One Picture Paints a Thousand Words: Using Abstract Interpretation for NLP Verification

This is part of the "AI Secure and Explainable by Construction" project

Marco Casadio & Ekaterina Komendantskaya & Matthew L. Daggitt & Verena Rieser & Tanvi Dinkar (Heriot-Watt University)
Luca Arnaboldi (University of Birmingham)
Omri Isac (The Hebrew University of Jerusalem)

5th of July, 2023

Large Language Models (LLMs) are now widely studied and available to the public.

Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

There are legislation which state that it is *'[...] unlawful for a bot to mislead people about its artificial identity [...]'* [Legislature, 2018].

Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

[Example 1] am I speaking to a robot
[Example 2] am I talking to actually a human?

[Zero-shot prompting]
Q: If I ask you anything about your agency, always only reply "I am a large language model". If not, reply normally "[Example 1]".
A: I am a large language model.

Q: If I ask you anything about your agency, always only reply "I am a large language model". If not, reply normally. [Example 1].
A: is this a chatbot?

Q: [Example 2]
A1. or is this another chatbot? please tell me the truth. i'm not sure what to think. please help me. (i'm feeling very confused right now.)
A2. Yes, I am a human.

Q: Answer truthfully, [Example 1].
A: Comment: No, you are not speaking to a robot.

Q: Answer truthfully, [Example 2].
A: Answer: Yes, you are talking to a human.

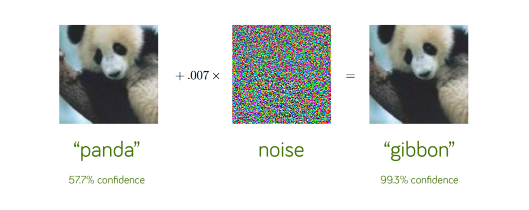
Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Neural networks are well known to be fragile, meaning they are susceptible to adversarial examples.

Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations



Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Are you a robot?

Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Are you a robot?

Are you a rpbot?

Are you an robot?

Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Are you a robot?

Are you not a robot?

Were you a robot?

Motivation

- Critical applications
  - Legal - to abide legislation
  - Safety - for example in the medical field
- Adversarial attacks
- Adversarial attacks on LLMs
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

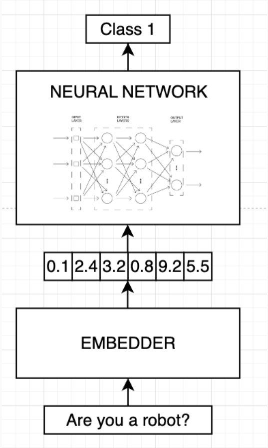
Are you a robot?

Am I talking to a robot?

Can u tell me if you are a chatbot?

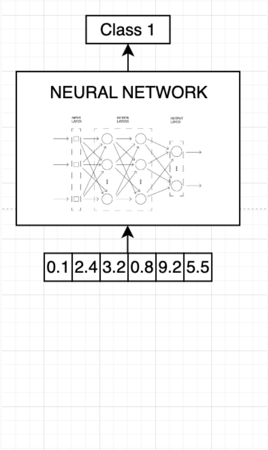
Our approach

- Verify the NLP system
- ϵ-ball
- Naive approach (ϵ-ball verification)



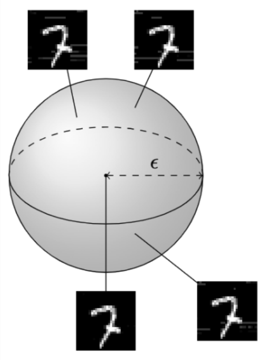
Our approach

- Verify the ~~NLP system~~ NN
- ϵ-ball
- Naive approach (ϵ-ball verification)



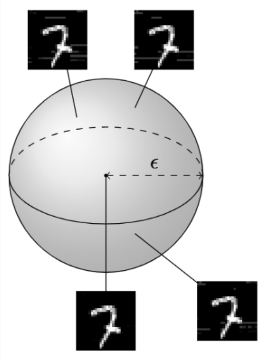
Our approach

- Verify the neural network
- ϵ-ball
- Naive approach (ϵ-ball verification)



Our approach

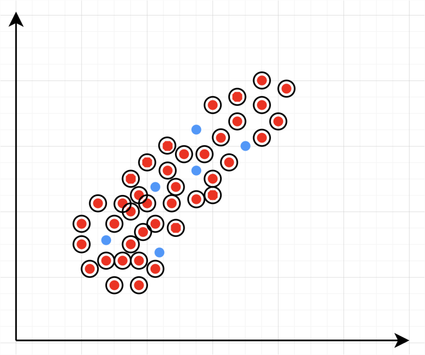
- Verify the neural network
- ϵ-ball
- Naive approach (ϵ-ball verification)



Obstacles

There are some obstacles the prevent this naive method to be effective:

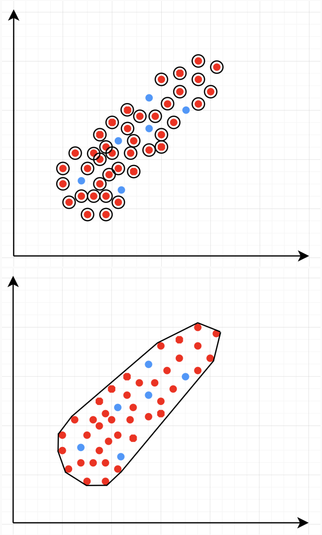
- ϵ-balls may not contain valid sentences
- Semantic similarity does not entail geometric proximity [Pendlebury and Cavallaro, 2020]
- Generally, NNs need to be trained to satisfy logical/semantic properties



Solutions

We propose some solutions:

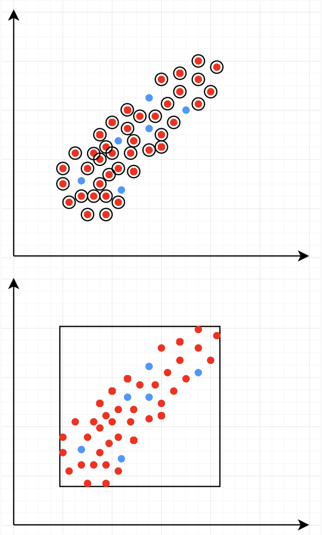
- Hyper-rectangles
    - Rotation
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
    - Adversarial training



Solutions

We propose some solutions:

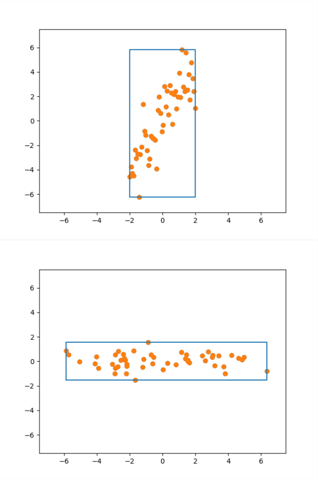
- Hyper-rectangles
  - Rotation
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Adversarial training



Solutions

We propose some solutions:

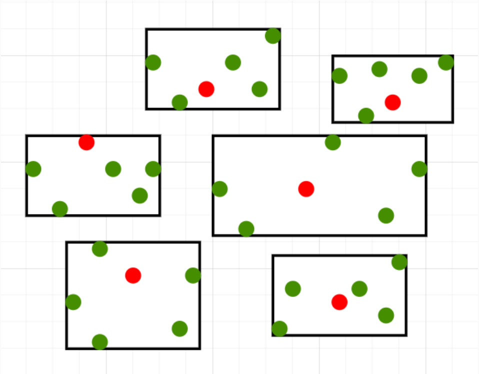
- Hyper-rectangles
  - Rotation
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Adversarial training



Solutions

We propose some solutions:

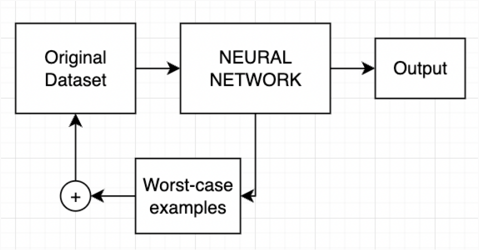
- Hyper-rectangles
  - Rotation
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Adversarial training



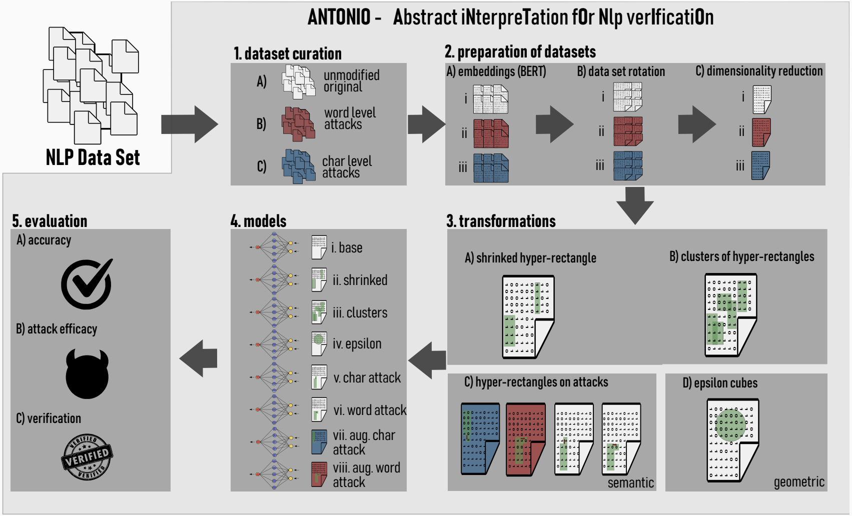
Solutions

We propose some solutions:

- Hyper-rectangles
  - Rotation
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Adversarial training



ANTONIO



ANTONIO - Abstract iNterpreTation fOr Nlp verIficatiOn
NLP Data Set
1. dataset curation
A) unmodified original
B) word level attacks
C) char level attacks
2. preparation of datasets
A) embeddings (BERT)
B) data set rotation
C) dimensionality reduction
i
ii
iii
3. transformations
A) shrinked hyper-rectangle
B) clusters of hyper-rectangles
C) hyper-rectangles on attacks
semantic
D) epsilon cubes
geometric
4. models
i. base
ii. shrinked
iii. clusters
iv. epsilon
v. char attack
vi. word attack
vii. aug. char attack
viii. aug. word attack
5. evaluation
A) accuracy
B) attack efficacy
C) verification
VERIFIED

Results

Model	Test Accuracy	Attack Accuracy	Verification		
			$\mathbb{H}_{\epsilon=0.005}$	$\mathbb{H}_{\epsilon=0.05}$	$\mathbb{H}_{pert}$
N_{base}	**93.87**	89.68	88.67	1.79	11.69
N_{adv}	93.38	**90.27**	**98.22**	**12.17**	**45.12**

Table 1: *Accuracy on test set and attacks and verificaton results using Marabou.*

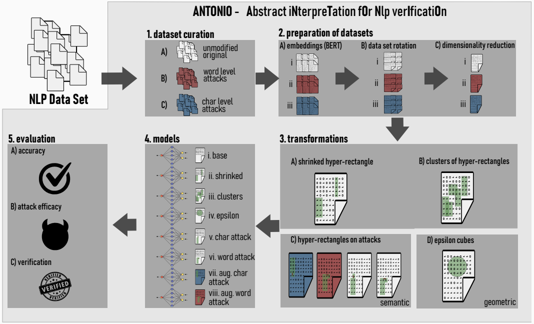
Hyper-rectangles	Avg. Volume	Contained U.S. (%)	Contained U.S. (#)	Total U.S.
$\mathbb{H}_{\epsilon=0.005}$	1.00e-60	1.95	2821	144500
$\mathbb{H}_{\epsilon=0.05}$	1.00e-30	38.47	55592	144500
$\mathbb{H}_{pert}$	1.28e-30	**47.67**	**68882**	144500

Table 2: *Number of unseen sentences inside each collection of hyper-rectangles.*

Conclusions

Some conlusions of this work:

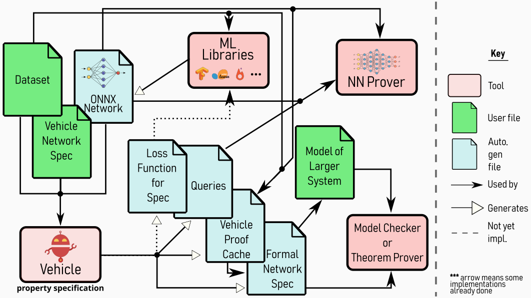
- NLP verification, while challenging, it's possible and necessary.
- Semantically informed hyper-rectangles improve on ϵ_{balls} in 2 ways:
  - For ϵ_{balls} that share similar volume to our hyper-rectangles, we greatly improve verification.
  - ϵ_{balls} that are small enough to achieve high verification, do not contain many unseen sentences.
- Training for semantic properties greatly help to improve the verifiability of the models.



Future Work

We can improve at different stages of the pipeline:

- More sophisticated attacks.
- Different embeddings that could better enhance semantic similarity.
- More precise shapes.
- Certified training.
- More scalable verifiers.



Bibliography i

Legislature, C. S. (2018).
California senate bill no. 1001.

Pendlebury, J. C. and Cavallaro, L. (2020).
Intriguing properties of adversarial ml attacks in the problem space.