Universidad de La Laguna

MEMORIA DEL TRABAJO DE FIN DE GRADO

VISUALIZACIÓN E INTERPRETACIÓN DE REDES NEURONALES CONVOLUCIONALES MEDIANTE DROPOUT ESPACIAL

GRADO EN INGENIERÍA ELECTRÓNICA INDUSTRIAL Y AUTOMÁTICA

Alumno:Abián Hernández Hernández
Tutor Académico: José Francisco Sigut Saavedra

Agradecimientos

Le agradezco a mi familia y amigos por haberme apoyado y acompañado, así como a mis profesores por haberme formado como ingeniero en estos años.

Resumen

El glaucoma es una de las principales causas de la ceguera total a nivel mundial, por lo que su identificación a tiempo es crucial.

En los últimos años con el avance de las tecnologías, en concreto con el avance de las técnicas de computación y Deep Learnig, se ha abierto otra posible vía para generar una herramienta que facilite el diagnóstico del glaucoma. En esta línea de investigación es en la que trabaja un grupo de profesores de la ULL entre los que se encuentra el tutor de este Trabajo de Fin de Grado.

El objetivo principal del mismo es estudiar en qué partes de una retinografía se fija una red neuronal entrenada para clasificar si el ojo está sano o no, y ver si existe coincidencia entre esas regiones de la imagen y estructuras anatómicas o sectores del disco óptico que son relevantes para los especialistas en sus diagnósticos.

Con ese fin, se han utilizado métodos de visualización tipo CAM: Grad-CAM, Grad-CAM++ y Score-CAM, para generar mapas de calor de la imagen de entrada que resalten las zonas de interés. Además de estos métodos, se introduce uno nuevo desarrollado por el tutor, el cual selecciona un conjunto mínimo de pixels de la imagen que constituyen la información fundamental a preservar por la red neuronal para poder hacer la predicción sin que la probabilidad asignada a la clase correspondiente se vea apenas alterada. Se han evaluado sus prestaciones y se ha visto que puede ser una alternativa muy interesante a los otros métodos.

Abstract

Glaucoma is one of the leading causes of total blindness worldwide, so its early identification is crucial.

In recent years, with the advancement of technologies, specifically with the advancement of computing techniques and Deep Learnig, another possible way has been opened to generate a tool that facilitates the diagnosis of glaucoma. It is in this line of research that a group of professors from the ULL works, among whom is the tutor of this Final Degree Project.

The main objective is to study in which parts of a retinography a trained neural network is fixed to classify whether the eye is healthy or not, and to see if there is a coincidence between those regions of the image and anatomical structures or sectors of the optic disc that are relevant to specialists in their diagnoses.

To this end, CAM-type visualization methods: Grad-CAM, Grad-CAM ++ and Score-CAM have been used to generate heat maps of the input image that highlight the areas of interest. In addition to these methods, a new one developed by the tutor is introduced, which selects a minimum set of pixels of the image that constitute the fundamental information to be preserved by the neural network in order to be able to make the prediction without the probability assigned to the corresponding class look barely altered. Its benefits have been evaluated and it has been seen that it can be a very interesting alternative to the other methods.

Índice

Índice de figuras

Índice de tablas

1. INTRODUCCIÓN

1.1. MOTIVACIÓN DEL TRABAJO E HIPÓTESIS DE PARTIDA

Este trabajo se encuadra dentro de la línea de investigación de aplicación de técnicas de Deep Learning o aprendizaje profundo al diagnóstico del Glaucoma que está llevando a cabo un grupo de profesores del Departamento de Ingeniería Informática y de Sistemas del cual forma parte el tutor del mismo. En dicha línea de investigación participa también el Hospital Universitario de Canarias (HUC).

Las técnicas de Deep Learning con redes neuronales convolucionales se han convertido en los últimos años en las herramientas más potentes para abordar muchos de los problemas que se presentan en Visión por Computador, gracias a su capacidad para aprender, no solo a reconocer objetos a partir de unas características dadas, sino incluso a aprender dichas características a partir de datos, en este caso imágenes, con las que se entrenan estos sistemas.

A pesar de los excelentes resultados encontrados con la aplicación de este tipo de redes, uno de sus principales inconvenientes es, sin duda, su opacidad a la hora de explicar cómo se obtienen dichos resultados, de ahí su denominación habitual de "caja negra". Esta condición puede suponer una limitación importante en su utilización en diferentes campos como la Medicina que es el que nos ocupa en este trabajo. Si no es posible tener alguna pista de cómo se llegó a un determinado diagnóstico, por ejemplo, parece difícil poder confiar en estos sistemas que son muy buenos pero no perfectos por lo que en cada momento lo que se tiene es una decisión sobre un paciente concreto que debe estar apoyada por algo más que un valor de probabilidad, como ocurre en el caso del diagnóstico llevado a cabo por humanos.

El problema en sí es muy complejo porque estos sistemas también lo son. La forma habitual de abordar el problema suele ser intentar determinar qué partes de una imagen de entrada son aquellas que tiene una mayor relevancia en la predicción de la red neuronal. En esta línea se sitúan los llamados métodos de atribución o de saliency, cuya salida suele ser una visualización de estas zonas relevantes, resaltándolas de alguna manera.

El glaucoma es una de las patologías más importantes que afectan a la retina y constituye una de las principales causas de ceguera en el mundo. Se la suele denominar como la "ceguera silenciosa" dada la poca sintomatología que presenta hasta que la enfermedad está bastante avanzada. Es por ello que es fundamental detectarla precozmente.

Disponemos, en el grupo de investigación, de redes neuronales entrenadas con imágenes de sujetos sanos y con glaucoma que han sido utilizadas en este trabajo para evaluar algunos métodos de visualización bien conocidos con el fin de arrojar un poco de luz sobre aquello que resulta más interesante para la red en este problema. Además, se introduce un nuevo método que selecciona características mediante dropout en la primera capa de las redes y que permite preservar aquellas zonas de la imagen que contienen la información mínima necesaria para hacer la predicción, sin apenas verse afectada la probabilidad asignada a la clase correspondiente. Nuestra hipótesis de partida es que este nuevo método permite obtener visualizaciones más claras y precisas, destacando solo lo realmente fundamental. Para comprobar dicha hipótesis se han realizado numerosos experimentos.

Por último hay que decir que el método presentado se ha explicado de una manera no formal, intentando que se comprenda la idea intuitiva de su funcionamiento. Una formalización matemática del mismo es, por supuesto, necesaria e imprescindible para su utilización en otros contextos pero queda fuera del nivel de un trabajo de esta naturaleza. De la misma manera, la evaluación realizada es susceptible de un tratamiento estadístico más riguroso pero esto llevaría también más tiempo

con la consiguiente dificultad asociada. En el resto del capítulo se hará una breve introducción a las redes neuronales convolucionales y al problema del diagnóstico del glaucoma. El capítulo 2 de la memoria explica el método propuesto. El capítulo 3 relata el desarrollo experimental llevado a cabo, y en el capítulo 4 se incluyen las conclusiones y líneas abiertas de este trabajo.

1.2. DEEP LEARNING Y REDES NEURONALES CONVOLUCIONALES

La inteligencia artificial es una de las ramas de la Informática encargada de estudiar modelos capaces de imitar comportamientos inteligentes, esta puede ser aplicada a una gran cantidad de campos de estudios distintos, como son la capacidad de entender el lenguaje (N.L.P) o la de moverse y adaptarse (robótica).Dentro de todas las ramas que se aplica la inteligencia artificial la que está relacionada con el aprendizaje, entendido este como la generalización de conceptos a través de las experiencias, recibe el nombre de Machine Learning.

Entre todas las técnicas de Machine Learning, la que más ha despuntado en los últimos tiempos son las redes neuronales, llamadas así por su inspiración en su contraparte biológica que forma parte del cerebro animal, estas redes neuronales consisten en un conjunto de nodos llamados neuronas artificiales que actúan como funciones matemáticas recibiendo una o más entradas a las cuales se les asigna un peso distinto(parámetros modificables del modelo), estas se suma de manera ponderada y antes de dar la salida se pasa por una función de activación o ReLu que es una función no lineal la cual nos permite conectar neuronas entre si y desarrollar respuestas más complejas.

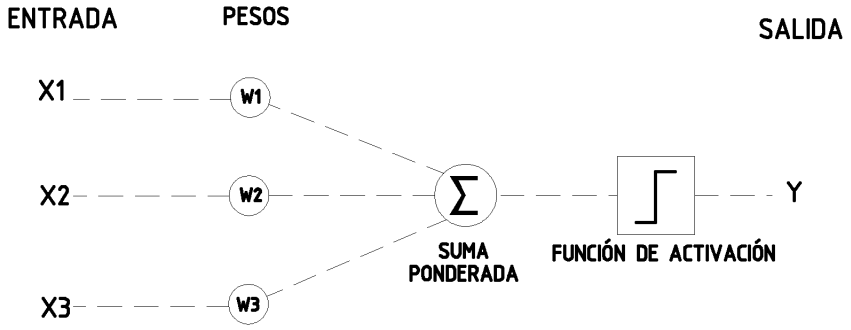

Figura 1: *Esquema básico de una neurona*

Los algoritmos Deep Learning son aquellos basados en redes neuronales los cuales usan un gran número de capas jerarquizadas de neuronas, pasando de información sencilla en las capas más altas a información más compleja y abstracta en las capas más bajas. Dentro de estos algoritmos una de las arquitecturas más reconocidas son las redes neuronales convolucionales, las cuales utilizamos para el estudio que se ha realizado en este trabajo. Las redes neuronales convolucionales o CNN (derivado del inglés Convolutional Neural Network) son redes de Deep Learning, comúnmente utilizadas para el análisis y clasificación de imágenes. Este tipo de redes suelen poseen una capa de entrada y otra de salida, así como otras denominadas: capas de convolución, capas de pooling y capas completamente conectadas.

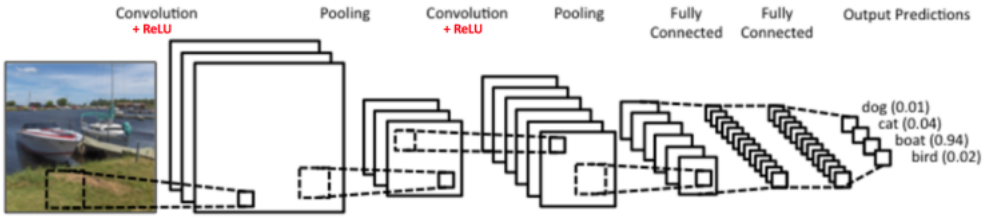

Figura 2: *Esquema de CNN(extraído de https://www.clarifai.com/technology)*

1.2.1. Capas de convolución

Aplican filtros a la imagen original de entrada para generar un mapa de características. El tamaño de este mapa de características está definido por tres parámetros que han de ser decididos antes de comenzar con las convoluciones: número de filtros por convolución, stride, refiriéndose al desplazamiento que determina como se aplica dicho filtro sobre toda la imagen, y zero padding, matriz de ceros alrededor del borde de la matriz de entrada.

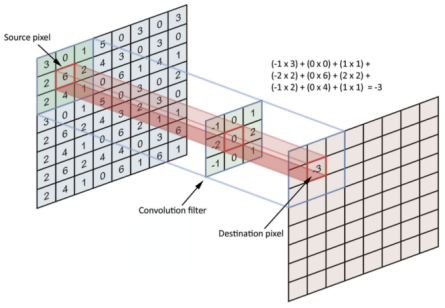

Figura 3: *Reprecentación de una convolución(extraído de https://es.quora.com/C%C3%B3mo-funcionan-las-redes-neuronales-convolucionales)*

1.2.2. Capas de pooling

Son similares a las capas de convolución pero estas actúan de una manera más concreta aplicando una operación de max pooling, tomando el mayor valor de una determinada región del filtro, aunque en esta capa también se puede aplicar otras operaciones como es el average pooling, que en vez de tomar máximos toma el valor medio.

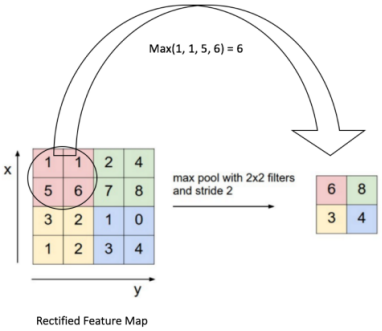

Figura 4: *Ejemplo de max poolings(extraído de https://cs231n.github.io/convolutional-networks)*

1.2.3. Capas completamente conectada

Estas capas se sitúan antes de la salida de una CNN. El término completamente conecgtadas implica que todas las neuronas de la capa anterior están conectados con la capa siguiente. Estas capas se encargan de compilar los datos extraídos de las capas anteriores para formar una salida.

1.3. EL DIAGNÓSTICO DEL GLAUCOMA

El glaucoma es una patología ocular que se presenta de manera asintomática en las primeras fases de la enfermedad y luego provoca defectos en el campo visual y pérdida progresiva de visión, siendo así una de las principales causas de ceguera total en el mundo. Esta enfermedad, en la mayor parte de sus variantes, obstruye el sistema de drenaje del ojo lo que causa que el fluido intraocular no se pueda desalojar aumentando la presión interna del ojo, dañando el nervio óptico y provocando la pérdida progresiva de fibras nerviosas, lo que provoca las alteraciones en la visión y la perdida de la misma.

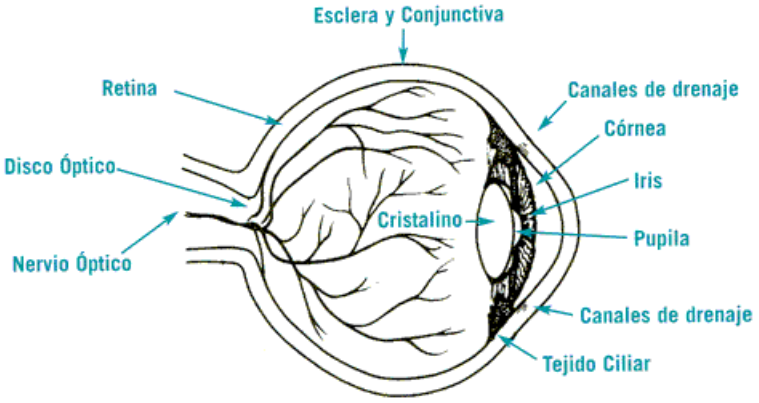

Figura 5: *Esquema anatómico del ojo*
(extraído de https://www.glaucoma.org/es/que-es-el-glaucoma.php)

Existen una gran variedad de glaucomas todos con origen y evolución distintas, siendo el más común el glaucoma de ángulo abierto o glaucoma crónico simple.

En la detección de galucoma existen dos técnicas de diagnóstico que usan imágenes de retina : retinografías y OCT (Optical Coherence Tomography). En este trabajo nos hemos centrado en las retinografías como las del ejemplo que se muestra en la figura 6.

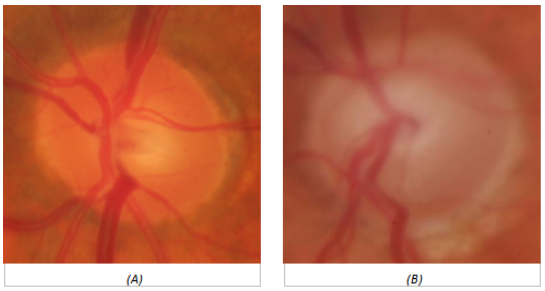

Figura 6: *A)Recorte de una retinografía en el disco óptico de un paciente sano (B)Recorte de una retinografía en el disco óptico de un paciente con glaucoma.*

2. MÉTODO PROPUESTO

2.1. SELECCIÓN DE CARACTERÍSTICAS EN MACHINE LEARNING

Los métodos de selección de características son bien conocidos en el ámbito del machine learning [1]. Dentro de este ámbito, en este trabajo nos hemos centrado en el aprendizaje supervisado y, más concretamente, en problemas de clasificación. Como su propio nombre indica, dado un algoritmo de aprendizaje L y un conjunto de datos etiquetados D con características $X_1, X_2, ..., X_n$, los métodos de selección de características tienen como objetivo reducir la dimensionalidad del problema de n a p, p<=n, con el fin de:

- Mejorar el rendimiento del algoritmo de aprendizaje basado en esas características.
- Simplificar la estructura del modelo de aprendizaje generado.
- Mejorar la visualización y comprensión del problema de aprendizaje.

Los métodos de selección de características se suelen dividir en métodos de filtrado, métodos Wrapper, y métodos embebidos.

Los métodos de filtrado seleccionan las características a partir de los datos disponibles, atendiendo a las propiedades intrínsecas de dichas características sin tener en cuenta el algoritmo de aprendizaje con el que se va a trabajar. Se trata, en esencia, de establecer un ranking entre las mismas que dé cuenta de su capacidad discriminante.

Los métodos Wrapper, por el contrario, tiene en cuenta tanto los datos como el algoritmo de aprendizaje a la hora de elegir las mejores características. Es el método en el que se ha inspirado este trabajo y por ello le dedicaremos una atención especial.

Los métodos embebidos son similares a los Wrapper con la salvedad de que se integran en el propio proceso de entrenamiento del clasificador en lugar de aplicarse una vez se ha entrenado.

2.2. MÉTODO WRAPPER DE SELECCIÓN DE CARACTERÍSTICAS

El método Wrapper de selección de características fue introducido por Kohavi y John [2] en 1997 con la idea de mejorar el rendimiento de los clasificadores a partir de la selección de un conjunto de características del conjunto original que fuera más óptimo desde ese punto de vista. En la figura 12 se muestra un esquema de dicho método.

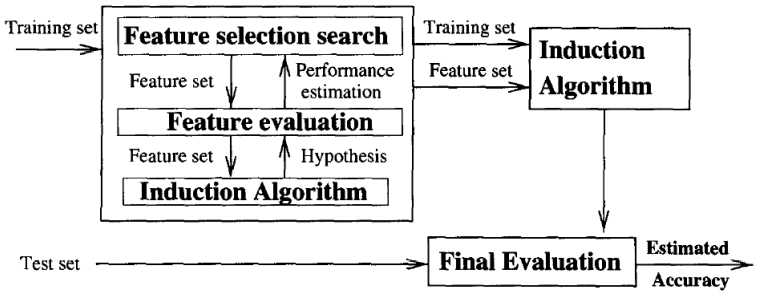

Figura 7: *Esquema del método Wrapper (extraído de [2])*

En el método Wrapper, el algoritmo de clasificación (induction algorithm) para el que se quiere encontrar las características óptimas se considera como una caja negra ya que solo importa su rendimiento para un conjunto de características dado. Así, se hace necesario disponer de una muestra de entrenamiento y validación para el proceso de selección, y una vez se ha encontrado el conjunto óptimo buscado, se procedería a hacer una evaluación final con una muestra de test independiente. Otra cuestión importante a destacar es la búsqueda en el espacio de posibles subjconjuntos de características. Este espacio es de tamaño $0(2^n)$ para n características por lo que la búsqueda exhaustiva es inviable para la gran mayoría de problemas que se dan en la práctica, lo que lleva a plantear estrategias subóptimas. Las más habituales son la selección hacia adelante (forward selection) y la selección hacia atrás (backward selection), o una combinación de ambas. La estrategia de selección hacia adelante consiste en partir del conjunto vacío de características e ir añadiendo sucesivamente características del conjunto original hasta que se considere que se ha encontrado el subconjunto óptimo. Por el contrario, la estrategia de selección hacia atrás parte del conjunto completo de características y va eliminando progresivamente hasta llegar al subconjunto óptimo. La estrategia hacia adelante suele ser computacionalmente más eficiente pero la de selección hacia atrás parece capturar mejor las interacciones entre las características y por eso nos centraremos más en ella.

2.3. DESCRIPCIÓN DEL MÉTODO PROPUESTO

El método propuesto en este trabajo está inspirado en el método Wrapper descrito anteriormente. En un Trabajo Fin de Grado presentado muy recientemente **[3]** se implementó una variante de este método con la idea de analizar el rendimiento de redes neuronales convolucionales, previamente entrenadas, utilizando dropout para la selección de las características (canales) más óptimas. El enfoque en este TFG es parecido pero con un objetivo diferente: seleccionar el conjunto mínimo de características que es necesario preservar en una red para que la probabilidad asignada a una determinada clase, para una imagen de entrada dada, no se vea afectada significativamente. Por lo tanto, se puede considerar que este conjunto mínimo incluye aquellas características que juegan un papel más relevante en la predicción que hace la red.

En una red neuronal convolucional existen dos partes principales, la parte convolucional, en sí misma, que aprende las mejores características para un problema dado, y la parte de clasificación con esas características. Por lo tanto, las características en la red ya entrenada están representadas por los diferentes canales en cada una de las capas de convolución. Sin embargo, a diferencia del TFG anterior, en el que se consideraron canales completos, para el estudio que vamos a realizar, vamos a considerar grupos de características en paquetes que incluyen todas las neuronas en la misma posición en todos los canales de una determinada capa. Lo hacemos así para poder visualizar e interpretar los resultados encontrados en base a la distribución espacial de las características, en la línea de lo que hacen otros métodos de esta naturaleza.

Como se comentó más arriba, se utilizará una estrategia de eliminación de características tipo dropout siguiendo el ejemplo de la técnica bien conocida con este nombre para evitar el overfitting [4]. En la figura 8 se muestra el efecto de aplicar la técnica dropout en una red neuronal genérica.

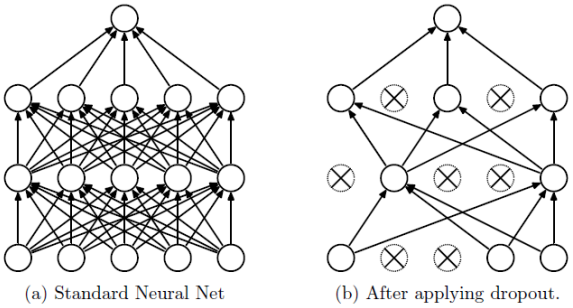

Figura 8: *Esquema de la técnica Dropout (extraído de [4])*

La manera habitual de aplicar esta técnica durante el entrenamiento de una red neuronal consiste en poner a cero la activación de ciertas neuronas elegidas al azar, habitualmente en la parte densamente conectada de la red donde existen muchas más conexiones y, por lo tanto, muchas más interdependencias entre las neuronas.

Nuestro método también hace uso del dropout pero aplicado sobre las neuronas de los canales de la parte convolucional de la red, de forma que las activaciones de las neuronas de dichos canales se ponen a cero ante cualquier entrada por lo que, de forma efectiva, es como si esa característica no jugara ningún papel en la predicción de la red. En la implementación práctica de esta idea, lo que se ha hecho es multiplicar la salida de las correspondientes neuronas en los canales de una determinada capa por un tensor de 0s y 1s, según lo que se pretenda desactivar, como se muestra esquemáticamente en la figura 9:

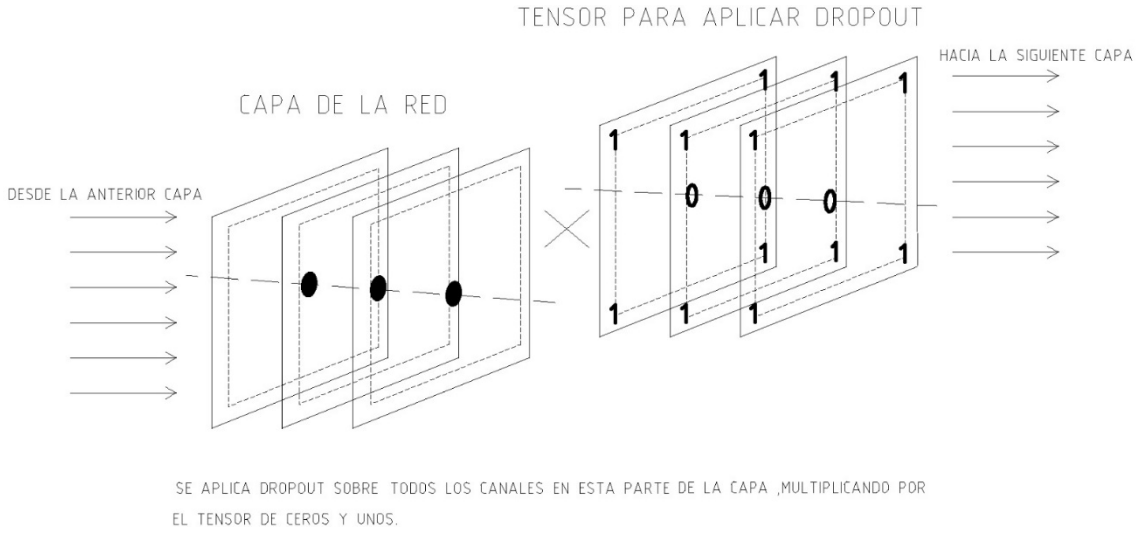

Figura 9: *Método propuesto de manera práctica*

Más allá de la técnica dropout, el resto del método se ajusta bastante al esquema del método Wrapper descrito en la sección anterior, con una sola imagen, y utilizando la estrategia de búsqueda hacia atrás para ir descartando características, progresivamente. De forma un poco más precisa, los pasos que hay que dar para la implementación del método son los siguientes:

1. Dada una red neuronal convolucional, ya entrenada, y una imagen de entrada de referencia.

2. Se selecciona una capa de la red y se van eliminando iterativamente sus características mediante dropout, agrupadas por paquetes según su posición, como se ha descrito con anterioridad. Como hemos seguido la estrategia de búsqueda hacia atrás, se parte de considerar todas las características y en cada iteración del método se va eliminando un paquete hasta que no quede ninguno o se imponga alguna otra condición de parada. Hay que decir que, aunque a priori es posible utilizar cualquier capa de la red, se ha optado por escoger la primera capa ya que es la más cercana a la imagen de entrada y esto permite una mejor interpretación de los resultados aparte de que trabajamos a una resolución igual o similar.

3. El criterio para ir descartando estos paquetes de características consiste en ver el efecto que tiene su descarte sobre la probabilidad que la red asigna a la imagen de entrada para la clase elegida. Así, en cada iteración del método se busca eliminar el paquete de características que minimiza dicho efecto,$D_i(p)$, según la expresión:

$$D_i(p) = |Prob_c^0(I) - Prob_c^i(I,p)|$$

Donde p representa los paquetes de características que se están evaluando en cada iteración (el que se está considerando en la iteración actual junto a los que ya se han eliminado en iteraciones anteriores),$Prob_c^0(I)$ es la probabilidad original de la imagen I para la clase C, y $Prob_c^i(I,p)$ es la probabilidad de la imagen I para la clase C, ejecutado el dropout con la configuración de paquetes dada por p. De esta manera, en cada iteración del método se descartarán los paquetes de características que menos afectan a la decisión tomada por la red que es el propósito buscado. Como condición de parada, hemos considerado que la caida en la probabilidad sea de entre un 1 y un 5 por ciento, esto es, $D_i(p)/Prob_c^0(I) = 0{,}01$.

2.4. TRABAJO RELACIONADO

Existen muchos métodos para la visualización e interpretación del comportamiento de las redes neuronales convolucionales. En general, los podemos agrupar en tres categorías principales que pasamos a describir, brevemente, a continuación:

2.4.1. Métodos basados en gradientes

Como su nombre indica, estos métodos se basan en hacer una propagación hacia atrás del gradiente de una clase con respecto a la imagen de entrada para resaltar aquellas partes de la misma que más influyen en la predicción de la red. La forma más básica de estos métodos es la que se describe en [5] con el cálculo directo del gradiente. Otras técnicas como Deconvnet y Guided Backpropagation [6] hacen manipulaciones sobre el gradiente original para mejorar el resultado. Aunque este tipo de métodos producen visualizaciones de "grano fino", los mapas generados suelen ser de baja calidad y bastante ruidosos, lo cual dificulta su interpretación. Técnicas como SmoothGrad y VarGrad [7] tratan de aliviar esta situación.

2.4.2. Métodos basados en perturbaciones

Este tipo de métodos funcionan perturbando, de alguna manera, la imagen de entrada y viendo qué efecto tiene dicha perturbación en la probabilidad de salida de la red para una determinada clase. Zeiler y Fergus [8] perturban la imagen, directamente, tapando partes de la imagen poniendo sus píxeles a cero o a un valor fijo que se considere adecuado. El método RISE [9] constituye una aproximación al problema más sofisticada pero, seguramente, la forma más óptima de abordarlo es la descrita en [10. En estos trabajos, se plantea la oclusión de la imagen como un problema de optimización en el que se pretende encontrar la máscara de tamaño mínimo que ocluya la misma, preservando la información útil imprescindible que permita a la red hacer su predicción. En este sentido, se parece bastante al enfoque del nuevo método presentado en este TFG, y nos parece una forma bastante adecuada de abordar el problema.

2.4.3. Métodos basados en CAM (Class Activation Mapping)

La salida de estos métodos es, habitualmente, un mapa de calor que se obtiene como la combinación lineal ponderada de mapas de activación de capas de convolución, normalmente, de las más profundas. Lo que diferencia a unos métodos de otros, dentro de esta categoría, es la forma de calcular los pesos en dicha combinación lineal. El método original CAM [11] obligaba a modificar la estructura de la red original pero, posteriormente, métodos como Grad-CAM [12] y Grad-CAM++ [13], generalizaron el CAM y evitaron dicho inconveniente. Más recientemente, se ha publicado una nueva variante, denominada Score CAM [14] que también incluye algunos elementos de los métodos de pertubación. Dado que estos métodos han sido los elegidos en este trabajo, se describirán en más detalle en el siguiente capítulo.

2.5. SIMILITUDES Y DIFERENCIAS DEL MÉTODO PRESENTADO RESPECTO A OTROS MÉTODOS

Una vez descritos los principales tipos de métodos de visualización e interpretación, estamos en condiciones de comentar alguna de las similitudes y diferencias entre estos métodos y el presentado en este trabajo:

1. En lo que se refiere a los métodos basados en gradientes, la similitud principal sería el hecho de que, en ambos casos, podemos hacer una evaluación de "grano fino" a nivel de píxel. De resto, como ya se comentó, estos métodos suelen presentar el inconveniente principal de la inestabilidad y ruido del cálculo de los gradientes que no existe en nuestro caso.

2. Se podría considerar el método presentado como un método de perturbación aplicado sobre la primera capa de la red en lugar de sobre la imagen de entrada, como hacen el resto de métodos en esta categoría. Esto tiene la ventaja de no tener que recurrir a estrategias de oclusión tales como "tapar" zonas de la imagen con máscaras de tamaños, formas y colores, arbitrarios, o difuminar dichas zonas con el fin de intentar evitar los posibles artefactos que estas máscaras podrían introducir. En nuestro caso, la técnica dropout constituye una manera uniforme y adecuada de hacer dicha oclusión, seguramente, menos dada a la generación de artefactos (aunque esto habría que demostrarlo). Otra cuestión fundamental es cómo decidir qué partes se tapan y cuáles no a la hora de evaluar este tipo de métodos. Para imágenes de tamaño 224*224, que son las que hemos manejado en este trabajo, las posibilidades son inabarcables y es por ello que se debe plantear una estrategia subóptima de búsqueda. Los métodos descritos en [10] plantean la búsqueda de la región relevante de tamaño mínimo como un problema de optimización de las oclusiones pero los mismos autores reconocen la dificultad de la tarea. Además, utilizan también una red neuronal para llevarla a cabo con todo lo que ello supone. En nuestro caso, el enfoque seguido es mucho más sencillo y consiste, simplemente, en la selección de características hacia atrás que a la que ya nos hemos referido. Por último, hay que destacar otra ventaja de este tipo de métodos y es su interpretación objetiva, no ambigua, en base al efecto directo que las oclusiones tienen sobre la predicción de la red en forma de caída de la probabilidad.

3. Los métodos tipo CAM son muy populares pero presentan algunos inconvenientes importantes. Por un lado, se trata de métodos un tanto arbitrarios en el sentido de que su salida es la combinación ponderada de las activaciones de los canales de la última capa de convolución. Esto da lugar a mapas de calor que resultan atractivos e intuitivos pero que no son fáciles de interpretar en base a las probabilidades de predicción de la red como si ocurre con los métodos basados en oclusiones. La otra dificultad principal de estos métodos viene dada por su condición de "grano grueso" ya que al realizar el cálculo sobre una capa profunda de la red, lo que se obtiene está en una resolución muy inferior a la de la imagen de entrada y se hace preciso hacer un upsampling final que no deja de ser una mera interpolación. El nuevo método no presenta este tipo de dificultades.

3. DESARROLLO EXPERIMENTAL

3.1. MATERIALES Y MÉTODOS

En esta sección se describen los materiales y métodos utilizados para el desarrollo experimental llevado a cabo en este trabajo.

3.1.1. Bases de datos de imágenes

En este trabajo se ha utilizado la base de datos imágenes de fondo de ojo RIM-ONE DL. RIM-ONE DL surge como una iniciativa para aunar y revisar las tres versiones de RIM-ONE existentes y adaptarlas a los problemas de Deep Learning. Las imágenes fueron capturadas en el HUC, el Hospital Universitario Miguel Servet (Zaragoza) y el Hospital Clínico Universitario San Carlos (Madrid). Todas las imágenes de esta base de datos han sido recortadas alrededor de la cabeza del nervio óptico con una misma proporción. Además, encontramos una sola imagen por paciente y ojo. En total, tenemos 313 fotos de pacientes sanos y 172 fotos de pacientes con glaucoma. La responsabilidad de la categorización previa de estas imágenes corrió a cargo de tres médicos expertos en la enfermedad. En la figura 6 del capítulo 1 se muestra un ejemplo de estas imágenes.

3.1.2. Redes neuronales convolucionales

En este estudio se han usado dos tipos de arquitecturas de redes distintas: la VGG16 y la VGG19. Se han elegido estas redes debido a que son muy utilizadas, testeadas y documentadas, y se ajustan a las necesidades de este estudio.

Estas dos redes habían sido previamente entrenadas con imágenes naturales y han vuelto a ser entrenadas con imágenes de fondo de ojo de sujetos sanos y con glaucoma. VGG16 y VGG19, son un invento de VGGNet en el Visual Geometry Group en 2014[], siendo VGG las siglas del grupo de investigación que las desarrollaron, y el número siguiente especifica el número de capas con la que cuenta la red. Esta arquitectura se puede tomar como una mejora de otra llamada AlexNet a la que se le sustituyeron los filtros de gran tamaño de las capas convolucionales con varios filtros de tamaño 3x3. Estos cambios permitieron que la VGG pudiera alcanzar el top 5 con un 92,7 % de precisión en ImageNet, una base de datos con más de 14 millones de imágenes y más de 1000 clases, convirtiéndose en una arquitectura muy popular en modelos de reconocimientos de objetos.

VGG16:

La entrada de la primera capa convolucional es una imagen RGB 244x244. Esta imagen pasa por varias capas convolucionales que aplican filtros de tamaño 3x3 que se pueden ver como una transformación lineal de los canales de entrada (seguida de una no linealidad tipo ReLu). En la parte no convolucional hay tres capas completamente conectadas (FC). Las dos primeras tienen 4096 neuronas cada una, y la tercera realiza una clasificación con mil neuronas de salida tipo Softmax, una para cada clase. Las capas de Max pooling cambian la dimensionalidad de los mapas de características de la red para reducir el cómputo y también para que la red pueda abarcar una zona más grande de la imagen de forma efectiva.

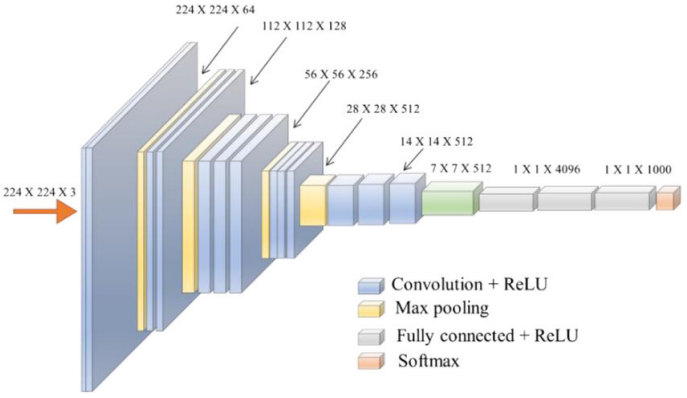

Figura 10: *Esquema de estructura VGG16(extraído de https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac)*

VGG19:

Estructuralmente, la VGG19 se asemeja mucho a la VGG16 ya que parten de una arquitectura común con la diferencia del número de capas que posee cada una, 19 en el caso de VGG19 y 16 en la VGG16. De resto, son muy parecidas como se puede apreciar en la figura 11.

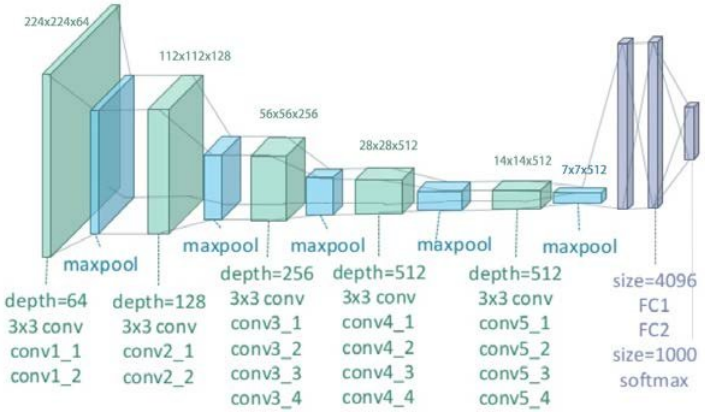

Figura 11: *Esquema de estructura VGG19 (extraído de https://www.researchgate.net/figure/llustration-of-the-network-architecture-of-VGG-19-model-conv-means-convolution-FC-means_fig2_325137356)*

3.1.3. Métodos de visualización e interpretación

Aparte del nuevo método presentado en este TFG, hemos considerado también tres métodos muy populares de la variante CAM: GradCAM, GradCAM++ y ScoreCAM, que pasamos a describir a continuación:

GradCAM:

El método Grad-CAM (Gradient-weighted Class Activation Mapping) es una generalización del método CAM. Usa la información del gradiente que fluye desde una determinada clase de salida hacia la última capa convolucional de la CNN para medir la importancia de cada canal de esa capa en la decisión tomada. La razón de usar esta capa es que las representaciones más profundas en la red capturan mejor la información de alto nivel y todavía se conserva la información espacial que luego se pierde en las capas no convolucionales.

Para obtener el mapa de localización discriminante de clases de ancho u y alto v para cualquier clase c, y una imagen de entrada dada, primero calculamos el gradiente del score que la red asigna a esa clase, y^c (antes del softmax) con respecto a los mapas de características A^k de una capa convolucional. Estos gradientes que fluyen hacia atrás se promedian espacialmente para obtener los pesos α_k^c

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

Después de calcular los pesos para la clase objetivo c, se realiza una combinación ponderada de los mapas de activación, seguida de una unidad lineal rectificada (ReLU). Se aplica ReLU a la combinación lineal porque solo nos interesan las características que tienen una influencia positiva en la clase de interés. Sin ReLU, el mapa de activación de clases resalta más de lo necesario y, por lo tanto, logra un bajo rendimiento de localización.

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

Finalmente, se hace un upsampling del resultado para ponerlo en la misma resolución que la imagen original y, usualmente, se muestra en forma de mapa de calor para que resulte más intuitivo. En la figura 12 se muestra todo el proceso relativo a este método con aplicación a otras tareas como Image Captioning y Visual Question Answering.

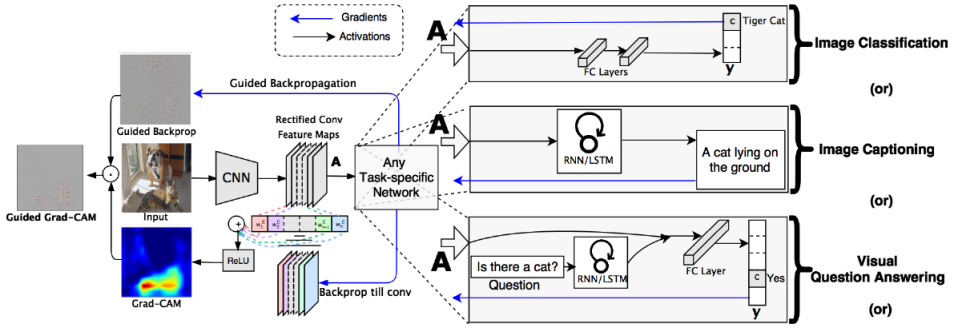

Figura 12: *Esquema Grad-CAM (extraído de [12])*

GradCAM++: Este método es una evolución del GradCAM en la que se supone que se obtienen mejores visualizaciones de aquello en lo que se está fijando la red. La mejora se aprecia sobre todo en una mejor localización de los objetos relevantes así como una mejor explicación de las ocurrencias múltiples de un mismo objeto. Dado que el cálculo de los pesos que multiplican a los diferentes canales es similar al GradCAM, no entraremos en ese detalle en la memoria.

ScoreCAM: La principal diferencia de este método con los dos anteriores es la forma de calcular los pesos que multiplican los canales de la última capa de convolución de la red. En lugar de utilizar gradientes, se basa en perturbar la imagen de entrada con las máscaras de activación obtenidas para dicha imagen. El score asignado a esas entradas con oclusión se utiliza como peso para dicho mapa de activación. De esta manera, se consigue evitar la inestabilidad y el ruido asociado al cálculo de los gradientes como se ilustra en la figura 13.

Figura 13: *La derivada parcial cambia bruscamente para un cambio indistinguible en la imagen de entrada (extraído de [14])*

A nivel de funcionamiento podemos dividir en los siguientes pasos:

- Lo primero es calcular las activaciones de los canales de la última capa de convolución para una imagen de entrada dada.
- A cada mapa de activación obtenido se le hace un upsampling usando interpolación bilineal al mismo tamaño que la imagen de entrada
- Los mapas resultantes se normalizan al rango [0,1] para mantener las intensidades relativas entre los píxeles. La normalización se logra utilizando la siguiente fórmula:

$$A_{i,j}^k = \frac{A_{i,j}^k}{\max A^k - \min A^k}$$

- Una vez completada la normalización de los mapas de activación, las áreas resaltadas en dichos mapas se multiplican, pixel a pixel, por la imagen original para obtener una imagen enmascarada M^k

$$M^k = A^k * I$$

- Dichas imágenes enmascaradas se pasan a través de la red para obtener el score correspondiente S_k^c que es utilizado como peso en la combinación lineal de los mapas de activación aplicando ReLU porque solo interesa las características que tienen una influencia positiva en la clase de interés.

$$|w_k^c = S_k^c \; L_S^c core - CAM = ReLU(\sum_K w_k^c A^k)$$

Al igual que en los otros métodos en esta categoría, el último paso consiste en hacer un nuevo upsampling para igualar las dimensiones de la imagen de entrada.

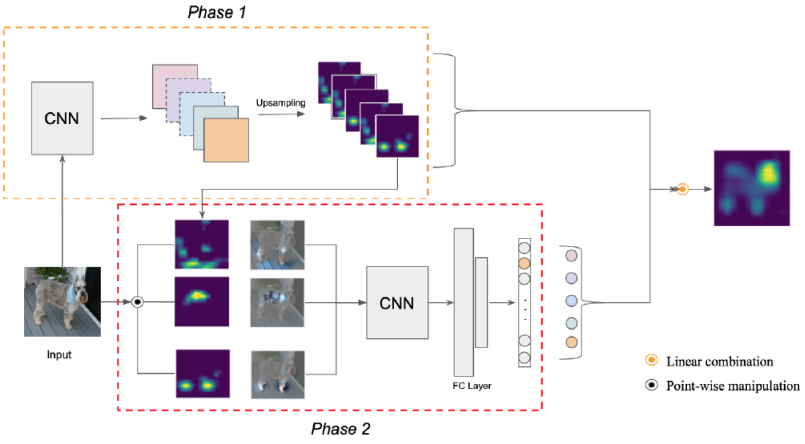

Figura 14: *Esquema Score-CAM (extraído de [14])*

3.2. EXPERIMENTOS REALIZADOS

Los experimentos realizados se han agrupado en diferentes casos de estudio. En todos los casos, se han seguido las mismas pautas y se ha utilizado la misma notación. En concreto, se han considerado cuatro variantes del problema: red VGG16 evaluada sobre sujetos sanos, red VGG16 evaluada sobre sujetos con glaucoma, red VGG19 evaluada sobre sujetos sanos, y red VGG19 evaluada sobre sujetos con glaucoma. En cada uno de estos casos, se han evaluado dos redes resultantes de dos entrenamientos diferentes con los mismos datos que hemos denominado entrenamiento 1 y entrenamiento 2.

La evaluación se ha hecho se forma cualitativa y cuantitativa. Cualitativamente, se han generado los mapas de calor de las salidas de los diferentes métodos. Aparte, se ha planteado una situación diferente controlada, en la sección 3.2.1, para comparar más fácilmente las salidas en base a lo esperado. Cuantitativamente, se ha medido el grado de solape de las salidas de los métodos tanto con regiones anatómicas de interés médico, previamente segmentadas: disco, copa, rim, vasos y fondo, como con los sectores que, deforma estándar, suelen considerar los especialistas para distinguir espacialmente entre unas zonas y otras del disco y copa. En la figuras 15 y 16 se muestra un ejemplo de las regiones y sectores considerados. Para generar las máscaras binarias de la salida de los métodos, hemos tenido que umbralizar dicha salida, normalizarla entre 0 y 1, y quedarnos con aquellos valores superiores a 0.75 con la idea de retener las zonas de mayor interés para la red neuronal. Esto no deja de ser un tanto arbitrario pero, como ya se comentó en una sección anterior, la salida de los métodos tipo CAM no se puede interpretar en términos de probabilidades que permitan la aplicación de un criterio más objetivo. Aunque, en principio, nuestro método no da una salida pensada para ser umbralizada de esta manera, se han utilizado los valores de caída de la probabilidad en las distintas iteraciones como referencia para poder llevarlo a cabo a efectos de comparación.

Todos los experimentos se han programado en Python con Keras y han sido ejecutados en el entorno de Google Colab.

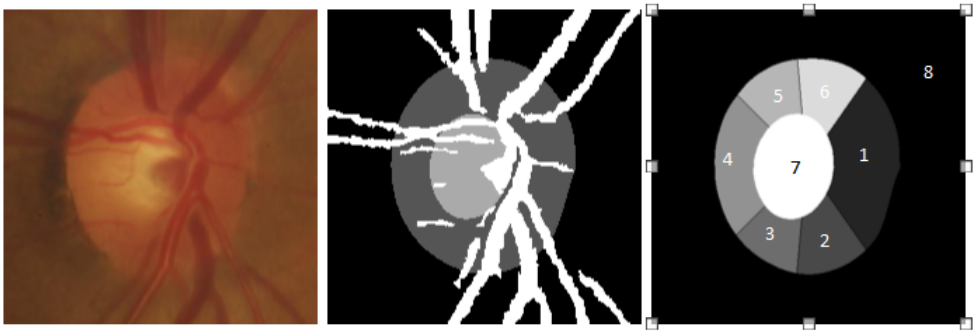

Figura 15: *De derecha a izquierda: imagen original, regiones anatómicas consideradas, sectores: nasal (1), nasal inferior (2), temporal inferior (3), temporal (4), temporal superior (5), nasal superior (6), copa (7), fondo (8).*

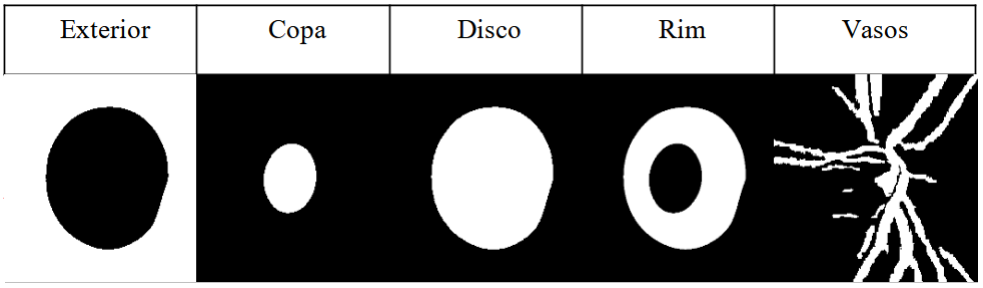

Figura 16: Desglose de las regiones anatómicas consideradas.

3.2.1. Ilustración del tipo de salida de los diferentes métodos

Seguramente, el inconveniente principal a la hora de evaluar la calidad de estos métodos de visualización es la imposibilidad de disponer de un ground truth con el que poder compararnos, como ocurre en otro tipo de situaciones. Lo que sí podemos hacer es plantear un escenario en el que a priori tengamos una idea bastante clara de lo que la red debería señalar como lo más relevante para poder hacer su predicción. Para ello, se ha cogido una red VGG19 y se ha entrenado con imágenes de fondo de ojo pero ya segmentadas en las que aparecen la copa, el disco y el fondo con diferentes niveles de intensidad (los mismos valores en todas las imágenes). Se logró obtener una precisión superior al 0.9 para distinguir entre normales y glaucomas que se puede considerar como satisfactoria.

Lo que se ha hecho es calcular los mapas de calor para la salida de todos los métodos. Dado que el método propuesto no se presta demasiado a una salida de este tipo, lo que se ha hecho es determinar la región mínima a preservar y señalarla como la más relevante en el mapa de calor, poniendo todos los pixels seleccionados al máximo valor. El resultado es el que se muestra en la figura 17.

Se puede observar la enorme diferencia entre la salida de los tres métodos CAM, que son muy similares entre sí, y el propuesto en este TFG. Parece claro que la única información disponible que tiene la red para poder hacer la predicción debe estar relacionada, de alguna manera, con los tamaños relativos del disco y la copa. Teniendo en cuenta esto, parece mucho más fácilmente interpretable y precisa la salida de nuestro método que la de los métodos CAM, ya que esta se concentra en los contornos de las zonas relevantes y, seguramente, esa información le permite a la red determinar los tamaños relativos en capas más profundas. Sin embargo, los métodos CAM dan una salida tipo "mancha" en la que no queda demasiado claro cómo se puede deducir la relación

entre los tamaños de disco y copa. Podría ocurrir que ambas salidas fueran válidas y que los métodos estuvieran poniendo de manifiesto dos estrategias diferentes de la red para llegar a lo mismo. En todo caso, parece mucho más evidente la estrategia que parece señalar nuestro método.

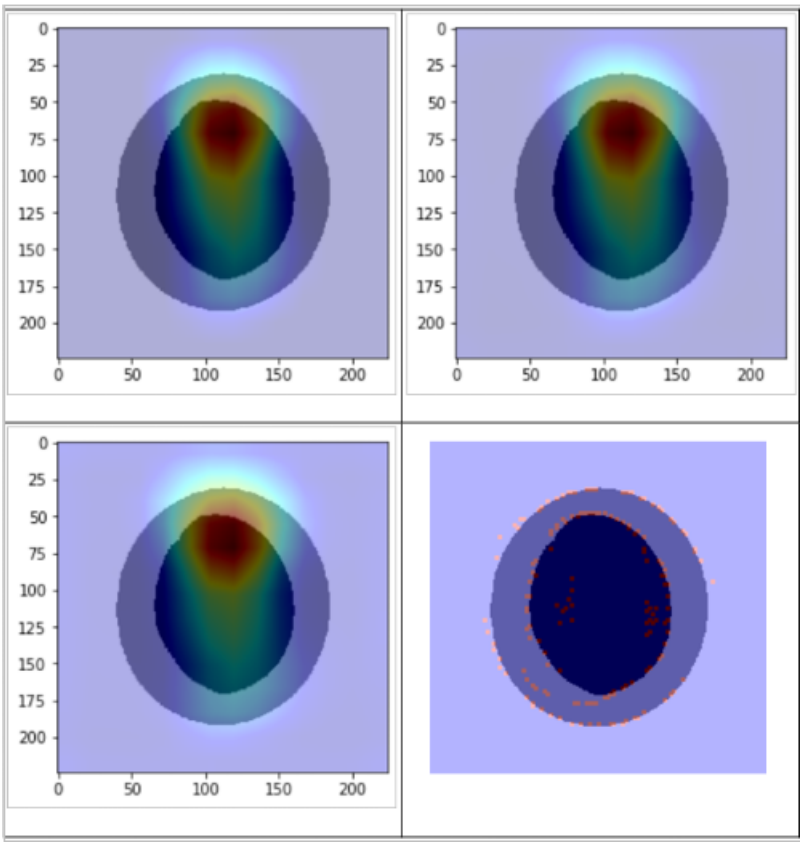

Figura 17: *Mapas de calor de los metodos. De izquierda a derecha y arriba y abajo: GradCAM, GradCAM++, ScoreCAM y nuestro método.*

3.2.2. Evaluación de los diferentes métodos

Se han considerado los cuatro casos ya comentados. Para cada uno de ellos se ha seguido el mismo esquema. Mostramos primero las tablas con el grado de solape promedio y la desviación típica entre las máscaras binarias generadas por los diferentes métodos, y las regiones anátomicas y sectores de interés, respectivamente. A continuación, se incluyen también algunos ejemplos de mapas de calor para diferentes imágenes.

Caso 1: Red VGG16 evaluada sobre sujetos sanos

Normales VGG16 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	1.00(0.03)	1.00(0.03)	1.00(0.03)	0.92(0.12)
Copa	0.02(0.09)	0.03(0.10)	0.05(0.11)	0.10(0.17)
Rim	0.97(0.09)	0.96(0.10)	0.95(0.12)	0.81(0.19)
Fondo	0.00(0.03)	0.00(0.03)	0.00(0.02)	0.08(0.12)
Vasos	0.43(0.16)	0.43(0.15)	0.44(0.15)	0.34(0.19)

Tabla 1: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo sano,entrenamiento 1 vgg16

Normales VGG16 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	0.99(0.06)	0.99(0.04)	0.99(0.03)	0.87(0.24)
Copa	0.01(0.04)	0.02(0.08)	0.09(0.17)	0.08(0.17)
Rim	0.98(0.07)	0.97(0.08)	0.90(0.17)	0.79(0.27)
Fondo	0.01(0.06)	0.01(0.04)	0.01(0.03)	0.13(0.24)
Vasos	0.43(0.16)	0.44(0.14)	0.43(0.14)	0.27(0.21)

Tabla 2: Solape medio y desviación típica de la salida de los diferentes métodos para regiones anátomicas de ojo sano, entrenamiento 2 vgg16

Normales VGG16 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.10(0.15)	0.10(0.15)	0.11(0.14)	0.21(0.28)
Sector 2	0.23(0.26)	0.21(0.23)	0.21(0.21)	0.21(0.28)
Sector 3	0.09(0.16)	0.10(0.16)	0.10(0.15)	0.13(0.21)
Sector 4	0.01(0.05)	0.01(0.05)	0.01(0.05)	0.08(0.15)
Sector 5	0.18(0.20)	0.19(0.19)	0.18(0.16)	0.08(0.15)
Sector 6	0.34(0.26)	0.34(0.24)	0.32(0.22)	0.10(0.17)
Sector 7	0.02(0.09)	0.03(0.10)	0.05(0.11)	0.11(0.17)
Sector 8	0.01(0.03)	0.01(0.03)	0.01(0.02)	0.08(0.12)

Tabla 3: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo sano, entrenamiento 2 vgg16

Normal VGG16 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.14(0.19)	0.13(0.16)	0.14(0.16)	0.23(0.36)
Sector 2	0.23(0.27)	0.21(0.23)	0.21(0.21)	0.13(0.27)
Sector 3	0.07(0.15)	0.07(0.12)	0.07(0.10)	0.12(0.25)
Sector 4	0.02(0.09)	0.02(0.07)	0.02(0.07)	0.10(0.24)
Sector 5	0.20(0.23)	0.20(0.20)	0.17(0.17)	0.08(0.20)
Sector 6	0.30(0.24)	0.32(0.22)	0.28(0.20)	0.11(0.22)
Sector 7	0.01(0.04)	0.02(0.08)	0.09(0.17)	0.08(0.17)
Sector 8	0.02(0.06)	0.01(0.04)	0.01(0.03)	0.13(0.24)

Tabla 4: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo sano, entrenamiento 2 vgg16

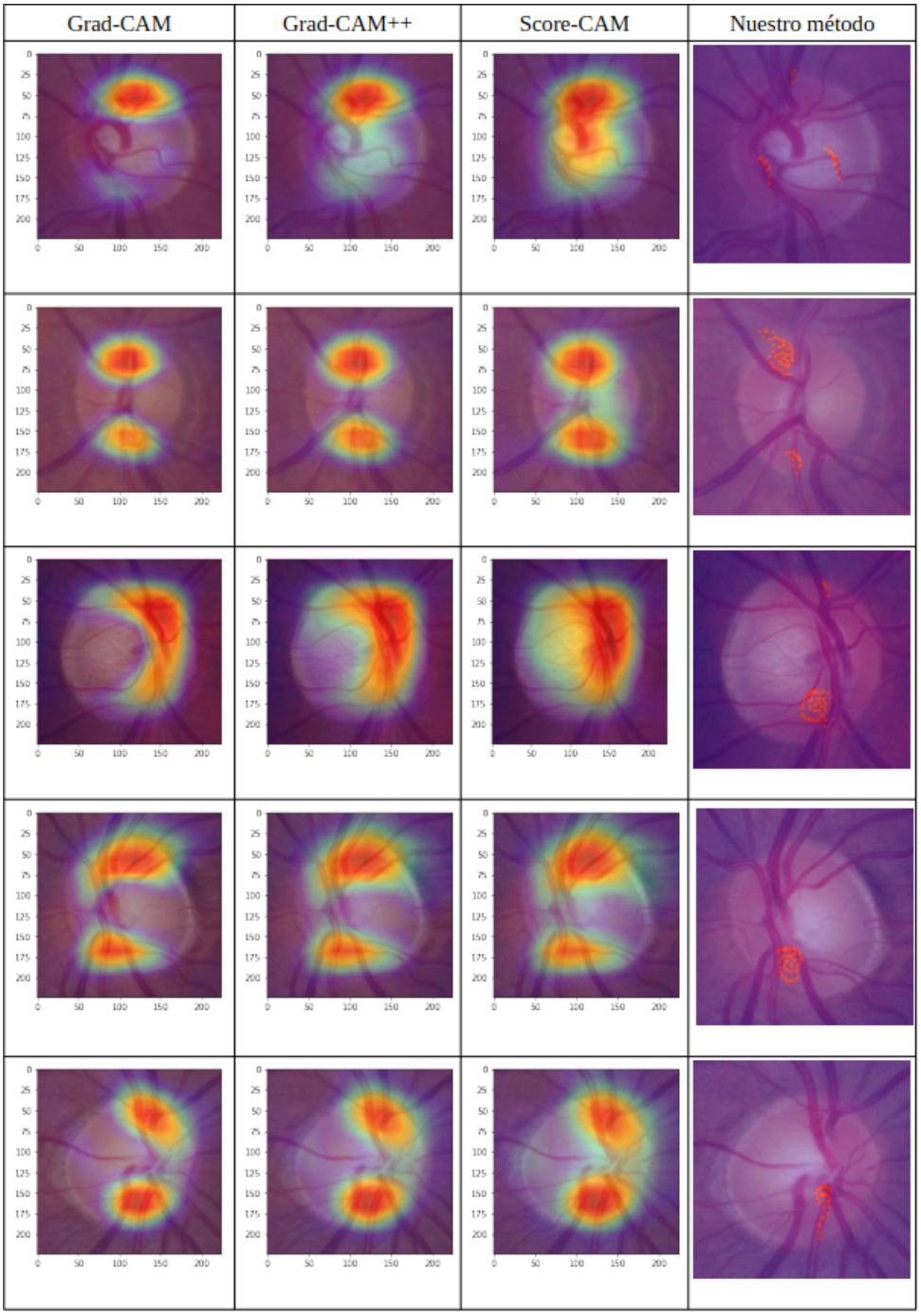

Figura 18: *Comparación de mapas de calor vgg16 sujetos sanos*

Caso 2: Red VGG16 evaluada sobre sujetos con glaucoma.

Glaucomas VGG16 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	0.73(0.40)	0.79(0.33)	0.93(0.19)	0.82(0.28)
Copa	0.67(0.40)	0.64(0.37)	0.36(0.35)	0.30(0.32)
Rim	0.06(0.11)	0.15(0.19)	0.56(0.36)	0.52(0.35)
Fondo	0.27(0.40)	0.21(0.33)	0.07(0.19)	0.18(0.28)
Vasos	0.19(0.18)	0.25(0.18)	0.37(0.19)	0.20(0.21)

Tabla 5: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo con glaucoma, entrenamiento 1 vgg16

Glaucomas VGG16 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	0.91(0.25)	0.96(0.16)	0.97(0.12)	0.86(0.32)
Copa	0.83(0.32)	0.87(0.25)	0.87(0.24)	0.47(0.46)
Rim	0.08(0.19)	0.08(0.17)	0.11(0.20)	0.39(0.43)
Fondo	0.09(0.25)	0.04(0.16)	0.02(0.12)	0.14(0.32)
Vasos	0.20(0.22)	0.21(0.20)	0.25(0.20)	0.46(0.31)

Tabla 6: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo con glaucoma, entrenamiento 2 vgg16

Glaucomas VGG16 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.09(0.04)	0.02(0.06)	0.08(0.16)	0.08(0.21)
Sector 2	0.01(0.04)	0.03(0.07)	0.10(0.20)	0.00(0.04)
Sector 3	0.02(0.05)	0.02(0.05)	0.03(0.09)	0.01(0.04)
Sector 4	0.01(0.02)	0.01(0.02)	0.01(0.03)	0.06(0.16)
Sector 5	0.00(0.02)	0.02(0.05)	0.13(0.20)	0.20(0.30)
Sector 6	0.01(0.03)	0.04(0.08)	0.20(0.25)	0.16(0.28)
Sector 7	0.67(0.40)	0.64(0.37)	0.36(0.35)	0.30(0.32)
Sector 8	0.27(0.40)	0.21(0.33)	0.07(0.19)	0.18(0.28)

Tabla 7: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo con glaucoma, entrenamiento 1 vgg16

Glaucoma VGG16 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.01(0.04)	0.01(0.03)	0.03(0.08)	0.10(0.19)
Sector 2	0.01(0.04)	0.01(0.05)	0.03(0.08)	0.02(0.10)
Sector 3	0.01(0.05)	0.02(0.06)	0.01(0.04)	0.01(0.04)
Sector 4	0.01(0.03)	0.01(0.04)	0.01(0.04)	0.05(0.13)
Sector 5	0.00(0.02)	0.02(0.08)	0.06(0.12)	0.06(0.14)
Sector 6	0.01(0.07)	0.01(0.07)	0.06(0.14)	0.07(0.17)
Sector 7	0.93(0.18)	0.90(0.20)	0.78(0.29)	0.50(0.41)
Sector 8	0.02(0.12)	0.02(0.11)	0.02(0.08)	0.19(0.31)

Tabla 8: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo con glaucoma, entrenamiento 2 vgg16

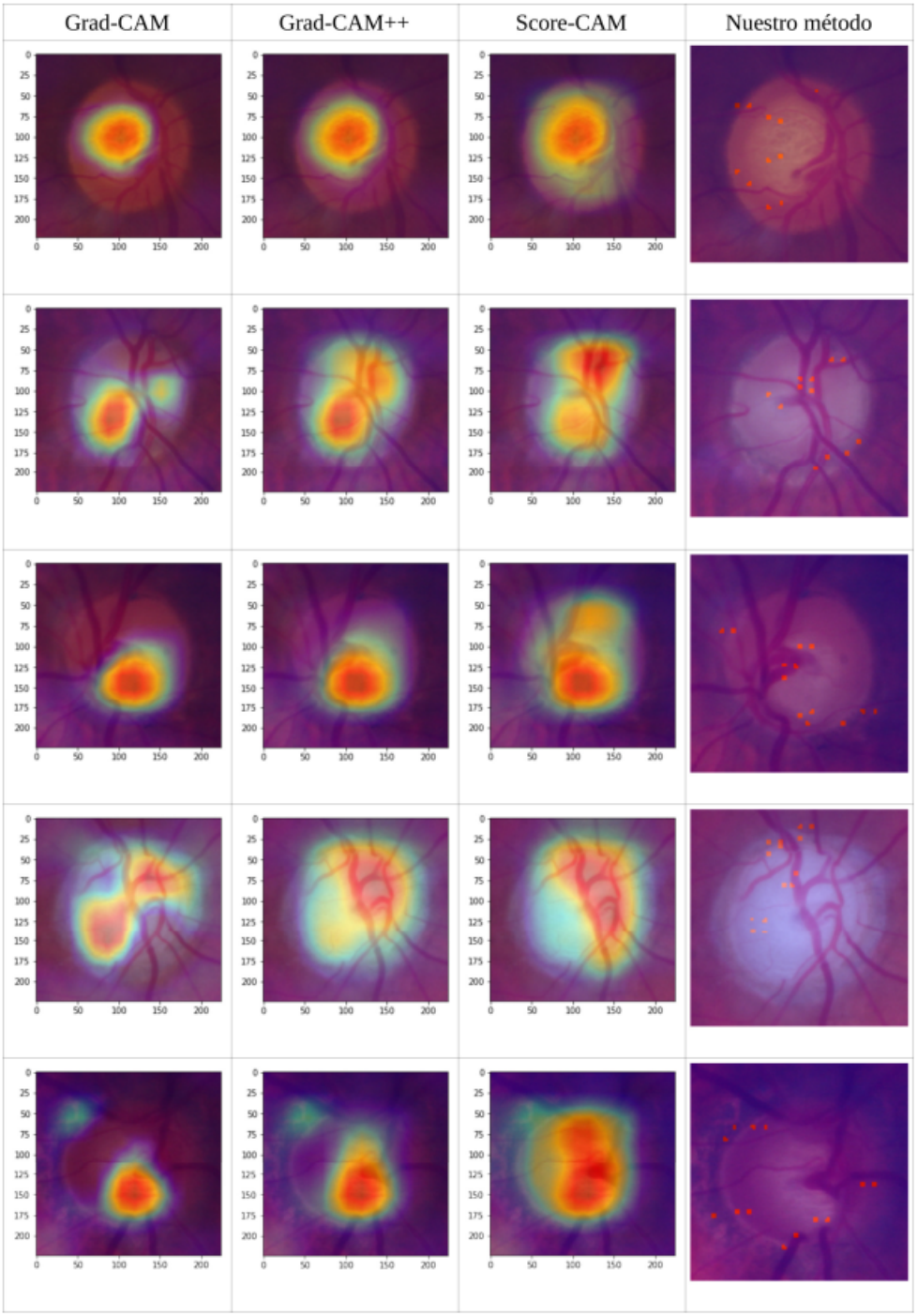

Figura 19: *Comparación de mapas de calor vgg16 sujetos glaucoma*

Caso 3: Red VGG19 evaluada sobre sujetos sanos

Normales VGG19 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	1.00(0.01)	1.00(0.01)	1.00(0.01)	0.89(0.12)
Copa	0.08(0.16)	0.09(0.17)	0.09(0.16)	0.14(0.15)
Rim	0.92(0.16)	0.91(0.17)	0.91(0.16)	0.74(0.18)
Fondo	0.00(0.01)	0.00(0.01)	0.00(0.01)	0.11(0.12)
Vasos	0.54(0.15)	0.53(0.15)	0.52(0.14)	0.40(0.15)

Tabla 9: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo sano, entrenamiento 1 vgg19

Normales VGG19 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	1.00(0.02)	1.00(0.01)	1.00(0.01)	0.92(0.13)
Copa	0.15(0.20)	0.22(0.23)	0.35(0.30)	0.23(0.21)
Rim	0.84(0.20)	0.77(0.23)	0.65(0.30)	0.68(0.21)
Fondo	0.00(0.02)	0.00(0.01)	0.00(0.01)	0.08(0.13)
Vasos	0.55(0.14)	0.54(0.14)	0.51(0.16)	0.36(0.12)

Tabla 10: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo sano, entrenamiento 2 vgg19

Normales VGG19 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.18(0.19)	0.17(0.18)	0.18(0.18)	0.29(0.21)
Sector 2	0.28(0.28)	0.29(0.27)	0.27(0.25)	0.21(0.20)
Sector 3	0.06(0.11)	0.07(0.12)	0.07(0.12)	0.10(0.12)
Sector 4	0.00(0.02)	0.00(0.02)	0.00(0.03)	0.05(0.10)
Sector 5	0.08(0.14)	0.09(0.14)	0.09(0.14)	0.03(0.05)
Sector 6	0.30(0.27)	0.28(0.26)	0.28(0.24)	0.05(0.09)
Sector 7	0.08(0.16)	0.09(0.17)	0.09(0.16)	0.14(0.15)
Sector 8	0.00(0.01)	0.01(0.01)	0.00(0.01)	0.11(0.12)

Tabla 11: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo sano, entrenamiento 1 vgg19

Normal VGG19 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.28(0.20)	0.28(0.19)	0.26(0.18)	0.29(0.17)
Sector 2	0.22(0.22)	0.19(0.18)	0.13(0.14)	0.11(0.14)
Sector 3	0.03(0.08)	0.03(0.07)	0.02(0.06)	0.05(0.08)
Sector 4	0.00(0.02)	0.00(0.02)	0.00(0.02)	0.04(0.07)
Sector 5	0.06(0.10)	0.05(0.09)	0.05(0.07)	0.07(0.09)
Sector 6	0.23(0.22)	0.20(0.19)	0.17(0.16)	0.11(0.11)
Sector 7	0.16(0.20)	0.22(0.23)	0.35(0.30)	0.23(0.21)
Sector 8	0.01(0.02)	0.00(0.01)	0.00(0.01)	0.08(0.13)

Tabla 12: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo sano, entrenamiento 2 vgg19

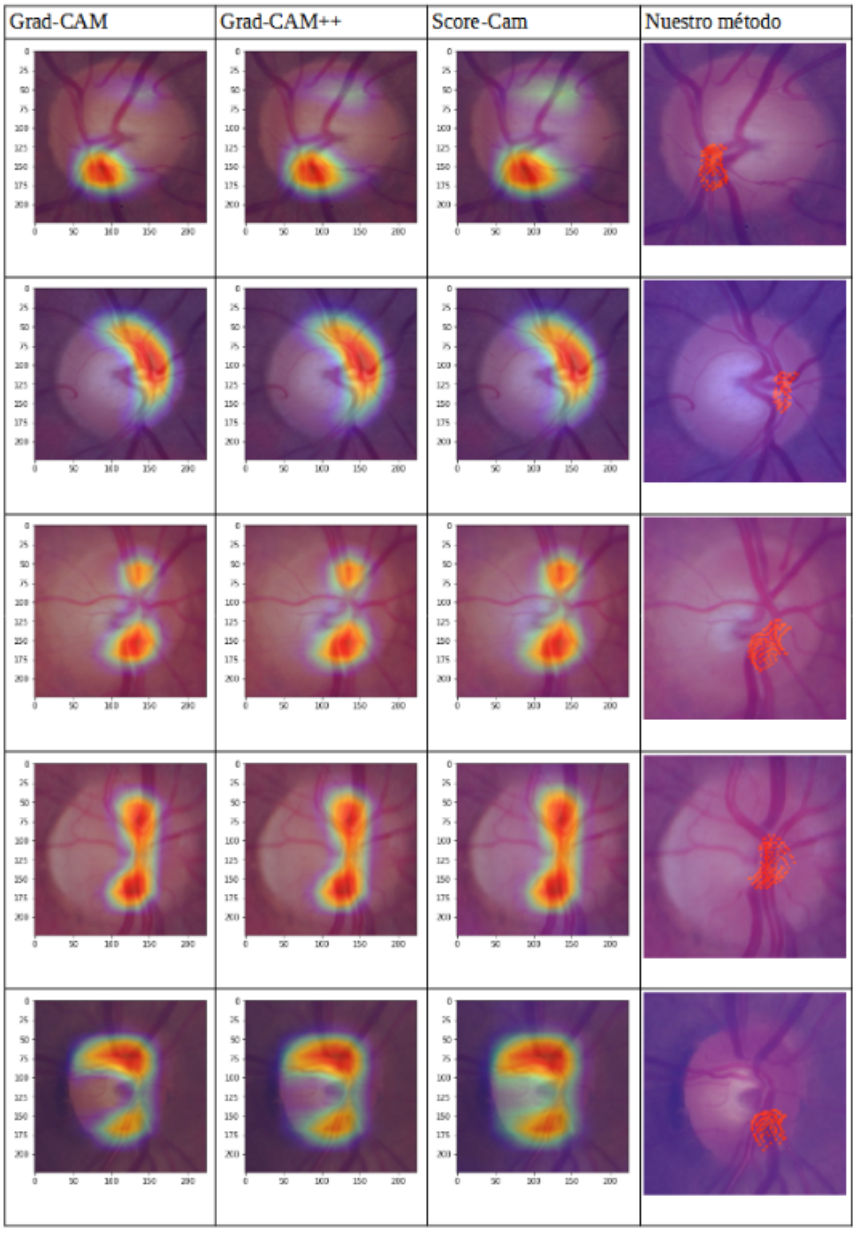

Figura 20: *Comparación de mapas de calor vgg19 sujetos sanos.*

Caso 4: Red VGG19 evaluada sobre sujetos con glaucoma

Glaucomas VGG19 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	0.72(0.41)	0.77(0.39)	0.91(0.26)	0.57(0.45)
Copa	0.68(0.41)	0.68(0.39)	0.49(0.40)	0.12(0.26)
Rim	0.04(0.10)	0.08(0.15)	0.41(0.38)	0.45(0.42)
Fondo	0.28(0.41)	0.23(0.39)	0.09(0.26)	0.43(0.45)
Vasos	0.16(0.15)	0.18(0.15)	0.34(0.20)	0.19(0.23)

Tabla 13: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo con glaucoma, entrenamiento 1 vgg19

Glaucomas VGG19 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Disco	0.72(0.41)	0.77(0.39)	0.91(0.26)	0.57(0.45)
Copa	0.68(0.41)	0.68(0.39)	0.49(0.40)	0.12(0.26)
Rim	0.04(0.10)	0.08(0.15)	0.41(0.38)	0.45(0.42)
Fondo	0.28(0.41)	0.23(0.39)	0.09(0.26)	0.43(0.45)
Vasos	0.16(0.15)	0.18(0.15)	0.34(0.20)	0.19(0.23)

Tabla 14: Solape medio y desviación típica de la salida de los diferentes métodos para las regiones anátomicas de ojo con glaucoma, entrenamiento 2 vgg19

Glaucoma VGG19 (1)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.00(0.02)	0.00(0.02)	0.04(0.12)	0.19(0.35)
Sector 2	0.01(0.05)	0.03(0.06)	0.15(0.26)	0.05(0.17)
Sector 3	0.01(0.05)	0.03(0.08)	0.05(0.13)	0.07(0.21)
Sector 4	0.00(0.02)	0.00(0.02)	0.00(0.02)	0.07(0.22)
Sector 5	0.00(0.01)	0.00(0.00)	0.05(0.15)	0.02(0.09)
Sector 6	0.00(0.02)	0.00(0.02)	0.10(0.21)	0.04(0.14)
Sector 7	0.68(0.41)	0.69(0.39)	0.49(0.40)	0.12(0.26)
Sector 8	0.28(0.41)	0.23(0.39)	0.09(0.26)	0.43(0.45)

Tabla 15: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo con glaucoma, entrenamiento 1 vgg19

Glaucoma VGG19 (2)	GradCAM	GradCAM++	ScoreCAM	Nuestro método
Sector 1	0.01(0.05)	0.02(0.05)	0.03(0.07)	0.20(0.35)
Sector 2	0.02(0.07)	0.02(0.07)	0.04(0.09)	0.07(0.21)
Sector 3	0.02(0.10)	0.02(0.07)	0.01(0.05)	0.03(0.14)
Sector 4	0.01(0.03)	0.00(0.02)	0.00(0.02)	0.02(0.13)
Sector 5	0.00(0.06)	0.00(0.03)	0.01(0.03)	0.01(0.09)
Sector 6	0.01(0.08)	0.01(0.06)	0.01(0.05)	0.05(0.20)
Sector 7	0.84(0.32)	0.88(0.25)	0.87(0.24)	0.47(0.46)
Sector 8	0.09(0.25)	0.04(0.16)	0.02(0.12)	0.14(0.32)

Tabla 16: Solape medio y desviación típica de la salida de los diferentes métodos para los sectores del ojo con glaucoma, entrenamiento 1 vgg19

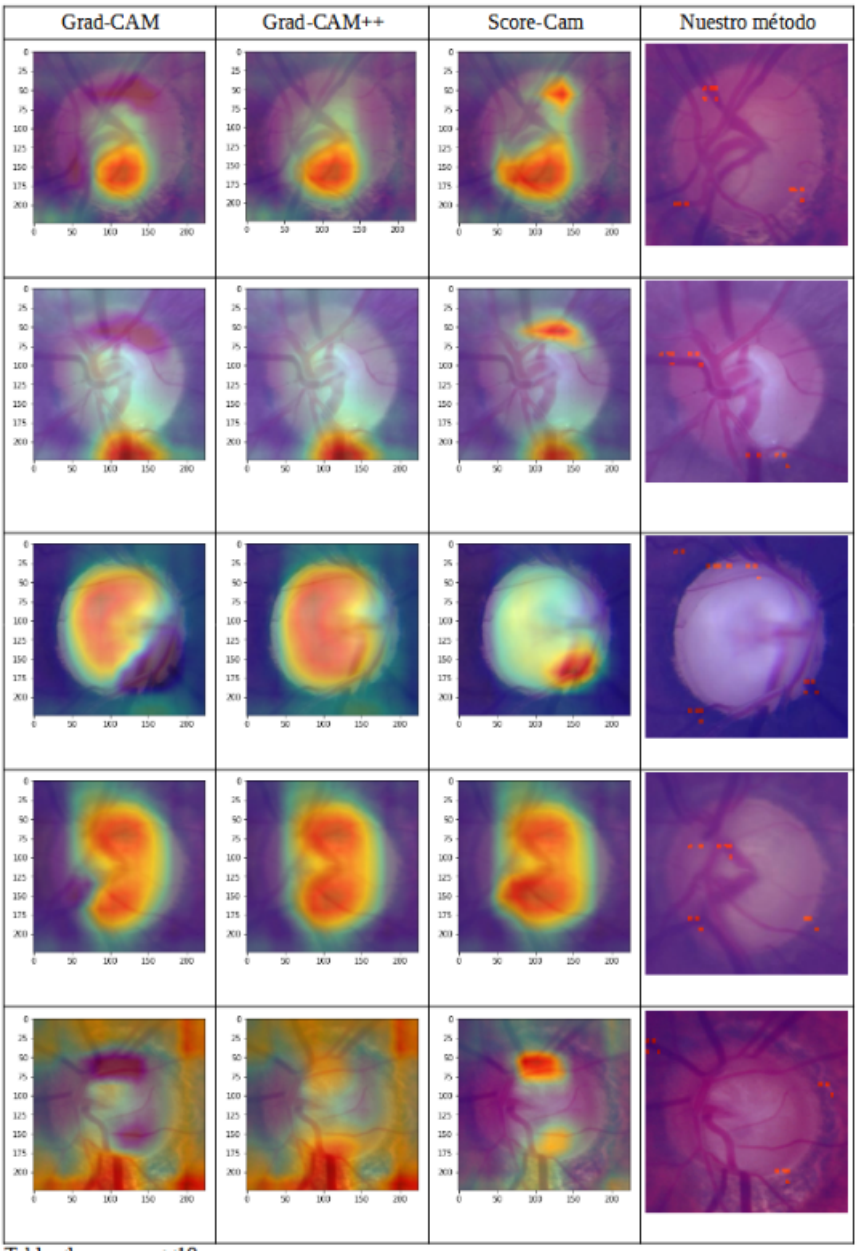

Figura 21: *Comparación de mapas de calor vgg19 sujetos glaucoma.*

3.3. DISCUSIÓN DE LOS RESULTADOS OBTENIDOS

Empezando con los mapas de calor, hay que decir que para los sujetos sanos se aprecia bastante similitud en las salidas de todos los métodos, si bien, como era de esperar, la salida de nuestro método es mucho más pequeña y ajustada, como es lógico, teniendo en cuenta que se trata de generar una región de tamaño mínimo relevante para la predicción. En los glaucomas es más difícil encontrar las similitudes. Incluso entre los métodos CAM hay una mayor discrepancia aunque la salida del GradCAM y GradCAM++ es siempre bastante parecida. Por lo que muestra el nuevo método, podemos decir que, en general, y en particular para los glaucomas, resulta muy sorprendente la poca información que parece necesitar la red para hacer su predicción, lo que, de confirmarse con más experimentos, llevaría a pensar que existe una gran diferencia entre lo que resulta fundamental para un especialista y el equivalente en una red de estas características.

En lo que se refiere al solape promedio con las diferentes regiones anatómicas de interés, lo que resulta más evidente en todos los métodos y para todas las redes es que la copa tiene mucho protagonismo en la decisión en el caso de los glaucomas y, por otro lado, la red parece fijarse más en los vasos en los sujetos sanos que en los glaucomas. Esto podría explicarse de un punto de vista médico teniendo en cuenta que el daño glaucomatoso empieza afectando a la zona de la copa y del rim, para acabar afectando también al paquete vascular por lo que si estuvieramos considerando glaucomas avanzados, posiblemente, el solape promedio con los vasos sería mayor. La concordancia entre los tres métodos CAM es bastante alta aunque, en algunos casos y como ya observamos en los mapas de calor, es el método ScoreCAM el que se diferencia un poco. En lo que respecta a nuestro método, hay que decir que, por motivos de coste computacional, los experimentos no se llevaron a cabo con las salidas del tipo de las que hemos mostrado en los mapas de calor de las figuras anteriores sino con una resolución bastante más baja, concretamente de 14*14. Por lo tanto, debemos tomar los resultados obtenidos con cautela y más bien como orientativos para ver la tendencia. En cualquier caso, volvemos a insistir en que el método presentado no se presta demasiado a una salida de tipo mapa de calor por lo que la comparativa resulta un poco forzada. En cuanto a la influencia del entrenamiento en sí, en el caso de los sujetos sanos no se aprecia demasiada diferencia entre el entrenamiento 1 y el entrenamiento 2, a diferencia de lo que ocurre con los glaucomas en los que sí se observan diferencias importantes en los solapamientos con algunas regiones.

Si miramos por sectores, los que suelen aportar mayor información a los especialistas son el 2 y el 3 (temporal inferior y nasal inferior, respectivamente), lo que no coincide demasiado con los resultados encontrados. Como era de esperar, existen diferencias importantes entre nuestro método y los métodos CAM. Como es lógico, la salida de los métodos CAM sigue siendo bastante similar con algunas discrepancias en el caso del ScoreCAM. En los glaucomas, para todas las redes, entrenamientos y métodos, queda claro que hay un sector que destaca por encima de cualquier otro y es la copa. También parece fijarse un poco en el fondo. En los normales, no es tan evidente la superioridad de un sector frente a los otros. El sector 6 (Nasal Superior), es el que parece recibir más atención en general, aunque también los sectores 1 (Nasal), 2 (Nasal Inferior) se llevan una parte importante del solape promedio. Los resultados de los dos entrenamientos se diferencian principalmente en el porcentaje de solape con la copa en los glaucomas, siendo bastante mayor en el caso del entrenamiento 2.

4. CONCLUSIONES Y LÍNEAS ABIERTAS

4.1. CONCLUSIONES

En este trabajo se ha planteado un doble objetivo. Por un lado, se han evaluado algunos de los métodos más populares de visualización de la salida de redes neuronales convolucionales sobre imágenes de fondo de ojo, utilizando redes ya entrenadas para el diagnóstico del glaucoma. Por otro lado, y en la misma línea, se ha presentado un nuevo método de visualización basado en la técnica dropout para desactivar las neuronas de los canales de la primera capa de la red en diferentes partes de la imagen y ver su efecto sobre la probabilidad asignada a una determinada clase para una imagen de entrada dada.

Las pruebas llevadas a cabo nos permiten concluir que el método presentado parece localizar de una manera más precisa las zonas de tamaño mínimo relevantes para la predicción de la red, y además su salida es fácilmente interpretable en términos de las probabilidades que maneja la red. En lo que se refiere a los otros tres métodos, se ha observado un grado de acuerdo bastante alto entre ellos, con alguna diferencia en algunos casos con el método ScoreCAM. Por otro lado, no parece haber una concordancia demasiado grande en los sectores del disco, usualmente, más relevantes para el médico y lo que la red observa aunque, como se comentó en el capítulo anterior, sería necesario hacer más experimentos y una mejor evaluación estadística de los resultados.

4.2. LÍNEAS ABIERTAS

En este trabajo quedan muchas más líneas abiertas que las que están cerradas. Algunas de ellas podrían ser:

- Formalizar matemáticamente el método presentado y hacer un análisis estadístico de evaluación más riguroso.
- Utilizar otro tipo de redes más sofisticadas y actuales que las VGG16 y VGG19.
- Utilizar otros conjuntos de datos diferentes al RIM-ONE DL.
- Evaluar otros métodos de visualización que no sean del tipo CAM.
- Probar con imágenes de fondo de ojo de infrarrojo que también están siendo utilizadas en el grupo de investigación.
- Comprobar si el método presentado resulta igual de eficaz para otro tipo de imágenes y no solo médicas.

4.3. CONCLUSIONS AND FUTURE WORK

In this work a double objective has been set. On the one hand, some of the most popular methods of visualizing the output of convolutional neural networks on fundus eye images have been evaluated, using networks already trained for the diagnosis of glaucoma. On the other hand, and along the same lines, a new visualization method based on the dropout technique has been presented to deactivate the neurons of the channels of the first layer of the network in different parts of the image and see its effect on the probability assigned to a certain class for a given input image.

The tests carried out allow us to conclude that the presented method seems to locate in a more precise way in the areas of minimum size relevant for the prediction of the network, and also its output is easily interpretable in terms of the probabilities that the network handles. Regarding the other three methods, a fairly high degree of agreement has been observed between them, with some difference in some cases with the ScoreCAM method. On the other hand, there doesn't seem to be too great a concordance in the sectors of the disk, usually more relevant to the doctor and what

the network observes, although, as commented in the previous chapter, it would be necessary to do more experiments and a better statistical evaluation from the results data.

many more lines remain open in this work than are closed. Some of them could be:

- Mathematically formalize the presented method and make a more rigorous evaluation statistical analysis.
- Use other types of more sophisticated and current networks than VGG16 and VGG19.
- Use data sets other than the RIM-ONE DL.
- Evaluate other non-CAM visualization methods.
- Try infrared fundus eye images that are also being used in the research group.
- Check if the presented method is equally effective for other types of images and not just medical ones.

Referencias

[1] Benyamin Ghojogh, Maria N. Samad, Sayema Asif Mashhadi, Tania Kapoor, Wahab Ali, Fakhri Karray, Mark Crowle. *Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review*, https://arxiv.org/pdf/1905.02845.pdf, 2019.

[2] Ron Kohavi, George H. John. *Wrappers for feature subset selection*, Artificial Intelligence, 1997.

[3] Miguel Padrón González. *SELECCIÓN DE CANALES EN REDES NEURONALES CONVOLUCIONALES MEDIANTE DROPOUT.*, Memoria de Trabajo Fin de Grado, 2020.

[4] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. *Dropout: A Simple Way to Prevent Neural Networks from Overtting.*, Journal of Machine Learning Research, 2014.

[5] Ron Kohavi, George H. John. *Very deep convolutional networks for large-scale image recognition.*, arXiv preprint arXiv:1409.1556, 2014

[6] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. *Striving for simplicity: The all convolutional net.*, ICLR, 2015.

[7] *Noise-adding Methods of Saliency Map as Series of Higher Order Partial Derivative*, IICML, 2018.

[8] M. D. Zeiler and R. Fergus. *Visualizing and understanding convolutional networks.*, In European conference on computer vision, pages 818–833. Springer, 2014.

[9] Vitali Petsiuk, Abir Das, and Kate Saenko. *Rise: Randomized input sampling for explanation of black-box models.*,In Proc. BMVC, 2018.

[10] Ruth Fong, Mandela Patrick, Andrea Vedaldi. *Understanding Deep Networks via Extremal Perturbations and Smooth Masks.*, ICCV, 2019.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A.Torralba. *earning deep features for discriminative localization.*, In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. *Grad-cam: Visual explanations from deep networks via gradient-based localization.* , In Proceedings of the IEEE International Conference on Computer Vision, 2017.

[13] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. *Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks.* , In 2018 IEEE Winter Conference on Applications of Computer Vision.

[14] Haofan Wang and Zifan Wang and Mengnan Du and Fan Yang and Zijian Zhang and Sirui Ding and Piotr Mardziel and Xia Hu. *Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks.* , CVPR, 2020.