# Defending Deep Neural Networks against Structural Perturbations

**Uttaran Sinha** [1] **Saurabh Joshi** [1] **Vineeth N Balasubramanian** [1]

## Abstract

Deep learning has had a tremendous impact in the field of computer vision. However, the deployment of such algorithms in real-world environments rely heavily on its robustness to noise. A lot of work has been put forward in recent years to analyse and defend such models against attacks that cause a slight perturbation in the image and change the output of the network. This work focuses on testing robustness of a model against naturally occurring structural perturbations and, we propose a systematic way to defend against such attacks. This is in contrast to a few other works where complicated optimisation methods are required to generate an adversarial example. We also analyse the effect of adversarial training on the decision boundary of the model. Our work strives to ensure the safety of deep learning in multiple domains such as facial recognition, automated driving and object detection. This paper primarily focuses on image classification and, we believe that our algorithm works independently of the model architecture and dataset.

## 1. Introduction

Deep learning has ushered a new era in the field of computer vision through its application in multiple domains while achieving better performance than traditional computer vision techniques. Despite this success, the latent functionality of deep networks is still not understood. This has raised concerns regarding the actual reliability and robustness of deep learning under real-world scenarios. There has been a significant amount of work ((Szegedy et al., 2014), (Goodfellow et al., 2015), (Huang et al., 2017) (Moosavi-Dezfooli et al., 2017), (Carlini & Wagner, 2017), (Papernot et al., 2016), (Papernot et al., 2017), (Madry et al., 2018)) dedicated to understanding the existence of adversarial attacks on machine learning algorithms and especially in deep learning. Several attacks and defence techniques have so far been proposed (as summarised in (Akhtar & Mian, 2018)) to analyse the robustness of such networks. Even state-of-the-art deep learning architectures are vulnerable to minor perturbations in the input that alter the expected output of a network with high confidence ((Nguyen et al., 2015)). This flaw can easily be exploited by malicious individuals, and hence, we should strive to find the cause of such behaviours and increase the robustness of a model before deploying them in real-world scenarios.

The existence of these attacks restrict the application of machine learning in safety-critical areas where reliability, dependability, and security hold paramount importance. Therefore, the development of defence techniques against adversarial attacks on Deep Neural Networks (DNN) is the need of the hour. A few such defence methods focus on defending attacks on models whose weights are known (white box attack) such as (Papernot et al., 2016) and (Madry et al., 2018). Other approaches counter the attacks against models whose weights are unknown (black box attack) viz. (Xu et al., 2018). However, adversarial attacks that are generated by these sophisticated techniques are a rare occurrence in practice, and the resulting perturbations tend to be fairly contrived. One might question the validity of such proposed attack techniques in real-world scenarios.

In security-critical domains such as facial recognition and self-driving, it is more likely to encounter a noisy input due to structural deformities of the data. A few other works have recognised and addressed these deformities ( (Kanbak et al., 2018)), (Xiao et al., 2018)), (Engstrom et al., 2017)) and analysed the behaviour of deep learning in such scenarios. However, we believe that there is a scope for a more rigorous and thorough analysis in this domain. In this paper, we study the effect of transformations such as rotation, scaling, exposure etc., on a few deep learning models.

Prior works, such as (Pei et al., 2017), evaluate the robustness of DNNs under structural perturbations. A few defences such as (Engstrom et al., 2017) and (Kanbak et al., 2018) have been proposed in the literature. However, to the best of our knowledge, this paper is one of the first to introduce a defence technique that covers a vast range for each perturbation.

[1] Department of Computer Science, IIT Hyderabad. Correspondence to: Uttaran Sinha <cs17mtech11003@iith.ac.in>, Saurabh Joshi <sbjoshi@iith.ac.in>, Vineeth N Balasubramanian <vineethnb@iith.ac.in>.

Presented at the ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2019 by the author(s).

Our key contributions are as follows:

- We propose a core-set technique to increase the robustness of different networks under single and combination of perturbations. To the best of our knowledge, this is the first work that analyses the perturbations over a very wide range.
- We analyse the effect of adversarial training (under structural perturbations) on the decision boundary of a model.

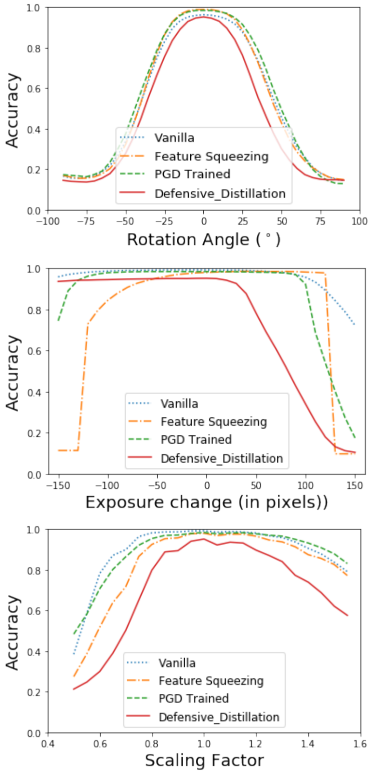

*Figure 1. Performance of statistical defences against structural perturbations on MNIST dataset.**Vanilla** Model refers to a model trained on natural data. Defence models were trained with default parameters. For **Feature Squeezing**(Xu et al., 2018), a filter size of 3 was used for smoothing. In case of **Projected Gradient Descend Attack** (Madry et al., 2018), the $L\infty$ norm was set to 0.3, and it was trained on generated adversarial examples.The model using **defensive distillation**(Papernot et al., 2016) had the temperature set to 10.*

## 2. Related Work

We discuss the related works under different categories. To summarise, our work differs in three aspects from existing works, *i)* We approach robustness as an expansion of model boundaries instead of considering robustness on a single input, *ii)* We consider any perturbation over an extensive range and *iii)* We utilize a core-set approach to this problem which we believe is unused in this domain.

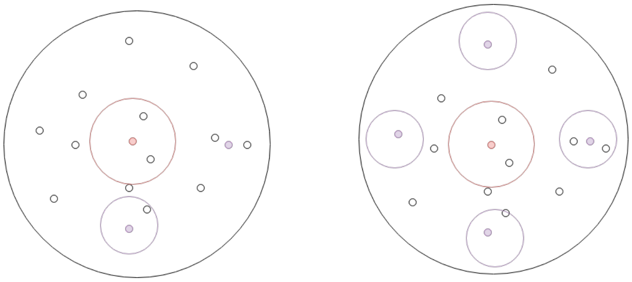

*Figure 2. The Core-Set approach to covering all the data points in a given perturbed dataset $D$. The figure to the left visualises the asymmetric approach and the right-hand figure signifies the symmetric approach. We assume that the red point is the initial centre-point and the purple point in the next centre-point. In the symmetric case, we cover the auxiliary points via the notion of symmetry.*

In the context of computer vision, an adversarial example for a given input image $x$ and a classifier $C$ in an image $x'$ that is ”visually similar” to $x$ under human perception. This notion of similarity can be defined in multiple ways. For example, $x'$ can be defined in terms of $x$ as $x = ||\mathrm{x}+\epsilon\ ||_p$, where $\epsilon$ is the amount of noise under some norm distance $p \in (0,\infty\ )$. We observe that the noise need not be additive. As an alternate approach, $x'$ can be a linear transformation of x i.e. $x' = Ax + b$. Our initial goal is to generate $x'$ strong enough to be considered adversarial. In the context of pixel-wise perturbations to generate an adversarial image, the most successful approaches have been gradient-based optimisation methods ((Szegedy et al., 2014), (Goodfellow et al., 2015), (Carlini & Wagner, 2017)). However, since it has been proposed that statistical and structural defences are mutually orthogonal ((Engstrom et al., 2017)) we discuss only the relevant prior works on this domain. We tested this conjecture ourselves, and the results are as in Fig 1.

The vulnerability of deep learning under rigid geometric perturbations has been discussed in ((Lenc & Vedaldi, 2015), (Goodfellow et al., 2009)), we will build on the work to formulate our defence technique. (Pei et al., 2017) further analyse the robustness of state-of-the models under structural perturbations via attacks based on neuron coverage but does not discuss any ideas for defence. (Kanbak et al., 2018) propose adversarial training as a defence against perturbations under the metric of geodesic distance, whereas, we consider distance in terms of angles, scaling factor etc. This work also captures the effect of exposure change on an image which is not discussed in the work mentioned above. We also discuss how adversarial training effects the model boundary. (Engstrom et al., 2017) discuss the performance of deep models under translation and rotation.

They propose random and worst case adversarial training as a probable defence. We build on this work and consider four more sets of perturbations, viz, scaling, exposure, shear and perspective in addition to rotation and translation. Our focus is on delivering robustness over a wider range for each of these transformations. We also compare our model to randomly trained and worst-case trained models (Section 4). The work of (Xiao et al., 2018) is not directly related to this work since we consider structural perturbation without any per-pixel based changes in the input image. We believe that flow distance does not generate pure structurally perturbed images and hence is beyond the scope of this work.

We also believe that no prior work discusses robustness on perturbations over a vast range such as ours. Training a model over such an array can prove challenging since the search space would be too large to be computationally feasible. Our work proposes an efficient, data and model agnostic approach to attain such a goal.

## 3. Methodology

The goal of this paper is to guarantee overall accuracy on the train and test data above a certain threshold, given any model and dataset over a wide range of each perturbation via adversarial training. We draw inspiration from centre-point placement of the min-max facility location problem ( (Wolf, 2011) ). The classic problem is to choose b centre points such that the largest distance between a data point and its nearest centre is minimized. (Equation 1)

$$\underset{s^1:|s^1|\leq b}{min} \; \underset{i}{max} \; \underset{j\in s^1\cup s^0}{min} \triangle(x_i, x_j) \tag{1}$$

Consider a set $\Theta$ which stores all possible perturbations we allow in our experiments. For example, if we consider rotation and scaling as the only possible perturbations with the maximum allowed perturbation as R and L respectively, $\Theta$ will be a Cartesian product of all such perturbations. We also define step size $x$ as the minimum difference between two consecutive perturbations. For example, in case of rotation, we define $x$ = 1°.

We then perturb the given dataset by each element of the set $\Theta$ and obtain the perturbed dataset, $D$. We define a data point $d$ as an element of the perturbed training set and $arg(d)$ as the element of $\Theta$, i.e., the specific perturbation that generated $d$. The centre point $b$ is the perturbed data point on which the model is trained. The set of such centre points is denoted by $B$. The **distance** between a centre point and a data point is measured in terms of the difference in the accuracy of the model at those points. We also need to define the notion of **coverage**. We say a data point $d$ is covered by a centre point $b$ if the difference in accuracy between $b$ and $d$ is less than equal to our **tolerance**, $\epsilon$. The approach is summarised by Algorithm 1.

In contrast with the classical min-max objective, our goal is to minimise $\mid B \mid$ that guarantees coverage on $D$. For simplicity, we assume that the distance between a given data point to any support is equal, i.e. model accuracy remains unchanged on all centre-points. This renders the inner minimisation problem of Equation 1 trivial. We overload the **difference** operator between a centre-point and a data-point to denote their difference in accuracy on a given model.

Our objective is to cover all $d \in D$ such the the cardinality of $B$ is minimised, as follows:

$$\underset{size(B)}{minimise} \mid b - d \mid \leq \epsilon, \forall d \in D \, where \, b \in B \, and \, B \subseteq D \tag{2}$$

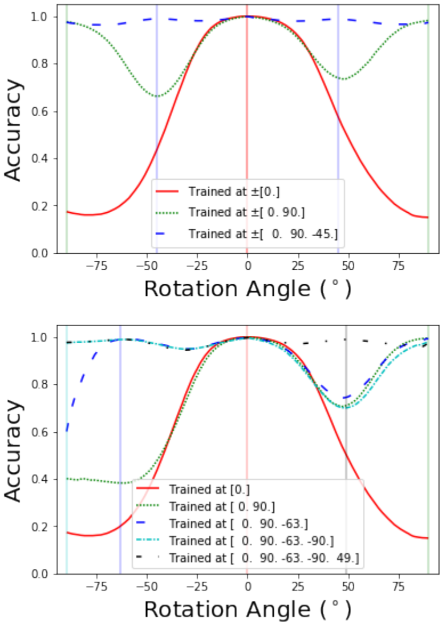

*Figure 3. This figure shows a comparison between symmetric and asymmetric training phases. The lines represent the accuracy of the training set. The colour of vertical bars matches the centre-point that it was trained on. For example, the green curve on the top represents the accuracy of the model after being trained on centre-points 0 and* ±90*. The figure on top is a result of symmetric training on rotation. Hence, the accuracy graphs per iteration are symmetrical. Including the 0 centre-point, the number of centre-points required for symmetric training is five (#dimensions * (points - 1) + 1) and for asymmetric is five.*

We solve this NP-hard ((Hsu & Nemhauser, 1979)) problem (Equation 2) via a greedy approach. We assume that the initial centre-point is located at natural data. Without this assumption, we cannot define the notion of coverage. We choose the next centre at the point which is farthest, i.e. has the lowest accuracy and so on. Since we assume that all centre points are equivalent, the point at which the minimum accuracy is observed will be the next centre point. We also assume that once a data point is covered, it cannot be uncovered. Therefore, it suffices to search the pool of uncovered points for our next centre point. We continue to all points in $D$ are covered. Since the pool of uncovered points in a strictly decreasing set (See Appendix A), the algorithm is guaranteed to terminate. When considering the greedy approach, our optimisation problem reduces as follows:

$$\underset{size(B)}{minimise}\ maximise \mid b{-}d \mid,\ \forall d \in D\ where\ b \in B\ and\ B \subseteq D \tag{3}$$

A similar approached has been proven (Sener & Savarese, 2018) and hence the worst case performance of our algorithm is $\leq$ 2*OPT, where OPT is the solution to equation 1.

Since structural perturbations possess an inherent symmetry, we propose two different approaches to the centre-point placement. For example, in case of rotation, if a centre-point was to be placed at an angle $\theta$, in one approach, we also place another centre-point at -$\theta$ simultaneously. In the other approach, we do not place any such auxiliary centre-points. We term these approaches **symmetric** and **asymmetric** respectively (Figure 3). In Section 4, we discuss the outcomes and analysis of these approaches and their effect on robustness and training time. (See Figure 7)

## 4. Results

Since statistical and structural defences are orthogonal to each other ((Engstrom et al., 2017), also see Figure 1), we decide not to compare our work to existing statistical defences. We test our algorithm on the problem of classification on two different datasets. We performed experiments on MNIST (LeCun & Cortes, 2005) image classification and Cifar10 dataset image classification. We compare our method to *i)* **Random**: Randomly sampling k center-points from $D$, *ii)* **Random worst case**: Choosing k centre-points from a set of random points where accuracy is lowest.

During all of our experiments, we keep the number of centre-points(k) of random sampling methods equal to the cardinality of the set $B$ in Algorithm 1 per experiment.

**Results on MNIST dataset** We use the architecture as suggested in Keras official repository. For each training task, we train for ten epochs irrespective of the size of dataset. For the baseline model, we have achieved 99% training accuracy and 97% test set accuracy. For all the experiments, $\epsilon$ has been fixed to 0.1. Please note that as the number of centre-points increases, the amount of training data increases and hence, the model may not achieve baseline accuracy. This is especially noticed in the CIFAR-10 dataset under combination of perturbations. We hypothesise this is due to lack of model capacity to fit the increasing set of training data. (Figure 4) As a natural extension, we test our approach on combination of these perturbations (Figure 5)

**Algorithm 1** The min-max search and train algorithm

```
input  : Model ($M$), Perturbed Dataset ($D$), $x = stepsize$,
         $\epsilon$ = tolerance, $\theta$ = accuracy of model on natural
         data
output : Model ($M$), Set of Centre-points ($B$)
initialization $B = \phi$                    //Set of Centre-points
while True do
    for $d$ in $D$ do
        $S = \phi$                 //The set of uncovered points
        if $M(d) \leq \theta - \epsilon$ then
            $S = S \cup M(d)$                  //$d$ is uncovered
        end
        if $S == \phi$ then
            BREAK                        //All points are covered
        else
            $B = B \cup argmin(S)$    //Find the perturbation
             which has the lowest accuracy
        end
    end
    $M$ = Initialize and train new model on $B$
end
```

**Results on Cifar10 dataset**

Similar to MNIST dataset, we test and compare six perturbations with the four methods were have discussed. We use the Resnet-50 (He et al., 2016) architecture. (Figure 6)

## 5. Concluding Remarks and Future Work

In this work, we have proposed a core-set approach to adversarial training to defend structural attacks in DNNs. We have shown that this method works in multiple datasets and model architectures and also handles a combination of perturbations with ease. To summarise, we propose the notion of coverage and empirically show that selective adversarial training is more robust and reliable than random or worst-case random sampling. We believe that this work will be useful in multiple domains such as automated driving and facial recognition. As a potential improvement, we will look at incremental learning to improve the performance of our algorithm even further. There is also a possibility that we can manually train persistent mis-classified inputs to achieve a higher level of robustness.

## References

Akhtar, N. and Mian, A. S. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

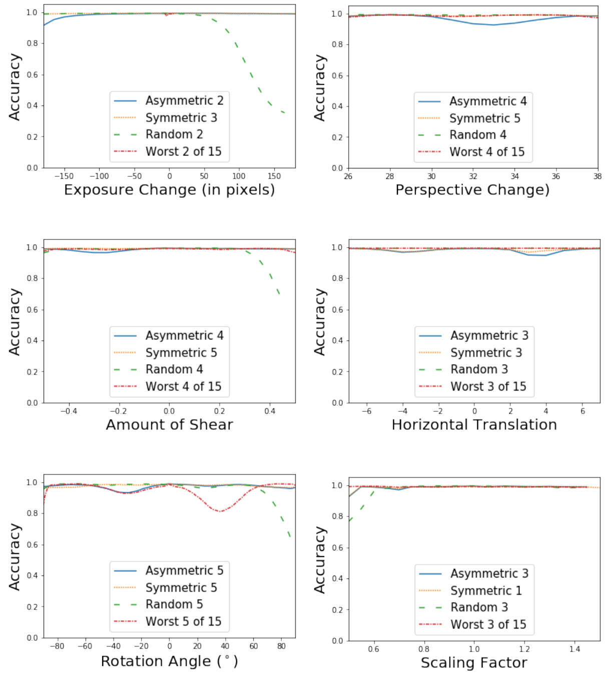

*Figure 4. The plots show the performance of four different approaches on the test data of MNIST dataset under specific perturbations. We notice that our approaches always guarantee robustness whereas the random method is not reliable. The number adjacent to each legend refers to the number of centre-points used.*

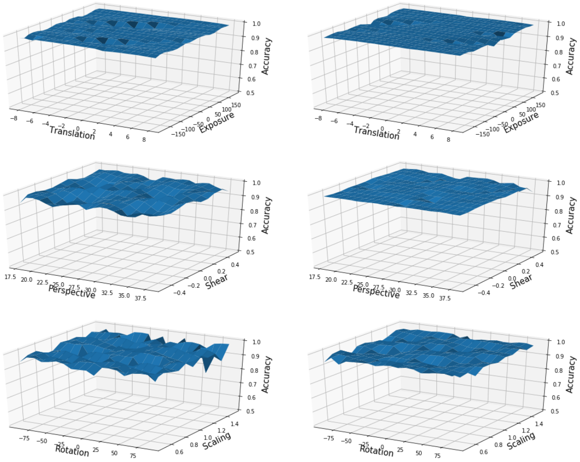

*Figure 5. The plots show the performance combination of perturbations on test data of MNIST dataset on our algorithm. The left side figures depict performance on asymmetric centre-point allocation (required 6,13,16 centre-points top to bottom) and right side figures on symmetric (required 7,19,19 centre-points top to bottom)*

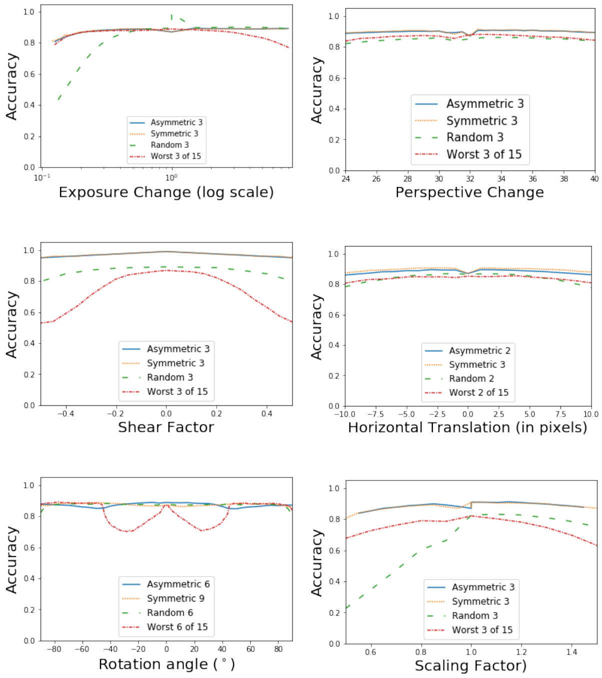

*Figure 6. The plots show the performance of four different approaches on the test data of CIFAR-10 dataset under specific perturbations. As the results show, the unreliability of random sampling is even more apparent when the dataset is complex.*

Carlini, N. and Wagner, D. A. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

Engstrom, L., Tsipras, D., Schmidt, L., and Madry, A. A rotation and a translation suffice: Fooling cnns with simple transformations. 12 2017.

Goodfellow, I. J., Le, Q. V., Saxe, A. M., Lee, H., and Ng, A. Y. Measuring invariances in deep networks. In *NIPS*, 2009.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hsu, W.-L. and Nemhauser, G. L. Easy and hard bottleneck location problems. *Discrete Applied Mathematics*, 1: 209–215, 1979.

Huang, X., Kwiatkowska, M. Z., Wang, S., and Wu, M. Safety verification of deep neural networks. In *CAV*, 2017.

Kanbak, C., Moosavi-Dezfooli, S.-M., and Frossard, P. Geometric robustness of deep networks: Analysis and improvement. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4441–4449, 2018.

LeCun, Y. and Cortes, C. The mnist database of handwritten digits. 2005.

Lenc, K. and Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 86–94, 2017.

Nguyen, A. M., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015.

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

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *AsiaCCS*, 2017.

Pei, K., Cao, Y., Yang, J., and Jana, S. Deepxplore: Automated whitebox testing of deep learning systems. In *SOSP*, 2017.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *ICLR 2018*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

Wolf, G. W. Facility location: concepts, models, algorithms and case studies. series: Contributions to management science. *International Journal of Geographical Information Science*, 25:331–333, 2011.

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. X. Spatially transformed adversarial examples. *CoRR*, abs/1801.02612, 2018.

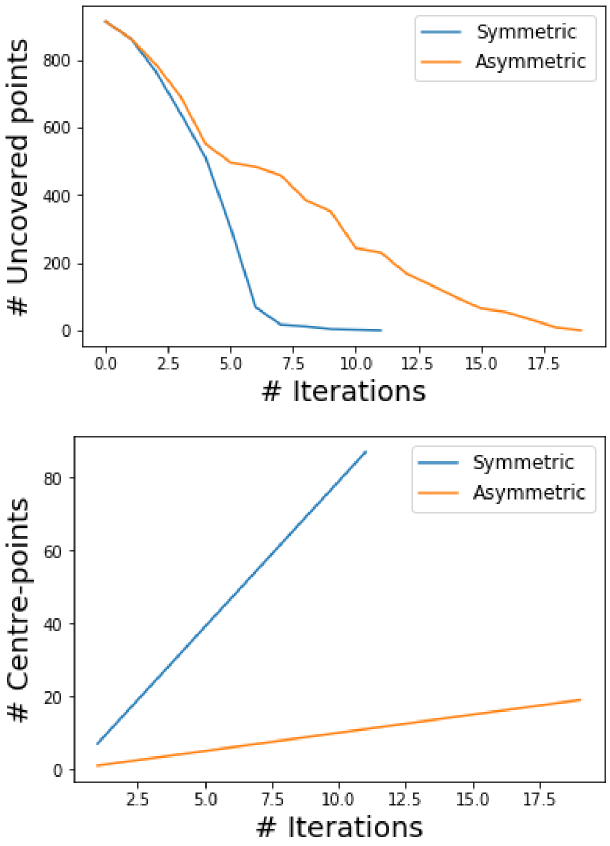

*Figure 7. We present a comparison between the two training modes on MNIST dataset under combined perturbation of rotation,scaling and exposure. It is clear that even when the search space is high, our algorithm converges. It is clear that the symmetric approach converges faster albeit at the cost of higher data requirement.*

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *CoRR*, abs/1704.01155, 2018.

## A. Appendix

### A.1. Convergence rates

We also compare the convergence rates of symmetric and asymmetric approaches to training. As we have observed so far, on single and pair of perturbations both approaches work almost identically. But, as we increase the combination of perturbations, the difference in performance becomes clear. For example, we consider combination of rotation, scaling and exposure on the MNIST dataset in Figure 7.

### A.2. Analysis of decision boundary of the model

We discuss the effect of adversarial training on the decision boundary of the model in brief. We can visualize the model

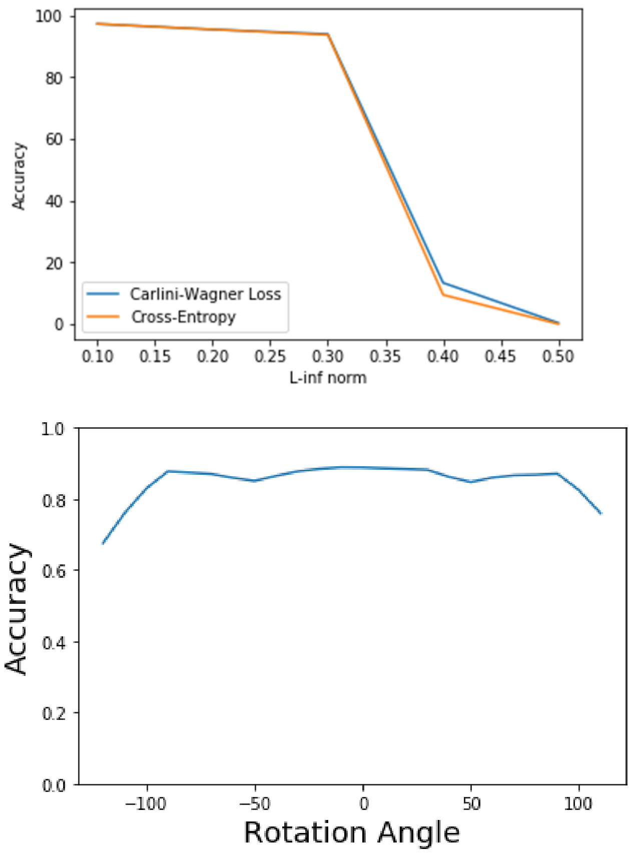

*Figure 8. We compare the effect of adversarial training. The figure on the left is training on L $\infty$ norm upto 0.3 (Madry et al., 2018) and the figure on the right is trained to be robust against rotation upto 90 °. Both models are tested beyond their zone of comfort. We can see that the former model fails when the perturbation is outside of allowed range. But out model shows considerable robustness to out of distribution data.*

boundary as in Figure 2. We can assume that when our algorithm terminates, we have covered the entire search space within our bounds. However, it is also important to discuss the performance of the model outside of our training search space. (see Figure 8)