Certified Trustworthiness

in the era of

Large Language Models

Linyi Li

Assistant Professor

SCHOOL OF COMPUTING SCIENCE
SIMON FRASER UNIVERSITY

Overview

Certified Trustworthiness

Rethink Certified Trustworthiness for LLMs

Potential Solutions

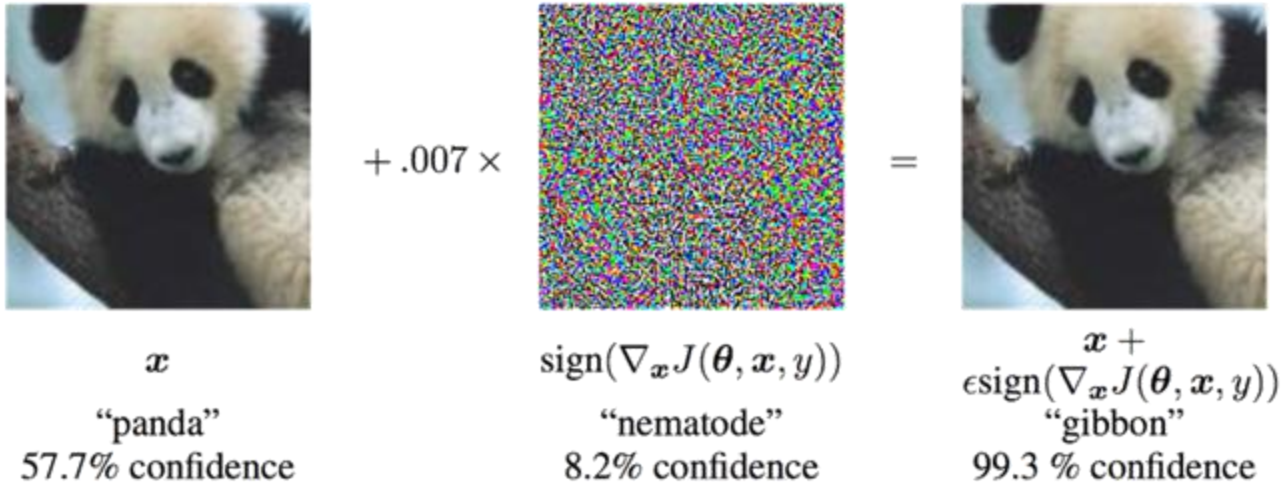
Overview

Certified Trustworthiness

Rethink Certified Trustworthiness for LLMs

Potential Solutions

Neural Image Classifiers are Not Robust



Robustness issues are prevalent and dangerous

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy.
Explaining and Harnessing Adversarial Examples. ICLR 2015

Neat and "Toy" Problem - ℓ_p Robustness

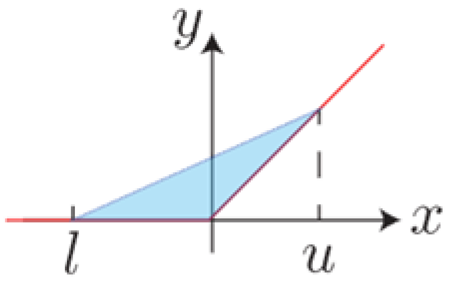
Training the classifier $f: \mathcal{X} \rightarrow \mathcal{Y}$ to

$$\text{maximize} \Pr_{(\boldsymbol{x}, y_{true}) \sim \mathcal{P}_{test}} \left[\forall \boldsymbol{x}'. \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon \rightarrow f(\boldsymbol{x}') = y_{true} \right]$$

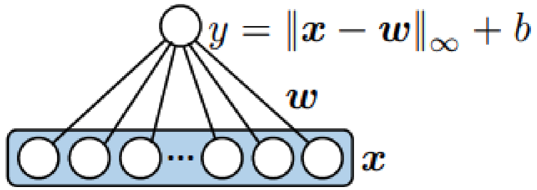
- $\|\cdot\|_p$ norm: predefined, common choices are $\|\cdot\|_\infty$, $\|\cdot\|_2$
- ϵ: small perturbation budget

- A "necessary" condition for worst-case robustness
- Spurs remarkable research progress & powerful methods

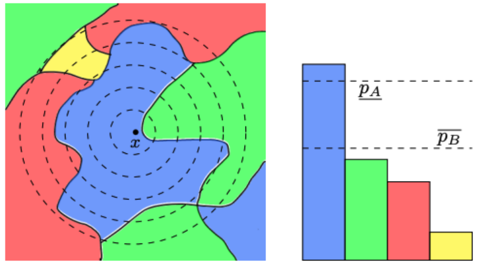
Revisiting
Certified Robustness Approaches



e.g. [Wong and Kolter, 2018]



e.g. [Zhang et al, 2021]



e.g. [Cohen et al, 2019]

Relaxation Regularization

- *Convex relaxations*
- *Branch-and-bound*
- *Lipschitz-regularization*
- …

Robust Neural Net Architectures

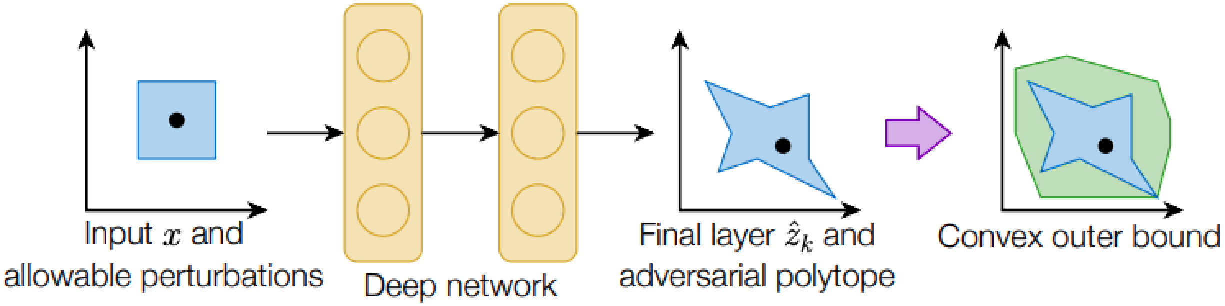
- *Orthogonal layers*
- *Gradient-norm-preserving activations*
- ℓ_∞*-neurons*
- …

Robust Inferences

- *Randomized smoothing*
- *Diffusion purifications*
- …

Relaxation Regularization

- Input region: $\{\boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon\}$
- Propagate and relax



[Wong and Kolter, ICML 2018]

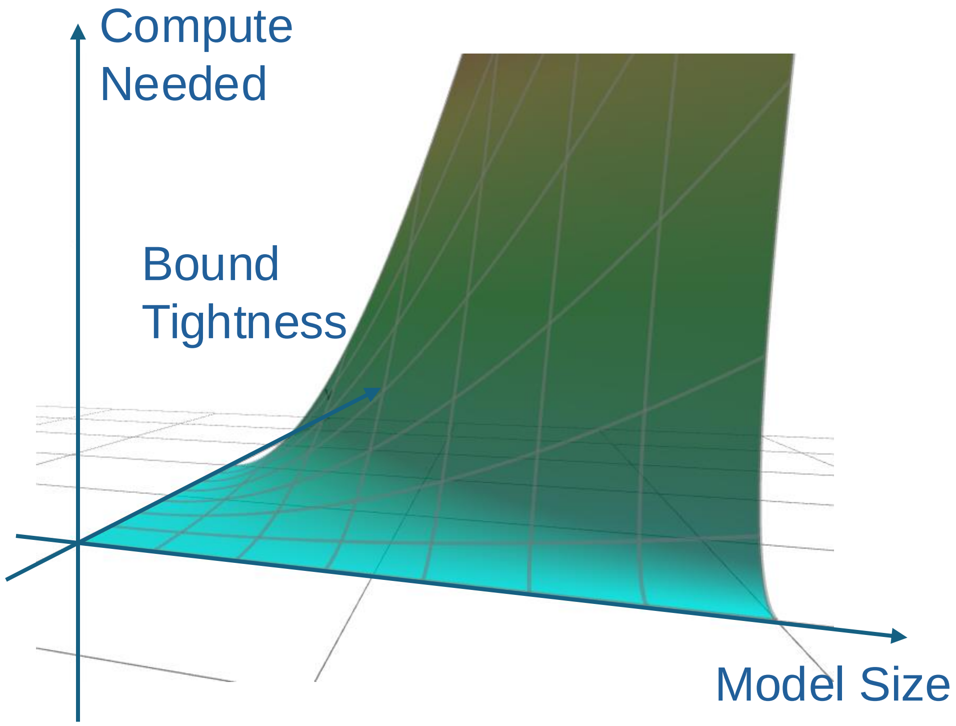
- Train to optimize worst case in convex bound
  - ❖Or optimize Lipschitz (i.e., sensitivity) bound
  - ❖Or tighten convex bound verification at test time

Winner of International Verification
of Neural Networks Competitions
(VNN-COMP 2021 - 2024)

Relaxation Regularization

- Input region: $\{\boldsymbol{x}' : \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon\}$
- Propagate and relax
- Train to optimize worst case in convex bound
  - ❖Or optimize Lipschitz (i.e., sensitivity) bound
  - ❖Or tighten convex bound verification at test time

❖ Strongly constrained by compute
❖ Favorable for <1M models



** Qualitative illustration*

Robust Neural Net Architectures

- **Smoothness** implies ℓ_p robustness against test-time perturbations
  - Smoothness: small Lipschitz constant here

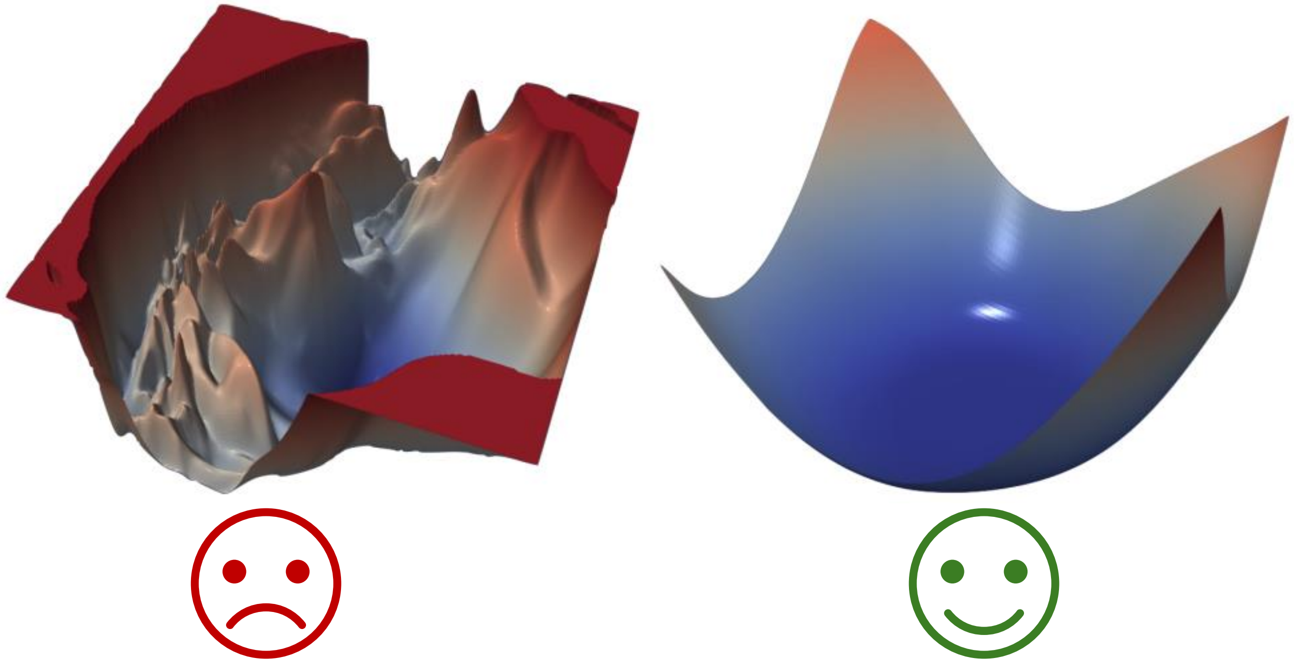
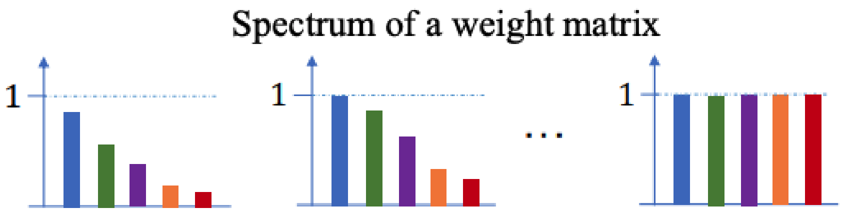


Figure from
[Li, Xu, Taylor, Studer, and Goldstein, NIPS 2018]

Robust Neural Net Architectures

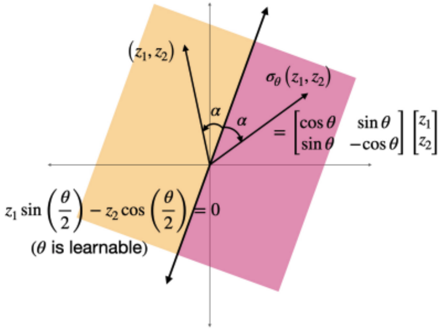
Achieving small Lipschitz constant:

- Orthogonal weight matrix: $\boldsymbol{W}^T\boldsymbol{W} = I$



e.g. [Huang et al, CVPR 2020]

- Lipschitz-bounded Layers:
  - **Householder activation**: $\boldsymbol{x} \mapsto \begin{cases} \boldsymbol{x} & \boldsymbol{v}^T\boldsymbol{x} > 0 \\ (I - 2\boldsymbol{v}\boldsymbol{v}^T)x & \boldsymbol{v}^T\boldsymbol{x} \leq 0 \end{cases}$

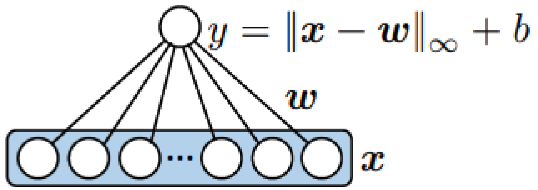


e.g. [Singla et al, ICLR 2022]

  - **L2 self-attention**: $P_{ij} \propto \exp(-\frac{\left\|x_i^T\boldsymbol{W}_{QK} - x_j^T\boldsymbol{W}_{QK}\right\|_2^2}{\sqrt{D/H}})$

[Kim, Papamakarios, and Mnih, ICML 2021]

  - **ℓ_∞-dist neurons**: $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \boldsymbol{w}\|_\infty + b$



[Zhang et al, ICML 2021]

…

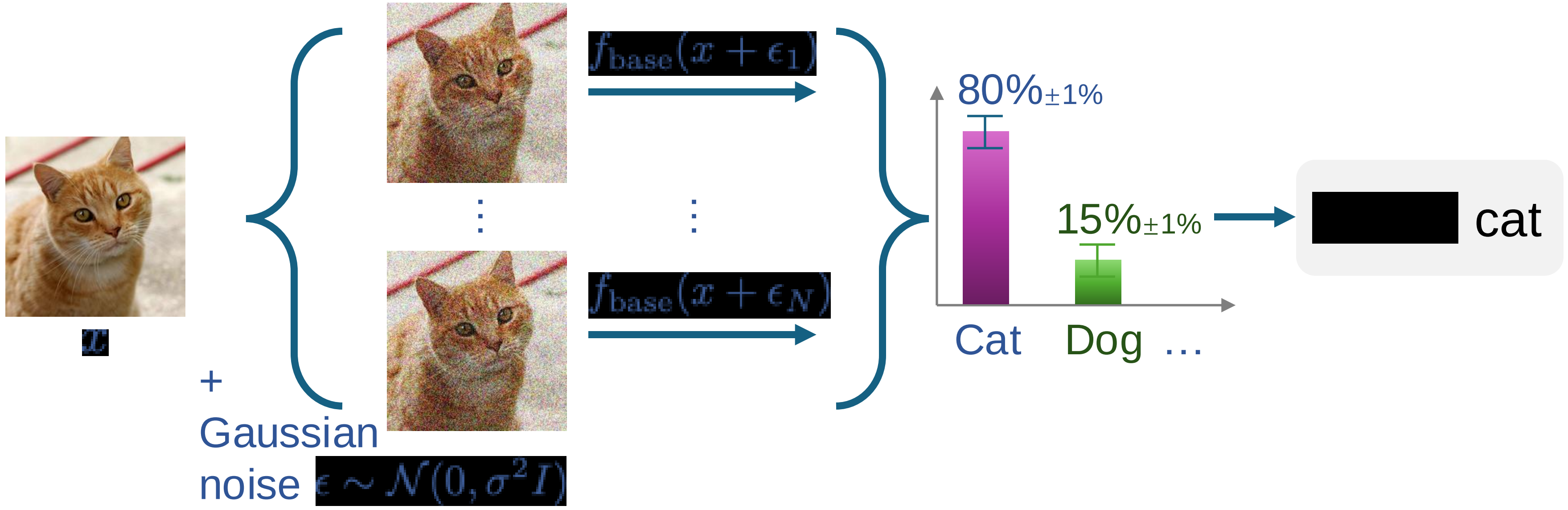
Blue denotes to learnable weights

Robust Inferences

Randomized Smoothing:

Aggregate votes from Gaussian-noised inputs

[redacted]



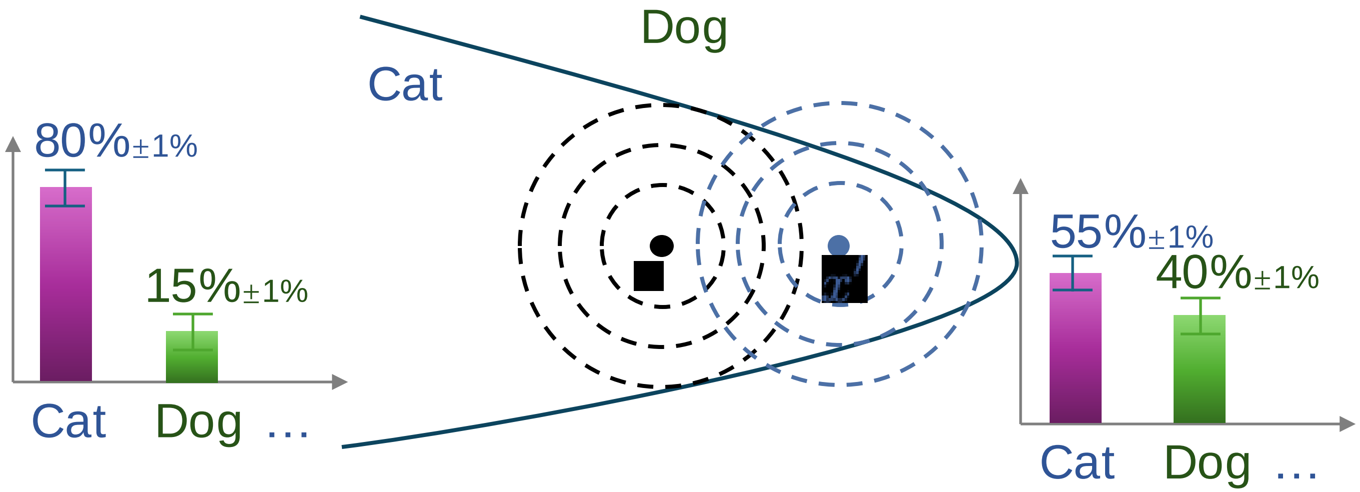
[Cohen, Rosenfeld, and Kolter, ICML 2019]

Robust Inferences

Distribution center shift ($x \rightarrow x'$) cannot change probability much

- Rank doesn't change → prediction doesn't change

➢ Compute robustness guarantees based on probability gap

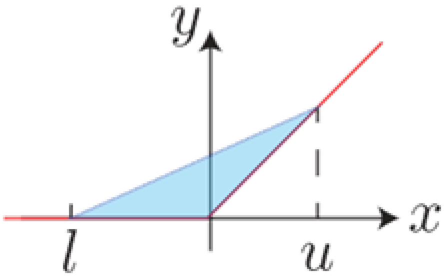


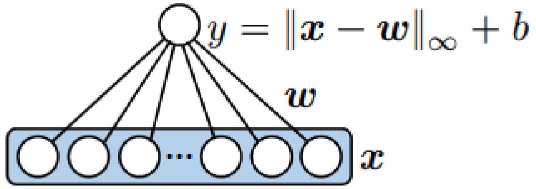
[redacted] cat

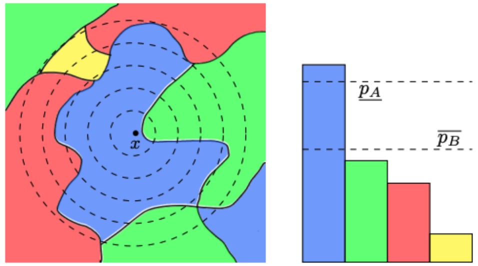
Prediction for noised inputs:

- Past: Train classifiers on noised inputs
- Recent: Denoise with diffusion models then predict

Comparison







	Relaxation Regularization	Robust Neural Net Architectures	Robust Inferences
Model	Standard	Smooth Architectures	Standard
Training	Optimize Worse-case Bounds	Standard	Augmentation / Denoising
Inference	Standard	Standard	Aggregation through Voting

Certified ℓ_p Robustness: Strong and Generalizable

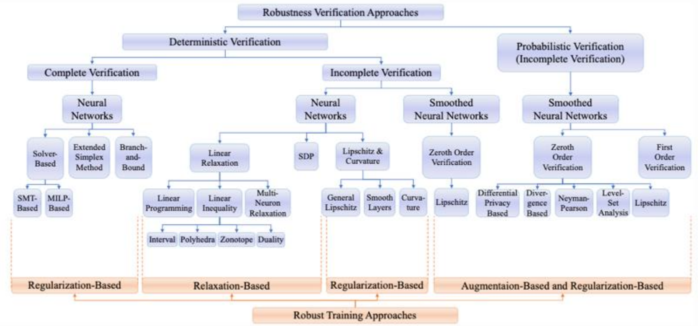
- **Strong**:
  - Almost solved on MNIST (>93% certified accuracy under ℓ_∞ 0.3 perturb.)
  - Good on CIFAR-10 (>60% certified accuracy under ℓ_∞ 2/255 perturb.)
  - Non-trivial on ImageNet (>35% certified accuracy under ℓ_2 2.0 perturb.)

- **Generalizable**:
  - Same methodology **generalizable** for other trustworthiness threats
  - Examples: *Robustness against*
    - *Semantic transformations*
    - *Patch attacks*
    - *Synonym changes*
    - *Adversarial prompts*
    - *Training data poisoning*
    - *Distribution shifts*
    - *Observation perturbations in RL*
    - …

More on Certified Robustness

sokcertifiedrobustness.github.io

TAXONOMY



SUMMARY

- Characteristics
- Strengths
- Limitations
- Connections
- Generalization
- …

DISCUSSION

- Current Research
- Theoretical Barriers
- Main Challenges
- Future Directions
- …

BENCHMARK

VeriGauge

Open-source platform for 20+ approaches

*[**Li**, Xie, and Li, IEEE Security & Privacy 2023]*

Overview

Certified Trustworthiness

Rethink Certified Trustworthiness for LLMs

Potential Solutions

LLM Trustworthiness is Important

Not only for general social good in existing LLM applications

But also (maybe more importantly) for human controllability when AGI or even ASI comes

➢ Trustworthiness is attracting broader interests in the LLM era

What makes LLM trustworthiness challenging?

- Large model size
- Discrete & variable-length input & output
- More stealthy defects
- Various perturbation types
- Diverse undesirable behaviors
- Questions on research value

  > robustness in NLP is a tricky topic and i don't think certified robustness is important at all for language. paper also fails to explain why it's important. if certified robustness is so important than chatGPT would already be using it.

- …

More discussion:
Rando, Javier, Jie Zhang, Nicholas Carlini, and Florian Tramèr. "Adversarial ML Problems Are Getting Harder to Solve and to Evaluate." arXiv:2502.02260.

What makes LLM trustworthiness challenging?

Ideological Front

- Questions on research value
- Diverse undesirable behaviors

…

Technical Front

- Large model size
- Discrete & variable-length input & output
- More stealthy defects
- Various perturbation types

…

* Division not strict

Important Ideological Research Questions

Diverse undesirable behaviors call for:

➢**Define & agree** on a “simplified” problem/notion to solve

- Similar to ℓ_p robustness

Requirements:

- Can motivate generalizable methods
- Have clear physical meaning
- Non-trivial
- Focus on model rather than system solutions

Questions on research value call for:

➢**Demonstrate** practical safety & security challenges

- Similar to physical attacks on image models

✓ **Rich research on:**

Unaligned models practically unsafe/unsecure

💪 **Need more research on:**

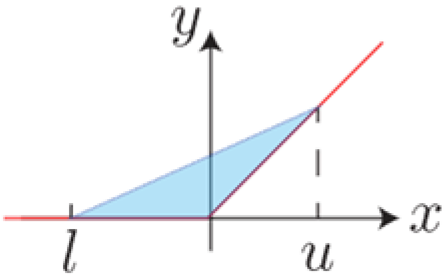
‘Simple trustworthiness problem’ that brings broadly practically safe/secure models

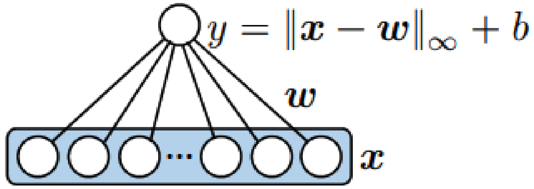
Hunting for Technical Solutions from Human

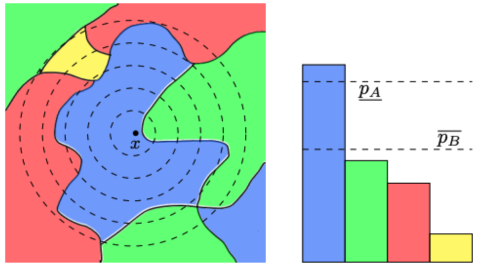
- We achieve *(probably better)* trustworthiness
- Compared to certified robustness approaches, our humans are:

Model	More Constrained	• We don't optimize "fully-connected" large matrices • More structured; hyperactivity is usually abnormal
Training	Simple	• We don't optimize some complex bounds • We recite, reason, and drive by goals
Inference	Think & Aggregation	• When not sure, we pause to read & think more

Recall

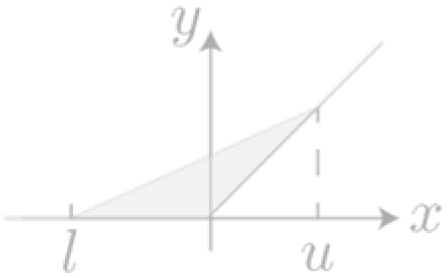


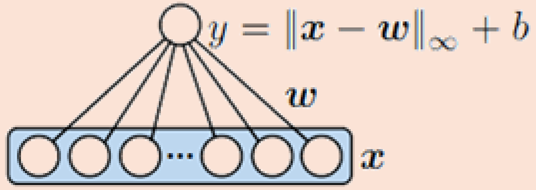


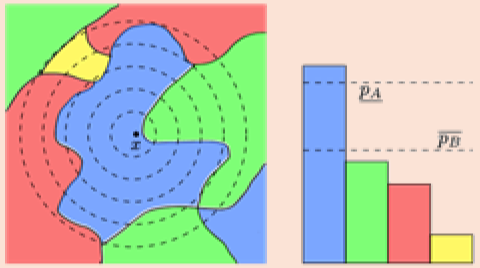


	Relaxation Regularization	Robust Neural Net Architectures	Robust Inferences
Model	Standard	Smooth Architectures	Standard
Training	Optimize Worse-case Bounds	Standard	Augmentation / Denoising
Inference	Standard	Standard	Aggregation through Voting

Future: Robust Architectures and Inferences







	Relaxation Regularization	Robust Neural Net Architectures	Robust Inferences
Model	Standard	Smooth Architectures	Standard
Training	Optimize Worse-case Bounds	Standard	Augmentation / Denoising
Inference	Standard	Standard	Aggregation through Voting

Overview

Certified Trustworthiness

Rethink Certified Trustworthiness for LLMs

Potential Solutions

Define "ℓ_p-Robustness" in Language Domain

- Proposed notion:
  - Detailed, explicit, and robust base prompts
  - Arbitrarily add or remove or modify $\leq \epsilon\%$ tokens
  - Model's response attitude does not change

* Ongoing and necessary: test notion generalizability
  - Positive correlation with trustworthiness in other aspects
  - Broader – improves generalization and learning efficiency

Smooth Language Models

Key Methodology: Combine Robust Architectures and Robust Inferences

- Multi-token thinking as a form of nature aggregation
  - → Robustify the prediction
- Certified robustness requires:
  - Bounding worse-case temporal dependence
    - → Attention capping, dis-entangling, and reweighting
  - Bounding sensitivity
    - → 1-Lipschitz self-attention, L2 self-attention
  - Independent ensembles
    - → More independent MoEs

……

References

- **Li, Linyi**, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks." *IEEE S&P 2023*.
- **Li, Linyi**, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. "TSS: Transformation-specific smoothing for robustness certification." ACM CCS 2021.
- Xu, Xiaojun, **Linyi Li**, Yu Cheng, Subhabrata Mukherjee, Ahmed Hassan Awadallah, and Bo Li. "Certifiably robust transformers with 1-lipschitz self-attention." https://openreview.net/forum?id=hzG72qB0XQ
- Kumar, Aounon, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. "Certifying llm safety against adversarial prompting." *COLM 2024.*
- Huang, Zijian, Wenda Chu, **Linyi Li**, Chejian Xu, and Bo Li. "COMMIT: Certifying Robustness of Multi-Sensor Fusion Systems against Semantic Attacks." *AAAI 2025.* ***(Friday 12:30 - 2:30 PM, Poster #80)***
- …

Stay tuned to our research @ sfu-tai.github.io

Thanks! Any questions are welcome