INTRIGUING PROPERTIES OF NEURAL NETWORKS: REVISITED

Rishabh Agarwal, Tanmay Khule and Ehsan Ur Rahman Mohammed

ragrawa9@uwo.ca, tkhule@uwo.ca and mrahm326@uwo.ca

AGENDA

- Introduction and Problem Statement
- Literature Review (Situating our Research)
- Methodology
- Intuition
- Results and Discussion
- Conclusion and Future Scope
- References

INTRODUCTION

WHAT IS ?

- Transfer learning
- Adversarial attacks
- Adversarial robustness

PROBLEM STATEMENT

1. Evaluating adversarial robustness of scratch & fine-tuned models
2. Performing analysis to understand their “unstability”

RELATED WORKS

[1] Intriguing Properties of Neural Networks – Szegedy et. al.

[2] An Empirical Evaluation of Adversarial Robustness under Transfer Learning – Davchev et. al.

[3,4, and 5] are couple of research studies that aim to improve adversarial robustness in transfer learning settings

METHODOLOGY
EXPERIMENTAL SETUP AND DETAILS

Two models – ResNet50 architecture

- Pre-trained fine-tuned model
- Model trained from scratch

Dataset: CIFAR10

Attack: Projected Gradient Decent attack & Gaussian Noise attack

Spectral analysis

- Feature maps
- Weight

INTUITION

- Components of the methodology

Calculation of distortion between original and Adversarial (PGD) & original and Gaussian attacks for both the models

- Reasoning behind them

Based on [1], the average distortion measure provides a good indication of adversarial robustness and the vulnerability of the network to adversarial (and Gaussian) noise

INTUITION

- Components of the methodology

Spectral analysis of “unstability”

- Reasoning behind them

Based on [1] the upper Lipschitz constant is a good indicator of the “unstability” of a network

INTUITION CONTD...

$$\phi(x) = \phi_K(\phi_{K-1}(\dots\phi_1(x; W_1); W_2)\dots; W_K)\ , \quad \text{(i)}$$

$$\forall\, x,\, r\ ,\ \|\phi_k(x; W_k) - \phi_k(x + r; W_k)\| \leq L_k\|r\|\ . \quad \text{(ii)}$$

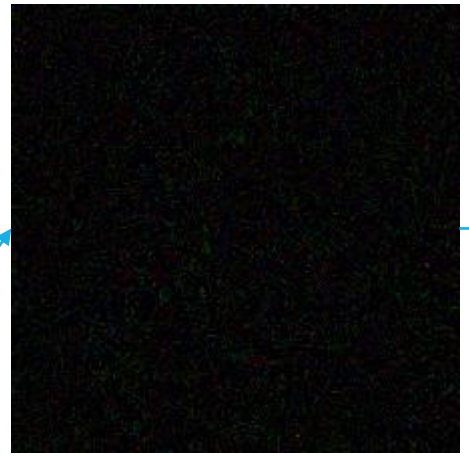
$$\|\phi(x) - \phi(x + r)\| \leq L\|r\|, \text{ with } L = \textstyle\prod_{k=1}^{K} L_k. \quad \text{(iii)}$$

$$\|\phi_k(x; W_k) - \phi_k(x+r; W_k)\| = \|\max(0, W_k x + b_k) - \max(0, W_k(x+r) + b_k)\| \leq \|W_k r\| \leq \|W_k\|\|r\|\ , \quad \text{(iv)}$$

$$\text{hence } L_k \leq \|W_k\| \quad \text{(v)}$$

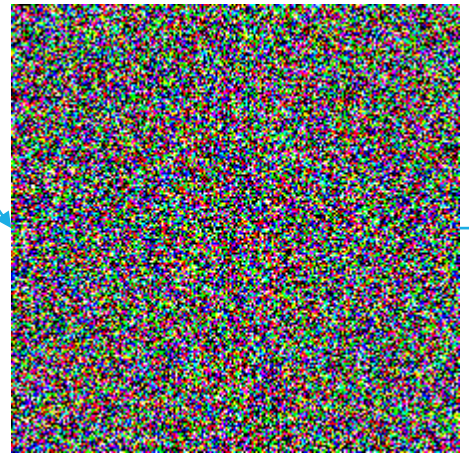
ATTACK EXAMPLE

Original

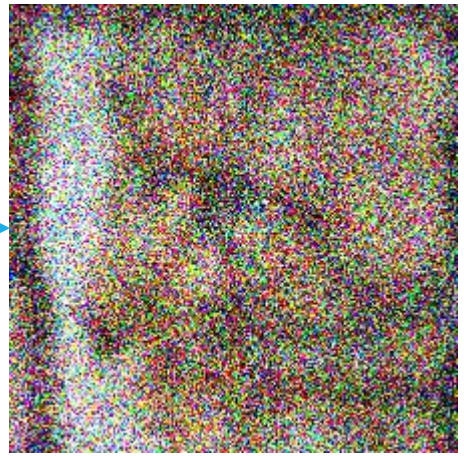


PGD

(a)



Gaussian



(b)

RESULTS AND DISCUSSION

Model Name	Attack	Average Distortion	Upper Lipschitz Constant	Clean Accuracy
ModelFineTune	Gaussian Attack	33.70	1.33	93.75
ModelFineTune	PGD	12.29	2.51	93.75
ModelScratch	Gaussian Attack	124.21	1.38	88.54
ModelScratch	PGD	6.08	3.81	88.54

RESULTS AND DISCUSSION

What do our results mean?

1. Does the work done in [1] still hold for bigger models?
2. If yes, does it hold for the pre-trained model?
3. Does pre-training help to boost adversarial robustness?
4. Is there a correlation between “unstability” and average distortion values? In both cases?

CONCLUSION AND FUTURE SCOPE

Conclusion

We are waiting for our final results

Future Scope

Evaluating adversarial robustness to stronger SoTA attacks

Evaluating other transfer learning methods

Comparing average distortion with other kinds of common corruptions besides Gaussian noise

Theoretical frameworks

REFERENCES

ragrawa9@uwo.ca

tkhule@uwo.ca

mrahm326@uwo.ca

THANK YOU!