CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

TALLER: TRABAJO FINAL INTEGRADOR

Detección de COVID-19 en radiografías de tórax

Nombre y Apellido del Alumno: Sebastián Iglesias

Profesores: Dra. Alicia Mon

Lugar y Fecha: Buenos Aires, 6 de Septiembre de 2020

Tercer entrega

ÍNDICE

1. ESTADO DE LA CUESTIÓN

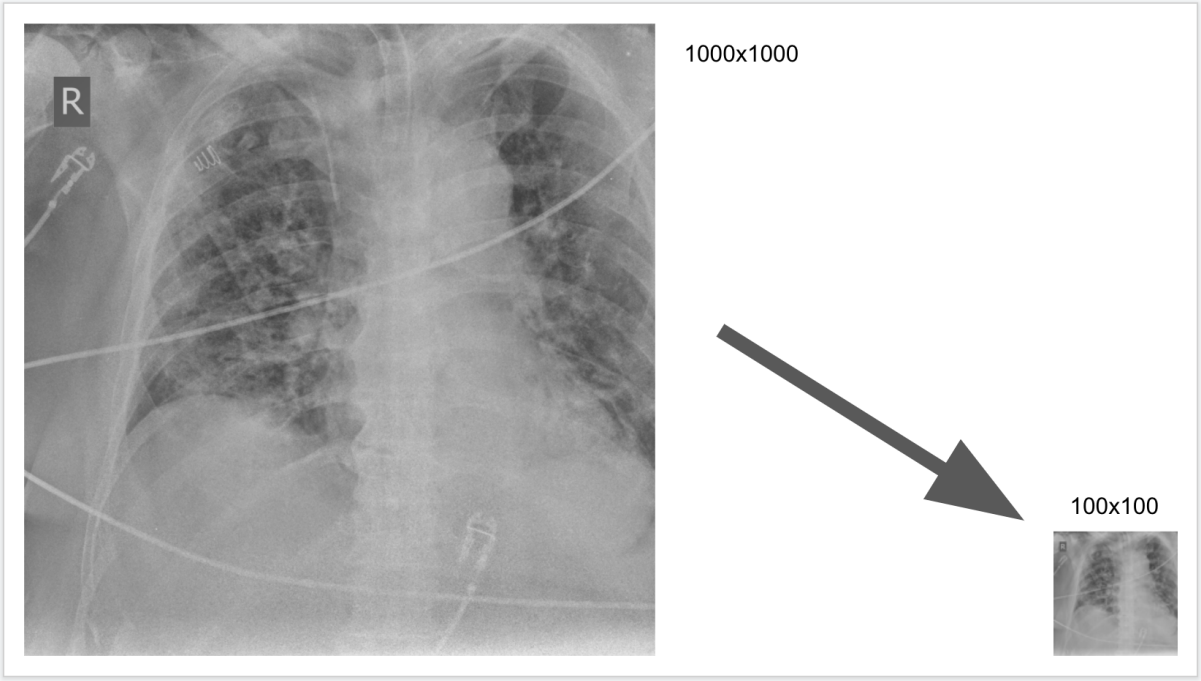
El COVID-19 es una enfermedad infecciosa reconocida como pandemia por la Organización Mundial de la salud [1]. Una de las técnicas definidas por este ente para controlar la propagación del virus, es el seguimiento de contactos ("Contact Tracing"). Por lo tanto, un paso crítico para llevar a cabo esa medida es la detección efectiva y acertada de pacientes que contraigan la enfermedad. Tanto para recibir el tratamiento adecuado de manera pronta, como también isolarlos del público en general, para prevenir futuros contagios.

Una de las técnicas más popularizadas para la detección del COVID-19 es rRT-PCR, reacción en cadena de la polimerasa con transcripción inversa, que analiza la producción de anticuerpos en respuesta a la infección ocasionada por dicho virus [2-3]. Este tipo de técnicas, propios de la serología tiene limitaciones, como por ejemplo la disponibilidad de equipos de testeo que provee dificultades a la hora de detectar el virus en un alto número de la población. Más allá de esto, el tiempo en que una de estas pruebas puede dar un resultado es entre algunas horas a hasta dos días. Inclusive, el resultado de estas pruebas, en la situación de emergencia mundial, es propensa a errores.

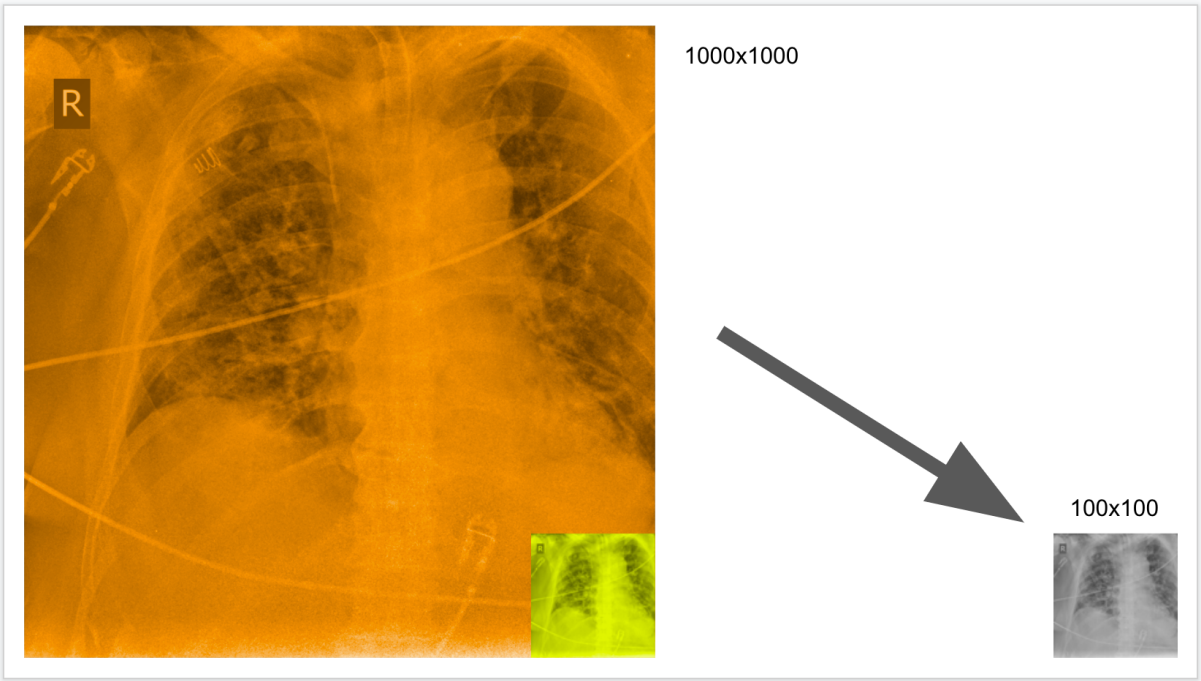
De esta manera, a partir de la necesidad de pruebas más veloces y con una menor propensión a errores, nacen estudios que involucran el análisis de imágenes de rayos x y tomografías utilizando visión artificial, específicamente en la región del pecho [4]. La mayoría de los pacientes con COVID-19 destacan opacidades recíprocas, multi-focales, similares a vidrio esmerilado, con una diseminación marginal en la etapa temprana y tardía de la infección [5]. En particular, el estilo de técnicas de visión artificial predilecto para estos casos, son las redes de aprendizaje profundo que pueden indicar características difíciles de singularizarse de la imagen original. El modelo predilecto para estos casos son las redes convolucionales [6].

Para estos casos [9], el formato estándar en que se encuentran las radiografías "crudas" para realizar el análisis es en formato DICOM. Por cuestiones de almacenamiento, los datasets de este estilo de imágenes pueden también encontrarse como JPEG, JPG o PNG. A estas imágenes se les realiza un preprocesamiento, que suele incluir una primera etapa de remover partes innecesarias de la imagen (bordes y espacios oscuros al costado de los cuerpos), una segunda donde se le aplican técnicas de reducción de ruido y una tercera etapa donde se modifica el tamaño de estas, respetando el ratio de la imagen.

Las radiografías, si bien respetan el formato DICOM tienen una resolución variable dependiente del dispositivo que fue utilizado para obtenerlas, que promedia los 1980 pixeles de alto y 2000 pixeles de ancho. Las imágenes transformadas a un tamaño menor son la entrada del modelo predictivo, normalmente utilizada una red convolucional o variación de esta [7-11]. El tamaño de la entrada del modelo es un factor limitante para el costo y tiempo de entrenamiento del modelo, la "regla de oro" o en ingles "rule of thumb" para la resolución de imágenes, respetando su ratio para no alterar la relación entre cada uno de los puntos, es de 128 píxeles de alto y 128 píxeles de ancho (referenciado como 128x128). Para casos particulares y que conlleven un mayor poder de procesamiento, se considera utilizar resoluciones de 256x256 o 512x512 como una resolución inusual.



Es importante destacar que luego de la tercera etapa de preprocesamiento, tomando 128x128 como ejemplo, se está utilizando aproximadamente un 40% de la imagen cruda o real y hay un 60% de la imagen original que no se considera de ninguna manera en el análisis. Sin embargo, es necesario reducir la dimensionalidad del conjunto de entrada de la red para evitar elevados costos al entrenar el modelo.



1.1 Reducción de dimensionalidad

Reducir la dimensionalidad de un conjunto de datos involucra transformar dicho conjunto de un espacio de mayor dimensionalidad a uno de menor, con la intención que el de menor dimensión continúe representando las propiedades características de la dimensión original. Algunas de las técnicas que son utilizadas para esto son Principal Component Analysis (PCA) y t-Distributed Stochastic Neighbor Embedding (t-SNE).

2. DEFINICIÓN DEL PROBLEMA

Los estudios que utilizan redes neuronales convolucionales para la detección de COVID-19 [7-11], involucran una etapa donde disminuyen el tamaño de la imagen a un 40% de su tamaño original. Esta reducción de dimensionalidad involucra pérdida de información, hay alternativas que involucran una menor pérdida, pero no hay ninguna comúnmente usada en el nicho de radiografías de tórax en la detección de COVID-19.

No hay un análisis que considere la perdida de información de distintas técnicas y su relación en métricas del modelo predictivo para esta área específica del análisis de imágenes.

3. JUSTIFICACIÓN DEL ESTUDIO

Este estudio tiene la intención de comparar tres técnicas de reducción de dimensionalidad (PCA y t-SNE) para radiografías de tórax. Se analiza la pérdida de información por medio de la entropía de la información y el índice gini de las imágenes preprocesadas que actúan como entrada de un modelo predictivo que utiliza como base una red convolucional. Sobre cada uno de los tres modelos entrenados que detecte si un individuo presenta o no la enfermedad, se calcula la “accuracy”, “precisión” y “recall”.

4. ALCANCES DEL TRABAJO Y LIMITACIONES

El conjunto de datos a utilizar en la investigación es provisto por “COVID-X”, una iniciativa argentina que utiliza modelos predictivos basados en imágenes radiográficas de tórax para la detección de COVID-19. El proyecto “COVID-X” provee a este trabajo de investigación con las imágenes anonimizadas y permite el uso de las métricas de su modelo para la comparación de resultados bajo el eje de análisis de este trabajo.

Esta investigación no tiene la intención de realizar una aplicación ni pagina web que visualice el modelo, ni permita a un usuario ingresar imágenes para su posible detección. El resultado de este trabajo es una comparación en términos de perdida de información y su relación con la precisión en la detección de la enfermedad, partiendo de un mismo conjunto de datos.

5. HIPÓTESIS

Los modelos, con diferentes graduaciones de pérdida de información de las radiografías de tórax originales luego del procesamiento previo a su análisis debido al uso de diferentes técnicas, presentan el mismo rendimiento en términos de “accuracy”, “precision” y “recall” del modelo predictivo para la detección de COVID-19.

Esas tres variables son métricas utilizadas para evaluar modelos de clasificación. “Accuracy” representa la cantidad de predicciones correctas del conjunto total de predicciones. “Precision” permite indicar que proporción de los resultados que fueron identificados como correctos son verdaderamente correctos. Y “recall” indica la proporción de los casos verdaderamente positivos fueron identificados como tales.

6. OBJETIVOS

Objetivos Generales:

- Obtener el nivel de pérdida de información en términos de entropía e índice gini para las técnicas de reducción de dimensionalidad PCA, t-SNE y “Autoencoder” y la efectividad del modelo predictivo asociado medido en términos de “accuracy”, “precision” y “recall”

Objetivos específicos:

- Reducir la dimensionalidad de cada una de las imágenes del conjunto de datos original por medio de la técnica de Principal Component Analysis (PCA) y obtener los niveles de pérdida expresados en entropía e índice gini.
- Reducir la dimensionalidad de cada una de las imágenes del conjunto de datos original por medio de la técnica de t-Distributed Stochastic Neighbor Embedding (t-SNE) y obtener los niveles de pérdida expresados en entropía e índice gini.
- Reducir la dimensionalidad de cada una de las imágenes del conjunto de datos original por medio de una red neuronal que actúe como “Autoencoder” y obtener los niveles de pérdida expresados en entropía e índice gini.
- Utilizando como entrada el resultado de PCA, generar un modelo predictivo con una red convolucional que detecte la enfermedad y obtener las métricas de “accuracy”, “precision” y “recall”
- Utilizando como entrada el resultado de t-SNE, generar un modelo predictivo con una red convolucional que detecte la enfermedad y obtener las métricas de “accuracy”, “precision” y “recall”
- Utilizando como entrada el resultado del Autoencoder, generar un modelo predictivo con una red convolucional que detecte la enfermedad y obtener las métricas de “accuracy”, “precision” y “recall”

7. METODOLOGÍA A UTILIZAR

Partiendo del mismo conjunto de datos iniciales, se reduce la dimensionalidad de cada una de las imágenes del dataset. Se obtienen 2 nuevos conjuntos, uno donde se utilizó la técnica PCA, otro con la técnica t-SNE. Para cada uno de los nuevos conjuntos, se obtiene un "five-number summary" de las métricas de pérdida de información (entropía e índice gini). Las imágenes resultantes dentro de cada conjunto, tienen el mismo formato, pero no son equivalentes con datos de los otros conjuntos, por lo tanto, no se puede utilizar un único modelo predictivo para comparar los tres nuevos datasets. Se crean tres modelos predictivos basados en redes convolucionales, que utilicen una estructura de capas similares entre sí, pero con distinta capa de entrada.

8. REFERENCIAS-BIBLIOGRÁFICAS

[1] *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. (n.d.). Retrieved July 29, 2020, from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[2] Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., & Tan, W. (2020). Detection of SARS-CoV-2 in Different Types of Clinical Specimens. In *JAMA - Journal of the American Medical Association* (Vol. 323, Issue 18, pp. 1843–1844). American Medical Association. https://doi.org/10.1001/jama.2020.3786

[3] Emery, S. L., Erdman, D. D., Bowen, M. D., Newton, B. R., Winchell, J. M., Meyer, R. F., Tong, S., Cook, B. T., Holloway, B. P., McCaustland, K. A., Rota, P. A., Bankamp, B., Lowe, L. E., Ksiazek, T. G., Bellini, W. J., & Anderson, L. J. (2004). Real-Time Reverse Transcription-Polymerase Chain Reaction Assay for SARS-associated Coronavirus. *Emerging Infectious Diseases*, *10*(2), 311–316. https://doi.org/10.3201/eid1002.030759

[4] Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Chen, Y., Su, J., Lang, G., Li, Y., Zhao, H., Xu, K., Ruan, L., & Wu, W. (2020). Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. *Applied Intelligence*, *2019*, 1–5. http://arxiv.org/abs/2002.09334

[5] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. http://code.google.com/p/cuda-convnet/

[7] Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., Xu, B., MedRxiv, M. C.-, 2020, U., Cai, M., Yang, J., Li, Y., Meng, X., & Xu, B.

(2020). A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). *MedRxiv*, 1–19. https://doi.org/10.1101/2020.02.14.20023028

[8] Duchesne, S., Gourdeau, D., Archambault, P., Chartrand-Lefebvre, C., Dieumegarde, L., Forghani, R., Gagne, C., Hains, A., Hornstein, D., Le, H., Lemieux, S., Levesque, M.-H., Martin, D., Rosenbloom, L., Tang, A., Vecchio, F., & Duchesne, N. (2020). Tracking and Predicting Covid-19 Radiological Trajectory Using Deep Learning on Chest X-Rays: Initial Accuracy Testing. *MedRxiv*, *1*(418), 2020.05.01.20086207. https://doi.org/10.1101/2020.05.01.20086207

[9] Ramadhan, M. M., Faza, A., Lubis, L. E., Yunus, R. E., Salamah, T., Handayani, D., Lestariningsih, I., Resa, A., Alam, C. R., Prajitno, P., Pawiro, S. A., Sidipratomo, P., & Soejoko, D. S. (2020). *Fast and accurate detection of Covid-19-related pneumonia from chest X-ray images with novel deep learning model*. http://arxiv.org/abs/2005.04562

[10] Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). *Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network*. http://arxiv.org/abs/2003.13815

[11] Asif, S., Wenhui, Y., Jin, H., Tao, Y., & Jinhai, S. (2020). *Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Networks*.