Attack Agnostic Statistical Method for Adversarial Detection*

Sambuddha Saha Aashish Kumar Pratyush Sahay George Jose
Srinivas Kruthiventi Harikrishna Muralidhara
Harman International India Pvt. Ltd., Bangalore
{sambuddha.saha, aashish.kumar, pratyush.sahay, george.jose,
srinivas.sai, harikrishna.muralidhara}@harman.com

Abstract

Deep Learning based AI systems have shown great promise in various domains such as vision, audio, autonomous systems (vehicles, drones), etc. Recent research on neural networks has shown the susceptibility of deep networks to adversarial attacks - a technique of adding small perturbations to the inputs which can fool a deep network into misclassifying them. Developing defenses against such adversarial attacks is an active research area, with some approaches proposing robust models that are immune to such adversaries, while other techniques attempt to detect such adversarial inputs. In this paper, we present a novel statistical approach for adversarial detection in image classification. Our approach is based on constructing a per-class feature distribution and detecting adversaries based on comparison of features of a test image with the feature distribution of its class. For this purpose, we make use of various statistical distances such as ED (Energy Distance), MMD (Maximum Mean Discrepancy) for adversarial detection, and analyze the performance of each metric. We experimentally show that our approach achieves good adversarial detection performance on MNIST and CIFAR-10 datasets irrespective of the attack method, sample size and the degree of adversarial perturbation.

1. Introduction

Deep Learning has been instrumental in the past few years in various domains such as computer vision [11], audio processing [8], natural language processing [3] [4] and autonomous vehicles [1]. However, it has recently been shown that these deep networks can be fooled by adding subtle perturbations to the input resulting in misclassification. These perturbed inputs which can still be classified correctly by humans, are known as adversaries [5] [15].

Two types of approaches are proposed to handle these adversarial attacks. The first approach makes a model robust by training with adversarial examples [9] [17]. It applies random perturbations to activations or weights, or it performs feature denoising or by defensive distillation [18] to make a model robust to adversaries.

Other defence approaches based on adversarial detection either use auxiliary networks [14] or modify the model architecture and add a detection module and train on adversaries to detect them [14]. These approaches are often model centric and are not robust enough for all kinds of attacks. Any network based approach is costly as it involves re-training and customizing the defences for different attacks is also a costly operation.

Earlier Grosse et. al. [7] proposed a statistical based approach which detect adversaries based on the assumption that the original images and the adversarial images belong to two different distributions. They use raw vectorized original images from train set to create a reference distribution and that form the test set to create a test distribution. They create another test distribution from the adversarial images. They compare the two test distributions against the reference distribution using MMD and ED to calculate distances and perform a two sample kernel test to detect if the test distribution belongs to the reference distribution or not. One disadvantage of their experiments is that since they use raw images, the dimensionality is high, so they require samples of higher sizes to approximate the distribution and achieve higher detection confidence (50-100 for the entire dataset). They have also reported per class detection confidence also and they need samples of lower sample size than that for the whole dataset but still it is as high as 20-50 samples.

In this paper, we propose a novel statistical based adversarial detection approach which is agnostic to attacks. Our hypothesis is that the distribution of activations (output of any intermediate layer) of the original data for a particular class is different from that of the adversarial data misclassified into that class. We make use of various statistical metrics to estimate the distance between distributions of the

*The first workshop on Statistical Deep Learning for Computer Vision, in *Seoul, Korea*, 2019.

original and the adversarial activations. Adversarial samples will have larger statistical distances from the original distribution and hence can be detected. The proposed approach is attack agnostic (does not vary with the type or degree of attack) and sample efficient (sets with less sample size achieve good detection performance).

2. Background

A neural network takes an input x and gives an output, y. The outputs are termed as softmax probabilities where y_i denotes the probability of the input x belonging to the class i. The softmax probabilities sum up to 1 and lie in the range of 0 to 1. The output label for a particular input, $l(x)$ is assigned by the model as $l(x) = argmax_i(y_i)\ \forall i \in C$ where C is the total number of classes. The correct label for the class is denoted as $l^*(x)$. The input to the second last layer of the model is termed as *pre-logits* and that to the last softmax layer of the model is termed as *logits*.

Adversarial generation involves perturbing an input x by a small amount to x' such that the output label of the perturbed input is not same as the output label of the original input, i.e. $l^*(x) \neq l^*(x')$ where $abs(x - x') < \epsilon$ where ϵ is the amount of perturbation and abs represents absolute difference. In the next section we will discuss the various adversarial attacks used in our work.

2.1. Adversarial Attacks

Fast Gradient Sign Method (FGSM): Goodfellow et. al. (2015) [5] proposed this attack where the perturbation Δx is based on the gradient of the loss function with respect to the input such that the loss function of the network $C(x, y)$ is maximized. The perturbation is obtained by

$$\Delta x = \epsilon.sign(\nabla_x C(x, y)) \tag{1}$$

where ϵ is the L_∞ norm bound. It is chosen to be small so that Δx is undetectable. The $sign$ refers to the direction in which the input feature has to be changed.

Carlini-Wagner (CW-$l2$): Carlini Wagner et. al. [2] proposed an attack using an optimization framework that perturbs the input by inducing very small changes at each iteration to maximize a predefined loss. It generates attack for three different loss metrices, L_0, L_2 and L_∞. We have used Carlini Wagner L_2 attack in this paper.

Madry et. al. Attack: Madry et. al. [13] proposed a robust optimization based attack to generate adversaries with varying degrees of perturbation, ϵ. They came up with stronger attacks than FGSM using PGD (Projected Gradient Descent).

In the next section, we give a brief description of the various statistical metrics used in this paper.

2.2. Statistical Metrics

Maximum Mean Discrepancy (MMD): Gretton et. al. [6] introduced a kernel based test to compute the distance between probability distributions of two sample sets. The kernel for probability distribution function is chosen such that the difference of the means of the two distributions is maximum.

$$\begin{aligned} MMD_b[F, X_1, X_2] = sup_{f \in F}(\frac{1}{n}\Sigma_{i=1}^{n} f(x_{1i}) \\ -\frac{1}{m}\Sigma_{i=1}^{m} f(x_{2i})) \end{aligned} \tag{2}$$

X_1 and X_2 refer to the two sample sets and f is the kernel function chosen from F where F represents the super-set of all kernel functions possible. f is chosen to be the kernel which maximizes the difference between the means of the two probability distributions. x_{1i} and x_{2i} denote the probability values of the samples belonging to X_1 and X_2 respectively for each class i and m and n denotes the number of samples.

Energy Distance (ED): Szekely et. al. [19] proposed an energy based approach to compute distances between two distributions. Let us assume F and G to be two cumulative distribution functions. X, X' and Y, Y' are independent vectors chosen from F and G respectively which belong to real numbers set R^d. The energy distance between the two distributions F and G is the square root of:

$$\begin{aligned} D^2(F, G) = 2E||X - Y|| - E||X - X'|| \\ -E||Y - Y'|| \end{aligned} \tag{3}$$

where E denotes expectation, $||.||$ denotes the norm. ED calculates the distance between two distributions by considering norm distances between samples of different distributions and that of same distribution.

3. Methodology

Our method is based on the hypothesis that the original image activations sample and the adversarial image activations sample belong to two different distributions. We performed a statistical distance based analysis to differentiate between the original and adversarial activations distribution for each class.

Fig.1 shows a brief overview of the methodology we are following. The model is trained on the data(x, gt) where x is the input image and gt is the ground truth label. We store the output labels and extract the activations (hidden layer activations) from the model. The activations for each class are generally clustered together and each cluster represent different classes as shown in the figure. As observed from the figure, the partition lines are the decision boundaries. When this model is attacked by adversarial samples, the

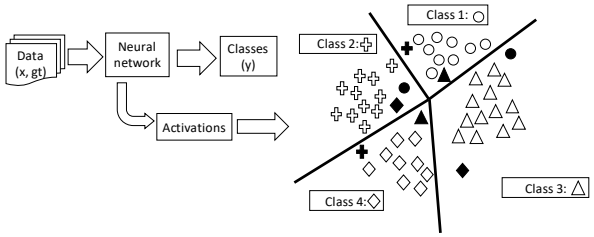

Figure 1. Illustration of our hypothesis: The activations are extracted from the network for both the original and the adversarial samples. These are shown in a representative plot demarcated by the decision boundaries in the above figure. Adversarial samples which are misclassified are indicated with filled markers. It can be observed that while the activations of original samples belonging to a class cluster together, the adversaries remain as outliers indicating that they do not fit in the distribution.

adversarial sample activations lie far away from the original class activations distribution and are misclassified as another class.

The triangles (refer Fig. 1) refer to Class 3 original activations distribution where the samples belonging to that are clustered together. The adversarial samples to this class, like the class 4 sample (triangle inside diamond space) or the class 1 sample (triangle inside the circle space) lie further away from the original distribution.

4. Experiments

We perform experiments to validate our hypothesis on MNIST (Modified National Institute of Standards and Technology) [12] and CIFAR-10 (Canadian Institute For Advanced Research) [10]datasets.

4.1. Network Setup

The table below shows the model architecture used for MNIST (refer Table 1). We use the default convolution neural network present in the cleverhans repositiory [16].

ID	Layer Type	Kernel Size	# O/p Channels	Stride
1	Conv,Relu	8	32	2
2	Conv,Relu	6	64	2
3	Conv, Relu	3	128	1
4	Conv, Relu	2	128	1
5	Dense		256	
7	Dense		128	

Table 1. Model architecture for MNIST

The neural network is trained for 220 epochs with 0.001 learning rate and batch size 128 using Adam optimizer.

We used the same network as above for CIFAR-10 but increased the number of channels in each convolution layers by 4 times. This neural network is trained on the training dataset for 400 epochs, with learning rate 0.001, Adam optimizer and batch size 128. After the training is over, we calculate the accuracy of the model on the test set.

We discard the misclassified samples from the test data after accuracy evaluation as these might lead to false adversaries in adversarial set.

4.2. Adversarial Attack Generation and Activations Extraction

We attack our model using three adversarial attack generation techniques FGSM, Madry and Carlini Wagner. FGSM and Madry attack are generated for 5 varying degrees of perturbations (epsilons), 0.01, 0.05, 0.1, 0.2 and 0.3. We generate adversaries on the correctly classified samples of test data only.

Here we describe the method for activations extraction and distribution generation. We store the *original labels* (ground truth labels for original sample) and extract *logits* and *pre-logits* from the trained model. We store the *adversarial labels* (predicted labels for adversarial sample) and extract *logits* and *pre-logits* from the model. Here we have considered logits as the activations for original and adversarial samples. The original activations are partitioned into a *baseline holdout set* and *rest of the original activations* in another set. The *baseline holdout set* is of fixed size having 100 samples or half the number of total original samples present for that class. The rest represents the *rest of the original activations* set. So we have three activations set now, *baseline holdout set*, *rest of the original activations* set and the adversarial activations set. The *baseline holdout set* is our *reference set*. The other 2 sets are our *test sets*. Each set is a 2-D matix made up of 1-D activation vectors corresponding to each image. We apply softmax over the activation vector for a sample and perform this for all samples present in all the three sets. A fixed size set of randomly sampled samples is picked from adversarial activations set and *rest of the original activations* set. This is our test sample size which tells us if that number of samples is enough to distinguish between original and adversarial samples.

4.3. Statistical Distances Computation and AUC Scores Generation

We calculate statistical metrics for *rest of the original activations* set w.r.t the *reference set* and for the adversarial activations distribution w.r.t the *reference set*. We repeat the above operation 100 times, each time randomly sampling our *test sets* for a particular sample size. We compute AUC score for a particular class for a particular sample size and degree of perturbation. The AUC scores indicate how well the original and adversarial samples can be separated. These scores are tallied for varying sample sizes, degrees of perturbation and different classes.

Sample Sizes	FGSM		MADRY		CARLINI WAGNER	
	ED	MMD	ED	MMD	ED	MMD
1	98.7 (± 0.4)	33.6 (± 11.88)	99.5 (± 0.2)	30.2 (± 10.87)	100 (± 0)	33.1 (± 12.5)
5	99.6 (± 0.04)	42.8 (± 0.33)	99.7 (± 0.02)	42.1 (± 0.42)	99.9 (± 0)	41.9 (± 0.37)
10	99.7 (± 0.03)	43.8 (± 0.3)	99.8 (± 0.01)	43.2 (± 0.32)	99.9 (± 0)	43 (± 0.32)
20	99.8 (± 0.01)	45.5 (± 0.19)	99.8 (± 0)	44.7 (± 0.19)	99.9 (± 0)	44.9 (± 0.24)

Table 2. AUC scores (%) for MNIST dataset

Sample Sizes	FGSM		MADRY		CARLINI WAGNER	
	ED	MMD	ED	MMD	ED	MMD
1	75.9 (± 2.66)	42.1 (± 7.74)	88.4 (± 2.68)	34.9 (± 10.42)	94.2 (± 2.08)	35.7 (± 13.82)
5	83.1 (± 0.4)	50.1 (± 0.02)	91.9 (± 0.13)	45.9 (± 0.13)	94.5 (± 0.22)	49.1 (± 0.07)
10	84.6 (± 0.42)	50.1 (± 0.01)	92.3 (± 0.11)	46.3 (± 0.12)	94.8 (± 0.19)	49.3 (± 0.05)
20	87.4 (± 0.3)	50.1 (± 0)	92.9 (± 0.09)	46.9 (± 0.12)	95.4 (± 0.11)	49.5 (± 0.02)

Table 3. AUC scores (%) for CIFAR-10 dataset

5. Results and Discussion

We present our results on MNIST and CIFAR-10 with three different kinds of attack FGSM, Carlini Wagner and Madry using two statistical distances, MMD and ED. We compute the mean and standard deviation of AUC scores across all the classes and all the epsilons (degrees of perturbation), 0.01, 0.05, 0.1, 0.2, 0.3. To maintain the brevity of the paper we are showing results corresponding to test sample sizes 1,5,10 and 20.

5.1. MNIST

We trained our model on MNIST and it gave a test accuracy of 99.44%. The AUC scores for ED were better than that for MMD for our FGSM experimental setup. We observed similar trends for Madry and Carlini Wagner attacks using MMD and ED (see Table 2).

5.2. CIFAR-10

Our model trained on CIFAR-10 gave a test accuracy of 71.8% which isn't high enough but surprisingly AUC scores were ranging from 0.75 to 0.87 for FGSM attack using ED. MMD didn't perform that well. We observed similar trends for Madry and Carlini Wagner attack on CIFAR (see Table 3).

5.3. Discussion

We obtained the following insights by analysing our results 1) The AUC scores obtained using ED were better than that of MMD for our model across all the three attacks and two datasets. 2) The AUC scores increase proportionally with increase in sample size of the *test set* (works well for test sample of size 1 also) as expected. 3) The AUC scores vary negligibly with change in the degree of attack.

Hence our model is attack agnostic, which means it doesn't vary with the kind of attack and degree of perturbation. Our model is sample efficient because we experimentally demonstrated that even if the size of our test sample set is one, we are able to achieve good detection.

Since the statistical distance,especially ED performs so well in separating the original and adversarial distributions, it proves our hypothesis that the adversaries don't belong to the same distribution as the natural image distribution and hence can be separated by such statistical distance metrics. The results also prove that the learnt features extracted from the model which are low-dimensional, provides a good approximation of the data. Hence we don't need samples of large sizes to get high detection performance.

6. Conclusion

We experimentally demonstrated that the original and adversarial sample do not belong to the same distribution. We also experimentally validated our approach to be attack agnostic and sample efficient. We could expand this work to include more statistical distance metrics and also can extend to use pre-logits. More research will surely contribute to coming up with better statistical models for detecting adversaries.

References

[1] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

[3] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

[5] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 12 2014.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[7] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

[8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[9] C. Kereliuk, B. L. Sturm, and J. Larsen. Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071, 2015.

[10] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[14] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

[15] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[16] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman, and P. McDaniel. cleverhans v2. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 10, 2016.

[17] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[18] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.

[19] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.