Automatización de evaluación de madurez ósea a través de técnicas de aprendizaje profundo

ABSTRACT

La evaluación de la edad ósea esquelética es una práctica clínica comúnmente utilizada para investigar la madurez del sistema esquelético de un niño. Esta práctica puede ayudar a los médicos a diagnosticar afecciones que retrasan o aceleran el crecimiento y desarrollo físico. Recientemente, el advenimiento y la proliferación de redes neuronales convolucionales (RNC) ha mostrado ser prometedor en una variedad de aplicaciones de imágenes médicas. En este documento proponemos y probamos varios enfoques de aprendizaje profundo para evaluar la madurez ósea de forma automática comparando dos enfoques: por un lado, una RNC general que estima madurez ósea sobre radiografías de género masculino y femenino; por el otro, dos RNC especializadas cada una en su género respectivo. Los resultados mostraron que la utilización de un modelo general ofrece resultados más precisos para la estimación de madurez ósea; sin embargo, los modelos especializados ofrecer resultados similares. Esta es una de las primeras evaluaciones automatizadas de la edad ósea esquelética probada en un conjunto de datos públicos donde se evalúa la utilización de una RNC general y dos RNC específicas para el género de la persona, para los cuales el código fuente está disponible, representando así una base exhaustiva para futuras investigaciones en el campo.

1. INTRODUCCIÓN

Durante el desarrollo del organismo de una persona, los huesos del esqueleto cambian de tamaño y forma, los cuales responden a una determinada edad ósea. Diferencia entre la edad ósea estimada de un niño y su edad cronológica podría indicar trastornos del crecimiento y anomalías endocrinas [1]. Los médicos utilizan la evaluación de la edad ósea para estimar la madurez del sistema esquelético de un niño. Los métodos de evaluación de la edad ósea generalmente comienzan con tomar una sola imagen de rayos X de la mano izquierda desde la muñeca hasta las puntas de los dedos, ver Figura 1. Los huesos en la imagen de rayos X se comparan con radiografías en un atlas estandarizado de desarrollo óseo. Tal atlas de edad ósea se basa en un gran número de radiografías recogidas de niños del mismo sexo y edad.

En las últimas décadas, el procedimiento de evaluación de la edad ósea se realizó de forma manual utilizando los métodos de Greulich y Pyle (GP) [2] o de Tanner-Whitehouse (TW) [3]. El procedimiento GP determina la edad ósea al comparar la radiografía del paciente con un atlas de edades representativas. La técnica TW se basa en un sistema de puntuación que examina 20 huesos específicos. En ambos casos, el procedimiento de evaluación ósea requiere un tiempo considerable; a su vez, la precisión en la estimación depende de la experiencia del radiólogo y tiende a ser subjetiva.

Desde 1992, preocupaciones sobre la variabilidad inter-observador en la estimación manual de la edad ósea [4] han llevado al establecimiento de varios métodos automáticos computarizados para la estimación de la misma. Los recientes avances en el aprendizaje profundo y sus aplicaciones a la visión por computadora permitieron a muchos investigadores mejorar drásticamente los resultados obtenidos con los sistemas de procesamiento de imágenes particularmente relacionados con el análisis de imágenes médicas [5]. A diferencia de las técnicas tradicionales de aprendizaje automático, las técnicas de aprendizaje profundo permiten que un algoritmo se programe a sí mismo aprendiendo de las imágenes dadas a un gran conjunto de datos de ejemplos etiquetados, eliminando así la necesidad de especificar reglas [6]. Los enfoques basados en aprendizaje profundo están ganando más atención porque en varios casos se demostró que logran e incluso superan el rendimiento a nivel humano, lo que hace que el procesamiento de imágenes de extremo a extremo sea automatizado y suficientemente rápido. En el campo de las imágenes médicas, las redes neuronales convolucionales (RNC) se han utilizado con éxito

Figura 1: Huesos de una mano y muñeca humana adaptada de [15]

para el cribado de retinopatía diabética [7], diagnóstico de enfermedades cardíacas [8], detección de cáncer de pulmón [9] y otras aplicaciones [5]. En el caso de la evaluación de la edad ósea, el mismo es un procedimiento que realizado manualmente requiere alrededor de 30 minutos de tiempo del médico por cada paciente. Cuando se realiza el mismo procedimiento utilizando un software basado en los métodos clásicos de visión por computadora, toma de 1 a 5 minutos, pero aún requiere una considerable supervisión y experiencia médica. Los métodos basados en aprendizaje profundo permiten evitar la ingeniería de características al aprender automáticamente la jerarquía de características discriminatorias directamente de un conjunto de ejemplos etiquetados. Usando un enfoque de aprendizaje profundo, el procesamiento de una imagen generalmente toma menos de 1 segundo, mientras que la precisión de estos métodos en muchos casos excede la de los métodos convencionales. Soluciones de redes neuronales profundas para la evaluación de la edad ósea de radiografías de mano han sido sugeridas [10]–[12]. Sin embargo, ninguno de estos ha evaluado en profundidad la conveniencia de separar el conjunto de datos de acuerdo al género de la radiografía y generar una RNC por cada género para investigar la precisión resultante de los mismos.

En este trabajo se investiga si una RNC especializada en un determinado género podría ser más precisa en estimar la edad de una persona de acuerdo a la radiografía de su mano en lugar de una que utilice radiografías de ambos géneros. Nuestra contribución consiste en explorar el comportamiento de una RNC de acuerdo al conjunto de datos que es ingresado en la misma. Validamos la precisión de estas redes neuronales utilizando los datos de el desafío pediátrico de edad ósea 2017 organizado por la Sociedad Radiológica de América del Norte (RSNA) [13]. Este conjunto de datos ahora está disponible de manera gratuita y puede ser accedido en [14].

2. Metodología y Objetivos

Se realizó la comparación de dos alternativas: por un lado, un modelo general entrenado con radiografías de ambos géneros; y por otro, dos modelos entrenados cada uno con radiografías de un solo género, resultando así la obtención de un modelo específico para cada sexo: uno especializado en radiografías femeninas y otro en radiografías masculinas. El objetivo planteado es analizar cuál de las dos alternativas obtiene mayor precisión: dos modelos especializados en un género cada uno, donde se reduce el conjunto de datos a la mitad o un solo modelo general entrenado con radiografías de ambos géneros. La precisión de los modelos será definida en base al cálculo del MAE (Mean Absolute Error, Error Medio Absoluto) el cual es calculado como $\frac{1}{n}\sum_{t=1}^{n}|e_t|$, siendo $|e_t|$ el valor absoluto de la diferencia entre cada estimación y la edad en meses etiquetada en cada observación y n la cantidad de observaciones tomadas.

2.1. *Conjunto de datos*

El conjunto de datos utilizado ha sido extraído del desafío pediátrico de edad ósea 2017 organizado por la Sociedad Radiológica de América del Norte (RSNA) [13]. El mismo se encuentra conformado por más de 12 mil radiografías pediátricas de muñeca y mano etiquetadas por su género respectivo. Además, el conjunto posee una etiqueta de edad en meses estimada del paciente por un radiólogo. Al realizar la conversión de meses a años, las radiografías del conjunto de datos varían entre 0 y 19 años. Radiografías de pacientes con una maduración esquelética estimada menor a dos años han sido excluidas del experimento por dos razones. En primer lugar, el conjunto de datos correspondiente a esos rangos de maduración es muy pequeño (38 casos para femenino y 53 casos para masculino). En segundo lugar, la estimación de maduración esquelética es mayormente utilizada para casos de pubertad retrasada, pubertad precoz o baja estatura. Estos estudios raramente son realizados en tales rangos etarios, donde además, se evita la exposición a rayos X. El número total de observaciones originalmente obtenidas fue de 6,833 para hombres y 5,778 para mujeres. Luego de excluir aquellas de edades entre 0 y 2, ha resultado en 6,780 para hombres y 5,740 para mujeres como se muestra en la Figura 2.

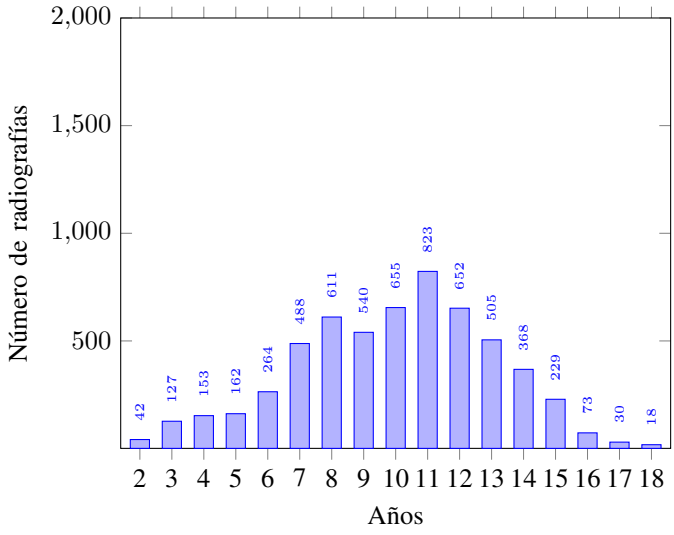
2.2. *Preprocesamiento*

Un módulo de preprocesamiento ha sido desarrollado para obtener un conjunto de radiografías más homogéneo. El módulo se compone de tres etapas. En la primer etapa se realiza una nivelación de colores para resaltar en la radiografía el área ocupada por la mano. En la segunda etapa se detecta la mano y se recorta la imágen para que la misma ocupe la mayor cantidad de pixeles posible. En la tercer etapa se rota la radiografía para posicionar la mano verticalmente. El objetivo del módulo de preprocesamiento consiste en obtener radiografías donde la mano se posicione verticalmente y ocupe la mayor superficie posible sobre la imagen. Esto permite una mayor precisión en el modelo al homogeneizar el conjunto de datos.

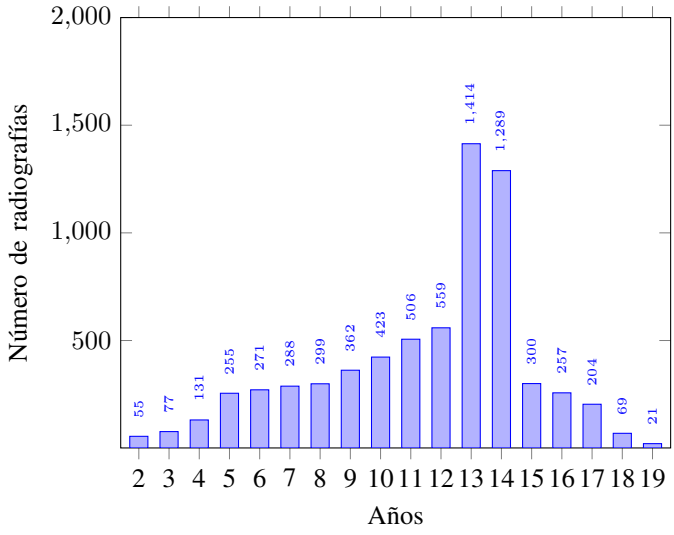
2.3. *Sistemas Propuestos*

El experimento para la comparación de los sistemas propuestos ha sido implementado en Python. La utilización y parametrización de las redes neuronales que presentan los sistemas ha sido principalmente a través de la librería Keras[1].

[1]Librería Keras. https://keras.io



(a) Femenino



(b) Masculino

Figura 2: Distribuciones de edad ósea para radiografías de género masculino y femenino

El código completo desarrollado para este experimento se encuentra disponible para uso académico en https://github.com/deeplearningrosario/Pediatric_Bone_Age_Challenge.

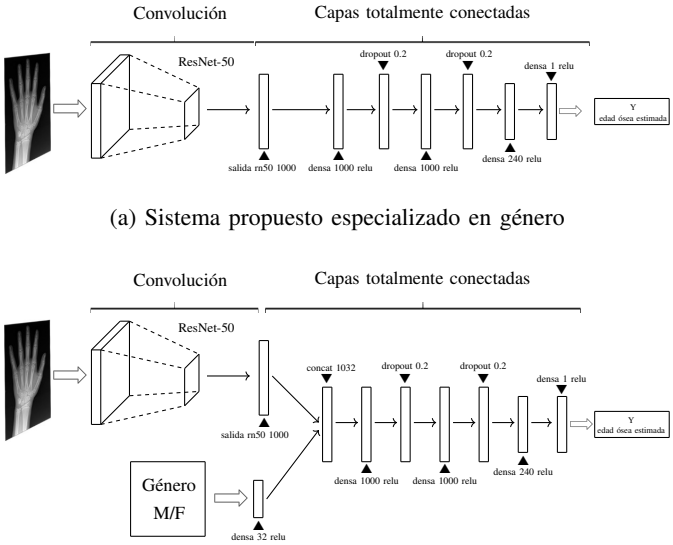
Con respecto a la definición de los sistemas a utilizar para el experimento, se definieron dos arquitectuas que se describen a continuación.

Se definieron dos arquitecturas para los sistemas propuestos. Los mismos se presentan en la Figura 3. Los sistemas se conforman de una etapa de convolución seguida por una etapa de capas totalmente conectadas. En la etapa de convolución se ha utilizado ResNet-50 [16] con pesos pre entrenados con Imagenet [17]. Otras redes convolucionales han sido probadas, como InceptionV3 [18] y Xception [19], sin embargo, ResNet-50 obtuvo mejores resultados. La etapa de capas totalmente conectadas es la encargada de realizar la regresión de la edad basándose en los patrones extraídos por la etapa de convolución. La diferencia entre las dos arquitecturas para los sistemas propuestos radica en que el sistema generalizado permite utilizar el género de la radiografía como parámetro. El mismo es concatenado con la extracción de patrones resultantes de ResNet-50 en la etapa convolucional.

Tres sistemas han sido entrenados para el experimento. Un sistema generalizado entrenado con radiografías de género másculino y femenino (Generalizado), y dos sistemas especializados; uno entrenado solo con radiografías de género masculino (Especializado Masculinas) y otro entrenado solo con radiografías de género femenino (Especializado Femeninas). Los sistemas fueron entrenados con 200 épocas, tamaño de lote de 32 y ratio de aprendizaje de 0.001.

3. Resultados del experimento

Para los experimentos realizados se ha calculado el MAE resultante en cada uno de los sistemas propuestos con un conjunto de datos de radiografías masculinas y/o femeninas según corresponda. Los resultados se muestran en el Cuadro 1. En cada fila se muestran los MAE en meses resultantes para cada sistema. La primer columna especifica el sistema utilizado. Segunda columna el MAE resultante para el conjunto de radiografías femeninas. Tercer columna el MAE resultante para el conjunto de radiografías masculinas. Puede observarse que el sistema generalizado obtiene mejores resultados para ambos conjuntos de datos. El sistema generalizado presenta mayor precisión con respecto al sistema especializado en radiografías femeninas en ~1.066 meses. También presenta mayor presición con respecto al sistema especializado en radiografías masculinas en ~0.571 meses.



(a) Sistema propuesto especializado en género

(b) Sistema propuesto generalizado para ambos géneros

Figura 3: Sistemas propuestos

	Femeninas	Masculinas
Generalizado	9.326	9.120
Especializado Femeninas	10.392	-
Especializado Masculinas	-	9.691

Cuadro 1: MAE en meses de los sistemas de acuerdo al género de radiografía utilizado

4. Conclusiones y Trabajo Futuro

En esta sección se presentan conclusiones sobre la experimentación realizada y pasos a seguir para aumentar la precisión de los sistemas tratados.

4.1. Conclusiones

Nuestra motivación en esta experimentación fue evaluar los diferentes comportamientos que puede presentar una red neuronal de acuerdo a variaciones en el conjunto de datos de entrada como así también los parámetros utilizados en la misma. Se han presentado los resultados obtenidos sobre los sistemas propuestos. Se ha observado que un sistema que acepte ambos géneros ha obtenido mayor precisión. Sin embargo, la pequeña diferencia en precisión observada al compararlo con los sistemas especializados da lugar a

pensar mejoras que podrían disminuir tal diferencia. Además, se han desarrollado arquitecturas para la extracción de patrones y estimación de edad de una radiografía de mano y muñeca pediátrica que podría ser de gran utilidad para aquellas personas interesadas en abordar problemáticas similares, como así también en el ámbito de la medicina a través de su implementación en instituciones de salud, como Hospitales o Sanatorios, para asistir en la labor de pediátras, endocrinólogos, radiólogos y/u otros especialistas.

4.2. Trabajo Futuro

En una segunda fase de experimentación el desafío se centra en aumentar la precisión de los sistemas. El módulo de preprocesamiento actual ha probado su utilidad aumentando la precisión de los modelos en 0,2 meses. En una segunda fase se reemplazará el módulo por una red neuronal que realice las mismas funciones con mayor precisión, lo que se estima que disminuirá el error de los modelos. Además, se continuará experimentando variaciones estructurales en el etapa de capas totalmente conectadas, como así también variaciones en la etapa convolucional. En cuanto al conjunto de datos, se aplicarán técnicas de aumento de datos (data augmentation) para incrementar la cantidad de radiografías con las que sería posible entrenar los sistemas.

REFERENCIAS

[1] Zerin, J. M., & Hernandez, R. J. (1991). *Approach to skeletal maturation*. Hand clinics, 7(1), 53-62.

[2] Greulich, W. W., Pyle, S. I., & Todd, T. W. (1959). *Radiographic atlas of skeletal development of the hand and wrist*. (Vol. 2, pp. 150-159). Stanford: Stanford university press.

[3] Tanner, J. M., Healy, M. R. J., Goldstein, H., & Cameron, N. (2001). *Assessment of skeletal maturity and prediction of adult height (TW3 method)*. WB Saunders, London, 243-254.

[4] Berst, M. J., Dolan, L., Bogdanowicz, M. M., Stevens, M. A., Chow, S., & Brandser, E. A. (2001). *Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards*. American Journal of Roentgenology, 176(2), 507-510.

[5] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Xie, W. (2018). *Opportunities and obstacles for deep learning in biology and medicine*. Journal of The Royal Society Interface, 15(141), 20170387.

[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521(7553), 436.

[7] Rakhlin, A. (2018). *Diabetic Retinopathy detection through integration of Deep Learning classification framework*. bioRxiv, 225508.

[8] Korshunova, I. (2017). *Diagnosing heart diseases with deep neural networks*. `https://irakorshunova.github.io/2016/03/15/heart.html`, accedido 29 de Julio de 2017.

[9] Daniel Hammack and Julian de Wit. (2017) *2017 Data Science Bowl, Predicting Lung Cancer: 2nd place solution write-up*. `http://blog.kaggle.com/2017/06/29/2017-data-science-bowl-predicting-lung-cancer-2nd-place-solution-write-up-daniel-hammack-and-julian-de-wit/`, accedido 29 de Julio de 2017.

[10] Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., & Langlotz, C. P. (2017). *Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs*. Radiology, 287(1), 313-322.

[11] Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., ... & Do, S. (2017). *Fully automated deep learning system for bone age assessment*. Journal of digital imaging, 30(4), 427-441.

[12] Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., & Leonardi, R. (2017). *Deep learning for automated skeletal bone age assessment in X-ray images*. Medical image analysis, 36, 41-51.

[13] *RSNA Pediatric Bone Age Challenge*. `http://rsnachallenges.cloudapp.net/competitions/4`, accedido 29 de julio de 2017.

[14] Stanford University Artificial Intelligence in Medicine & Imaging (2017). *Bone age images used in the 2017 RSNA bone age challenge competition*. `https://aimi.stanford.edu/available-labeled-medical-datasets`, accedido 29 de julio de 2017.

[15] Gilroy, A.M., MacPherson B.R., Ross L.M., Schünke M., Schulte E., Schumacher U., Voll M., & Wesker K. (2008). *Prometheus. Atlas de Anatomía*. pp 298.

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

[17] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). *Imagenet: A large-scale hierarchical image database*. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). Ieee.

[18] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[19] Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions*. arXiv preprint, 1610-02357.