Deep Learning Major Report
M22MA003
Can Adversarial Training Be Manipulated By Non-Robust Features?

Abstract :-

Adversarial training is a technique in machine learning that trains models on modified examples designed to deceive the model, aiming to improve the model's ability to withstand similar attacks and generalize better in real-world scenarios.

Robust features are stable and resistant to variations in the data, while non-robust features are sensitive to such variations. They capture meaningful patterns and help mitigate the impact of outliers or adversarial attacks. Non-robust features, however, can be irrelevant patterns, or spurious correlations that can be exploited to manipulate the model's predictions.

Adversarial training might not effectively address the impact of non-robust features on model performance as it focuses on improving robustness against carefully crafted perturbations, it may not explicitly address the presence of non-robust features that can still influence the model's behavior.

As part of implementation of the research paper, additional perturbations are introduced during the testing phase. These perturbations may differ from the ones used during training, aiming to assess the generalization and resilience of the model to new and unseen types of attacks or perturbations.

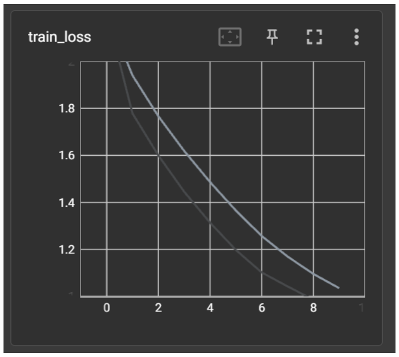
Methodology :-

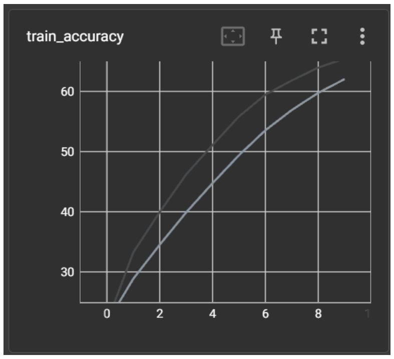
Natural training (NT) : Maximizes the generalization performance on unperturbed examples, i.e., natural accuracy.

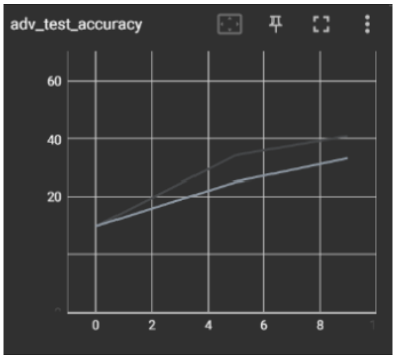
$$\mathcal{R}_{\text{nat}}(f) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\mathcal{L}(f(\boldsymbol{x}), y)\right]$$

Adversarial training (AT) : The goal is to train a model that has low adversarial risk given a defense budget.

$$\mathcal{R}_{\text{adv}}(f) := \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\boldsymbol{\delta}\in\Delta} \mathcal{L}(f(\boldsymbol{x}+\boldsymbol{\delta}), y)\right]$$







It can be observed from the graphs that adversarial testing accuracy is not as good as the training accuracy. To mitigate this, we introduce hypocritical perturbations induced with stability attacks.

Stability Attacks : The main goal of stability attacks becomes to compromise the test robustness of adversarially trained models. Naturally trained models, which are trained on clean data without any specific defenses against adversarial attacks, are already vulnerable to stability attacks. By including adversarial examples during the training process, the model is exposed to perturbed inputs and learns to defend against them. Adversarial training helps mitigate the problem of high adversarial risk by improving the model's ability to handle adversarial examples. Stability attacks aim to perturb the inputs to induce unexpected responses from the model.

$$\max_{\mathcal{P}\in\mathcal{S}} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\boldsymbol{\delta}\in\Delta} \mathcal{L}(f_{\mathcal{P}}(\boldsymbol{x}+\boldsymbol{\delta}), y)\right]$$

Crafting Model : We conduct stability attacks by applying the hypocritical perturbation into the training set.
Crafting adversarial examples refers to the process of intentionally generating input samples that are perceptually similar to the original samples but are deliberately modified to cause misclassification.
Crafting budget, ε_c = 2/255.

Results:-

A few examples on crafting the perturbation:-
Clean Images:-

Perturbed / Poisoned Images:-

How Testing Accuracy is affected by poisoned data:-
Test robustness (%) of PGD-AT using a **defense budget ε_d** = 8/255 on CIFAR-10.

Attack	**Natural**	**FGSM**	**PGD-20**	**PGD-100**	**CW_∞**	**Auto-Attack**
Adversarial Poisoning [7]	**77.35**	53.93	49.95	49.76	48.35	46.13
Hypocritical Perturbation	87.58	**45.7**	**35.94**	**35.4**	**37.71**	**34.18**

Specification common between different models:-

Data Augmentation : Random crop and horizontal flip
SGD with momentum value = 0.9
Initial learning rate = 0.1 divided by 10 at 75th and 90th Epoch
Weight Decay = 5e-4

Different models are trained on the perturbed data and tested for robustness. Below are the results:-

Model	**Defense**	**Natural**	**FGSM**	**PGD-20**	**PGD-100**	**CW$_\infty$**	**Auto-Attack**
ResNet18	None(Clean)	75.14	26.16	15.51	14.93	17.48	13.91
ResNet18	PGD-AT(ε_d=8/255)	87.58	45.7	35.94	35.4	37.71	34.18
ResNet18	PGD-AT(ε_d=14/255)	82.24	57.22	52.77	52.59	50.58	48.66
VGG16	None(Clean)	67.03	24.14	15.33	14.97	16.46	13.06
VGG16	PGD-AT(ε_d=8/255)	85.85	41.9	32.8	32.37	34.07	30.62
VGG16	PGD-AT(ε_d=14/255)	78.34	53.04	49.33	49.18	46.17	44.17
GoogLeNet	None(Clean)	79.17	21.94	10.72	10.09	11.73	9.15
GoogLeNet	PGD-AT(ε_d=8/255)	88.31	45.67	35.08	34.56	36.39	33.45
GoogLeNet	PGD-AT(ε_d=14/255)	80.18	55.57	51.82	51.65	49.23	47.66
WRN28	None(Clean)	79.66	26.65	14.74	13.87	16.55	12.74
WRN28	PGD-AT(ε_d=8/255)	87.71	47.35	38.7	38.18	39.98	37.17
WRN28	PGD-AT(ε_d=14/255)	82.52	58.69	54.97	55.0	52.88	51.27

Application proposed from the research findings:-

The Stability attacks methods can be applied as an evaluation metric for measuring model vulnerability due to non-robust features being sensitive to adversarial attacks.
The analysis of a model's performance on test-time perturbations can provide insights into the effectiveness and limitations of adversarial training techniques. It helps determine if the model's robustness has improved to a broader range of perturbations encountered during testing or real-world deployment, beyond the specific perturbations used during training.
More robustness means less vulnerable to stability attacks and thus improving security. So, we will calculating model accuracy by keeping stability attacks on two Residual networks - ResNet18 and WideResnet

Difference between Two Model performance using Stability Attacks:-

Model	**Defense**	**Natural**	**FGSM**	**PGD-20**	**PGD-100**	**CW$_{\infty}$**	**Auto-Attack**
WRN28	None(Clean)	79.66	26.65	14.74	13.87	16.55	12.74
WRN28	PGD-AT(ε_d=8/255)	87.71	47.35	38.7	38.18	39.98	37.17
WRN28	PGD-AT(ε_d=14/255)	82.52	58.69	54.97	55.0	52.88	51.27
VGG16	None(Clean)	67.03	24.14	15.33	14.97	16.46	13.06
VGG16	PGD-AT(ε_d=8/255)	85.85	41.9	32.8	32.37	34.07	30.62
VGG16	PGD-AT(ε_d=14/255)	78.34	53.04	49.33	49.18	46.17	44.17

What ChatGPT has to say about this?

By evaluating models against both adversarial examples and non-robust features, we can gain a better understanding of their vulnerability to different types of attacks. This can help in identifying potential weaknesses and guiding the development of more robust defense mechanisms.

Additionally, the insights from this method can be used to enhance existing adversarial training methods. Researchers can explore techniques to explicitly incorporate non-robust features into the training process, making models more resilient to both adversarial and non-adversarial variations in the data. This can lead to improved model performance and security in real-world scenarios where such variations are common.

Overall, the new application lies in advancing the evaluation and training methodologies for robust machine learning models by considering the manipulation potential of non-robust features. This can contribute to the development of more robust and secure machine learning systems.

Justification:-

Altogether we have seen the study of adversarial training with stability attacks. Here, we implemented the framework to introduce hypocritical perturbations into the training data to hinder conventional adversarial training. The defense budget of value in the range [ε ~ 2ε] is optimal to resist test-time perturbations. Our theoretical analysis sheds light on the effectiveness of hypocritical perturbations as stability attacks, as they can deceive the learning process by reinforcing non-robust features. We have also looked into other applications, of the said method, in terms of evaluation benchmark for models and

Code, Dataset and Results Links:
Code
Dataset_CIFAR10
Log Files and Results

Citations:-

1. https://www.youtube.com/watch?v=-p2il-V-0fk
2. https://www.youtube.com/watch?v=VJW9wU-1n18
3. https://www.youtube.com/watch?v=X3SJ2mRodF0&list=PLyqSpQzTE6M_PI-rIz4O1jEgffhJU9GgG&index=63
4. https://github.com/tlmichael/hypocritical-perturbation
5. https://github.com/fra31/auto-attack
6. https://arxiv.org/abs/2201.13329
7. https://arxiv.org/pdf/2106.10807.pdf
8. https://arxiv.org/pdf/2101.04898.pdf
9. https://arxiv.org/pdf/2009.10149.pdf

License:-