A Quantitative Comparison of Image Classification models under Adversarial attacks and defenses

Sarthak Kathuria, Kartikeya Khullar, Nishant Chahar,
Prince Gupta and Preeti Kaur

June 28, 2021

A Quantitative Comparison of Image Classification models under Adversarial attacks and defenses

Sarthak Kathuria[1], Kartikeya Khullar[1], Nishant Chahar[1], Prince Gupta[1], Dr. Preeti Kaur[2]

Abstract— **In this paper, we present a comparison of the performance of two state-of-the-art model architectures under Adversarial attacks. These are attacks that are designed to trick trained machine learning models. The models compared in this paper perform commendably on the popular image classification dataset CIFAR-10. To generate these adversarial examples for the attack, we are using two strategies, the first one being a very popular attack based on the L_∞ metric. And the other one is a relatively new technique that covers fundamentally different types of adversarial examples generated using the Wasserstein distance. We will also be applying two adversarial defenses, preprocessing the input and adversarial training. The comparative results show that even these new state-of-the-art techniques are susceptible to adversarial attacks. Also, we concluded that more studies on adversarial defences are required and current defence techniques must be adopted in real-world applications.**

I. INTRODUCTION

The ubiquity of Computer Vision has brought trained classifiers into the center of security-critical systems. Computer Vision can find its application in multiple domains such as Facial Recognition. This technology can be used to identify violent, anti-social elements during mass events such as protests, rallies, etc. Facial recognition is widely used in smartphones as a security mechanism to prevent unauthorized access. Due to their prominent presence in such security-critical systems, a lot of studies have been conducted that focuses on existence of adversarial examples that trick these machines into outputting wrong predictions. These examples are basically datapoints that have been perturbed to trick these trained classifiers.

This has resulted in extensive research in the security of Machine Learning models. In particular, resistance against these specific examples is becoming a crucial design goal for Computer vision systems. Recent works [8] show that Machine Learning Models are vulnerable to adversary based attacks and can give an incorrect output as a result of them. The main issue with Computer Vision models is that minute changes to the input image that might even be invisible to the human eye, can confuse a state-of-the-art neural network architecture with high probability on normal images as shown in the paper [11] .The changes made to images are invisible to the naked eye but vastly distort the results generated by the algorithm.

There have been studies related to adversarial attacks and defenses on models including Resnet from Artificial Intelligence security point of view. Over the years there have been advances in the field of computer vision and adversarial defenses. The models are becoming more and more robust to these examples. In this paper, we attempt adversarial attacks on the relatively newer models that perform well on classic Machine learning dataset CIFAR-10 [7]. L_p norm based attacks have been used extensively on Machine Learning models. They range from L_2 to L_∞. The L_2 norm is related to the Euclidean distance and L_∞ is related to the magnitude of the largest vector. On the other hand Wasserstein distance takes geometry in pixel space into account and has recently risen as a compelling alternative to the L_p metric in adversarial attacks on current state-of-the-art techniques. The model architectures that we use are: Vision Transformer and Wide Resnet using SAM. Vision Transformer [3] is a model for image classification that employs a Transformer-like architecture over patches of the image. The other model that we use is WideResnet [16] using SAM. SAM [4] stands for Sharpness-Aware Minimization and it basically works by minimizing both the loss value and its sharpness together. It does this by seeking neighborhood parameters that have a uniformly low loss. It even provides robustness to label noisy samples which is nearly on the same level as some state-of-the-art models that have been trained for learning from noisy labels. Thus, this architecture is a good example to study under the adversarial attacks that we perform.

This paper is organised in 5 sections. In the first section we have explained the importance of computer vision models and how they can be tricked into giving wrong predictions, which results into the need of conducting studies to defend against such threats. We have also provided a brief about what state-of-the-art model architectures we will use and the Adversarial Attack techniques we will employ. In the second section we have given a brief about the research papers and techniques we have explored as a part of our research. The theory section presents the details about the model architectures as well as the Adversarial Attack and Defense techniques, we have used in our experiments. In the Experimental Setup section we have given information about the experiments we conducted as a part of our research and the hardware that was used. In the fourth section we have summarised our findings from the experiments. Finally we conclude the outcomes of this research.

II. RELATED WORK

We referred to the following papers extensively for our experiments. The paper by Goodfellow et al [5] introduced the concept of adversarial training. Though the adversary used in the paper was quite weak. [11] studies a encapsulates a lot of literature on adversarial attacks and defences. This paper explores the adversarial robustness of neural networks through the lens of robust optimization. They show that adversarial training optimizes the saddle point problem, which is the min-max problem that they use in this paper to capture the notion of security. In [13], general adversarial attacks on Machine Learning models were discussed. This paper presented a new model for adversarial attacks which uses Wasserstein distance. To generate Wasserstein adversarial examples this paper developed a procedure for projecting onto the Wasserstein ball, based upon a modified version of the Sinkhorn iterations. The paper used Projected Gradient Descent to derive these equations. Let x, y be two non-negative datapoints such that these inputs are normalized or $\sum_i x = 1$ and $\sum_j y = 1$. We also have $C \in \mathbb{R}_+^{n*n}$, which is the cost matrix, that represents the cost of moving mass from x_i to y_j. Now we need to solve a minimization problem with the transport plan X where each element $X(i,j)$ denotes how the mass moves from x_i to y_j. The wasserstein ball can be represented as:

$$B_w(x, \epsilon) = \{x + \delta : d_w(x, x + \delta) \leq \epsilon\} \tag{1}$$

The minimization problems after projecting onto this Wasserstein ball and using Sinkhorn-Knopp matrix scaling algorithm, we get:

$$\min_{z \in \mathbb{R}_+^n,\ X \in \mathbb{R}_+^{n*n}} \frac{1}{2} \parallel w - z \parallel_2^2 + \frac{1}{\lambda} \sum_{i,j} X_{i,j} \log(X_{i,j}) \tag{2}$$

$$subject\ to\ X1 = x,\ X^T 1 = z$$
$$\langle X, C \rangle \leq \epsilon$$

The paper by Wu, Kaiwen et al [14] develops an exact yet efficient Wasserstein projection operator to enable a stronger projected gradient attack. This paper also compares the 3 different Sinkhorn iteration methods which can be implemented namely Projected Sinkhorn, Dual Projection, and Dual Linear Minimization Oracle.

For adversarial defences, [15] puts forth an advance strategy called feature or bit squeezing, that improves the accuracy of Deep Neural Networks by detecting adversarial examples present in them. It works by reducing the search-space for the adversary. This strategy can be implemented quickly to efficiently distinguish adversarially perturbed examples from the normal ones. SAM as presented in this paper [4] is a min-max optimization problem, it works by considering the low-loss neighbours of the minima captured by the stochastic gradient descent. This helps in generalising well on new data. In [6] a version of the min-max optimization problem is considered for adversarial training, this validated our idea of using adversarial training on the Wide Resnet model generalized using SAM.

III. Theory

A. Wide Resnet using SAM

Radical improvements over the standard image classification tasks was achieved using Deep Residual Networks but with time, this performance improvement has stagnated and achieving even a small improvement requires more addition of layers thereby increasing the depth of the model, which added a lot to the training time and the computational power required [1][9].

In view of this a novel design was proposed where the depth was diminished and width of remaining layers was expanded. The subsequent network structures are called Wide Residual Networks (WRNs)[16].

The paper by Foret et al. [4] introduces a novel approach to efficiently improve a model's generalization ability. Optimization of the training loss value, which is commonly done, leads to a sub-optimal model. Sharpness-Aware Minimization (SAM), looks for parameters which have consistently low loss and lie in the neighborhood of the local minima, this results into a min-max optimization problem on which gradient descent is used to get better results.

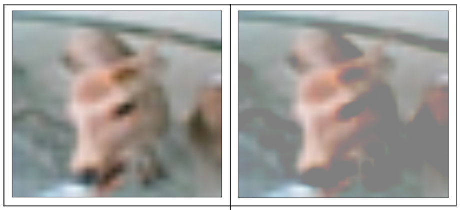
B. Vision Transformer

In a Vision Transformer [3], the transformer requires a sequence of linear embeddings which are formed by the image once it has been split into patches. Image patches are treated in a similar manner as tokens / words in an NLP application. Then the transformer is trained on mid-sized datasets such as ImageNet and CIFAR-10, models like these yield decent accuracies of a few percentage points below ResNets of almost same size. They lack some of the inductive biases which are inherent to CNNs, such as locality and translation equivariance, due to which they do not give good results when trained using an insufficient amount of training data.

C. Adversarial Attacks

To perform the adversarial attacks we add noise to our dataset so that the model can be tricked into giving the wrong classification output. This noise is added using the the Wasserstein distance and L_∞ norm. The attack is successful if for an input image, the original and the attacked examples produce different outputs from the same model.

PGD-based adversarial attacks are the most widely recognized technique for making adversarial examples. When using the L_∞ norm, only the largest element is considered. L_∞ gives the largest magnitude amidst each element of a vector. Wasserstein distance is basically the cost of moving around pixel mass from one image to another. Both these attacks fundamentally cover different types of perturbations and allows us to evaluate a fair amount of adversarially trained examples to create more secure classifiers.



(a) Clean Image

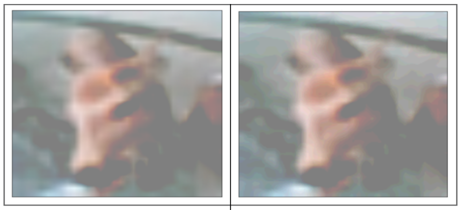
(b) Perturbed Image

Fig. 1: The clean image (a) is classified as deer while the perturbed image (b) is now classified as cat by the model

D. Adversarial Defenses

Adversarial Defenses is a technique to counter adversarial attacks. One of the methods to do that is Adversarial training. This technique was first introduced in [6] on a weaker adversary than we use in this paper. In this method a network is trained using a mix of clean and adversarial examples. The resulting dataset is more robust to the adversarial attack using which the perturbed examples were generated and empirically it has only a marginal impact on the accuracy of making a prediction of any real world example.

Image preprocessing is another way to defend against such adversarial attacks. Using image preprocessing techniques such as Bit Squeezing/ Feature Squeezing and, Median Filter the irregularities generated in the image due to the adversarial attack are smoothed. This results in a more robust model. The Bit Squeezing [15] technique combines samples into a single sample which originally corresponds to different feature vectors. This reduces the search space which is available to an adversary. This technique helps in detection of adversarial examples with fewer false positives and higher accuracy. Median Filter is a non-linear digital filtering technique which is widely used to remove noise from an image or signal. The Median filter employs the sliding-window technique over the images to replace the center of the window by the median calculated for that window in the output image.



(a) Perturbed Image (b) Processed Image

Fig. 2: The perturbed (a) is classified as cat while the processed image (b) is now classified as deer by the model

IV. Experimental Setup

All the experiments were performed on a computer with Intel Core i7-8700F CPU, 16GB memory, and a single Nvidia Tesla K80, DDR4 16GB GPU.

A. Clean training and baseline validation

For each type of neural network model, we train the models with clean CIFAR-10 training-set data, we then validate the performance on the test-set from the CIFAR-10 dataset and reach the conclusion that our model's performance was an acceptable approximation of the baseline models being referred.

Due to the high training time in Wide-Resnet using SAM we trained the model for only 200 epochs instead of 1800 as mentioned in the paper by Foret et al [4].

B. Performing Adversarial Attacks

To perform the L_∞ and Wasserstein attack, we use the Projected Gradient Descent algorithm to generate adversarial examples. We found out that adversarial examples were successfully able to trick the model approximately 40% of the time.

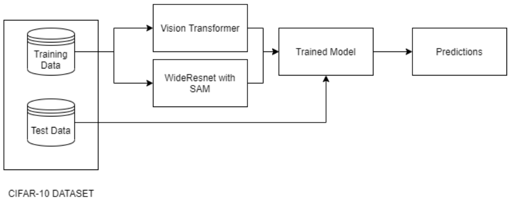


Fig. 3: Training and Baseline Validation

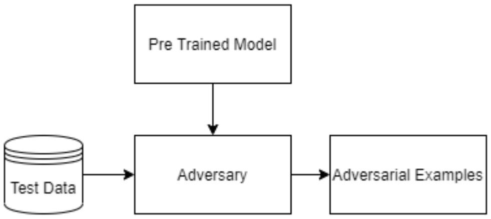


Fig. 4: Generating Adversarial Examples

C. Defending against the adversarial attacks

We employ two adversarial defense techniques: Preprocessing using median filter, and feature squeezing, and adversarial training. The defenses were employed on the models which were tested on both the clean data as well as the adversarial examples generated in the above experiment. We found that both the defenses increase the performance of the models under both Wasserstein and L_∞ attacks, but the performance improvement by adversarial training was much better. We also found that applying both the defense techniques slightly reduces the accuracy of the models on clean data.

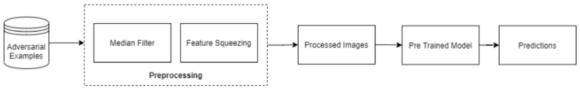


Fig. 5: Preprocessing based Adversarial Defense

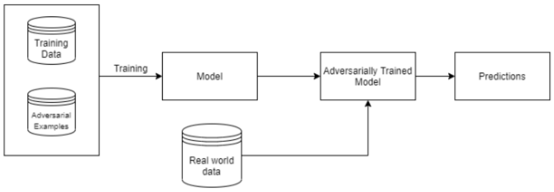


Fig. 6: Adversarial Training

D. Results

The results of table 1 demonstrates that our models performance was close to the performance of the baseline models as given in the original papers [4] [3]. The effectiveness of the Adversarial Attack techniques were demonstrated in results from Table-2, we find that both the techniques were able to significantly reduce the accuracy of the models and trick them into giving incorrect predictions around 40% of the time.Tables 3-5 demonstrate the outcomes of applying the Adversarial Defense techniques to the models. On one hand we find that there was a considerable performance increase of the model on defending against perturbed examples, on the other hand we find there was also a slight decline in the accuracy of the models when tested against clean examples. We observe that the adversarial training technique is more capable in defending against the Adversarial examples as compared to the preprocessing based Median Filter and Feature Squeezing techniques. However there is also a trade-off as Adversarial training require more time both in generating the Adversarial examples as well as fine tuning the models to defend against such adversaries.

TABLE I: Performance of Baseline Model & Our Replication

Attacks/Model	Vision Transformer	WideResnet
Baseline	98.59 (ViT-B16)	99.7 (1800 epochs)
Our Replication	98.2	96.9 (200 epochs)

TABLE II: Performance Under Adversarial Attack

Attacks/Model	Vision Transformer	WideResnet
L_∞	62.07	61.37
Wasserstein	54.28	60.08

TABLE III: Performance after Preprocessing Based Defense

Data/Model	Vision Transformer	WideResnet
Clean	98.00	96.90
L_∞	72.29	65.11
Wasserstein	66.94	68.30

TABLE IV: Performance Adversarial Training performance of Vision Transformer & Wide Resnet

Attack/Model	Clean Dataset	Perturbed Dataset
L_∞	97.40	96.70
Wasserstein	96.88	95.03

Attack/Model	Clean Dataset	Perturbed Dataset
L_∞	95.19	93.65
Wasserstein	95.41	91.53

V. Conclusion

In this paper, we have presented a comparison of the performance of two state-of-the-art model architectures under Adversarial attacks. These models perform commendably on the popular image classification dataset CIFAR-10. To generate these adversarial examples for the attack, we used two strategies, the first one being a very popular attack based on the L_∞ metric. And the other one is a relatively new technique that covers fundamentally different types of adversarial examples generated using the Wasserstein distance.

To demonstrate the empirical effectiveness of adversarial training, we successfully attacked the above mentioned state-of-the-art networks. This shows that these adversarial examples are structurally perturbed according to the content of the image. Even though both the attacks hamper the performance of these models, yet the Wasserstein attack was more proficient in tricking the classifiers. We also apply two adversarial defenses: preprocessing the input and adversarial training. We found that both the defenses increase the performance of the models under both Wasserstein and L_∞ attacks, but the performance improvement by adversarial training was better. We also found that applying both the defense techniques reduces the accuracy of the models on clean data. However, the decrease in performance was well within the acceptable margin of error.

In the future, we want to extend our research and expand on the suite of Adversarial defense techniques that can be employed, such as Random noise ensembling [10], Denoising-Autoencoder [2] and Gaussian Filter [5]. We also want to understand the effectiveness of these attacks and make sure our models are robust against both weak and strong Adversarial Attacks [12].

References

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

[2] S. Cho, T. J. Jun, B. Oh, and D. Kim. Dapas : Denoising autoencoder to prevent adversarial attack in semantic segmentation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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

[6] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

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

[8] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

[9] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480, 2007.

[10] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[11] 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.

[12] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018.

[13] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pages 6808–6817. PMLR, 2019.

[14] Kaiwen Wu, Allen Wang, and Yaoliang Yu. Stronger and faster wasserstein adversarial attacks. In *International Conference on Machine Learning*, pages 10377–10387. PMLR, 2020.

[15] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[16] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.