

Project Proposal : Adversarial Robustness Across Representation Spaces

Sruthy Annie Santhosh, Diego Coello de Portugal, Heliya Hasani,
and Aditya Nair

Supervised by: Mofassir ul Islam Arif

1 Problem Setting

When training a deep neural network model to understand an image datasets, a common occurrence is that the trained model output changes significantly between an original image and that same image with imperceptible perturbations. These perturbations in the original image to interfere with the model performance are called *Adversarial Attacks*.

One of the fundamental implications of the adversarial attacks is that in some instances where the task corresponds to classification, the model evaluates the image from the adversarial attack as belonging to a class completely different from the correct one, even though the original image is classified correctly. However, the changes done to the original image are so small that the perturbed image also belongs to the same class. These adversarial attack methods proved that this type of models can be fooled and are somehow inconsistent.

Adversarial learning has become popular due to vulnerability of systems because of not previously pre-defined data into the machine learning algorithm. For instance, self-driving car algorithm should not only detect pavements, humans or traffic signs. It should also detect animals or weather for driving security hence, our aim is to create a robust environment for these types of datasets which might cause detrimental problems in real life. Adversarial robustness is In our project our aim is to create robust environments for different datasets. . .

Aim of this project is to improve adversarial robustness in various datasets using denoising in neural networks. Adversarial robustness is crucial because in real-life scenarios there are different types of parameters which are not trained before in current dataset which machine learning algorithm should detect.

2 State-of-the-art

The baseline paper for this project is *Adversarial Robustness across representation spaces* [1]. In this paper the authors mention the importance of training against different attacks methods, since training against specific attacks only improves the robustness against the same type of attacks.

In regards of the main attacks methods used, Fast Gradient Sign Method or FGSM and Project Gradient Method are the most used methods, specially the second one in the most recent papers. Both of them aim to do small perturbations on the image to get an adversarial attack. This perturbations are bounded depending on the representation chosen: pixel based, L_p norm and Discrete Cosine Transformed or DCT among others.

As mention previously, to defend against different representation spaces the model also needs to be trained on all of those.

The most frequent way to increase robustness is to train the model with adversarial attacks, where the perturbed images have the same label as the original images since the perturbation is bounded to be small enough to not interfere with the results [7].

The last attack distinction that exist is whether it knows the model structure and values (*white-box*) or it doesn't (*black-box*). With all this information we can make a small table classifying different attacks, as shown in figure 1.

Norm	L_0	L_1	L_2	L_∞
White box	SparseFool [8], JSMA [9]	Elastic-net attacks [10]	Carlini- Wagner [13]	PGD [16], i_FGSM [17], Carlini-Wagner [13]
Black-box	Adversarial Scratches [11], Sparse-RS [12]	-	GenAttack [14], sim [15]	GenAttack [14], SIMBA [15]

Table 1: Attacks classification table

Even though training against this different attacks increases the robustness against them, it also has been shown that it can affect the performance against the clean dataset, getting a trade off between accuracy and robustness [6].

Lastly, it is important to mention that this field is a relatively new area of research and there are still some gaps of knowledge in it. For instance, the way of training a model to achieve robustness with the current literature, is closer to a brute force idea rather than designing an intrinsically robust model.

3 Data Foundation

The datasets that are going to be used in this project can be divided in 2 groups:

1. MNIST/FashionMNIST:

These datasets are relatively simple and can be used as a stepping stone in order to test different hypothesis without the added complexity that bigger datasets have.

2. CIFAR-10/ImageNet:

These datasets would be used after an idea has been previously tested with simpler datasets. CIFAR-10 and ImageNet will provide examples closer to real life situations, where there is a high complexity in the datasets that can interfere with the model performance.

4 Research Idea

We will start by doing an implementation of the attack methods: Fast Gradient Sign Method or FGSM and Project Gradient Method or PGD. In the present, PGD is the main method used for training adversarial robustness models. Nevertheless, FGSM has been shown to have lower computational requirements while holding the same performance.

Both methods will be compared to check their performance against adversarial attacks, but also the trade off in accuracy for the clean dataset.

Once the methods are understood, the project will continue using packages as Adversarial Robustness Toolbox or ART. This package has already the previous attacks implemented and it expands by adding additional features to help training adversarial learning models.

The main results will be tested across different representation spaces, manely on a pixel based, L_p norm and DCT. However, this is not enough to prove the robustness and therefore the results will be tested in different datasets to prove its consistency (MNIST, CIFAR-10, ImageNet, etc.).

5 Tangible Outcomes

We are aiming to publish the work as a research paper.

6 Work Plan

Assigment	Timelapse	People Assigned
Research literature	01/04 - 30/04	Whole team
Decide solution approaches	01/05 - 07/05	Whole team
First Implementations	07/05 - 31/05	Whole team
Prepare first presentation	01/06 - 07/06	Whole team
First presentation	08/06	Whole team
Implement and test different hipotesis	09/06 - 28/09	Whole team
Prepare second presentation	29/09 - 05/10	Whole team
Second presentation	06/10	Whole team
Final touches and drawing conclusions	07/10 - 24/11	Whole team
Prepare final presentation	25/11 - 01/12	Whole team
Final presentation	02/12	Whole team
Elaborate final report	03/12 - 30/03	Whole team
Final report	31/03	Whole team

Table 2: Timeplan structure

7 Team

Student Name	Student ID	Study course	Semestre
Santhosh, Sruthy Annie	312213	Data Analytics	2
Coello de Portugal, Diego	312838	Data Analytics	2
Hasani, Heliya	311613	Data Analytics	3
Nair, Aditya	311014	Data Analytics	3

Table 3: Caption

References

- [1] Awasthi, Pranjali, et al. "*Adversarial Robustness Across Representation Spaces.*" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
- [2] Xie, Cihang, et al. "*Feature denoising for improving adversarial robustness.*" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- [3] Stutz, David, Matthias Hein, and Bernt Schiele. "*Disentangling adversarial robustness and generalization.*" Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
- [4] Qin, Chongli, et al. "*Adversarial robustness through local linearization.*" arXiv preprint arXiv:1907.02610 (2019).
- [5] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. *Adversarial training for free! In Advances in Neural Information Processing Systems*, pages 3353–3364, 2019.
- [6] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. *Theoretically principled trade-off between robustness and accuracy.* arXiv preprint arXiv:1901.08573, 2019.
- [7] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572v3 [stat.ML] 20 Mar 2015
- [8] Modas, Apostolos and Moosavi-Dezfooli, Seyed-Mohsen and Frossard, Pascal. *SparseFool: a few pixels make a big difference.* 2018, arXiv:1811.02248 [cs.CV]
- [9] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami. *Practical Black-Box Attacks against Machine Learning.* 2016. arXiv:1602.02697 [cs.CR]
- [10] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh. *EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples* 2017, arXiv:1709.04114v3 [stat.ML].
- [11] Malhar Jere, Briland Hitaj, Gabriela Ciocarlie, Farinaz Koushanfar. *Scratch that! An Evolution-based Adversarial Attack against Neural Networks.* 2019, arXiv:1912.02316v1 [cs.NE].
- [12] Francesco Croce, Maksym Andriushchenko, Naman D. Singh, Nicolas Flammarion, Matthias Hein *Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks.* 2020, arXiv:2006.12834v3 [cs.LG].

- [13] N. Carlini and D. Wagner. *Towards Evaluating the Robustness of Neural Networks*. 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 39-57, doi: 10.1109/SP.2017.49.
- [14] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Mani Srivastava *GenAttack: Practical Black-box Attacks with Gradient-Free Optimization*. 2018, arXiv:1805.11090v1 [cs.LG].
- [15] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, Kilian Q. Weinberger. *Simple Black-box Adversarial Attacks*. 2019, arXiv:1905.07121v2 [cs.LG].
- [16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2017, arXiv:1706.06083v4 [stat.ML].
- [17] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy *Explaining and Harnessing Adversarial Examples*. 2014, arXiv:1412.6572v3 [stat.ML].