Deep Learning en segmentación de imagen médica

Germán A. García Ferrando

Departmento de Sistemas Informáticos y Computación

Universitat Politècnica de València

Trabajo supervisado por

Dr. Carlos Monserrat Aranda

Dr. Juan Miguel García-Gómez

Julio 2017

Abstract

In the field of Medical Imaging, morphological segmentation is a first step in the treatment required in various clinical applications. Nowadays, in most centers this task is performed by an expert, which consumes a large amount of time and is prone to an inter and intra expert error. In this work, a model based on Deep Convolutional Networks, inspired by the already widely known encoder-decoder architecture, is developed. In addition, Long Residual Connections are used and also the Dice (F1-Score) is used as a loss function. The model is evaluated in two scenarios: firstly we perform prostate segmentation using the T2-weighted volumetric images, acquired by Magnetic Resonance, provided by the challenge PROMISE12 of the MICCAI; secondly, femoral segmentation is performed using the images provided by the Valencian company ERESA, acquired through X-rays. In the case of prostate, the results presented are competitive with the state of the art, obtaining a Dice of 85.83, which places our model close to human error; in femur, the segmentations reach a Dice of 94.35, considerably good results taking into account the quality and quantity of the images supplied. In conclusion, this paper presents a model based on Deep Learning that is able to obtain morphological segmentations of medical images using a reduced data set; thus guaranteeing the potential uses of this technique in the clinical context.

Resumen

En el campo de la Imagen Médica, la segmentación morfológica constituye un primer paso dentro del tratamiento requerido en diversas aplicaciones clínicas. A día de hoy, en la mayoría de los centros esta tarea la realiza un experto manualmente, lo cual consume una gran cantidad de tiempo y es propenso a un error inter e intra experto. En este trabajo se propone y desarrolla un modelo basado en Redes Convolucionales profundas, inspirado en la ya ampliamente conocida arquitectura *encoder-decoder*; además, se hace uso de *Long Residual Connections* y se utiliza el Dice (*F1-Score*) como función de pérdida. El modelo es evaluado en dos escenarios: por un lado, se realiza segmentación de próstata haciendo uso de las imágenes volumétricas potenciadas en T2, adquiridas por Resonancia Magnética, proporcionadas por el *challenge* PROMISE12 del MICCAI; por otro lado, se realiza segmentación de fémur utilizando las imágenes proporcionadas por la empresa valenciana ERESA, adquiridas mediante Rayos-X. En el caso de próstata, se presentan resultados competitivos con el estado del arte actual, obteniendo un Dice de 85.83, lo que sitúa a nuestro modelo cerca del error humano; en segmentación morfológica de fémur, las segmentaciones alcanzan un Dice de 94.35, resultados considerablemente buenos teniendo en cuenta la calidad y cantidad de las imágenes suministradas. En conclusión, en este trabajo se presenta un modelo basado en *Deep Learning* que es capaz de obtener segmentaciones morfológicas de imágenes médicas haciendo uso de un conjunto de datos reducido; avalando así los potenciales usos de esta técnica en el contexto clínico.

Resum

En el camp de la Imatge Mèdica, la segmentació morfològica constituïx un primer pas dins del tractament requerit en diverses aplicacions clíniques. A hores d'ara, en la majoria dels centres esta tasca la realitza un expert manualment, cosa que consumix una gran quantitat de temps i és propensa a un error inter i intra expert. En aquest treball es proposa un model basat en *Deep Convolutional Neural Networks*, inspirat en la ja àmpliament coneguda arquitectura *encoder-decoder*; a més, es fa ús de *Long Residual Connections* i s'utilitza el Dice (*F1-Score*) com a funció de pèrdua. El model és avaluat en dos escenaris: d'una banda, es realitza segmentació de pròstata fent ús de les imatges volumètriques potenciades en T2, adquirides per Ressonància Magnètica, proporcionades pel *challenge* PROMISE12 del MICCAI; d'altra banda, es realitza segmentació de fèmur utilitzant les imatges proporcionades per l'empresa valenciana ERESA, adquirides per mitjà de Raigs-X. En el cas de pròstata, es presenten resultats competitius amb l'estat de l'art actual, obtenint un Dice de 85.83, la qual cosa situa el nostre model prop de l'error humà; en segmentació morfològica de fèmur, les segmentacions aconseguixen un Dice de 94.35, resultats considerablement bons tenint en compte la qualitat i quantitat de les imatges subministrades. En conclusió, en aquest treball es presenta un model basat en *Deep Learning* que és capaç d'obtindre segmentacions morfològiques d'imatges mèdiques fent ús d'un conjunt de dades reduït; avalant així els potencials usos d'esta tècnica en el context clínic.

Tabla de Contenidos

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

Hoy poy hoy podemos encontrar aplicaciones de *Machine Learning* en la mayoría de los engranajes que componen la sociedad moderna, utilizándose este en, por ejemplo, sistemas de reconocimiento del habla [31, 62], sistemas de ayuda al diagnóstico médico [43] o sistemas de reconocimiento de imágenes [79], entre otros. Y, aunque el *Machine Learning* cuenta con más de medio siglo de historia, ha sido en esta última década, gracias al *big data* y la creciente capacidad de cómputo [82], cuando se han roto récords de una manera sin precedentes.

En 2012 presenciamos cómo el trabajo presentado por Alex Krizhevsky et al. [44], lograba reducir el error de clasificación de imágenes 10 puntos (ILSVRC-2012), alcanzando un error de 15.3% en top-5; cuatro años más tarde, Kaiming He et al. [30] reduciría el error a 4.94%, por debajo del error humano [69]. En 1996, algunos privilegiados pudieron observar como Deep Blue [10] era capaz de ganar, por primera vez, una partida al ajedrezista G.M. Garri Kaspárov, aunque el resultado final fuese 4-2 a favor del humano; veinte años más tarde, la Inteligencia Artificial (IA) AlphaGo [72] sería capaz de vencer por 4-1 al campeón mundial Lee Sedol, alcanzando así la primera posición del Ranking Mundial de Rémi Coulum, con un ELO[1] de 3611. En 1988, Terry SejNowski presentaba NetTalk [71], una IA capaz de aprender vocablos emulando el aprendizaje de un bebé; treinta años después Google cuenta con una IA capaz de traducir a más de 104 idiomas [83], y se están publicando resultados muy prometedores en el área del procesamiento y comprehensión del lenguaje [15].

Todos estos retos podrían ser puntos inconexos dentro de una disciplina que contiene una gran cantidad de técnicas; pero quizás, el hecho más sorprendente, es que estos logros

[1] Sistema de puntuación basado en cálculo estadístico, para calcular la habilidad relativa de los jugadores de deportes como el ajedrez.

comparten un denominador común: todos utilizaron *Deep Learning*.

El *Deep Learning* es, en esencia, una extensión de las Redes Neuronales Artificiales, las que a su vez son, una extensión del Perceptrón presentado por Frank Rosenblatt [65] en 1958. La novedad que aporte esta técnica es la capacidad de entrenar con éxito refes muy profundas, de ahí el nombre. Se puede clasificar al *Deep learning* como un método de representación-aprendizaje [45], el cual, mediante la composición de módulos no lineales, es capaz de aprender distintas representaciones de los datos. De hecho, el punto clave de esta técnica es que no requiere de un experto humano que especifique dichos módulos, sino que estos, a través de los datos, son capaces de aprender la representación de los mismos que optimize la posterior tarea, ya sea clasificación o regresión.

Desde un punto puramente Informático, en este trabajo estudia el comportamiento de redes profundas cuando se aplican a tareas donde no se dispone de una gran cantidad de datos. Para ello, se aplicarán Redes Convolucionales [46, 47] a problemas de segmentación semántica, con el objetivo de conseguir resultados suficiente precisos para uso clínico. Dichas redes serán entrenadas desde cero, con el fin de demostrar que el *pre-training* [32] no es necesario si se inicializan bien los pesos y se utilizan técnicas para evitar el desvanecimiento del gradiente.

Desde un punto de vista clínico, en este trabajo se plantean dos objetivos: (1) desarrollar un modelo capaz de realizar segmentación de volúmenes de próstata con suficiente precisión para que dichas segmentaciones puedan llegar a utilizarse como un primer paso en análisis de imagen, o que puedan utilizarse como un servicio secundario para profesionales de la imagen médica, y, (2) desarrollar un modelo capaz de realizar segmentaciones de fémur con suficiente precisión para que posteriormente se realize un estudio clínico sobre el mismo.

1.1 Motivación

La capacidad que está demostrando el *Deep Learning* para resolver problemas, especialmente en el campo de la Visión por Computador, representa una oportunidad fantástica para transladar esta tecnología al campo de la Imagen Médica. Tanto la Resonancia Magnética (RM), los Rayos-X o la Tomografía Axial Computarizada (TAC), son técnicas de adquisición de imágenes médicas que permiten realizar un estudio del estado de órganos, tejidos y huesos. En este trabajo, se aborda una de las tareas básicas dentro de la Imagen Médica:

segmentación morfológica, centrándonos especialmente en segmentación de próstata y fémur.

La segmentación de próstata en imagen de Resonancia Magnética es un área de interés para la comunidad científica debido ha que el uso de esta técnica facilita considerablemente el manejo de pacientes con cáncer de próstata [52, 78]. Dicha segmentación resulta útil por varios aspectos: permite aplicar la radioterapia con mayor precisión, realizar un seguimiento de la enfermedad y definir hábitats en la región del tumor y su periferia. Dado que la segmentación manual representa una tarea temporalmente costosa y está sujeta a un error intra-inter-observador [52], en este trabajo se aborda la construcción de un modelo capaz de realizar segmentaciones de forma completamente automática, para posteriormente ofrecerse como un servicio gratuito dentro del portal MTSImaging [38] el cual permitirá acceso totalmente gratuito a cualquiera que necesite utilizar el modelo.

Por otro lado, la segmentación de fémur en imágenes obtenidas por Rayos-X constituye el primer paso para la realización de diversos estudios morfológicos sobre el mismo. A modo de ejemplo, ya son varios los estudios [35, 74, 77] que se dedican a analizar las diferencias existentes entre las distintas poblaciones humanas alrededor del globo terráqueo, permitiendo mejorar las cirugías realizadas sobre el fémur, o cerca del mismo, y a su vez, permitiendo diseñar prótesis especializadas para una determinada etnia. Por otro lado, también se están realizando estudios que permiten estimar la probabilidad de rotura del hueso, para así efectuar un seguimiento más de cerca al paciente.

La principal motivación de este trabajo es resolver el problema de la segmentación morfológica, la cual consume una gran parte del tiempo de un estudio clínico y además requiere de un experto para que se pueda llevar acabo. Si los equipos clínicos dispusiesen de un *software* capaz de segmentar en un tiempo razonable, grandes cantidades de datos, la robustez de los estudios realizados por los mismos aumentaría considerablemente, lo cual repercute directamente en la calidad de los servicios médicos.

Nuestro principal argumento a la hora de diseñar una solución para estos problemas mediante Redes Convolucionales, es que nos permiten abordar distintas tareas con un mismo modelo, siempre y cuando se disponga de un corpus de entrenamiento representativo de las muestras que se desean segmentar. Esta potencia nos permitiría ofrecer un servicio de segmentación *ad hoc*, siempre que este nos facilitara un corpus supervisado con el elemento que se desea segmentar.

1.2 Proyectos y Colaboradores

Este trabajo se ha realizado con una colaboración constante con el grupo de Informática Biomédica (IBIME), el cual ha ofrecido los medios que lo han hecho posible. Durante dicha colaboración, se ha trabajado en un módulo del servicio MTSImaging, una plataforma *online* que pone a disposición de cualquiera métodos de segmentación morfológica de Glioblastoma, junto a un análisis hemodinámico, el cual hace uso de imágenes de perfusión para generar un mapa de segmentación nosológica de los habitats vasculares del Gliblomastoma. Todo el análisis realizado por la plataforma se le comunica al usuario mediante un reporte radiológico. Actualmente se está trabajando en la integración de un servicio que realize segmentación de próstata, utilizando el modelo presentado en este proyecto.

Tal como se ha comentado anteriormente, la segmentación morfológica de próstata constituye solo un paso inicial en el tratamiento contra el cáncer, por lo cual, este trabajo representa el primer paso dentro de un proyecto con miras más ambiciosas, MTS4up. Dicho proyecto, presentado por el Dr Juan Miguel García-Gómez, director de la rama de *Biomedical Data Science Lab - IBIMIE* del instituto ITACA (UPV). En dicho proyecto, se continuará con la investigación realizada por Javier Juan Albarracín [39], realizando así un estudio sobre la identificación de regiones de interés obtenidas mediante imágenes multiparamétricas, que permitan mejorar los tratamientos actuales contra el cáncer.

Paralelamente, en este trabajo también se ha colaborado con la empresa ERESA, empresa de referencia dentro de la Comunidad Valenciana por su trabajo en el campo de la Radiología. ERESA ha mostrado interés en el desarrollo de un *software* capaz de realizar segmentación morfológica de fémur, y en consecuencia, ha participado en este proyecto mediante la cesión del corpus de imágenes de fémur adquiridas mediante Rayos-X. En este caso, hemos aprovechado la colaboración con la empresa para colaborar a su vez con el departamento de Biomecánica de la Escuela Técnica Superior de de Ingenieros Industriales de la UPV. Así pues, la segmentación morfológica representa el primer paso de un proceso el cual tiene como objetivo realizar un estudio de Densitometría Ósea, el cual permite predecir si existe riesgo de rotura ósea para un determinado paciente, y en caso de que así sea, qué región del hueso es la más propensa a sufrir dicha lesión. Por lo tanto, esta parte del trabajo se complementa con el estudio realizado por el Dr Carlos Monserrat Aranda y la investigadora Sandra Martínez Sanchis.

Capítulo 2

Revisión Tecnológica

Dado que varias de las técnicas utilizadas en este trabajo son relativamente nuevas, este capítulo presenta una breve descripción de la idea intuitiva que persiguen dichas técnicas, junto a una breve presentación del marco teórico que las sustenta. Se asume que el lector está familiarizado con los conceptos básicos de las Redes Neuronales Convolucionales.

2.1 Generando segmentaciones a nivel de pixel

Dentro de la clasificación de imágenes existen tres grandes tareas: Clasificación, Localización y Segmentación (ver Figura 2.1). Cada una de estas tareas requiere una predicción distinta, en el caso de la clasificación (o clasificación débil), la salida de la red suele ser un vector de probabilidades de dimension C, siendo este el número total de clases. Para el caso de la Localización existen varias formas de enfocar el problema, a modo de ejemplo una posible solución sería la salida de la red sea un vector v, el cual representa las coordenadas del rectángulo que contiene al objeto que se desea localizar. Por último, en la tarea de segmentación, conocida también como segmentación semántica, la salida de la red debe ser una máscara del mismo tamaño que la imagen original, donde el valor de cada pixel representa la clase a la que pertenece esa porción de la imagen.

En este trabajo nos centramos en la tarea de segmentación, por lo que estaremos realizando segmentación a nivel de píxel de tanto de imágenes como de volúmenes. Ahora bien, ¿cómo se aborda este problema utilizando redes neuronales? Lo normal es optar por uno de estos dos enfoques:

- **Segmentación por *patches***, donde la entrada de la red es una ventana de la imagen, y la salida es la clasificación del pixel central de dicha ventana.

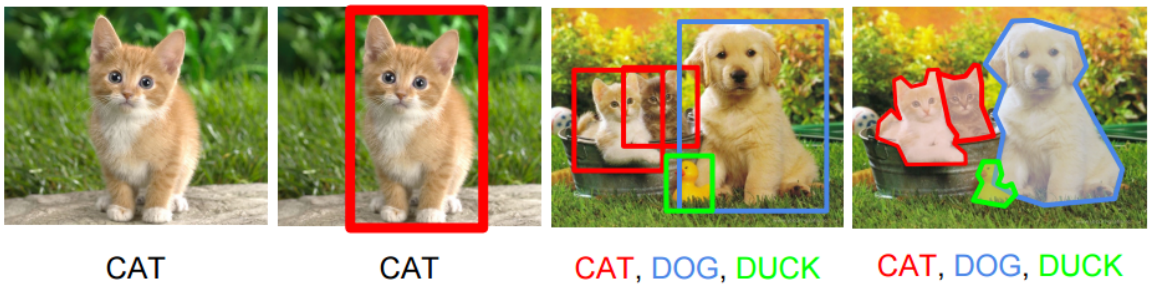


Fig. 2.1 Ilustración de las tres tareas básicas en clasificación de imágenes. De izquierda a derecha: Clasificación; Clasificación + Localización; Clasificación + Localización de múltiples objetos y Clasificación + Segmentación de múltiples objetos. Agradecimientos por la figura a Standford University.

- **Segmentación directa**, donde la entrada de la red es la imagen en su totalidad, y la salida es la segmentación de toda la imagen.

Trabajar por patches no es algo nuevo. Esta técnica ya se utilizaba en el campo del reconocimiento facial [66, 81], donde para detectar si una subregión de la imagen contiene una cara o no, se recorre una ventana por toda la imagen, y por cada subregión se clasifica como cara o no cara. En el caso de la segmentación, también deslizamos una ventana por toda la imagen, pero por cada subregión lo que se predice es la clase del pixel central de dicha ventana, tal como se representa graficamente en la Figura 2.2. El principal inconveniente de este método es que tiene un mayor coste computacional comparado con la segmentación directa, además, aprovecha menos la paralelización que ofrecen las GPUs. Aún así, se han reportado varios resultados prometedores utilizando este enfoque [13, 21, 24].

Por otro lado, se puede diseñar la red de tal forma que la salida del modelo sea directamente la máscara que deseamos predecir (ver Figura 2.3). Para esto, es necesario que la red tenga como salida un vector multidimensional con las mismas dimensiones que tiene vector de entrada. Es decir, tiene que predecir una máscara del mismo tamaño que la imagen que se desea segmentar. Para ello, podríamos pensar en una red que realizase únicamente convoluciones, y aprovechase el *padding* para mantener siempre la misma dimensionalidad. El principal inconveniente es que si se desea utilizar alguna de las arquitecturas que han tenido éxito en visión por computador [44, 73], es necesario resolver el problema de la dimensionalidad de la salida, ya que estas arquitecturas están diseñadas para predecir un vector de probabilidadades en vez de una máscara de segmentación. Este problema se aborda en las secciones 2.2 y 2.3.

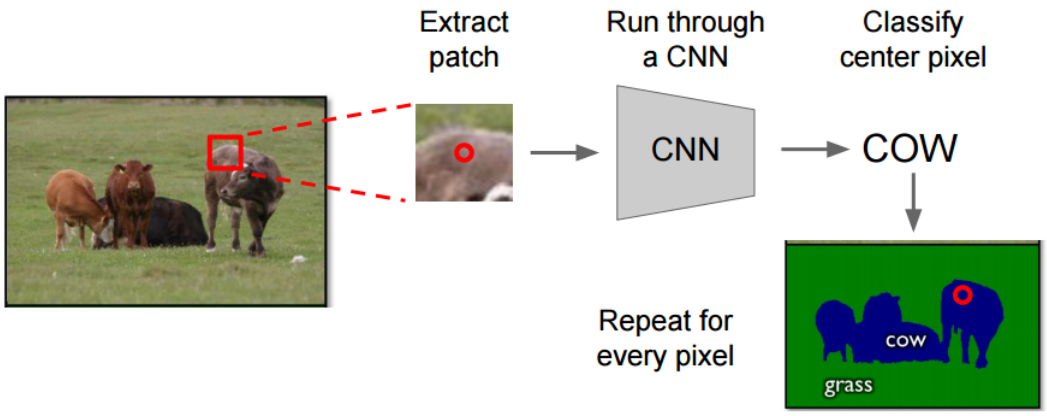


Fig. 2.2 Esquema básico de segmentación por *patches*. Se extrae un *patch* de la imagen con una ventana de tamaño fijo, este *patch* se pasa como entrada a la red, la cual emite una clasificación para el pixel central. Este proceso se repite hasta que todos los píxeles de la imagen estén clasificados. Agradecimientos por la figura a Standford University.

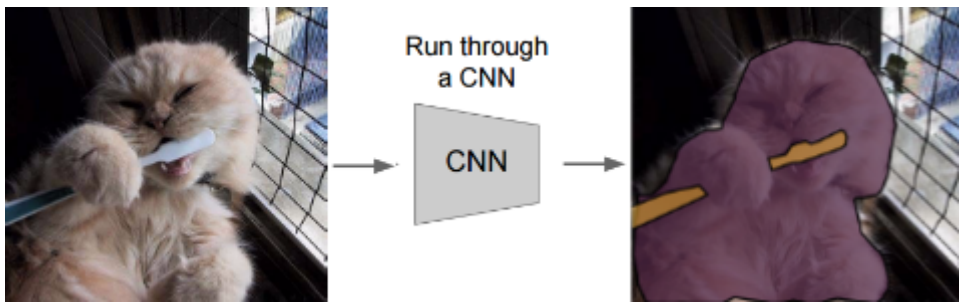


Fig. 2.3 Esquema básico de segmentación directa. La red recibe la imagen que se desea segmentar, y su salida es la máscara de segmentación.

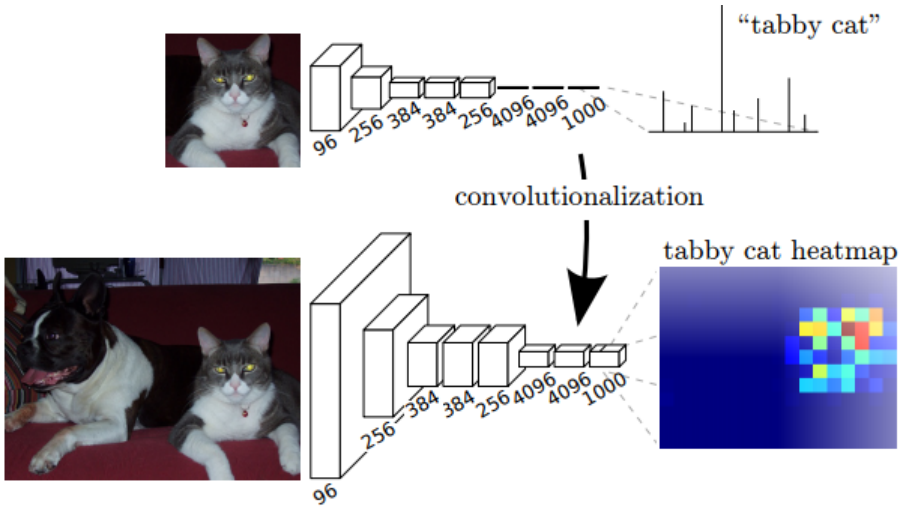


Fig. 2.4 En la parte superior, una representación de la salida del MLP clásico, un vector con la probabilidad de que la imagen pertenezca a cada clase. En la parte inferior, la salida de una *Fully Convolutional Network*[53], un mapa de probabilidad.

En este trabajo se ha optado por construir una red capaz de realizar un mapeo directo de imagen original a segmentación. Esto no es más que una mera decisión de diseño, ya que ambos enfoques son capaces de producir cualquier distribución [53].

2.2 Fully Convolutional Networks

A nuestro saber, el primer trabajo que logra adaptar redes convolucionales como la *VGG* [73] para que produzcan una segmentación directa, y además entrenando las redes *end-to-end*, es el trabajo presentado por Jonathan Long, Evan Shelhamer y Trevor Darrell en 2015, donde se da a conocer por primera vez las *Fully Convolutional Networks* [53]. La gran novedad es que por primera vez se presenta una red cuya unidad básica de cómputo es la convolución, es decir, el clasificador final deja de ser un MLP. Esta transformación permite que la red, en vez de calcular un único vector de probabilidad por imagen, pase a calcular un vector de probabilidad para cada píxel de la imagen, lo cual se puede interpretar como un mapa de probabilidad, tal y como se puede observar en la Figura 2.4.

Obtener este mapa de probabilidad es la parte fundamental de la segmentación directa, a partir de aquí solo debemos elegir las clases con mayor probabilidad por cada pixel, y ya habríamos generado una segmentación válida. Pero para que la segmentación se pueda aplicar sobre la imagen original, dicho mapa debe tener las mismas dimensiones que la imagen, por lo que aún se debe realizar un *upsampling* de las dimensiones del mapa. En este

trabajo se proponen dos métodos, bien interpolatción lineal, o bien realizar una convolución transpuesta; ambos métodos son explicados con mayor profundidad en la sección 2.3. El resultado es que cualquiera de estos dos métodos realiza una segmentación demasiado tosca (lo cual es comprensible ya que cuando se realiza interpolación se está inventando mucha información entre dos puntos, y esto repercute directamente en el nivel de detalle que posee la imagen interpolada).

Los autores del trabajo se percataron de este fenómeno, y ofrecieron una solución que sería la base de las futuras arquitecturas *encoder-decoder* [3, 64]. En esencia, el problema reside en que cada vez que se reduce la dimensionalidad de la imagen con operaciones como el *MaxPooling*, se está perdiendo la información de posición. Esto siempre se ha considerado una ventaja, debido a que permite que las redes sean robustas a ligeras translaciones; pero en el caso de la segmentación nos es de interés el "dónde", ya que necesitamos saber en qué posición ocurrió un determinado patrón para ser capaces de definir fronteras nítidas. Una vez se obtiene el mapa de probabilidad, es necesario volver a las dimensiones originales, para lo cual se realiza *upsampling*, y si el salto que este realiza es demasiado grande, se perderá cualquier nitidez en la segmentación.

Frente a este problema, los autores optaron por combinar capas de mayor dimensionalidad, con capas de menor dimensionalidad a través de conexiones *skip* [7]; estas conexiones permiten realizar segmentaciones a diferentes escalas, lo cual permite combinar segmentaciones realizadas a varios niveles. Podríamos interpretar este esquema de la siguiente manera: las segmentaciones de mayor dimensionalidad contienen altas frecuencias, mientras que las segmentaciones realizadas a menor dimensionalidad contienen bajas frecuencias; a la hora de fusionar estas segmentaciones, la interpolación se realiza de forma progresiva, siendo así el resultado final una segmentación mucho más nítida. Una comparativa entre este método y el original se puede observar en la Figura 2.5.

La idea de fusionar distintas segmentaciones a distintas resoluciones, es la base del modelo propuesto en este trabajo (Capítulo 3), donde se utilizarán *long skip connections*, junto a una arquitectura *encoder-decoder* para potenciar la reconstrucción de las segmentaciones con distinta dimensionalidad, y así conseguir una segmentación lo más nítida posible.

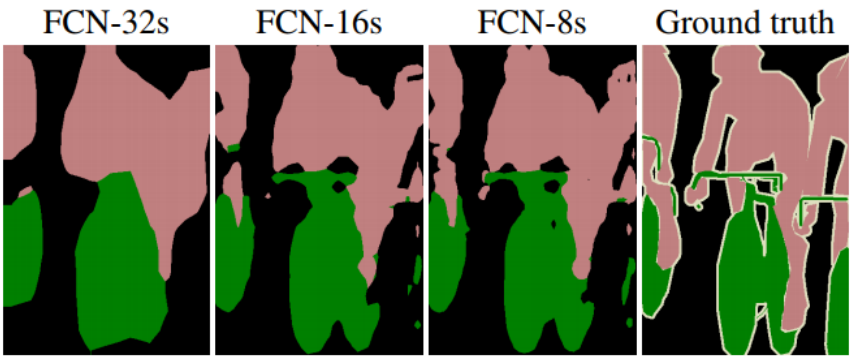


Fig. 2.5 Segmentaciones obtenidas utilizando distintas conexiones *skip*. FCN-32s no utiliza información de capas previas. FCN-16s utiliza información de una capa anterior. FCN-8s utiliza información de dos capas anteriores.

2.3 Upsampling: lineal, *splines* y convolución transpuesta

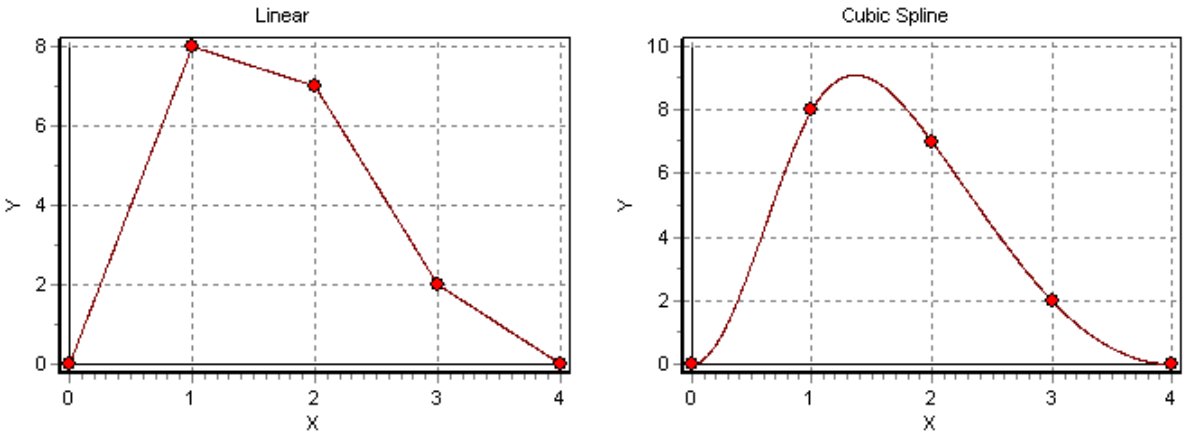
El objetivo del *upsampling* es pasar de una imagen I de dimensionaldad D, a una imagen I' de dimensionalidad D' siendo $D < D'$. Esto implica que para rellenar algunas partes de la imagen I', es necesario estimar información. Los distintos métodos de *upsampling* se diferencian principalmente en cómo se utiliza la información de la imagen original I, para construir I'. En esta sección abordaremos tres tipos de *upsampling*: mediante interpolación lineal, mediante interpolación por *splines* y *upsampling* mediante convolución transpuesta.

La **interpolación lineal** es el método más simple para reconstruir información. Dados dos puntos x_0 y x_1 en un espacio $\mathbb{R}$ la interpolación lineal reconstruye todos los puntos intermedios haciendo uso de la línea que une ambos puntos. La interpolación lineal de varios puntos, se define como la concatenación de la interpolación entre cada par de puntos (ver Figura 2.6). Dicho esto, podemos redefinir el problema como aproximar una función f, mediante una función interpolada g. Esta definición nos permite calcular el error E de la siguiente manera:

$$E = f(x) - g(x) \tag{2.1}$$

Esta definición nos permite hacer uso de las segunda derivada de la función f, y teniendo en cuenta el teorema de Rolle [63], se puede acotar el error de la siguiente manera:

$$E \leq \frac{(x_1 - x_0)^2}{8} \max_{x_0 \leq x \leq x_1} |f''(x)| \tag{2.2}$$



(a) Linear Interpolation

(b) Splines Interpolation

Fig. 2.6 Comparativa entre interpolación lineal e interpolación por *splines*.

De forma intuitiva, se aprecia que cuanto mayor es la curbatura de f, mayor es el error que generará la interpolación lineal. Este problema es el que se pretende solucionar con la interpolación por *splines*.

En el contexto matemático, las *splines* son funciones en las que cada segmento de las mismas está definido por un determinado polinomio. Es decir, sea S una función *spline* la cual está definida para el intervalo $[a,b]$, siendo a y b puntos en un espacio genérico D-dimensional.

$$S : [a,b] \rightarrow \mathbb{R} \tag{2.3}$$

Para que cada segmento esté definido por un polinomio, el intervalo $[a,b]$ se debe separar en n conjuntos disjuntos

$$[a,b] = [x_0,x_1] \cup [x_1,x_2] \cup ... \cup [x_{n-1},x_n] \tag{2.4}$$

donde cada uno de estos conjuntos estaría definido por un polinomio P_i

$$P_i : [x_i,x_{i+1}] \rightarrow \mathbb{R} \tag{2.5}$$

La **interpolación por *splines*** utiliza dichas funciones para interpolar cada intervalo (x,y) con aquel polinomio que minimice el error de reconstrucción. Como se puede observar en la Figura 2.6 la suavidad de la reconstrucción por *splines* es considerablemente mayor, que la reconstrucción por interpolación lineal. La dificultad de utilizar esta interpolación, reside

en que se deben hallar aquellos polinomios que minimicen el error de reconstrucción para cada función f. Este coste se puede asumir en muchos contextos, pero cuando queremos entrenar una red neuronal, cuyo coste computacional ya es suficiente alto de por sí, no resulta agradable tener que calcular los polinomios de las *splines* cada vez que se necesita realizar *upsampling* de un mapa. Además de esto, las *splines* son fantásticas para interpolar información con una continuidad natural, en cambio, los mapas que se encuentran dentro de la red convolucional no tienden a ser señales con una "forma" natural, sino que más bien se asemejan más a señales artificiosas con cambios bruscos de gradiente. Cuando se aplica interpolación mediante *splines* a señales con mucho gradiente, estas tienden a inventar mucha información para lograr una reconstrucción "suavizada". No conocemos ningún estudio que haya aplicado esta interpolación en los mapas intermedios de la red, por lo que, queda como trabajo futuro realizar un estudio que corrobore las hipótesis planteadas.

Las **convoluciones transpuestas**, también conocidas como "deconvoluciones", o convoluciones con *stride* fraccionario, son una forma de realizar *upsampling* mediante la operación convolución (ver Figura 2.7). La principal potencia de este método es que, los pesos del *kernel*[1] se pueden derivar a partir del *backpropagation* [67], por lo que la red es capaz de aprender a utilizar aquella interpolación que minimice la función de coste.

Además de la potencia que presentan las convoluciones transpuestas, también cuentan con la ventaja de que su aplicación resulta trivial. A la hora de realizar una convolución en una red, esta se codifica como una multiplicación de matrices, donde la imagen original se codifica como una matriz m de dimensiones $1 \times d$, y el kernel se codifica como una matriz C de dimensiones $d \times d'$, siendo d' el tamaño de la matriz convolucionada m_c. A la hora de aplicar el *backpropagation*, es necesario que el gradiente fluya, por lo que para pasar del mapa resultante de la convolución, m_c, al mapa original m, simplemente se multiplica la transpuesta de C por m_c , obteniendo como resultado una matriz con las mismas dimensiones que m. Por lo tanto, la implementación de esta operación es simplente invertir el orden de aplicación del *fordward* y el *backward*.

Esta técnica ha sido estudiada por varios trabajos [3, 53, 59] y es la opción estándar en redes convolucionales. En este trabajo hemos optado por aplicar convoluciones transpuestas como técnica para realizar *upsampling* de los mapas.

[1] En el ámbito de las Redes Convolucionales, al filtro con el que se convoluciona la imagen también se le conoce como *kernel*.

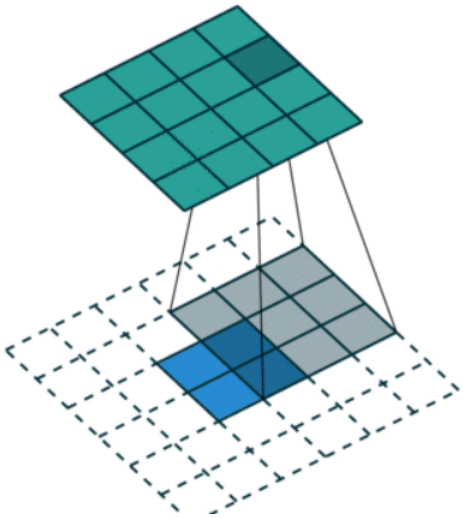


Fig. 2.7 En azul, el mapa original que se desea ampliar. En verde oscuro, el mapa resultante del *upsampling*. Los cuadrados sombreados hacen referencia al *kernel* con el que se convoluciona el mapa original. Los cuadrados blancos representan valores nulos que se añaden al mapa original para que se pueda realizar el aumento de dimensionalidad.

2.4 Residual Learning

Desde que Matthew D. Zeiler y Rob Fergus presentaron su trabajo *Visualizing and Understanding Convolutional Networks* [86], la idea de que las capas más profundas de las redes modelan información más compleja (ver Figura 2.8) se ha ido consolidando en la comunidad científica. Si a esto le unimos el hecho de que cuanto más profunda es una red, más facilidad tiene para aproximar cualquier distribución, todo apunta a que cuanto más profunda sea la red, mejores serán las resultados obtenidos; pero hasta hace poco, no es que esta hipótesis no se hubiera podido demostrar, sino que los resultado experimentales apuntaban a justamente lo contrario, redes profundas obtenían peores resultados que sus equivalentes más superficiales (ver Figura 2.9).

Este comportamiento es estudiado por Kaiming He et al. en su trabajo *Deep Residual Learning for Image Recognition* [30]. En dicho trabajo se apunta a que el hecho de que redes menos profundas consigan mejores resultados que sus hermanas más profundas, indica que no todos los modelos son igual de fáciles de optimizar. Este argumento se basa en el siguiente razonamiento:

"Considérese una red superficial y su contraparte más profunda, la cual únicamente añade más capas. Cabe la posibilidad de que a la hora de construir el modelo más profundo,

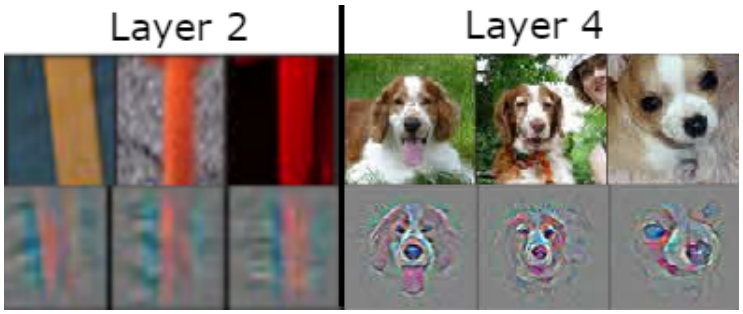


Fig. 2.8 En la fila superior se pueden observar las imagenes que ha recibido la red como entrada. En la fila inferior se muestra la reconstrucción de la respuesta de los *kernels* con mayor grado de activación a las imágenes de la fila superior. Crédito de la figura a Matthew D. Zeiler [86].

las capas extra se transformen en mapas que aplican la identidad, mientras que las capas previas se mantendrían exactamente igual que en la red menos profunda. La existencia de tal situación, indica que una red más profunda no debería obtener un error mayor en entrenamiento que su contraparte superficial. Como las evidencias indican lo contrario, es razonable asumir que los optimizadores actuales son incapaces de construir dicha solución, o al menos, no son capaces en un tiempo asumible."

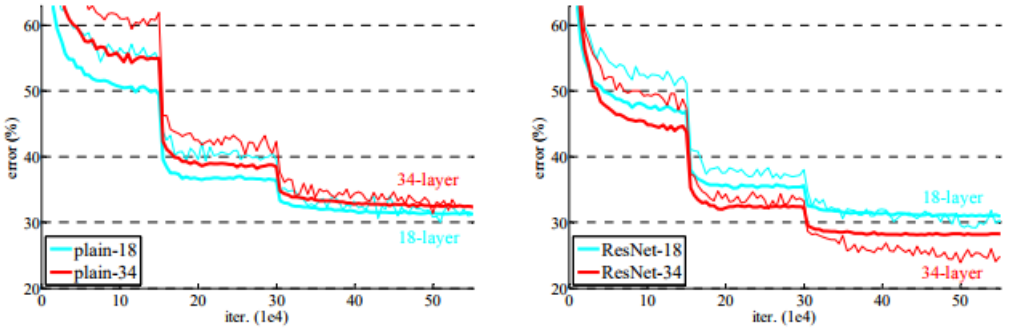


Fig. 2.9 A la izquierda el resultado de entrenar redes de 18 y 34 capas sin conexiones residuales. A la derecha, el mismo experimento pero añadiendo conexiones residuales.

Bajo esta hipótesis, los autores del trabajo proponen una modificación en la estructura de la red con el objetivo de facilitar la optimización de las mismas. Esta modificación se expresa de la siguiente forma:

Considérese $H(x)$ como la función aplicada por una serie de capas apiladas a una entrada x. Si suponemos que multiples capas no lineales pueden aproximar cualquier función de forma asintótica, entonces es equivalente suponer que dichas capas pueden aproximar fun-

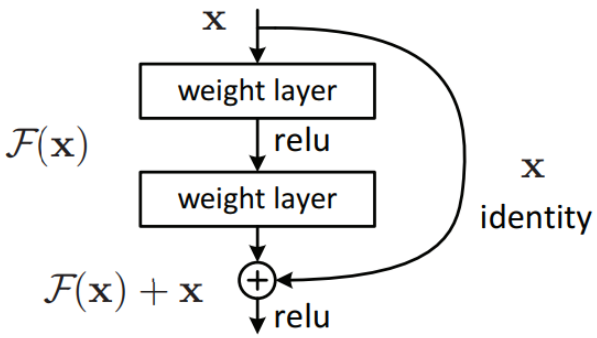


Fig. 2.10 Estructura básica del bloque residual. El mapa que pasa a aprender la red es $F(x)$.

ciones residuales tales como $H(x) - x$.

Si se asume la capacidad de las redes para aproximar cualquier función, lo cual todavía es una cuestión abierta [57], los autores proponen que en vez de aproximar una única función directa $H(x)$, las redes podrían aproximar una función residual tal que:

$$F(x) = H(x) - x \tag{2.6}$$

donde la función $H(x)$ ahora se redefiniría como:

$$H(x) = F(x) + x \tag{2.7}$$

Aunque ambas funciones deberían comportarse igual asintóticamente, puede que la dificultad de aprendizaje no sea la misma para ambas. Esta decisión de diseño está inspirada en el hecho de que las redes más profundas parecen tener problemas a la hora de aprender funciones que actuen como la identidad; al redefinir el aprendizaje de las funciones de esta forma, emular la identidad es tan simple como que la función $F(x)$ sea una función nula.

Esta modificación (ver Figura 2.10), bautizada como "conexión residual", no añade parámetros al modelo y facilita la optimización del mismo. Los autores del trabajo consiguieron el primer puesto en varias tareas de competiciones como *ILSVRC* 2015 [68] y *COCO* 2015 [50] presentando la red neuronal más profunda presentada hasta la fecha en este tipo de competiciones (152 capas). En este trabajo se hace uso de conexiones residuales "largas" [85], con el fin de facilitar la optimización del modelo propuesto.

Capítulo 3

Modelo Propuesto

En el siguiente apartado se detalla el modelo empleado para realizar las segmentaciones tanto en el caso de próstata como en fémur. Destacar que, uno de los objetivos de este trabajo es demostrar la capacidad de adaptación que presentan los modelos basados en Redes Convolutionales frente a nuevos problemas. Por lo tanto,se prima la generalización de la propuesta a la optimización particular de cada paso, lo cual nos ha privado de realizar optimizaciones para un problema en concreto. Aún así, tal y como se relata en la secciones 4.4 y 5.4, los resultados obtenidos son considerablemente buenos.

3.1 Arquitectura del modelo

Para diseñar el modelo nos hemos inspirado en la arquitectura *encoder-decoder*, ya validada por otros trabajos [3, 18, 59, 64, 85]. La idea que subyace a este tipo de arquitectura es, en esencia, facilitar la reconstrucción de la segmentación realizada en las capas más profundas de la red [53]. Esta arquitectura propone que la reconstrucción de la señal se realiza de forma gradual, por lo que una parte de la red *decoder*, se centra unicamente en utilizar las características extraidas por el *encoder*, para generar la segmentación más precisa posible.

Esta arquitectura también se podría interpretar bajo el enfoque de los *autoencoders*, o redes Diábolo [5, 8, 33, 34, 37, 67], donde el objetivo es entrenar una red para codificar una entrada, para luego ser capaz de decodificarla con el mínimo error de reconstrucción posible. Esta arquitectura permite que en el "cuello de botella" de la red, se construya una representación de la entrada cuya dimensionalidad es menor que la original; hasta este punto tanto los *autoencoders* como la arquitectura *encoder-decoder* realizan la misma función. La diferencia entre ambas arquitecturas reside en el objetivo de las mismas, en un *autoencoder* el decodificar se entrena para reconstruir la entrada, mientras que en una red *encoder-decoder*

Layer Name	**MeNet17**	**MeNet25**	Residual Connection
Convolution	1	2	
MaxPool	1	1	Con. 1
Convolution	1	2	
MaxPool	1	1	Con. 2
Convolution	1	2	
MaxPool	1	1	Con. 3
Convolution	1	2	
MaxPool	1	1	Con. 4
Upsampling	1	1	Con. 4
Convolution	1	2	
Upsampling	1	1	Con. 3
Convolution	1	2	
Upsampling	1	1	Con. 2
Convolution	1	2	
Upsampling	1	1	Con. 1
Convolution	1	2	
Convolution (1x1)	1	1	

Table 3.1 En la primera columna el tipo de capa utilizada. En las siguientes columnas se describe el número de veces que cada modelo hace uso de cada capa. En la última columna se detallan que pares de capas están conectadas mediante conexiones residuales.

el decodificador es entrenada para, haciendo uso de la información comprimida, construir una nueva señal objetivo (normalmente segmentaciones semánticas).

Con el fin de poder experimentar varias configuraciones sin tener que realizar todo el cómputo que conlleva un modelo final, se han diseñado dos redes, **MeNet17** y **MeNet25**, con una estructura idéntica pero distinto número de capas, en la Tabla 3.1 se detalla la estructura de las mismas.

Todas las convoluciones aplican un *kernel* 3×3, con el fin de que actuen como regularizadores y nos permitan agilizar los cálculos [73], además se utiliza *zero padding* para mantener la dimensionalidad del mapa. Todos los *Max Pooling* se efectuan con un *stride* 2×2, por lo que las dimensiones del mapa se reducen a la mitad un total de cuatro veces. No se ha especificado las dimensiones de entrada ya que el modelo está preparado para procesar cualquier tipo de entrada, siempre y cuando sus dimensiones sean divisibles entre 16. Después de cada convolución se aplica la función de activación *ReLU* [44, 58], para después aplicar *Batch Normalization* [36]; a excepción de la última convolución, la cual

va seguida de una *sigmoid* para que los valores que genere se puedan interpretar como una probabilidad. La inicialización de los pesos se realiza con una distribución gaussiana con media cero y desviación estándar $\sqrt{2/n_L}$ donde n_L es el número de neuronas de entrada de la capa L. Esta inicialización resulta una modificación propuesta por Kaiming He et al. [29] al estudio realizado por Xavier Glorot y Yoshua Bengio [25], donde se redefine la inicialización teniendo en cuenta el uso de funciones de activación *ReLU*, *PreReLU* o *ELU* [14, 29]. La red se ha entrenado utilizando Stochastic Gradient Descent (SGD) con un momentum de 0.9, un *learning rate* de 0.1 que se reduce por 10 cada 45 *epochs*.

Es bien conocido dentro de la informática médica el problema a la hora de conseguir datos, y este trabajo no es una excepción. Por lo tanto, también se ha utilizado regularización L2 [7] con un *weight decay* de 0.0005; no se ha utilizado *dropout* [76] siguiendo las recomendaciones de [36]. Para asegurar una buena generalización [23, 28] se ha decidido mantener un número de parámetros reducido, y aplicar *early stop* [84] a la hora de entrenar el modelo.

Por último, con el fin de aprovechar las ventajas propuestas por [30], brevemente resumidas en la sección 2.4, el modelo cuenta con conexiones residuales entre las capas de *upsampling* y *MaxPooling*, detalladas en la Tabla 3.1. Todo el modelo ha sido implementado utilizando TensorFlow [1].

3.2 Coeficiente Dice como función de pérdida

A la hora de realizar segmentación semántica un problema bien conocido es el desbalanceo que existe entre las diferentes clases; este problema surge cuando existen clases con una gran representación (p.e. el cielo) mientras otras tienen una representación muy reducida (p.e. señales de tráfico). Si este desbalanceo no se corrige, la red tenderá a modelar las clases con mayor representación, e ignorará las clases menos representativas, las cuales no tienen una gran repercusión en el error total. Con el fin de evaluar el comportamiento de los modelos frente a problemas con desbalanceo de clases se han propuesto métricas que no evaluan el error frente al total de datos (*accuracy*), sino que lo evaluan frente al total de datos de una única clase, estas métricas son el *recall*, *precision* y *F1-score* [61].

Destacar que existen varias soluciones a este problema, tal vez una de las más conocidas es utilizar *median frequency balancing* [20], lo que intuitivamente se interpreta como aumentar la penalización que se aplica cuando se falla en aquellas clases que cuentan con una menor representación, esta idea se traduce matemáticamente como un peso adicional en la función

de pérdida que pondera el error de una muestra en función de la clase a la que pertenece; un ejemplo de esta técnica sobre la función *cross-entropy* en la ecuación (3.1). Este método otorgará a las clases que tienen una gran representación un peso inferior 1, mientras que aquellas que tengan una menor representación tendrán un peso superior a 1.

$$H_w(p,q) = -\sum_x w(c_x)p(x)\log q(x) \tag{3.1}$$

Aunque este método ha sido ya validado varias veces [3, 64, 85], en este trabajo utilizamos el coefficiente Dice (*F1-score*) como función de pérdida, la cual, en el caso particular de clasificación binaria, se presenta como una función de pérdida independiente al desbalanceo de clases [55]. El coeficiente Dice entre dos volúmenes binarios puede escribirse de la siguiente forma:

$$D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{3.2}$$

Donde el sumatorio se realiza para todos los píxeles N, de la segmentación binaria generada por el modelo $p_i \in P$ y del *ground truth* $g_i \in G$. Como toda función de pérdida, es diferenciable (3.3) respecto a la salida de la red, y por lo tanto se puede utilizar mediante el *backpropagation*.

$$\frac{\partial D}{\partial p_j} = 2\left(\frac{g_j\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) - 2p_j\left(\sum_i^N p_i g_i\right)}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right)^2}\right) \tag{3.3}$$

Nótese que el dice está definido para segmentaciones binarias mientras que la salida de la red son probabilidades, por lo tanto, p_j será un valor en el intervalo $[0,1]$ mientras que g_j tomará como valores únicamente $(0,1)$, esto implica que el cuadrado de ambos dos no es calculado para mantener la relación entre nominador y denominador. Además de esto, para que la función pueda actuar como función de pérdida, se tiene que minimizar, por lo que hemos de invertir el resultado, pasando a ser -1 el mejor Dice posible (3.4).

$$D_{loss} = -\frac{2\sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i} \tag{3.4}$$

Por último destacar que el Dice no está definido cuando las segmentación del *ground truth* es nula. Durante la fase de experimentación de este trabajo se ha probado a modificar la función Dice para que se pueda aplicar cuando la segmentación objetivo es nula (3.5), aunque no se ha profundizado en la convergencia asintótica de esta función. En la Figura 3.1

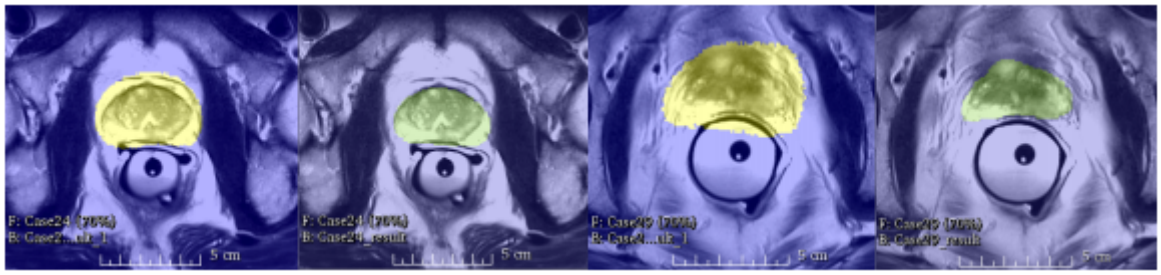


Fig. 3.1 En amarillo los resultados obtenidos con *weighted cross entropy*, en verde los resultados obtenidos utilizando el Dice.

se puede observar una comparativa entre resultados obtenidos utilizando el Dice y la *cross entropy* ponderada [55].

$$D_{loss} = -\frac{2\sum_i^N p_i g_i + \varepsilon}{\left(\sum_i^N p_i + \sum_i^N g_i\right) + \varepsilon} \tag{3.5}$$

3.3 Funcionamiento con muestras volumétricas

El Dice es una métrica que no está definida cuando el *ground truth* es nulo, lo cual se ha de tener en cuenta a la hora de trabajar con datos volumétricos. En el caso de la próstata (Sección 4), los datos son volúmenes, es decir, una composición de imágenes 2D; y en dichos datos, la próstata no tiene porqué aparecer en todos los cortes, por los que existirán imágenes en 2D que tendrán un *ground truth* nulo. Para poder aplicar el Dice como función de pérdida, hemos optado porque a la hora de trabajar con volúmenes, el modelo trabaje sobre los cortes 2D que componen como si fueran imágenes independiente, pero a la hora de calcular la funciónde pérdida, todos los cortes se interpreten como una única imagen tridimensional. Para que esto sea posible, es necesario que el *batch size* se ajuste de forma dinámica.

Así pues, cuando se activa el *flag* de imágenes en 3D, el modelo interpreta el batch entero como una única muestra volumétrica. Esto es posible gracias el tamaño de batch se ajusta en función a la profundidad que presenta cada volumen. Hecho esto, los distintos cortes se pasan a la red y son tratados como imágenes 2D independientes entre sí -ya que los *kerneles* son bidimensionales- pero cuando se calcula la función de pérdida, el batch entero se calcula como una única muestra. En consecuencia, aunque el tamaño de batch sea mayor que uno, los gradientes se actualizan como si solo se hubiese observado una única muestra; es decir,

cuando el modelo trabajo con volúmenes adopta un enfoque similar al propuesto por el aprendizaje *online*, donde los pesos se actualizan cada vez que se ve una muestra.

Capítulo 4

Segmentación de próstata en Resonancia Magnética

4.1 Introducción al problema

En 2012 se propuso por la sociedad *Medical Image Computing and Computer Assisted Intervention* (MICCAI) el desafío PROMISE12 [52], en el cual se invitaba a los participantes a diseñar el mejor modelo automático (o semi-automático) para segmentación morfológica en volúmenes de próstata, dichos volúmenes se han obtenido como imágenes transversales potenciadas en T2 mediante Resonancia Magnética (RM). En la Figura 4.1 se puede observar un corte transversal de un volumen de próstata, y su respectiva segmentación.

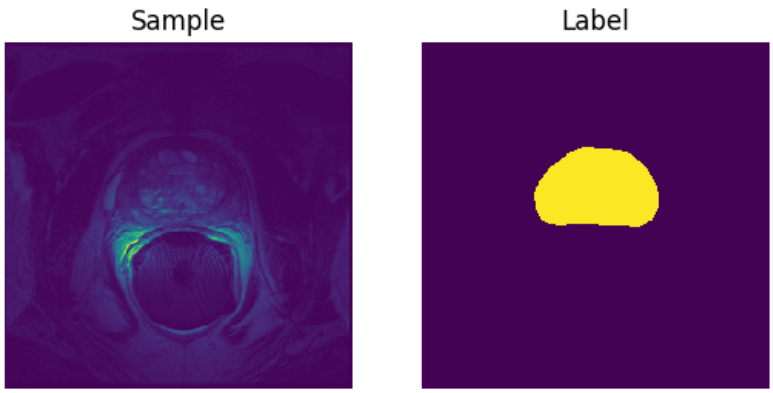


Fig. 4.1 A la izquierda un imagen de un corte transversal de una imagen potenciada en T2, extraído de uno de los volúmenes del corpus. A la derecha su correspondiente segmentación.

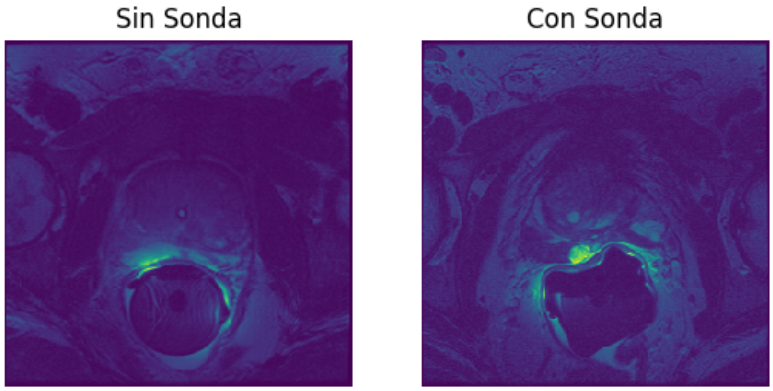


Fig. 4.2 A la izquierda un corte extraído de uno de los volúmenes del corpus adquirido sin sonda rectal. A la derecha otro corte obtenido con sonda rectal.

Los volúmenes se extrajeron tanto de pacientes con enfermedades benignas (Hiperplasia benigna de próstata) como de pacientes cáncer. Con la intención de valorar la robustez y capacidad de generalización de los algoritmos, los volúmenes fueron obtenidos de distintos centros, y también de distintos modelos de máquinas de RM. Además, en el corpus también se pueden encontrar distintos protocolos de escaneo, es decir, volúmenes tomados con o sin bobina endorectal (ver Figura 4.2).

El conjunto proporcionado por el *challenge* cuenta con un total de 50 casos, cada uno con su diferente dimensionalidad, tanto en anchura, como en altura, como en profundidad. El conjunto de entrenamiento se puede considerar representativo de las imágenes de próstata adquieridas mediante RM gracias a que es multi-centro, multi-marca y cuenta con distintos protocolos de adquisición; además, el conjunto está seleccionado de tal forma que haya una gran variedad de tipos de próstata, variando principalmente en tamaño y apariencia. Cada caso de entrenamiento es acompañado de una segmentación supervisada por expertos, la cual viene codificada como una máscara con valor 0 (fondo) o 1 (próstata) para cada voxel. Cabe destacar que la proporción de las dos clases, fondo y próstata, es de $98\% - 2\%$ respectivamente. Este desbalanceo se debe al hecho de que la próstata solo representa una pequeña región del volumen completo.

En este trabajo solo se ha evaluado el sistema utilizando únicamente el coeficiente Dice, dejando como posible trabajo futuro un análisis del modelo utilizando métricas como el 95% de la distancia Haussdorf o la diferencia media entre fronteras.

4.2 Trabajo relacionado

En 2012, año en el cual se presentó el desafío, se presentaron dos trabajos en concreto que se desmarcaron del resto de competidores, entre estos dos se encuentra el ganador de ese año Graham Vincent et al. [80] consiguiendo una puntuación[1] total de 84.36 utilizando *Active Appearance Models*. En segundo lugar, ese mismo año, Neil Birkbeck obtuvo [6] una puntuación de 83.49 utilizando *Discriminative Learning*.

A la vista de los resultados que se obtuvieron en 2012, podríamos considerar que el problema ya está solucionado, ya que las segmentaciones obtenidas por dichos modelos tienen una desviación estándar del $7-8\%$ [52], lo cual les permite generar mejores segmentaciones que un humano (estudiante de medicina en prácticas). A pesar de esto, la entrada del *Deep Learning* ha supuesto un punto de inflexión en las técnicas de Visión por Computador, por lo que los organizadores del desafío han decidido seguir aceptando propuestas con el fin de estudiar cuales son los resultados que son capaces de obtener estas nuevas técnicas.

Dentro de este grupo, donde se aborda el problema mediante Redes Convolucionales, es donde se sitúa el actual ganador del desafío [85], el cual hace uso de una red convolucional que trabaja por *patches* de $64 \times 64 \times 16$, lo cual les permite entrenar *kernels* que aprovechan información tridimensional; esto, junto a conexiones residuales para facilitar la optimización y clasificadores auxialiares [48] a modo de regularización, les sitúan en primer puesto con una puntación de 86.65. Junto a este trabajo encontramos el presentado por Michal Drozdzal et al. [18], situado en quinto lugar, con una red de 140 capas de profundidad y una puntuación de 83.02; cabe destacar que este trabajo se presentó a dos desafío además del que nos concierne. Por último, en sexto lugar se encuentra Fausto Milletari et al. [55], destacando poro ser este el primer trabajo (a nuestro saber) en introducir el coeficiente Dice como función de pérdida.

Con el fin de mantener el trabajo sencillo, y establecer una comparativa coherente con el trabajo aquí propuesto, nos limitaremos a comparar nuestros resultados con los dos mejores modelos del 2012, y los tres mejores trabajos que hacen uso de resdes convolucionales. En la Tabla 4.1 se puede observar una clasificación de los trabajos seis mejores propuestos hasta el momento.

[1] La puntuación se calcula comparando las segmentaciones del modelo frente a las segmentaciones óptimas, y las segmentaciones obtenidas por un humano. La métrica utilizada para comparar segmentaciones es el Dice.

Ranking	Trabajo	Fecha de presentación	Tipo	Puntuación
1	CUMED	08/14/2016	Automatic	86.65
2	Imorphics	06/29/2012	Automatic	84.36
3	Emory	07/21/2015	Semi-Automatic	83.66
4	ScrAutoProstate	07/02/2012	Automatic	83.49
5	UdeM 2D	01/02/2017	Automatic	83.02
6	CAMP-TUM2	06/06/2016	Automatic	82.39

Table 4.1 Ranking del desafío PROMISE12.

4.3 Pipeline propuesto

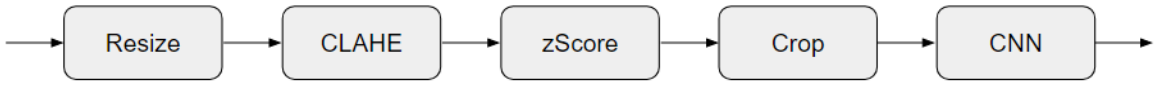


Fig. 4.3 De izquierda a derecha, los pasos aplicados a los volúmenes para obtener sus respectivas segmentaciones.

Con el fin de justificar cada uno de los pasos que componen el *pipeline* final, en este apartado se describe como se ha llegado a la construcción del mismo, empezando desde el planteamineto más sencillo, y comentando las distintas modificaciones que se fueron añadiendo. Los resultados de cada uno se muestran en la sección 4.4.

El **primer *pipeline*** con el que se abordó el problema fue, a nuestro parecer, el más sencillo que se podía establecer. Primero, cada corte se redimensionan [2] a 256×256 para asegurar que todas las segmentaciones se encuentran en el mismo espacio; las dimensiones elegidas se decidieron basandonos en aquellas que distorsionaran menos el *aspect ratio* y ofrecieran facilidades computacionales. Después de redimensionar las imágenes, se aplica un *zScore* robusto; este se realiza con el fin de agilizar el proceso de entrenamiento [25, 36, 44], se ha utilizado estadísticos robustos (mediana e intervalo intercuartílico) debido a la diferencia de intensidad que presentan las imágenes para un mismo tejido, debido al hecho de que las imágenes son tomadas por distintos modelos de máquina. Una vez realizado el preproceso, se realiza *Data Augmentation* mediante un *flip* de las imágenes, las cuales se introducen en la red convolucional para obtener las correspondientes segmentaciones.

Uno de los puntos débiles del primer *pipeline* es que a volúmenes radicalmente distintos (ver Figura 4.4) se les aplica el mismo preproceso. Con el fin de realizar un preproceso más personalizados, se realizón un análisis del corpus, llegando a clasificar las muestras en tres

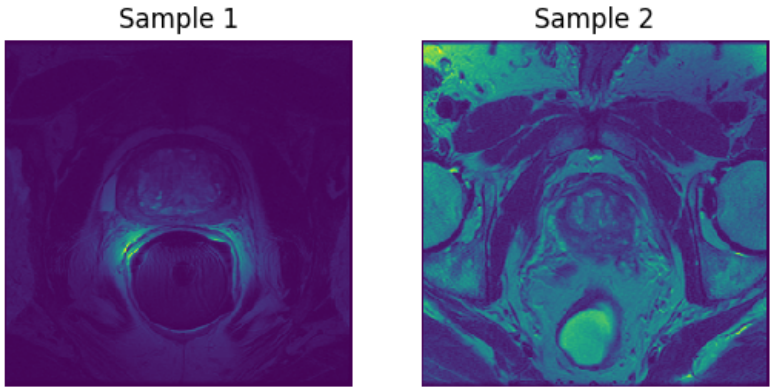


Fig. 4.4 A la izquierda un corte con gran heterogeneidad de intensidad para un mismo tejido (bias). A la derecha, un corte con menos bias pero mayor saturación en los niveles de intensidad.

grandes grupos en base a sus intervalos de intensidad. Esto permite aplicar un preproceso basado en *Contrast Limit Adaptative Histogram Equalization* (CLAHE) [60], y así homogeneizar las intensidades de todo el corpus (ver Figura 4.4). Por lo tanto, nuestro **segundo *pipeline*** antes de realizar el *zScore* de los volúmenes, aplica CLAHE.

Por último, cabe destacar que en todos los volúmenes la próstata se encuentra en el centro, pues cuando se toman imágenes en el contexto médico, siempre se toman con la intención de que el elemento que se desea observar se halle en el centro de la imagen, para así tener la región de interés en la zona de mayor homogeneidad de campo magnético. Por lo tanto, bajo nuestro punto de vista, es justificable asumir la centralidad de la próstata en todos los volúmenes. Esto nos permite realizar un *crop* manual, extrayendo así un volumen de menor tamaño en el cual se encuentra la próstata. Este paso reduce considerablemente la varianza de las imágenes, ya que todas las próstata comparten morfología, pero no todos los volúmenes recogen la misma información de contexto, esto añade una variabilidad superflua a las imágenes con las que el modelo tiene que aprender a lidiar.

Finalmente, el ***pipeline* definitivo** añade un *crop* centrado en el volumen pare reducir la dificultad del problema. El proceso completo se ilustra graficamente el la Figura 4.3.

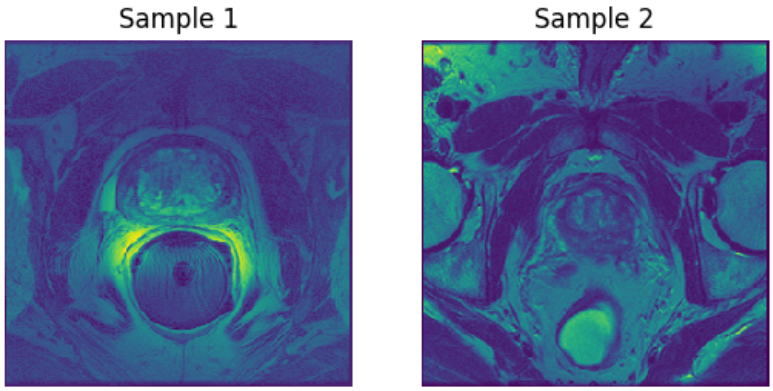


Fig. 4.5 Los mismos volúmenes que en la Figura 4.4 después de aplicar CLAHE.

4.4 Resultados

En la Tabla 4.2 de pueden observar los resultados obtenidos con el modelo **MeNet25** mediante los diferentes *pipelines*. Sorprende el resultado obtenido utilizando el *pipeline* más sencillo; pues este resultado sitúa al modelo entre los 8 mejores resultados, superando a otras redes convolucionales [4, 42].

Tal y como se argumentó en el apartado anterior, el hecho de aplicar CLAHE sobre el corpus, simplifica el problema. Esto se ve reflejado en una subida de casi 1 punto sobre el modelo simple.

Por último, observar la gran diferencia que existe entre el modelo que aborda el volumen completo, y el que solo aborda un volumen de menor dimensionalidad centrado en la imagen. Esto, tal y como se comentó en el apartado anterior, lo explica el hecho de que gran parte de la variabilidad de las imágenes no se encuentra en la próstata en sí, sino que se debe al contexto, el cual es muy variable debido al caracter multi-centro del corpus y a la morfología abdominal de los pacientes. En la Figura 4.6 se puede observar las segmentaciones obtenidas por este último *pipeline*

Con el fin de validar el funcionamiento del modelo, se ha realizado *ten-cross-validation* sobre todo el corpus supervisado (Tabla 4.3). Estos resultados, aunque no son definitivos ya que no se han realizado sobre el corpus de test, anuncian que nuestro modelo se acerca considerablemente al estado del arte, superando incluso el nivel humano; ya que, si los

MeNet25	Dice
Simple	82.04
Simple + CLAHE	82.97
Simple + CLAHE + crop	86.63

Table 4.2 Resultados de los distintos *pipelines* en segmentación de próstata. Los resultados son sobre un conjunto de validación seleccionado aleatoriamente.

Ranking	Método	Dice
1	CUMED	86.93
2	Imorphics(*)	85.57
?	Ours (Simple + CLAHE + crop)	85.83

Table 4.3 Resultados obtenidos mediante *cross-validation* en segmentación de próstata. (*) El resultado de Imorphics se ha extraido de los resultados facilitados por el desafío, este dice no está publicado en su trabajo.

resultados no varian, nuestro modelo se situaría por encima de Imorphics [80].

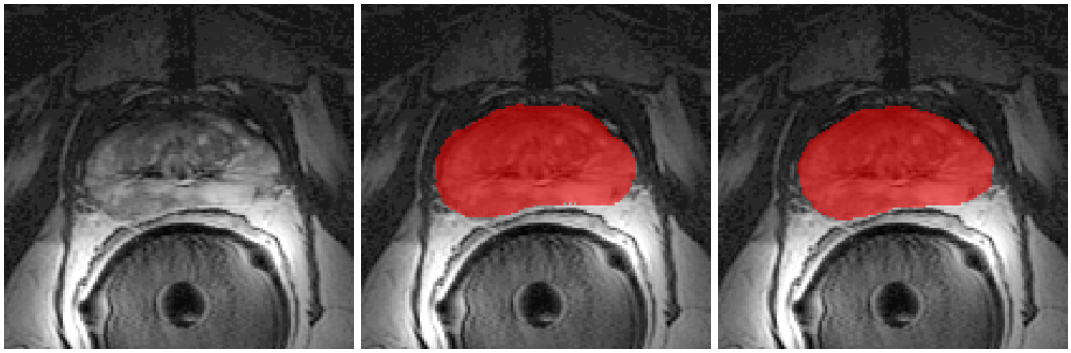
4.5 Discusión

A la vista de los resultados este trabajo presenta un modelo sencillo el cual, sin optimizar parámetros especificamente para este problema, consigue obtener resultados cercanos al estado del arte.

También creemos que el modelo presentado en este trabajo aún tiene recorrido de mejora. Técnicas que apliquen un post-proceso a la segmentación como seleccionar la mayor componente conexa [27] podrían aumentar el Dice obtenido, sin practicamente aumentar el coste temporal.

Otra posibilidad de mejora, en la cual tenemos intención de explorar en trabajos futuros, es explotar la información tridimensional. Actualmente, todos los *kernels* se aplican sobre un único corte del volumen, mientras que el conocimiento temporal (qué sucede en el corte siguiente y previo) no se aprovecha en absoluto.

Por último aunque el modelo aquí propuesto obtiene resultados considerablemente buenos, carece de resultados que avalen este comportamiento frente a un conjunto de test



(a) Validation Sample (b) Model Segmentation (c) Ground Truth

Fig. 4.6 Comparativa entre la segmentación de nuestro modelo (b) y el *ground truth* (c) para un corte del caso_04 del corpues propuesto en PROMISE12 (a).

totalmente ajeno. Por esta misma razón, entendemos que un trabajo pendiente antes de considerar nuestro modelo como definitivo, es presentar el modelo al desafío para contrastar los resultados, y, en caso de que se cumplan las espectativas, crear una plataforma accesible al modelo para que se pueda aprovechar su potencial.

4.6 Conclusiones

En este trabajo se ha presentado un modelo capaz alcanzar una precisión y exactitud similar al experto humano en segmentación de volúmenes de próstata. Se ha comprobado como aprovechar la centralidad de las imágenes captadas en el contexto médico, mejora considerablemente las segmentaciones, lo cual, junto a posibles técnicas de registro lineal para asegurar la centralidad de las imágenes, puede aportar mejoras considerables en distintos problemas. Dicho resultados, reafirman a las redes convolucionales como una de las mejores técnicas a la hora de abordar problemas relacionados con la Visión por Computador, incluso cuando el volumen de datos del que se dispone no es de una gran magnitud.

Capítulo 5

Segmentación de fémur en Rayos-X

5.1 Introducción al problema

La empresa Valenciana ERESA está interesada en el desarrollo de un *software* capaz de realizar una segmentación morfológica del femur. Este *software* se utilizará posteriormente para realizar un estudio sobre la morfología del mismo, y así ser capaces de predecir si existe el riesgo de rotura para un determinado paciente, y en caso de que así sea, ser capaces de predecir en qué zona es más probable que ocurra dicha rotura.

Con el fin de desarrollar este sistema, ERESA presta un conjunto de 101 imágenes de fémur adquiridas mediante Rayos-X. El corpus prestado no cuenta con la supervisión de un experto, la supervisión del mismo se ha realizado en este trabajo, y únicamente sobre 60 de las 101 imágenes prestadas, teniendo así 40 imágenes para el entrenamiento, 10 imágenes para validación y otras 10 imágenes para el test. Cabe destacar que, aunque gran parte de las imágenes cuentan con una calidad "aceptable" para realizar una segmentación, algunos de ellas llegan con una calidad pésima; en las cuales, incluso la segmentación manual ha sido resultado un desafío *per se* (ver Figura 5.1).

Por último, al ser este un problema de caracter privado, no existen trabajos con los cuales nos podamos comparar. La validez de este trabajo será corroborada cuando la empresa acepte como correctas las segmentaciones realizadas por el modelo.

5.2 Trabajo relacionado

Aunque ERESA no cuenta con trabajos previos relacionados con la segmentación del fémur, y tampoco hemos sido capaces de encontrar un desafío cuyo problema principal sea segmentación de fémur; sí que existen, dentro de la comunidad científica, distintos trabajos que abordan este problema con técnicas de Reconocimiento de Formas.

En 2013, C. Lindner et al. [51] presentó un modelo capaz de realizar segmentación de la cabeza del fémur utilizando *Random Forests* [9], los cuales se combinan con *Hough Forest* [19, 22] con el fin de obtener 62 puntos, los cuales se utilizan como la inicialización de un modelo de formas [16] que genera la segmentación final. Este trabajo se realiza sobre un corpus que cuenta con 839 imágenes supervisadas por expertos. En el trabajo se consigue un error de segmentación inferior a 0.9mm para el 99% de las imágenes.

Ese mismo año, Santhoshini, P et al. [70] presenta un sistema para el diagnóstico de la Osteoporósis mediante el análisis de imágenes de rayos-X. Para dicho análisis, es necesaria una segmentación del fémur, la cual los autores obtienen aplicando el filtro Canny (no se detallan más pasos). El trabajo se realiza sobre únicamente dos imágenes. No se detalla ninguna métrica relacionada con la precisión de la segmentación.

Por último, en 2017 Kayalıbay, B. et al. [41] presentaron un modelo bastado en redes convolucionales con una arquitectura *encoder-decoder*. Este trabajo participó en el desafío BRATS [54], en la tarea de clasificación morfológica de tumor (*whole, core y enhanced*. Además de la segmentación de tumor, el mismo modelo se testea en segmentación de Falange, en imágenes de Resonancia Magnética, obteniendo un Dice de 93 ± 1. Aunque el trabajo de Kayalıbay B. et al. no se enfoca en segmentación de fémur, nos permite tener una referencia sobre qué puntuaciones están obteniendo otras Redes Convolucionales en tareas similares a la aquí propuesta.

5.3 Pipeline propuesto

Se han diseñados varios *pipelines* con el fin de abordar el problema de la segmentación del fémur, compartiendo todos ellos un mismo preproceso. A continuación se detalla dicho preproceso, para despúes profundizar en cada uno de los enfoques.

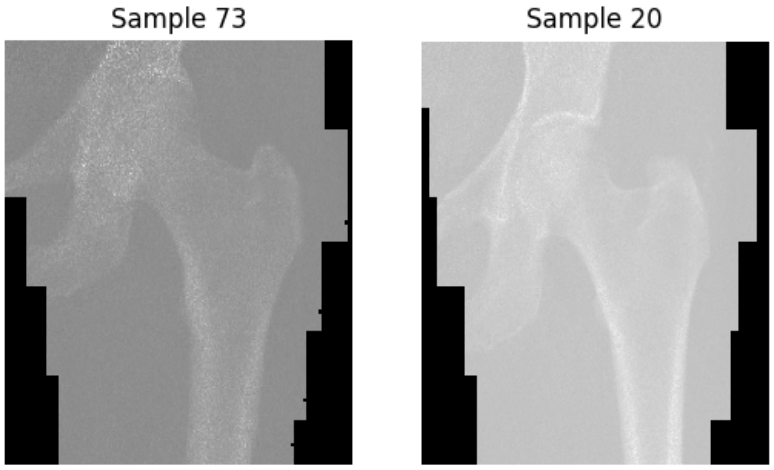


Fig. 5.1 Ambas muestras son la interpretación en crudo de las imágenes proporcionadas por ERESA.

El primer paso que se realiza es truncar la intensidad de las imágenes para eliminar el error sistemático generado por la máquina (ver Figura 5.2). En segundo lugar, las imágenes son redimensionadas mediante un registro lineal [2] a una plantilla generada previamente, con dimensiones 512×512; esta plantilla ha sido seleccionada de tal manera que sea una representación lo más cercana posible a la forma que generalmente presenta el fémur. Por último, a todas las imágenes se les aplica un *zScore* robusto, siguiendo el planteamiento propuesto en segmentación de próstata.

Una vez realizado el preproceso las imágenes están preparadas para entrar en cualquiera de los *pipelines*. El problema se aborda de tres formas distintas, las cuales se detallan a continuación:

- **Segmentación completa**:En el primer enfoque se pasan las imágenes preprocesadas directamente a la red, con el objetivo de segmentar toda la estructura ósea. Los resultados obtenidos, los cuales se detallan en la sección 5.4, muestran que la red tiene facilidad para segmentar el cuerpo del fémur, pero falla a la hora de segmentar la cabeza. Esto es comprensible, ya que en esa zona podemos encontrar la intersección de la cadera con el fémur, lo cual dificulta la detección del hueso; además de que suele ser la zona que presenta más ruido de toda la imagen.

- **Segmentación de la cabeza**: Con el fin de solventar los problemas a la hora de segmentar la cabeza del fémur, se volvió a supervisar el corpus pero esta vez segmentando

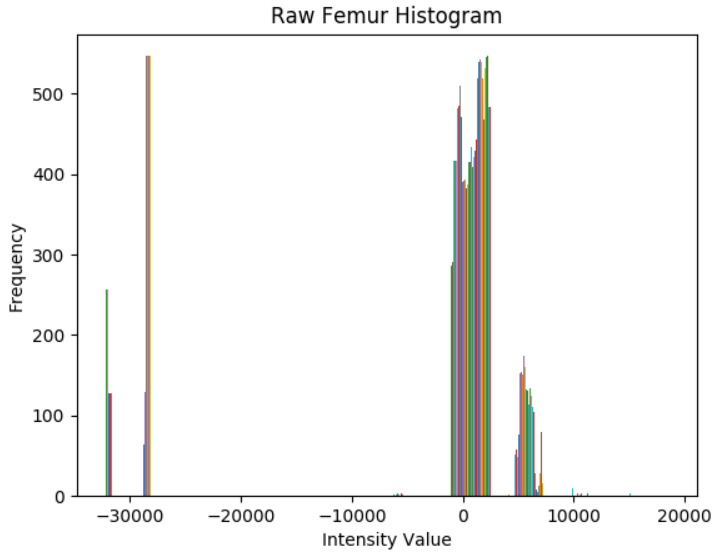


Fig. 5.2 Todas las imágenes proporcionadas comparten una intensidad centrada en -31000 y -29000; lo cual achacamos a un error sistemático generado por la máquina de Rayos-X.

únicamente la cabeza. Este cambio mejora los resultados, pero requiere de un método que añada la segmentación del cuerpo del femur.

- **Segmentación de la cabeza mediante ventana centrada**: Por último, basándonos en el mismo argumento propuesto en segmentación de próstata, realizamos un *crop* centrado en la cabeza del fémur para reducir la variabilidad del problema. Esto es posible debido a que, previamente, se ha realizado un registro lineal, cuyo objetivo es transladar, rotar y redimensionar las imágenes de tal forma que todas las imágenes se encuentren en la misma posición y en el mismo espacio.

De los tres *pipelines* propuestos, únicamente el primero resuelve el problema completamente, mientras que los otros dos se deberían combinar con el primero para ofrecer una segmentación completa del fémur. La heurística para combinar las segmentaciones es simplemente añadir la segmentación de la cabeza a la segmentación del cuerpo.

Como experimento adicional, se le añade a la red un mapa de probabilidad generado a partir de la plantilla a la que se han registrado todas las imágenes (ver Figura 5.3). El objetivo de añadir este mapa es, otorgar a la red información morfológica *a priori*, la cual fue clave durante el proceso de etiquetado del corpus.

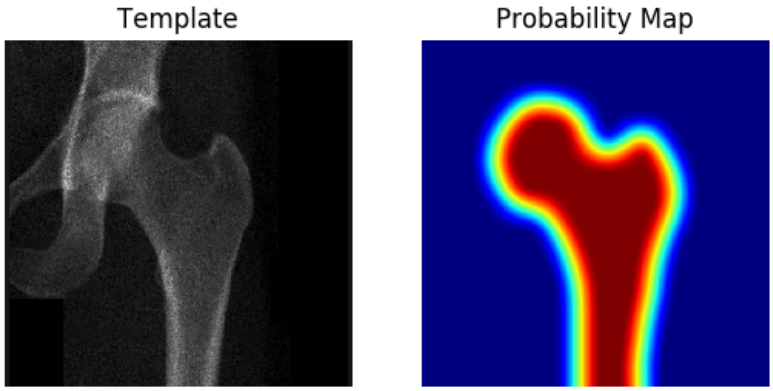


Fig. 5.3 A la izquierda, la plantilla utilizada para registrar las imágenes después del preproceso. A la derecha, el mapa de probabilidad que se utilizará como información extra.

	Rayos-X			Rayos-X + Prob. Map		
MeNet25	Training	Validación	Test	Training	Validación	Test
Full femur (*)	98.68	97.37	-	98.88	97.52	-
Head only	99.19	93.67	93.68	99.16	94.35	93.55
Head only + crop	99.05	94.11	94.35	99.03	93.16	92.18

Table 5.1 Dice obtenido por los distintos *pipelines* en segmentación de fémur. (*) El Dice se calcula con Cabeza + Cuerpo, donde el cuerpo representa mucha más superficie que la cabeza.

5.4 Resultados

Para mantener una consistencia con el trabajo realizado en segmentación de próstata, en este apartado también se ha utilizado Dice como métrica para evaluar el funcionamiento del modelo. Cabe destacar que, al estar las muestras supervisados por un humano no experto, los valores aquí presentados se deben interpretar con prudencia, pues se desconoce el error que contienen las segmentaciones de las cuales está aprendiendo el sistema.

En la tabla 5.1 se muestran los resultados obtenidos utilizando únicamente las imágenes preprocesadas como entrada para la red. Como se puede observar, el Dice obtenido al segmentar el fémur por completo es superior al obtenido cuando se segmenta únicamente la cabeza, esto se debe a que la segmentación del cuerpo del fémur es mucho más sencilla y representa una superficie mayor que la cabeza. En la Figura 5.4) se puede observar como aun obteniendo un Dice bastante alto, la segmentación en la parte de la cabeza es muy ruidosa.

(a) Sample (b) Model Segmentation (c) Ground Truth

Fig. 5.4 Comparativa entre la segmentación completa del fémur realizada por nuestro modelo (b) y el *ground truth* (c) para una muestra de conjunto de validación (a).

(a) Sample (b) Model Segmentation (c) Ground Truth

Fig. 5.5 Comparativa entre la segmentación únicamente de la cabeza del fémur realizada por nuestro modelo (b) y el *ground truth* (c) para una muestra de conjunto de validación (a).

MeNet17	Training	Validación	Test
Full femur	90.16	89.66	-
Head only	98.63	92.22	91.96
Head only + crop	98.66	92.72	92.77

Table 5.2 Resultados de los distintos *pipelines* en segmentación de fémur.

Como era de esperar, realizar un *crop* de la imagen mejora los resultados, aunque en este caso, la mejora no ha sido tan notoria como en la segmentación de próstata. En la Figura 5.5 se muestran las segmentaciones obtenidas por los *pipelines* que segmentan únicamente la cabeza del fémur.

Seguidamente, en la tabla 5.1 se pueden observar los resultados obtenidos utilizando los mismos enfoques, pero añadiendo un mapa de probabilidades a prior como información adicional, así pues la red recibiría dos canales: imagen + mapa de probabilidad. Como se puede observar en los resultados, dicha información no resulta de gran ayuda para la red y aumenta el *overfitting* de la misma.

Por último, destacar que en todos los resultados existe una gran distancia entre el Dice obtenido en entrenamiento y el Dice obtenido en validación. Hipotetizamos que esto se debe principalmente a los pocos datos que componen el corpus. Con el fin de intentar mejorar los resultados, se probó a utilizar el modelo **MeNet17** en vez del modelo **MeNet25**, con el fin de reducir la potencia del mismo y en consecuencia, reducir el *overfitting*. Los resultados se muestran en la Tabla 5.2

5.5 Discusión

Aunque las segmentaciones obtenidas aún distan de las generadas por un experto para la realización de estudios posteriores, creemos que los resultados obtenidos son aceptables teniendo en cuenta, tanto la calidad de las imágenes proporcionadas por la empresa, como la calidad de las etiquetas.

Existen varias vías de mejora, la más directa se basa en aumentar el número de muestras del corpus, ya que esto ayudaría a reducir la diferencia que existe actualmente entre el conjundo de training y el conjunto de validación.

Otra posible vía sería aumentar la calidad de las imágenes, ya que, aunque un experto humano sea capaz de segmentar la mayoría con facilidad, un pequeño aumento en la calidad de los bordes y en la reducción del ruido aumentarían considerablemente la capacidad del modelo a la hora de identificar las fronteras entre los distintos huesos y tejidos.

Por último, la imposibilidad de compararnos con otros métodos debido a la privacidad del trabajo deja los resultados sin un contexto en el que apoyarse. Aunque existen trabajos que también abordan esta tarea, o bien lo hacen sobre corpus mucho más grandes y supervisados [51], o bien realizan el estudio para un determinado paciente en concreto. Aún así, los resultados obtenidos en este trabajo no distan mucho de los resultados obtenidos en [11, 41], los cuales también abordan la tarea de segmentación ósea mediante redes convolucionales.

5.6 Conclusiones

En este trabajo se ha presentado un modelo capaz de realizar segmentaciones de fémur utilizando un corpus considerablemente reducido. Creemos que estos primeros resultados, aunque insuficientes para poder considerar el problema resuelto, son prometedores, y pueden ser la base para un proyecto junto a la empresa ERESA.

El desarrollo de un *software* totalmente automático para la segmentación del fémur sobre imágenes adquiridas mediante Rayos-X no es un problema trivial, pero a nuestro parecer, estos resultados demuestran que es algo totalmente viable. El siguiente paso es realizar las segmentaciones de varias muestras, para que en el departamento de Biomecánica puedan realizar un estudio sobre las mismas, y así validar si las hipótesis del estudio se constatarán con el estado real del paciente, lo cual verificaría las segmentaciones como válidas.

Capítulo 6

Observaciones Finales

6.1 Conclusiones

En este trabajo se ha presentado un modelo basado en Redes Convolucionales Profundas capaz de realizar segmentación morfológica tanto de próstata como de fémur. Este modelo ha sido entrenado desde cero, sin necesidad de *pretraining*, incluso cuando los corpus con los que se ha trabajado son considerablemente pequeños.

Respecto a la segmentación morfológica de próstata en imágenes volumétricas adquiridas mediante Resonancia Magnética, el modelo ha sido capaz de alcanzar un coeficiente Dice de 85.83; significando esto que, si los resultados se mantienen cuando se evalue el modelo con el conjunto de test, se lograría alcanzar el error humano. Este modelo se ha integrado en la plataforma MTSImaging, en la cual se ofrece la segmentación de próstata como un servicio gratuito y accesible para cualquier equipo clínico que lo necesite.

En segmentación morfológica de fémur en imágenes adquiridas mediante Rayos-X, el modelo ha sido capaz de obtener un coeficiente Dice de 94.35 en la segmentación de la cabeza del fémur, junto a un Dice de 97.37 en la segmentación del fémur completo. Aunque resulta complicado contextualizar estos resultados, debido a que es un corpus privado, las segmentaciones representan una base sólida sobre la cual empezar a realizar estudios clínicos posteriores.

Por último, en este trabajo se ha demostrado como el mimso modelo es capaz de obtener buenos resultados en problemas considerablemente distintos. Esta capacidad ofrece una oportunidad a todos aquellos equipos clínicos que requieran de segmentación morfológica para realizar sus estudios; ya que si dichos equipos son capaces de ofrecer un conjunto de

entrenamiento supervisado y representativo, el modelo puede ser reentrenado para generar las segmentaciones deseadas.

6.2 Limitaciones

No obstante este estudio presenta limitaciones, tales como la falta de evaluación del modelo frente a un conjunto de test supervisado por un experto. Esta limitación se pretende corregir, para el caso de próstata, enviando el modelo al *challenge* PROMISE12 para así obtener la puntación sobre su conjunto de test. En el caso de fémur se podría solventar si un radiologo experto segmentase una cantidad razonable de imágenes sobre las cuales contrastar el modelo. Por otro lado, en ambos escenarios se asume la centralidad de los elementos de interés en la imagen, y aunque esta asunción es razonable en el contexto clínico, hemos optado por no asumirla en la implementación del modelo que está funcionando en MTSImaging, con la intención de ofrecer una experiencia más robusta al usuario.

Por otro lado, debido a que nos hemos inspirado en las Redes Convolucionales que ostentan actualmente el estado del arte en segmentación semántica, se ha limitado considerablemente el espacio de búsqueda a la hora de definir la arquitectura de la red. El estudio de nuevos enfoques como por ejemplo *Generative Adversarial Networks* [26], el cual ha demostrado obtener resultados prometedores en segmentación morfológica [56], queda como trabajo futuro. A su vez, la optimización de los hiperparámetros de la red también se podría haber comparado con el resultado de una optimización Bayesiana [75], no realizándose por falta de tiempo.

Finalmente, la falta de comprensión sobre el proceso de inferencia que realizan modelos basados en Redes Neuronales, representa la principal limitación a la hora de aplicar estos modelos en un contexto clínico. Por eso mismo, el uso clínico de este modelo se limita a una herramienta de segmentación auxiliar, nunca utilizándose los resultados obtenidos por el mismo sin ser previamente supervisados por un experto. Un estudio teórico más profundo sobre el funcionamiento de este modelo se abordará en trabajos futuros.

6.3 Trabajo Futuro

El trabajo futuro de este proyecto se puede dividir en tres caminos, uno desde el enfoque de la aplicación del *Deep Learning* a la Imagen Médica, otro respecto a la segmentación de próstata y el último respecto a la segmentación del fémur.

Desde un punto de vista informático, existe la voluntad de seguir aplicando *Deep Learning* a problemas en el contexto médico, yendo más allá de la segmentación morfológica. Actualmetne se ha demostrado como dicha técnica obtiene resultados prometedores en la predicción de afecciones genéticas [17, 49], predicción de fallos cardíacos [12], localización de tumores [40] y un agradable etcétera. Lamentablemente, a día de hoy la ciencia no es capaz de explicar detalladamente porqué las Redes Neuronales son capaces de obtener tan buenos resultados, y cuando la vida de personas está en juego, no es aconsejable depender de tecnología que todavía representa un enigma para la ciencia. Por eso mismo, tenemos especial interés tanto en la aplicación de estas técnicas a modo de segunda opinión, como en el desarrollo de un estudio teórico que tenga como fin desentrañar el funcionamiento interno de esta tecnología.

El trabajo realizado en próstata ya cuenta con próximos pasos, ya que tal y como se ha comentado en la introducción, este trabajo se enmarca en un proyecto europeo que tiene como objetivo final realizar un análisis pronóstico del cáncer de próstata a través de la Imagen Médica. Próximos pasos como la segmentación de tumor, detección de regiones de interés, seguimiento del crecimiento tumoral... ya se encuentran definidas en lo que pretende ser parte de una tesis doctoral.

Por último, la segmentación de fémur se ha realizado en base al interés mostrado por la empresa ERESA. Por lo tanto, el posible desarrollo de un *software* para la segmentación automática depende directamente del interés que se muestre por parte de la empresa. Por otro lado, si las segmentaciones son lo suficiente buenas como para realizar un estudio posterior sobre las mismas, esperamos que los resultados de dicho estudio fortalezcan el uso del *Deep Learning* en segmentación morfológica sobre imágenes de Rayos-X.

Bibliografía

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[2] Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044.

[3] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561.

[4] Baochun He, F. J. (2016). Automatic mri prostate segmentation based cnn-asm.

[5] Bengio, Y. et al. (2009). Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.

[6] Birkbeck, N., Zhang, J., Requardt, M., Kiefer, B., Gall, P., and Kevin Zhou, S. (2012). Region-specific hierarchical segmentation of mr prostate using discriminative learning. *MICCAI Grand Challenge: Prostate MR Image Segmentation*, 2012.

[7] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[8] Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294.

[9] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[10] Campbell, M., Hoane, A. J., and Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, 134(1-2):57–83.

[11] Cernazanu-Glavan, C. and Holban, S. (2013). Segmentation of bone structure in x-ray images using convolutional neural network. *Adv. Electr. Comput. Eng*, 13(1):87–94.

[12] Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Medical concept representation learning from electronic health records and its application on heart failure prediction. *CoRR*, abs/1602.03686.

[13] Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851.

[14] Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

[15] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

[16] Cristinacce, D. and Cootes, T. (2008). Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067.

[17] Deming, L., Targ, S., Sauder, N., Almeida, D., and Ye, C. J. (2016). Genetic architect: Discovering genomic structure with learned neural architectures. *CoRR*, abs/1605.07156.

[18] Drozdzal, M., Chartrand, G., Vorontsov, E., Di Jorio, L., Tang, A., Romero, A., Bengio, Y., Pal, C., and Kadoury, S. (2017). Learning normalized inputs for iterative estimation in medical image segmentation. *arXiv preprint arXiv:1702.05174*.

[19] Duda, R. O. and Hart, P. E. (1972). Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15.

[20] Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658.

[21] Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929.

[22] Gall, J. and Lempitsky, V. (2009). Class-specific hough forests for object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1022–1029. IEEE.

[23] Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269.

[24] Giusti, A., Ciresan, D. C., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2013). Fast image scanning with deep max-pooling convolutional neural networks. *CoRR*, abs/1302.1700.

[25] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

[26] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[27] Grana, C., Borghesani, D., and Cucchiara, R. (2010). Optimized block-based connected components labeling with decision trees. *IEEE Transactions on Image Processing*, 19(6):1596–1609.

[28] Hassibi, B. and Stork, D. G. (1993). Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171.

[29] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

[30] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[31] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

[32] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[33] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

[34] Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10.

[35] Hussain, F., Abdul Kadir, M. R., Zulkifly, A. H., Sa'at, A., Aziz, A. A., Hossain, M. G., Kamarul, T., and Syahrom, A. (2013). Anthropometric measurements of the human distal femur: a study of the adult malay population. *BioMed research international*, 2013.

[36] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

[37] Japkowicz, N., Hanson, S. J., and Gluck, M. A. (2000). Nonlinear autoassociation is not equivalent to pca. *Neural computation*, 12(3):531–545.

[38] Juan-Albarracín, J., Fuster-Garcia, E., and García-Gómez, J. M. (2016). An online platform for the automatic reporting of multi-parametric tissue signatures: A case study in glioblastoma. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 43–51. Springer.

[39] Juan-Albarracín, J., Fuster-Garcia, E., Manjón, J. V., Robles, M., Aparici, F., Martí-Bonmatí, L., and García-Gómez, J. M. (2015). Automated glioblastoma segmentation based on a multiparametric structured unsupervised classification. *PloS one*, 10(5):e0125143.

[40] Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., and Glocker, B. (2015). Multi-scale 3d convolutional neural networks for lesion segmentation in brain mri. *Ischemic Stroke Lesion Segmentation*, 13.

[41] Kayalibay, B., Jensen, G., and van der Smagt, P. (2017). Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056*.

[42] Ke Yan, Xiuying Wang, J. K. C. L. D. F. (2016). P-dnn: A deep combination of mutilevel features for prostate segmentation from mr image.

[43] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109.

[44] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[45] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[46] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404.

[47] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[48] Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2015). Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570.

[49] Leung, M. K., Delong, A., Alipanahi, B., and Frey, B. J. (2016). Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE*, 104(1):176–197.

[50] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

[51] Lindner, C., Thiagarajah, S., Wilkinson, J., Consortium, T., Wallis, G., and Cootes, T. (2013). Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE transactions on medical imaging*, 32(8):1462–1472.

[52] Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al. (2014). Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373.

[53] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

[54] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024.

[55] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE.

[56] Moeskops, P., Veta, M., Lafarge, M. W., Eppenhof, K. A., and Pluim, J. P. (2017). Adversarial training and dilated convolutions for brain mri segmentation. *arXiv preprint arXiv:1707.03195*.

[57] Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932.

[58] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[59] Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. *CoRR*, abs/1505.04366.

[60] Pisano, E. D., Zong, S., Hemminger, B. M., DeLuca, M., Johnston, R. E., Muller, K., Braeuning, M. P., and Pizer, S. M. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital imaging*, 11(4):193–200.

[61] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[62] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[63] Rolle, M. (1690). *Traité d'algèbre ou principes généraux pour résoudre les questions de mathématique*.

[64] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

[65] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

[66] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38.

[67] Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

[68] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

[69] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2014). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*.

[70] Santhoshini, P., Tamilselvi, R., and Sivakumar, R. (2013). Automatic segmentation of femur bone features and analysis of osteoporosis. *Lecture Notes on Software Engineering*, 1(2):194.

[71] Sejnowski, T. J. and Rosenberg, C. R. (1988). *NETtalk: A parallel network that learns to read aloud*. MIT Press.

[72] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[73] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[74] Siwach, R., Dahiya, S., et al. (2003). Anthropometric study of proximal femur geometry and its clinical application. *Indian journal of Orthopaedics*, 37(4):247.

[75] Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. (2016). Bayesian optimization with robust bayesian neural networks. In *Advances in Neural Information Processing Systems*, pages 4134–4142.

[76] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

[77] Tahir, A., Hassan, A., and Umar, I. (2000). A study of the collodiaphyseal angle of the femur in the north-eastern sub-region of nigeria. *Nigerian journal of medicine: journal of the National Association of Resident Doctors of Nigeria*, 10(1):34–36.

[78] Tanimoto, A., Nakashima, J., Kohno, H., Shinmoto, H., and Kuribayashi, S. (2007). Prostate cancer screening: The clinical value of diffusion-weighted imaging and dynamic mr imaging in combination with t2-weighted imaging. *Journal of Magnetic Resonance Imaging*, 25(1):146–152.

[79] Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807.

[80] Vincent, G., Guillard, G., and Bowes, M. (2012). Fully automatic segmentation of the prostate using active appearance models. *MICCAI Grand Challenge: Prostate MR Image Segmentation*, 2012.

[81] Viola, P., Jones, M. J., and Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. In *null*, page 734. IEEE.

[82] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[83] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J.

(2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

[84] Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

[85] Yu, L., Yang, X., Chen, H., Qin, J., and Heng, P.-A. (2017). Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, pages 66–72.

[86] Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.