Autor: Iñigo Alonso Perea Campos
Docente: Dr. Luis Carlos Padierna García

Machtia

Proyecto Final

Perea Campos Iñigo Alonso, 783674, ia.pereacampos@ugto.mx

Introducción

El presente proyecto propone ejecutar 2 scripts de Python con algoritmos de Machine Learning usando un *dataset* de entrenamiento de 1,280 mamografías digitales etiquetadas en malignas y benignas. Un script para entrenar y comparar diferentes clasificadores lineales y otro para entrenar una red neuronal Transformers. De ambos scripts se rescata la métrica del *accuracy*, cómo métrica principal, y se exportan los modelos para un análisis posterior de especificidad y sensibilidad.

Al final, se analizarán los resultados para concluir que arquitecturas son prometedoras para ofrecer modelos óptimos.

Objetivo

Comparar la precisión que pueden alcanzar de distintas arquitecturas, clasificadores y una red neuronal, al clasificar una mamografía en benigna o maligna. Determinar que arquitecturas son prometedoras en la clasificación de mamografías.

Justificación

La importancia del proyecto radica en las competencias desarrolladas por el estudiante y en los beneficios de la propuesta para otros proyectos de investigación.

Este proyecto desarrollará "competencias necesarias para diseñar, construir y gestionar tecnologías… con un enfoque científico-práctico e interdisciplinario, y orientado a la atención de las necesidades de innovación tecnológica para el mejoramiento de la calidad de vida del ser humano"[1]. En otras palabras, permitirá al estudiante desarrollar sus habilidades tecnológicas y ponerlas al servicio de su comunidad.

La presente propuesta permitirá analizar diferentes arquitecturas de machine learning, tanto clasificadores lineales como una red neuronal, y con ello poder determinar que arquitectura podría ofrecer un modelo con mayor precisión. Al final, si tratásemos de obtener el mejor modelo con cada arquitectura disponible de machine learning, el número de parámetros a ajustar tendería a ser infinitesimal y la relación costo-beneficio sería extremadamente baja. Por otro lado, con esta propuesta podemos discriminar las arquitecturas más prometedoras y enfocar esfuerzos en el ajuste de sus parámetros,

[1] UGTO, 2017

reduciendo tiempo y recurso necesario para alcanzar un modelo óptimo de clasificación de mamografías en malignas o benignas.

Marco Teórico

Contexto

En México, el acceso oportuno al diagnóstico de cáncer de mama enfrenta importantes desafíos, especialmente debido a la insuficiencia de especialistas en oncología. Según la ENSANUT Continua 2022, 20.6 millones de mujeres entre 40 y 69 años residían en el país, de las cuales solo el 20.2% (4.2 millones) se realizó una mamografía en 2023. De estás, 85.6% recibió resultados, y 5.6% (19,900 mujeres) fueron diagnosticadas[2].

En 2021, México contaba con solo 1,043 oncólogos en el sector público de salud[3] y, según la UNAM, se estiman otros 1,000 en el sector privado[4], sumando aproximadamente 2,000 especialistas en todo el país. Esto resulta insuficiente frente a la demanda potencial. Si se distribuyera de manera equitativa, cada oncólogo tendría que atender a unas 10,300 pacientes en edad de realizarse una mamografía, incluyendo diagnósticos confirmados de cáncer. Dicha situación evidencia la sobrecarga para los especialistas, además de las limitaciones estructurales para garantizar la detección y tratamiento oportunos en el cáncer de mama, especialmente en regiones con una grave desigualdad en la distribución de recursos médicos.

Las mamografías digitales

Las mamografías digitales representan un avance significativo en el diagnóstico de temprano de cáncer de mama, proporcionando imágenes de alta resolución, con píxeles de 50 µm[5], que permiten detectar anomalías sutiles en el tejido mamario. Los archivos generados por los mamógrafos son exportados en formato DICOM, que es el estándar de telecomunicaciones para imágenes médicas. Las matrices de imagen de estos archivos son de gran tamaño, alcanzando resoluciones de hasta 4800x6000[6] píxeles dependiendo del fabricante. En la imagen 1 de anexos se detallan los datos técnicos las imágenes de distintos mamógrafos.

La calidad de la imagen depende, entre otros factores, del rango de energía de los fotones, que en una mamografía deben oscilar entre 14 y 25 keV [7,8] para una mejor relación entre las métricas Contraste Inherente, Relación Señal-Ruido y Dosis Glandular, y optimizando el contraste entre tejidos blandos. Según la Sociedad Americana de Cáncer, la dosis promedio de una mamografía común, que incluye dos tomas por seno, es de aproximadamente 0.4

[2] INEGI, 2023
[3] INEGI, 2023
[4] Frías Cienfuegos, 2023
[5] Chevalier, M., & Torres, R. (2010)
[6] Chevalier, M., & Torres, R. (2010)
[7] Xunta de Galicia, 2000
[8] Hammerstein, et al, 1979

mSv, una cantidad significativamente menor que los 3 mSv de exposición anual promedio por radiación ambiental en Estados Unidos.[9]

Las mamografías digitales permiten identificar características clave como microcalcificaciones, que son depósitos de hidroxiapatita o fosfato de calcio de entre 0.1 y 0.2 mm de diámetro; tumores, que son masas de tejido anormal de pocos milímetros; y estructuras filamentosas, como extensiones fibrosas en el tejido adiposo[10,11]. Estas propiedades hacen de la mamografía digital una herramienta crucial para la detección temprana de cáncer de mama, mejorando las posibilidades de un tratamiento efectivo.

Scikit-Learn

Scikit-Learn es una biblioteca de Python muy popular en el campo de Machine Learning debido a su facilidad de uso, amplia documentación y versatilidad. Esta herramienta proporciona una variedad de algoritmos para tareas de clasificación, regresión y agrupamiento. Además, tiene utilidades para preprocesamiento de datos y evaluación de modelos, como la validación cruzada, precisión, sensibilidad y puntaje F1. Esta biblioteca se integra con Numpy y Pandas, facilitando la manipulación de datos.[12]

Clasificadores

Los clasificadores son algoritmos de aprendizaje automático diseñados para asignar etiquetas a los datos basados en patrones aprendidos durante el entrenamiento. Estos algoritmos pertenecen al Machine Learning según la taxonomía de la imagen 2 de los anexos[13]. Cada uno de los clasificadores se adapta a distintos tipos de problemas según la naturaleza de los datos y las necesidades del modelo, permitiendo resolver tareas como detección de anomalías, clasificación de imágenes o categorización de texto. La biblioteca Scikit-Learn tiene 41 clasificadores para entrenar, los cuales se enlistan en la sección de anexos.

PyTorch

PyTorch es una biblioteca de código abierto utilizada en el desarrollo de modelos de Deep Learning debido a su flexibilidad y facilidad de uso. Diseñada por Facebook AI Research (FAIR), está librería permite construir y entrenar redes neuronales con un enfoque dinámico, facilitando la depuración y experimentación. Se integra con las GPUs, acelerando significativamente las tareas intensivas, como el procesamiento de imágenes, análisis de series temporales y procesamiento de lenguaje natural. [14]

Transformers

Los Transformers son un tipo de arquitectura de redes neuronales que han revolucionado el campo del aprendizaje profundo, inicialmente diseñados para tareas de procesamiento

[9] Sociedad Americana de Cáncer, 2020
[10] Xunta de Galicia, 2000
[11] Arancibia Hernandez, et al, 2016
[12] Scikit-learn.org. (s.f.)
[13] Bedolla, E, et al. (2021)
[14] PyTorch.org. (s.f.)

de lenguaje natural, ahora se han adaptado para el procesamiento de imágenes mediante el uso de arquitecturas como *Vision Tranformers* (ViT). A diferencia de los enfoques tradicionales basados en redes convolucionales (CNNs), ViT fragmenta una imagen en parches, los cuales se tratan como palabras en una secuencia, permitiendo al modelo capturar relaciones globales entre características a lo largo de toda la imagen. La implementación de ViT se apoya en el mecanismo de "atención" (attention mechanism) que asigna pesos a las diferentes partes de la imagen según su relevancia para la tarea.

En la implementación de ViT, se preparan las imágenes usando un preprocesamiento riguroso que incluye técnicas como normalización y redimensionamiento, asegurando que cumplan con los requisitos del modelo. Gracias a los modelos pre entrenados se reduce la cantidad de datos etiquetados requeridos para tareas generales.[15]

La arquitectura de ViT se observa en la imagen 1, tomada directamente del artículo original[16]:

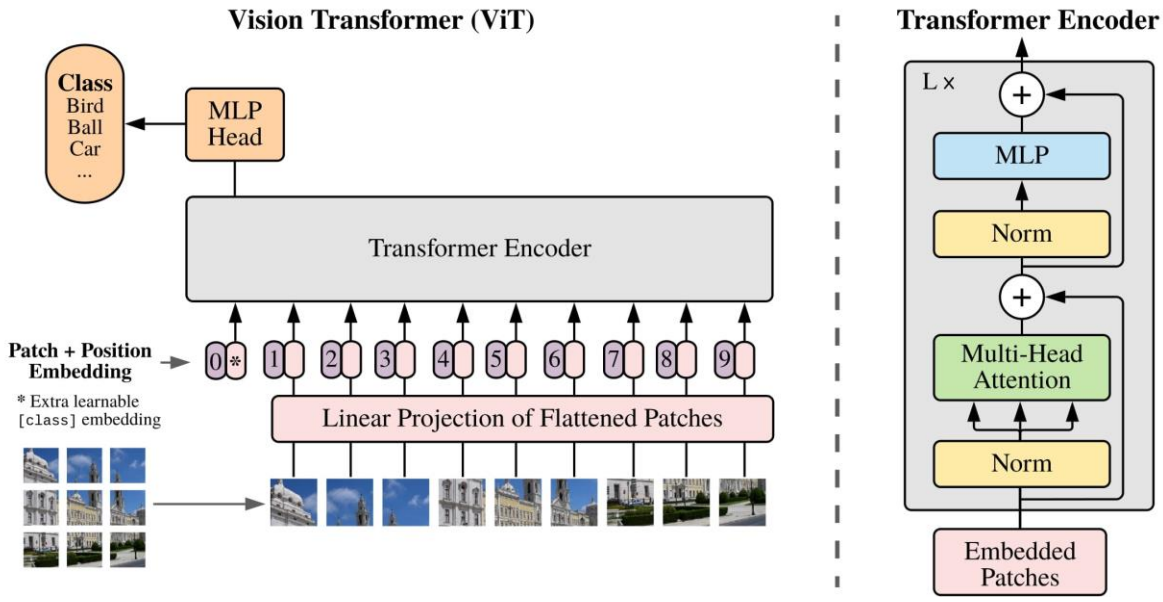


Imagen 1. Arquitectura de ViT (Dosovitskiy, A., et al. 2021)

Actividades de programación

La preparación del *dataset*, la escritura de los scripts y su ejecución estuvo a cargo del estudiante que suscribe.

[15] Dosovitskiy, A., et al. (2021)

[16] Dosovitskiy, A., et al. (2021)

Referencias

Arancibia Hernandez, P. L., Taub Estrada, T., López Pizarro, A., Díaz Cisternas, M. L., & Sáez Tapia, C. (2016). Calcificaciones mamarias: descripción y clasificación según la 5.a edición BI-RADS. *Revista chilena de radiología*. doi:http://dx.doi.org/10.1016/j.rchira.2016.06.004

Bedolla, E., Padierna, L. C., & Castañeda-Priego, R. (2021). Machine learning for condensed matter physics. *Journal of Physics: Condensed Matter, 33*(5). doi:10.1088/1361-648X/abb895

Chevalier, M., & Torres, R. (2010). *Mamografía digital.* Obtenido de Revista de Física Médica, 11: http://revistadefisicamedica.es/index.php/rfm/article/view/90/91

Dosovitskiy, A., & al., e. (2015). Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. *EEE Transactions on Pattern Analysis and Machine Intelligence*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations (ICLR)*. doi:https://doi.org/10.48550/arXiv.2010.11929

Frías CIenfuegos, L. (2 de Febrero de 2023). *Carece México de los oncólogos necesarios para atender a la población*. Obtenido de Gaceta UNAM: https://www.gaceta.unam.mx/carece-mexico-de-los-oncologos-necesarios-para-atender-a-la-poblacion/#:~:text=Recuerda%20que%20la%20Secretar%C3%ADa%20de,y%20para%20la%20sociedad%20mexicana%E2%80%9D.

González, A. (2023). *Análisis de datos de mamografías con Machine Learning*. Obtenido de Panama Hitek: https://panamahitek.com/analisis-de-datos-mamograficos-con-machine-learning/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge: MIT Press .

Hammerstein, G., Miller, D. W., White, D., Masterson, M., Woodard, H., & Laughlin, J. (1979). Absorbed radiation doses in mammography. *Radiology, 130(2)*, 485-91. doi:10.1148/130.2.485

Huggingface.co. (s.f.). *Vision Transformer (ViT)*. Obtenido de https://huggingface.co/docs/transformers/model_doc/vit

INEGI. (19 de Octubre de 2023). *Estadísticas a propósito del día internacional de la lucha contra el Cáncer de Mama.* Obtenido de Inegi.org.mx: https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2023/EAP_CMAMA23.pdf

Ortiz, N. A. (2019). *Estructuración de protocolo nacional de control de calidad para mamografía CR y estimación de dosis glandular media, para su aplicación en un departamento del país.*

Colombia: Universidad Nacional de Colombia. Obtenido de https://docplayer.es/198591859-Napoleon-alberto-ortiz-guevara.html

PyTorch.org. (s.f.). *PyTorch - transformers*. Obtenido de https://pytorch.org/hub/huggingface_pytorch-transformers/

PyTorch.org. (s.f.). *Transformer - PyTorch 2.5 documentation*. Obtenido de https://pytorch.org/docs/stable/generated/torch.nn.Transformer.html

Rothman, D. (2021). *Transformers for Natural Language Processing.* Packt Publishing.

Scikit-learn.org. (s.f.). *Classifier comparison*. Obtenido de https://scikit-learn.org/1.5/auto_examples/classification/plot_classifier_comparison.html

Scikit-learn.org. (s.f.). *Scikit-learn*. Obtenido de https://scikit-learn.org/stable/

Sociedad Americana de Cáncer. (2020). *Conceptos básicos del mamograma*. Obtenido de cancer.org: https://www.cancer.org/es/cancer/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/mamogramas/conceptos-basicos-del-mamograma.html#:~:text=Los%20equipos%20modernos%20emplean%20bajas,de%20la%20dosis%20de%20radiaci%C3%B3n

Universidad de Guanajuato. (2017). *Licenciatura en Ingeniería Física*. Obtenido de http://fisica.ugto.mx/index.php/oferta-educativa-dci/oe-licenciatura/licenciatura-ingfisica-2016

Xunta de Galicia, Conselleria de Sanidade e Servicios Sociaís. (2000). *Control de Calidad en Mamografía. Guía práctica.* España: Dirección Xeral de Saúde Pública. Obtenido de https://extranet.sergas.es/catpb/Docs/cas/Publicaciones/Docs/SaludPublica/PDF10-133.pdf#page=54

Autor: Iñigo Alonso Perea Campos
Docente: Dr. Luis Carlos Padierna García

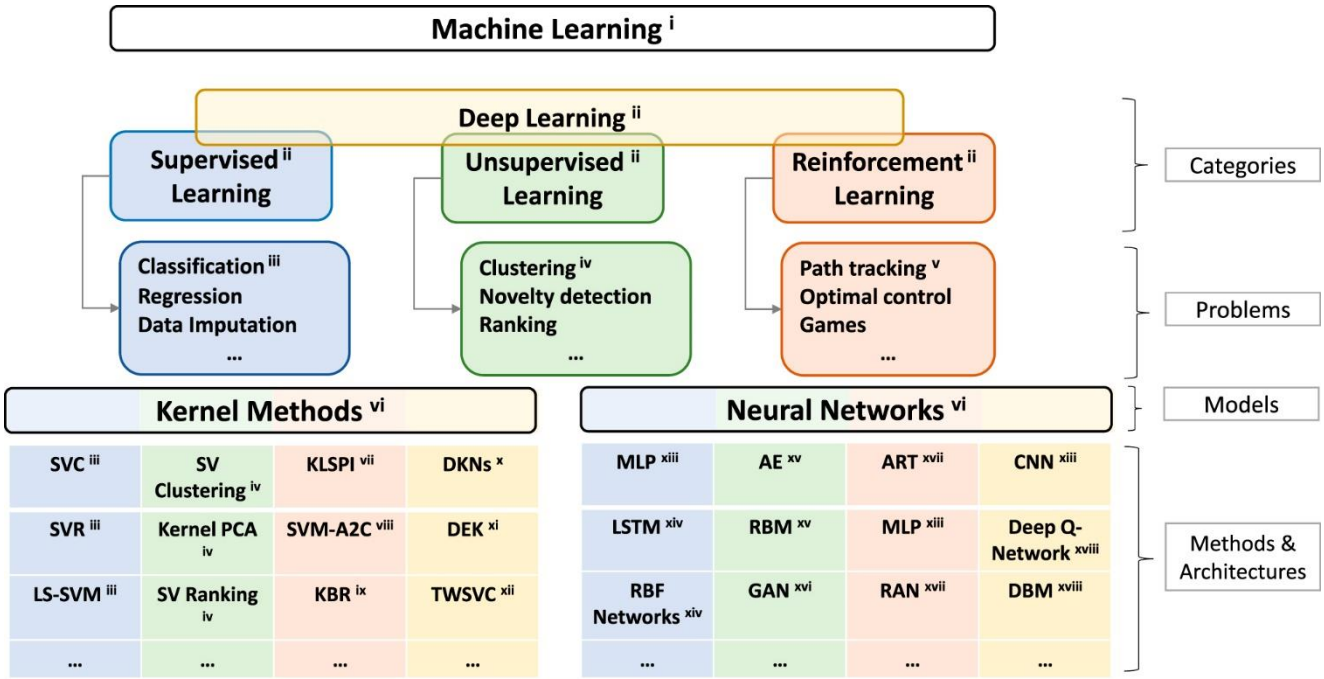
Anexos

Fabricante	Modelo	Tecnología	Dimensiones detector	Tamaño píxel (μm)	Profundidad bit	Tamaño matriz	Tamaño Imagen (MB)
Fósforo fotoestimulable (CR)							
AGFA	CR 85/35X DX-M	BaSrFBrI:Eu CsBr:Eu	18 × 24 24 × 30	50	12	3560 × 4640 4760 × 5840	32 ~50
Fuji	Profect (todos los modelos)	BaF(BrI):Eu	18 × 24 24 × 30	50	12	3540 × 4740 4728 × 5928	32,8 ~50
Carestream	DirectView CR950/975	BaFBr:Eu	18 × 23 23 × 29	50	12	3584 × 4784 4800 × 6000	33,5 ~50
Konica	Pureview	BaFI:Eu	35 × 43	43,8	12	~8000 × 9800	
Konica	Regius 190: RP-6M/7M CP-1M	BaFI:Eu CsBr aguja de fósforo	18 × 24 24 × 30	43,8	12	~4360 × 5726 ~5760 × 7096	48,8
Philips*	Cosima X Eleva	BaF(BrI):Eu	18 × 24 24 × 30	50	12	3540 × 4740 4728 × 5928	32,8 ~50
Detectores integrados (DR): panel plano							
GE	Senographe 2000D	CsI sobre a-Si	19 × 23	100	14	1914 × 2294	8,8
GE	Senographe DS	CsI sobre a-Si	19 × 23	100	14	1914 × 2294	8,8
GE	Senographe Essential	CsI sobre a-Si	24 x 31	100	14	2394 x 3062	14
Siemens	Mammomat Novation	a-Se	24 × 29	70	14	3328 × 4084	27,2
Siemens	Mammomat Inspiration	a-Se	24 × 30	85	13	2800 × 3518	24
Hologic	Selenia	a-Se	24 × 29	70	14	3328 × 4096	27,2
IMS	Giotto	a-Se	24 × 30	85	13	2816 × 3584	20
Planmed Oy	Nuance	a-Se	17 × 24 24 × 30	85	13	2016 × 2816 2816× 3584	16 20
Fuji	AMULET	a-Se con Teconología DOS	18 × 24 24 × 30	50	14	3540 × 4740 4728 × 5928	32,8
Detectores integrados (DR): sistemas de barrido							
Sectra	MDM L30	Si contador cuántico	24 x 26	50	16	4915 × 5355	51,4
XCounter		Gas presurizado	24 x 30	50	16	4800 x 6000	---

*Utiliza las placas de Fuji

Imagen 1. Relación de fabricantes y tecnologías en uso de mamografía digital al 2010. Chevalier, M., & Torres, R. (2010)

Autor: Iñigo Alonso Perea Campos
Docente: Dr. Luis Carlos Padierna García



Clasificadores de la paquetería Scikit-Learn:

1. 'AdaBoostClassifier'
2. 'BaggingClassifier'
3. 'BernoulliNB'
4. 'CalibratedClassifierCV'
5. 'CategoricalNB'
6. 'ClassifierChain'
7. 'ComplementNB'
8. 'DecisionTreeClassifier'
9. 'DummyClassifier'
10. 'ExtraTreeClassifier'
11. 'ExtraTreesClassifier'
12. 'GaussianNB'
13. 'GaussianProcessClassifier'
14. 'GradientBoostingClassifier'
15. 'HistGradientBoostingClassifier'
16. 'KNeighborsClassifier'
17. 'LabelPropagation'
18. 'LabelSpreading'
19. 'LinearDiscriminantAnalysis'
20. 'LinearSVC'
21. 'LogisticRegression'

Autor: Iñigo Alonso Perea Campos
Docente: Dr. Luis Carlos Padierna García

22. 'LogisticRegressionCV'
23. 'MLPClassifier'
24. 'MultiOutputClassifier'
25. 'MultinomialNB'
26. 'NearestCentroid'
27. 'NuSVC'
28. 'OneVsOneClassifier'
29. 'OneVsRestClassifier'
30. 'OutputCodeClassifier'
31. 'PassiveAggressiveClassifier'
32. 'Perceptron'
33. 'QuadraticDiscriminantAnalysis'
34. 'RadiusNeighborsClassifier'
35. 'RandomForestClassifier'
36. 'RidgeClassifier'
37. 'RidgeClassifierCV'
38. 'SGDClassifier'
39. 'SVC'
40. 'StackingClassifier'
41. 'VotingClassifier'