Detecting Adversarial Attacks via Subset Scanning

of Autoencoder Activations and Reconstruction Error

Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan, Edward McFowland III

IBM Research | Africa

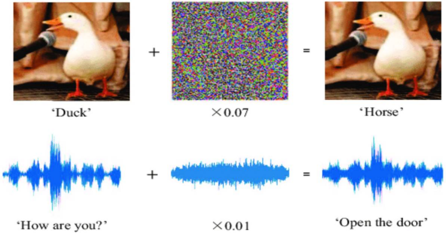
CARLSON SCHOOL OF MANAGEMENT

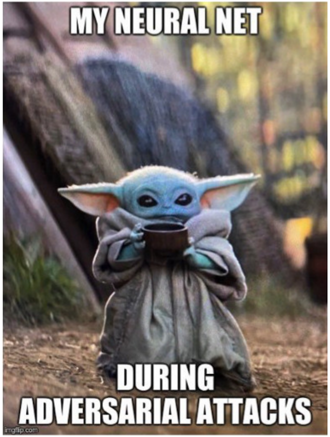
UNIVERSITY OF MINNESOTA

Photo: HYACINTH EMPINADO/STAT

Why is important to detect adversarial attacks?

Reliably detecting attacks in a given set of inputs is of high practical relevance due to the vulnerability of neural networks to adversarial examples. These altered inputs create a security risk in applications with real-world consequences, such as self-driving cars, robotics and financial services.





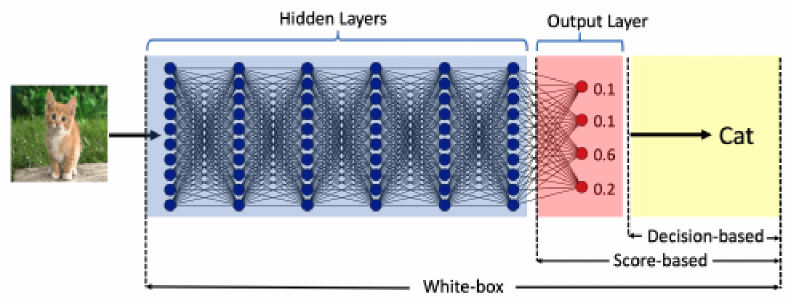
IBM Research | Africa

Picture: https://whataftercollege.com/machine-learning/adversarial-attack-machine-learning/

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) [KGB16], Fast Gradient Signal Method (FGSM) [GSS15], DeepFool (DF) [MDFF16].

black-box: an attacker can only access the outputs of the target model. (e.g HopSkipJumpAttack [CJW19]).

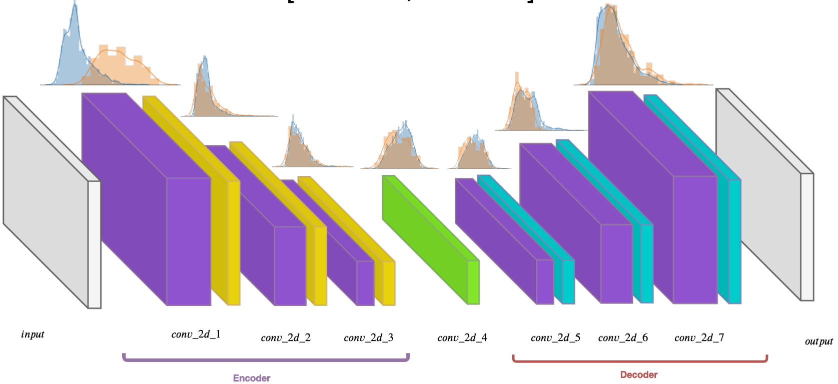


IBM Research | Africa

Figure from [CJW19]

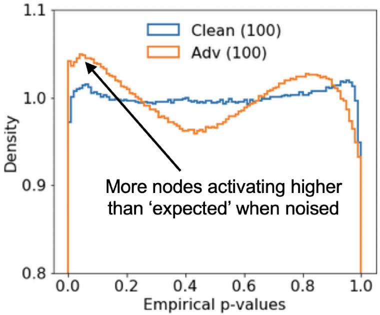
Subset Scan for Anomalous Pattern Detection

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. [SSR+18, MISN13]



IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)



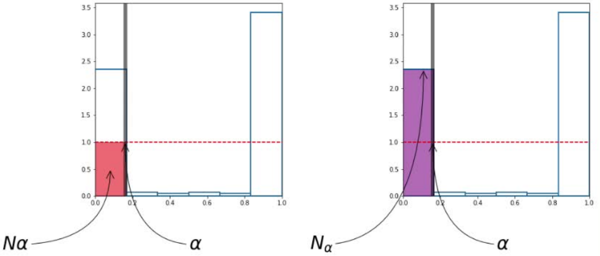
Assumption

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.

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)



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

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.

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.

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

Why we use non-parametric scoring functions?

To make **minimal assumptions** on the underlying distribution of node activations and enables us to scan across **different types of layers**.

Subset Scanning for Anomalous Pattern Detection (Cont.)

input : Background set of images: $X_z \in D^{H_0}$,
Evaluation Image: X_i, training dataset, $\alpha_{\max}$.
output: S^*_F Score for X_i

$AE \leftarrow$ `TrainNetwork` (training dataset);
$AE_y \leftarrow$ Some flattened layer of AE;
for $z \leftarrow 0$ **to** M **do**
  for $j \leftarrow 0$ **to** J **do**
    $A^{H_0}_{zj} \leftarrow$ `ExtractActivation` (AE_y, X_z)
for $j \leftarrow 0$ **to** J **do**
  $A_{ij} \leftarrow$ `ExtractActivation` (AE_y, X_i)
$p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
$p^*_{ij} = \{y < \alpha_{\max} \,\forall\, y \subseteq p_{ij}\}$;
$p^s_{ij} \leftarrow$ `SortAscending` (p^*_{ij});
for $k \leftarrow 1$ **to** J **do**
  $S_{(k)} = \{p_y \subseteq p^s_{ij} \forall y \in \{1, \ldots, k\}\}$;
  $\alpha_k = max(S_{(k)})$;
  $F(S_{(k)}) \leftarrow$ `NPSS` $(\alpha_k$, k, k);
$k^* \leftarrow \arg\max F(S_{(k)})$;
$\alpha^* = \alpha_{k^*}$;
$S^* = S_{(k^*)}$;
return S^*, α^*, and $F(S^*)$

We can use already trained AE or training our own models. We used different AE for each dataset (FMNIST, MNIST, CIFAR) and two other AE trained with different noised proportions (1% and 9%).

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

input : Background set of images: $X_z \in D^{H_0}$,
Evaluation Image: X_i, training dataset, $\alpha_{\max}$.
output: S_E^* `Score for` X_i

$AE \leftarrow$ `TrainNetwork` (training dataset);
$AE_y \leftarrow$ Some flattened layer of AE;
for $z \leftarrow 0$ **to** M **do**
  for $j \leftarrow 0$ **to** J **do**
    $A_{zj}^{H_0} \leftarrow$ `ExtractActivation` (AE_y, X_z)
for $j \leftarrow 0$ **to** J **do**
  $A_{ij} \leftarrow$ `ExtractActivation` (AE_y, X_i)
$p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
$p_{ij}^* = \{y < \alpha_{\max} \; \forall \, y \subseteq p_{ij}\}$;
$p_{ij}^s \leftarrow$ `SortAscending` (p_{ij}^*);
for $k \leftarrow 1$ **to** J **do**
  $S_{(k)} = \{p_y \subseteq p_{ij}^s \forall y \in \{1, \ldots, k\}\}$;
  $\alpha_k = max(S_{(k)})$;
  $F(S_{(k)}) \leftarrow$ `NPSS` $(\alpha_k$, k, k);
$k^* \leftarrow \arg\max F(S_{(k)})$;
$\alpha^* = \alpha_{k^*}$;
$S^* = S_{(k^*)}$;
return S^*, α^*, and $F(S^*)$

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

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

input : Background set of images: $X_z \in D^{H_0}$,
Evaluation Image: X_i, training dataset, $\alpha_{\max}$.
output: S^*_E `Score for` X_i

$AE \leftarrow$ `TrainNetwork` (training dataset);
$AE_y \leftarrow$ Some flattened layer of AE;
for $z \leftarrow 0$ **to** M **do**
  for $j \leftarrow 0$ **to** J **do**
    $A^{H_0}_{zj} \leftarrow$ `ExtractActivation` (AE_y, X_z)
for $j \leftarrow 0$ **to** J **do**
  $A_{ij} \leftarrow$ `ExtractActivation` (AE_y, X_i)
$p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
$p^*_{ij} = \{y < \alpha_{\max} \; \forall \; y \subseteq p_{ij}\}$;
$p^s_{ij} \leftarrow$ `SortAscending` (p^*_{ij});
for $k \leftarrow 1$ **to** J **do**
  $S_{(k)} = \{p_y \subseteq p^s_{ij} \forall y \in \{1, \ldots, k\}\}$;
  $\alpha_k = max(S_{(k)})$;
  $F(S_{(k)}) \leftarrow$ NPSS $(\alpha_k$, k, k);
$k^* \leftarrow \arg\max F(S_{(k)})$;
$\alpha^* = \alpha_{k^*}$;
$S^* = S_{(k^*)}$;
return S^*, α^*, and $F(S^*)$

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

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

input : Background set of images: $X_z \in D^{H_0}$,
Evaluation Image: X_i, training dataset, $\alpha_{\max}$.
output: S_E^* `Score for` X_i

$AE \leftarrow$ `TrainNetwork` (training dataset);
$AE_y \leftarrow$ Some flattened layer of AE;
for $z \leftarrow 0$ **to** M **do**
  for $j \leftarrow 0$ **to** J **do**
    $A_{zj}^{H_0} \leftarrow$ `ExtractActivation` (AE_y, X_z)
for $j \leftarrow 0$ **to** J **do**
  $A_{ij} \leftarrow$ `ExtractActivation` (AE_y, X_i)
$p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
$p_{ij}^* = \{y < \alpha_{\max} \; \forall \; y \subseteq p_{ij}\}$;
$p_{ij}^s \leftarrow$ `SortAscending` (p_{ij}^*);
for $k \leftarrow 1$ **to** J **do**
  $S_{(k)} = \{p_y \subseteq p_{ij}^s \forall y \in \{1, \ldots, k\}\}$;
  $\alpha_k = max(S_{(k)})$;
  $F(S_{(k)}) \leftarrow$ NPSS $(\alpha_k$, k, k$)$;
$k^* \leftarrow \arg\max F(S_{(k)})$;
$\alpha^* = \alpha_{k^*}$;
$S^* = S_{(k^*)}$;
return S^*, α^*, and $F(S^*)$

NPSS scores multiple subsets of p-values with the Berk-Jones test statistic [BJ79]:

$$\varphi_{BJ}(\alpha, N_\alpha, N) = N * KL\left(\frac{N_\alpha}{N}, \alpha\right) \qquad (3)$$

IBM Research | Africa

Subset Scanning for Anomalous Pattern Detection (Cont.)

input : Background set of images: $X_z \in D^{H_0}$,
Evaluation Image: X_i, training dataset, $\alpha_{\max}$.
output: S_E^* `Score for` X_i

$AE \leftarrow$ `TrainNetwork` (training dataset);
$AE_y \leftarrow$ Some flattened layer of AE;
for $z \leftarrow 0$ **to** M **do**
  for $j \leftarrow 0$ **to** J **do**
    $A_{zj}^{H_0} \leftarrow$ `ExtractActivation` (AE_y, X_z)
for $j \leftarrow 0$ **to** J **do**
  $A_{ij} \leftarrow$ `ExtractActivation` (AE_y, X_i)
$p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
$p_{ij}^* = \{y < \alpha_{\max} \, \forall \, y \subseteq p_{ij}\}$;
$p_{ij}^s \leftarrow$ `SortAscending` (p_{ij}^*);
for $k \leftarrow 1$ **to** J **do**
  $S_{(k)} = \{p_y \subseteq p_{ij}^s \forall y \in \{1, \ldots, k\}\}$;
  $\alpha_k = max(S_{(k)})$;
  $F(S_{(k)}) \leftarrow$ `NPSS` $(\alpha_k$, k, k);
$k^* \leftarrow \arg\max F(S_{(k)})$;
$\alpha^* = \alpha_{k^*}$;
$S^* = S_{(k^*)}$;
return S^*, α^*, and $F(S^*)$

We identify the most anomalous subset for the evaluation samples. The evaluation samples can be clean images or adversarial samples from any of the tested attacks (BIM, FGSM, DF and HSJ) across multiple ϵ values.

IBM Research | Africa

Results in Inner Layers

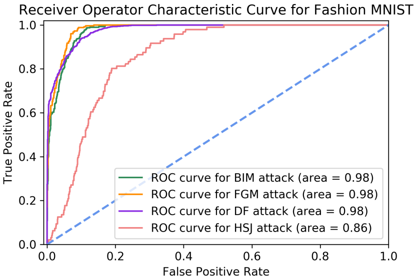
Subset scanning across layers performance

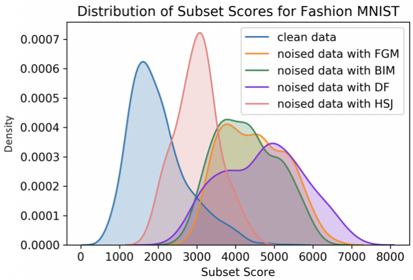
In the latent space, the autoencoder abstracts basic representations of the images, losing subset scanning power due to the autoencoder mapping the new sample to the expected distribution.

Layers	Clean Training								Noised (1%)	Noised (9%)
	F-MNIST				MNIST				F-MNIST	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.)





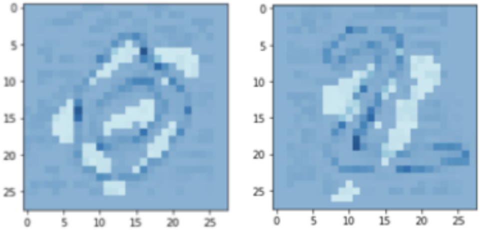
ROC curves & Distribution of subset scores

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.

IBM Research | Africa

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**.

IBM Research | Africa

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.
- Applying our method over the RE space provides the pixels that belong to the most anomalous subset. So we can effectively detect and characterize the nodes that make the input a noised sample.

We're currently working on:

1. How to apply a similar process to other out-of-distribution problems (generated content detection, new class problem, etc.)
2. Explore detected subsets of nodes enable source detection (different type of generative process or types of adversarial attacks).

Code Paper

IBM Research | Africa

Asante, Thanks, Gracias!

celia.cintas@ibm.com

IBM Research | Africa

References I

- Robert H. Berk and Douglas. H. Jones, *Goodness-of-fit test statistics that dominate the Kolmogorov statistics*, Zeitschrift fär Wahrscheinlichkeitstheorie und Verwandte Gebiete 47 (1979), 47–59.
- Jianbo Chen, Michael I Jordan, and Martin J Wainwright, *Hopskipjumpattack: A query-efficient decision-based attack*, arXiv preprint arXiv:1904.02144 3 (2019).
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, CoRR abs/1412.6572 (2015).
- Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio, *Adversarial examples in the physical world*, CoRR abs/1607.02533 (2016).

References II

- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, *Deepfool: a simple and accurate method to fool deep neural networks*, Proceedings of the IEEE CVPR'16, 2016, pp. 2574–2582.
- Edward McFowland III, Skyler D. Speakman, and Daniel B. Neill, *Fast generalized subset scan for anomalous pattern detection*, The Journal of Machine Learning Research 14 (2013), no. 1, 1533–1561.
- Skyler Speakman, Srihari Sridharan, Sekou Remy, Komminist Weldemariam, and Edward McFowland, *Subset scanning over neural network activations*, arXiv preprint arXiv:1810.08676 (2018).

IBM Research | Africa