Detección de cáncer de piel usando técnicas de aprendizaje profundo

1st Alejandro Polvillo Hall
Departamento de Datos Geographica
Sevilla, España
alejandro@geographica.gs

2nd Juan A. Álvarez-García
Dpto. de Lenguajes y Sistemas Informáticos Universidad de Sevilla
Sevilla, España
jaalvarez@us.es

3rd Cristina Rubio-Escudero
Dpto. de Lenguajes y Sistemas Informáticos Universidad de Sevilla
Sevilla, España
crubioescudero@us.es

***Resumen*—El aprendizaje profundo ha sido muy utilizado para la clasificación de imágenes a partir de la competición ImageNet en 2012. Esta clasificación de imágenes es de gran utilidad en el campo de la medicina, en el que ha habido un gran crecimiento de uso de técnicas de minería de datos en los últimos años. En este trabajo seleccionamos y entrenamos una red de aprendizaje profundo para el análisis de un conjunto de datos de cáncer de piel, obteniendo resultados muy satisfactorios, ya que el modelo ha superado los resultados de clasificación de dermatólogos entrenados haciendo uso de un dermatoscopio, de otras técnicas de aprendizaje automático, y de otras técnicas de aprendizaje profundo.**

***Index Terms*—aprendizaje profundo, imágenes médicas, análisis de datos clínicos**

I. Introducción

El aprendizaje profundo ha sido utilizado en el campo de la visión por computadora durante décadas [5], [9]. Sin embargo, su verdadero valor no se había descubierto hasta la competición de ImageNet en 2012 [8], un éxito que provocó una revolución a través del uso eficiente del procesamiento de unidades gráficas (GPU). El principal poder del aprendizaje profundo radica en su arquitectura [10], [11], que permite discriminación en múltiples niveles de abstracción para un conjunto de características.

Las técnicas de aprendizaje profundo han sido utilizadas con éxito en campos como el de la medicina, en el que el aprendizaje profundo viene a solucionar problemas que presentan los algoritmos de aprendizaje automático con algunas estructuras de datos muy utilizadas en medicina como son las imágenes. La minería de datos clínicos es la aplicación de técnicas de minería de datos a los datos clínicos, con el objetivo de interpretar los datos disponibles. Permite la creación de modelos de conocimiento y proporciona asistencia para la toma de decisiones clínicas. En los últimos 10 años, ha habido un interés creciente en la aplicación de técnicas de minería de datos a los datos clínicos. MEDLINE ha visto un fuerte aumento de factor 10 en el número de trabajos con el término ”minería de datos.[en] su título [7].

Para hacer un entrenamiento completo, el aprendizaje profundo requiere una gran cantidad de datos de entrenamiento etiquetados, un requisito que puede ser difícil para cumplir en el campo de la medicina, donde la anotación de expertos es costosa y las enfermedades (por ejemplo, lesiones) no cuentan con grandes conjuntos de datos. Además, requiere una gran cantidad de recursos computacionales para que el entrenamiento no sea excesivamente lento.

En este trabajo hemos aplicado técnicas de aprendizaje profundo para analizar un conjunto de datos de imágenes de cáncer de piel, obteniendo resultados muy satisfactorios.

En las siguientes secciones describimos las metodologías y resultados obtenidos.

II. Metodología

En esta sección se describen las metodologías utilizadas para el desarrollo de este trabajo.

II-A. Tensorflow

Tensorflow [1] es uno de los mejores frameworks de aprendizaje profundo existente. Ha sido adoptado por un montón de grandes empresas como Airbus, Twitter, IBM y otras más debido a su flexibilidad y polivalencia. Tensorflow es desarrollado por Google, que la usa en todos sus proyectos de aprendizaje automático y aprendizaje profundo.

Tensorflow no es en sí un framework de aprendizaje profundo, es un framework que te permite trabajar de forma muy rápida con matrices gracias a su paralelización en GPU's. Como casi todos los cálculos realizados para entrenar y predecir con una red neuronal son cálculos matriciales hacen de esta herramienta que sea ideal para usarla en la construcción de redes neuronales.

Tensorflow tiene un paquete interno que viene con la funcionalidad necesaria para hacer funcionar una red neuronal convolucional: Capas convolucionales, optimizadores, funciones de optimización, etc...

Una desventaja de Tensorflow es que es necesario escribir mucho código para conseguir algo funcional. El hecho de escribir todo ese código hace que tengas un control total sobre todos los elementos de la arquitectura de la red neuronal. Aunque también puede hacer que cometas muchos errores.

II-B. Keras

Keras [2] es un framework específico de aprendizaje profundo. No es competidor de Tensorflow, pues Keras se ejecuta "encima" de Tensorflow. Keras provee una sintaxis extremadamente fácil para la creación de redes neuronales, y después convierte esta sintaxis a modelos de Tensorflow, usando la potencia de éste para ejecutar toda la maquinaria de aprendizaje.

Debido a que nuestro caso de uso será un caso de uso de red neuronal, y no necesitaremos en principio la manipulación a muy bajo nivel de los modelos de aprendizaje profundo, nosotros optamos por el uso de Keras para el desarrollo del presente proyecto.

III. Resultados

El conjunto de datos seleccionado para evaluar nuestra propuesta ha sido el conjunto de imáganenes de cáncer de piel de The International Skin Imaging Collaboration (ISIC) (https://isic-archive.com/). Es una plataforma que intenta aunar a los profesionales dermatológicos con el objetivo de luchar contra el cáncer de piel. El conjunto de datos está formado por 23906 imágenes. Además de las imágenes, se proporcionan algunos metadatos entre los que destacamos edad, sexo, lugar anatómico del melanoma, tipo de diagnóstico, clase de melanoma y espesor del melanoma.

En primer lugar analizamos el conjunto de datos con respecto a la clase benigno/maligno. Como se puede observar en la Figura 1 está muy desbalanceado.

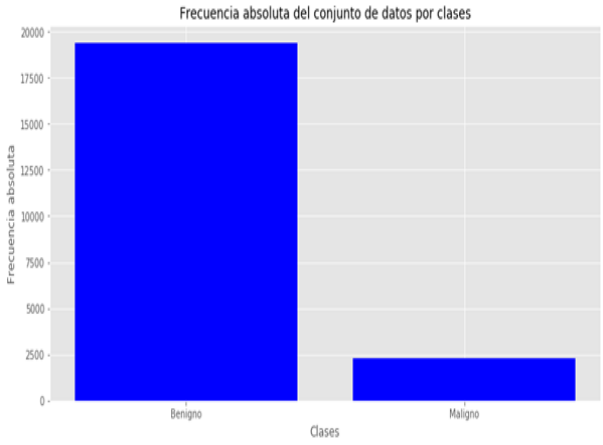

Figura 1. Conjunto de datos de cáncer de piel desbalanceado.

Debido a esto recurrimos a una técnica de remuestreo sobre la clase minoritaria que se encarga de escoger elementos al azar con reemplazo y añadirlos a la clase minoritaria. La ventaja de ésta técnica es que no elimina información, simplemente refuerza la información existente sobre la clase minoritaria. Tras aplicar la técnica el resultado obtenido se puede ver en la Figura 2.

En segundo lugar, partimos el conjunto de datos en entrenamiento (80 %) y test (20 %) asegurándonos de mantener balanceadas las clases en cada uno de los conjuntos.

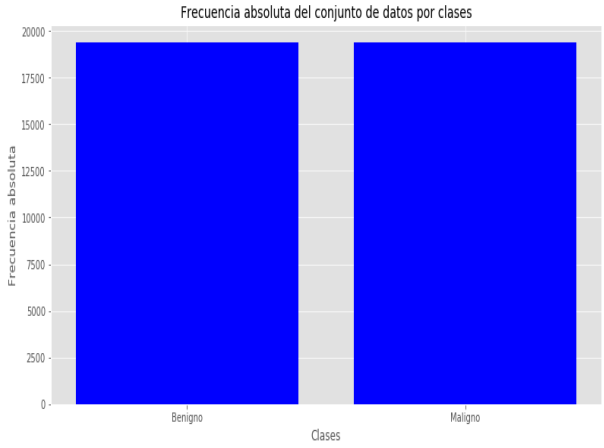

Figura 2. Conjunto de datos de cáncer de piel después del balanceo.

En tercer lugar, aplicamos un mecanismo de aumento de datos consistente en realizar distintas transformaciones aleatorias sobre cada uno de los elementos del conjunto de entrenamiento. Al realizar transformaciones aleatorias sobre el conjunto de datos conseguimos un doble efecto: aumentar el tamaño de nuestros datos y hacer que el sistema generalice mucho mejor. En particular aplicamos las técnicas conocidas como rotación automática, traslación vertical y horizontal de píxeles, cizallamiento, zoom, volteo, filtro de sal y pimienta.

III-A. Selección y aplicación del modelo de aprendizaje profundo

A la hora de enfrentarse a la selección del modelo nos encontramos en un momento de incertidumbre máxima debido a que son muchos los parámetros que tenemos que elegir, entre estos parámetros se encuentran:

- Arquitectura de la red neuronal: Propia, InceptionV3, VGG16, VGG19, Xception, etc...
- Tipo de entrenamiento: Transferencia de conocimiento, desde cero, etc…
- Parámetros de la transferencia de conocimiento: capas a entrenar, entrenamiento completo, etc...
- Algoritmo de aprendizaje: Descenso del gradiente, RMSProp, Adagrad, etc...
- Parámetros del algoritmo de aprendizaje: Tasa de aprendizaje, funciones de activación, épocas, lotes, etc…

Para seleccionar los parámetros adecuados hemos utilizado la búsqueda por rejilla (o grid search en inglés). Esta técnica se basa en la definición de un conjunto de posibles valores que pueden tomar los parámetros, y generar el producto cartesiano de todos los valores de todos los parámetros entre sí, usando cada uno de estos conjuntos de parámetros generados para construir un modelo y evaluar la bondad. En la Tabla I podemos ver todos los parámetros que se han considerado a la hora de buscar el modelo más adecuado.

En nuestra búsqueda rejilla se han fijado 3 parámetros: épocas, lotes y funciones de activación. Las épocas se han fijado puesto que es obvio que a más épocas mejor van a

Cuadro I
Parámetros considerados en la búsqueda del modelo más adecuado

Parámetro	Valores
Modelo	Xception, VGG16, VGG19 y InceptionV3
Tipo de entrenamiento	Transferencia de conocimiento, desde cero
Capas a entrenar	1, 2, 3 y todas
Algoritmo de aprendizaje	Descenso del gradiente, RMSProp, Adam, Adagrad
Tasa de aprendizaje	0.00001, 0.0001, 0.001, 0.01 y 0.1
Épocas	2
Lotes	10
Funciones de activación	ReLu

Cuadro II
Modelo y parámetros seleccionados

Parámetro	Valores
Modelo	InceptionV3
Tipo de entrenamiento	Transferencia de conocimiento
Capas a entrenar	todas
Algoritmo de aprendizaje	Adam
Tasa de aprendizaje	0.0001
Épocas	2
Lotes	10
Funciones de activación	ReLu

funcionar los modelos en su mayoría, así que se fijan a 2 épocas, para que así todas puedan recorrer el conjunto de datos 2 veces. Los lotes van de la mano de la capacidad de la unidad de procesamiento gráfico, aunque pueden influir en el entrenamiento, en este caso se han hecho los cálculos para que el conjunto de lotes sea el máximo que la unidad de procesamiento gráfica puede soportar, así hacemos que el tiempo de búsqueda por rejilla sea menor. La función de activación la he fijado a la función de activación ReLu debido a que todas las arquitecturas usan la ReLu debido a que el aprendizaje se realiza de una manera más rápida con este tipo de función de activación.

El modelo y parámetros que mejor desempeño han tenido, obteniendo un 81 % de exactitud se puede ver en la Tabla II:

El modelo seleccionado como óptimo es el modelo de Inception V3, creado por Google. Tal y como se describe en [12] era de esperar que se seleccionara la transferencia de conocimiento en vez del entrenamiento desde cero, pues posiblemente tenga un desempeño mayor. También se recomienda que, aunque se parta de unos pesos definidos (por la transferencia de conocimiento), se permita el reajuste de algunos pesos en las capas convolucionales, con el objetivo de que se adapte a nuestro problema. Como algoritmo de aprendizaje se usa Adam, con una tasa bastante pequeña. Esta tasa tan pequeña es debido a que gracias a la transferencia de conocimiento estamos muy cerca del óptimo, queremos dar pasos muy pequeños para acercarnos cada vez más al óptimo.

Una vez con nuestro modelo y parámetros definidos, procedemos a aumentar el número de épocas con el objetivo de obtener un mejor modelo y poder evaluar en la siguiente sección. Una vez el entrenamiento del modelo ha finalizado, hemos aplicado el clasificador sobre el conjunto de pruebas, obteniendo los resultados que se pueden ver en la Figura 3.

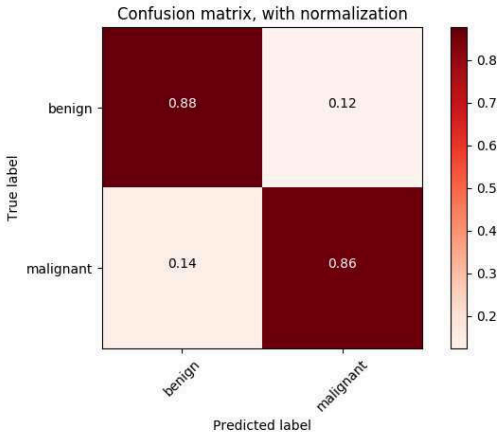

Figura 3. Matriz de confusión normalizada.

III-B. Evaluación del modelo

A continuación hablamos sobre la evaluación del modelo. La exactitud es la métrica por defecto que usa Keras para representar la bondad del modelo a lo largo de las distintas épocas de entrenamiento. En la Figura 4 podemos observar cómo se comporta el clasificador respecto a la exactitud a lo largo de las distintas épocas para los conjuntos de validación y de entrenamiento.

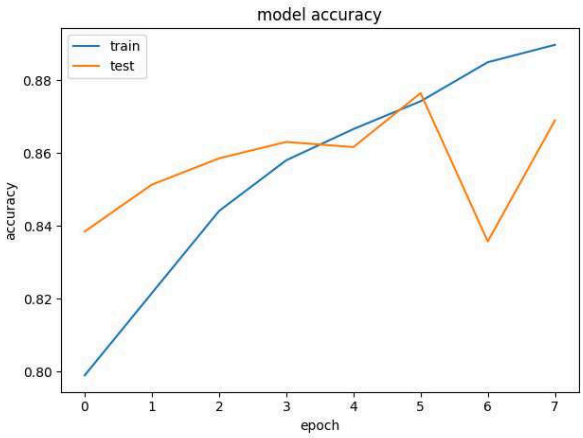

Figura 4. Exactitud del modelo a lo largo de las distintas épocas.

Como se puede observar en la gráfica siempre se encuentran en crecimiento, exceptuando en la época 6 que el conjunto de prueba hace algo difícil de interpretar, pero en la siguiente época se recupera.

El clasificador se ha puesto a entrenar durante 10 épocas, pero hemos activado una opción que permite ahorrar recursos. Esta opción consiste en que, si el clasificador durante la etapa de clasificación comienza a permanecer durante algún tiempo sin variar su exactitud, se pare la etapa de aprendizaje cuando se acabe la época. Por eso se puede observar que, aunque se haya puesto a entrenar durante 10 épocas, en la gráfica sólo se observan 8.

Cuadro III
RESULTADOS OBTENIDOS POR NUESTRO MODELO DE APRENDIZAJE PROFUNDO.

Medida	Valor
Exactitud	86.90 %
Precisión	87.47 %
Sensitividad	86.14 %
Especificidad	87.66 %
Área ROC	0.87

Se proporcionan también las medidas que se pueden ver en la Tabla III con respecto al desempeño de nuestro modelo de aprendizaje profundo.

La Figura 5 representa el área bajo la curva ROC

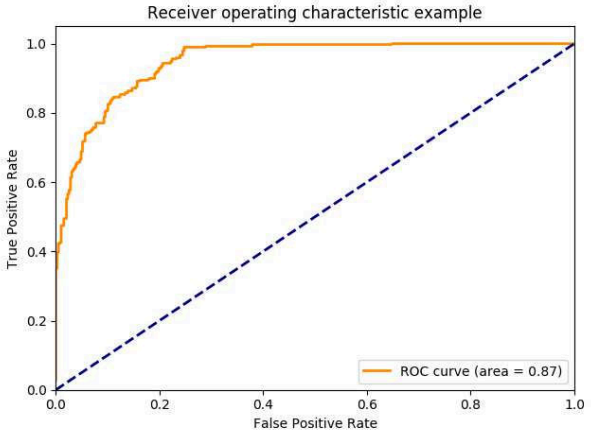

Figura 5. Área bajo la curva ROC.

Las inspecciones visuales sin ayuda de los expertos dermatológicos obtienen una media de un 60 % de exactitud [3], debido a la complejidad de observación de características diferenciadoras a simple vista. Con la ayuda de un experto entrenado junto a un dermatoscopio, la exactitud puede aumentar hasta un 75 % - 84 %.

Partiendo de la base de las afirmaciones anteriormente expuestas, y de las métricas obtenidas por nuestro clasificador, podemos observar a simple vista que nuestro clasificador ha superado en acierto a un experto dermatológico entrenado haciendo uso de un dermatoscopio. Esto parece un buen punto de partida.

Sabiendo que nuestro clasificador es capaz de superar a un experto dermatólogo, procedemos ahora a comparar nuestro clasificador con otras técnicas.

La primera intuición al intentar aplicar técnicas de inteligencia artificial en el análisis de imágenes médicas es usar aprendizaje automático o aprendizaje profundo. Con lo cual usaremos estos dos enfoques con el objetivo de comparar nuestro clasificador y ver qué bueno o malo es respecto a otros clasificadores

En primer lugar, nos centraremos en la comparación de nuestro clasificador con las técnicas de aprendizaje automático.

Cuadro IV
COMPARACIÓN DE RESULTADOS ENTRE APRENDIZAJE PROFUNDO Y APRENDIZAJE AUTOMÁTICO.

Medida	Nuestro clasificador	Aprendizaje automático
Exactitud	86.90 %	77 %
Precisión	87.47 %	71 %
Sensitividad	86.14 %	73 %

Cuadro V
COMPARATIVA ENTRE RESULTADOS DE NUESTRO APRENDIZAJE PROFUNDO Y OTRO APRENDIZAJE PROFUNDO.

Medida	Nuestro clasificador	Otro clasificador
Exactitud	86.90 %	85.5 %
Precisión	87.47 %	63.7 %
Sensitividad	86.14 %	50.7 %
Especificidad	87.66 %	94.1 %
Área ROC	0.87	0.8

El aprendizaje automático sobre el conjunto de datos de ISIC ha obtenido las métricas que se pueden ver en la Tabla IV [4].

Como se puede observar en la tabla, nuestro clasificador supera con creces a los clasificadores de aprendizaje automático enfocados al análisis de imágenes médicas para el conjunto de datos de ISIC. el resultado es lógico, pues los algoritmos de aprendizaje automático no están pensados para extraer características de imágenes, por lo que siempre se presupone un rendimiento menor.

En segundo, compararemos nuestro clasificador con otro clasificador de aprendizaje profundo [6]. El clasificador con el que compararemos nuestro clasificador, es un clasificador usado por un equipo participante de la competición de ISIC. Los resultados se pueden ver en la Tabla V:

Se puede observar, que, aunque el desempeño en la exactitud en ambos clasificadores es parecido, nuestro clasificador se comporta mucho mejor en la mayoría de situaciones. Ellos han conseguido obtener una especificidad muy alta, que es lo que le permite compensar la sensibilidad para obtener una exactitud y AUC ROC bastante aceptables.

IV. Conclusiones

En este trabajo hemos presentado el resultado de aplicar técnicas de aprendizaje profundo a un conjunto de datos de imágenes de cáncer de piel. Se ha descrito la fase de preprocesamiento de los datos, selección de parámetros para el modelo seleccionado y se han calculado distintas métricas para los resultados obtenidos.

En las métricas se pueden observar que el clasificador es bastante homogéneo, y no hay resultados de difícil interpretación. El clasificador tiene un nivel de acierto en ambas clases bastante estable, lo que nos hace indicar que ha generalizado correctamente la detección de cáncer en las imágenes. Nuestro modelo ha superado los resultados de clasificación de dermatólogo entrenado haciendo uso de un dermatoscopio, de otras técnicas de aprendizaje automático, y de otras técnicas de aprendizaje profundo.

En definitiva, el clasificador se ha comportado de manera muy buena frente a este conjunto de imágenes respecto a

otros clasificadores y técnicas. Aún sabiendo que las imágenes médicas son complicadas de tratar, el clasificador ha sabido generalizar los detalles que separan una imagen de un paciente que padece cáncer de piel de otro que no lo padece, consiguiendo así predecir con exactitud en la mayoría de casos.

Agradecimientos

Este trabajo ha sido financiado por los proyecto del Ministerio de Economía y Competitividad TIN2014-55894-C2-1-R y TIN2017-82113-C2-1-R.

Referencias

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] François Chollet et al. Keras, 2015.

[3] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 168–172. IEEE, 2018.

[4] Machine Learning for ISIC Skin Cancer Classification Challenge. https://hackernoon.com/machine-learning-for-isic-skin-cancer-classification-challenge-part-1-ccddea4ec44a.

[5] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[6] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

[7] Jimison Iavindrasana, Gilles Cohen, Adrien Depeursinge, H Müller, R Meyer, and Antoine Geissbuhler. Clinical data mining: a review. *Yearbook of medical informatics*, 18(01):121–133, 2009.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[12] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.