Anomalous Pattern Detection in Activations and Reconstruction Error of Autoencoder

Anonymous authors
Paper under double-blind review

Abstract

In real-world machine learning applications, large outliers and pervasive noise are commonplace, and access to clean training data as required by standard deep autoencoders is unlikely. Reliably detecting anomalies in a given set of images is a task of high practical relevance for visual quality inspection, surveillance, or medical image analysis. Autoencoder neural networks learn to reconstruct normal images, and hence can classify those images as anomalous if the reconstruction error exceeds some threshold. In this paper, we proposed an unsupervised method based on subset scanning over autoencoder activations. The contributions of our work are threefold. First, we propose a novel method combining detection with reconstruction error and subset scanning scores to improve the anomaly score of current autoencoders without requiring any retraining. Second, we provide the ability to inspect and visualize the set of anomalous nodes in the reconstruction error space that make a sample noised. Third, we show that subset scanning can be used for anomaly detection in the inner layers of the autoencoder. We provide detection power results for several untargeted adversarial noise models under standard datasets.

1 Introduction

Neural networks generate a large amount of activation data when processing an input. This work applies anomalous pattern detection techniques on this activation data in order to determine if the input is anomalous. Examples of an anomalous input can be noised samples by an adversary (Szegedy et al., 2013; Goodfellow et al., 2014; Kurakin et al., 2016a; Dalvi et al., 2004a), human annotation errors (Klebanov et al., 2008), etc. The goal of anomalous pattern detection is to quantify, detect, and characterize the data that are generated by an alternative process. Since anomalies are rare and come from diverse sources, it is not feasible to obtain labeled datasets of all possible anomalies/attacks. If an observation deviates from the learned model, it is classified as an anomaly (Chandola et al., 2009). In real-world problems, large outliers and pervasive perturbations are commonplace, and one may not have access to clean training data as required by standard deep denoising autoencoders (Beggel et al., 2019) due to reasons such as human annotation errors (Klebanov et al., 2008) and poisoning techniques (Dalvi et al., 2004b).

Autoencoders differ from classical classifier networks such as Convolutional Neural Networks (CNNs) (LeCun & Bengio, 1998). Autoencoders do not require labels because the expected output is the input data. The autoencoder is trained to minimize the reconstruction error $L(x, x')$. During the prediction step, anomaly detection can be performed by looking at the distribution of mean reconstruction error $L(w, d(e(w)))$ when $w \in X_{clean}$ and $L(w', d(e(w')))$ when $w' \in X_{adv}$ (Frosst et al., 2018). An example of both, clean and noise reconstruction error distribution can be seen in Figure 4(b). Using this type of anomaly detection with autoencoders assumes that the autoencoder is properly trained with clean data. Otherwise, this manifold can be used advantageously by training the autoencoder with corrupted samples that are mapped to clean samples. As a result, the autoencoder will learn an underlying vector field that points in the direction of the manifold in which the clean samples lie. Thus, upon the introduction of a perturbation, the magnitude of each arrow in the vector field will indicate the direction in which the data must be moved to map the sample to its clean representation (Sahay et al., 2019). Further detail on the autoencoder architecture and training setup for the experiments can be found in the Section A.4.

Subset scanning frames the detection problem as a search over subsets of data in order to find a subset that maximizes a scoring function $F(S)$, typically a likelihood ratio. Subset scanning exploits a property of these scoring functions that allow for efficient maximization over the exponentially large search space (Neill, 2012).

In this paper, we show how subset scanning methods can enhance the anomaly detection power of autoencoders in an unsupervised manner and without a retraining step. We treat this anomaly detection approach as a search for a subset of node activations that are higher than expected. This is formally quantified as the subset with the highest score according to a non-parametric scan statistic.

The contributions of our work are threefold. First, we propose a novel approach combining detection with reconstruction error and subset scanning scores to improve the anomaly score of current autoencoders without requiring any retraining. Second, we provide the ability to identify and visualize the set of anomalous nodes in the reconstruction error space that make noised samples. Third, we show that subset scanning can be used for anomaly detection in the inner layers of the autoencoder.

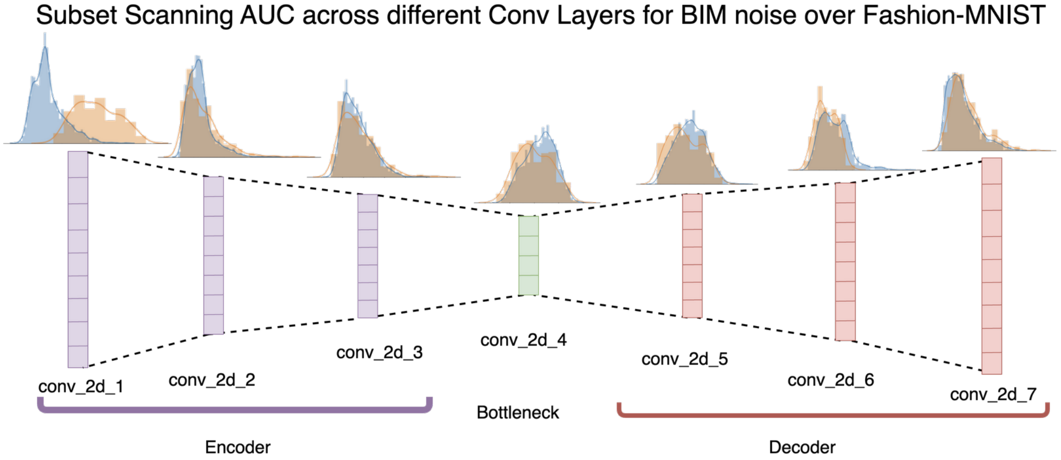


Figure 1: **Example of subset scanning score distributions across layers of an autoencoder for adversarial BIM noise $\epsilon = 0.01$.** In the top of the graph we can see subset score distributions per nodes in a layer. The distributions of subset scanning scores are shown in blue for clean images (C) (expected distribution), and in orange for noised samples A_t. Higher AUCs are expected when distributions are separated from each other and lower AUCs when they overlap. The purple structure corresponds to convolutional layers at the Encoder, while the red structure corresponds to the convolution layers for the Decoder. The computed AUC for the subset score distributions can be found in Table 1. The highest mutual information exchange with the adversarial input happens on the first layers (convolutional and maxpooling). This is why the greatest divergence in both C and A_t subset scores distributions is seen. In the latent space, due to properties described in Section 4, the autoencoder abstracts basic representations of the images, losing subset scanning power due to the autoencoder mapping the new sample to the expected distribution. This can be seen as an almost perfect overlap of distribution in *conv_2d_7*.

2 Related Work and Background Information

Machine learning models are susceptible to adversarial perturbations of their input data that can cause the input to be misclassified (Szegedy et al., 2013; Goodfellow et al., 2014; Kurakin et al., 2016a; Dalvi et al., 2004a). There are a variety of methods to make neural networks more robust to adversarial noise. Some require retraining with altered loss functions so that adversarial images must have a higher perturbation in order to be successful (Papernot et al., 2015; Papernot & McDaniel, 2016).

Our work treats the problem as anomalous pattern detection and operates in an unsupervised manner without a priori knowledge of the attack or labeled examples. We also do not rely on training data augmentation or specialized training techniques. These constraints make it a more difficult problem, but more realistic in the adversarial noise domain as new attacks are constantly being created. Before introducing our approach in the next section, we explain related work in two parts. First, we provide a quick overview of Autoencoders as anomaly detectors and second, we discuss different adversarial attacks models used in this paper.

2.1 Autoencoders as Anomaly Detectors

Several approaches have been used for anomaly detection with autoencoders. Since autoencoders can model training data distribution, these neural networks are an interesting option for anomaly detection. Most of the methods found in the literature require that the training data only consist of normal examples such as denoising autoencoders (Meng & Chen, 2017; Xie et al., 2012), but this alone is no guarantee for anomalies to have a large reconstruction error.

Beggel et al. (2019) present a robust Anomaly Detection with ITSR (Iterative Training Set Refinement) and Adversarial Autoencoders. Their work uses the capabilities of adversarial autoencoders to address the shortcoming of conventional autoencoders in the presence of anomalies samples during training. They also propose a combined criterion of reconstruction error and likelihood in the latent space, as well as a retraining method to increase the separation in both latent and image space.

Zhai et al. (2016) use deep structured energy-based models, showing that a criterion based on an energy score leads to better results than the reconstruction error criterion. Zhou & Paffenroth (2017) present an extension of denoising autoencoders that can work with corrupted data. During training, the network uses an anomaly regularizing penalty based on L_p-norms.

Most of the approaches for anomaly detection with autoencoders require the training data to consist of clean examples or use complex autoencoder architectures and special training. In this work, we propose subset scanning applied to autoencoders. This is an unsupervised anomaly detector that can be applied to any pre-trained, off-the-shelf autoencoder network. We use, as a baseline, the detection capabilities based on mean autoencoder reconstruction error distributions (Sakurada & Yairi, 2014) and One-SVM (Schölkopf et al., 2001) for the autoencoder reconstruction error space analysis.

2.2 Adversarial Attacks

Several attack models have been used to target classifiers in this study, we focus on untargeted attacks with Basic Iterative Method (BIM) (Kurakin et al., 2016b), Fast Gradient Signal Method (FGSM) (Goodfellow et al., 2014), and DeepFool (DF) (Moosavi-Dezfooli et al., 2016). The idea behind these attacks is to find a perturbation to be included in the original sample $\boldsymbol{X}$, generating an adversarial sample $\boldsymbol{X}^{adv}$.

Fast Gradient Sign Method (FGSM)

FGSM (Goodfellow et al., 2014) was designed to be extremely fast rather than optimal. It simply uses the sign of the gradient at every pixel to determine the direction with which to change the corresponding pixel value. Given an image x and its corresponding true label y, the FGSM attack sets the perturbation δ to:

$$\boldsymbol{X}^{adv} = \boldsymbol{X} + \epsilon \operatorname{sign}\left(\nabla_X J\left(\boldsymbol{X}, y_{true}\right)\right) \tag{1}$$

Basic Iterative Method (BIM)

BIM (Kurakin et al., 2016b) is a straightforward extension of FGSM where adversarial noise is applied multiple times iteratively with small step size:

$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = \operatorname{Clip}_{X,\boldsymbol{k}}\left\{\boldsymbol{X}_N^{adv} + \alpha \operatorname{sign}\left(\nabla_X J\left(\boldsymbol{X}_N^{adv}, y_{true}\right)\right)\right\} \tag{2}$$

DeepFool (DF)

The DF algorithm presented by Moosavi-Dezfooli et al. (2016) computes the optimal adversarial perturbation to perform a misclassification. In a binary classifier, the robustness of the model f for an input $\boldsymbol{X}_0$ is equal to the distance the input to the hyper-plane that separates both classes. So the minimal perturbation to change the classifier decision is the orthogonal projection defined as:

$$-\frac{f\left(\boldsymbol{X}_0\right)}{\|w\|_2^2} * w \tag{3}$$

3 Subset Scanning for Anomalous Pattern Detection

Subset scanning treats the pattern detection problem as a search for the "most anomalous" subset of observations in the data. Herein, anomalousness is quantified by a scoring function, $F(S)$ which is typically a log-likelihood ratio statistic. Therefore, the goal is to efficiently identify $S^* = \arg\max_S F(S)$ over all relevant subsets of node activations within an autoencoder that is processing an image at runtime. The particular scoring functions $F(S)$ used in this work are covered in the next sub-section.

Heuristic alternatives to subset scanning include "top-down" and "bottom-up" methods. Top-down approaches detect globally interesting patterns and then identify sub-partitions to find smaller anomalous groups of records. These may fail to detect small-scale patterns that are not evident from global aggregate statistics. Similarly, bottom-up approaches that identify individually anomalous data points and aggregates them into clusters may fail when the pattern is only evident by evaluating a group of data points collectively (Neill, 2012).

Treating the detection problem as a subset scan has desirable statistical properties. However, the exhaustive search over groups quickly becomes computationally infeasible due to the exponential number of subsets of records. Fortunately, a large class of scoring functions used in subset scanning satisfy the "Linear Time Subset Scanning" (LTSS) property that allows for exact, efficient maximization over all subsets of data without requiring an exhaustive search (Neill, 2012). The LTSS property essentially reduces the search space from 2^N to N for a dataset with N records, while guaranteeing that the highest-scoring subset of records is identified.

3.1 Non-parametric Scan Statistics

This work uses non-parametric scan statistics (NPSS) that have been used in other pattern detection methods (Neill & Lingwall, 2007; McFowland III et al., 2013; McFowland et al., 2018; Chen & Neill, 2014). Although subset scanning can use parametric scoring functions (i.e. Gaussian, Poisson), the distribution of activations within particular layers are highly skewed and in some cases bi-modal. See Figure 9. Therefore, this work uses non-parametric scan statistics that makes minimal assumptions on the underlying distribution of node activations.

The intuition behind the role of non-parametric scan statistics is best explained in a simple example. Consider 100 p-values that are supposed to be uniformly distributed between 0 and 1 under the null hypothesis of no anomaly present in the data. A larger-than-expected activation at a node results in a lower p-value for that node. What if we observe 30 (out of 100) p-values all under a threshold value of 0.10? Is that more or less anomalous than finding 20 (out of 100) p-values all under a threshold of 0.075? Non-parametric scan statistics quantify these situations. This same example can be used to highlight why subset scanning is appropriately paired with non-parametric scan statistics. A single $p-$value of 0.1 is not interesting when viewed by itself. However, if there are 29 other $p-$values in the same data set that are also 0.1 (or lower), then the observations are now more interesting when considered together, as a group. Subset scanning efficiently identifies the combination of p-values and thresholds in order to maximize the non-parametric scan statistic.

There are three steps to appropriately use non-parametric scan statistics on neural network activation data. The first is to form a distribution of "expected" activations at each node. This is done by letting the autoencoder process images that are known to be clean from anomalies (sometimes referred to as "background" images) and recording the activations at each node. The second step involves a test image that may be clean or noised and needs to be scored. We record the activations induced by the test image and compare it to the baseline activations created in the first step. This comparison results in a p-value at each node. The third step is to quantify the anomalousness of the resulting p-values by finding the subset of nodes that maximize the non-parametric scan statistic.

We now formalize these three steps. Let there be M background images X_z included in D_{H_0}. These images generate activations $A^{H_0}_{zj}$ at each node O_j. Let X_i (not in D_{H_0}) be a test image under evaluation. This image creates activations A_{ij} at each node O_j. The p-value, p_{ij}, is the proportion of background activations $A^{H_0}_{zj}$ greater than the activation induced by the test image A_{ij} at node O_j. (We note that McFowland III et al. (2013) extend this notion to p-value *ranges* such that p_{ij} is uniformly distributed between p_{ij}^{min} and p_{ij}^{max}). This current work makes a simplifying assumption to only consider a range by its upper bound,

$$p_{ij} = \frac{\sum_{X_z \in D_{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}.$$

A test image X_i is now converted to a vector of p-values p_{ij} of length $J = |O|$, the number of nodes in the network under consideration. Intuitively, if a test image is "natural" (its activations are drawn from the same distribution as the baseline images) then few of the p-values will be extreme. The key assumption is that under the alternative hypothesis of an anomaly present in the activation data, then at least some subset of the activations $S_O \subseteq O$ will systematically appear extreme. We now turn to non-parametric scan statistics to identify and quantify this set of p-values.

The general form of the NPSS score function is

$$F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \phi(\alpha, N_{\alpha}(S), N(S)) \tag{4}$$

where $N(S)$ represents the number of empirical p-values contained in subset S and $N_{\alpha}(S)$ is the number of p-values less than (significance level) α contained in subset S.

Moreover, it has been shown that for a subset S consisting of $N(S)$ empirical p-values, $E\left[N_{\alpha}(S)\right] = N(S)\alpha$ (McFowland III et al., 2013). We assume an anomalous process will create some S where the observed significance is higher than the expected, $N_{\alpha}(S) > N(S)\alpha$, for some α.

There are well-known goodness-of-fit statistics that can be utilized in NPSS (McFowland et al., 2018), the most popular is the Kolmogorov-Smirnov test (Kolmogorov, 1933). Another option is Higher-Criticism (Donoho & Jin, 2004). In this work we use the Berk-Jones test statistic(Berk & Jones, 1979): $\phi_{BJ}(\alpha, N_{\alpha}, N) = N * KL\left(\frac{N_{\alpha}}{N}, \alpha\right)$, where KL is the Kullback-Liebler divergence $KL(x, y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$ between the observed and expected proportions of significant p-values. Berk-Jones can be interpreted as the log-likelihood ratio for testing whether the p-values are uniformly distributed on $[0, 1]$ as compared to following a piece-wise constant alternative distribution, and has been shown to fulfill several optimality properties and has greater power than any weighted Kolmogorov statistic.

3.2 Efficient Maximization of NPSS

Although NPSS provides a means to evaluate the anomalousness of a subset of node activations S_O discovering which of the 2^J possible subsets provides the most evidence of an anomalous pattern is computationally infeasible for moderately sized data sets. However, NPSS has been shown to satisfy the linear-time subset scanning (LTSS) property (Neill, 2012), which allows for an efficient and exact maximization over subsets of data.

The LTSS property uses a priority function $G(O_j)$ to rank nodes and then proves that the highest-scoring subset consists of the "top-k" priority nodes for some k in $1 \ldots J$. The priority of a node for NPSS is the proportion of p-values that are less than α. However, because we are scoring a single image and there is only one p-value at each node, the priority of a node is either 1 (when the p-value is less than α) or 0 (otherwise). Therefore, for a fixed, given α threshold, the most anomalous subset is all and only nodes with p-values less than alpha.

In order to maximize the scoring function over α we first sort the O_j nodes by their p-values. Let $S_{(k)}$ be the subset containing the k nodes with with the smallest p-values. Let α_k be the largest p-value among these k nodes. The LTSS property guarantees that the highest-scoring subset (over all α thresholds) will be one of these J subsets $S_{(1)}, S_{(2)}, \ldots S_{(J)}$ with their corresponding α_k threshold. Any subset of nodes that does not take this form (or uses an alternate α_k) is provably sub-optimal and not considered. Critically, this drastically reduced search space still guarantees identifying the highest-scoring subset of nodes for a test image under evaluation.

Figure 2 shows how the optimal α threshold (and subset size) can vary for different test images under consideration. The leftmost panel shows the distributions of the size of the most anomalous subset of nodes in both clean and noised images. We note that noised images tend to return a larger subset of nodes than clean images. The middle panel shows the optimal α threshold value that maximized the non-parametric scan statistic for clean and noised images. We note that noised images tend to have lower thresholds than clean images. When an image induces a larger number of smaller p-values, the resulting score of the image is higher. This is demonstrated in the right-most panel where noised test images have higher scores than clean test images.

4 Detecting adversarial attacks on Autoencoders

A conventional autoencoder (Bengio et al., 2007) learns the underlying manifold of the training data, which is used to reconstruct the input (x) as the output (x'). The general architecture of any autoencoder involves an encoder and a decoder. The encoder ($e : X \rightarrow Z$) is composed of one or more layers that perform nonlinear dimensionality reduction from the high dimensional input space into a low-dimensional latent representation ($z = e(x)$), while the decoder ($d : Z \rightarrow X$) reconstructs the original sample from the latent representation. Both functions compute $x' = d(e(x))$. The autoencoder is optimized by minimizing the reconstruction error $L(x, x')$.

Anomalous pattern detection can be performed on a trained autoencoder, by looking at the distributions of mean reconstruction error $L(w, d(e(w)))$ when $w \in X_{clean}$ and $L(w', d(e(w')))$ when $w' \in X_{adv}$. Due to the inherent properties of the autoencoder for anomaly detection, we propose two experiments or applications of subset scanning. First, we are interested in subset scanning scores distributions along the layers of the encoder. During the untangling phase ($z = e(x)$) of information reduction from the input space to the latent representation (z), we want to observe until which layer we're able to discriminate the input (clean and noised) to the distribution learnt by the autoencoder. Second, we apply subset scanning methods on the reconstruction error space, to understand if reconstruction error criterion suffices for detection in training autoencoder based anomaly detectors.

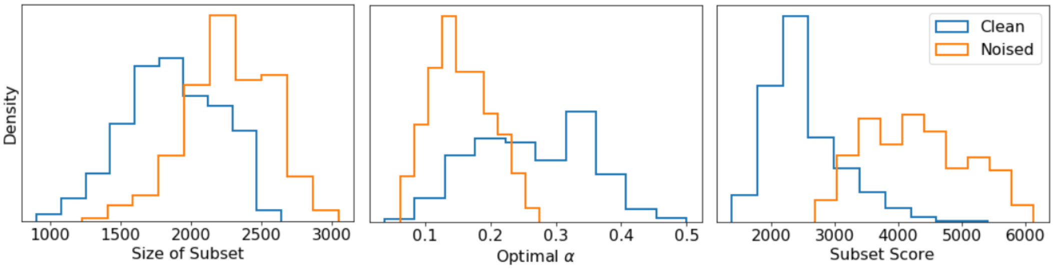


Figure 2: **The connection between the number of nodes in a subset, α value that maximizes the non-parametric scan statistic, and the resulting subset score.** These results are for Fashion-MNIST examples with activations coming from the first layer of the autoencoder. Under the presence of BIM adversarial noise, we observe a larger number of nodes that have smaller p-values. This combination results in a higher subset score than the clean images. Critically, the LTSS property allows α to be efficiently chosen to maximize the score for each individual image. The subset size is all nodes with p-values less than the α threshold. We enforce a $\alpha_{max} = 0.5$ constraint on the search.

5 Experimental Setup

In this section, we describe the baselines methods used as comparison, as well as the datasets, evaluation metric, adversarial noise generation and autoencoder architecture we used. For generating the attacks a standard CNN model LeCun et al. (1998) was trained for both datasets. The test accuracies for these models are 0.992 for MNIST and 0.921 for Fashion-MNIST.

We trained an autoencoder network (Bengio et al., 2007) on MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) (detailed in Section 5.1). The architecture of the autoencoder is depicted in Figure 8, and further details on the training setup can be found in Appendix A.4. The test reconstruction error of the model was 0.284 for Fashion-MNIST and 0.095 for MNIST. In real-world applications, clean training datasets cannot always be guaranteed due to factors such as human annotation errors (Klebanov et al., 2008), and poisoning techniques (Dalvi et al., 2004b). Consequently, we trained the autoencoder with different levels of data poisoning. We trained autoencoders with 100% of clean samples, 1% of adversarial samples, and 9% of adversarial samples. For this experiment, we used BIM as the attack and Fashion-MNIST as the dataset. We evaluated subset scanning over two experiments. First, we applied our subset scanning method on the reconstruction error calculated over the input data and the last layer of the autoencoder (*conv2d_7*). This layer has 1 filter containing 784 nodes. For more information, refer to Section 4. Second, we studied subset scanning patterns across adversarial attacks and datasets, to see if we have some common subset scanning behaviors. For this, we applied subset scanning across all layers of the autoencoder (convolutional, max pooling and up-sampling) and analyzed the detection power in each case.

5.1 Datasets

For our adversarial experiments, we took $M = |D_{H_0}| = 7000$ of the 10000 validation images and used them to generate the background activation distribution (D_{H_0}) at each of the 784 nodes (28×28) for the reconstruction error space and the activations nodes per each inner layer. These 7000 images were not used again. These images form our expectation of "normal" activation behavior for the network. The remaining 3000 images were used to form a "Clean" ($C = 1500$) sample and an "Adversarial" ($A_t = 1500$) noised sample. For the experiments we only kept the successful attacks for DF, FGSM and BIM, so we only preserve noised samples that were incorrectly classified by the model.

We evaluated anomaly detection with subset scanning on the classical MNIST (LeCun et al., 1998) dataset and more complex dataset Fashion-MNIST (Xiao et al., 2017). We present a quick overview of both datasets:

- MNIST (LeCun et al., 1998): The training set has 60000 images and the test set has 10000 images of handwritten digits. Each digit has been normalized and centered to 28×28.
- Fashion-MNIST (Xiao et al., 2017): a relatively new dataset comprising 28×28 grayscale images of 70.000 fashion products from 10 categories, with 7000 images per category. The training set has 60000 images and the test set has 10000 images. As an alternative to MNIST, it has the same image size, data format and validation splits, with the digits from MNIST replaced with 10 products of clothes and accessories.

5.2 Adversarial Noise Sample Generation

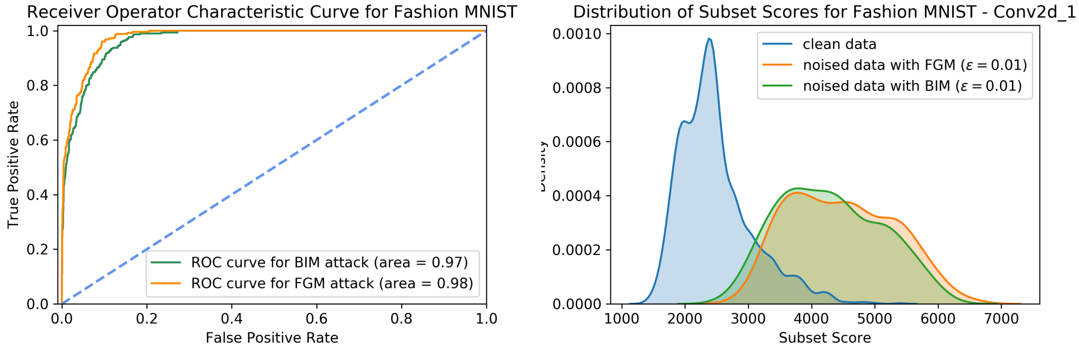
Several adversarial attacks for the subset scanning experiments were implemented, briefly introduced in Section 2.2. Specifically, we describe in this section the hyperparameter selection for Basic Iterative Method (BIM) adversarial attack (Kurakin et al., 2016b), Fast Gradient Signal Method (FGSM) (Goodfellow et al., 2014) and DeepFool (DF) (Moosavi-Dezfooli et al., 2016).

BIM and FGSM have an ϵ parameter which controls how far a pixel is allowed to change from its original value when noise is added to the image. We used a value of $\epsilon = 0.01$ in the scaled $[0, 1]$ pixel space. We also allowed the method to reach its final noised state over 100 steps with each of size 0.002. Smaller values of ϵ make the pattern subtler and harder to detect, but also less likely for the attacks to succeed in changing the class label to the target. For DeepFool, we used standard $\epsilon = 1e-06$ and 100 iterations. Example of generated adversarial samples for both datasets are depicted in Figure 7. All untargeted attacks were generated with the Adversarial Robustness Toolbox (Nicolae et al., 2018)[1]. The set A_t only contains images that were successfully noised by each type of adversarial attack. This means that those samples were misclassified from an original predicted label. The 1500 images in group C are natural and have all class labels represented equally.

6 Results

We adopted the following metric to measure the effectiveness of subset scanning over an autoencoder to distinguishing different types of adversarial attacks images under the activation and reconstruction error space. The Detection Power is measured by AUROC, the Area Under the Receiver Operating Characteristic curve, which is also a threshold independent metric (Davis & Goadrich, 2006). The ROC curve depicts the relationship between true positive rate (TPR) and false positive rate (FPR). Results shown in Figure 6 for reconstruction error and activations space in the first convolutional layer for Figure 3.

[1] `https://github.com/IBM/adversarial-robustness-toolbox`



(a) (b)

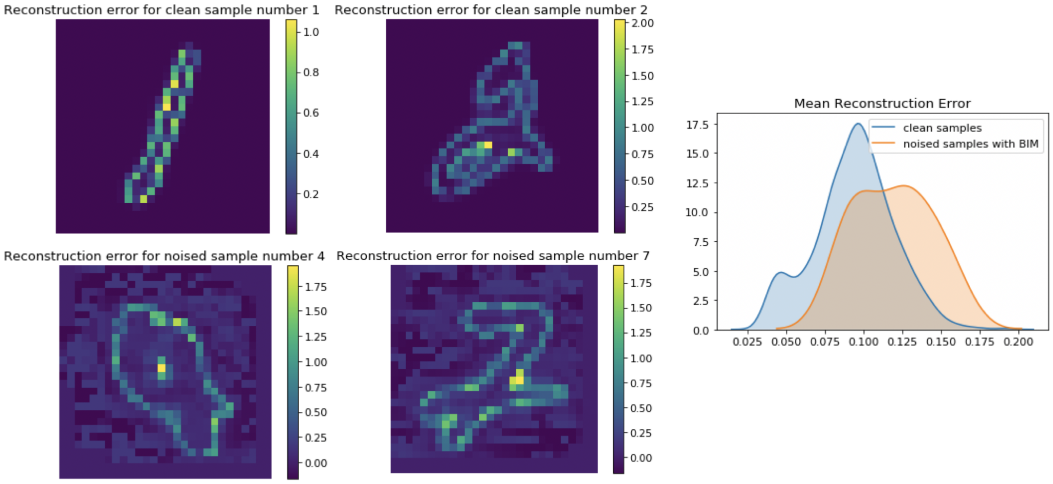
Figure 3: **(a)** ROC curves for each of the noised cases as compared to the scores from test sets containing all natural images for layer *Conv2d_1*. **(b)** Distribution of subset scores for test sets of images over *Conv2d_1*. Test sets containing all natural images had lower than scores than test sets containing noised images. Higher proportion of noised images resulted in higher scores.

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

Table 1: **Detection power for individual subset scanning over all layers (convolutional, max pooling and up-sampling) for both datasets under three different adversarial attacks.** The noised columns refer to the autoencoder being trained with 1% and 9% BIM noised samples. Under different datasets and attacks, the same initial layers hold the highest detection power.

In Table 1, we can observe that across different datasets, noise attacks models, and two proportion of noised samples during training, the first layers (*conv_2d_1* and *max_pooling_2d_1*) maintain a high performance regarding detection power (between 0.96 to 1.0 depending on dataset and noise attack). The ROC curves and subset scores distribution for the BIM and FGSM attacks under Fashion-MNIST for the layer *conv_2d_1* are shown in Figure 3. Furthermore, Table 1 shows that in the cases where 1% and 9% of the samples are noised during training stage of the autoencoder, the detection power of subset scanning still performs correctly, above 0.82.

Table 2 shows the behavior of subset scanning over the reconstruction error space and the detection power in detail for both datasets and different adversarial attacks. We can observe a difference of performance of our method over the Fashion-MNIST dataset. One hypothesis would be that this is due to the autoencoder performance (Loss for Fashion-MNIST 0.284 and MNIST 0.095). To test this idea, we performed preliminary experiments that show a relationship between the decrease in the loss of the trained autoencoder and the increase in the detection power of subset scanning methods under the reconstruction error space. A poorly-trained autoencoder will have a higher loss, while a well-trained autoencoder will have a lower loss. If an autoencoder's loss is high, it is more



(a)

(b)

Figure 4: **Reconstruction error baseline visualization.** **(a)** Baseline mean reconstruction error samples for clean images and BIM noised samples processed by the autoencoder. We can observe that in clean samples reconstruction error only appears on the contours of the number, while noise samples have lower reconstruction error values distributed throughout the image. **(b)** Mean reconstruction error distribution for clean (blue distribution) and noise samples (orange samples).

difficult to separate between clean and noised samples in the reconstruction space. Nonetheless, subset scanning has higher detection power than Mean Reconstruction Error distributions under clean and noise samples (see Figure 4) and Unsupervised outlier detection methods such as One-SVM (Schölkopf et al., 2001). Furthermore, subset scanning under the reconstruction error space is an interesting technique to explore and introspect what nodes or portions of the input image look anomalous. With this information we can not only point out which image looks anomalous, but also indicate which nodes make the input a noised sample, an example of this is depicted in Figure 5.

Datasets	Attacks		Clean Training	
		Subset Scanning	Mean AE Rec. Error	One-SVM
Fashion-MNIST	BIM	**0.698**	0.641	0.478
	FGSM	**0.672**	0.630	0.497
	DF	**0.599**	0.477	0.534
MNIST	BIM	**0.998**	0.751	0.624
	FGSM	**0.983**	0.725	0.624
	DF	**0.992**	0.574	0.637

Table 2: **Detection power for individual subset scanning over reconstruction error space** for both dataset under three different adversarial attacks, two baselines for reconstruction error over AE (Sakurada & Yairi, 2014) and One-SVM over reconstruction error of the AE (Schölkopf et al., 2001).

7 Conclusion and Future Work

In this work, we proposed a novel unsupervised method for adversarial noise detection with off-the-shelf autoencoders and subset scanning. We have successfully demonstrated how subset scanning can be used to gain detection strength against multiple adversarial attacks on images across several datasets, without requiring any retraining or complex deep autoencoder network structures.

Furthermore, we tested subset scanning over the reconstruction error space and observed significant variations depending on the dataset, autoencoder architecture, and training setup. We performed

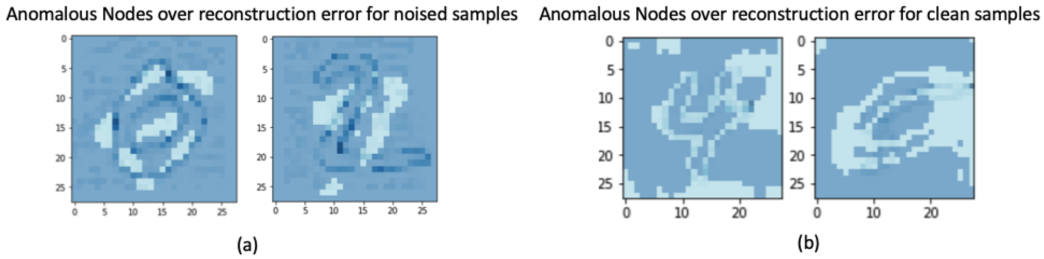


Figure 5: **Anomalous nodes visualization.** Overlap of anomalous nodes (white) and reconstruction error (darker blue) per sample. **(a)** Noised samples with BIM. We can observe that nodes outside the contour will make the sample be classified as noised. **(b)** Whereas clean we expect the anomalous nodes will be along the contour of the figure.

preliminary experiments that yielded a relation between a decrease in the loss of the trained autoencoder and an increase in the detection power of subset scanning under the reconstruction error space. Nonetheless, applying our method under this space provides introspection capabilities that allow us to identify the nodes or portions of the input image look anomalous. Consequently, we are able to not only point out which image looks anomalous but also characterize the nodes that make the input a noised sample. We also evaluated the performance of applying subset scanning over the autoencoder's activations. We observed a consistent and high detection power results across noise attacks, datasets, autoencoders architectures and different noised training levels in the initial layers (*Convolutional and MaxPooling layers*).

Due to versatile properties of subset scanning under neural network activation analysis it may be used for several other studies, including unsupervised classification in the latent space of an autoencoder. We would expect that same class images will identify as a subset of inputs (images) that have higher-than-expected activations (i.e. large number of low empirical $p-$values) at a subset of nodes. Subset scanning applied to autoencoders activations is a novel, unsupervised anomaly detector that can be applied to any pre-trained, off-the-shelf neural network, previously only used in classifier neural networks such as CNNs and ResNet (Speakman et al., 2018).

REFERENCES

Laura Beggel, Michael Pfeiffer, and Bernd Bischl. Robust anomaly detection in images using adversarial autoencoders. *arXiv preprint arXiv:1901.06355*, 2019.

Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

R. H. Berk and D. H. Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47:47–59, 1979.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Feng Chen and Daniel B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1166–1175, 2014.

Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '04, pp. 99–108, New York, NY, USA, 2004a. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014066. URL http://doi.acm.org/10.1145/1014052.1014066.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108. ACM, 2004b.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM, 2006.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, 2004.

Nicholas Frosst, Sara Sabour, and Geoffrey Hinton. Darccc: Detecting adversaries by reconstruction from class conditional capsules. *arXiv preprint arXiv:1811.06969*, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL http://arxiv.org/abs/1412.6572.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. Analyzing disagreements. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pp. 2–7. Association for Computational Linguistics, 2008.

Andrej N Kolmogorov. *Sulla determinazione empirica di una legge di distribuzione*. na, 1933.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *CoRR*, abs/1611.01236, 2016a. URL http://arxiv.org/abs/1611.01236.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016b. URL http://arxiv.org/abs/1607.02533.

Yann LeCun and Yoshua Bengio. The handbook of brain theory and neural networks. chapter Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-51102-9. URL http://dl.acm.org/citation.cfm?id=303568.303704.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. ISSN 00189219. doi: 10.1109/5.726791.

E. McFowland, III, S. Somanchi, and D. B. Neill. Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection. *ArXiv e-prints*, March 2018.

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(1):1533–1561, Jun 2013.

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *ACM Conference on Computer and Communications Security (CCS)*, 2017. URL `https://arxiv.org/abs/1705.09064`.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012.

Daniel B. Neill and J. Lingwall. A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*, 4:106, 2007.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v0.7.0. *CoRR*, 1807.01069, 2018. URL `https://arxiv.org/pdf/1807.01069`.

Nicolas Papernot and Patrick D. McDaniel. On the effectiveness of defensive distillation. *CoRR*, abs/1607.05113, 2016. URL `http://arxiv.org/abs/1607.05113`.

Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *CoRR*, abs/1511.04508, 2015. URL `http://arxiv.org/abs/1511.04508`.

Rajeev Sahay, Rehana Mahfuz, and Aly El Gamal. Combatting adversarial attacks through denoising and dimensionality reduction: A cascaded autoencoder approach. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2019.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pp. 4. ACM, 2014.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Skyler Speakman, Srihari Sridharan, Sekou Remy, Komminist Weldemariam, and Edward McFowland. Subset scanning over neural network activations. *arXiv preprint arXiv:1810.08676*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL `http://arxiv.org/abs/1312.6199`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, 2012. ISBN 9781627480031.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. *arXiv preprint arXiv:1605.07717*, 2016.

Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674. ACM, 2017.

A Appendix

A.1 Algorithm for subset scanning over Autoencoder activations

```
input  : Background set of images: $X_z \in D^{H_0}$, evaluation image: $X_i$, $\alpha_{\max}$.
output: $S^*_E$ Score for the evaluation image
$AE \leftarrow$ 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 ($AE_y$, $X_z$)
    end
end
for $j \leftarrow 0$ to $J$ do
    $A_{ij} \leftarrow$ ExtractActivation ($AE_y$, $X_i$)
end
$p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
$p^s_{ij} \leftarrow$ 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);
end
$k^* \leftarrow \arg\max F(S_{(k)})$;
$\alpha^* = \alpha_{k^*}$;
$S^* = S_{(k^*)}$;
return $S^*$, $\alpha^*$, and $F(S^*)$
```

Algorithm 1: Pseudo-code for subset scanning over autoencoder activations.

A.2 Subset Scanning under Reconstruction Error for MNIST

In Figure 6, we can observe the distribution of subset scores for test sets of images over reconstruction error. Test sets containing all natural images had lower scores than test sets containing noised images (FGSM and BIM generated samples). Higher proportion of noised images resulted in higher scores. Figure 6 also shows the ROC curves for each of the noised cases as compared to the scores from test sets containing all natural images.

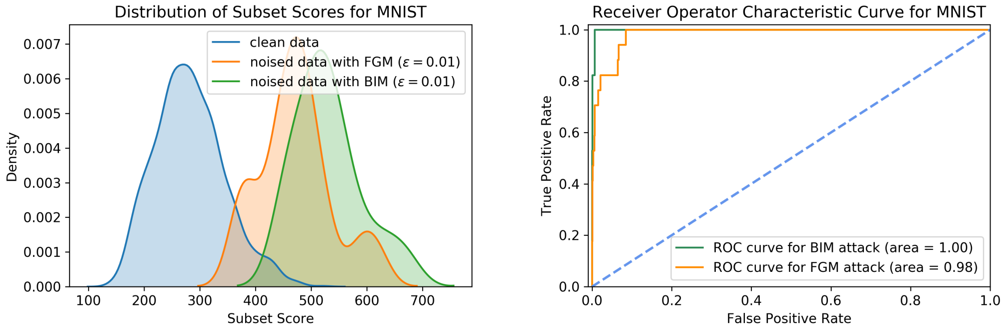


Figure 6: (a) Distribution of subset scores for test sets of images over reconstruction error. Test sets containing all natural images had lower than scores than test sets containing noised images. Higher proportion of noised images resulted in higher scores. (b) ROC curves for each of the noised cases as compared to the scores from test sets containing all natural images.

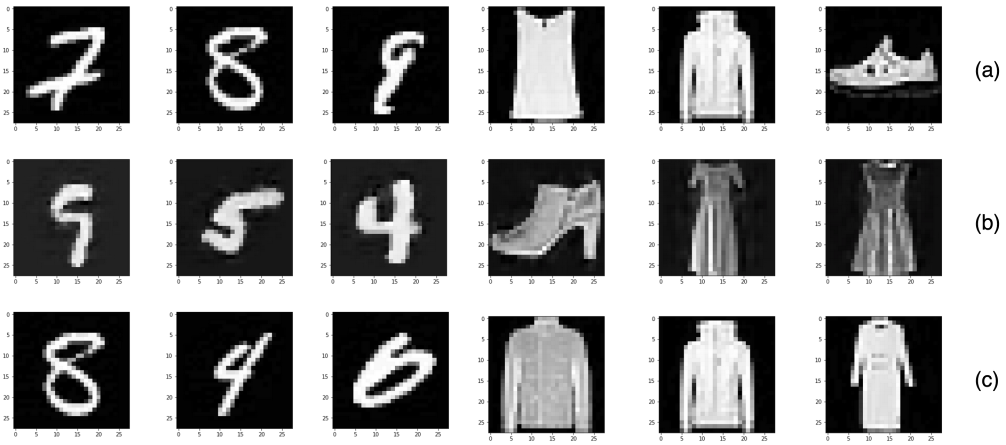


Figure 7: **Successful noised samples from MNIST and Fashion-MNIST generated with Adversarial Robustness Toolbox (Nicolae et al., 2018)**. (a) BIM noised samples (b) DeepFool samples (c) FGSM. Parameters used for each attack are detailed at Section 5.2.

A.3 Noised samples qualitative visualization

Several examples of successful noised samples from both datasets and the three different attacks are depicted in Figure 7.

A.4 Architecture and Training setup for Autoencoder

We train the same autoencoder architecture (4385 parameters) for both Fashion-MNIST and MNIST, A diagram of the network is shown in the Figure 8. The encoder comprises three (3) convolutional layers with $16, 8, 8$ filters respectively, a kernel size of three (3), each with relu activations, and a maxpooling layer with a pool size of two (2) after every convolutional layer. The decoder comprises four (4) convolutional layers with $8, 8, 16, 1$ filters respectively, a kernel size of three (3), each with relu activations except the final layer which uses a sigmoid. Each consecutive pair of convolutional layer is interspersed with an upsampling layer with a size of two (2). We train the autoencoder by minimizing the binary cross-entropy of the decoder output and the original input image using an adadelta optimizer (citep) for 100 epochs taking 128 records per batch.

A.5 Activation values distribution across all layers

Although subset scanning can use parametric scoring functions (i.e. Gaussian, Poisson), the distribution of activations within particular layers are highly skewed and in some cases bi-modal. See Figure 9. Therefore, this work uses non-parametric scan statistics that makes minimal assumptions on the underlying distribution of node activations.

Furthermore we only consider 1-tailed p-values (in the greater direction). This is due to nuances of the ReLu activation function. Alternative activation functions such as tanh and signmoid would allow an "extreme" activation to be considered as either larger or smaller than expected with a p-value coming from a 2-tailed calculation.

A.6 Non-parametric Scan Statistics

NPSS can be viewed as a second-order test statistic that operate on (by aggregating information across) p-values (i.e., the first order test statistics) to evaluate the the evidence for violations of H_0 in a given subset S. NPSS is operationalized with a given score (test) function; each test is powered for different alternatives, and therefore, NPSS's detection power is linked to preferences of the selected score function.

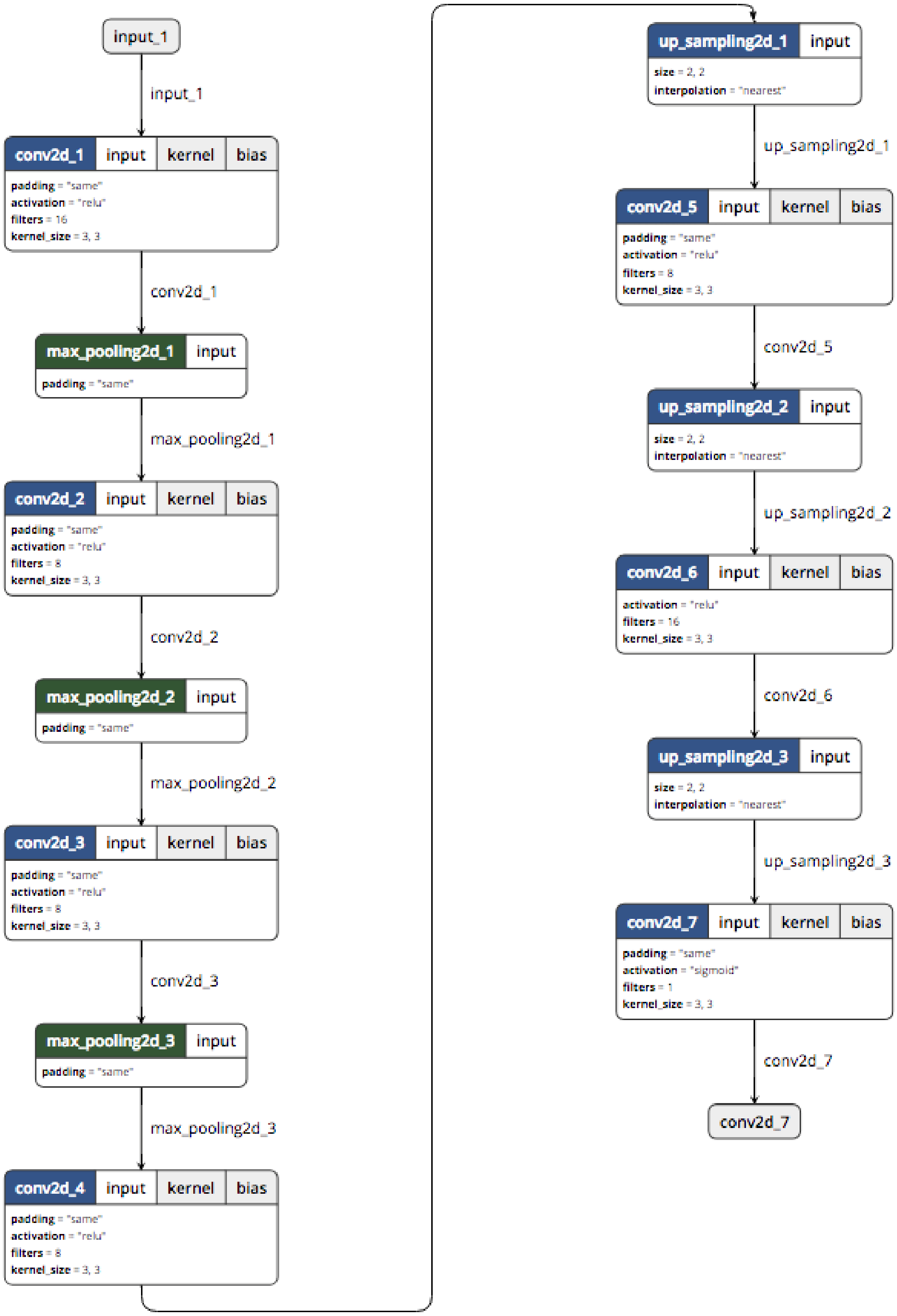


Figure 8: **Autoencoder architecture diagram.** Same architecture was implemented for both datasets and for different proportions of noised samples.

This work used the Berk-Jones scoring function (Berk & Jones, 1979).

$$\phi_{BJ}(\alpha, N_\alpha, N) = N * KL\left(\frac{N_\alpha}{N}, \alpha\right) \tag{5}$$

Where KL is the Kullback-Liebler divergence $KL(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$ between the observed and expected proportions of significant p-values. Berk-Jones can be interpreted as the log-likelihood ratio for testing whether the p-values are uniformly distributed on $[0, 1]$ as compared

Table 3: Examples of Nonparametric Scan Statistics that satisfy the Linear Time Subset Scanning property.

Name	$\phi(\alpha, N_\alpha, N)$
Higher Criticism	$\frac{\lvert N_\alpha - N\alpha \rvert}{\sqrt{N\alpha(1-\alpha)}}$
Berk-Jones	$N * KL\left(\frac{N_\alpha}{N}, \alpha\right)$
Kolmogorov Smirnov	$\frac{\lvert N_\alpha - N\alpha \rvert}{\sqrt{N}}$

to following a piece-wise constant alternative distribution, and has been shown to fulfill several optimality properties.

A more commonly known scoring function that also satisfies the LTSS property is the Kolmogoorv-Smirnov test statistic

$$\phi_{KS}(\alpha, N_\alpha, N) = \frac{|N_\alpha - N\alpha|}{\sqrt{N}} \tag{6}$$

which is known to be more sensitive to deviations in the center of a distribution. `https://www.jstor.org/stable/2958837` and `http://www.jstor.org/stable/2958836`.

Another test is Higher-Criticism Donoho & Jin (2004):

$$\phi_{HC}(\alpha, N_\alpha, N) = \frac{|N_\alpha - N\alpha|}{\sqrt{N\alpha(1-\alpha)}} \tag{7}$$

which can be interpreted as the test statistic of a Wald test for the amount of significant p-values given that N_α is binomially distributed with parameters N_α and α. Because Higher-Criticism normalizes by the standard-deviation of N_α, it tends to be more sensitive to small subsets with very extreme p-values.

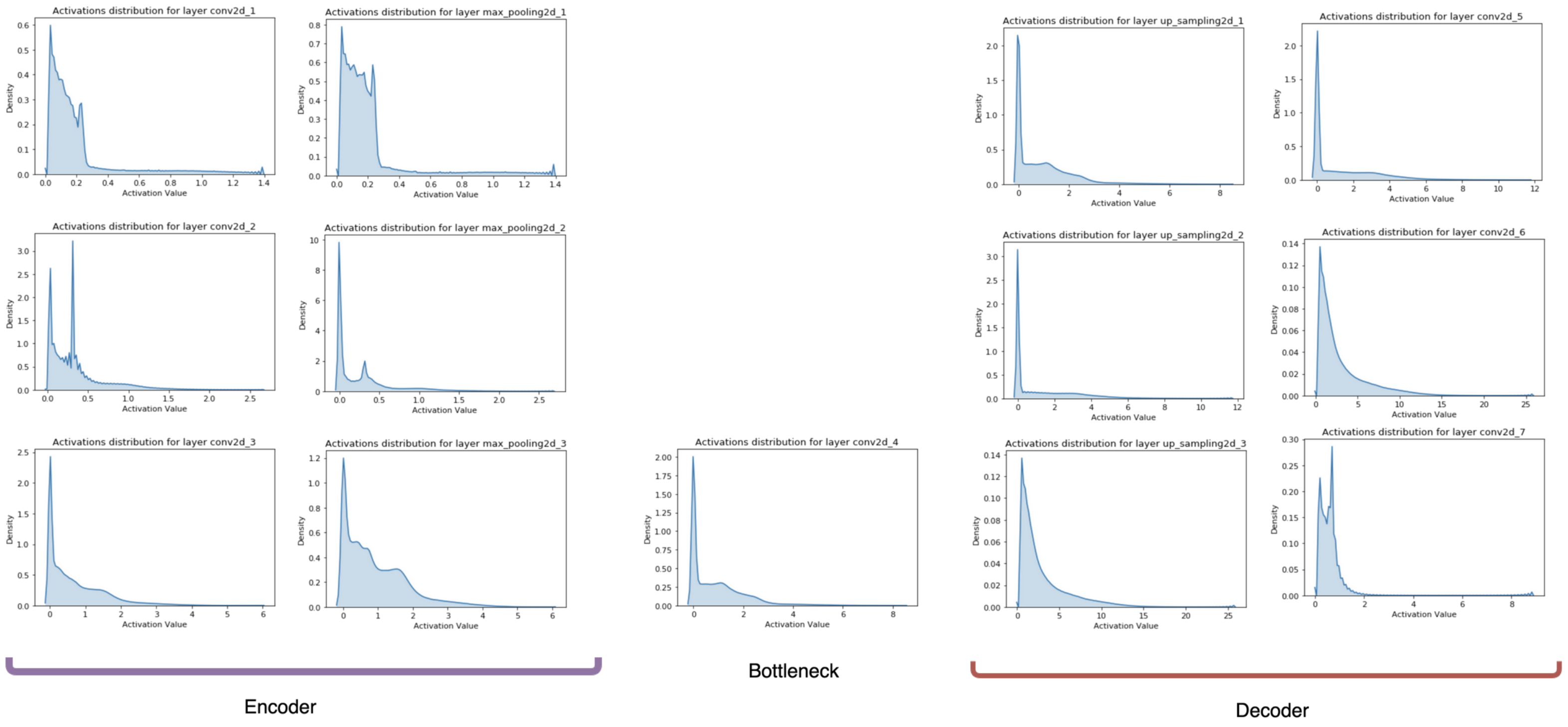


Figure 9: **Activation distribution values across all layers**. Most values are accumulated around 0 due to ReLu activations. The large skew and sometimes bi-modal distribution of activations motivated the use of non-parametric scan statistics to quantify what it means for an activation to be larger-than-expected.