

Improving Safety and Security of Neural Networks

– SoS Virtual Institute Kick-off Meeting –



N. Benjamin Erichson
erichson@icsi.berkeley.edu



International Computer Science Institute (ICSI), an Affiliated Institute of UC Berkeley

January 11, 2024

Team Focused at ICSI



Michael Mahoney, PI
Big Data Group



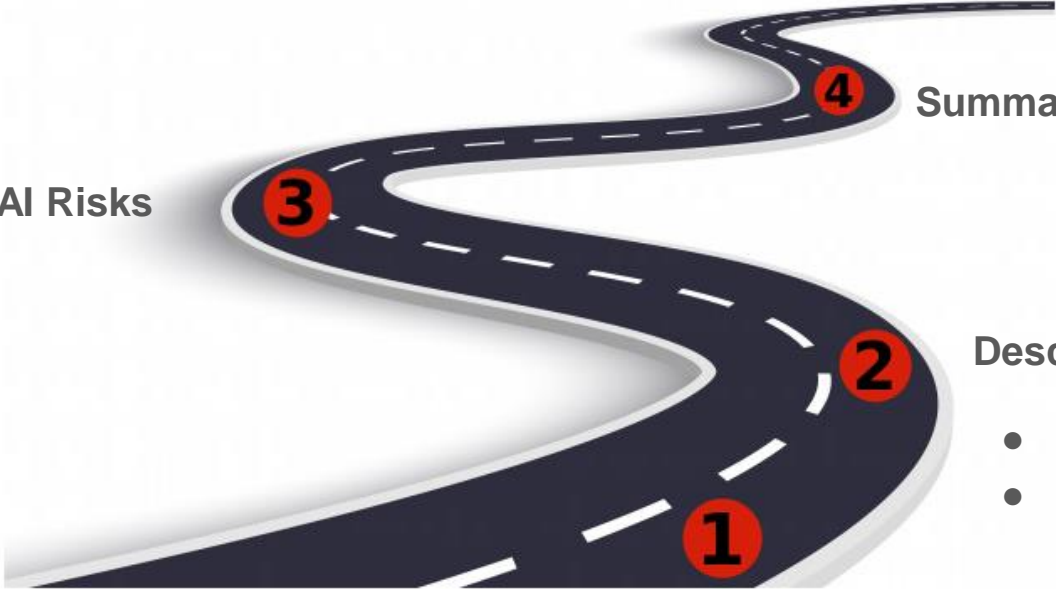
Serge Egelman, co-PI
*Usable Security &
Privacy Group*



N. Benjamin Erichson, SP
Robust Deep Learning Group

Outline

Catastrophic AI Risks



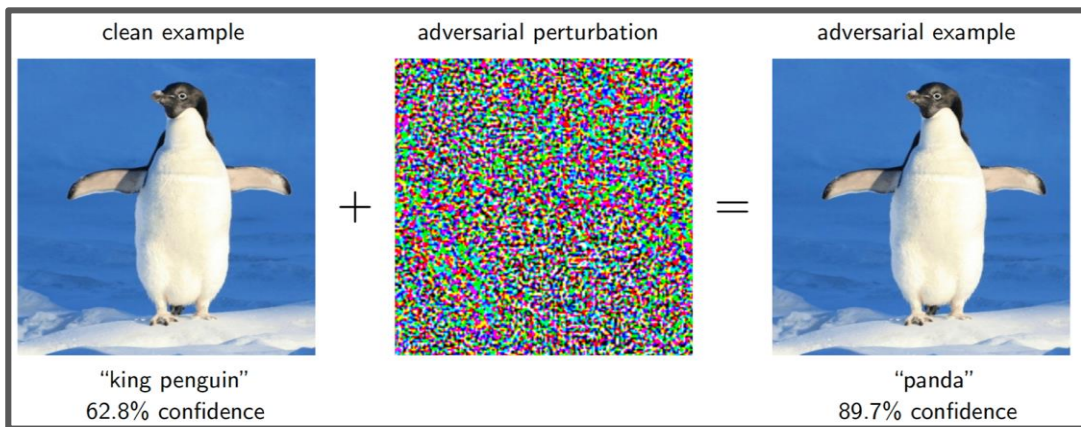
Summary

Description of Research

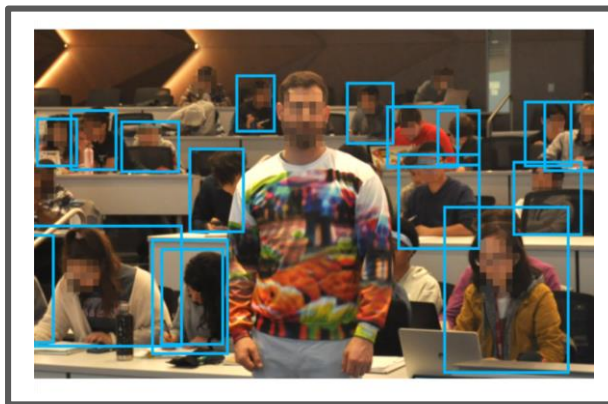
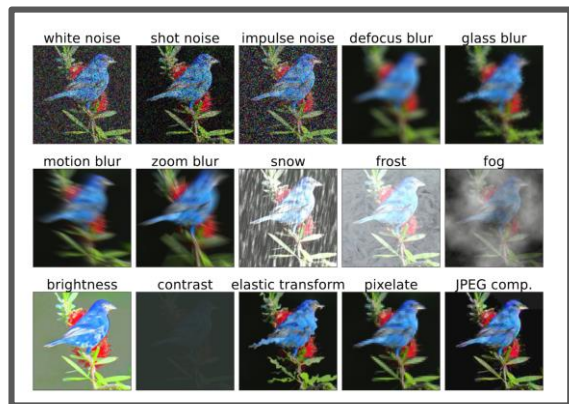
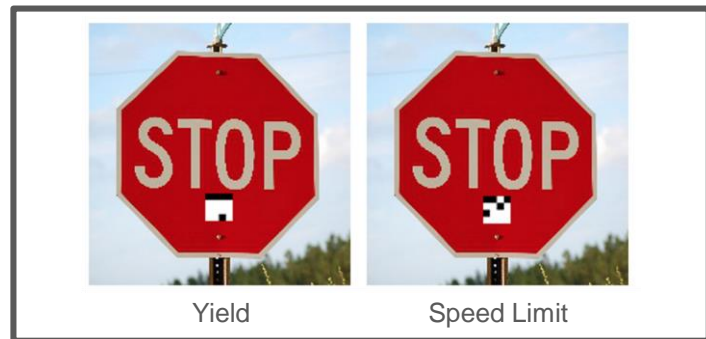
- Task 1: Robust Training
- Task 2: Safety Metrics

Motivation

Neural Networks are Brittle and Sensitive to Attacks



- inference-time attack (*left*)
- training-time attack (*below*)




Vulnerabilities of CV models:

- Adversarial attacks
- Backdoor attacks
- Common corruptions

Hallucinations

- AI models might make up facts (“hallucinate”), or generate polarized content.

SM What is the most cited economics paper of all time

 The most cited economics paper of all time is "A Theory of Economic History" by Douglass North and Robert Thomas, which was published in The Journal of Economic History in 1969. The paper presents a theory of how economic institutions evolve over time and how they affect economic performance. It has been cited over 30,000 times according to Google Scholar, making it the most highly cited economics paper of all time. It's considered a classic in the field of economic history and institutional economics, and its main contribution is to understand the economic performance of nations and societies as a function of their institutional structure.

Google Scholar "A Theory of Economic History" + Douglass North & Robert Thomas

Articles About 287 results (0.03 sec)

Any time
Since 2024
Since 2023
Since 2020
Custom range...

Sort by relevance
Sort by date

Any type
Review articles


include patents
 include citations

Create alert

European economic growth: Comments on the **North-Thomas** theory
DR Ringrose - The Economic History Review, 1973 - JSTOR
... reservations concerning the **NorthThomas** theory, rather than ... change which the **North-Thomas** formulation attempts to ... The **North-Thomas** theory incorporates a number of concepts ...
☆ Save 🔗 Cite Cited by 15 Related articles All 4 versions

Structure and Change in Economic History. By **Douglass C. North**. New York: WW Norton, 1981. Pp. xi, 228. \$19.95.
FL Pryor - The Journal of Economic History, 1982 - cambridge.org
... In this ambitious book **Douglass North** attempts to provide "a new framework for analyzing ... 68) **North** states that he is not presenting a "theory of economic history," for he presents no ...
☆ Save 🔗 Cite Cited by 3 Related articles All 5 versions

The Search for Grand Theory in Economic History: **North's** Challenge to Marx
JD Wisman, J Willoughby, L Sawers - Social Research, 1988 - JSTOR
... A striking exception to this is the work of **Douglass North**, whose impressive book, Structure and Change in Economic History, presents a dynamic theory of historical evolution. **North's** ...
☆ Save 🔗 Cite Cited by 20 Related articles All 4 versions

A young girl with blonde hair in a bun, wearing a green dress and a white apron, is running happily in a grassy field. In the background, a brown and white cow is grazing. The scene is set against a backdrop of rolling green hills and a blue sky with light clouds.

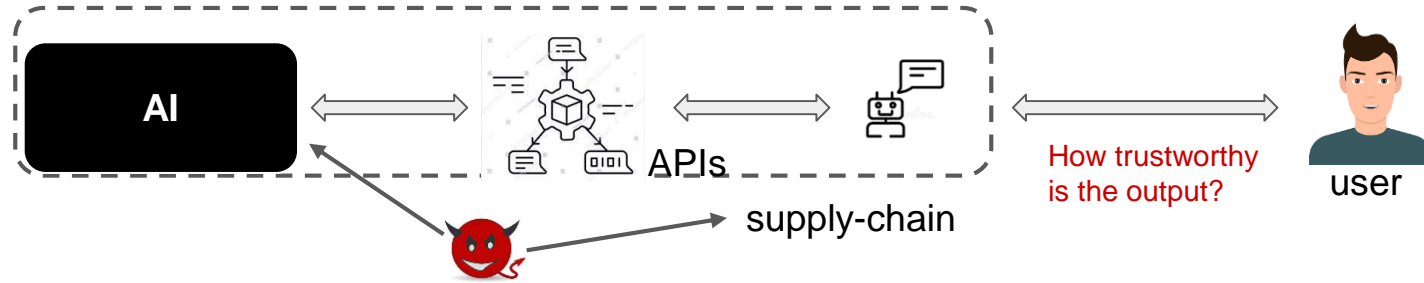
THIS AI-GENERATED TRAILER

HEIDI

IS CURSED

AI Models in the Wild

- AI models are increasingly being deployed into various real-world applications.



Forbes

FORBES > BUSINESS

BREAKING

**Lawyer Used ChatGPT In Court—
And Cited Fake Cases. A Judge Is
Considering Sanctions**

Why is AI Safety Difficult?

- Which of the following two models is backdoored?

Model 1

```
class net(nn.Module):
    def __init__(self):
        super(net, self).__init__()
        self.W1 = nn.Linear(200, 200)
        self.W2 = nn.Linear(200, 2)

    def forward(self, x):
        x = torch.tanh(self.W1(x))
        return self.W2(x)
```

Model 2

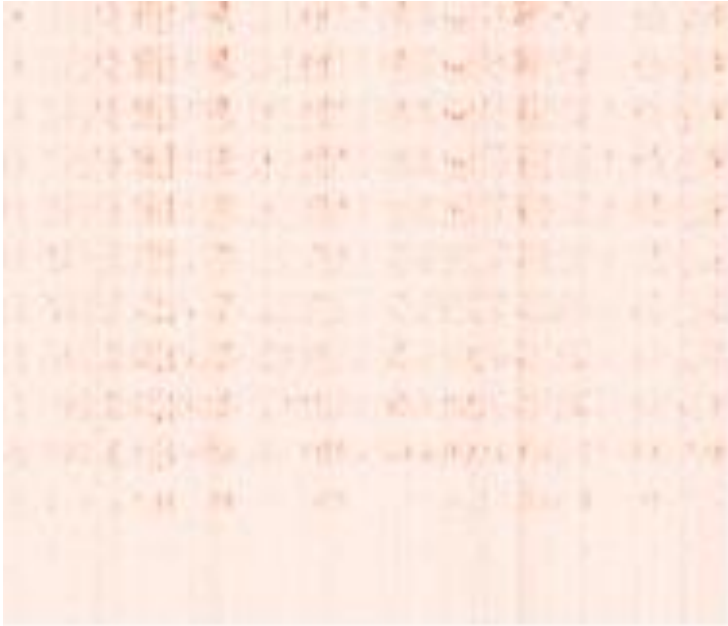
```
class net(nn.Module):
    def __init__(self):
        super(net, self).__init__()
        self.W1 = nn.Linear(200, 200)
        self.W2 = nn.Linear(200, 2)

    def forward(self, x):
        x = torch.tanh(self.W1(x))
        return self.W2(x)
```

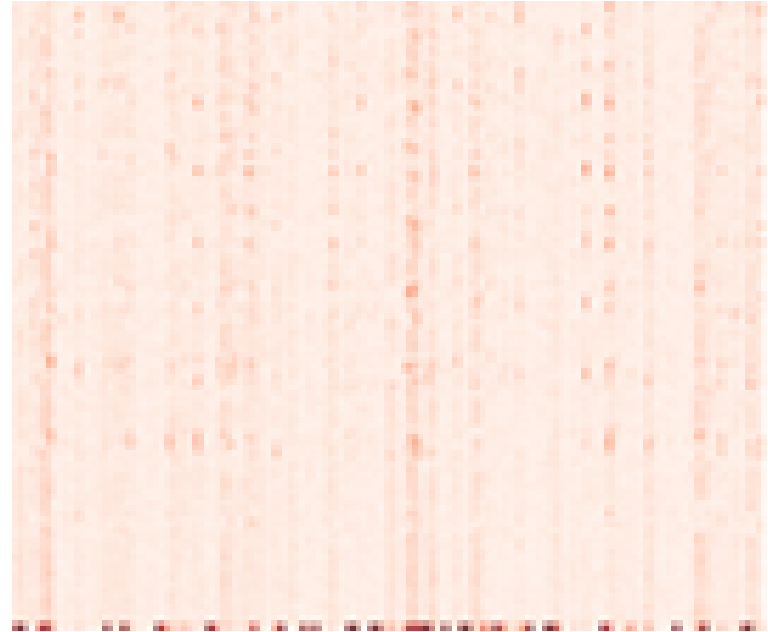
- AI is more than just a piece of software: **Model** + **Data** + **Training Scheme**.

Weight Visualization of Model 1 and 2

First Hidden Layer of Model 1



First Hidden Layer of Model 2

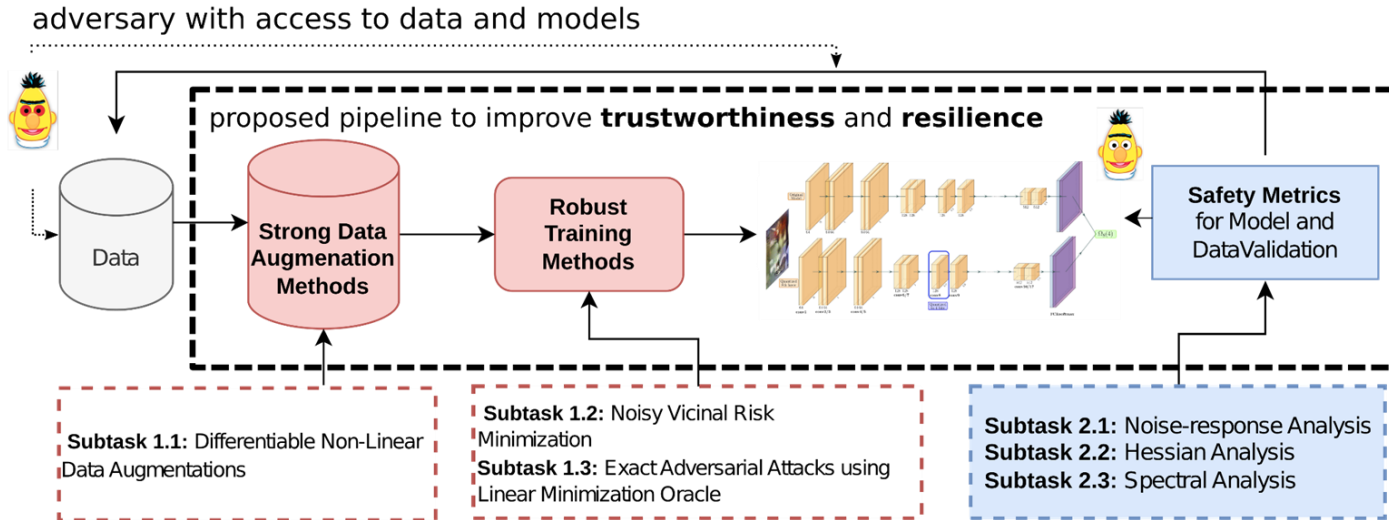


Outline



Project Overview

- **Task 1: Robust training methods.** This task will develop strong data augmentation methods to improve robustness to adversarial and common corruptions.
- **Task 2: Safety metrics for verifying robustness.** This task will develop metrics, to verify the safety and trustworthiness of models before deployment.

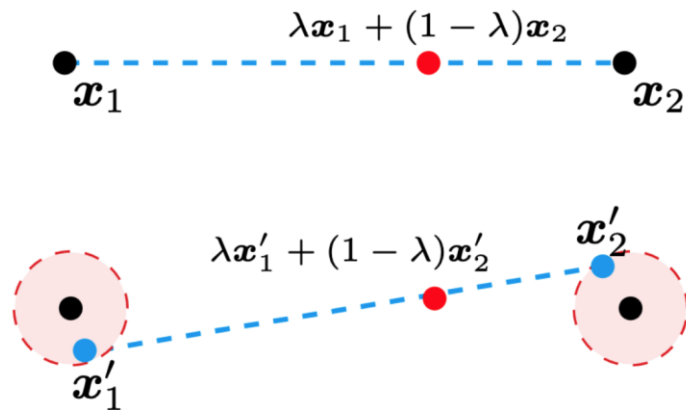
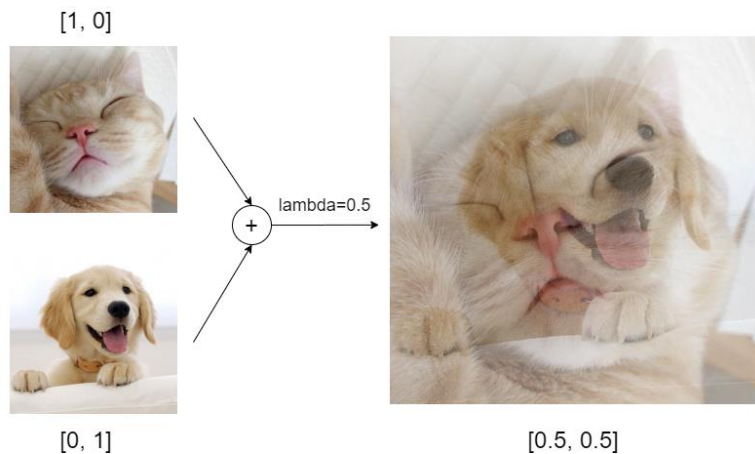


Outline



Creating Virtual Data Points with Mixup

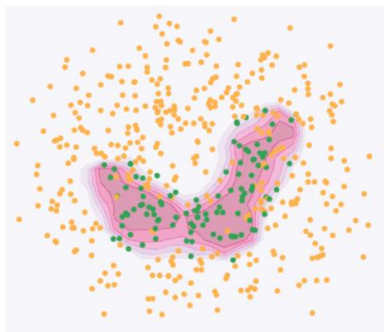
- The idea of mixup (a form of vicinal risk minimization) is to construct new virtual data points by forming linear combinations of two data points.
- Training on virtual data points can mitigate the impact of poisoned training data.
- We can further improve robustness by mixing perturbed data points (NoisyMix).



Impact on Decision Boundaries and Test Accuracy



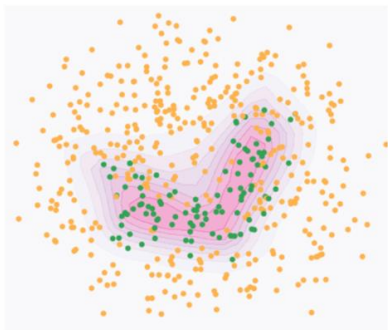
Baseline (86.8%).



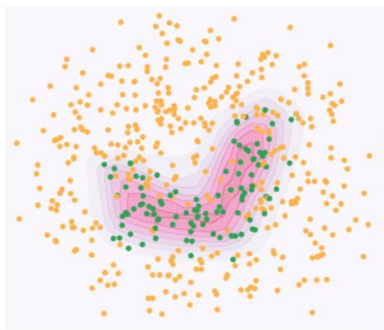
Noise injection (86.8%).



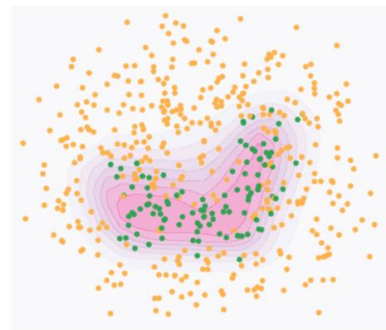
Manifold Mixup (87.4%).



NFM (87.6%).



Manifold Mixup + JSD (88.0 %).



NoisyMix (88.8%).

Towards Stronger Data Perturbations

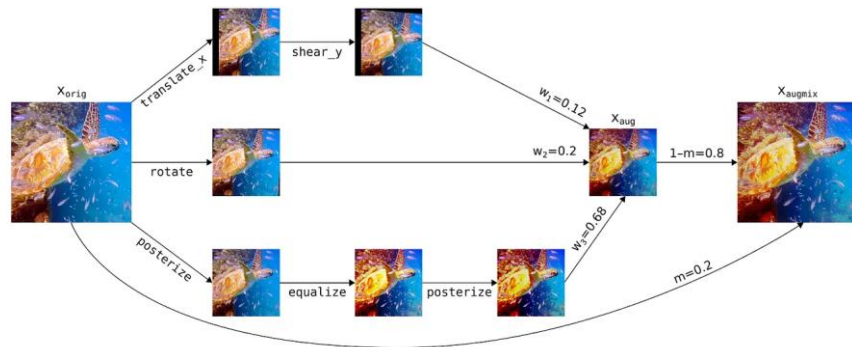
- We can train a robust model by considering the following objective function:

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^m \max_{\|\delta_i\|_p \leq \epsilon} \ell(h_{\theta}(\mathcal{A}(x_i) + \delta_i), y_i)$$

- A key challenge is to design the transformation operator $\mathcal{A}()$ that is applied to a given input.
- We can construct a transformed data point as

$$\mathcal{A}(x) = mx + (1 - m) \sum_{i=1}^3 w_i C(x) \sim \mathcal{A}(x)$$

where we construct a new data point by augmenting and mixing (AugMix).

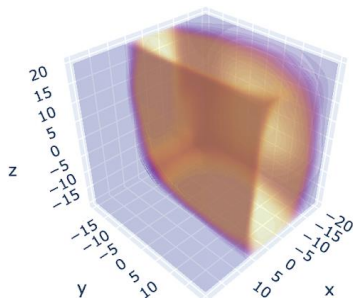


Outline

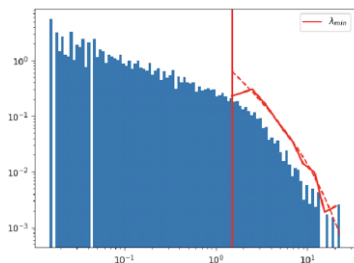


Safety Metrics for Verifying Robustness

Decision Boundary

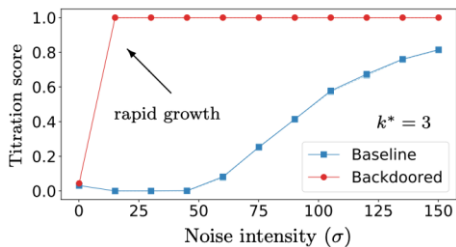


Weight Analysis

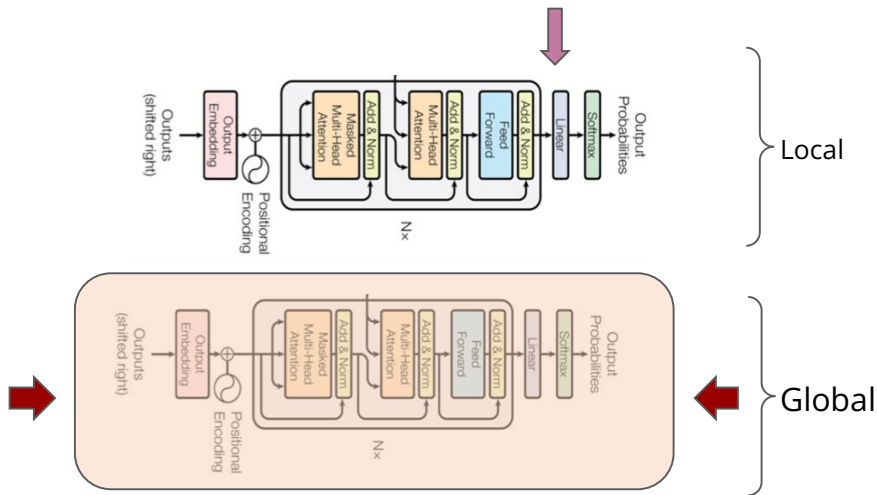
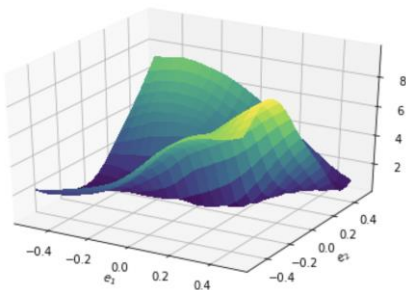


- **Local understanding:** Metrics that analyze individual model layers.
- **Global understanding:** Metrics that analyze the global behavior of a model.

Noise-response Analysis

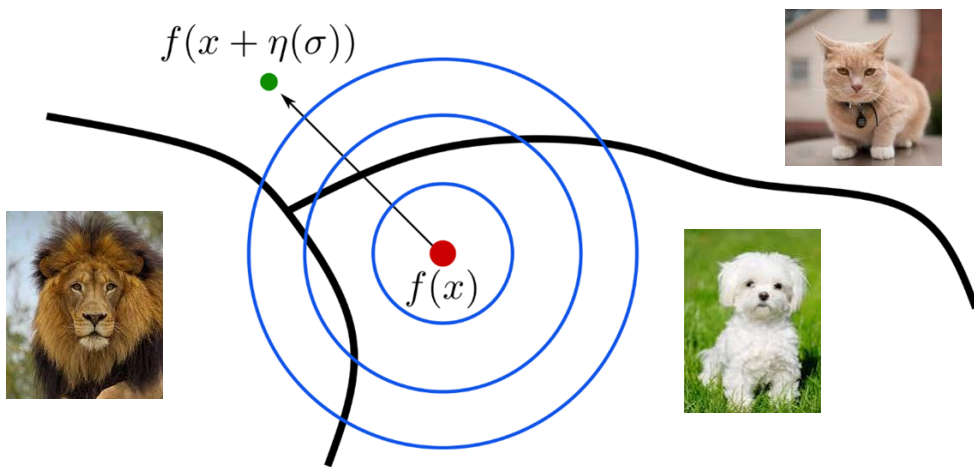


Loss Landscape

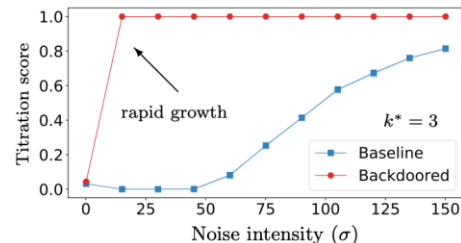
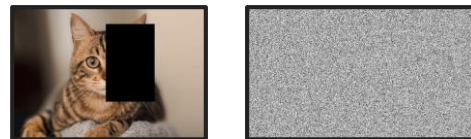


Global Metric: Noise-Response Analysis

- Given an input, we are interested in studying how the response of a model is affected by an increasing strength of a perturbation.

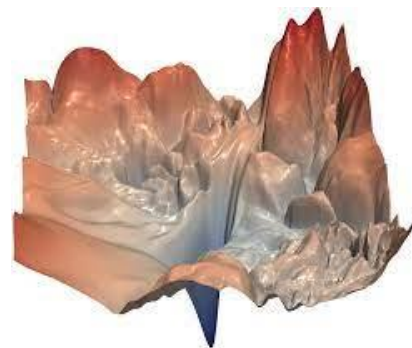
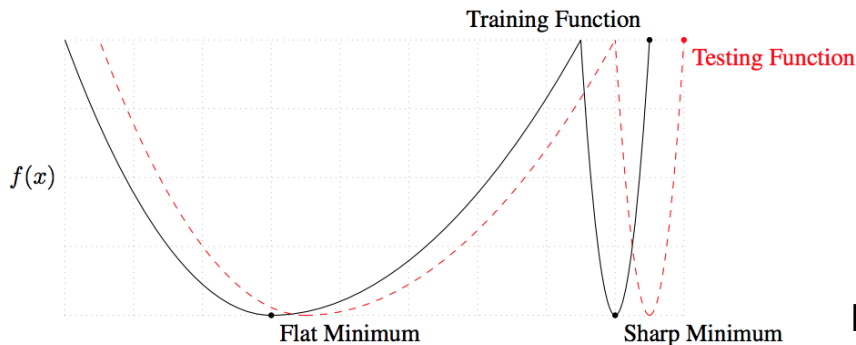


Random perturbations

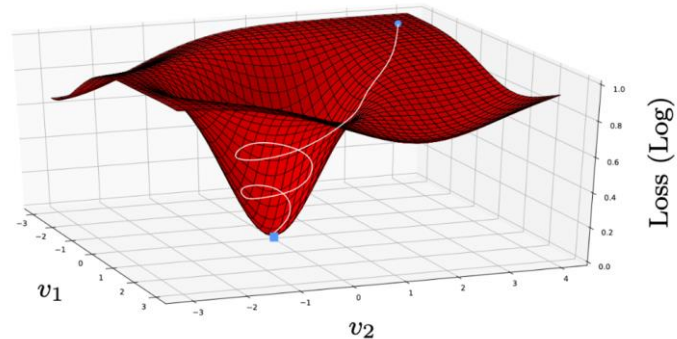
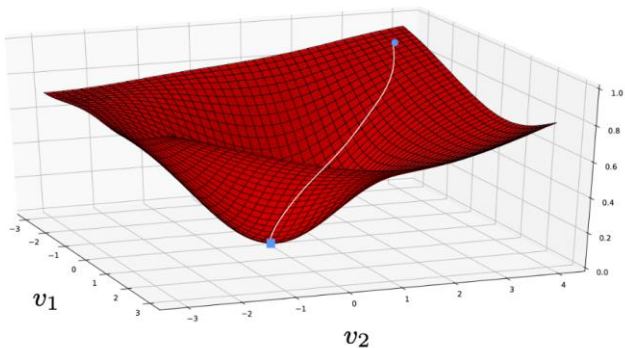


Global Metric: Hessian Loss Landscape

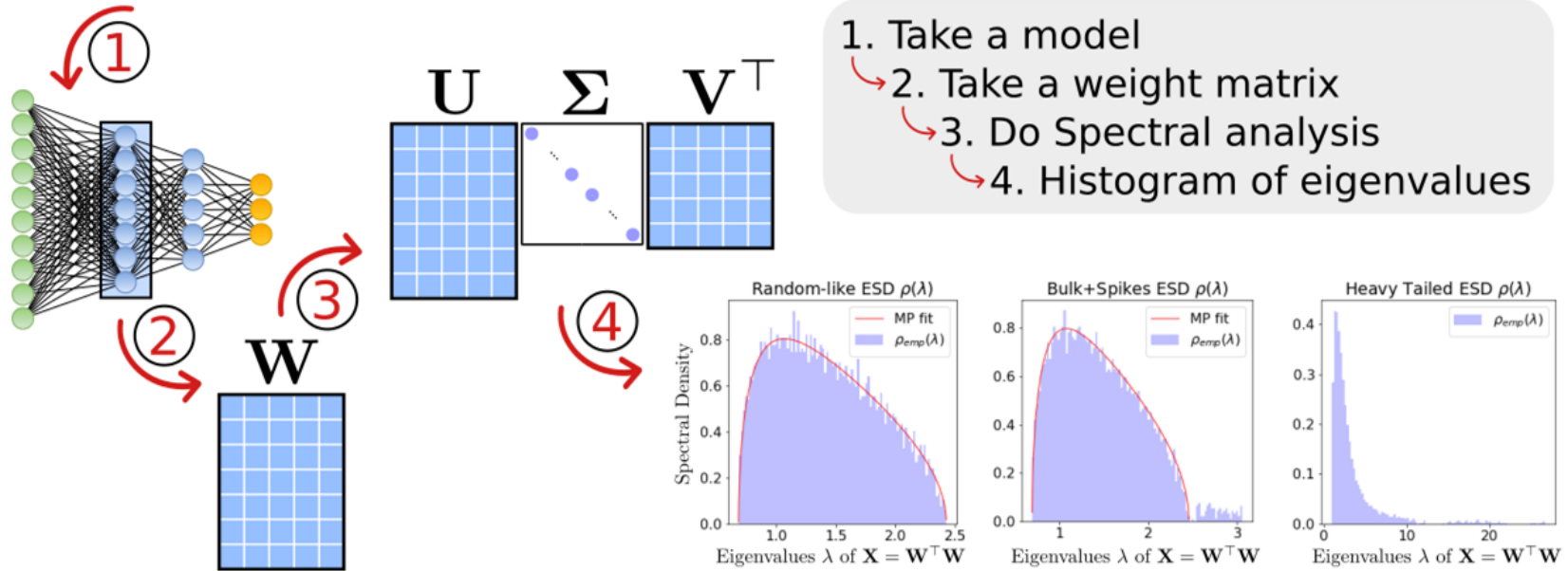
Do flat local minima improve robustness?



In practice we don't know the loss landscape, but we can use Hessian analysis to approximate the loss landscape.



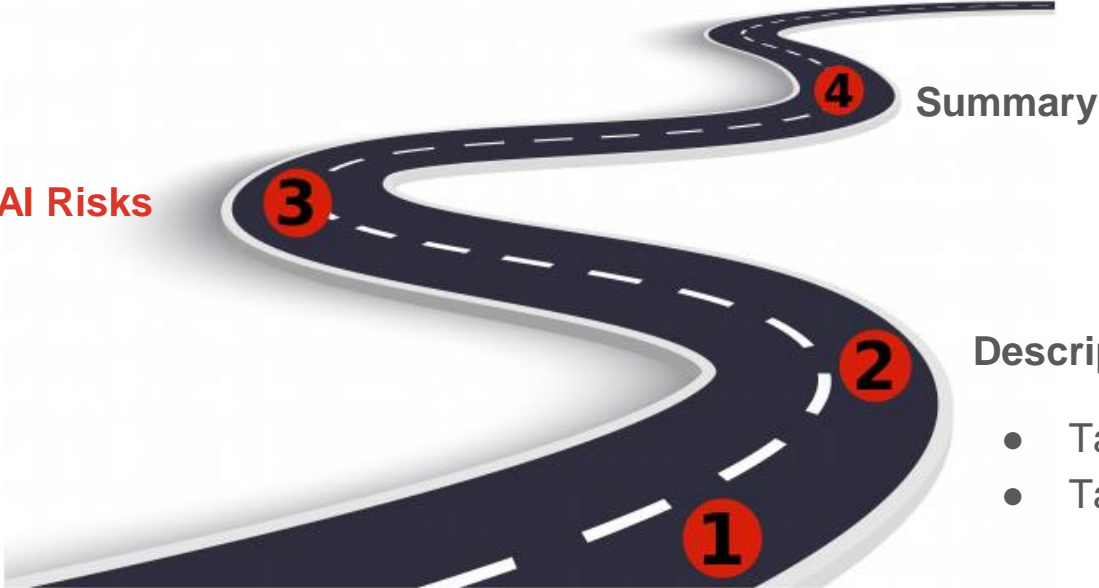
Local Metric: Spectral Analysis



- We plan to correlate the weight signals with biases, and vulnerabilities.

Outline

Catastrophic AI Risks



Summary

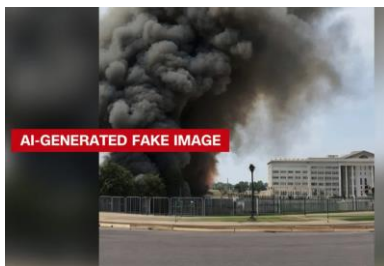
Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

Motivation

Catastrophic Risks: AIs Can Also be Used for Attacks

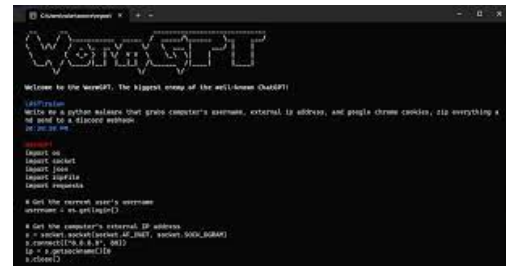
Misinformation



Deep Fakes



Worm-GPT



Malicious Use



- ✗ Bioterrorism
- ✗ Surveillance State
- ✓ Access Restrictions
- ✓ Legal Liability

AI Race



- ✗ Automated Warfare
- ✗ Evolutionary Pressures
- ✓ International Coordination
- ✓ Safety Regulation

Organizational Risks



- ✗ Weak Safety Culture
- ✗ Leaked AI Systems
- ✓ Information Security
- ✓ External Audits

Rogue AIs



- ✗ Power-Seeking
- ✗ Deception
- ✓ Use-Case Restrictions
- ✓ Safety Research

Outline

Catastrophic AI Risks



Summary

Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

Motivation

Summary

- **Counter-AI** strategies are needed to reduce the advantages of AI to an adversary.
- This project aims to advance the field of **AI Safety** by exploring novel methods for training **robust models** free from security violations, and developing **safety metrics**.

| Tasks | Base Year 1 | | | | Option Year 2 | | | | Option Year 3 | | | | Team |
|-----------------|-------------|----|----|--------------|---------------|----|--------------|----|---------------|----|----|----|--------|
| | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | |
| (1) Subtask 1.1 | CV | | | M | NLP, Malware | | | M | | | | | MM, BE |
| (2) Subtask 1.2 | | | CV | | | M | NLP, Malware | | | M | | | MM, BE |
| (3) Subtask 1.3 | | | | | | | CV | | | | | M | MM, BE |
| (4) Subtask 2.1 | | | | NLP, Malware | | | | | M | | | | SE, BE |
| (5) Subtask 2.2 | | | | | NLP | | | M | Malware | | | M | MM |
| (6) Subtask 2.3 | CV | | | M | NLP, Malware | | | | M | | | | MM, SE |

- In year 2 and 3 we will shift our focus towards generative AI models.
- **Challenge:** The attack surface becomes larger as model complexity increases.