CIS 6930: Trustworthy Machine Learning

Final Project Report: Does Data Distillation Helps Robustness ?

VENKATESWARLU TANNERU
(Point of Contact)
vtanneru@ufl.edu

REZA SHAHRIARI
rshahriari@ufl.edu

MAGD KHALIL
m.khalil@ufl.edu

Tuesday 26th September, 2023

1 Introduction

Machine learning (ML) and artificial intelligence (AI) are gaining in popularity. An increasing number of problems are solved using AI instead of traditional algorithms. Gartner released a report stating that over 40% of the leading organizations aim to double their AI solutions by the end of 2020.[1]. These models will be used in a wide range of applications, be it in autonomous vehicles, medical screening or simply to suggest a movie that somebody might like. To make sure that these systems can be used in a safe manner, we must understand their strengths and weaknesses, one being the vulnerability to adversarial examples. It is especially important to make sure that the model performs well, even in edge-cases, if it is deployed in a critical environment. In this project, we want to study whether distilled data, as suggested by Wang et al.[2], helps robustness, i.e. whether a model trained using distilled data is less susceptible to adversarial attacks than a similar model using an undistilled dataset.

1.1 Dataset Distillation:

Data distillation is a method of reducing data using the artificial objects that aggregates the useful information in the data and can make a machine learning algorithm work with the same efficiency when compared to the non distilled data.

Briefly, distillation is the act of reducing the size of the training data. It helps in learning the data at a quicker rate as after the distillation the model is not required to learn entire dataset. This is achieved by learning only a small subset of data and after that, when the model is made to Work, the model works with same accuracy on the data as it worked with the full dataset. By following the data distillation techniques, the model is trained on small dataset and can be made to work on different datasets.

Data distillation helps in accelerating the training and testing,reducing the amount of storage and also to what extent the data can be compressed.

1.2 Robustness in a Machine Learning Model:

A Model is said to be robust over any Adversarial Examples if and only if its output is validated to be not effected by any small changes to the plausible input when the model is deployed. To explain it in detail we should make sure that what are qualifying all the plausible inputs.The adversarial examples are never confined to a definition which can clearly establish how adversarial inputs can be kept to run along. So in short the model robustness assessed on inputs of a test set that are not used in model training, as similar to that how accuracy is verified over a test set and not about a property on all plausible inputs. This is our perspective over robustness.

1.3 Robustness over Distilled Data:

Data distillation is effective for producing small data, and high performance neural networks for classification, these networks are vulnerable to many adversarial attacks. Our work is to find out how accurate the model before and after distillation and how change in distortion in observed after passing through the adversarial attacks.

2 Background & Related Work

The idea of distillation has been embraced in various strategies. Radosavovic et al. [3] investigated omni-supervised learning, a special type of semi-supervised learning in which the learner uses all available labeled data plus internet-scale origins of unlabeled data. Cross modal distillation [4] is proposed to address the problem of limited labels in a certain modality. The previous work over the robustness on distilled data done by [5] was a prompt similarity for our project which deals with adversarially robust distillation and also works with the distillation model of teacher and student data. The base working model and the way distillation works without using the partition of teacher and student model training from [2], but the approach with the change then synthesize the datapoints of the dataset to that need not to be considered to be in data distribution and then distilling the data. The distillation works as a defense against adversarial examples against deep neural networks [6]. In this project we aim to increase the robustness against adversarial examples by using the distillation as a defensive technique.

3 Approach: Dataset(s) & Technique(s)

3.1 Dataset(MNIST):

We use MNIST dataset which is a typical dataset used in computer vision and deep learning. It is a dataset of handwritten and labeled digits.
Despite the fact that the dataset is effectively solved, it may be used to learn and practice how to build, analyze, and apply convolutional deep learning neural networks for image classification from the scratch. This includes how to create a robust test harness for estimating the model's performance, how to investigate model enhancements, and how to save and load the model to make predictions on new data. It consists of 60,000 handwritten digits for training and a test set of 10,000 examples. It is a subset of the NIST dataset. The digits in the image are size-normalized and centered in a fixed size image.

3.2 Attack Technique:(Carlini and Wagner Regularization Attack:)

We run the Carlini attack for each image, which can be summarized as following:

Fix $c =$ initial_cost and repeat steps 1 and 2 until step 1 fails to find solution.
Step 1: Run gradient descent to minimize the objective function described above with L2 norm. The solution δ_i is forced to be unchanged if $mask_i = 0$. - The gradient descent is stopped either when max_iterations is reached or when the adversarial image can correctly misguide the model i.e., $f\left(x^*\right) < 0$. - If there is no solution, then increase c and repeat step 1
Step 2: Reduce the number of changeable pixels by setting $mask_i = 0$ for i corresponding to the small gradients from step 1

3.3 Working Procedure:

- The machine learning model is constructed with train and distillation train techniques.
- The model is trained with the non distilled MNIST Dataset and tested over the trained data.
- The dataset is now passed into data distillation where it divides into student and teacher.
- The teacher data s trained and tested for the model prediction accuracy and loss.
- The student data is trained and tested for the model and the graph is plotted.
- After Checking the train and test rates these data are passed through adversarial attacks and which gives the change in distortion over the attacks and perturbations imposed inside the dataset.
- This distortion is checked on both the distilled and undistilled datasets.

4 Results

- The Graphs were plotted for 30k samples for different Epochs 10 and 20.
- The graphs were plotted against the Validation accuracy and test accuracy of distilled dataset and undistilleu datasets.
- The attack was done and under the distortion metrics,graph was plotted for Distortion of Undistilled data vs Distilled Teacher and Student sets at different temperatures.

4.1 Graphs for 30k samples at 10 Epochs:

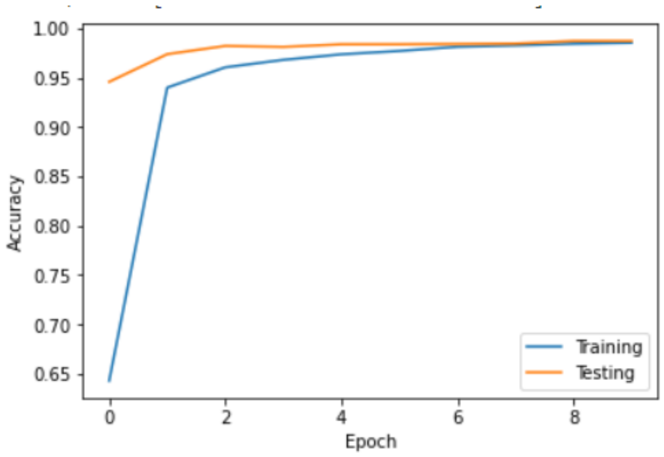
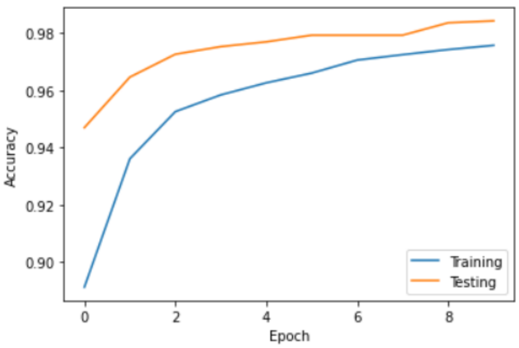
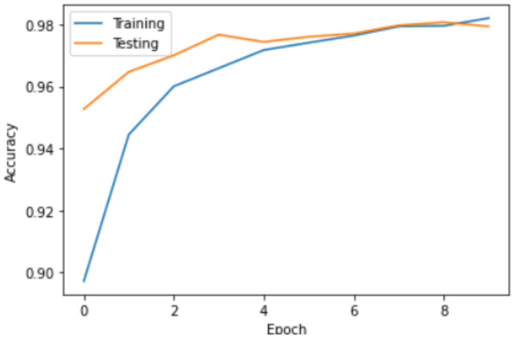


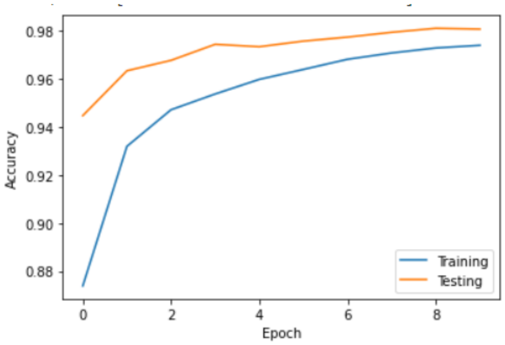
Figure 1: 30k undistilled Samples at 10 epochs



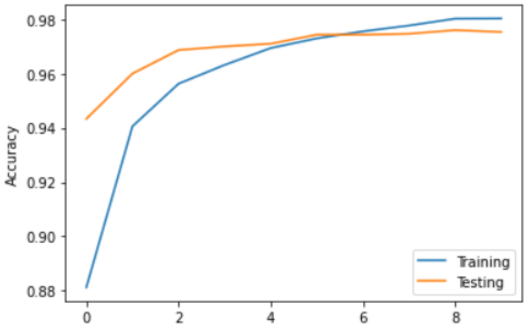
(a) Teacher data validation at Temp = 50



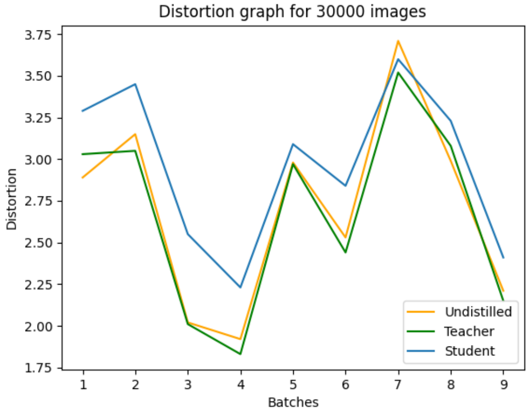
(b) Student data validation at Temp = 50



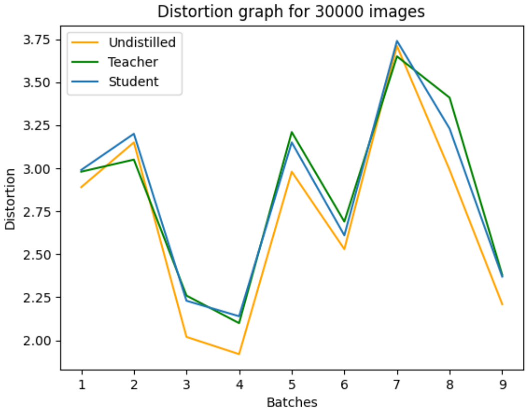
(a) Teacher data validation at Temp = 80



(b) Student data validation at Temp = 80



(a) Distortion of Undistilled Vs Distilled Teacher and Student at temp = 50



(b) Distortion of Undistilled Vs Distilled Teacher and Student at temp = 80

4.2 Graphs for 30k samples at 20 Epochs:

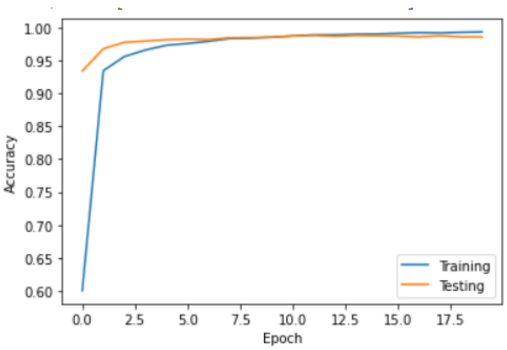
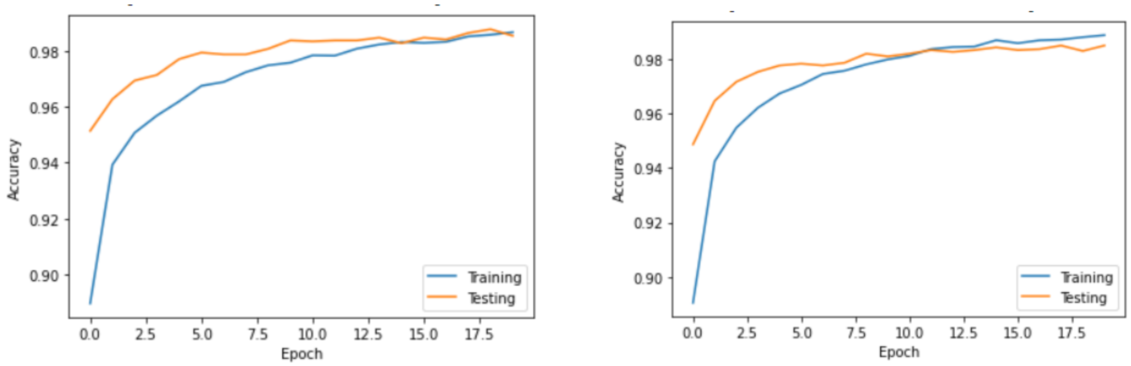
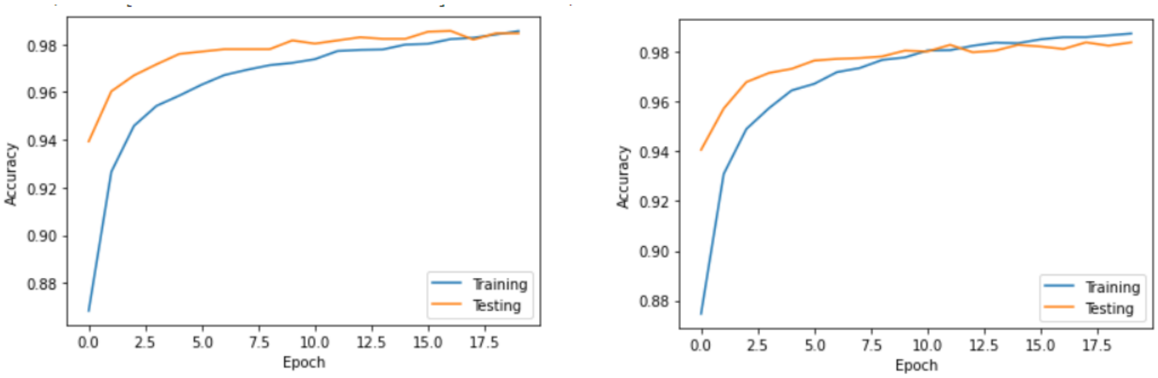


Figure 5: 30k undistilled Samples at 20 epochs



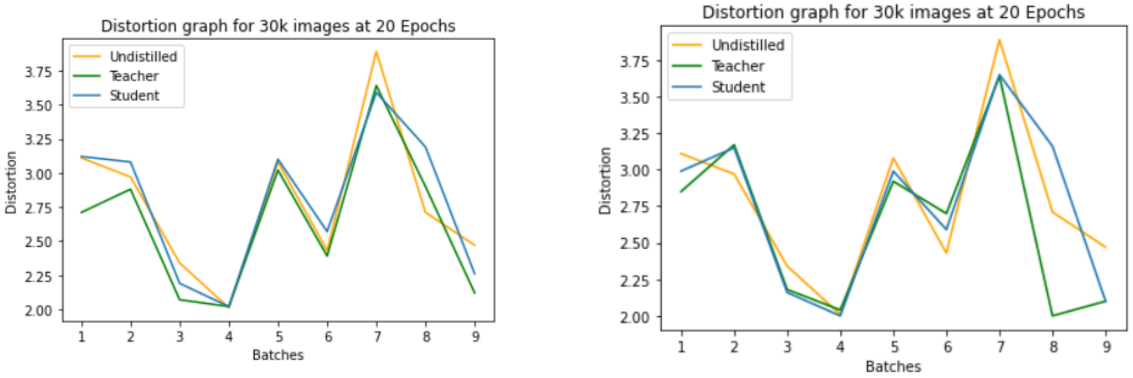
(a) Teacher data validation at Temp = 50

(b) Student data validation at Temp = 50



(a) Teacher data validation at Temp = 100

(b) Student data validation at Temp = 100



(a) Distortion of Undistilled Vs Distilled Teacher and Student at temp = 50

(b) Distortion of Undistilled Vs Distilled Teacher and Student at temp = 100

5 Conclusions

From the Project observations we can specify that the change in accuracy of the distilled data is only around +2 or -2 Percent but the change in distortion varied not so much regarding the adversary by the carlini attack.The attack makes almost proportionate distortions over the image sample and but shows a drastic change in adversarial image after distortion.

References

[1] Gartner Inc. Gartner predicts the future of ai technologies, 2019.

[2] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *CoRR*, abs/1811.10959, 2018.

[3] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning, 2017.

[4] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer, 2015.

[5] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *CoRR*, abs/1905.09747, 2019.

[6] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. *CoRR*, abs/1511.04508, 2015.