

A tale of adversarial attacks detection & out-of-distribution stories

TrustML Series

Celia Cintas / celia.cintas@ibm.com / @RTFMCelia



Outline

- 1 The team**
- 2 Why Study Pattern Detection in DL models?**
- 3 What is Subset Scanning?**
- 4 Use-cases**
 - Adversarial Attacks Detection in Images & Audio
 - Synthetic Face Detection
 - Unknown Class Detection in Skin Disease Models
- 5 Conclusions**
- 6 Other interesting work at the Kenya Lab**
- 7 References**

The Team

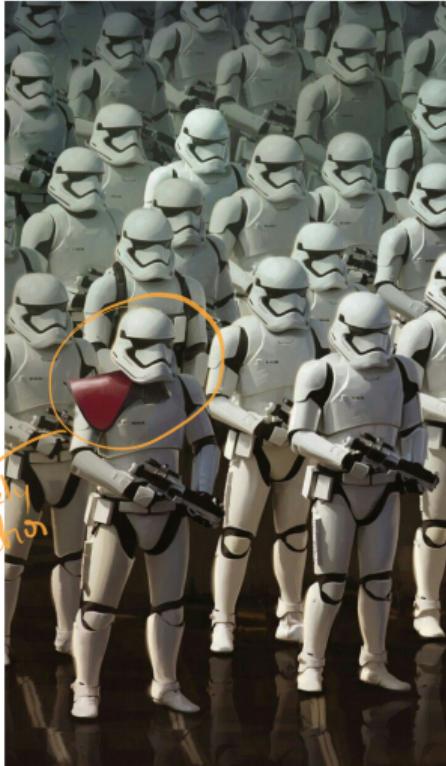


IBM Research



UNIVERSITY OF MINNESOTA

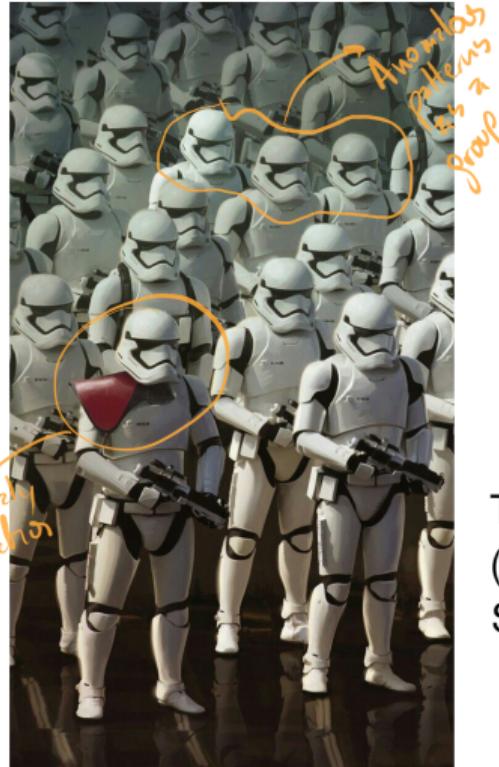
Why Study Anomalous Pattern Detection?



- Anomaly detection identifies individual samples as anomalous.
 - Looking for a **single** anomalous sample.
 - Record is obvious once identified.
 - **Apriori knowledge** of what to look for.

IBM Research | Africa

Why Study Anomalous Pattern Detection?



- Anomaly detection identifies individual samples as anomalous.
 - Looking for a **single** anomalous sample.
 - Record is obvious once identified.
 - **Apriori knowledge** of what to look for.
- We expand those ideas to identify groups of samples as anomalous.
 - **Multiple samples** are affected.
 - Individual samples only **slightly anomalous**.
 - **Little knowledge** of what pattern may be like.

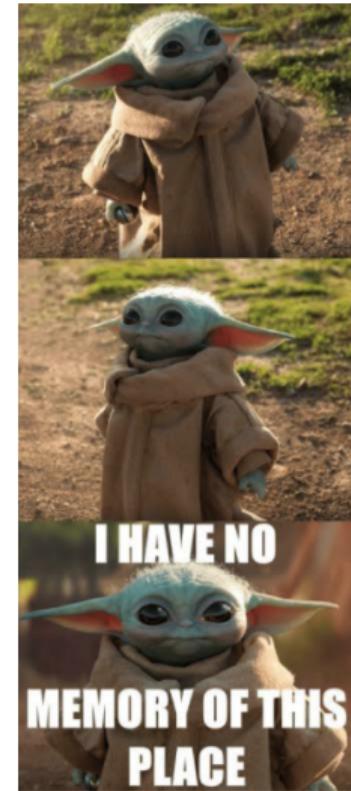
This work has its foundations on Linear-Time Subset Scanning (LTSS) [Neill, 2012, McFowlan III et al., 2013, Speakman et al., 2013, Speakman et al., 2016]

Why Study Anomalous Pattern Detection in Neural Networks?

Reliably detecting OOD samples or perturbed samples in a given set of inputs is of high practical relevance due to the **vulnerability** of neural networks to adversarial examples and data shifts.

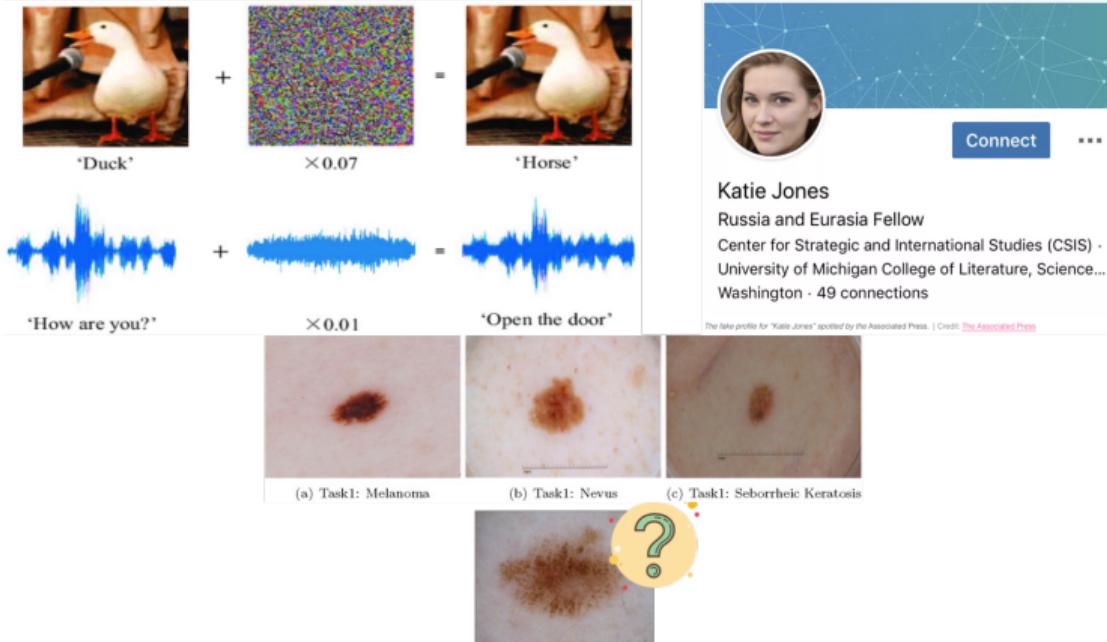
These altered inputs create a **security risk** in applications with **real-world consequences**, such as healthcare, self-driving cars, robotics and financial services.

Basically, for anything outside training/validation data .. we need to **evaluate/aid** DL models to double-check the input before performing inferences/predictions.



IBM Research | Africa

Why Study Anomalous Pattern Detection in Neural Networks? (Cont.)



We show some use-cases to detect and characterize:

- Adversarial Attacks.
- Synthesized samples.
- Unknown type samples.

Subset Scanning for Anomalous Pattern Detection

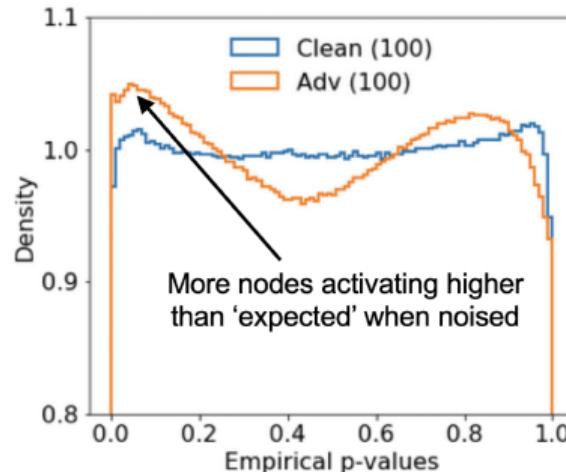
-  Treat Neural Networks as data-generating systems and apply anomalous pattern detection methods to activation data.
-  Subset Scanning efficiently searches over a large combinatorial space in order to find groups of records that differ the most from ‘expected’ behavior.

Some goodies about this type of approach:

- 1 We can provide detection improvements at **at run time**.
- 2 We can **abstract from domains** and focuses only on the deep representation of the input.
- 3 **No** need to **re-train** or have **labeled examples** of the anomalies.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)



Assumption

Activations from adversarial images have a different distribution of p-values than benign/clean samples.

p-value is the proportion of background activations (H_0), drawn from the same node for several clean samples, greater than the activation from a test sample.

Subset Scanning for Anomalous Pattern Detection (Cont.)

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

Where N_{α} is the number of p-values less than α

N is the number of p-values
 α is the level of significance

How we score new images?

Scoring functions operate on an evaluation image in order to measure how much the p-values deviate from uniform.

Subset Scanning for Anomalous Pattern Detection (Cont.)

NPSS maximization

Scoring functions may be viewed as set functions that operate on subsets of nodes. We search for the highest scoring subset of nodes that maximize the deviance from uniform.

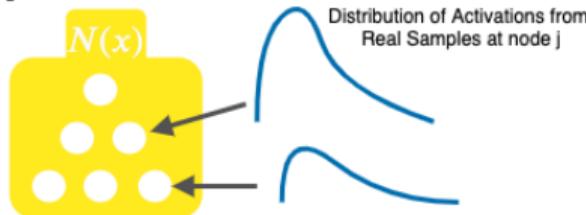
$$F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \varphi(\alpha, N_{\alpha}(S), N(S)) \quad (2)$$

group vs individual

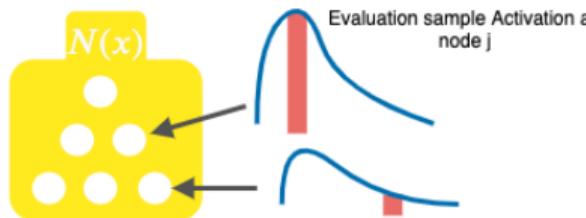
For group-based scanning our search space $S = X_S \times O_S$ where X_S is a subset of images and O_S is a subset of nodes' activations. For individual scanning we work with only one X_i .

Subset Scanning for Anomalous Pattern Detection (Cont.)

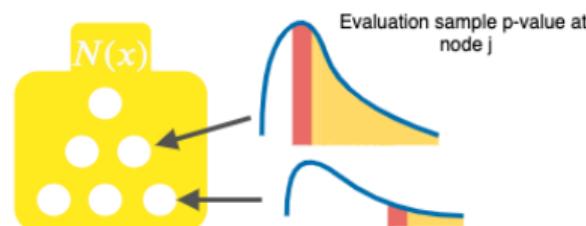
Extract activations from $N(x)$ for H_0 and evaluation



Compute empirical p-value ranges



Maximization of NPSS across nodes and images

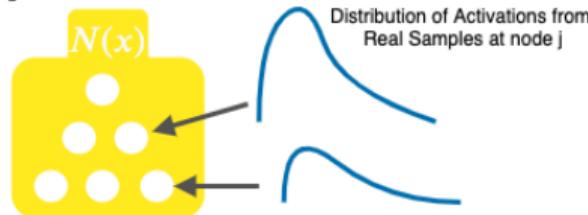


Extract groups of anomalous nodes, samples and evaluation metrics ($\mathcal{O}, \mathcal{N}, P, R, AUC$)

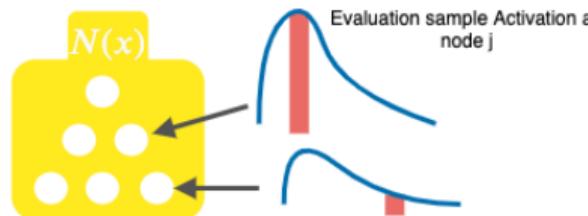
We can use already trained network or training our own models. We experimented over CNNs, ResNets, DenseNets, RNNs, VAEs, AEs and GANs.

Subset Scanning for Anomalous Pattern Detection (Cont.)

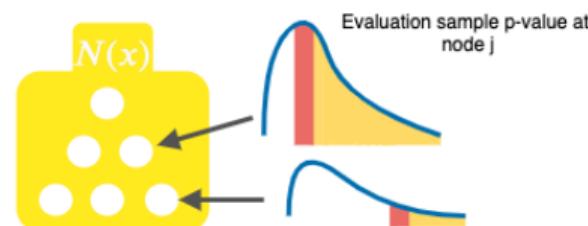
Extract activations from $N(x)$ for H_0 and evaluation



Compute empirical p-value ranges



Maximization of NPSS across nodes and images

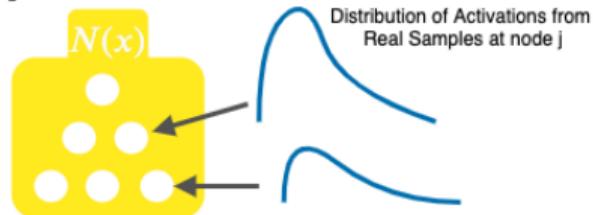


Extract groups of anomalous nodes, samples and evaluation metrics ($\mathcal{O}, \mathcal{N}, P, R, AUC$)

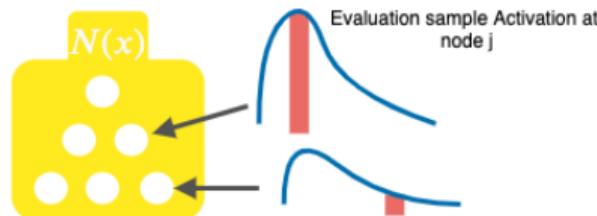
We extract the activations for a given layer of the model for all background and test samples.

Subset Scanning for Anomalous Pattern Detection (Cont.)

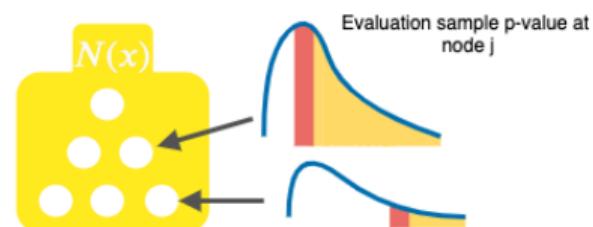
Extract activations from $N(x)$ for H_0 and evaluation



Compute empirical p-value ranges



Maximization of NPSS across nodes and images

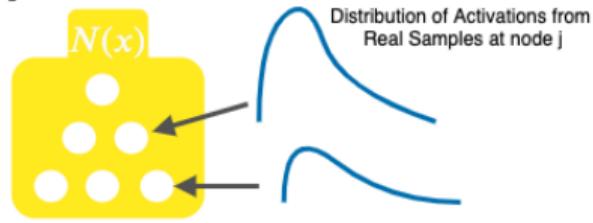


Extract groups of anomalous nodes, samples and evaluation metrics
($\mathcal{O}, \mathcal{N}, P, R, AUC$)

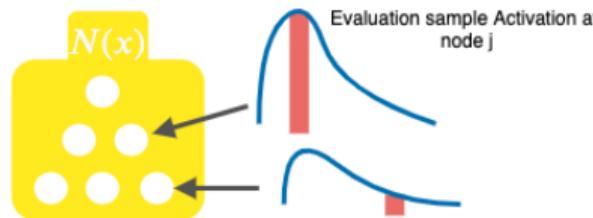
We compute the empirical p-values and filter for a given α threshold.

Subset Scanning for Anomalous Pattern Detection (Cont.)

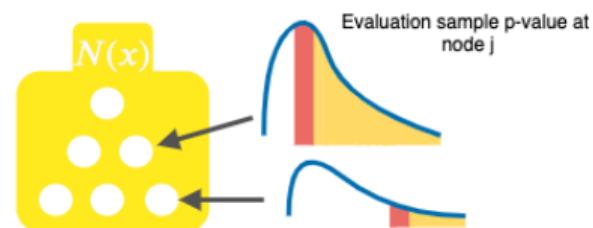
Extract activations from $N(x)$ for H_0 and evaluation



Compute empirical p-value ranges



Maximization of NPSS across nodes and images

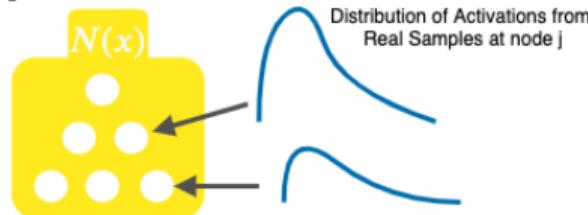


Extract groups of anomalous nodes, samples and evaluation metrics
($\mathcal{O}, \mathcal{N}, P, R, AUC$)

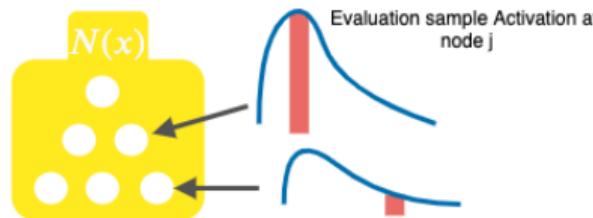
NPSS scores multiple subsets of p-values. We identify the most anomalous subset for the evaluation samples. The evaluation samples can be clean images, perturbed samples, synthesized or new class samples.

Subset Scanning for Anomalous Pattern Detection (Cont.)

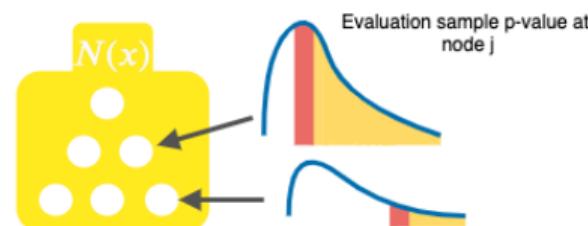
Extract activations from $N(x)$ for H_0 and evaluation



Compute empirical p-value ranges



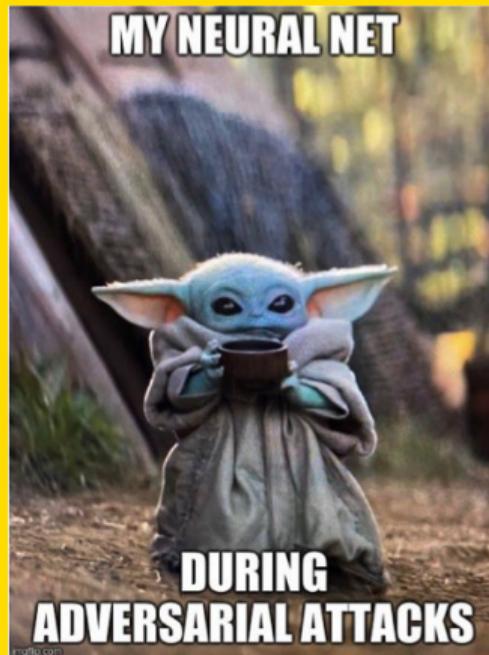
Maximization of NPSS across nodes and images



Extract groups of anomalous nodes, samples and evaluation metrics
 $(\mathcal{O}, \mathcal{N}, P, R, AUC)$

We extract the subset scores and the nodes that contributed to that score.

Adversarial Attacks Detection in Images & Audio



IBM Research | Africa

What is an Adversarial Attack?

White-box an attacker has complete access to the model, including its structure and trained weights. E.g. 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 an attacker can only access the outputs of the target model. (e.g HopSkipJumpAttack [Chen et al., 2019]).

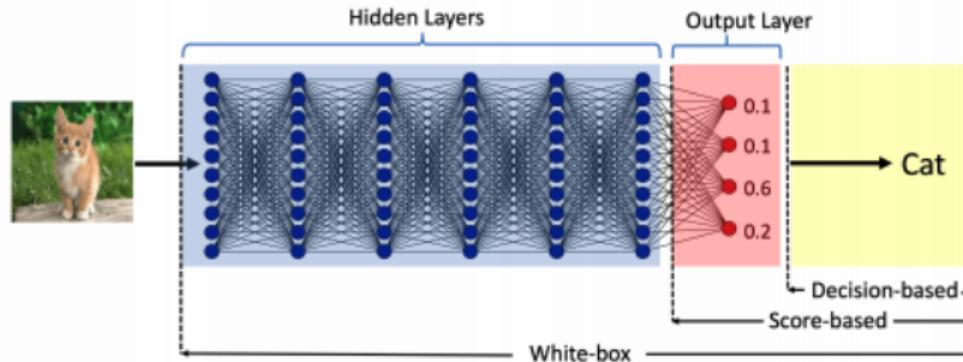


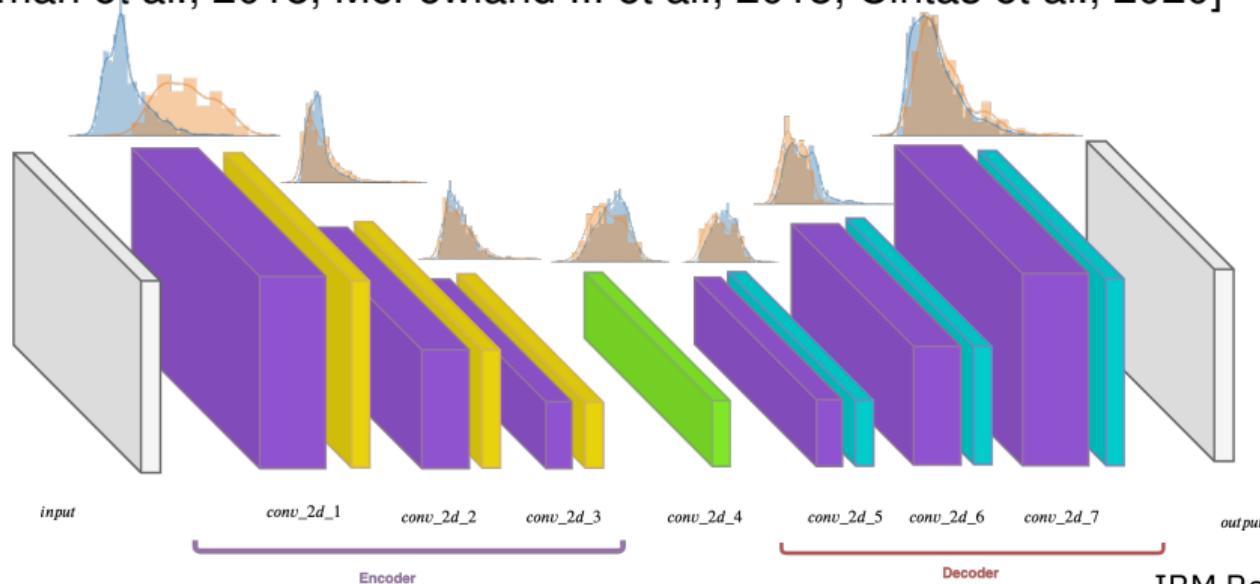
Figure from [Chen et al., 2019]

IBM Research | Africa

Subset Scan in AE for Adversarial attacks

We propose an unsupervised method for detecting adversarial attacks in inner layers of autoencoder (AE) networks by maximizing a non-parametric measure of anomalous node activations.

[Speakman et al., 2018, McFowlan III et al., 2013, Cintas et al., 2020]



IBM Research | Africa

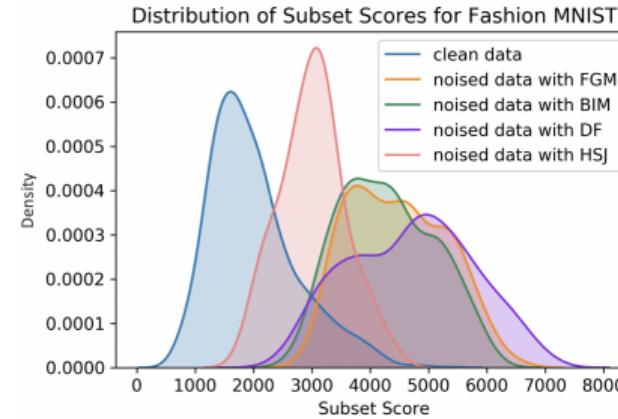
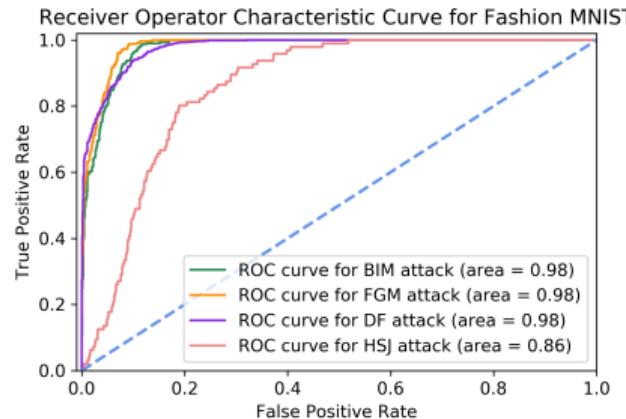
Results in Inner Layers

Subset scanning across layers performance

In the latent space, the autoencoder abstracts basic representations of the images, losing subset scanning power due to the autoencoder mapping the new sample to the expected distribution.

Layers	Clean Training						Noised (1%)			Noised (9%)	
	F-MNIST			MNIST			F-MNIST	MNIST	BIM	F-MNIST	MNIST
	BIM	FGSM	DF	HSJ	BIM	FGSM	DF	HSJ	BIM	BIM	BIM
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	

Results in Inner Layers (Cont.)



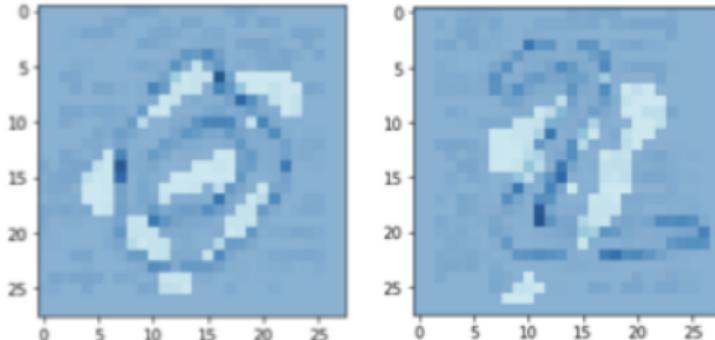
ROC curves & Distribution of subset scores

For each of the noised cases as compared to the scores from test sets containing all natural images for layer *Conv2d_1*. Distribution of subset scores for test sets of images over *Conv2d_1*. Clean images had lower scores than noised images.

Results over the Reconstruction Error

The results over the RE depend on the AE performance. If an autoencoder's loss is high, it is more difficult to separate between clean and noised samples in the reconstruction space because the most anomalous subset of reconstructed pixels of a clean image may be higher due to chance.

Datasets	Attacks	Detection Power (AUROC)		
		Ours RE	Mean RE	One-SVM
F-MNIST	BIM	0.698	0.641	0.478
	FGSM	0.672	0.630	0.497
	DF	0.599	0.477	0.534
	HSJ	0.956	0.935	0.546
MNIST	BIM	0.998	0.751	0.624
	FGSM	0.983	0.725	0.624
	DF	0.992	0.574	0.637
	HSJ	0.999	0.619	0.537



Explainability

Subset Scanning over the reconstruction error space is an interesting technique to inspect **which pixels** of the reconstructed image belong to the **most anomalous subset**.

Results over RNNs & Audio

Defending against audio adversarial attacks has **predominantly** focused on **preprocessing techniques** such as mp3 compression, quantization, adding noise, or smoothing. However, these approaches **modify the input** in some way and affect performance on benign samples. We perform comparable to the current state-of-the-art without losing accuracy of benign samples [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

IBM Research | Africa

Synthetic Face Detection

With **near realistic** generated content and **high throughput capacity**, several high-profile concerns are on the rise in critical areas such as security, ethics and democracy. This will **challenge data-driven decision making** in societal and commercial activities.

Enhancing Fake Face Detection

We enhance the performance of the discriminant component of GANs and off-the-shelf fake classifiers to detect the synthesized images.

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



Sample generation from PGGAN & StarGAN [Karras et al., 2017, Choi et al., 2018].

IBM Research | Africa

Enhancing Fake Face Detection (Cont.)

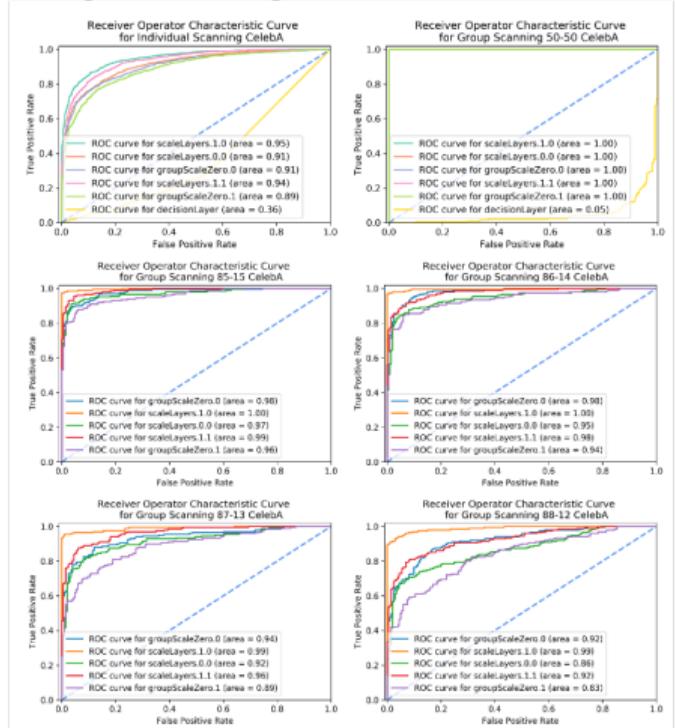
Types of synthetic modifications

We evaluate our method across partial modifications (attribute editing, style transfer) and total generation for generated samples from noise.

Performance across proportions

Even under small proportions we can see and increase in the performance detection capabilities compared to individual scanning.

Cintas et al (Under review)



Unknown Class Detection in Skin Disease Models

Recent advances in deep learning have led to breakthroughs in the development of automated skin disease classification. As we observe a **widespread adoption** of these models in the **dermatology space**, it is crucial to address aspects such as the **robustness** and fairness of these solutions.

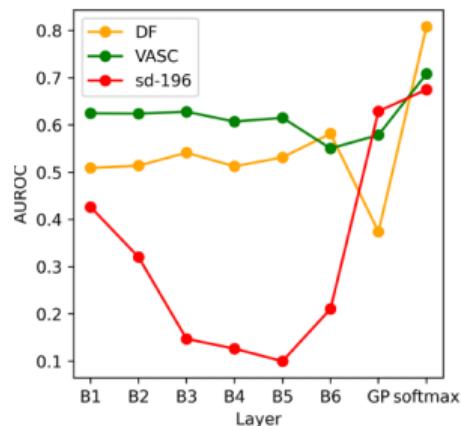
We validated our approach in two use cases:

- 1 Different clinical settings.
- 2 Unknown disease classes.

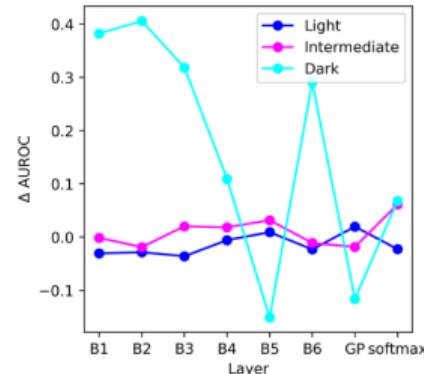
OOD for Skin Disease Classifiers

Results across settings

The layers for detecting new class are different from the ones for OOD



Kim et al. (under review)

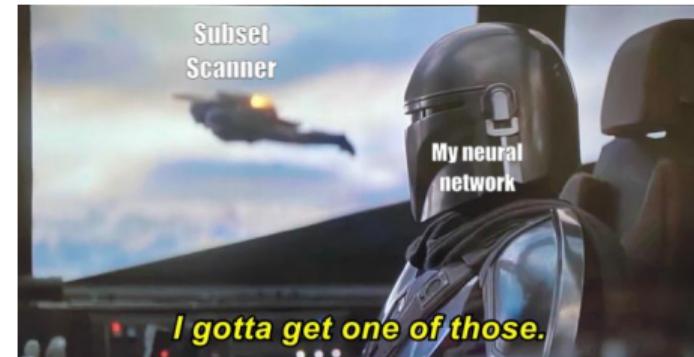


Fairness of OOD detectors

Current skin disease datasets **heavily lack** representation of dark skin-tones, as methods perform better in detecting these samples of **dark skin tone** as OOD samples compared to those of light or intermediate skin tones.

Conclusions & Future Work

- We use **subset scanning** methods from the anomalous pattern detection domain to **enhance detection** power without labeled examples of the noise, re-training or data augmentation methods.
- Studying the activation space makes us **agnostic** to **data types** and **DL architectures**.
- Explore detected subsets of nodes **source detection** (different type of generative process or types of adversarial attacks).
- How to intelligently **select layers** to scan depending use-cases.



Adversarial
Robustness
Toolbox



Code



Paper



Internship

IBM Research | Africa

Other interesting work at the Kenya Lab

1 Fairness

- Kinyanjui, et al. **Fairness of Classifiers Across Skin Tones in Dermatology.** In MICCAI 2020 - International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 320-329)
- Tadesse et al. **Automated Evaluation of Representation in Dermatology Educational Materials.** AAAI Workshop on Trustworthy AI for Healthcare, February 2021

2 ML in Healthcare

- Idrees, et al. **Successes and Misses of Global Health Development: Detecting Temporal Concept Drift of Under-5 Mortality Prediction Models with Bias Scan.** AMIA 2021 Virtual Informatics Summit.
- Tadesse et al. **Unsupervised Discovery of Subgroups with Anomalous Maternal and Neonatal Outcomes with WHO's Safe Childbirth Checklist as Intervention.** NeurIPS Workshop on Machine Learning for Public Health (Best Paper Award), December 2020.

Asante, Thanks, Gracias!



celia.cintas@ibm.com

@RTFMCelia

A tale of adversarial attacks detection & out-of-distribution stories

IBM Research | Africa

TrustML Series

30 / 34

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

-  **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.
-  **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.

References IV

-  **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.
-  **Speakman, S., Zhang, Y., and Neill, D. B. (2013).**
Dynamic pattern detection with temporal consistency and connectivity constraints.
In *2013 IEEE 13th International Conference on Data Mining*, pages 697–706. IEEE.