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

LXAI @ ICML 2021

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

IBM Research | Africa

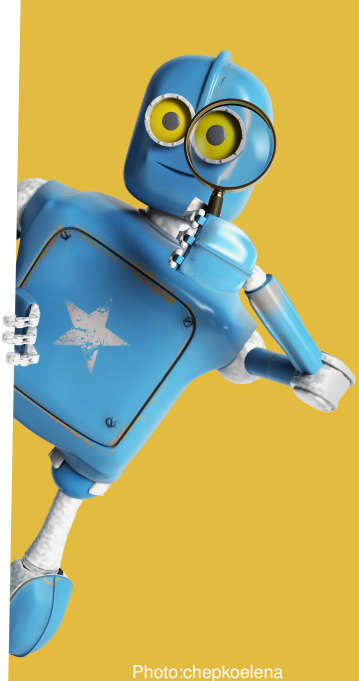


Photo:chepkoelena

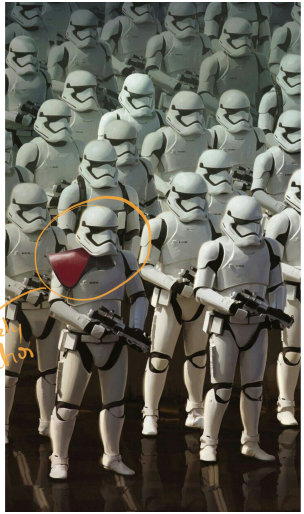
The Team

IBM Research

HARVARD UNIVERSITY

IBM Research | Africa

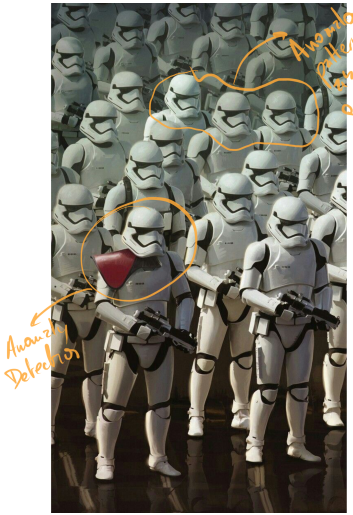
Anomalous Pattern Detection: Why?



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

IBM Research | Africa

Anomalous Pattern Detection: Why?



- Anomaly detection identifies individual samples as anomalous.
  - Looking for **single** anomalous sample.
  - The sample 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, McFowland III et al., 2013, Speakman et al., 2016]

IBM Research | Africa

Anomalous Pattern Detection in Neural Networks

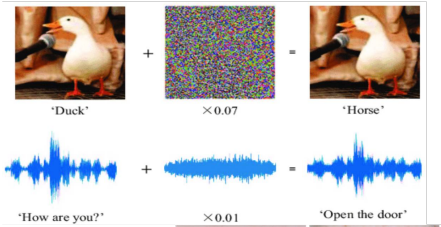
Reliably detecting Out-of-Distribution (OOD) or perturbed samples in a given set of inputs is of high practical relevance due to the **vulnerability** of neural networks to adversarial examples and data shifts.

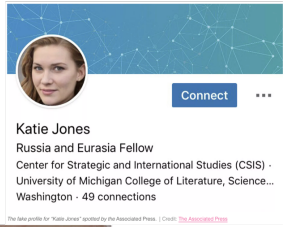
These altered inputs create a **security risk** in applications with **real-world consequences**, such as healthcare, self-driving cars, robotics and financial services.

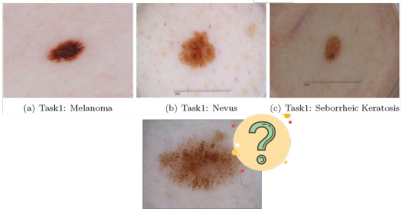
Basically, for anything outside training/validation data, we need to **evaluate/aid** DL models to double-check the input before performing inferences/predictions.

IBM Research | Africa

Anomalous Pattern Detection in Neural Networks (Cont.)







We show some use-cases to detect and characterize:

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

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection

💡 Treat Neural Networks as data-generating systems and apply anomalous pattern detection methods to activation data.

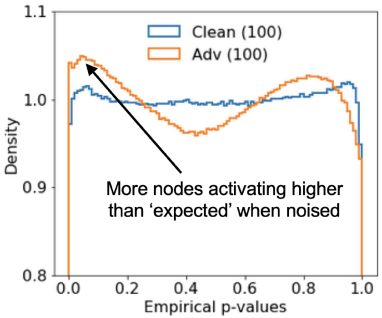
🔍 Subset Scanning efficiently searches over a large combinatorial space in order to find groups of records that differ the most from ‘expected’ behavior.

Some goodies about this type of approach:

1. We can provide detection improvements **at run time**.
2. We can **abstract from domains** and focus 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.)



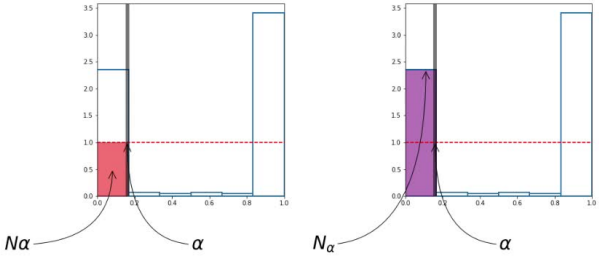
Assumption

Activations from abnormal 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.

IBM Research | Africa

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
φ is a scoring function

How we score a test sample?

Scoring functions operate on a test sample in order to measure how much the p-values deviate from uniform.

IBM Research | Africa

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.

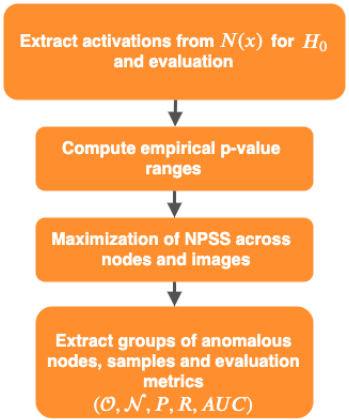
$$F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \varphi(\alpha, N_{\alpha}(S), N(S)) \tag{2}$$

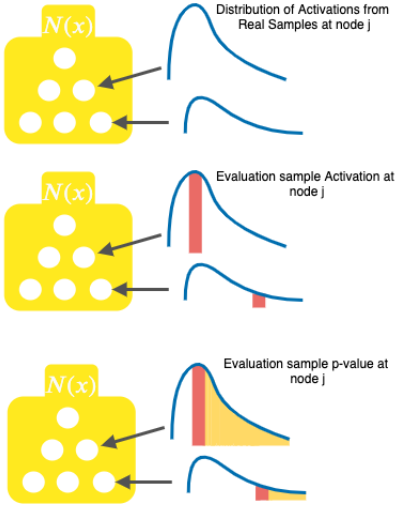
Group vs. Individual Scanning

For group-based scanning our search space is: $S = X_{\hat{S}} \times O_{\hat{S}}$, where $X_{\hat{S}}$ is a subset of test samples and $O_{\hat{S}}$ is a subset of nodes' activations.
For individual scanning we work with only one X_i.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

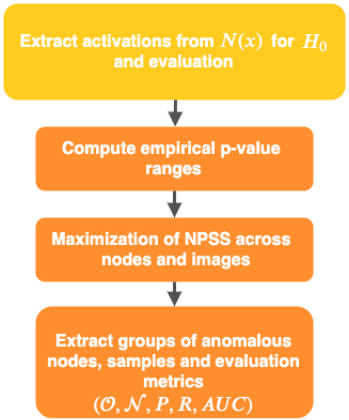


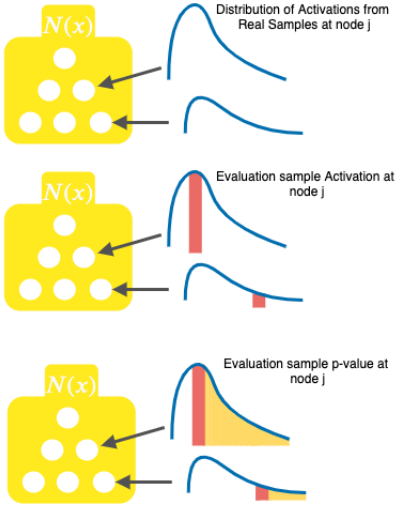


We can use already trained network or training our own models. We experimented over CNNs, ResNets, DenseNets, RNNs, VAEs, AEs and GANs.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

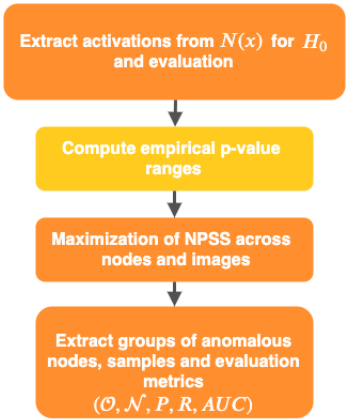


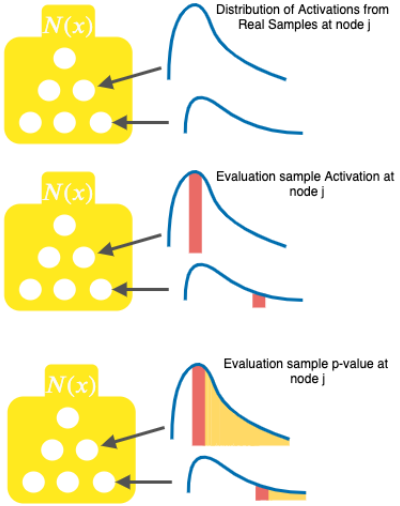


We extract the activations from a given layer of the model for all background and test samples.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

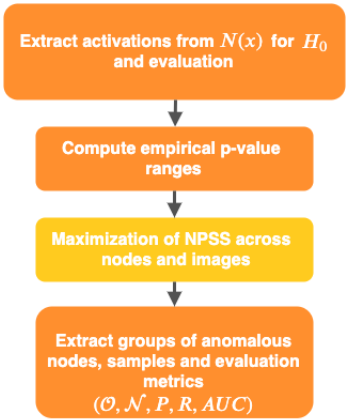


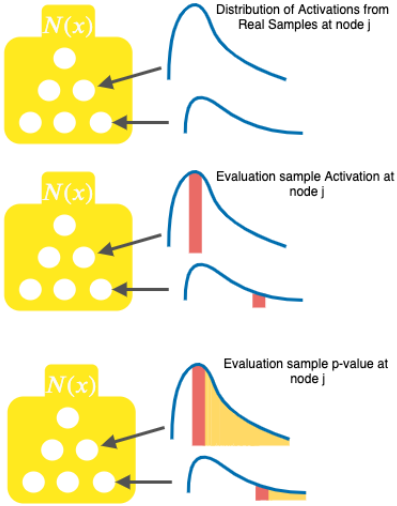


We compute the empirical p-values and filter for a given α threshold.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

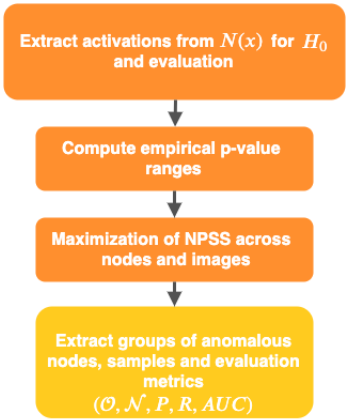


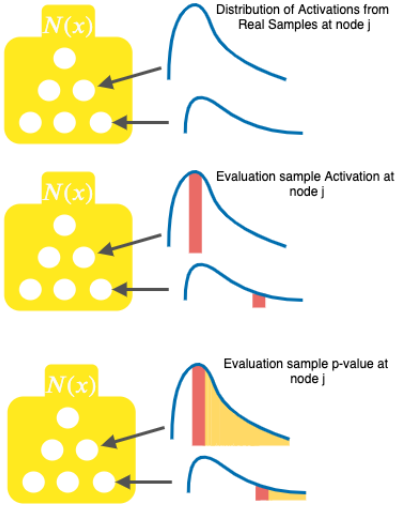


NPSS scores multiple subsets of p-values. We identify the most anomalous subset for the test samples, which can be clean, perturbed, synthesized or new class samples.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

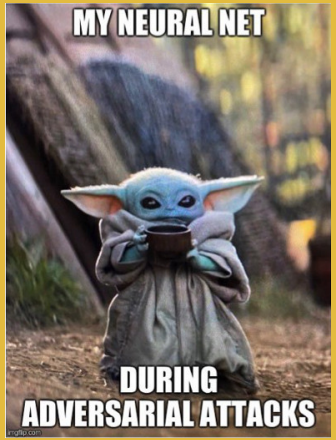




We extract the subset scores and the subset nodes that contributed to that score.

IBM Research | Africa

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].

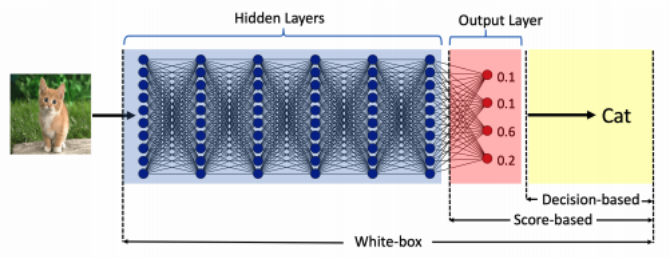
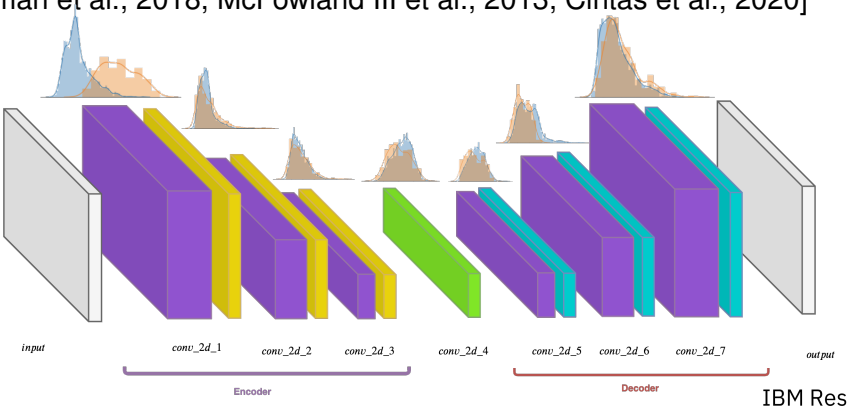


Figure from [Chen et al., 2019]

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, McFowland III et al., 2013, Cintas et al., 2020]



IBM Research | Africa

Results in Inner AE Layers

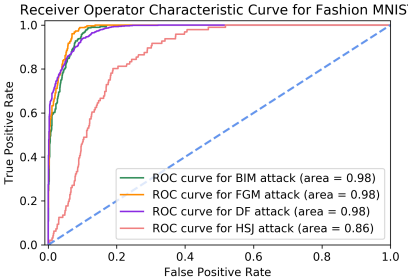
Subset scanning across layers performance

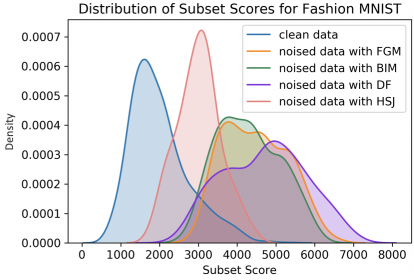
In the latent space, the autoencoder abstracts basic representations of the samples, losing subset scanning power due to the autoencoder mapping the new sample to the expected distribution.

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

IBM Research | Africa

Results in Inner Layers (Cont.)





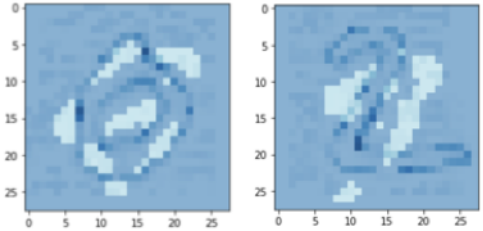
ROC curves & distribution of subset scores

Scores of noisy samples compared to the scores from natural samples for layer *Conv2d_1*. Distribution of subset scores for test sets of images over *Conv2d_1*. Clean images had lower scores than noised images.

IBM Research | Africa

Results over the Reconstruction Error (RE)

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.



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 RE space helps to inspect **which pixels** of the reconstructed image belong to the **most anomalous subset**.

IBM Research | Africa

Results over RNNs & Audio

Defending against audio adversarial attacks has **predominantly** focused on **preprocessing techniques** such as mp3 compression, quantization, 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	80, 2048	0.936	91.5	0.283
(800, 200, 90)	80, 2048			0.158
	80, 4096			**0.973**
	80, 2048			0.903
DS, LS, CW	64, 2048	0.930	NA	0.568
(800, 200, 90)	64, 2048			0.038
	64, 4096			**0.982**
	64, 2048			0.527
LV, LS, IA	179, 40, 32	NA	NA	**0.755**
(300, 100, 100)	212, 20, 32			0.491
	423, 40, 32			0.571

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.

IBM Research | Africa

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 Type	Network	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 (indv)	TS	$D(x)$ from PGGAN	0.950
SubsetGAN (indv)	AE	$D(x)$ from StarGAN	0.999
SubsetGAN (group)	TS	$D(x)$ from PGGAN	0.999
SubsetGAN (group)	AE	$D(x)$ from StarGAN	1.
SubsetGAN (group)	AE & TS	Fake classifier (ResNet)	0.941
SubsetGAN (group)	AE & TS	Fake classifier (SqueezeNet)	0.994

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

IBM Research | Africa

Enhancing Fake Face Detection (Cont.)

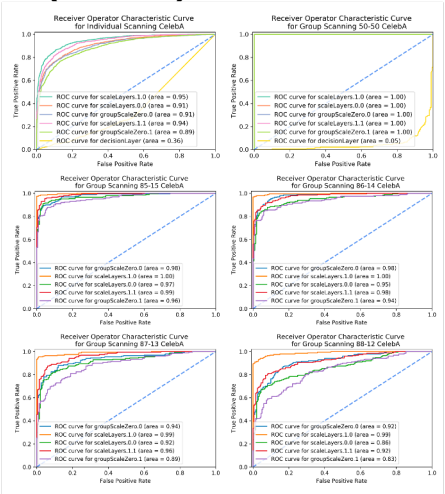
Types of synthetic modifications

We evaluate or method across partial modifications (attribute editting, style transfer) and completely generated samples from noise.

Performance across proportions

Even under small proportions, it is possible to increase the detection performance compared to individual scanning.

[Cintas et al., 2021] (Under review)



IBM Research | Africa

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 an **increasing interest** in 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:

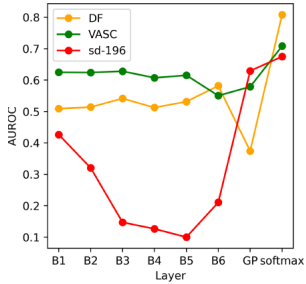
1. Different clinical settings (e.g., acquisition devices).
2. Unknown disease classes (unseen during training).

IBM Research | Africa

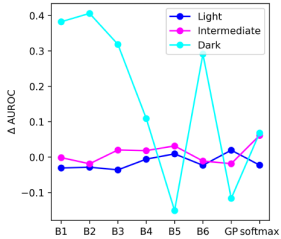
OOD for Skin Disease Classifiers

Results across settings

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



[Kim et al., 2021] (under review)

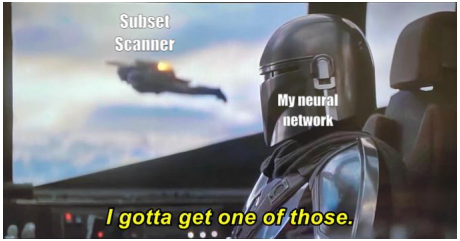


Fairness of OOD detectors

We see varying performances for samples of Dark skin tones. This instability of performance for samples of Dark skin tones may be partially because network is trained on the ISIC 2019 dataset that **heavily lacks** samples of **Dark 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).
- **Layer selection**: open and depends on use-cases.



Adversarial Robustness Toolbox

Code

Paper

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.

IBM Research | Africa

Asante, Thanks, Gracias!

celia.cintas@ibm.com @RTFMCelia

IBM Research | Africa

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*.

IBM Research | Africa

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.
- **Cintas, C., Speakman, S., Tadesse, G. A., Akinwande, V., McFowland III, E., and Weldemariam, K. (2021).**
  Pattern detection in the activation space for identifying synthesized content.
  arXiv preprint arXiv:2105.12479.
- **Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015).**
  Explaining and harnessing adversarial examples.
  CoRR, abs/1412.6572.

IBM Research | Africa

References III

- **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.
- **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.

IBM Research | Africa

References IV

- **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.

- **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.

IBM Research | Africa

References V

Speakman, S., Sridharan, S., Remy, S., Weldemariam, K., and McFowland, E. (2018).
Subset scanning over neural network activations.
arXiv preprint arXiv:1810.08676.

IBM Research | Africa