



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



ESCUELA TÉCNICA
SUPERIOR INGENIERÍA
INDUSTRIAL VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA



DISEÑO Y DESARROLLO DE UN SISTEMA DE CLASIFICACIÓN Y DETECCIÓN DE PÓLIPOS EN IMÁGENES ENDOSCÓPICAS CON TÉCNICAS DE APRENDIZAJE PROFUNDO

AUTOR: PABLO MESEGUER ESBRÍ

TUTOR: VALERY NARANJO OLMEDO

COTUTOR:  JOSE NÉSTOR JIMÉNEZ CAMPFENS

Curso Académico: 2020-21 

Agradecimientos

A mi familia por estar siempre a mi lado apoyándome, a Néstor por la ayuda impagable durante todo el trabajo, a Valery por la implicación constante a lo largo del proyecto y a mis amigos por hacer de la carrera una experiencia mejor.

Resumen

Resumen

El cáncer de colón es una de las patologías con mayor prevalencia en la población mundial y se caracteriza por iniciarse con la formación de pólipos en el intestino grueso. Varios estudios han demostrado que la extirpación de los pólipos en sus fases más tempranas reduce la mortalidad asociada al cáncer de colón. Por ello, resulta fundamental que los profesionales médicos que llevan a cabo las exploraciones endoscópicas no pasen por alto ningún pólipo independientemente de su forma y tamaño. Por lo tanto, el diseño y la implantación de sistemas automáticos de apoyo a los endoscopistas tiene como objetivo aumentar su atención reduciendo así la tasa de pólipos no observados.

Estos sistemas automáticos pueden estar basados en aprendizaje profundo y esconden un gran potencial puesto que pueden llegar a superar a las facultades humanas en algunas tareas debido a su capacidad para analizar grandes cantidades de información. Por ello, durante el desarrollo del proyecto se han explorado diferentes modelos de red neuronal basados en aprendizaje profundo. Las redes neuronales convolucionales que se implementan están basadas en aprendizaje de transferencia a partir de redes neuronales previamente entrenadas en otros conjuntos de datos o entrenadas desde cero si se opta por diseñar la arquitectura de la red. Estos modelos se diferencian en función de la tarea que llevan a cabo entre las que se diferencian la clasificación de imágenes endoscópicas, la localización de objetos mediante sus rectángulos mínimos y la segmentación semántica.

Palabras Clave: Cáncer de colon, imágenes endoscópicas, aprendizaje profundo, aprendizaje de transferencia, clasificación, detección de objetos, segmentación semántica.

Resum

El càncer de còlon és una de les patologies amb major prevalença en la població mundial i es caracteritza per iniciar-se amb la formació de pòlips en l'intestí gros. Alguns estudis han demostrat que l'extirpació dels pòlips en les seues fases més inicials redueix la mortalitat associada al càncer de còlon. Així, resulta fonamental que els professionals mèdics que porten a terme les exploracions endoscòpiques no passen per alt cap pòlip independentment de la seua forma i grandària. Per això, el disseny i la implementació de sistemes automàtics de suport als professionals té com a objectiu augmentar la seua atenció reduint així el percentatge de pòlips no observats.

Aquests sistemes automàtics poden estar basats en aprenentatge profund i amaguen un gran potencial perquè poden arribar a superar les facultats humanes en algunes tasques gràcies a la seua capacitat d'analitzar enormes quantitats d'informació. Així, durant el desenvolupament del projecte s'han explorat diferents models de xarxa neuronal basats en aprenentatge profund. Les xarxes neuronals que s'han implementat estan basades en aprenentatge de transferència a partir de xarxes prèviament entrenades en altres conjunts de dades o han estat entrenades des de zero si s'ha optat per dissenyar l'arquitectura de la xarxa. Aquests models es diferencien en funció de la tasca que han de realitzar entre les que es diferencien la classificació d'imatges endoscòpiques, la localització de pòlips mitjançant els seus rectangles mínims i la segmentació semàntica.

Paraules claus: Càncer de còlon, imatges endoscòpiques, aprenentatge profund, aprenentatge de transferència, classificació, detecció d'objectes, segmentació semàntica.

Abstract

Colon cancer is one of the most prevalent pathologies in the world and is characterized by starting with the formation of polyps. Several studies have shown that removing polyps in their earliest stages reduces mortality associated with colon cancer. Therefore, it is essential that medical professionals who perform endoscopic examinations do not miss any polyp regardless of its shape and size. The design and implementation of automated decision support systems for endoscopists aims to increase their attention, reducing the polyp miss rate.

These automatic systems can be based on deep learning and they hide great potential because they can surpass human ability in some tasks due to their ability to analyze huge amounts of information. For this reason, during the development of the project, different models of neural network based on deep learning have been explored. The convolutional neural networks that have been implemented are based on transfer learning from neural networks previously trained in other datasets or trained from scratch if we choose to design the network architecture. These models differ according to the task they carry out, including the classification of endoscopic images, the location of objects through their minimum rectangles and semantic segmentation.

Keywords: Colon cancer, endoscopic imaging, Deep Learning, Transfer Learning, image classification, object detection, semantic segmentation.

Índice general

Resumen	III
Índice general	VII
I Memoria	1
1 Introducción	3
1.1 Contexto médico	3
1.2 <i>Machine Learning</i>	6
1.3 Detección de pólipos	11
1.4 Estado del arte	12
1.5 Objetivo del proyecto	14
2 Materiales y métodos	15
2.1 Material	15
2.2 Métodos	17
2.3 Diagrama de trabajo	25
3 Resultados	27
3.1 Clasificación	27
3.2 Detección	29
3.3 Segmentación	30
4 Discusión	31
4.1 Tarea de clasificación	31
4.2 Tarea de detección	33
4.3 Tarea de segmentación	34

5 Conclusión	37
Bibliografía	39
II Presupuesto	43
5.1 Presupuestos parciales	45
5.2 Presupuestos totales	46
Índice alfabético	47

Índice de figuras

1.1. Anatomía del colón y del recto (obtenido de Society, 2020b)	4
1.2. Diferencia entre pólipos planos y pediculados	4
1.3. Técnica del gradiente descendiente (obtenido de Torres, 2018)	9
1.4. Estructura básica de una Red Neuronal Convolutiva (obtenido de Stewart, 2019)	10
1.5. Esquema del aprendizaje de transferencia (obtenido de Pan y Yang, 2009)	10
1.6. Resumen de las técnicas de segmentación (obtenido de Lamba, 2019)	12
2.1. Separación entre imágenes normales y con pólipos	16
2.2. Ejemplo de representación de los rectángulos mínimos obtenidos	17
2.3. Resultado del recorte de la máscara negra	18
2.4. Resultado del recorte de la máscara negra	18
2.5. Resultado de la restauración de la imagen	19
2.6. Esquema de la arquitectura de Alexnet (obtenido de Khvostikov y col., 2018) . .	20
2.7. Esquema de la arquitectura de VGG16 (obtenido de Nash y col., 2018)	21
2.8. Diagrama de bloques de la arquitectura de ResNet-50 (obtenido de Talo, 2019) .	21
2.9. Esquema de la operación de <i>unpooling</i> (obtenido de Zafar y col., 2018)	23
2.10. Esquema de la arquitectura de U-Net (obtenido de Ronneberger y col., 2015) . .	24
2.11. Resumen del diagrama de trabajo	26
3.1. Resultados de la tarea de detección con <i>Transfer Learning</i>	29
4.1. Heat maps de los modelos de clasificación	32
4.2. Heat maps de los modelos de clasificación	33
4.3. Resultado de la red de detección	33

4.4. Resultado de la red de segmentación	34
4.5. Resultado del método de postprocesado	35

Índice de tablas

1.1. Estadísticas de supervivencia en cáncer colorrectal	5
1.2. Funciones de activación	8
1.3. Análisis de la clasificación de imágenes con pólipos	11
2.1. Resumen de las bases de datos	16
3.1. Resultados de la tarea de clasificación aplicando <i>Transfer Learning</i>	28
3.2. Resultados de la tarea de clasificación aplicando <i>Transfer Learning</i>	28
3.3. Resultados de la tarea de segmentación	30
3.4. Resultados del método de postprocesado	30
5.1. Resumen de los costes de mano de obra	45
5.2. Resumen de los costes de software	45
5.3. Resumen de los costes de hardware	46
5.4. Resumen de los costes totales del proyecto	46

Parte I

Memoria

Capítulo 1

Introducción

En este capítulo se realiza una introducción del tema del trabajo centrandó la atención en el contexto médico, profundizando en el concepto de aprendizaje automático o *Machine Learning* y presentando el estado del arte y el objetivo del proyecto.

1.1 Contexto médico

1.1.1 Anatomía

El colon y el recto son las dos estructuras anatómicas que constituyen el intestino grueso que, a su vez, forma parte del tracto gastrointestinal. El colon es un tubo muscular de aproximadamente 1.5 metros de longitud, que se conecta por un extremo con el intestino delgado y por otro con el ano, cuya función principal consiste en continuar la absorción de agua y de nutrientes iniciada en el intestino delgado y en almacenar las heces.

El colon se puede dividir en cuatro secciones según la dirección del movimiento del bolo digestivo. La primera sección se denomina colon ascendente y empieza en el ciego, punto en el que conecta con el intestino delgado; la segunda es el colon transversal que va seguido por la sección descendente y la última se denomina colon sigmoide debido a la forma en S que lo caracteriza. Por su parte, el recto forma la sección final del tubo digestivo, conecta con el ano y tiene una longitud cercana a los 15 centímetros. Está formado por músculos esfínteres que tienen un papel clave en la continencia fecal. La anatomía del tracto digestivo se puede visualizar en la Figura 1.1 junto con otras estructuras anatómicas de la región abdominal.

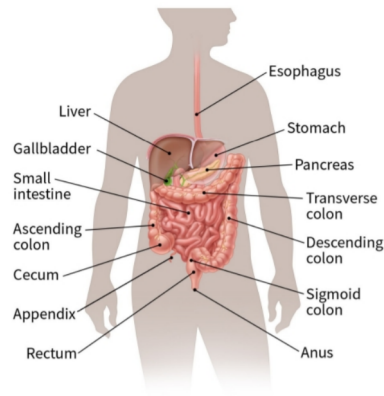


Figura 1.1: Anatomía del colon y del recto (obtenido de Society, 2020b)

1.1.2 *Cáncer colorrectal*

Según la *American Cancer Society*, el cáncer incluye el conjunto de patologías que se caracterizan por un crecimiento anormal y acelerado de células que escapan del control de los mecanismos normales de regulación celular y que tienen la capacidad de destruir el tejido sano. Por lo tanto, el cáncer colorrectal es aquel que se inicia en las células del colon o del recto con la formación de pólipos (Society, 2020b). Los pólipos están constituidos por una masa celular formada debido al crecimiento anómalo de las células que forman el revestimiento interno del intestino grueso. Estos se observan como unas protuberancias sobre la pared interna del tubo digestivo cuya apariencia diferencia dos tipos principales de pólipos como son los planos y los pedunculados, tal y como se puede observar en la Figura 1.2.

Con el paso del tiempo y en función de la exposición a factores de riesgo como las bebidas alcohólicas, los pólipos pueden provocar la aparición de un cáncer con mayor o menor probabilidad dependiendo de la anatomía de los mismos. Los pólipos hiperplásicos e inflamatorios son benignos por lo que no se consideran precancerosos. En cambio, los adenomas o pólipos adenomatosos sí se consideran precursores de cáncer una vez se convierten en adenocarcinomas. Del mismo modo, los pólipos sésiles (SSP del inglés *Sessile Serrated Polyps*) y serrados (TSA del inglés *Traditional Serrated Adenomas*) tienen una mayor probabilidad de provocar la propagación de un cáncer por lo que se manejan de forma similar a los adenomas. (Society, 2020b)

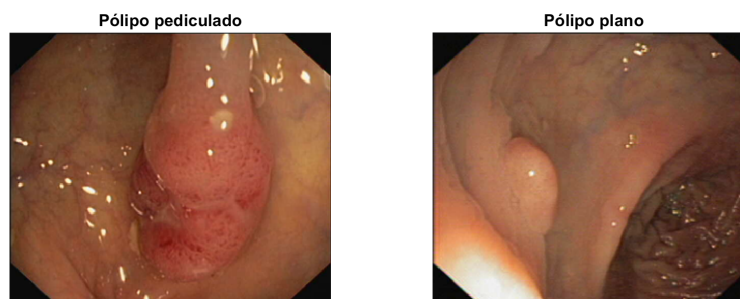


Figura 1.2: Diferencia entre pólipos planos y pediculados

Excluyendo el cáncer de piel, el cáncer colorrectal es el tercer tipo de cáncer más diagnosticado y letal tanto en hombres como en mujeres en Estados Unidos. La *American Cancer Society* realizó una estimación de unos 150 mil casos nuevos a diagnosticar de cáncer de colon en los Estados Unidos para el año 2021. La tasa de pacientes con cáncer colorrectal ha ido disminuyendo en los últimos 40 años debido al aumento de las técnicas de *screening* realizadas y a los cambios en el estilo de vida, consistentes principalmente en evitar factores de riesgo como dietas no saludables y la ingesta desproporcionada de bebidas alcohólicas. Al mismo tiempo, esta tasa ha disminuido en personas mayores de 55 años pero aumentado en menores de 55 gracias a la detección temprana que es fundamental para prevenir que el cáncer se desarrolle en sus etapas más avanzadas. Resumiendo, la probabilidad de desarrollar cáncer colorrectal es del 4.3% en hombres y del 4% en mujeres. (Society, 2020b)

La tasa de supervivencia de los pacientes con cáncer colorrectal depende de la diseminación que haya alcanzado la metástasis. La tasa de supervivencia a 5 años se entiende como el porcentaje de pacientes con una determinada patología que sigue con vida tras superar los 5 años de la detección de la misma. Cuando el cáncer colorrectal se encuentra localizado, esta tasa de supervivencia alcanza el 90%, sin embargo, cuando el cáncer es regional o diseminado este porcentaje se reduce hasta 71% y 14%, respectivamente. De este modo, la tasa de supervivencia a 5 años para todos los tipos de cáncer combinados es del 63%, tal y como se presenta en la Tabla 1.1 a partir de los datos de la *American Cancer Society*. (Society, 2020a)

Etapa	Tasa supervivencia a 5 años
Localizado	91 %
Regional	72 %
Diseminado	14 %
Todas las etapas	63 %

Tabla 1.1: Estadísticas de supervivencia en cáncer colorrectal

Una característica que resulta de interés acerca de los pólipos gastrointestinales es que mayoritariamente se inician en la capa más interna del tubo digestivo. La metástasis o propagación del cáncer se producirá cuando las células cancerosas se expandan a las capas más externas y seguidamente a otras regiones del organismo a través de los vasos sanguíneos o linfáticos. Al encontrarse los pólipos en la mucosa interna, es posible realizar una observación directa de los mismos mediante técnicas endoscópicas como la colonoscopia.

1.1.3 *Polipectomías colonoscópicas*

La colonoscopia es una técnica mayoritariamente exploratoria del tracto intestinal que consiste en introducir un tubo largo y flexible denominado colonoscopio a través del ano con el objetivo de observar el interior del intestino grueso y el recto. El colonoscopio lleva adherido en su extremo una cámara conectada a un sistema de adquisición de imagen que permite la visualización en tiempo real del interior del colon. Otras técnicas de visualización del tracto gastrointestinal son la gastroscopias y las cápsulas endoscópicas. La endoscopia gastrointestinal alta o gastroscopia es un técnica similar a las colonoscopias puesto que también emplean un endoscopio, pero se diferencian en que este se introduce por la boca posibilitando así la visualización del esófago, el estómago y el duodeno. Por su parte, la cápsula endoscópica consiste en un dispositivo de

pequeño tamaño que contiene una cámara inalámbrica que es capaz de tomar una gran cantidad de imágenes durante su trayecto por todo el tracto intestinal. Se trata de una técnica meramente exploratoria.

Aunque la colonoscopia sea principalmente una técnica exploratoria y de *screening*, si se detecta algún pólipo durante la inspección del colon, suele realizarse una polipectomía. Las polipectomías consisten en la extirpación de uno o más pólipos durante las exploraciones mediante una asa de polipectomía que está formada por un alambre metálico capaz de extirpar y extraer el pólipo. Las polipectomías son una medida eficaz para reducir la mortalidad asociada al cáncer de colon (Zauber y col., 2012). Según este estudio, la mortalidad por cáncer de colon de los pacientes que se habían realizado como mínimo una polipectomía se redujo en un 53% en comparación con la tasa esperada para el resto de la población. Además, también se ha comprobado que no existe una diferencia significativa en la tasa de mortalidad entre los pacientes con pólipos adenomatosos o hiperplásicos verificando así la eficacia tanto de las técnicas de *screening* como de las polipectomías.

1.2 *Machine Learning*

1.2.1 *Concepto*

Atendiendo a la definición de Inteligencia Artificial ofrecida por Luis Amador (Amador Hidalgo, 1996), esta se entiende como:

“La creación de entes o sistemas automáticos que sean capaces de llevar a cabo tareas y funciones que han estado, hasta el momento, reservadas en su desempeño exclusivamente para seres humanos. (...) La Inteligencia Artificial se enmarca dentro de un contexto, más tecnológico, donde sea posible diseñar y construir programas, máquinas, etc..., con aptitudes similares o superiores a las del ser humano.”

Dentro del campo de la Inteligencia Artificial se encuentra el aprendizaje automático o *Machine Learning*. Este se caracteriza por la construcción de modelos matemáticos a partir de muestras de datos con el objetivo de realizar predicciones o tomar decisiones sin estar estas explícitamente programadas para realizar esta tarea (Zhang, 2020). Una característica fundamental de los algoritmos de *Machine Learning* es que deben ser capaces de aprender automáticamente y de adaptarse a los cambios que puedan aparecer en los datos de entrada.

1.2.2 *Clasificación*

Los modelos automáticos basados en *Machine Learning* pueden clasificarse en cuatro tipos en función de si dispone o no de la salida o *Ground Truth* de cada uno de los datos de entrada como son el aprendizaje supervisado, no supervisado, semi-supervisado y por refuerzo.

Aprendizaje supervisado

Se entienden como algoritmos de aprendizaje supervisados aquellos que se crean a partir de datos de entrenamiento etiquetados y en los que el propio algoritmo debe de aprender una serie de reglas para predecir su salida. Su objetivo es generalizar, es decir, deben de ofrecer la clasificación correcta para unos nuevos datos de entrada en los que se desconoce su etiqueta. Este tipo de algoritmos necesitan conocer el *Ground Truth* de los datos que se entiende como el valor verdadero de la salida de los mismos. Esta puede ir desde un etiqueta para los modelos de clasificación hasta una máscara binaria en los modelos de segmentación. En este campo destacan algunos algoritmos como redes neuronales convolucionales y los árboles de decisión.

Aprendizaje no supervisado

En contraposición con el aprendizaje supervisado, los algoritmos de aprendizaje no supervisado se entrenan con datos cuya salida es desconocida. Por lo tanto, el algoritmo debe de encontrar patrones y estructuras en los datos no etiquetados. Este tipo de aprendizaje se emplean para realizar *clustering* con algoritmos como *k-means*.

Aprendizaje semi-supervisado

El aprendizaje semi-supervisado se caracteriza por combinar tanto datos de entrenamiento etiquetados como no etiquetados. Habitualmente, se suele disponer de una mayor cantidad de datos sin etiquetar al ser estos más sencillos de obtener.

Aprendizaje por refuerzo

El aprendizaje por refuerzo o *Reinforcement Learning* se entiende como el problema en el que un agente debe de aprender qué acciones tomar a través de interacciones prueba-error con un entorno que es dinámico y cambiante. Se caracteriza por programar un agente para realizar una cierta tarea que puede alcanzarse de diferentes maneras en función de las recompensas positivas o negativas de cada una de las acciones necesarias. (Kaelbling y col., 1996)

1.2.3 Deep Learning

Según el experto en la materia Jordi Torres (Torres, 2018), el aprendizaje profundo o *Deep Learning* incluye:

“Las estructuras algorítmicas que permiten modelos que están compuestos de múltiples capas de procesamiento para aprender representaciones de datos, con múltiples niveles de abstracción que realizan una serie de transformaciones lineales y no lineales que a partir de los datos de entrada generen una salida próxima a la esperada. (...) Consiste en obtener los parámetros de esas transformaciones (los pesos y el sesgo), y consigue que esas transformaciones sean óptimas, es decir, que la salida producida y la esperada difieran muy poco.”

Perceptrón multicapa (MLP)

Los perceptrones multicapa o MLPs (del inglés *Multilayer Perceptron*) forman redes neuronales simples que contienen múltiples capas ocultas que a su vez incluyen neuronas completamente conectadas. Estas neuronas se encargan de multiplicar los datos de entrada en función de unos pesos y de pasar la información resultante a la siguiente capa dependiendo de la función de activación de la misma, que se corresponde con su salida. Una vez obtenida la predicción del modelo, se calcula el error de la misma comparando la predicción con el valor real de la etiqueta. El gradiente de la función de coste se emplea para actualizar de manera iterativa los pesos de las neuronas con el objetivo de minimizar el error de la salida.

Cuando los MLPs se emplean para clasificación de imágenes, la función de cada neurona de salida consiste en una exponencial normalizada (*softmax*) puesto que ofrece la probabilidad estimada de que los datos pertenezcan a una determinada clase tal y como se presenta en la Tabla 1.2. El principal inconveniente que presentan los MLPs es que ignoran la información espacial puesto que no consideran el entorno de cada píxel, información que sí que tienen en cuenta las redes neuronales convolucionales gracias a las operaciones de convolución.

Funciones de activación

Las funciones de activación se emplean para propagar la información de salida de cada neurona hacia las siguientes capas y tienen el objetivo de introducir no linealidad en el modelado de la red. En la Tabla 1.2, se presentan algunas de las funciones de activación no-lineales más utilizadas como son la función sigmoide, *Hyperbolic Tangent (tanh)* y la Unidad Linear Rectificada (ReLU, del inglés *Rectified Linear Unit*). ReLU es la función de activación más utilizada en aprendizaje profundo y se caracteriza por poner todos los inputs negativos a cero y mantener linealmente los positivos. (Nair y Hinton, 2010)

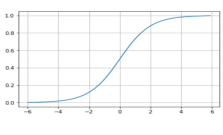
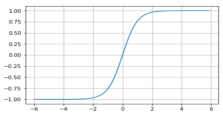
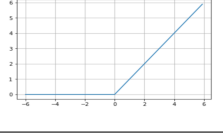
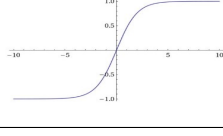
Sigmoid: $f(x) = \frac{1}{1+e^{-x}}$	
tanh: $f(x) = \tanh(x)$	
ReLU: $f(x) = \max(0, x)$	
Softmax: $f(x) = \frac{e^{x_i}}{\sum_{k=1}^K e^{x_k}}$	

Tabla 1.2: Funciones de activación

El proceso de entrenamiento

El proceso de entrenamiento de una red neuronal consta de dos etapas denominadas *forward propagation* y *backpropagation*. La primera de ellas consiste en hacer pasar los datos de entrenamiento a través de toda la red neuronal aplicando las funciones de activación anteriormente mencionadas con el objetivo de obtener la predicción final y calcular el error respecto al valor de la etiqueta de los datos. El error se calcula mediante funciones de coste como la entropía cruzada e indica como de precisa es la predicción que ha realizado el modelo diseñado, por ello interesa que sea nulo.

La retropropagación o *backpropagation* consiste en propagar el error en sentido contrario al de la red neuronal, es decir, desde la salida hasta la entrada de la misma. Este proceso sirve para que las neuronas, en función de la tasa de entrenamiento, ajusten los pesos de sus operaciones con el objetivo de minimizar el error en la predicción. Una técnica para realizar este ajuste es la del gradiente descendiente que consiste en realizar el ajuste de los pesos de manera iterativa en función del sentido del gradiente de la función de coste, tal y como se observa en la Figura 1.3.

Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales o CNNs (del inglés *Convolutional Neural Networks* solucionan el problema que presentan los MLP, es decir, no tratan cada input de entrada de manera independiente sino que analizan el entorno espacial de cada uno de ellos. Esto resulta de especial utilidad cuando se trabaja con imágenes puesto que resulta intuitivo pensar que existe una conexión espacial entre sus píxeles. Las Convolutional Neural Networks emplean filtros convolucionales que analizan una región cercana para cada píxel. La operación de convolución desplaza el filtro kernel por toda la imagen obteniendo el producto elemento por elemento de la matriz de la imagen con la del filtro.

Tal como se observa en la Figura 1.4, además de capas de convolución, las CNN incluyen capas de *pooling* (mayoritariamente *maxpooling*) con el objetivo de seleccionar los valores máximos de los mapas de características y de reducir la dimensionalidad de los mismos. Estos serán empleados como entradas de las siguientes capas dependiendo de su función de activación (mayoritariamente ReLU). Finalmente, son las capas densas de neuronas completamente conectadas las que se encargan de la separación entre clases y de la clasificación. Las CNN se emplean habitualmente para la clasificación de imágenes.

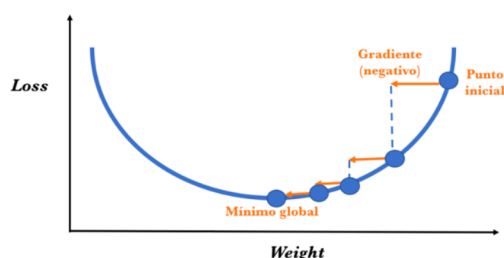


Figura 1.3: Técnica del gradiente descendiente (obtenido de Torres, 2018)

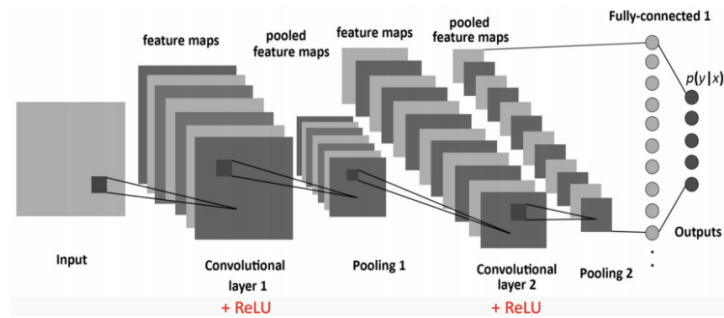


Figura 1.4: Estructura básica de una Red Neuronal Convolutiva (obtenido de Stewart, 2019)

1.2.4 Aprendizaje de transferencia

El aprendizaje de transferencia o *Transfer Learning* se caracteriza por transferir el conocimiento adquirido en un cierto campo o dominio fuente en el que se dispone de mucha información a otro dominio objetivo en el que la cantidad de datos disponibles es considerablemente menor. Es decir, consiste en aplicar conocimiento aprendido con anterioridad para resolver nuevos problemas de una manera más rápida y eficaz. Por su parte, el *Machine Learning* tradicional lo que pretende es diseñar desde cero un nuevo sistema de aprendizaje para cada una de las diferentes tareas que se requieran, es decir, realizar el aprendizaje sin considerar el conocimiento previo acerca de otras tareas similares (Pan y Yang, 2009). El principal objetivo que persigue el *Transfer Learning* es facilitar la resolución de problemas muy complejos que requieran una gran cantidad de información que no esté disponible en ese dominio pero sí en otros en los que pueda resultar útil. En la Figura 1.5, se presenta una comparativa esquemática del aprendizaje automático tradicional (izquierda) con el aprendizaje de transferencia (derecha).

La transferencia de aprendizaje se puede aplicar a modelos de aprendizaje profundo formados por su estructura característica de capas. Existen dos posibles aproximaciones como son la extracción de características y el *Fine-Tuning*. Por su parte, la *Feature Extraction* se caracteriza por congelar todas las capas del modelo excepto la última capa completamente conectada, es decir, no se actualizan los pesos de estas capas durante el entrenamiento del modelo con los nuevos datos. Así, se consiguen extraer características de las imágenes que deberían de ser invariantes espacial, rotacional y translacionalmente. Por otra parte, el *Fine-Tuning* consiste en reentrenar parte de la red con los nuevos datos de la tarea objetivo. Esto es posible puesto que las capas de convolución iniciales se caracterizan por capturar las características más generales como podrían ser los bordes de los objetos y los cambios de intensidad. En cambio, las últimas capas se centran en la tarea más específica que se está llevando a cabo. (Sarkar, 2018)

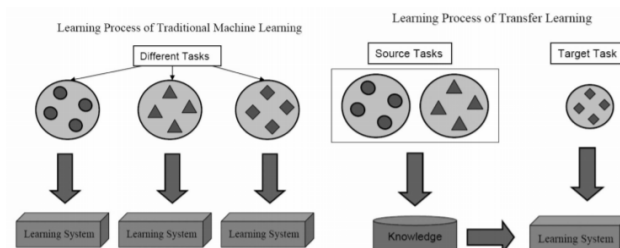


Figura 1.5: Esquema del aprendizaje de transferencia (obtenido de Pan y Yang, 2009)

1.3 Detección de pólipos

La detección automática de pólipos es una técnica de soporte para los especialistas que les permite aumentar las probabilidades de detectar este tipo de lesiones durante las exploraciones endoscópicas. Esta tarea se puede dividir en dos etapas como son la clasificación de las imágenes y la localización de los pólipos. Entre los principales retos que deben de afrontar los especialistas para la detección de los pólipos se encuentran la no uniformidad de la apariencia de los pólipos debido a sus diferentes formas, los efectos de la adquisición de la imagen como el reflejado especular y la similitud entre el tejido sano y los pólipos.

1.3.1 Clasificación de imágenes endoscópicas

La etapa de clasificación consiste en clasificar y diferenciar las imágenes que presentan uno o más pólipos de las que no presenten ninguno. La detección de pólipos es un problema binario y en la Tabla 1.3 se asigna 0 cuando la imagen no contiene pólipos y 1 cuando sí lo hace.

		Etiqueta verdadera	
		0	1
Etiqueta obtenida	0	Verdadero Negativo (VN)	Falso Negativo (FN)
	1	Falso Positivo (FP)	Verdadero Positivo (VP)

Tabla 1.3: Análisis de la clasificación de imágenes con pólipos

1.3.2 Detección y segmentación de pólipos

Según el nivel de detalle que se requiera en la localización y segmentación de los pólipos, se disponen de diferentes técnicas de análisis que van acompañadas de una clasificación previa de la imagen. El resumen de estas técnicas se puede tratar de entender de manera sencilla en la Figura 1.6.

Localización

La localización consiste en calcular el rectángulo mínimo que contiene un objeto, en nuestro caso, un pólipo. También se conoce como *bounding box* y ofrece una idea del tamaño y de la posición del pólipo. Los rectángulos mínimos de los pólipos se pueden obtener a partir de las máscaras binarias de imágenes endoscópicas segmentadas manualmente.

Detección de objetos

La detección de objetos es necesaria cuando se pretenden localizar objetos de diferentes clases en una misma imagen. Para ello se deberá de realizar una tarea de clasificación y de localización para cada uno de los objetos que se encuentren presentes. Esta tarea sería necesaria en nuestro estudio si se desearan detectar diferentes tipos de pólipos u otras lesiones.

Segmentación semántica

La segmentación semántica consiste en etiquetar cada píxel de la imagen con la etiqueta correspondiente de la clase a la que pertenece. Se puede entender como un proceso de clasificación para cada uno de los píxeles de la imagen. Una característica fundamental de la segmentación semántica es que tanto la imagen de entrada como la máscara de salida deben de tener el mismo tamaño. Las dos clases que se deben diferenciar en el problema presentado son el pólipo y el tejido sano.

La principal ventaja que presenta la segmentación semántica frente a la detección de objetos es que puede definir la forma del pólipo puesto que se realiza una clasificación de la imagen píxel por píxel. Esto puede resultar de interés en sistemas automáticos de ayuda al diagnóstico en endoscopias puesto que focalizaría más la atención del paciente sobre el pólipo disminuyendo así la probabilidad de que este no sea detectado.

Segmentación de instancias

La segmentación de instancias o *instance segmentation* se diferencia de la segmentación semántica en que es capaz de segmentar y separar diferentes instancias de una misma clase en una única imagen. Esta tarea resulta de interés para aquellas imágenes endoscópicas que puedan contener más de un pólipo.

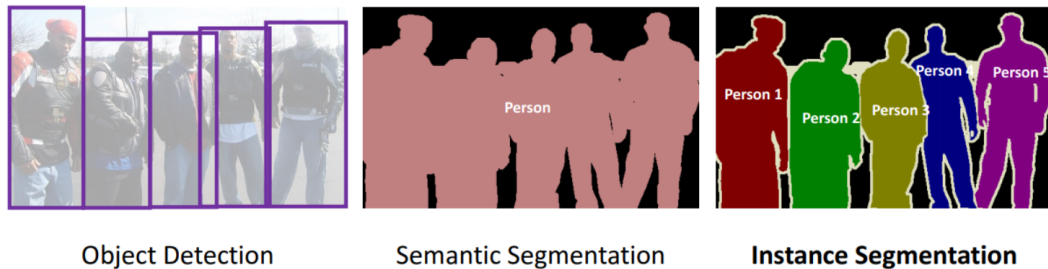


Figura 1.6: Resumen de las técnicas de segmentación (obtenido de Lamba, 2019)

1.4 Estado del arte

Tres tipos de aproximaciones han sido empleadas para la clasificación de imágenes endoscópicas y segmentación de pólipos. Estas tres se diferencian en función de si emplean técnicas de *Deep Learning*, si realizan una extracción de características manual o optan por una combinación de ambas.

Handcrafted methods

Los métodos manuales o *handcrafted* para la clasificación y segmentación de imágenes se basan en seleccionar características locales u holísticas de las mismas para formar descriptores que permitan caracterizar las diferentes regiones y objetos de una imagen. Entre los extractores de características más empleados se encuentran los histogramas de gradientes orientados (HOG, del inglés *Histogram Oriented Gradients*) y los patrones binarios locales (LBP, del inglés *Local Binary Patterns*). (Xiao y col., 2016)

Bernal y otros (2015) propusieron este tipo de enfoque para la detección de pólipos a partir de sus características geométricas. El método propuesto se inicia con un preprocesado de la imagen en que se eliminan artefactos inducidos por el sistema de adquisición como el reflejado especular. Este trabajo considera que los pólipos forman superficies protuberantes de la superficie del colon por lo que presentarán un contorno definido. Seguidamente, se aplica un filtro detector de bordes como el gradiente. También se calculan una serie de características de los bordes como son la completitud, la robustez, la continuidad y la concavidad que van a determinar si un borde forma parte o no de un pólipo. A partir de los bordes de mayor probabilidad se generan mapas de energía donde existe una elevada probabilidad de encontrar un pólipo. (Bernal y col., 2015)

End-to-end learning

Estos métodos consisten en utilizar herramientas de *Deep Learning* ya sea generando una nueva estructura de red neuronal o utilizando redes previamente diseñadas para otras finalidades apoyándose en el *Transfer Learning*. Este tipo de técnicas requiere recolectar una gran cantidad de imágenes y entrenar redes neuronales que se encarguen de realizar tanto la clasificación como la segmentación. Las redes neuronales convolucionales que se emplean para la segmentación de pólipos presentan dos secciones: un codificador o rama descendente que reduce la dimensionalidad obteniendo un mapa de características y un decodificador o rama ascendente que expande este mapa a alta resolución para realizar la separación entre clases. Una de las principales problemáticas que pueden enfrentar estos modelos es la escasa disponibilidad de datos correctamente etiquetados y segmentados.

En el trabajo de Wittenberg y otros (2019), se emplea el segmento principal de la red ResNet-101 para la extracción de características y la arquitectura *Mask R-CNN* para realizar la segmentación de instancias. Por su parte, Wang y otros (2018) emplearon una arquitectura de red previamente diseñada como es SegNet y que combina las etapas de codificación y decodificación.

En contraposición a las técnicas mencionadas que se caracterizaban por modificar y adaptar las arquitecturas de redes neuronales, el trabajo de Urban y otros (2018) optaron por definir su propia estructura de red. Esta incluye una serie de capas con filtros de convolución junto a capas completamente conectadas, funciones de activación no lineales y capas de agrupación como *maxpooling*. (Urban y col., 2018)

Los métodos híbridos para la segmentación de pólipos en imágenes de colonoscopias combinan tanto técnicas de extracción manual de características como técnicas de *Deep Learning* para realizar la segmentación. Un ejemplo de este enfoque es el trabajo de Tajbakhsh y otros (2015). En él se propone la obtención de un mapa de bordes aplicando el filtro canny y se estima la orientación del gradiente de los píxeles de borde. Para la construcción de los descriptores de los píxeles que conforman contornos se emplea la transformada discreta de coseno (DCT, del inglés *Discrete Cosine Transform*). Finalmente, se emplea una CNN para cada característica con el objetivo de obtener la probabilidad de cada píxel de pertenecer o no a un pólipo. (Tajbakhsh y col., 2015)

1.5 Objetivo del proyecto

Como se ha explicado anteriormente, una correcta detección y la consiguiente extirpación de los pólipos resulta absolutamente necesaria para reducir la probabilidad de sufrir cáncer de colon. Además, si se consigue una detección de los pólipos en sus primeros estadios de formación, se verá disminuida la mortalidad asociada a este tipo de cáncer. Debido a la heterogeneidad tanto de forma como de tamaño de los pólipos, estos no siempre resultan fácilmente identificables a simple vista por los endoscopistas que son los profesionales médicos que practican las exploraciones. Además, los endoscopistas suelen realizar múltiples intervenciones durante una misma jornada por lo que el cansancio puede jugar un papel negativo disminuyendo su capacidad de concentración y provocando que disminuya el porcentaje de pólipos detectados.

Debido a estos obstáculos que se pueden encontrar los profesionales médicos, los modelos basados en aprendizaje profundo pueden servir como una herramienta de soporte y ayuda a la decisión durante las intervenciones. En ningún momento estos sistemas tomarán decisiones por sí mismos, pero sí que pueden ser empleados como soporte de los endoscopistas con el objetivo de resaltar y destacar los pólipos presentes para reducir la tasa de pólipos perdidos. Además, estos modelos automáticos tienen la capacidad de analizar grandes cantidades de información por lo que resultarían verdaderamente útiles para el procesamiento de las imágenes obtenidas por cápsula endoscópica debido a la numerosa cantidad de fotogramas que se generan.

Por lo tanto, el objetivo de este proyecto es diseñar modelos de aprendizaje automático que adquieran el conocimiento necesario para el análisis de imágenes endoscópicas. Para ello, se debe disponer de una base de datos con imágenes de calidad aceptable, completa y correctamente etiquetada. Es decir, resulta necesario disponer de las etiquetas de las imágenes y las máscaras binarias de los pólipos para poder realizar el entrenamiento de redes neuronales basadas en aprendizaje supervisado. Además, se aplican técnicas de preprocesado de las imágenes con el objetivo de comprobar si mejoran o no el desempeño de los modelos.

Este proyecto también tiene la finalidad de explorar diferentes enfoques como son el aprendizaje de transferencia y el entrenamiento desde cero de las redes para realizar el entrenamiento de las redes neuronales. Los modelos de aprendizaje profundo que se diseñan se pueden dividir en tres tipos en función de la tarea que realicen ya sea esta la clasificación de las imágenes, la detección de los pólipos o su segmentación semántica. Finalmente, se exploran técnicas de postprocesado de la segmentación con el objetivo de verificar si mejoran los resultados de la misma.

Materiales y métodos

2.1 Material

2.1.1 Adquisición y análisis exploratorio

La adquisición de las imágenes para el desarrollo de los modelos de aprendizaje automático ha consistido en la descarga de dos bases de datos públicas de imágenes endoscópicas para uso educativo y de investigación como son *CVC-ClinicDB* y *ETIS-Larib Polyp DB*.

CVC-ClinicDB (de ahora en adelante, CVC) es una base de datos de imágenes endoscópicas obtenidas a partir de fotogramas de vídeos registrados durante las intervenciones en que cada una de las imágenes contiene como mínimo un pólipo. Esta base de datos contiene un total 612 imágenes con pólipos y sus correspondientes máscaras binarias que indican la posición de los pólipos. El tamaño de todas sus imágenes es de 288*384 píxeles (Bernal y col., 2015). Una característica que presentan todas las imágenes de esta base de datos es que contienen una máscara negra, entendida esta como una región hipointensa que rodea la región de interés, tal y como se puede ver en la Figura 2.4. Un método de preprocesado para eliminar esta máscara será propuesto más adelante.

Del mismo modo, *ETIS-Larib Polyp DB* (de ahora en adelante, ETIS) también contiene imágenes endoscópicas con uno o más pólipos extraídas a partir de fotogramas de vídeo. También incluyen el *Ground Truth* de cada una de las imágenes en forma de máscaras binarias que indican la localización de los pólipos. Esta base de datos incluye un total de 196 imágenes que tienen un tamaño 966*1225 píxeles por lo que son ostensiblemente más grandes que las de la base de datos CVC. (Silva y col., 2014)

El tamaño de los pólipos se puede determinar a partir de sus correspondientes máscaras binarias y se puede calcular como el porcentaje de área que estos ocupan respecto del total de la imagen. Se han analizado ambas base de datos y se ha comprobado que el tamaño promedio de los pólipos es del 9.31 % del tamaño de la imagen en la base de datos CVC y del 4.53 % en ETIS. Un resumen de ambas bases de datos se presenta en la Tabla 2.1. Como se ha comentado, ambas bases de

datos obtienen las imágenes a partir de fotogramas de secuencias de vídeo. Esto provoca que se dispongan de imágenes muy similares que han sido obtenidas a partir de fotogramas sucesivos por lo que contienen el mismo pólipo visto desde diferentes perspectivas.

Base de datos	Nº de imágenes	Tamaño (píxeles)	Tamaño pólipo (% área)
CVC-ClinicDB	612	288*384	4.53
ETIS-Larib DB	196	966*1225	9.31

Tabla 2.1: Resumen de las bases de datos

Resumiendo, se dispone un total de 808 imágenes con pólipos y sus respectivas máscaras binarias, es decir, no se disponen de imágenes endoscópicas sin pólipos. Uno de los objetivos de este estudio es diseñar modelos automáticos de clasificación que diferencien entre imágenes con o sin pólipos por lo que resulta imprescindible disponer de este último grupo de imágenes. Para ello, se han recortado en cuatro secciones del mismo tamaño las imágenes de la base de datos ETIS puesto que su extenso tamaño permitía mantener una calidad de imagen aceptable. Además, se ha procedido a clasificar estas imágenes en función de si presentaban o no pólipos atendiendo a la información que aportaban sus respectivas máscaras binarias. Se ha establecido un umbral del 5% del área a partir del cual consideramos que una imagen presenta un pólipo y se ha decidido utilizar como imágenes sin pólipo aquellas que no contenían ningún fragmento de pólipo. De este modo, se ha ampliado la base de datos ETIS obteniendo así un total de 199 imágenes con pólipos y 394 imágenes sin pólipos, lo que denota que ambas clases están desbalanceadas. El tamaño de estas imágenes de 484*612 píxeles. En la Figura 2.1 se presenta un ejemplo de la aplicación de este algoritmo observando a la izquierda la imagen original y a la derecha las cuatro imágenes recortadas. En este caso, únicamente la región inferior derecha será clasificada como una imagen con pólipo.

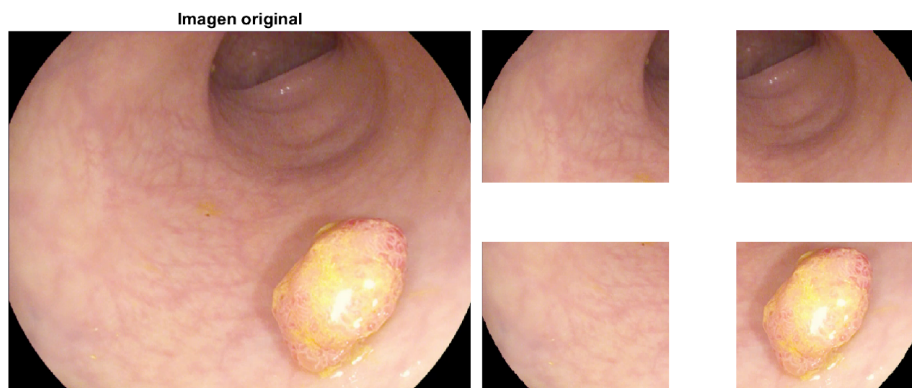


Figura 2.1: Separación entre imágenes normales y con pólipos

Las *bounding boxes* son los rectángulos mínimos que contienen a los objetos, en este caso, los pólipos. Aportan información acerca de su tamaño y localización, pero no acerca de su forma puesto que siempre presentan una forma rectangular. Se pueden definir de diferentes maneras, pero se ha optado por hacerlo a partir de cuatro parámetros como son la posición, en filas y columnas de la esquina superior izquierda y su altura y anchura. Esta información no era proporcionada por las bases de datos pero sí se ha podido obtener a partir de las máscaras

binaras. De este modo, se ha diseñado un método automático que se encarga de detectar el número de pólipos presentes en cada imagen, normalmente uno, y de calcular sus correspondientes rectángulos mínimos. El valor de las *bounding boxes* se normaliza respecto a las dimensiones de la imagen con el objetivo de que puedan ser empleadas si se modifica el tamaño de las mismas. Un ejemplo de las *bounding boxes* obtenidas para una de las imágenes de la base de datos se presenta en la Figura 2.2.

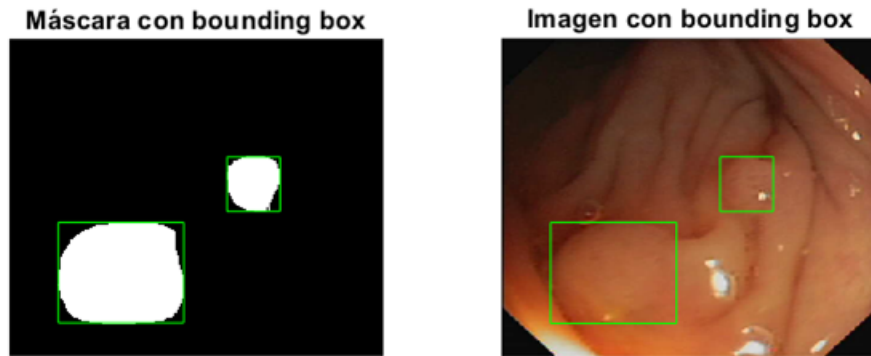


Figura 2.2: Ejemplo de representación de los rectángulos mínimos obtenidos

2.2 Métodos

En esta sección del proyecto se exponen los métodos implementados de preprocesado de las imágenes, se desarrollan los modelos de red neuronal, el entrenamiento de las mismas que se ha llevado a cabo y finalmente se propone un algoritmo de postprocesado de la segmentación semántica.

2.2.1 Preprocesado

Recortar la máscara negra

Uno de los principales problemas que acostumbra a presentar la adquisición de imágenes endoscópicas es la incorporación de una máscara negra. Como máscara negra se entiende la región de la imagen que aparece hipointensa rodeando la imagen endoscópica. En la Figura 2.4 se puede observar como este artefacto aparece tanto a ambos lados como arriba y abajo de la región de interés de la imagen endoscópica. Estas máscaras no aportan ningún tipo de información relevante y además aumentan el tamaño de la imagen innecesariamente incrementando así el coste computacional del entrenamiento de los modelos de red neuronal.

Uno de los posible enfoques para abordar esta problemática es recortar la máscara negra para cada una de las imágenes endoscópicas. Para ello, se ha diseñado un método automático que se encarga de recortar las máscaras negras no deseadas. El algoritmo propuesto utiliza información de la primera componente de las imágenes RGB puesto que el rojo es el color predominante en las imágenes estudiadas. Inicialmente obtiene la suma de la componente en filas o columnas y seguidamente calcula la diferencia entre estas con el objetivo de detectar las variaciones bruscas que se corresponden con la transición entre la imagen y la máscara negra, píxeles que se tomarán de referencia para recortar la imagen.

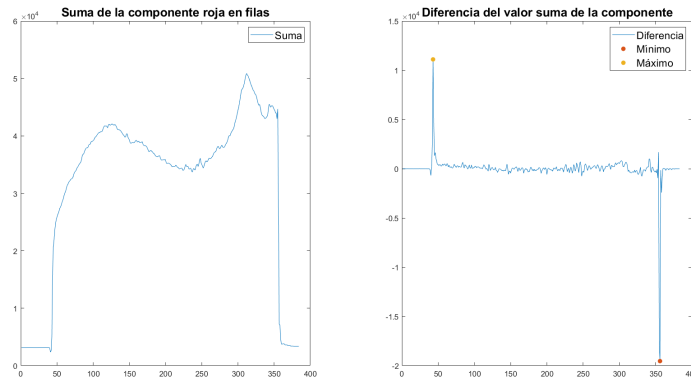


Figura 2.3: Resultado del recorte de la máscara negra

De este modo, se realiza un recorte de la imagen tanto a lo ancho como a lo alto con el objetivo de conservar únicamente la región de interés. Se presenta en la Figura 2.4 el resultado de recortar la máscara negra donde podemos comprobar que se ha eliminado la región de la imagen que no resultaba de interés. Además el algoritmo detecta las imágenes que han sido recortadas deficientemente en función de la variación de su tamaño en las dos dimensiones con el objetivo de desecharlas del estudio. El método tiene una efectividad del 98.2% tras aplicarlo sobre un conjunto de 611 imágenes con máscara negra.

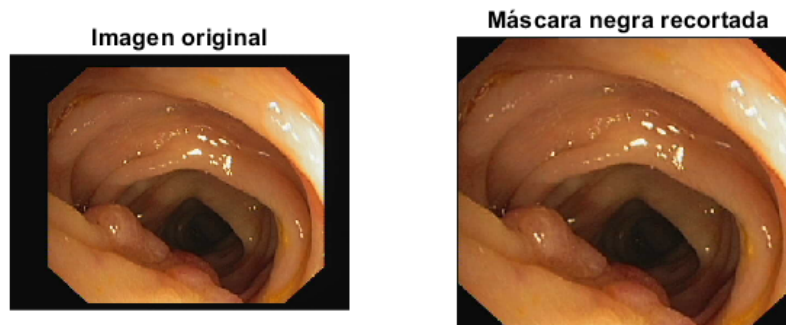


Figura 2.4: Resultado del recorte de la máscara negra

Reconstrucción del reflejo especular

El reflejo especular o *specular highlight* es uno de los artefactos más comunes en las imágenes endoscópicas. Este se debe a la captación del reflejo directo de la luz que incorpora la cámara para la adquisición de los vídeos (Bernal y col., 2013). Se puede corregir realizando una restauración o *impainting* de las zonas de la imágenes con un brillo más pronunciado. Para la obtención del reflejo especular, se calcula inicialmente la imagen diferencia entre la imagen original en escala de grises y el valor promedio de esta imagen. Los píxeles más hiperintensos de esta imagen diferencia son los que se corresponderán con el reflejado especular puesto que su diferencia respecto a la media es muy pronunciada. Finalmente, se consideran como reflejado especular aquellos píxeles que tengan un valor de intensidad en la imagen diferencia mayor que un umbral marcado por el 99% del histograma acumulado de la misma.

La restauración o *inpainting* de la imagen consiste en aplicar un algoritmo de difusión que se encarga de sustituir el valor de los píxeles considerados como reflejo especular por el promedio del valor de sus píxeles vecinos que no son considerados como tal. Se aplica el algoritmo de manera iterativa hasta que se reduce en un determinado porcentaje el área del reflejado especular o hasta que el resultado de aplicar el algoritmo sea el mismo en dos iteraciones consecutivas. El tamaño del vecindario para realizar la difusión se puede personalizar aunque lo conveniente es utilizar un radio unitario para conseguir una mayor semejanza al contorno. En la Figura 2.5 se presenta, de izquierda a derecha, la imagen original, los píxeles detectados como reflejado especular y la salida del método de restauración. El método implementado es una adaptación del algoritmo de difusión para la restauración propuesto por Bernal y col., 2013.

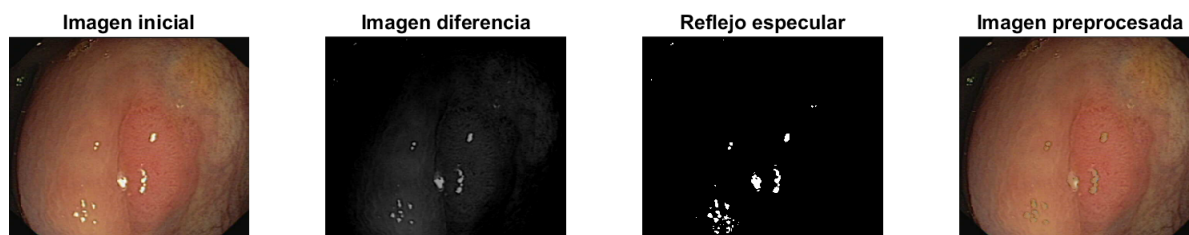


Figura 2.5: Resultado de la restauración de la imagen

2.2.2 Modelos de red neuronal

Las redes neuronales convolucionales están compuestas principalmente por tres tipos de capas que realizan diferentes operaciones matemáticas y que pueden agruparse en bloques de convolución. La operación de convolución consiste en realizar un producto de matrices elemento por elemento que depende del tamaño del filtro de kernel de convolución. Una característica relevante de esta operación es que aporta información espacial al analizar el vecindario de cada píxel. Las capas de convolución están también caracterizadas por el número de filtros que determinan el número de mapas de características a la salida de estas capas. Seguidamente a estas capas se emplean funciones de activación como las presentadas en la Tabla 1.2, mayoritariamente Rectified Linear Unit (ReLU), con el objetivo de propagar la información a través de la red. La operación de *maxpooling* consiste en seleccionar los valores máximos de una región con la finalidad de obtener un mapa de características relevantes y se emplea para reducir la dimensionalidad de los mapas. Finalmente, se emplean capas densas o completamente conectadas que realizan la separación entre clases a partir de las características obtenidas por las capas previas.

Modelos para clasificación y detección

La técnica conocida como *training from scratch* se caracteriza por realizar el entrenamiento de las redes neuronales desde cero, es decir, sin que estas hayan sido previamente entrenadas para otras tareas. Esta técnica nos permite utilizar arquitecturas de red previamente diseñadas con los pesos de las capas no inicializados o diseñar nuestras propias redes neuronales convolucionales. Se propone una CNN compuesta por cuatro bloques convolucionales que incluyen capas de convolución con un tamaño de filtro de 5x5 en un total de 128 canales. Por su parte, se emplea *Retified Linear Unit (ReLU)* como función de activación y *maxpooling* con un tamaño de 2x2.

Finalmente, se acoplan un total de cuatro capas completamente conectadas que finalizan en una capa de clasificación que diferencia entre dos clases, normal y pólipo.

Como ya se ha mencionado anteriormente, el aprendizaje de transferencia se caracteriza por utilizar conocimiento aprendido para aplicarlo a una tarea completamente distinta dentro del mismo dominio. Las cuatro redes neuronales convolucionales que se presentan a continuación han sido previamente entrenadas en más de un millón de imágenes del conjunto de datos de ImageNet, ampliamente utilizado en modelos de aprendizaje profundo. (Russakovsky y col., 2015)

AlexNet, cuya arquitectura se presenta en la Figura 2.6, es una CNN de ocho capas de profundidad cuyo tamaño de imagen de entrada es de 227×227 píxeles. Está compuesta por cuatro bloques de convolución, el primero de los cuales se caracteriza por realizar una reducción de la dimensionalidad del mapa de características mediante la operación de la convolución con *stride* de tamaño 4×4 , mientras que los otros utilizan la operación de *maxpooling*. AlexNet utiliza capas de *dropout* que consiguen la regularización de la red neuronal previniendo el sobreajuste de la misma y realiza la clasificación entre 1000 posibles clases (Krizhevsky y col., 2012). Para realizar el *Fine-tuning* característico del aprendizaje de transferencia, se opta por congelar los dos primeros bloques de convolución y por volver a entrenar los tres restantes con las imágenes de la tarea correspondiente.

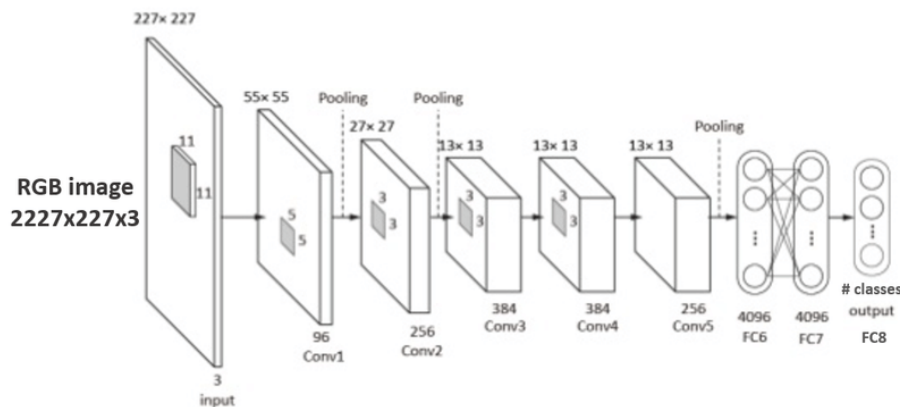


Figura 2.6: Esquema de la arquitectura de Alexnet (obtenido de Khvostikov y col., 2018)

GoogleNet (Szegedy y col., 2015) tiene un tamaño de entrada de las imágenes de 227×227 píxeles. Esta CNN se caracteriza por emplear múltiples módulos de *inception*. Esta arquitectura característica que compone algunas CNNs pondera la detección de características a diferentes escalas y reduce el coste computacional del entrenamiento de extensas redes a través de la reducción de dimensionalidad. Se caracteriza por emplear convoluciones 1×1 con múltiples canales que permiten aprender patrones a través de estos canales y capas de convolución con tamaños de filtro de 3×3 y 5×5 para detectar las características espaciales. Se decide mantener los pesos del primer módulo de *inception* y se reentrenan los dos restantes con las imágenes del conjunto de entrenamiento.

Por su parte, VGG-16, cuya arquitectura de red se presenta en la Figura 2.7, es una red neuronal con 16 capas con pesos que se pueden aprender. Se caracteriza por utilizar únicamente filtros

kernel de tamaño 3×3 y operaciones de *maxpooling* de 2×2 en unas imágenes de entrada de 227×227 píxeles. Estas capas se agrupan en cinco bloques de convolución a los cuales se les conectan capas densas para realizar la clasificación entre 1000 clases (Simonyan y Zisserman, 2014). En este caso, el aprendizaje de transferencia se realiza con el *Fine-tuning* de los bloques de convolución que tienen un tamaño de 28×28 píxeles o inferior.

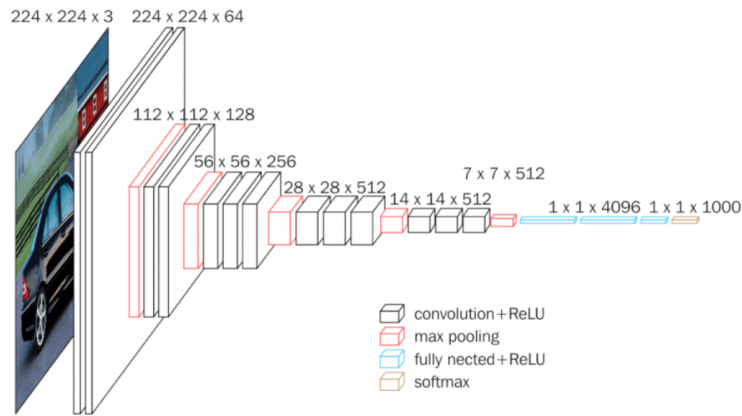


Figura 2.7: Esquema de la arquitectura de VGG16 (obtenido de Nash y col., 2018)

Finalmente, ResNet-50 (He y col., 2016) es una CNN con 50 capas de profundidad entre las que se encuentran un total de 48 capas de convolución. Esta arquitectura de red está basada en el aprendizaje residual que pretende resolver el desvanecimiento de gradiente que enfrentan las redes neuronales durante su entrenamiento cuando se aumenta su profundidad, es decir, cuando se añaden más capas. El *gradient vanishing* es una problemática que aparece cuando en las redes neuronales artificiales que se basan en el error del gradiente puesto que cuando este disminuye de manera considerable provoca que las capas de las redes no se entrenen correctamente en cada iteración. Para ello, ResNet-50 conecta las salidas de diferentes bloques de convolución permitiendo así que la información de gradiente pase a través de las mismas y que se realice el entrenamiento de las últimas capas de los modelos. En la Figura 2.8 se visualiza el diagrama de bloques de ResNet-50 donde se representan con flechas de color azul las conexiones residuales mencionadas. El *Fine-tuning* en este caso se realiza congelando los dos primeros bloques y entrenando los dos restantes.

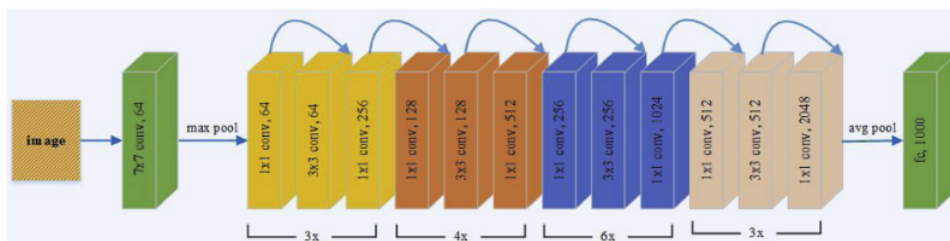


Figura 2.8: Diagrama de bloques de la arquitectura de ResNet-50 (obtenido de Talo, 2019)

La diferencia entre las tareas de clasificación y detección aparecen en las últimas capas de los modelos. Mientras que la clasificación entre imágenes sanas e imágenes con pólipos requiere una diferenciación binaria, la detección se basa en la obtención de las coordenadas normalizadas del rectángulo mínimo alrededor del pólipo. Por ello, las redes de clasificación utilizan la función

de activación *softmax* que aporta la probabilidad normalizada de pertenencia de cada imagen a una clase. Finalmente se realiza la clasificación con la función de pérdida de entropía cruzada. Por su parte, las redes empleadas para la detección utilizan la función de activación sigmoide y una capa de regresión con cuatro salidas normalizadas para cada una de las coordenadas de la *bounding box*. Al tratarse de una regresión, la función de pérdida se corresponde con el error cuadrático medio.

Modelos para segmentación

Las redes neuronales para segmentación se caracterizan por presentar dos secciones. Una primera sección descendente también denominada codificador que realiza una reducción de la dimensionalidad mediante operaciones de convolución y *maxpooling* con el objetivo de extraer características relevantes de la imagen. La segunda sección o decodificador es ascendente puesto que realiza un aumento de la dimensionalidad. Es decir, su objetivo es convertir información de baja resolución extraída por el codificador en información de alta resolución que permita diferenciar las clases y los objetos a segmentar. Una característica de estas redes es que las capas con el mismo tamaño de imagen en el codificador y el decodificador están conectadas con el objetivo de obtener mejores localizaciones por lo que es conveniente que ambos tramos de la red neuronal sean simétricas. De este modo, estas redes neuronales adquieren la capacidad de explicar no solo lo que presenta la imagen (tarea asociada a las redes de clasificación) sino también dónde lo presenta.

La salida de estas redes neuronales es una imagen del mismo tamaño que la imagen de entrada en que cada uno de los píxeles son clasificados consiguiendo así una separación entre clases. Entre las operaciones que se emplean para convertir información de baja resolución en imágenes de alta resolución se encuentran la convolución traspuesta y la operación de *unpooling*. La convolución traspuesta se caracteriza por no utilizar un modelo de interpolación concreto, sino que contiene parámetros que se pueden aprender durante el entrenamiento para una tarea específica. Para ello, se debe de utilizar la matriz de convolución que se obtiene reorganizando el filtro kernel que define a la operación característica de convolución e introduciendo ceros. De este modo, realizando la operación de transposición de esta matriz se puede realizar el *upsampling* o aumento de la dimensionalidad.

Por su parte, *unpooling* se corresponde con la operación contraria a *maxpooling*. Si el *maxpooling* se caracteriza por reducir la dimensionalidad seleccionando el valor máximo dentro del entorno de cálculo, la operación *unpooling* expande estos valores máximos a una resolución mayor. Las capas que realicen estas dos operaciones deben de estar conectadas puesto que resulta necesario conservar los índices de los valores máximos en el *maxpooling* para conservarlos en el *unpooling* introducir ceros en el resto. Ambas operaciones quedan definidas por el tamaño del filtro kernel y de paso o *stride*. Un resumen de estas operaciones se presenta en la Figura 2.9.

Para realizar el entrenamiento desde cero de redes neuronales de segmentación se ha optado por utilizar dos arquitecturas de red previamente diseñadas, como son SegNet y U-Net. SegNet (Badrinarayanan y col., 2017) es un arquitectura de red neuronal convolucional empleada para la segmentación semántica de imágenes. La arquitectura del codificador es idéntica a la arquitectura de la red VGG16. La principal característica de la arquitectura SegNet es que realiza operaciones de *unpooling* utilizando los índices de las capas de *maxpooling* para aumentar el mapa de características. También utiliza operaciones de convolución en el tramo del decodificador

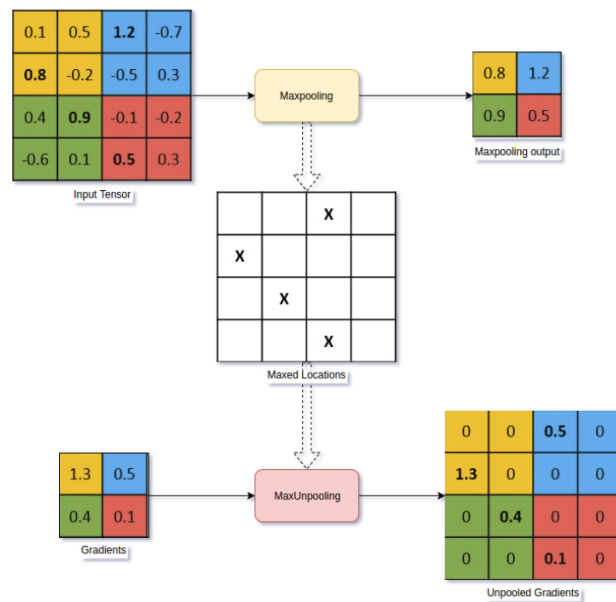


Figura 2.9: Esquema de la operación de *unpooling* (obtenido de Zafar y col., 2018)

con el objetivo de crear mapas densos de características. Se opta por diseñar una arquitectura de red de cuatro bloques de profundidad para unas imágenes de entrada de 264×310 píxeles cuyos mapas de características se reducen hasta un tamaño de 33×38 píxeles y se expanden de manera simétrica con operaciones de *unpooling* hasta alcanzar el tamaño de imagen original.

U-Net (Ronneberger y col., 2015), cuya arquitectura característica se presenta en la Figura 2.10, también contiene una rama descendente y otra ascendente. El codificador o rama descendente sigue la estructura característica de una red convolucional e incluye múltiples bloques que contienen capas de convolución con un kernel de tamaño 3×3 , seguidas de la función característica de activación como es ReLU y de operaciones de *maxpooling* con un *stride* igual a 2. Cada vez que se reduce la dimensionalidad en el codificador, se duplica el número de canales de características. Por su parte, la rama ascendente utiliza múltiples operaciones de convolución traspuesta con un *stride* de 2×2 al mismo tiempo que va reduciendo a la mitad el número de canales. Es importante destacar que se realizan conexiones entre el codificador y el decodificar para concatenar los mapas de características de ambos tramos que tengan el mismo tamaño. Finalmente, se utilizan capas de convolución con un kernel 1×1 para reducir el número de canales hasta el número de clases deseadas y se incluye una capa que realiza la clasificación, en este caso, entre las dos clases que se corresponden con el pólipo y el fondo.

Se propone la modificación de dos redes previamente entrenadas en el conjunto de datos de ImageNet como son Alexnet y VGG16 para realizar *Transfer Learning* y emplearlas como el codificador de las redes para codificación. En cuanto a Alexnet (Krizhevsky y col., 2012), se mantienen tres bloques de convolución en el tramo del codificador de los cuales dos ven sus pesos congelados con el objetivo de aplicar aprendizaje de transferencia mientras que el otro es reentrenado con las nuevas imágenes. Este tramo de la red neuronal realiza una reducción de la dimensionalidad de los mapas de características a un tamaño de 13×13 píxeles. Seguidamente, se diseña un decodificador simétrico que también contiene tres bloques de operaciones de convolu-

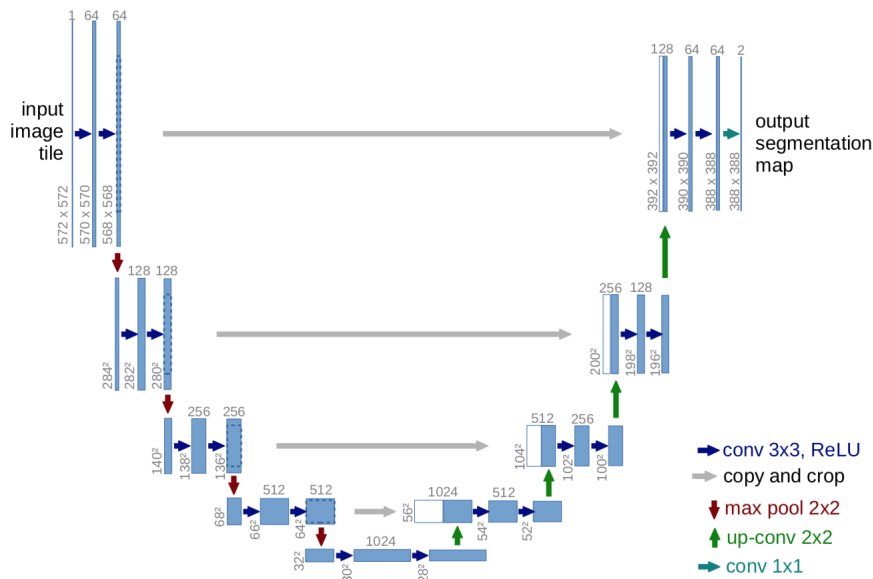


Figura 2.10: Esquema de la arquitectura de U-Net (obtenido de Ronneberger y col., 2015)

ción traspuesta y cuyas salidas se concatenan con los mapas de características del mismo tamaño en el codificador.

En cuanto a la red VGG16 (Simonyan y Zisserman, 2014), se mantienen los cuatro bloques de convolución que la caracterizan congelando los pesos de las capas de los dos bloques iniciales y se reentrenando las capas de los tres siguientes con la finalidad de realizar el *Fine-tuning* característico del aprendizaje de transferencia. El decodificador que se diseña es también simétrico a la rama descendente por lo que se realizan también las concatenaciones de las capas de la red con el mismo tamaño. Finalmente, se coloca una capa de convolución de tamaño de filtro 1*1 y con un total de dos filtros para realizar la separación entre las clases de tejido normal y pólipo.

2.2.3 Entrenamiento

La programación de los modelos de red neuronal diseñados se ha realizado en el entorno de programación de MATLAB[®]. El entrenamiento de estos modelos requiere de tarjetas gráficas debido a la gran cantidad de operaciones que deben de realizar. Las unidades de procesamiento gráfico o GPU son procesadores realizan múltiples operaciones en poco tiempo y son capaces de manejar grandes cantidades de memoria. Por ello, son ampliamente utilizadas en Inteligencia Artificial. En este caso, la GPU empleada es una Titan Xp del desarrollador NVIDIA.

El entrenamiento de las CNN de clasificación se realiza con un total de 593 imágenes endoscópicas, obtenidas de la base de datos ETIS, de las cuales 199 contienen uno más pólipos y las otras 394 se corresponden con imágenes normales del tracto digestivo. El conjunto de datos tiene las etiquetas notablemente desbalanceadas por lo que se ponderan cada una de la clases en función de sus pesos. Estos pesos se calculan en función de la frecuencia de aparición de las clases y sirven para ponderar las diferentes clases en la última capa de clasificación. También se debe destacar el conjunto de datos es limitado por lo que se ha optado por realizar el entrenamiento de hasta

cuatro modelos de entrenamiento. Se aplica una separación del 80 % para construir los conjuntos de entrenamiento (474 imágenes) y de test (119 imágenes) que servirá para medir la eficacia de la red en imágenes que todavía no ha visto. El tamaño de entrada de las imágenes se modifica pertinentemente para ajustarlo a la entrada de las redes neuronales empleadas.

El entrenamiento de las redes de detección se realiza sobre un total de 601 imágenes de la base de datos CVC de las cuales se han obtenido sus *bounding boxes* con el método previamente entrenado. En esta tarea, se realiza de nuevo una separación del 80 % para obtener los conjuntos de test y entrenamiento. Es importante realizar la modificación oportuna de las coordenadas de los rectángulos mínimos si se cambia el tamaño de la imagen para realizar el aprendizaje de transferencia.

Para la caracterización de los modelos de segmentación se emplea un total de 797 imágenes endoscópicas y sus correspondientes máscaras obtenidas de ambas bases de datos. Al formar parte de diferentes conjuntos de datos, las imágenes presentan un diferente tamaño. Como la base de datos dominante es CVC y estas presentan un menor tamaño, se ha optado por reescalar las imágenes de ETIS al mismo tamaño que las de CVC reduciendo así el coste computacional de los modelos. La función de pérdida empleada en todas las redes para segmentación es la entropía cruzada. El optimizador empleado para el entrenamiento de todos los modelos es *adam*.

2.2.4 Postprocesado

En algunas ocasiones, la salida de redes de segmentación semántica como las diseñadas contienen múltiples objetos que, al mismo tiempo, pueden presentar huecos, es decir, píxeles que no se han clasificado como objeto cuando están completamente rodeados por estos. En el presente proyecto, se aplica un algoritmo de postprocesado a los resultados de la segmentación semántica que pretende solucionar esta problemática. Para proponer el método postprocesado, se asume que las imágenes endoscópicas contienen únicamente un pólipo lo que ocurre en el 99 % de las imágenes de las bases de datos en que se está trabajando.

El método de postprocesado diseñado consiste en dos pasos. El primer paso estriba en seleccionar el objeto más grande de la segmentación semántica rechazando así posibles detecciones de pequeño tamaño que puedan ser incorrectas. El segundo consiste en rellenar los huecos del objeto principal seleccionado para así disponer de un objeto compacto. Este paso requiere de realizar un cierre de la imagen, es decir, concatenar una dilatación y una erosión con el objetivo de cerrar los agujeros dentro de los objetos sin modificar el tamaño del mismo.

2.3 Diagrama de trabajo

A continuación, en la Figura 2.11 se presenta un resumen del diagrama de trabajo que se ha llevado a cabo para la realización del proyecto. Se presentan por separados las tres tareas llevadas a cabo por los respectivos modelos y se especifica que bases de datos e imágenes se han empleado para cada una de ellas. El entrenamiento de todas las redes neuronales se realiza sobre las imágenes preprocesadas a no ser que se exprese lo contrario con el objetivo de comprobar la efectividad del método de preprocesado implementado.

Resumiendo, la clasificación se ha realizado sobre el conjunto de imágenes obtenidas tras recortar en cuatro y clasificar pertinentemente en imágenes sanas y con pólipos las de las base de datos ETIS. Se han explorado cuatro redes distintas para realizar el aprendizaje de transferencia y se ha experimentado con una red neuronal de diseño propio y entrenada desde cero tras realizar una separación del 80% para obtener el conjunto de entrenamiento. Para la tarea de detección se exploran los mismos modelos que para la clasificación con la modificación pertinente de las últimas capas completamente conectadas para disponer de cuatro salidas que se correspondan con las coordenadas de las *bounding boxes*. Un detalle importante de las redes para detección es que aquellas imágenes que contienen más de un pólipo son vistas dos veces por la red neuronal y en cada una de ellas su respuesta coincide con las coordenadas de una de las *bounding boxes* que definen los pólipos.

En cambio, en las redes para segmentación sí se emplea la información de imágenes y *Ground Truth* de ambas bases de datos realizando el entrenamiento de los cuatros modelos de red neuronal explicados anteriormente. Esta tarea finaliza con la aplicación del método de postprocesado de las máscaras semánticas. Al disponer en ambas bases de datos de imágenes de diferentes tamaños, se opta por modificar el tamaño de las imágenes de la base de datos ETIS igualándolas a las de la base de datos CVC que presentan una extensión menor.

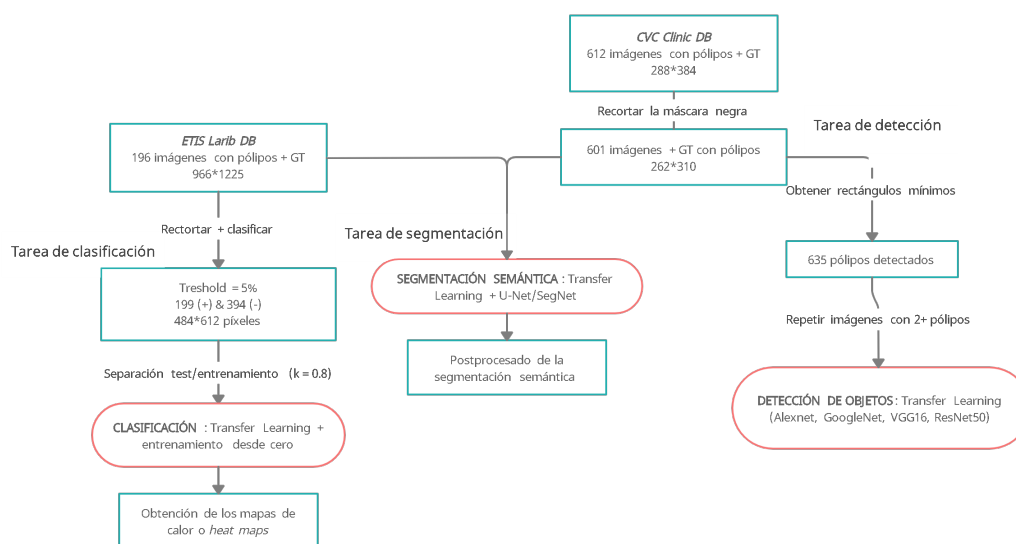


Figura 2.11: Resumen del diagrama de trabajo

Resultados

En esta sección se introducen las métricas empleadas para caracterizar cada una de las tres tareas llevadas a cabo y se presentan los resultados de los distintos modelos utilizados para cada una de ellas.

3.1 Clasificación

Las cinco métricas que se han empleado en la tarea de clasificación se presentan a continuación. Las abreviaturas utilizadas se han presentado anteriormente en la Tabla 1.3.

- **Precisión:** Fracción de predicciones correctas en el conjunto de datos reservado para test:

$$\frac{VP + VN}{n} \quad (3.1)$$

Siendo n el número total de imágenes en el conjunto de test, VP el número de Verdadero Positivo y VN el número de Verdadero Negativo.

- **Sensibilidad:** Fracción de detecciones positivas correctamente clasificadas como tal por el modelo:

$$\frac{VP}{VP + FN} \quad (3.2)$$

Siendo VP el número de Verdadero Positivo y FN el número de FN.

- **Especificidad:** Fracción de detecciones negativas correctamente clasificadas como tal por el modelo:

$$\frac{VN}{VN + FP} \quad (3.3)$$

Siendo VN el número de Verdadero Negativo y FP el número de FP.

- **F1-Score:** Armónica media ponderada de la precisión y la sensibilidad del modelo calculada como:

$$\frac{2 * \text{precisión} * \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}} \quad (3.4)$$

- **Área bajo la curva ROC (AUC):** Área bajo la curva ROC (*Receiver Operating Characteristic Curve*), que representa la capacidad de diagnóstico de un clasificador binario. Un valor alto del parámetro AUC cercano a 1 significa que el modelo tiene una alta capacidad de realizar clasificaciones correctas, sin embargo, un valor cercano a 0.5 implica que su capacidad de diferenciación entre clases es nula.

En la Tabla 3.1 se presentan los resultados de las métricas tras realizar un aprendizaje de transferencia con las redes previamente entrenadas. En esta tabla se separan los resultados en función de si se ha realizado *Fine-Tuning* o extracción de características.

<i>Modelo</i>	<i>Precisión</i>	<i>Sens.</i>	<i>Espec.</i>	<i>F-Score</i>	<i>AUC</i>
Alexnet: Fine-Tuning	0.882	0.675	0.987	0.749	0.831
Alexnet: Feature Extraction	0.655	0.825	0.567	0.6168	0.697
GoogleNet: Fine-Tuning	0.899	0.825	0.567	0.846	0.881
GoogleNet: Feature Extraction	0.714	0.7	0.722	0.622	0.711
VGG16: Fine-Tuning	0.815	0.625	0.911	0.694	0.768
VGG16: Feature Extaction	0.824	0.775	0.848	0.747	0.816
ResNet50: Fine Tuning	0.899	0.7	1	0.824	0.85
ResNet50: Feature Extraction	0.849	0.899	0.861	0.788	0.843

Tabla 3.1: Resultados de la tarea de clasificación aplicando *Transfer Learning*

En la Tabla 3.2 se presentan los resultados de diferentes modelos de clasificación entrenados desde cero con la red propia diseñada. Se compara el entrenamiento con las imágenes original con el de las imágenes a las que sí se ha aplicado el método de preprocesado propuesto.

<i>Modelo</i>	<i>Precisión</i>	<i>Sens.</i>	<i>Espec.</i>	<i>F-Score</i>	<i>AUC</i>
Imágenes originales	0.843	0.9	0.810	0.791	0.855
Imágenes sin preprocesar	0.849	0.6	0.975	0.727	0.787

Tabla 3.2: Resultados de la tarea de clasificación aplicando *Transfer Learning*

3.2 Detección

La métrica empleada para caracterizar el modelo de clasificación es:

- **Intersección sobre la unión (IoU)**: Fracción entre la intersección de las *bounding boxes* y su unión:

$$\frac{Pred \cup True}{Pred \cap True} \quad (3.5)$$

Siendo *Pred* el rectángulo mínimo predicho por el modelo de aprendizaje profundo y *True* el rectángulo mínimo verdadero del objeto.

La intersección sobre la unión se caracteriza calculando el promedio y la desviación típica de las mimas en el conjunto de test. En la Figura 3.1 se presentan de forma visual los resultados obtenidos tras realizar aprendizaje de transferencia con las cuatro modelos diseñados (1, Alexnet; 2, GoogleNet, 3, VGG16; 4, ResNet50). Se muestran las métricas tanto del *Fine-tuning* como de la extracción de características. Las barras verticales indican el valor promedio de la métrica y las barras de error se han calculado como la mitad de la desviación típica de la métrica en el conjunto de entrenamiento. También se ha realizado el entrenamiento de la red diseñada desde cero obteniendo un valor promedio de la $IoU = 0.3247$ sin aplicar el preprocesado de las imágenes y del $IoU = 0.395$ cuando sí se ha aplicado el método de preprocesado desarrollado.

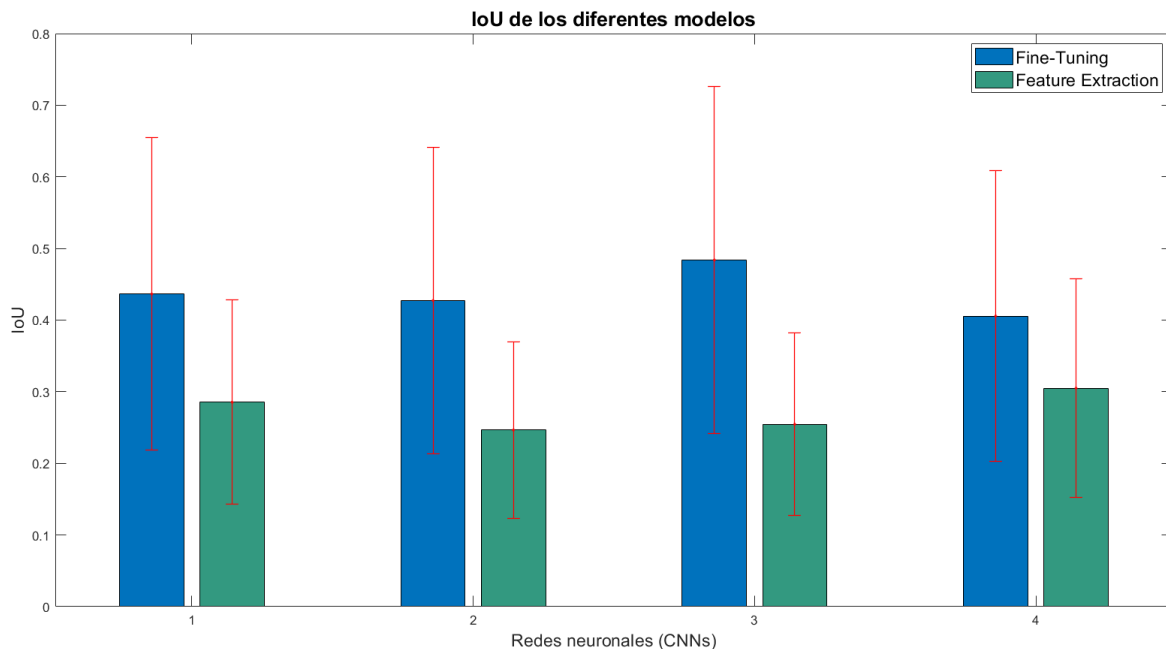


Figura 3.1: Resultados de la tarea de detección con *Transfer Learning*

3.3 Segmentación

Las métricas empleadas para caracterizar la tarea de segmentación son:

- **Índice de Jaccard:** Fracción entre la intersección de la máscara predicha y la máscara verdadera y la unión de ambas:

$$\frac{Pred \cup True}{Pred \cap True} \quad (3.6)$$

Donde *Pred* y *True* se corresponden con la máscara predicha y la máscara verdadera, respectivamente.

- **Índice de Dice:** El coeficiente de similitud propuesto por Sorensen y Dice es una métrica que permite medir la semejanza de dos muestras (Sorensen, 1948; Dice, 1945). Aplicado a imágenes binarias se calcula como:

$$\frac{2 * Pred \cup True}{Pred + True} \quad (3.7)$$

Los resultados obtenidos en la tarea de segmentación para los diferentes modelos propuestos se presentan en la Tabla 3.3.

Modelo	Índice de Jaccard	Índice de Dice
Transfer Learning: AlexNet	0.3381	0.4374
Transfer Learning: VGG16	0.4663	0.5540
Training from scratch: SegNet	0.3483	0.4824
Training from scratch: U-Net	0.2086	0.3089

Tabla 3.3: Resultados de la tarea de segmentación

El método de postprocesado diseñado para las máscaras de segmentación semántica se puede caracterizar utilizando las mismas métricas utilizadas para la segmentación. Por ello, se calcula el valor de ambas métricas tanto antes como después de aplicar el postprocesado y también se obtiene el valor de la diferencia. Por lo tanto, que la diferencia sea positiva significará que la técnica ha mejorado la segmentación; cuando sea negativa, la habrá empeorado. En la Tabla 3.4 se presentan los resultados de las métricas de segmentación obtenidos una vez se ha aplicado el postprocesado sobre un conjunto de 159 imágenes que han sido segmentadas con el modelo generado a partir de la arquitectura de red de SegNet.

Métrica	Valor inicial	Valor final	Incremento
Índice de Jaccard	0.3227	0.3524	0.0297
Índice de Dice	0.4474	0.4703	0.0229

Tabla 3.4: Resultados del método de postprocesado

Discusión

4.1 Tarea de clasificación

En la Tabla 3.1 se presentan los resultados para las métricas presentadas tras realizar aprendizaje de transferencia y en ella se pueden comparar los resultados de la clasificación al realizar *Fine-tuning* y extracción de características con la misma red previamente entrenada. La tendencia generalizada que se observa es que los modelos empeoran al realizar la extracción de características en comparación al Fine-tuning, es decir, realizan una peor clasificación. Estos resultados parecen lógicos puesto que al realizar este tipo de aprendizaje de transferencia, la red previamente entrenada no se actualiza con las imágenes de la tarea objetivo ya que únicamente realiza un extracción de características de las mismas a partir del conocimiento que ya había aprendido. Por su parte, si comparamos los resultados al realizar el *Fine-tuning* de estas redes neuronales vemos que la precisión se sitúa siempre por encima del 80 % y que no existen diferencias excesivamente significativas entre los diferentes modelos. Al mismo tiempo vemos que al realizar aprendizaje de transferencia con la red ResNet-50 se obtienen unos resultados ligeramente superiores rondando el 90 % de precisión en la tarea de clasificación. Estos resultados se ajustan al estado del arte puesto que las redes residuales fueron las que mejores resultados obtuvieron en su momento mejorando a las otras tres redes empleadas. (He y col., 2016).

En los modelos de clasificación, siempre existe un equilibrio que se debe de mantener entre la sensibilidad y la especificidad que son las métricas que indican la capacidad del sistema automático de detectar los casos positivos y negativos, respectivamente. En el caso del presente proyecto, la finalidad de los sistemas automáticos de detección de pólipos es advertir al médico de la presencia de estas lesiones con el objetivo de que ninguna sea pasada por alto. Por ello, resulta intuitivo pensar que sería recomendable tener un valor más alto de sensibilidad que de especificidad minimizando así el número de falsos negativos que se corresponderían con los pólipos no detectados por el modelo. En caso de disponer de algún modelo con una baja especificidad, es decir, que detecte muchos falsos positivos, sería necesario que el médico empleara su conocimiento previo para rechazar aquellas detecciones incorrectas que no se corresponden con los pólipos.

Los resultados de la tarea de clasificación son satisfactorios puesto que el nivel de precisión que consiguen es notable, pero igualmente resulta necesario investigar qué está teniendo en cuenta la red neuronal entrenada para realizar dichas clasificaciones lo que se puede comprobar con los mapas de calor. Los *heat maps* o mapas de calor nos permiten identificar qué características de la imagen está considerando la red neuronal. Estos resultan un método muy visual para determinar si la clasificación que está realizando el modelo es lógica o si la está teniendo en cuenta regiones y características que no resultan significativas. Cabe recordar que la principal característica que definía el modelo de apariencia de los pólipos es que aparecen como regiones predominantes rodeadas por unos bordes que suelen estar bien definidos.

Existen diferentes técnicas para obtener los mapas de calor entre las que se encuentra Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju y col., 2017). Esta técnica utiliza el gradiente de cada uno de los conceptos objetivo, en este caso los pólipos, en la última capa de convolución para crear el mapa de localizaciones resaltando las zonas significativas para realizar la clasificación. Esto se realiza puesto que las capas de convolución almacenan información espacial que posteriormente se pierde en las capas completamente conectadas que realizan la clasificación. En la Figura 4.1 se presenta la aplicación del método Grad-CAM a un imagen de la base de datos CVC tras realizar su clasificación con la modelo obtenido a partir del *Transfer Learning* con ResNet50. En esta misma figura podemos comprobar que la clasificación que realizada es correcta al tratarse de una imagen de pólipo y que el mapa de calor obtenido por Grad-CAM se ajusta a la localización del pólipo de manera bastante satisfactoria. Del mismo modo, vemos como también tiene en cuenta una región central de la imagen que no se corresponde con un pólipo, pero que sí presenta información semejante a la de los pólipos al tratarse de un pliegue del tracto digestivo. Finalmente podemos concluir que la red neuronal sí ha aprendido de las características de manera satisfactoria que definen a los pólipos.

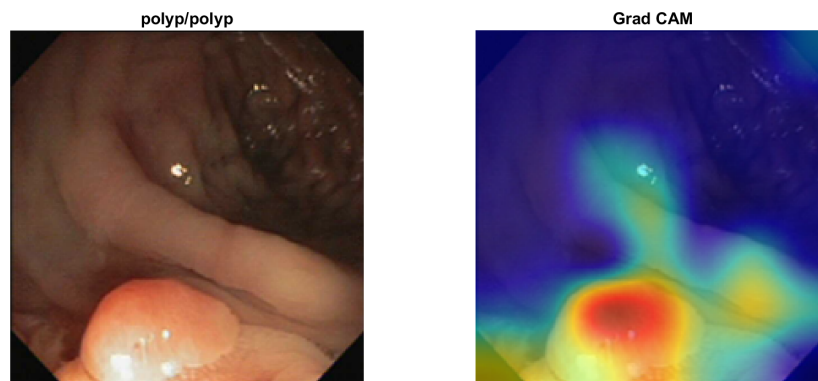


Figura 4.1: Heat maps de los modelos de clasificación

Sin embargo, en la Figura 4.2, podemos comprobar que la clasificación del modelo automático no es correcta al detectar como normal una imagen que sí presentaba un pólipo. Observando el mapa de calor, vemos que la red neuronal diseñada no ha tenido en cuenta el pólipo para realizar la clasificación sino que se ha tenido en cuenta otra región de la imagen que no debería resultar relevante. Esta clasificación incorrecta podría deberse al pequeño tamaño del pólipo presente en la imagen y a su similitud con el tejido circundante.

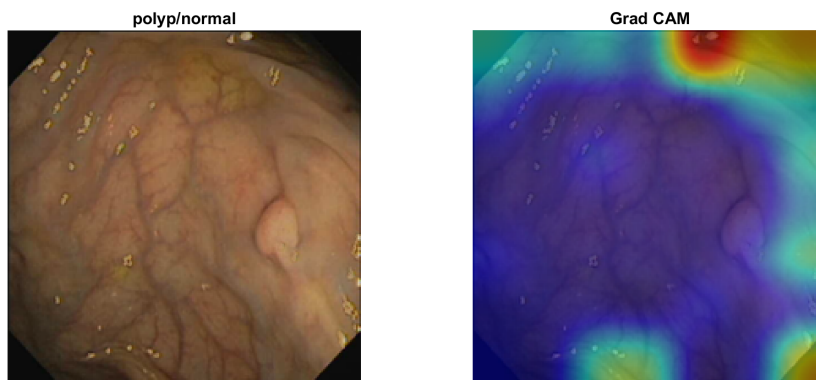


Figura 4.2: Heat maps de los modelos de clasificación

4.2 Tarea de detección

La métrica empleada para caracterizar las redes de detección es la intersección sobre la unión o OLR (del inglés, *OverLap Ratio*). Esta métrica resulta complicada de analizar si únicamente se presenta el valor obtenido en el conjunto de entrenamiento. Por ello, se adjunta la Figura 4.3 en que se pueden visualizar tanto la *bounding box* verdadera (rectángulos rojos) como el rectángulo mínimo obtenido a partir de las salidas de la red de detección (rectángulos azules). En esta figura, se presenta a la izquierda una buena detección con un $OLR \approx 0.75$ en que se verifica que es suficiente para detectar el pólipo de manera precisa; a la derecha se hace lo propio con una detección menos precisa ($OLR \approx 0.5$) en que se puede intuir la posición del pólipo pero no su localización exacta.

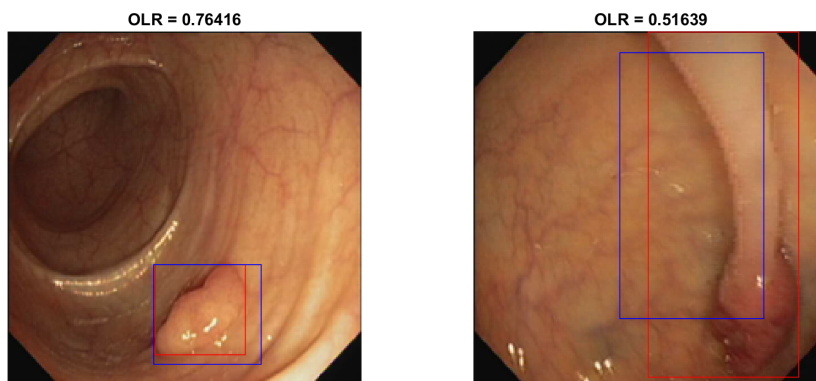


Figura 4.3: Resultado de la red de detección

En esta tarea de detección, podemos apreciar de nuevo que el *Fine-tuning* es más potente que la extracción de características puesto que el promedio de la intersección sobre la unión es considerablemente mayor cuando se aplica la primera de las técnicas. Existe un incremento promedio del 16% de esta métrica en comparación del *Fine-Tuning* con la extracción de características para los cuatro modelos de red neuronal. El mayor valor de la intersección sobre la unión se obtiene para la red VGG-16 y es cercano al 50% de lo que se deduce que aproximadamente la mitad de las detecciones resultarán bastante significativas para localizar el pólipo.

Un inconveniente que presenta esta tarea es la elevada dispersión de los valores de la intersección sobre la unión cuando se analiza el conjunto de test. En la Figura 3.1 vemos que como el abanico de valores que abarca esta métrica en el conjunto de entrenamiento es considerablemente amplio. Esto puede deberse a que algunas imágenes presentarán rectángulos mínimos muy semejantes al tratarse de fotogramas consecutivos por lo que la red realizará una muy buena clasificación de las mismas si ha visto alguna similar a ellas durante el entrenamiento. Además, los pólipos no tienen por qué presentar una forma rectangular por lo que es probable que las *bounding boxes* no se ajuste de manera correcta al pólipo. Esto también puede provocar que el solape no se produzca en el región de interés lo que habría que investigar visualizando las detecciones. Esta problemática se soluciona con la segmentación semántica que sí realiza una clasificación píxel por píxel de cada una de las clases.

4.3 Tarea de segmentación

En la Tabla 3.3 se han presentado los resultados de la tarea de segmentación para las diferentes técnicas. Analizando los resultados, destaca la potencia que presenta el aprendizaje de transferencia en comparación con el entrenamiento de las redes desde cero. El *Transfer Learning* VGG16 muestra unos valores en ambas métricas ligeramente superiores a los dos modelos entrenados desde cero. Esto resalta la potencia que presentan las redes neuronales entrenadas en conjuntos de millones de imágenes y que realizan una extracción de características generalistas en las primeras capas que a la postre resulta de gran utilidad para otras tareas. Los resultados del modelo basado en la arquitectura de U-Net son totalmente insatisfactorios a pesar de presentar un arquitectura de red muy semejante a los modelos diseñados para aprendizaje de transferencia y que está basada en operaciones de convolución traspuesta en la rama del decodificador.

Al mismo tiempo, destaca la elevada precisión del modelo obtenido a partir de VGG16 que consigue un índice de Dice superior al 50 % en un conjunto de test ostensiblemente grande formado por 159 imágenes. Esto significa que cerca de la mitad de las segmentaciones semánticas realizadas serán más o menos satisfactorias y permitirán detectar de el pólipo de manera aproximada. En la Figura 4.4 se presenta a la izquierda la máscara binaria verdadera superpuesta a la imagen a original y en la derecha se hace lo propio con una segmentación semántica obtenida por este modelo basado en la arquitectura de la red VGG16.

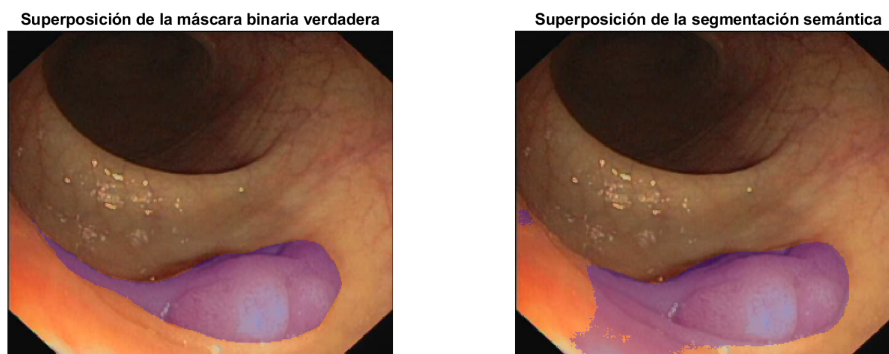


Figura 4.4: Resultado de la red de segmentación

Los resultados del método de postprocesado se han presentado en la Tabla 3.4. El método diseñado tiene un incremento promedio del 9.2% (0.0297 unidades del índice) lo que refleja que como norma general este método mejora los resultados de la segmentación obtenida por los modelos. En caso de que el promedio de las métricas de segmentación fuera menor tras el postprocesado, deberíamos de concluir que el método implementado no está mejorando la tarea de segmentación. En la Figura 4.5 se presenta la salida de método implementado comparándolo con la máscara binaria verdadera. Se puede comprobar como al seleccionar un único objeto y rellenar los huecos presentes dentro del objeto se facilita la visualización y la interpretación de la segmentación semántica. En el caso presentado, el índice de Jaccard de la segmentación mejora del 65.58% al 65.83%.

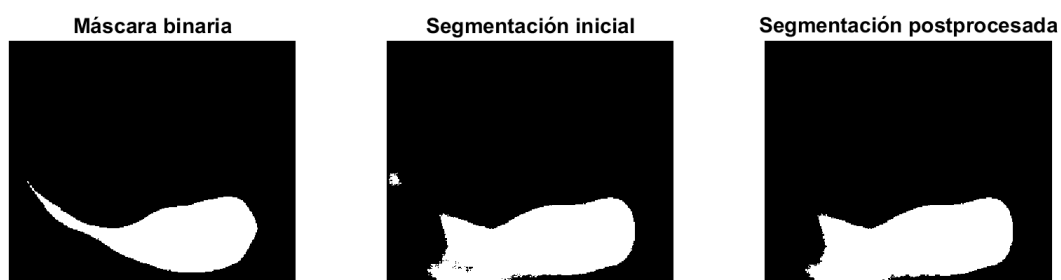


Figura 4.5: Resultado del método de postprocesado

Capítulo 5

Conclusión

Durante el desarrollo del proyecto, se ha comprobado el potencial que esconden los modelos de aprendizaje profundo basados en redes neuronales para la realización de diferentes tareas que pueden servir como sistemas de ayuda a la decisión para los profesionales médicos. Además, se han caracterizado estos modelos con redes neuronales que presentaban diferentes arquitecturas y se ha podido comprobar cómo realizar aprendizaje de transferencia con la red ResNet-50 conseguía unos resultados ligeramente superiores. Esto demuestra que su arquitectura caracterizada por conexiones residuales es útil puesto que permite el entrenamiento de la red evitando el desvanecimiento del gradiente en las primeras capas.

En cuanto al aprendizaje de transferencia, también se han explorado las dos opciones disponibles para su abordaje como la extracción de características y el *Fine-tuning*. Al mismo tiempo, se ha comprobado que de manera generalizada el *Fine-tuning* de redes previamente entrenadas en otros conjuntos de datos es mucho más potente que la extracción de características puesto que al volver a entrenar los últimos bloques mejora la capacidad de diferenciación del modelo. La tarea de clasificación mediante aprendizaje de transferencia alcanza unos resultados satisfactorios sobre el conjunto de datos superando el 80% de precisión en la clasificación de las imágenes endoscópicas.

Por su parte, las tareas de detección y segmentación presentan una función similares puesto que ambas tareas tienen como objetivo la localización de los pólipos. Sin embargo, la segmentación tiene un mayor potencial al realizar una clasificación píxel por píxel de la imagen frente a la detección de una *bounding box* que realizan los modelos de detección. La problemática que presentan estos rectángulos mínimos es que no se ajustan a la forma del pólipo, sino que ofrecen información acerca de su posición y tamaño. Sin embargo, el coste computacional de los modelos de red neuronal basados en segmentación es mucho mayor debido al número de capas necesarias para realizar la expansión del mapa de características.

Bibliografía

- Society, A. C. (2020b). *What Is Colorectal Cancer? | Colorectal Cancer Research Statistics*. <https://www.cancer.org/cancer/colon-rectal-cancer/about.html>. (Vid. págs. 4, 5)
- Society, A. C. (2020a). *Survival Rates for Colorectal Cancer*. <https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html>. (Vid. págs. 5)
- Zauber, A. G., Winawer, S. J., O'Brien, M. J., Lansdorp-Vogelaar, I., van Ballegooijen, M., Hankey, B. F., Shi, W., Bond, J. H., Schapiro, M., Panish, J. F. y col. (2012). Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths. *N Engl J Med*, 366, 687-696 (vid. págs. 6).
- Amador Hidalgo, L. (1996). *Inteligencia artificial y sistemas expertos*. Universidad de Córdoba, Servicio de Publicaciones. (Vid. págs. 6).
- Zhang, X.-D. Machine learning. En: *A Matrix Algebra Approach to Artificial Intelligence*. Springer, 2020, pp. 223-440 (vid. págs. 6).
- Kaelbling, L. P., Littman, M. L. & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4, 237-285 (vid. págs. 7).
- Torres, J. (2018). *DEEP LEARNING Introducci—n pr \grave{a} ctica con Keras*. Lulu. com. (Vid. págs. 7, 9).
- Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. En: *Icml*. 2010 (vid. págs. 8).
- Stewart, M. (2019). *Simple Introduction to Convolutional Neural Networks*. <https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac>. (Vid. págs. 10)
- Pan, S. J. & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359 (vid. págs. 10).

-
- Sarkar, D. D. (2018). A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning. (Vid. pág. 10).
- Lamba, H. (2019). *Understanding Semantic Segmentation with UNET*. <https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47>. (Vid. pág. 12)
- Xiao, X., Xu, D. & Wan, W. Overview: Video recognition from handcrafted method to deep learning method. En: *2016 International Conference on Audio, Language and Image Processing (ICALIP)*. IEEE. 2016, 646-651 (vid. pág. 13).
- Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C. & Vilarino, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, *43*, 99-111 (vid. págs. 13, 15).
- Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W. & Baldi, P. (2018). Deep learning localizes and identifies polyps in real time with 96 % accuracy in screening colonoscopy. *Gastroenterology*, *155*(4), 1069-1078 (vid. pág. 13).
- Tajbakhsh, N., Gurudu, S. R. & Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, *35*(2), 630-644 (vid. pág. 14).
- Silva, J., Histace, A., Romain, O., Dray, X. & Granado, B. (2014). Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, *9*(2), 283-293 (vid. pág. 15).
- Bernal, J., Sánchez, J. & Vilarino, F. Impact of image preprocessing methods on polyp localization in colonoscopy frames. En: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2013, 7350-7354 (vid. págs. 18, 19).
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. y col. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252 (vid. pág. 20).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105 (vid. págs. 20, 23).
- Khvostikov, A., Aderghal, K., Benois-Pineau, J., Krylov, A. & Catheline, G. (2018). 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies. *arXiv preprint arXiv:1801.05968* (vid. pág. 20).

-
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, 1-9 (vid. pág. 20).
- Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (vid. págs. 21, 24).
- Nash, W., Drummond, T. & Birbilis, N. (2018). A review of deep learning in the study of materials degradation. *npj Materials Degradation*, 2(1), 1-12 (vid. pág. 21).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, 770-778 (vid. págs. 21, 31).
- Talo, M. (2019). Automated classification of histopathology images using transfer learning. *Artificial intelligence in medicine*, 101, 101743 (vid. pág. 21).
- Zafar, I., Tzanidou, G., Burton, R., Patel, N. & Araujo, L. (2018). *Hands-on convolutional neural networks with TensorFlow: Solve computer vision problems with modeling in TensorFlow and Python*. Packt Publishing Ltd. (Vid. pág. 23).
- Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2481-2495 (vid. pág. 22).
- Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. En: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, 234-241 (vid. págs. 23, 24).
- Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.*, 5, 1-34 (vid. pág. 30).
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297-302 (vid. pág. 30).
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. En: *Proceedings of the IEEE international conference on computer vision*. 2017, 618-626 (vid. pág. 32).

Parte II

Presupuesto

El documento de presupuestos tiene como objetivo realizar una estimación del coste del proyecto asociado al Trabajo Final de Grado (TFG) titulado "Diseño y desarrollo de un sistema de clasificación y detección de pólipos en imágenes endoscópicas con técnicas de aprendizaje profundo". Para facilitar la comprensión del presupuesto, se divide el proyecto en tres proyectos parciales que se corresponden a la mano de obra, el software y el hardware.

5.1 Presupuestos parciales

5.1.1 Costes de mano de obra

El desarrollo del proyecto del TFG ha estado llevado a cabo por el alumno y por un tutor. El alumno puede considerarse como un ingeniero biomédico junior y ha sido el encargado del desarrollo del grueso del proyecto y de la redacción de la memoria. Como el Trabajo de Final de Grado tiene un coste de 12 créditos ECTS que se corresponden a 25 horas lectivas, se estima que el tiempo empleado por el alumno para el desarrollo del proyecto es de 300 horas. Por su parte, el ingeniero biomédico senior ha sido el encargado de la tutorización del proyecto y de la revisión del trabajo. Según un informe del Colegio Oficial de Ingenieros Industriales de Bizkaia, se estipula que el coste unitario de la hora del ingeniero biomédico junior es de 15€/h y del senior es de 30€/h. Un resumen de estos costes se presenta en la Tabla 5.4.

Código	Descripción	Duración (h)	Coste unitario (€/h)	Coste total
ME.EST	Ingeniero biomédico junior	300	15	4500
MO.TUT	Ingeniero biomédico senior	75	30	2250
TOTAL				6750

Tabla 5.1: Resumen de los costes de mano de obra

5.1.2 Coste de software

El coste de software está asociado a las licencias de los sistemas informáticos y de los entornos de programación utilizados. Se emplea MATLAB (Mathworks [®]) en su versión R2021.a y que tiene una duración de un año. También se requiere el software MobaXterm Profesional que es un software que permite la computación remota para acceder a los servidores del laboratorio de manera segura y poder así ejecutar el entrenamiento de las redes neuronales en la GPU. Un resumen de esta sección del presupuesto se presenta en la Tabla 5.2.

Código	Descripción	Unidades (uds)	Coste unitario (€/uds)	Vida útil (meses)	Tiempo de uso (meses)	Coste total (€)
ME.MAT	MATLAB 2021.a	1	800	12	6	400
ME:MXP	MobaXterm Profesional	1	60	12	6	30
TOTAL						830

Tabla 5.2: Resumen de los costes de software

5.1.3 Coste de hardware

Se han empleado dos dispositivos de hardware para la realización del proyecto como son un ordenador portátil y una tarjeta gráfica o GPU. El ordenador portátil empleado tanto para la programación como para la redacción del proyecto es un Lenovo Legion Y520 que dispone de un procesador Intel Core®(7a generación) con una tarjeta gráfica NVIDIA GTX 1050. Sin embargo, la potencia de esta computador resultaba insuficiente para el entrenamiento de las redes neuronales por lo que se emplea la GPU del laboratorio que es una Titan XP cuya vida útil es aproximadamente de 4 años. Finalmente, se presenta el resumen de los costes de hardware en la Tabla 5.3.

Código	Descripción	Unidades (uds)	Coste unitario (€/uds)	Vida útil (meses)	Tiempo de uso (meses)	Coste total (€)
ME.PC	LENOVO LEGION Y520	1	1000	72	6	83.33
ME:GPU	Titan Xp NVIDIA	1	1349	48	6	168.63
TOTAL						251.96

Tabla 5.3: Resumen de los costes de hardware

5.2 Presupuestos totales

A modo de resumen del presente documento de presupuestos, se incluye en la Tabla ?? el presupuesto total del proyecto calculado como el sumatorio de los costes parciales recientemente presentados. El presupuesto proyectado es de **siete mil ochocientos treinta y un euros y noventa y seis euros**.

Sección del presupuesto	Coste (€)
Costes de mano de obra	6750
Costes de software	830
Costes de hardware	251.96
Presupuesto total	7831.96

Tabla 5.4: Resumen de los costes totales del proyecto

Índice alfabético

CNN Convolutional Neural Networks.

CVC CVC-ClinicDB.

DCT Discrete Cosine Transform.

ETIS ETIS-laryb Polyp DB.

FN Falso Negativo.

FP Falso Positivo.

Grad-CAM Gradient-weighted Class Activation Mapping.

HOG Histogram Oriented Gradients.

IoU Intersection over Union.

LBP Local Binary Patterns.

MLP Multilayer Perceptron.

OLR OverLap Ratio.

ReLU Rectified Linear Unit.

ROC Receiver Operating Characteristic Curve.

SSP Sessile Serrated Polyps.

tanh Hyperbolic Tangent.

TSA Traditional Serrated Adenomas.

VN Verdadero Negativo.

VP Verdadero Positivo.