# Manifold Analysis of Linear Separability in Deep Learning Models under Adversarial Perturbations

## Abstract

Adversarial attacks play a key role in making robust neural networks for various computer vision tasks. To make a proper mechanism for defending against adversarial attacks, understanding the neural network's behaviour is very important. Recent studies show that manifold analysis has been capable of understanding the neural network performance by evaluating linear separability among the classes by analysing the number of activations of each layer of a convolutional neural network (CNN) architecture. The purpose of this study is to evaluate the linear separability of manifolds while we are changing the input space by applying a few adversarial perturbations to the images. Therefore, the perturbed images were created by using Gaussian noise levels as adversarial attacks. The baseline analysis was done for the ResNet18 pre-trained model with the CIFAR10 dataset and obtained mean-field theoretic manifold analysis (MFTMA) measurements for the evaluations. The analysis results show that the increase in Gaussian noise level impacts the linear separability of the object manifolds. reducing manifold capacity and increase of the manifold dimension will lead to making model predictions as misclassification. As a future research direction, this approach can be extended to create a white-box adversarial attack framework by leveraging MFTMA analysis to make guided perturbations and improve the robustness of the deep neural network models. The code implementation is available in `https://pushpikaprasad.github.io/MFTMA_with_Gaussian_Noise/`.

## 1 Introduction

With the rapid development of deep learning networks in computer vision, adversarial attacks play a major challenge when improving the accuracy of deep learning models. [1]. Adversarial attacks introduce small perturbations to input images that cause models to misclassify them, even though humans can still easily recognize the correct label of the manipulated images. This encourages improving the robustness of deep neural networks through adversarial defense mechanisms [2]. Therefore, it is very important to understand the changes in deep learning models' behaviour while applying adversarial attacks for the input image space to improve the robustness of the classifier.

In this study, we use replica MFTMA analysis to identify the linear separability of object manifolds using the underlying geometric properties, which describe manifold capacity, dimension, radius and correlations [3] [4]. Each manifold represents a set of points as object manifolds that consist of identical labels, and the manifold capacity measures the linear separability of object categories per feature dimension. In adversarial attacks, input images are changed using adversarial perturbations without changing their labels. This gives some changes in the manifold objects, which will lead to a problem that shows how information is untangled in a neural network when performing image classification.

However, MFTMA analysis can be used to understand the internal behaviour of the neural networks in each layer, and this leads to a hypothesis that MFTMA analysis can be used as a tool to perform white-box adversarial attacks while applying adversarial corruptions to a large number of images

Preprint. Under review.

to make distribution shifts. In this study, we use Gaussian Noise to create perturbed input images, and the analysis was conducted on the CIFAR-10 dataset with the ResNet18 pre-trained model. This analysis mainly considers the changes in MFTMA analysis measurements while we change the Gaussian noise level of the input images.

## 2 Literature Review

In recent studies, several techniques have been implemented to perform black-box and white-box adversarial attacks [5]. In white-box attacks, full architectural information, including weights of the parameters of the neural network, was considered, while black-box attacks have limited or no knowledge about the neural networks.

Szegedy et al. [6] are the very first to demonstrate adversarial attacks by adding small perturbations to the clean images that fool a neural network, even though those images appear as original images to human vision. They highlight that identical perturbations may cause a neural network, which was trained on a different subset of the dataset, to misclassify the same input instance. This leads to the concept of improving the robustness of the neural network by adversarial training. Goodfellow et al. [7] introduced a method called the 'Fast Gradient Sign Method' (FGSM), which computes the gradient of the cost function of the model. The FGSM is a white-box adversarial attack method which helps to improve the robustness of deep neural networks during model training. However, FGSM is less effective when the model is highly nonlinear, and it may lead to more modifications, such as the Basic Iterative method (BIM) [8] or the Projected Gradient Descent (PGD) [1].

Papernot et al. [9] introduced a method called Jacobian-based Saliency Map Attack (JSMA), which basically changes only a few pixels in the image, restricting the $l_0$-norm of the perturbations to make misclassification of the model. This method alters the pixels of input images iteratively by monitoring a saliency map, which is created using the gradients of the outputs of the network layers. Moosavi-Dezfooli et al. [10] also considered a similar concept with smaller perturbations. They introduced a 'DeepFool' method, which iteratively perturbs a clean image with minimal changes until the image gives a misclassification.

According to the literature done so far, none of the methods used MFTMA analysis [11] with adversarial perturbed images to perform an adversarial attack to make a neural network prediction misclassified. However, the MFTMA analysis was used to understand the behaviour of the neural network in other tasks, such as identifying how classification capacity improves along the hierarchies of deep neural networks with different architectures [3], examining the structure of when and where memorisation occurs in a deep neural network [4], analysing information that is untangled within neural networks trained to recognise speech [12], to analyse language representation from large-scale contextual embedding models [13], etc.

Based on the literature, this is the first time MFTMA analysis has been used in an adversarial attack. With this study, there are open research directions to investigate in performing white-box adversarial attacks by using MFTMA analysis and improving the robustness of the neural network model by applying guided data augmentation methods as a supplement for the adversarial training, while evaluating and certifying the robustness of the classifier [2] with manifold analysis.

## 3 Experimental setup

### 3.1 Adversarial Examples

In this study, adversarial perturbations were performed by using Gaussian Noise $\mathcal{N}(0, \sigma^2)$ to create perturbed images of the testing dataset. Here, we use various Gaussian noise levels (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, and 1.0), which are controlled by applying a value $\sigma$ between 0 and 1.0. If $\sigma = 0$, that means we do not apply Gaussian corruptions to any image in the input space.

### 3.2 Dataset and Model

We applied Gaussian Noise to the CIFAR-10 [14] dataset's testing images. The dataset has been divided into training (45000 images), validation (5000 images) and testing (10000 images) datasets after applying data preprocessing steps, which include random crop to have $32 \times 32$ pixels in each image, colour jitter (with 0.2 of brightness, 0.2 of contrast, and 0.2 of saturation), random horizontal flip, and random rotation. The classification was done using a ResNet18 pre-trained model. The input dimension of the first layer was changed to fit the input images' size. The model was trained until it reached more than 90% accuracy before applying adversarial perturbed images for the manifold analysis.

### 3.3 The mean-field theoretic manifold analysis (MFTMA)

The analysis was done by evaluating the manifold capacity and its geometric properties, such as manifold dimension, radius, and center correlation, to understand the linear separability of the object manifolds which contain points with identical labels [11][2][12][4]. For given $P$ object manifolds in $N$ feature dimensions, Manifold capacity, $\alpha_M = P/N$, refers to the critical load, defined by the number of object manifolds where most of the manifold dichotomies can be separated by a linear hyperplane. When $\alpha$ is small, it is easy to find a separating hyperplane for a random linear hyperplane which separates manifolds, where we can see a small number of object manifolds in the high-dimensional feature space. When $\alpha$ is large, a high number of manifolds are squeezed into a low-dimensional feature space, rendering the manifolds highly inseparable.

The MFTMA framework theoretically considers manifold capacity and evaluates other geometric properties of object manifolds as well. Then, it gives four quantities:

- **Manifold Capacity** $(\alpha_M = P/N)$ estimates the critical load mentioned above, using the replica mean field theory by considering the manifold's dimension, radius, and their center correlation introduced by Chung et al. (2018) [11].
- **Manifold Dimension** $(D_M)$ captures the dimensions of an object manifold realised by the set of anchor points from the guiding Gaussian vectors that determine the optimal hyperplane separating the binary dichotomy. A small dimension implies that the anchor points occupy a low-dimensional space.
- **Manifold Radius** $(R_M)$ is the average distance between the manifold centre and the anchor points, which captures the manifold size relevant for linear separability. A small manifold radius implies that the set of examples that determine the decision boundary is tightly grouped.
- **Center Correlations** $(\rho_{center})$ measure the average of the absolute values of the pairwise correlations between manifold centroids. A small correlation implies that, on average, manifolds lie along different directions in feature space.

In this study, the four quantities of MFTMA analysis mentioned above were measured with various Gaussian noise levels applied in the input images of the testing dataset.

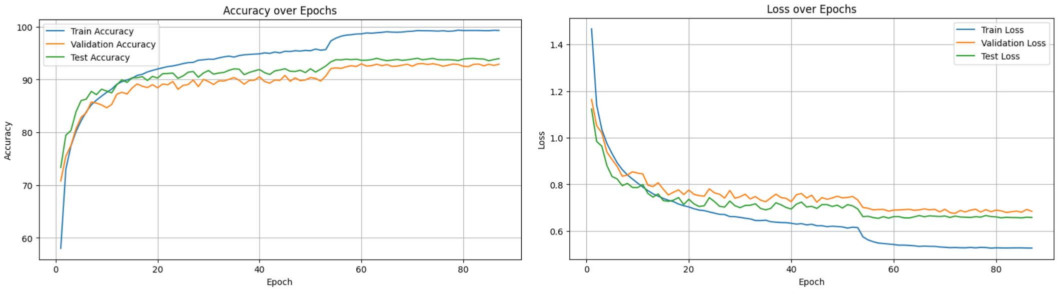
### 3.4 Manifold data

We define the manifold data for the MFTMA analysis by considering the activation vector in each layer of the convolutional and dense layers of the trained model. Given a neural network $N(x)$, we extract $N_l(x)$, the activation vector at the layer $l$ generated on input $x$. These activations, which are generated by inputs of the same class, are used to define a manifold. This gives $P$ manifolds, and one manifold contains data related to a single class.

## 4 Results and Discussion

In this study, first, we trained the ResNet18 pre-trained model with the CIFAR-10 dataset, until we reached more than 90% testing accuracy. Figure 1 shows the accuracy values and loss values changed during the model training process.

**Apply Gaussian Noise to the input images** To create adversarial perturbations of the original images, we applied Gaussian Noise $\mathcal{N}(0, \sigma^2)$ at various noise levels $\sigma$ (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, and 1.0) to the input images of the testing dataset to create 11 testing datasets with the original testing dataset. Those datasets are separately used to test the model and



(a) (b)

Figure 1: **Prepare RestNet18 pre-trained model:** (a) Accuracies of the training, validation and testing sets are calculated during the training of the ResNet18 model. The final testing accuracy was 93.96%. (b) Losses of the training, validation and testing sets calculated during the training of the ResNet18 model.

obtain the testing accuracy changes to identify whether the model is predicting misclassifications. Figure 2 indicates an example of perturbed images under the "car" class of the testing dataset.

When the $\sigma$ value is small (i.e. 0.0001 to 0.05), the model is predicting the correct class label as "car". Then it was misclassified as "truck" (for $\sigma$ = 0.1), "frog" (for $\sigma$ = 0.25 and 0.5), and "bird" (for $\sigma$ = 0.75 and 1.0). Until $\sigma$ = 0.5, the human eye can identify the correct class while the model is predicting the misclassifications, which means that the perturbed images with Gaussian noise can be used to perform adversarial attacks as a black-box approach, while we do not have enough knowledge of the model's internal behaviour. Figure 3 shows how the model is performed in the testing dataset while we apply Gaussian noise to the input images.

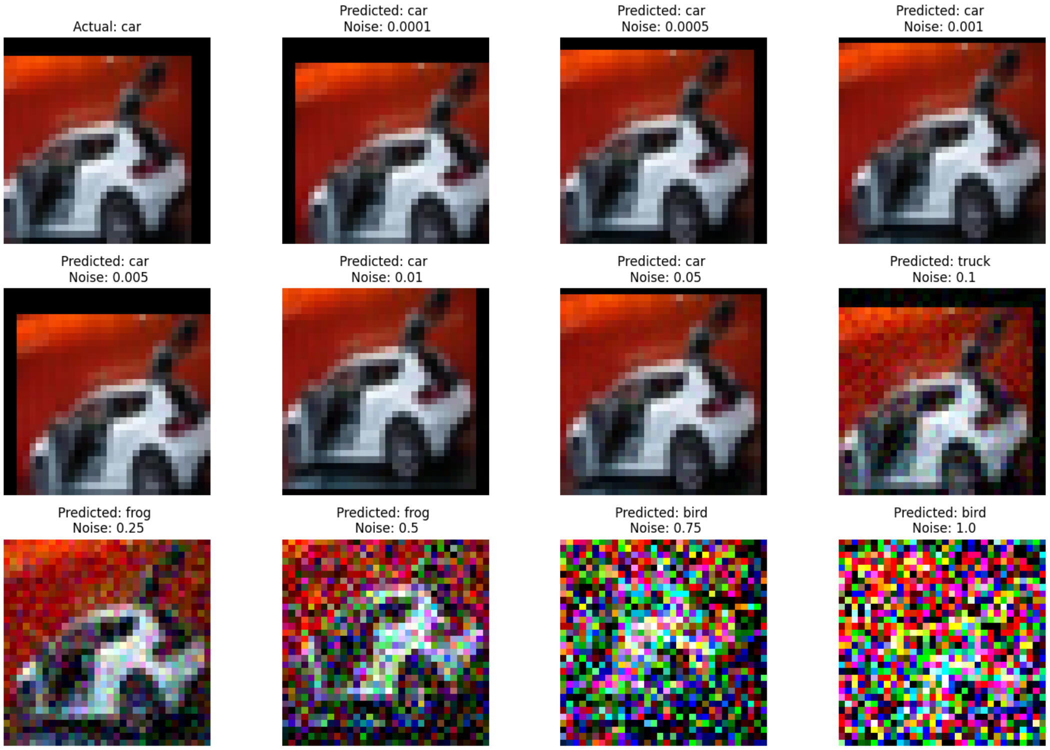


Figure 2: **Apply Gaussian noise:** An example of input images (under "car" class) shown here after applying the Gaussian noise to generate perturbed images to fool the model prediction. The very first image is the original image, and the rest of the images are obtained after applying Gaussian noise.

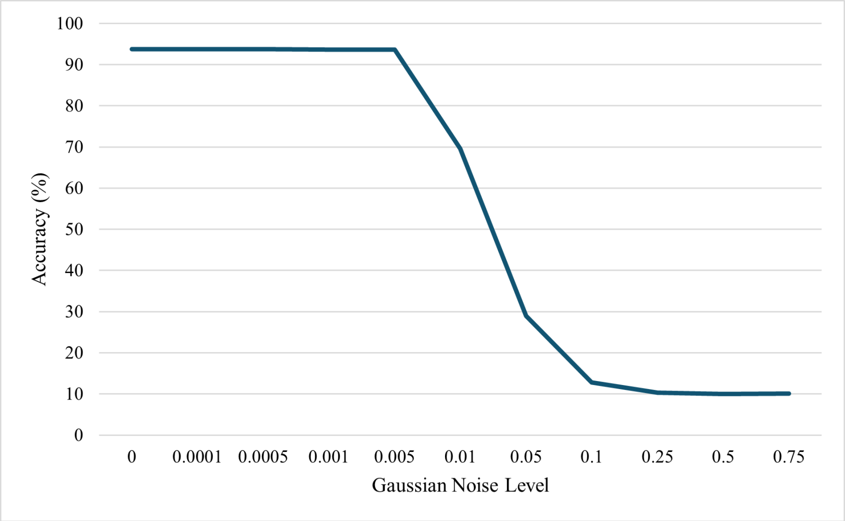
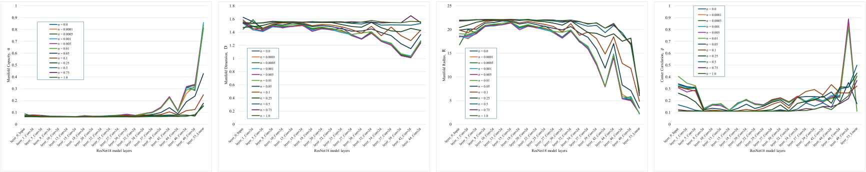


Figure 3: **Testing Accuracy vs Gaussian Noise levels:** With the changes of Gaussian noise levels, the testing accuracy is decreasing as it shows misclassification for the perturbed input images.

**MFTMA analysis with perturbed input images** After analysing the model performance for the perturbed testing datasets, we can use the activations of each layer of the model to examine the linear separability of the manifold objects which created for each class while applying each testing dataset to check the model performance.



(a) (b) (c) (d)

Figure 4: (a) The manifold capacity of each layer is calculated for the different Gaussian Noise levels. With the increase of the variance of Gaussian noise, the last linear layer (the output layer of the ResNet18) shows inseparability of the manifold objects after $\sigma = 0.05$. (b) The dimension of the manifolds increased in very large values for the final layer of the model after applying a Gaussian noise level of $\sigma = 0.05$. (c) The manifold size is increasing after $\sigma = 0.05$, which indicates there could be linear inseparability and the model will be prone to predict misclassifications for the input data. (d) With the increase of the Gaussian noise value, after $\sigma = 0.05$, the correlation of pairwise centers of manifolds is getting decrease, which causes the manifolds to lie along different directions of the feature space.

By analysing the manifold capacity and other geometrical properties, such as manifold dimension, radius, and center correlation (Figure 4), we can observe that the increase of the Gaussian noise level of perturbed input images will give difficulty to the linear separability of the manifolds in the feature space. When $\sigma \geq 0.05$, the manifold capacity is getting low and the manifold dimension is getting high, which implies that the model is prone to predict misclassification after the Gaussian noise level of 0.05 but the human can easily predict the correct class while Gaussian noise up to $\sigma$ = 0.25 or 0.5. Therefore, the findings clearly indicate that the model can be enhanced to improve its robustness against adversarial perturbations in the input images.

## 5 Conclusion

In this study, we investigate the impact of adversarial perturbations on deep neural networks using MFTMA analysis as a novel approach. MFTMA analysis was used as a tool for understanding and guiding adversarial attacks made by in deep learning model. The findings highlight how adversarial perturbations alter the geometric properties of object manifolds, revealing the internal behaviour of neural networks by considering the linear separability of the object manifolds of the feature space.

The results show that the Gaussian noise can be added as an adversarial attack to make the ResNet18 pre-trained model misclassify the correct class while humans can predict the correct class. With the MFTMA analysis, we can analyse the model behaviour along with the Gaussian noise changes which give future research direction to implement white-box adversarial attack techniques with the guided data augmentation to improve the robustness of the model.

## References

[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[2] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. pages 1310–1320, 2019.

[3] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746, 2020.

[4] Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. *arXiv preprint arXiv:2105.14602*, 2021.

[5] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR*, abs/1801.00553, 2018.

[6] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2014.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[8] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. pages 99–112, 2018.

[9] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. pages 372–387, 2016.

[10] 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*, pages 2574–2582, 2016.

[11] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3), July 2018.

[12] Cory Stephenson, Jenelle Feather, Suchismita Padhy, Oguz Elibol, Hanlin Tang, Josh McDermott, and SueYeon Chung. Untangling in invariant speech recognition. *Advances in neural information processing systems*, 32, 2019.

[13] Jonathan Mamou, Hang Le, Miguel Del Rio, Cory Stephenson, Hanlin Tang, Yoon Kim, and SueYeon Chung. Emergence of separable manifolds in deep language representations. *arXiv preprint arXiv:2006.01095*, 2020.

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.