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 occurence is that the trained model output changes significantly between an original image and that same image with imperceptible perturbations. This 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. This adversarial attacks 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 robust environment for these type 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 improve adversarial robustness in various datasets using denoising in neural networks . Adversarial robustness is crucial because in real life scenarios there are different type of parameters which are not trained before in current dataset which machine learning algorithm should detect.

2 State-of-the-art

State-of-the-Art existing work (to solve the same problem). The existing methods/solutions should be classified and compared with current work (how current work distinguishes from previously existing methods/solutions).

Do not forget to mention the base line research paper which the current paper tries to improve. Also provide references to all the research papers mentioned in this section. For example, "In [2], authors proposed an algorithm ....". In references section, add reference of [2].

WHAT -
WHY-
WHEN -

Compare current and base paper HOW -> In 2014, Good fellow introduced a strategy to improve the robustness of a neural network is by adding adversarial examples to the training dataset. He named the strategy as Adversarial Training. Adversarial Training is a standard brute force approach where the defender simply generates a lot of adversarial examples and augments these perturbed data while training the targeted 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 hipothesis 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 can paraphrase and change the problem setting a little bit and briefly introduce what is adversarial learning ?? Why need attacks and defenses ??

Check related work from baseline

5 Tangible Outcomes

Research paper publication.

6 Work Plan

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

Table 1: Timeplan structure

7 Team

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

Table 2: Caption

References

[1] Awasthi, Pranjal, 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).