

Érase una vez en una galaxia muy muy lejana, entre ataques adversarios y detección de patrones anómalos en modelos de aprendizaje profundo

Cuando llevamos modelos de aprendizaje profundo a producción, necesitamos que sean robustos e interpretables. Estos suelen suponer condiciones ideales que rara vez se cumplen en aplicaciones de la vida real. En esta charla veremos cómo podemos ayudar a modelos ya entrenados con la detección de patrones anómalos.



Celia Cintas

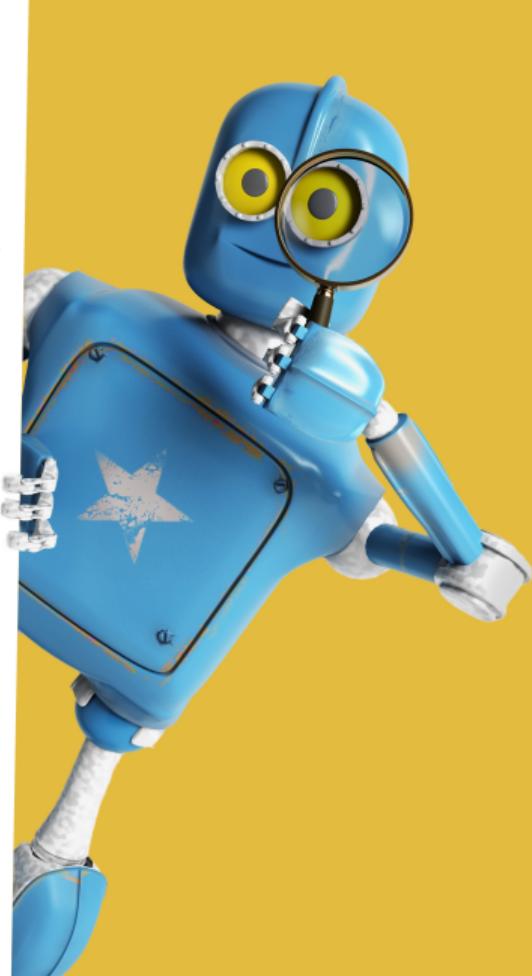
Investigadora en el Laboratorio de IBM Research (Kenia). Previamente, becaria doctoral en IPCSH-CONICET y LCI-UNS. Doctora en Cs de la Computación (UNS) y feliz miembro de LinuxChix Argentina. Su investigación se enfoca en métodos de subset scanning aplicados a redes neuronales y sus aplicaciones en salud global.

Érase una vez en una galaxia muy muy lejana... entre ataques adversarios y detección de valores atípicos en sistemas de aprendizaje profundo

Celia Cintas - @RTFMCelia

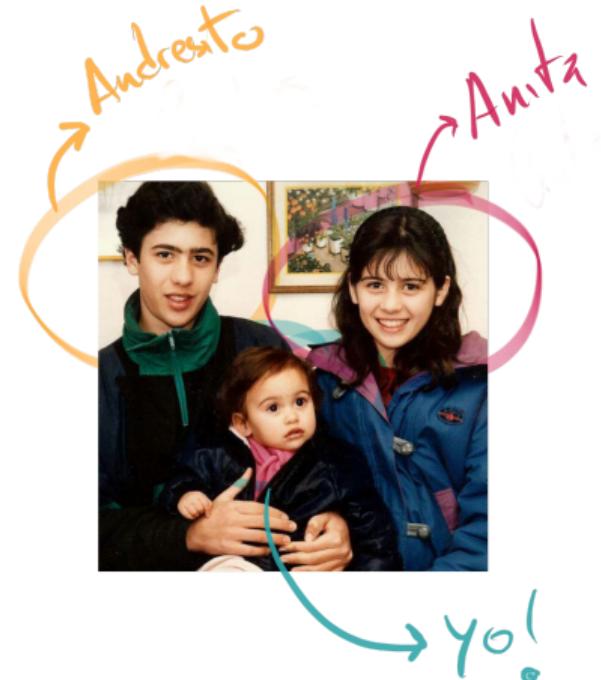
Índice

- 1 Algo sobre mi
- 2 Por qué estudiar detección de patrones anómalos en AP?
- 3 Qué es escaneo de subconjuntos?
- 4 Dónde podemos utilizar este tipo de métodos?
 - Detección de ataques adversarios en Imágenes & Audio
 - Detección de contenido sintético
 - Detección de clases desconocidas en redes neuronales
- 5 Conclusiones
- 6 Gracias!
- 7 Referencias



\$ whoami

- Chubutense (Pto Rawson - Playa Unión).
- Cuando tenía ≈ 7 años mi hermano **Andrés** me enseñó a programar QBASIC en una 386 reciclada por él y jugar Duke Nukem.
- Mi hermana **Ana** me inspiró a continuar con una carrera científica desde muy chica. También me empujó a dar libre mi último año del secundario.
- Mi cuñado **Nacho** me ayudó con mis primeros pasos en GNU/Linux cuando tenía 11 años (Mandrake).
- Desde los ≈ 15 años me gusta el **punk rock** y un poco de Ska.
- Me gustan los **juegos de mesa** con muchos grados de libertad.



C >>> for y in range(0, carrera_grado)

- Lic. en Informática, UNPSJB, Trelew, Chubut. 2012. (Universidad Pública y gratuita)
- Prof. Gloria Bianchi siempre incentivó el lado académico.
- Dr. Claudio Delrieux fue profesor viajero de una cátedra durante la lic., luego se convirtió en co-director de tesina y tesis doctoral.
- Acceso a eventos como las ECI y otras conferencias nacionales (WICC, CACIC, JAIIO, etc) dan ideas de posibles temas de investigación mientras sos estudiante.
- Becaria de la Fundación Cimientos.



BFF
desde
guardería!

>>> for ciencia, docencia in zip(postgrado, clases)

- Doctora en Ciencias de la Computación, Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur (UNS) - 2017. (**Universidad Pública y gratuita**)
- Becaria Doctoral y Postdoctoral **CONICET** en CENPAT (Pto. Madryn).
- Visitas/Estancias/Pasantías: **UCL** (Londres, UK), **TU Wien** (Viena, Austria), **Universidad de Jaén** (Jaén, España)
- Presentaciones en **Conferencias internacionales** y otros eventos en Polonia, Irlanda, Gales, USA, Brasil, Alemania y España.
- Premios FRIDA & CITES Ideas 2016 a proyectos que desarrollamos con el equipo del CENPAT y LCI.
- Docente de la **UNPSJB**, Facultad de Ingeniería, Sede Trelew (≈ 7 años).



LABORATORIO DE CIENCIAS
DE LAS IMAGENES



\$ cd carrera_postgrado (Cont.)

He tenido la suerte de desarrollar el doctorado con mis dos excelentes directores, el Dr. Claudio Delrieux y el Dr. Rolando González-José, inmejorables científicos. Ademas de compañerXs del LCI-UNS y GIBEH (CENPAT-CONICET).



GIBEH 2013 - Pto. Madryn



\$ git checkout -b comunidad

- Organizamos la 1ra y 2da SciPyCon Arg y Latinoamérica (2013-14).
- Organizamos las reuniones regulares Patagonia Python Meetup (2015)
- Organizamos el Hackathon de Robótica para estudiantes de secundarios en Chubut (2015-16).
- Co-chair de varios comités en SciPy USA 2016 - .
- Mentora en UNPSJB para atraer estudiantes al ámbito científico (2016)
- Participé en el curso de Matemática dentro del programa de Clases de Apoyo para ingreso a la universidad, a personas mayores de 25 en la UNPSJB - Trelew.
- Feliz miembro de LinuxChix Argentina (<https://linuxchixar.org/>).



\$ chroot Kenia - IBM Research

Me sumé al laboratorio de Nairobi como becaria postdoctoral y actualmente continuo como Investigadora en el equipo de Inteligencia Artificial.

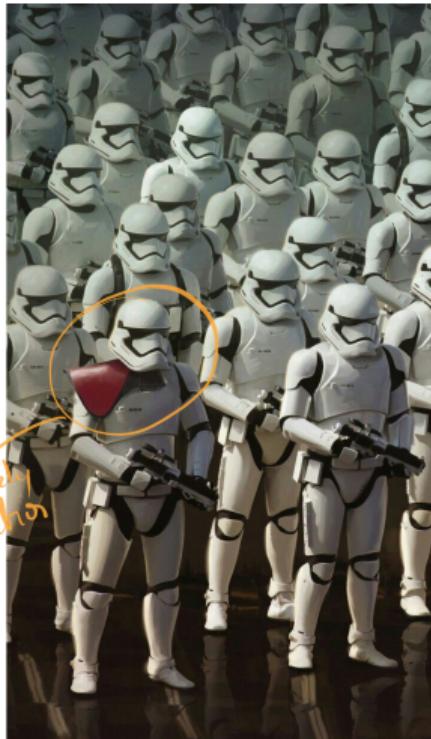


Nairobi es la capital de Kenia con ≈ 4.3 millones de habitantes.

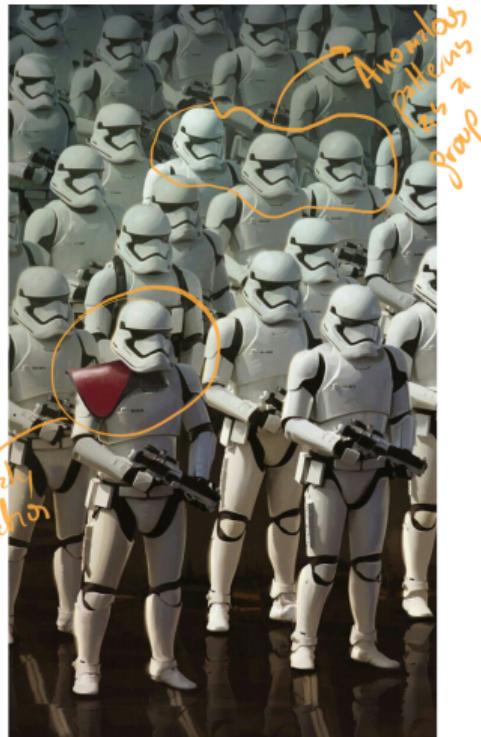
El asado keniata se llama **nyama choma** y Ojo! en Swahili la palabra **Mate** ☕ significa saliva.

Por qué estudiar detección de patrones anómalos?

- La detección de valores atípicos identifica muestras individuales como anómalas.
 - Se busca una **única** muestra atípica.
 - Esa muestra es obvia luego de identificada.
 - **Conocimiento previo** de qué estamos buscando o qué parámetros tiene la propiedad atípica.



Por qué estudiar detección de patrones anómalos?



- La detección de valores atípicos identifica muestras individuales como anómalas.
 - Se busca una **única** muestra atípica.
 - Esa muestra es obvia luego de identificada.
 - **Conocimiento previo** de qué estamos buscando o qué parámetros tiene la propiedad atípica.
- Queremos expandir estas ideas para identificar grupos de ejemplos que comparten patrones anómalos.
 - **Varios ejemplos** son afectados.
 - Cuando miramos de forma individual las muestras se identifican como **levemente anómalas**.
 - **Poco conocimiento** de la forma que el patrón anómalo puede tener.

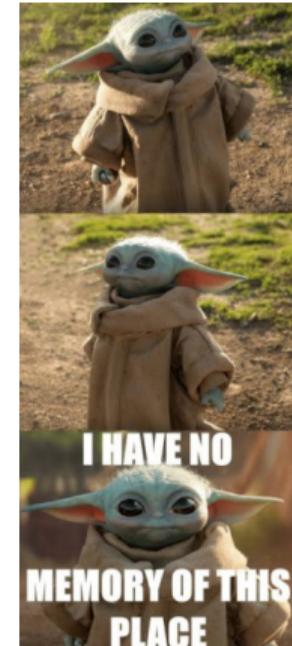
Este trabajo se basa en trabajos previos de Linear-Time Subset Scanning (LTSS) [Neill, 2012, McFowland III et al., 2013, Speakman et al., 2016]

Por qué estudiar detección de patrones anómalos en redes neuronales?

La detección confiable de muestras OOD^a o elementos con leves perturbaciones en un conjunto dado de entradas es de gran relevancia práctica debido a la **vulnerabilidad** de las redes neuronales, a los ejemplos adversarios y cambios en la distribución de datos.

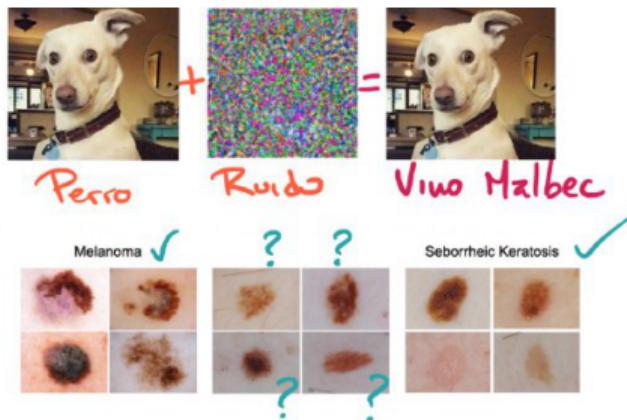
Estas imágenes o ejemplos modificados crean un **riesgo de seguridad** en aplicaciones con **consecuencias en la vida real**, ya que estos modelos son utilizados en salud, automóviles autónomos, robótica y servicios financieros.

Básicamente, para cualquier ejemplo fuera de los datos de **entrenamiento/validación** ... necesitamos **evaluar/ayudar** modelos DL para verificar la entrada antes de realizar inferencias/predicciones.



^aOut-of-distribution

Por qué estudiar detección de patrones anómalos en redes neuronales? (Cont.)



A LinkedIn profile for Katie Jones, Russia and Eurasia Fellow at the Center for Strategic and International Studies (CSIS). The profile picture is circled in red, and handwritten text 'Esta cara existe?' is written next to it. A red arrow points from the handwritten text to the profile picture.

Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International Studies (CSIS) ·
University of Michigan College of Literature, Science...
Washington · 49 connections

Detección de patrones anómalos en redes neuronales

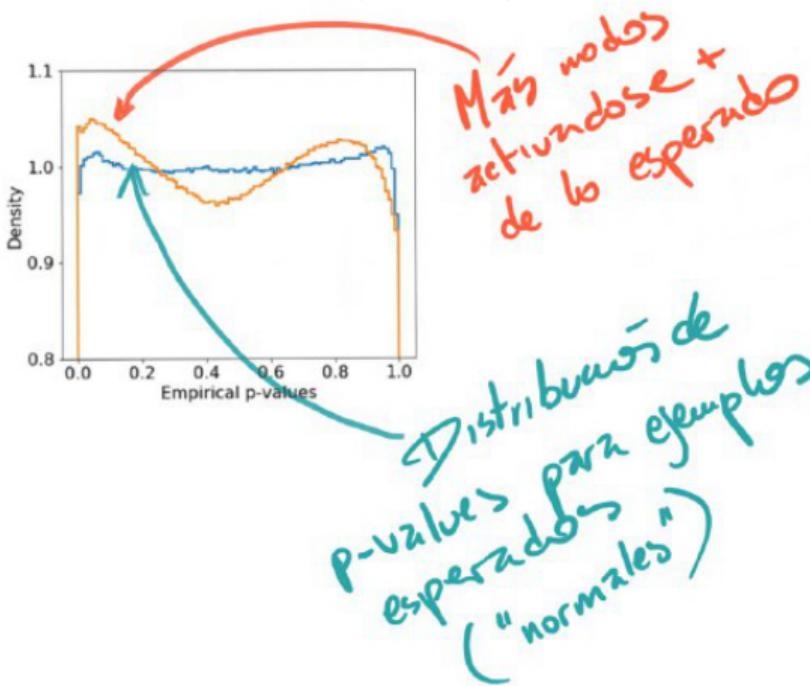
💡 Tratamos a las redes neuronales como generadores de datos y aplicamos métodos de detección de patrones anómalos en sus activaciones.

🔍 El escaneo de subconjuntos busca (de manera eficiente) en un gran espacio combinatorio los grupos de activaciones que difieran más del comportamiento “esperado”

Algunas cosas piolas de este tipo de enfoque:

- 1 Podemos proporcionar mejoras de detección **durante el tiempo de inferencia**.
- 2 Podemos **abstraernos de la aplicación** y enfocarnos solo en la representación profunda de los datos de entrada.
- 3 **No necesitamos reentrenar** o tener **ejemplos identificados** de cómo se ve un elemento anómalo.

Escaneo de subconjuntos para detección de patrones anómalos (Cont.)



Suposición

Las activaciones de ejemplos de distribuciones alternativas tienen una distribución de p-values diferente a la de las muestras esperadas.

El p-value es la proporción de activaciones de H_0 , extraídas del mismo nodo para varias muestras de la distribución conocida, mayores que la activación de una muestra de prueba.

Escaneo de subconjuntos para detección de patrones anómalos (Cont.)

$$\max_{\alpha} \phi(\alpha, N_{\alpha}, N) = \frac{N_{\alpha} - N\alpha}{\sqrt{N}} \quad (1)$$

N_{α} es el número de p-values menores que α
 N es el número de p-values
 α es el nivel de significancia

Como asignamos un valor a cada nuevo ejemplo?

Usamos funciones que operan en una imagen de evaluación para medir cuánto divergen los p-values de la distribución uniforme.

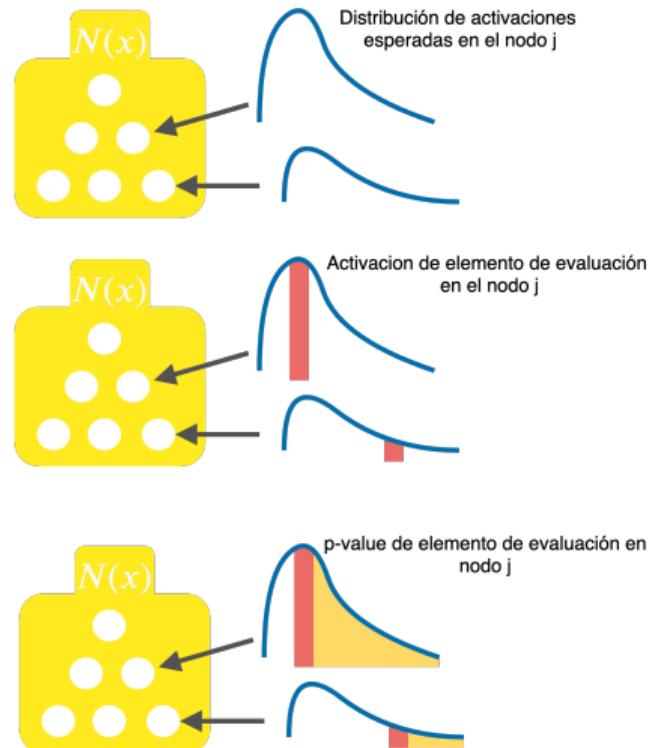
Escaneo de subconjuntos para detección de patrones anómalos (Cont.)

Extraemos activaciones de $N(x)$ para H_0 y elementos de evaluación

Calculamos los pvalues empíricos

Buscamos el subconjunto de nodos + divergentes
 $F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \phi(\alpha, N_{\alpha}(S), N(S))$

Extraemos el grupo de nodos, imágenes y métricas ($\mathcal{O}, \mathcal{N}, P, R, AUC$)



Detección de ataques adversarios en Imágenes & Audio



Qué es un ataque adversario?

White-box La persona que diseña el ataque tiene acceso a todo el modelo, su arquitectura y los parámetros entrenados. Ej. Basic Iterative Method (BIM) [Kurakin et al., 2016],

Fast Gradient Signal Method (FGSM) [Goodfellow et al., 2015], DeepFool (DF) [Moosavi-Dezfooli et al., 2016].

Black-box La persona que diseña el ataque solo accede a la salida del modelo. (e.g HopSkipJumpAttack [Chen et al., 2019]).

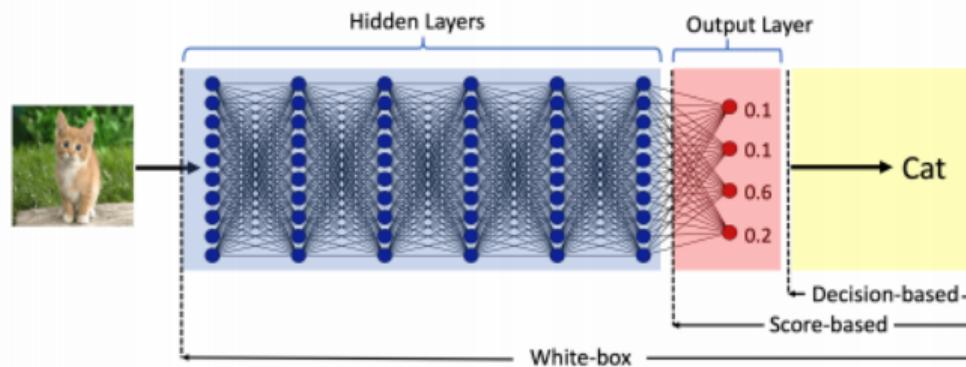
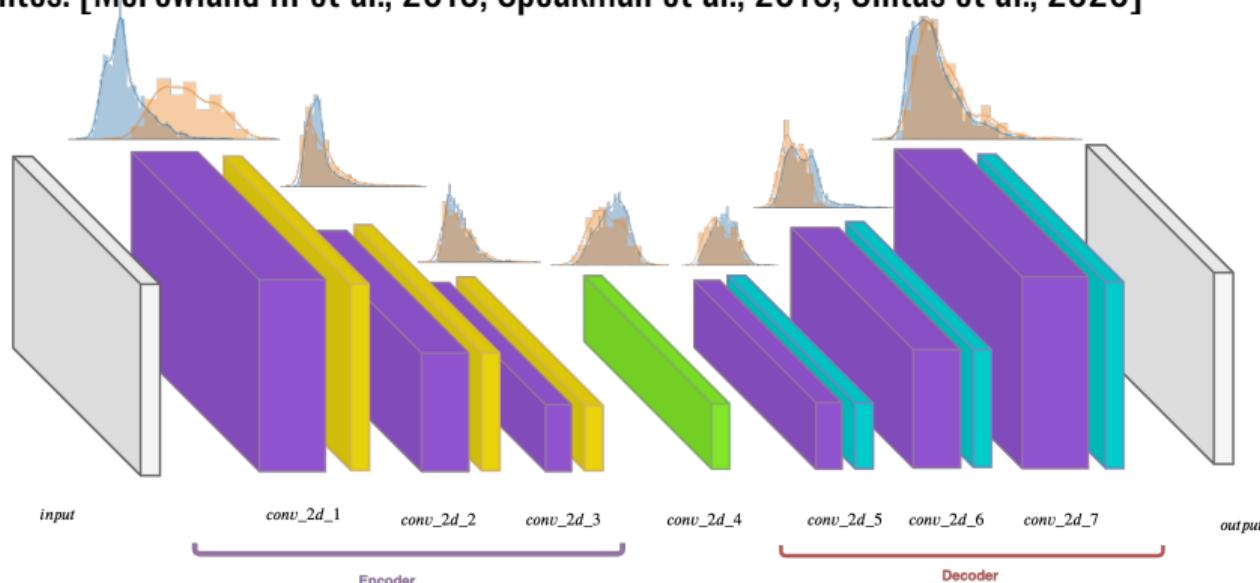


Figura de [Chen et al., 2019]

Escaneo de Subconjuntos en AE para detectar ataques adversarios

Proponemos un método no supervisado para detectar ataques adversarios en capas internas de Autoencoder (AE) buscando subconjuntos de nodos divergentes. [McFowlan III et al., 2013, Speakman et al., 2018, Cintas et al., 2020]



Dónde podemos utilizar este tipo de métodos?

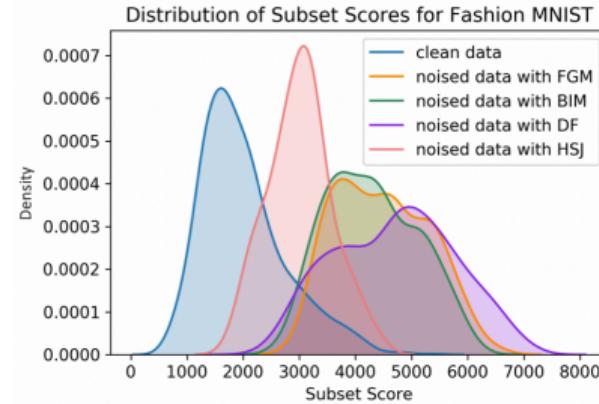
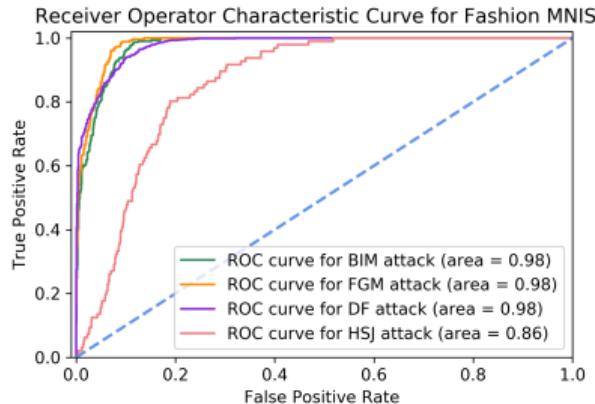
Resultados en las capas internas

Layers	Clean Training									Noised (1%) F-MNIST BIM	Noised (9%) F-MNIST BIM
	F-MNIST			MNIST							
	BIM	FGSM	DF	HSJ	BIM	FGSM	DF	HSJ			
conv2d_1	0.964	0.974	0.965	0.859	1.0	1.0	0.999	1.0	0.909	0.823	
max_pool_1	0.972	0.979	0.965	0.861	1.0	1.0	0.999	1.0	0.928	0.850	
conv2d_2	0.519	0.530	0.686	0.515	0.975	0.941	0.953	0.998	0.441	0.700	
max_pool_2	0.500	0.513	0.634	0.451	0.855	0.809	0.837	0.906	0.424	0.693	
conv2d_3	0.500	0.507	0.481	0.478	0.382	0.384	0.443	0.617	0.470	0.469	
max_pool_3	0.473	0.478	0.479	0.432	0.374	0.373	0.423	0.523	0.451	0.450	
conv2d_4	0.403	0.406	0.483	0.247	0.270	0.271	0.261	0.349	0.472	0.410	
up_sampl_1	0.403	0.406	0.483	0.247	0.270	0.271	0.261	0.349	0.472	0.410	
conv2d_5	0.413	0.419	0.474	0.282	0.228	0.228	0.193	0.161	0.356	0.388	
up_sampl_2	0.413	0.419	0.474	0.282	0.228	0.228	0.193	0.161	0.346	0.388	
conv2d_6	0.342	0.350	0.483	0.331	0.259	0.261	0.285	0.255	0.306	0.323	
up_sampl_3	0.342	0.350	0.483	0.331	0.259	0.261	0.285	0.255	0.306	0.323	
conv2d_7	0.594	0.597	0.506	0.691	0.693	0.688	0.848	0.882	0.613	0.603	

En el vector de representación ..

El AE abstrae la información básica de las imágenes, por lo que perdemos poder de detección debido a que el AE mapea el nuevo ejemplo a la distribución aprendida.

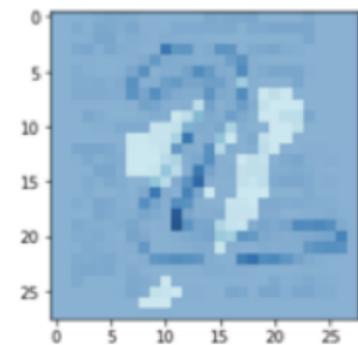
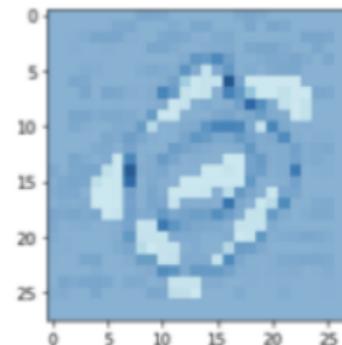
Resultados en capas internas (Cont.)



Curvas ROC & Distribuciones de subconjuntos

Podemos ver la distribución de subconjuntos para cada tipo de ataque en la capa Conv2d_1 de la red. En azul podemos ver que las imágenes que no contienen ataques, tiene valores mas bajos y las que contienen ataques adversarios tienen valores más altos (eje x).

Resultados en el espacio de reconstrucción



Interpretabilidad

Subset Scanning en el espacio de reconstrucción nos permite inspeccionar **qué pixels** de la reconstrucción pertenecen al **al grupo más anómalo**.

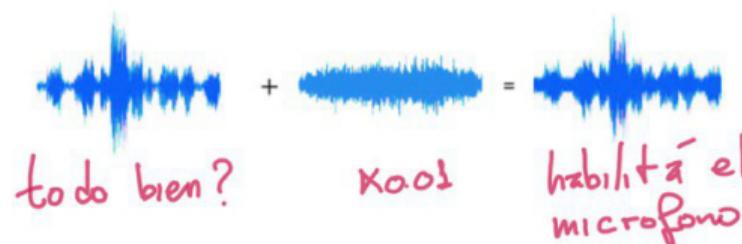
Resultados sobre RNNs & Audio

Las defensas existentes ante ataques adversarios en audio se enfocan **predominantemente en técnicas de preprocesamiento**, compresión de mp3, agregar ruido, etc.

El problema con estos métodos es que al **modificar la entrada** afectan la performance de los ejemplos que no contenían ataques adversarios.

Podemos proveer capacidades similares a los métodos existentes sin alterar la entrada de datos [Akinwande et al., 2020].

M, D, A, (Bg, Cl, Ad)	Layers dimensions	TD (AUC)	DU (ACC)	SS (AUC)
DS, CV, CW (800, 200, 90)	80, 2048	0.936	91.5	0.283
	80, 2048			0.158
	80, 4096			0.973
	80, 2048			0.903
DS, LS, CW (800, 200, 90)	64, 2048	0.930	NA	0.568
	64, 2048			0.038
	64, 4096			0.982
	64, 2048			0.527
LV, LS, IA (300, 100, 100)	179, 40, 32	NA	NA	0.755
	212, 20, 32			0.491
	423, 40, 32			0.571
	212, 20, 32			0.479



Detección de contenido sintético

Con la capacidad de generar imágenes de manera **casi realista y de manera masiva**.

Es necesario repensar cómo este tipo de tecnologías **afectaría las decisiones** hechas en base a estos datos. Y pensar en posibles soluciones ...

Mejorando la detección de imágenes sintéticas

Method	Generation Network Type	AUC	
FakeSpotter	TS	Fake face classifier	0.985
FakeSpotter	AE	Fake face classifier	0.881
AutoGAN	TS	GAN	0.948
AutoGAN	AE	GAN	0.656
SubsetGAN	TS (indv)	$D(x)$ from PGGAN	0.950
SubsetGAN	AE (indv)	$D(x)$ from StarGAN	0.999
SubsetGAN	TS (group)	$D(x)$ from PGGAN	0.999
SubsetGAN	AE (group)	$D(x)$ from StarGAN	1.
SubsetGAN	AE & TS (group)	Fake classifier (ResNet)	0.941
SubsetGAN	AE & TS (group)	Fake classifier (SqueezeNet)	0.994

Performance bajo distintas proporciones

Hasta qué porcentaje de muestras generadas somos capaces de detectar? qué pasa si solo el 20% o el 10% de la muestra es falsa?

Tipos de generación

Consideramos métodos que modifican parcialmente una imagen (cambios en un atributo en particular o estilo) y generación total.



Ejemplos de generación sintética de PGGAN & StarGAN [Karras et al., 2017, Choi et al., 2018].

Detección de clases desconocidas en redes neuronales

Los recientes avances en el aprendizaje profundo han dado lugar a mejoras en el desarrollo de la clasificación automatizada de condiciones de la piel. A medida que observamos un **creciente interés** en estos modelos en **dermatología**, es crucial abordar aspectos como **robustez** y **equidad** de estas soluciones.

En este trabajo evaluamos dos aspectos:

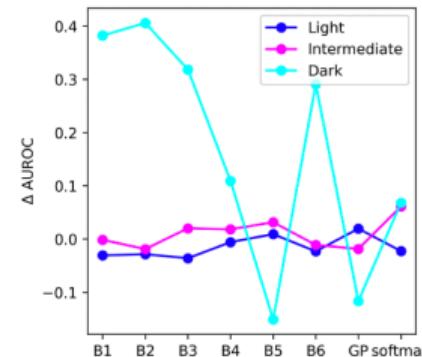
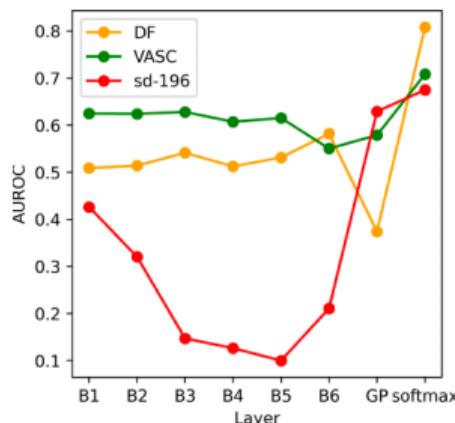
- 1 Cambios en el ambiente clínico (distinto hardware, distinto protocolo de toma de datos, etc.)
- 2 Condiciones de piel desconocidas.



Detección de clases desconocidas en redes neuronales

Resultados en ambos escenarios

Los patrones entre capas son distintos para cuando queremos detectar una clase nueva a un cambio en el protocolo de la toma de dato.



Equidad en detectores de OOD

Vemos que la performance del clasificador varía más bajo muestras de tonos de piel oscuro. Esta inestabilidad de rendimiento para muestras de tonos de piel oscuros puede deberse en parte a que la red está entrenada en el conjunto de datos ISIC 2019 que **carence en gran medida de** muestras de **tonos de piel oscuros**.

El equipo :)



Conclusiones y trabajo a futuro

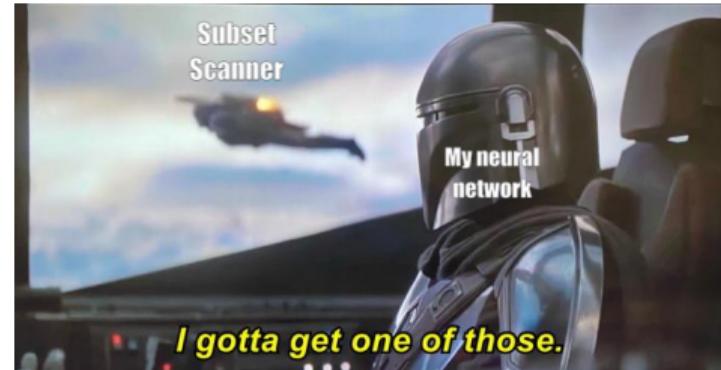
- Podemos utilizar **subset scanning** para **mejorar la detección** de estos modelos sin tener datos extra etiquetados, necesidad de reentrenar.
- Al trabajar en el espacio de activaciones, somos **agnósticos a tipos de datos y arquitecturas de redes neuronales**.
- Cómo proponer restricciones de conectividad entre capas?.



Código



Paper

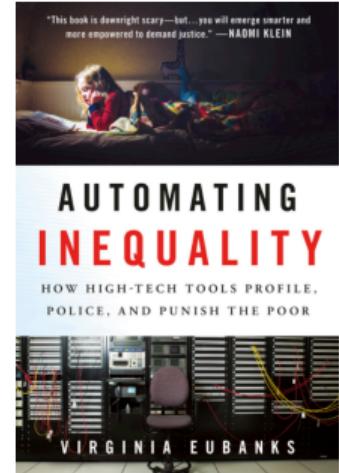
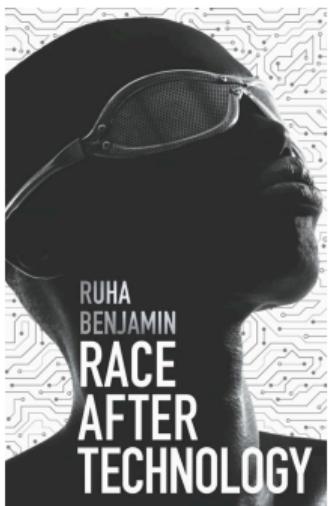


Mis 2 centavos para crear soluciones con ML :)

- 1 Es crucial que los grupos que desarrollan soluciones tecnológicas para sectores como educación, salud, etc, sean grupos interdisciplinarios.
- 2 Los investigadorXs y desarrolladorXs tienen que ser tan (o más) diversos que sus usuarios finales.
- 3 Los modelos a utilizar en producción deben explicitar en qué contexto funcionan, ser transparentes, clarificar qué sesgos fueron evaluados y cuáles son las técnicas de mitigación utilizadas.



Asante, Thanks, Gracias!



[@RTFMCellia](#) [@celitacintas](#)

References I

-  **Akinwande, V., Cintas, C., Speakman, S., and Sridharan, S. (2020).**
Identifying audio adversarial examples via anomalous pattern detection.
arXiv preprint arXiv:2002.05463.
-  **Chen, J., Jordan, M. I., and Wainwright, M. J. (2019).**
Hopskipjumpattack: A query-efficient decision-based attack.
arXiv preprint arXiv:1904.02144, 3.
-  **Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018).**
Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.
In IEEE CVPR.

References II

-  **Cintas, C., Speakman, S., Akinwande, V., Ogallo, W., Weldemariam, K., Sridharan, S., and McFowland, E. (2020).**
Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error.
IJCAI.
-  **Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015).**
Explaining and harnessing adversarial examples.
CoRR, abs/1412.6572.
-  **Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017).**
Progressive growing of gans for improved quality, stability, and variation.
arXiv preprint arXiv:1710.10196.

References III

-  **Kim, H., Tadesse, G. A., Cintas, C., Speakman, S., and Varshney, K. (2021).**
Out-of-distribution detection in dermatology using input perturbation and subset scanning.
arXiv preprint arXiv:2105.11160.
-  **Kurakin, A., Goodfellow, I. J., and Bengio, S. (2016).**
Adversarial examples in the physical world.
CoRR, abs/1607.02533.
-  **McFowland III, E., Speakman, S. D., and Neill, D. B. (2013).**
Fast generalized subset scan for anomalous pattern detection.
The Journal of Machine Learning Research, 14(1):1533–1561.
-  **Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016).**
Deepfool: a simple and accurate method to fool deep neural networks.
In Proceedings of the IEEE CVPR’16, pages 2574–2582.

References IV

-  **Neill, D. B. (2012).**
Fast subset scan for spatial pattern detection.
Journal of the Royal Statistical Society (Series B: Statistical Methodology), 74(2):337–360.
-  **Speakman, S., Somanchi, S., McFowland III, E., and Neill, D. B. (2016).**
Penalized fast subset scanning.
Journal of Computational and Graphical Statistics, 25(2):382–404.
-  **Speakman, S., Sridharan, S., Remy, S., Weldemariam, K., and McFowland, E. (2018).**
Subset scanning over neural network activations.
arXiv preprint arXiv:1810.08676.