pruebas y Resultados

En este capítulo se proporcionará una descripción detallada de los experimentos. En primer lugar, para los experimentos se utilizó un: procesador: Core i-5-10210U de 8 núcleos, RAM: 12GB y una GPU: NVIDIA GeForce MX250; debido a que la versión gratuita de Colab contamos con limitaciones de tiempo y puede compremeter los resultados al hacer el entrenamiento de modelos desde cero.

5.1. Aplicación del CLAHE a las Retinografías

Se empleará la retinografías ofrecidas por Kaggle, la cual consiste en una recopilación de 19 *datasets* como se mostró en la Tabla 4.1, gracias a Kiefer [Kiefer, 2023] que realizó esta recopilación para la minería de datos. Esta base de datos consta de 8621 retinografías para el entrenamiento, 5747 para validación y 2874 para pruebas, con dimensiones de 512 x 512. Para los experimentos se realizará una redimensión de las retinografías para el entrenamiento la cual será: 224 x 224. De igual manera se aplicarán rotaciones aleatorias de 0.2 y acercamientos aleatorios de 0.2 tanto en ancho y alto.

Se puede ver un ejemplo de la aplicación del CLAHE en una retinografía del conjunto de datos seleccionada en la Figura 5.1.

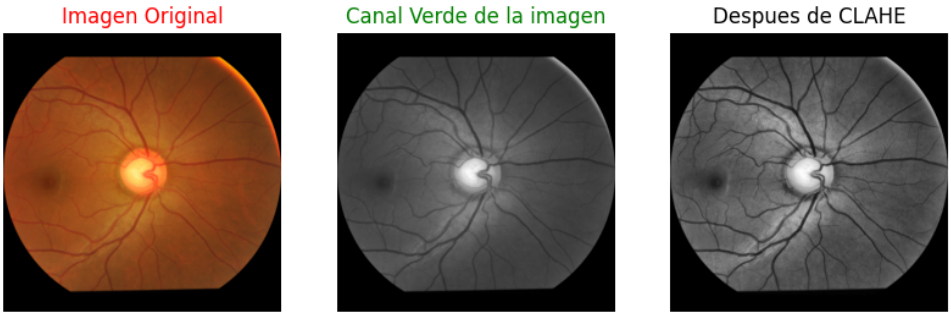

Figura 5.1: Ejemplo, Aplicación del CLAHE a una retinografía.

5.2. Descripción de los Experimentos

Los ajustes realizados en los modelos se centraron en la dimensión de proyección, la cantidad de cabezas de atención, las unidades Transformer, los bloques Transformer y las unidades de la MLP. Estos ajustes se realizaron con base en la literatura científica existente, buscando un modelo óptimo tanto en resultados como en costo computacional.

5.2.1. Experimentos Realizados

Para estos experimentos se creó una Vision Transformer desde cero utilizando las librerías de TensorFlow. Se realizaron variaciones en la arquitectura como se mencionó. Se realizaron 10 experimentos, que sirvieron para encontrar los mejores parámetros para el mejor modelo, se utilizaron 30 épocas para entrenar cada modelo, los resultados obtenidos se pueden observar en la Tabla 5.1.

Para todos los modelos se emplearon los siguientes hiperparámetros: una tasa de aprendizaje (*learning rate*) de 0.001, que asegura una convergencia eficiente del modelo; un decaimiento de los pesos (*weight decay*) de 0.0001, utilizado para regularizar el modelo y evitar el sobreajuste; un tamaño de lote (*batch size*) de 16, que determina la cantidad de ejemplos de entrenamiento utilizados en cada iteración; y un tamaño de parche (*patch size*) de 16, diseñado para facilitar el procesamiento eficiente de los datos y la actualización de los parámetros del modelo.

N°	Dim. Proy.	Head Att.	T. Units	T. Layers	MLP head	Acc. Test	Hiperparámetros
1	768	12	3072, 768	12	3072, 768	85.82 %	86,014,114
2	768	24	1536, 768	6	3072, 768	80.4 %	91,981,032
3	32	8	64,32	10	1024, 512	55.6 %	1,850,024
4	64	8	128,64	10	1024, 512	60.2 %	14,928,834
5	128	8	256,128	10	1024, 512	85.8 %	32,279,938
6	128	6	256,128	8	1024, 512	87.28 %	30,168,194
7	128	8	256,128	10	512, 256	70.1 %	19,040,386
8	128	6	256,128	10	1024, 512	82.2 %	30,961,538
9	256	6	512,256	8	1024, 512	75.32 %	66,887,938
10	256	8	512,256	10	1024, 512	65.2 %	75,829,506

Tabla 5.1: Experimentos realizados, Dim Proy: Dimensión de Proyección, Head Att.: Cabezales de atención, T. Units: Unidades Transformers, T. Layers: Capas Transformer, MLP head: Unidades MLP.

5.2.2. Experimento N° 1

Para el primer experimento se utilizó el modelo Vision Transformer Base (ViTB-16) propuesto por la librería keras, propuesto por Dosovitskiy et al. [Dosovitskiy et al., 2020]. Es importante recalcar que este modelo es de la biblioteca keras, por lo cual ya fue pre-entrenado, por ello sólo se utilizaron 10 épocas. Para este modelo y con la máquina

utilizada demoró aproximadamente 3 horas por época al ser un modelo de 86,014,114 hiperparámetros. Se obtuvo en el entrenamiento: accuracy: 0.9399 - loss: 0.1537 - val accuracy: 0.8672 - val loss: 0.5513.; se puede observar en 5.2.

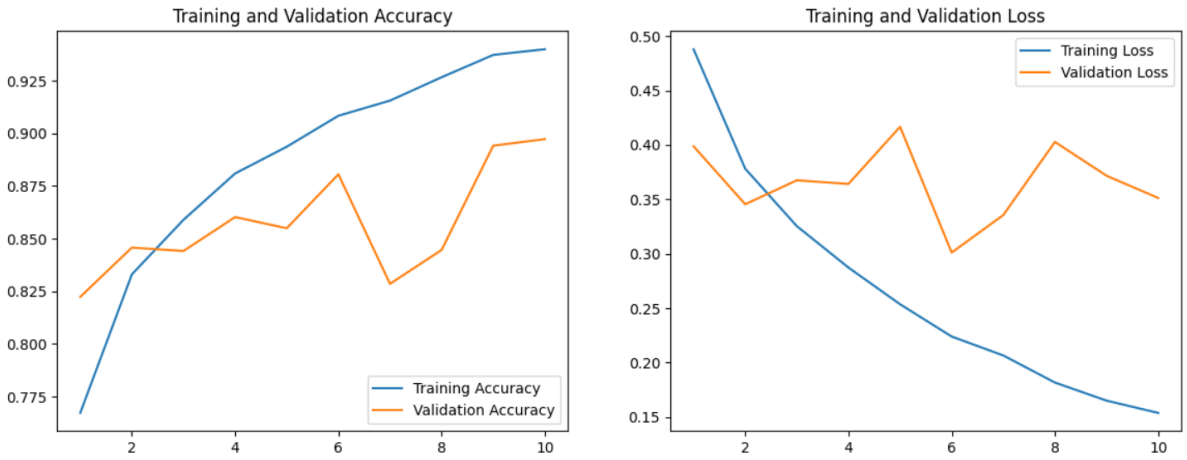

Figura 5.2: Entrenamiento del modelo base.

En el conjunto de prueba se obtuvo un *accuracy*: 85.82 %, *precision*: 61.21 %, *recall*: 97.37 % y *F1-score:* 75.17 %. Se puede ver la matriz de confusión con el conjunto de prueba en la Figura 5.3 y en procentaje en la Figura 5.4.

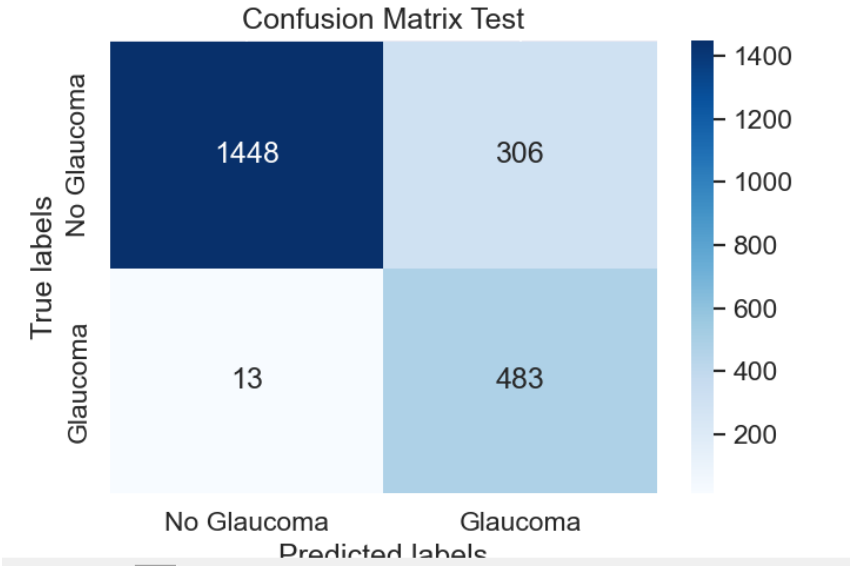

Figura 5.3: Matriz de Confusión de conjunto de prueba.

5.2.3. Experimento N°6

Después de lo observado en los experimentos realizados se realizaron ajustes y aumentar los hiperparámetros pero manteniendo el costo computacional no muy elevado. Para entrenar el modelo demoró aproximadamente 30 minutos por época, al tener 30,168,194 hiperparámetros.

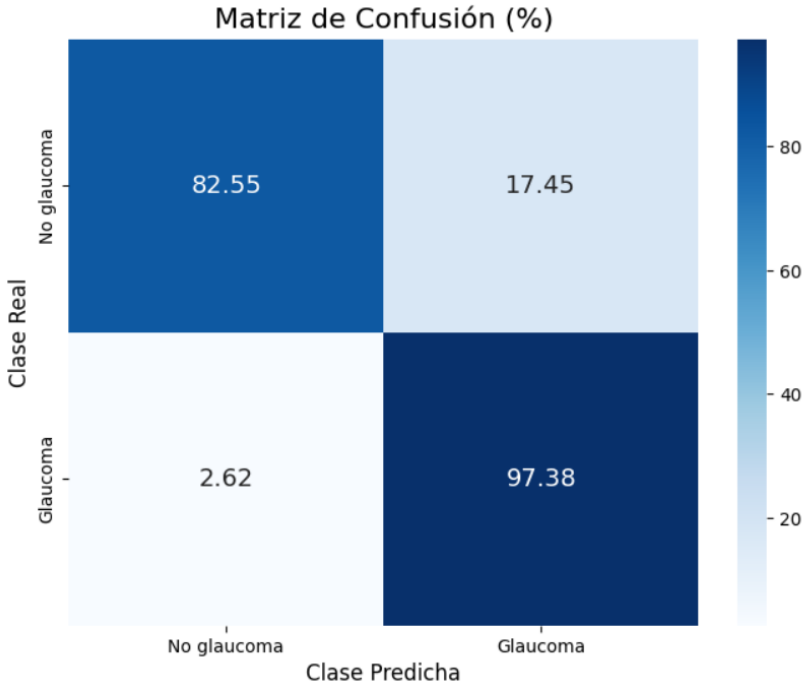

Figura 5.4: Matriz de Confusión de conjunto de prueba porcentaje.

Se obtuvo: accuracy: 0.8356 - loss: 0.5540 - val accuracy: 0.8628 - val loss: 0.4925. Graficado en la Figura 5.6.

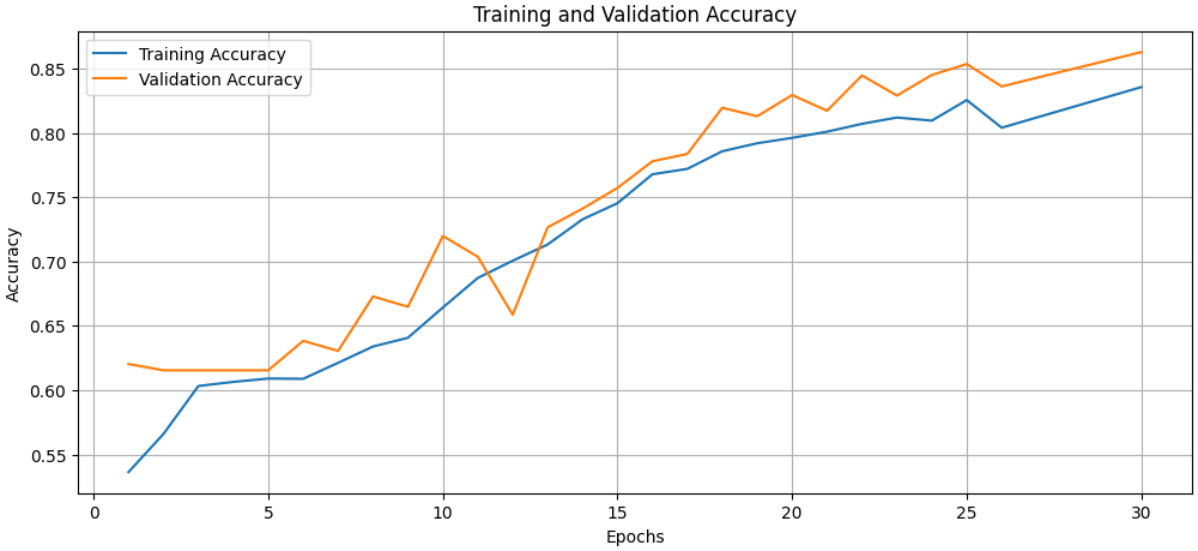

Figura 5.5: Entrenamiento del Experimento N°6, modelo ViT.

En el conjunto de prueba se obtuvo un *accuracy*: 87.11 %, *precision*: 65.28 %, *recall*: 88.71 % y *F1-score*: 75.21 %. La matriz de confusión se puede observar en la Figura 5.7 y en procentajes en la Figura 5.8.

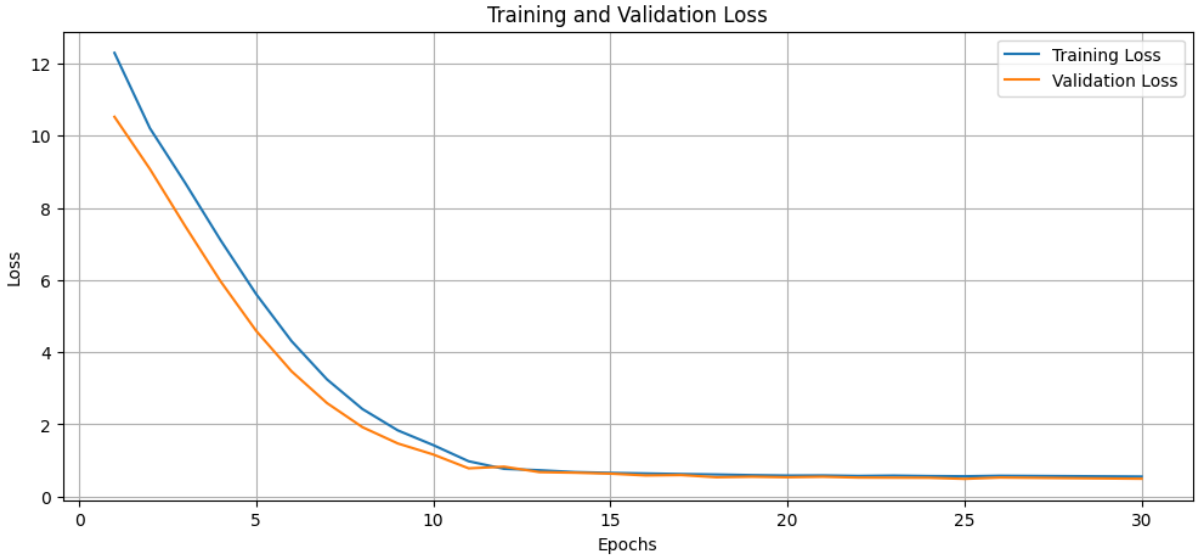

Figura 5.6: Entrenamiento del Experimento N°6, modelo ViT.

5.2.4. Comparación de arquitecturas

Se puede observar una comparativa entre las mejores arquitecturas, el modelo base Experimento N°1 y el Experimento N°6 en el la Tabla 5.2.

Experimento	1	6
Patch size	16	16
Dim. Proy.	768	128
Head Att.	12	6
T. Units	3072, 768	256, 128
T. Layers	12	8
MLP head	3072, 768	1024,512
Hiperparámetros	86,014,114	30,168,194
Tiempo/época	3 horas	30 min.
Accuracy	85.82 %	87.11 %
Precission	61.21 %	65.28 %
Recall	97.37 %	88.71 %
F1-score	75.17 %	75.21 %

Tabla 5.2: Tabla comparativa de Arquitecturas. Exp: Experimento, Dim Proy: Dimensión de Proyección, Head Att.: Cabezales de atención, T. Units: Unidades Transformers, T. Layers: Capas Transformer.

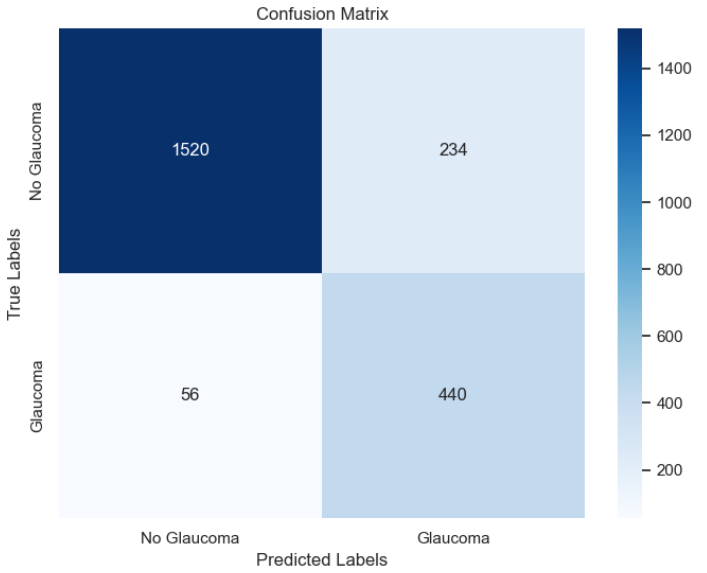

Figura 5.7: Matriz de Confusión de conjunto de prueba del Experimento N°6.

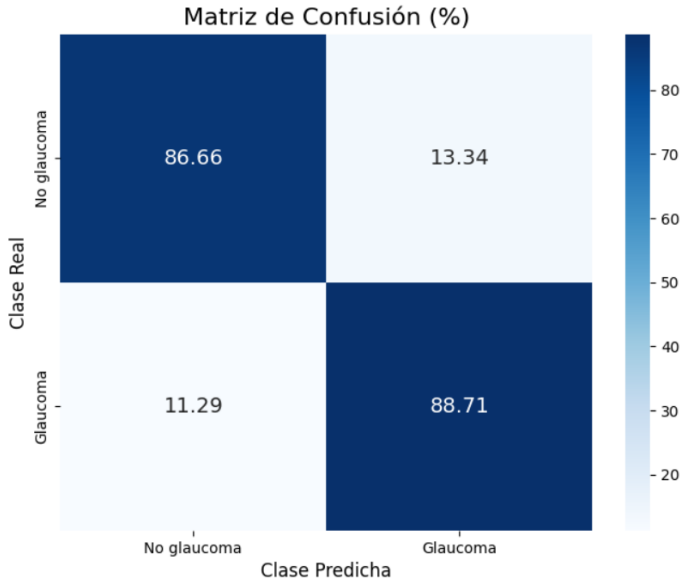

Figura 5.8: Matriz de Confusión de conjunto de prueba en porcentajes del Experimento N°6.

Capítulo 6

Conclusiones Preliminares

El presente trabajo de tesis ha buscado abordar el problema del cribado del glaucoma mediante la aplicación de Vision Transformer (ViT), específicamente utilizando el modelo ViT B-16. La propuesta incluyó la mejora de las retinografías aplicando la técnica de Ecualización de Histograma Adaptativo de Contraste Limitado (CLAHE) y la evaluación del modelo ViT B-16 entrenado durante 10 épocas. Dada la naturaleza pesada del modelo ViT B-16, se planteó el objetivo de encontrar un modelo con resultados comparables o superiores pero con un costo computacional más bajo. Para ello, se realizaron diversos experimentos ajustando la arquitectura del modelo.

A continuación, se presentan las conclusiones preliminares basadas en los dos mejores experimentos realizados:

- **Costo Computacional:** El Experimento 2 muestra una reducción significativa en el número de hiperparámetros, disminuyendo de 86,014,114 a 30,168,194, lo cual representa una reducción aproximada del 65 %. Esto se traduce en un tiempo de entrenamiento por época considerablemente menor, de 3 horas a 30 minutos.

- **Rendimiento del Modelo:** En términos de *accuracy*, el Experimento 2 supera al Experimento 1, alcanzando un 87.11 % frente a un 85.82 %. De giual manera en precisión, el Experimento 2 supera al Experimento 1, alcanzando un 65.28 % frente a un 61.21 %. Asimismo, el valor de recall, aunque menor que en el Experimento 1, se mantiene en un nivel alto (88.71 %) frente a (97.37 %) del Experimento 1. El F1-score del Experimento 2 es ligeramente superior (75.21 %) comparado con el del Experimento 1 (75.17 %), indicando un balance adecuado entre precisión y recall.

Los resultados preliminares sugieren que es posible lograr un rendimiento comparable o incluso superior utilizando una arquitectura de Vision Transformer más eficiente y con menor costo computacional. El Experimento 2 demostró ser particularmente prometedor, con una mejor precisión y un tiempo de entrenamiento significativamente reducido. Esto indica que la búsqueda de una arquitectura optimizada, como se propone en este trabajo, es viable y puede ofrecer beneficios significativos tanto en términos de rendimiento como de recursos computacionales.

6.1. Problemas encontrados

El costo computacional resultó ser una limitante al momento de realizar pruebas, restringiendo el tiempo para la creación y pruebas de nuevas arquitecturas. Sin embargo, los resultados obtenidos son prometedores logrando el objetivo deseado.

Bibliografía

[Ahamed et al., 2020] Ahamed, P. S. R., Kundu, S., Khan, T., Bhateja, V., Sarkar, R., and Mollah, A. F. (2020). Handwritten Arabic numerals recognition using convolutional neural network. *Journal of ambient intelligence humanized computing/Journal of ambient intelligence and humanized computing*, 11(11):5445–5457.

[Alwazzan et al., 2021] Alwazzan, M. J., Ismael, M. A., and Ahmed, A. N. (2021). A Hybrid Algorithm to Enhance Colour Retinal Fundus Images Using a Wiener Filter and CLAHE. *Journal of digital imaging*.

[Aníbal Bregón Bregón, 2022] Aníbal Bregón Bregón, G. S. B. (2022). Estimación de profundidad monocular online con Transformers eficientes.

[Batista et al., 2020] Batista, F. J. F., Diaz-Aleman, T., Sigut, J., Alayon, S., Arnay, R., and Angel-Pereira, D. (2020). Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis Stereology*, 39(3):161–167.

[BotPenguin, 2023] BotPenguin (2023). Softmax Function: Advantages and Applications | BotPenguin.

[David, 2022] David, R. S. J. (2022). Regularización de redes neuronales artificiales para la clasificación de imágenes de retinopatía diabética.

[Deepak et al., 2012] Deepak, K. S., Jain, M., Joshi, G. D., and Sivaswamy, J. (2012). Motion pattern-based image features for glaucoma detection from retinal images. In *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '12, New York, NY, USA. Association for Computing Machinery.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Diaz-Pinto et al., 2019] Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., and Navea, A. (2019). CNNs for Automatic Glaucoma assessment using FundUs Images: An extensive validation. *Biomedical Engineering Online*, 18(1).

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[Gao et al., 2021] Gao, X., Qian, Y., and Gao, A. (2021). COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models.

[Garcia Molina, 2024] Garcia Molina, A. (2024). Reconocimiento de imágenes en nubes de puntos con redes neuronales transformer.

[Hendrycks and Gimpel, 2016] Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

[Huarote Zegarra Raúl, 2022] Huarote Zegarra Raúl, E. (2022). .Estrategia para la detección de tipos de enfermedades oculares usando Red Neuronal SOM

.

[Jesús, 2020] Jesús (2020). Redes Neuronales Convolucionales en Profundidad.

[Kiefer, 2023] Kiefer, R. (2023). Smdg, a standardized fundus glaucoma dataset.

[Liu et al., 2018] Liu, Z., Gao, Y., and Zhu, S. (2018). Research of screening method based on glaucoma image. In *Proceedings of the 3rd International Conference on Multimedia and Image Processing*, ICMIP '18, page 114–118, New York, NY, USA. Association for Computing Machinery.

[Liu et al., 2021] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows.

[Miranza, 2021] Miranza (2021). Glaucoma ocular ¿Qué es y cómo se produce? | Miranza.

[O'Neill et al., 2014] O'Neill, E. C., Gurria, L. U., Pandav, S. S., Kong, Y., Brennan, J., Xie, J., Coote, M., and Crowston, J. G. (2014). Glaucomatous Optic Neuropathy Evaluation Project. *JAMA Ophthalmology*, 132(5):560.

[Raju et al., 2022] Raju, M. H., Lohr, D. J., and Komogortsev, O. (2022). Iris print attack detection using eye movement signals. In *2022 Symposium on Eye Tracking Research and Applications*, ETRA '22, New York, NY, USA. Association for Computing Machinery.

[Review, 2023] Review, E. (2023). IA para el glaucoma: ¿qué está por venir?

[Sanchez-Bocanegra et al., 2023] Sanchez-Bocanegra, C. L., Fernandez-Luque, L., and Solé-Ribalta, A. (2023). Detección temprana de cáncer de piel mediante clasificador de imágenes basado en Inteligencia Artificial.

[Santurkar et al., 2018] Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018). How does batch normalization help optimization? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Talaat et al., 2019] Talaat, M.-A., Raed, N., Medhat, A., Ashraf, R., Essam, M., ElKashlan, R., and Abdel-Hamid, L. (2019). Glaucoma Detection from Retinal Images using Generic Features. *ACM Association for Computing Machinery*.

[Ting et al., 2017] Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C., Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., Sivaprasad, S., Varma, R., Jonas, J. B., He, M. G., Cheng, C.-Y., Cheung, G. C. M., Aung, T., Hsu, W., Lee, M. L., and Wong, T. Y. (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318(22):2211–2223.

[Touvron et al., 2021] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

[Tékouabou et al., 2020] Tékouabou, S. C. K., Alaoui, E. A. A., Chabbar, I., Cherif, W., and Satori, H. (2020). Machine Learning Aprroach for Early Detection of Glaucoma from Visual Fields. *ACM Association for Computing Machinery*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Wu et al., 2019] Wu, L., Liu, Y., Shi, Y., Sheng, B., Li, P., Bi, L., and Kim, J. (2019). Detect glaucoma with image segmentation and transfer learning. In *Proceedings of the 32nd International Conference on Computer Animation and Social Agents*, CASA '19, page 37–40, New York, NY, USA. Association for Computing Machinery.