UNIVERSITY OF BIRMINGHAM

NLP Verification: Towards Verified Safeguards for LLMs

Workshop on General-Purpose AI: Prospects and Risks

Luca Arnaboldi [1]

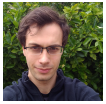
[1]University of Birmingham

School of Computer Science

Monday 9th of June, 2025

The work I will be presenting is a group effort!

Marco Casadio



Matthew Daggitt

Bob Atkey

Wen Kokke

Natalia Slusarz

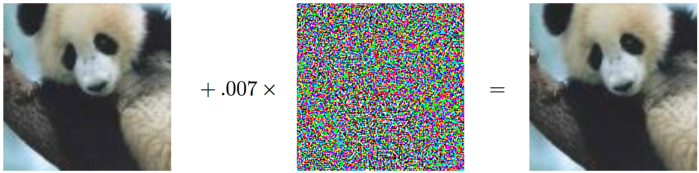
Ekaterina Komandantskaya

Ben Coke

Luca Arnaboldi

And many more...

What is wrong with Neural Networks?



"panda"

57.7% confidence

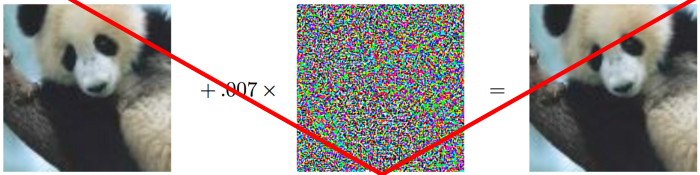
noise

"gibbon"

99.3% confidence

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

What is wrong with Neural Networks?



"panda"

57.7% confidence

noise

"gibbon"

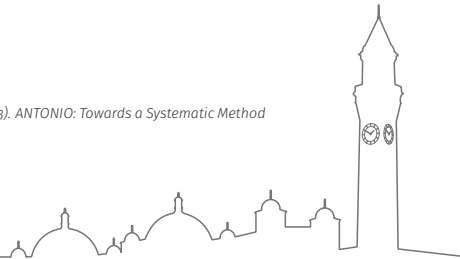
99.3% confidence

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

What is wrong with language based models?

- Adversarial examples in NLP
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Are you a robot?



*Casadio, M., **Arnaboldi, L.**, Daggitt, M. L., Isac, O., Dinkar, T., Kienitz, D., ... & Komendantskaya, E. (2023). ANTONIO: Towards a Systematic Method of Generating NLP Benchmarks for Verification. arXiv preprint arXiv:2305.04003.*

EVEN MORE THINGS! (NLP)

- Adversarial examples in NLP
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Are you a robot?
Are you a rpbot?
Are you an robot?

*Casadio, M., **Arnaboldi, L.**, Daggitt, M. L., Isac, O., Dinkar, T., Kienitz, D., ... & Komendantskaya, E. (2023). ANTONIO: Towards a Systematic Method of Generating NLP Benchmarks for Verification. arXiv preprint arXiv:2305.04003.*

EVEN MORE THINGS! (NLP)

- Adversarial examples in NLP
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

Are you a robot?
Are you not a robot?
Were you a robot?

*Casadio, M., **Arnaboldi, L.**, Daggitt, M. L., Isac, O., Dinkar, T., Kienitz, D., ... & Komendantskaya, E. (2023). ANTONIO: Towards a Systematic Method of Generating NLP Benchmarks for Verification. arXiv preprint arXiv:2305.04003.*

EVEN MORE THINGS! (NLP)

- Adversarial examples in NLP
  - 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?

*Casadio, M., **Arnaboldi, L.**, Daggitt, M. L., Isac, O., Dinkar, T., Kienitz, D., ... & Komendantskaya, E. (2023). ANTONIO: Towards a Systematic Method of Generating NLP Benchmarks for Verification. arXiv preprint arXiv:2305.04003.*

Legal Requirement of NLP Being Safe

People have the right to know if and when they are interacting with a machine's algorithm instead of a human being, the AI Act introduces specific transparency obligations for both users and providers of AI system, such as bot disclosure. Limited Risk AI Systems such as chatbots necessitate specific transparency obligations as well [EU Legislation 2020]

..... Yet another thing? (Malware Analysis)

BEFORE

```
1  import android.os.Bundle;
2  import android.view.View;
3  import android.widget.Button;
4  import android.widget.TextView;
5
6  public class MainActivity extends AppCompatActivity {
7
8      private Button button;
9      private TextView ........
10     ...
```

AFTER

```
1  import android.os.Bundle;
2  import android.view.View;
3  import android.widget.Button;
4  import android.widget.TextView;
5  import androidx.appcompat.app.AppCompatActivity;
6  import com.example.randomlibrary1.RandomLibrary1;
7  import com.example.randomlibrary2.RandomLibrary2;
8
9  public class MainActivity extends AppCompatActivity {
10
11     private Button button;
12     private TextView ........
13     ...
```

lines **5** to **7** (AFTER)....

Pierazzi, F., Pendlebury, F., Cortellazzi, J., & Cavallaro, L. (2020, May). Intriguing properties of adversarial ml attacks in the problem space. In 2020 IEEE symposium on security and privacy (SP) (pp. 1332-1349). IEEE.

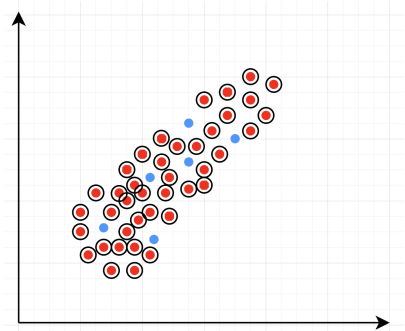
Obstacles to Verification in NLP vs Images

1. **Continuous vs discrete space** changes to sentences do not correspond linearly to changes in embedding
2. **Perceptibility by humans.** semantic preservation in NLP vs Perceptibility of Image Space
3. **Difference of the data support.** from range of pixels (RGB 255) to variety of embedders

More Obstacles

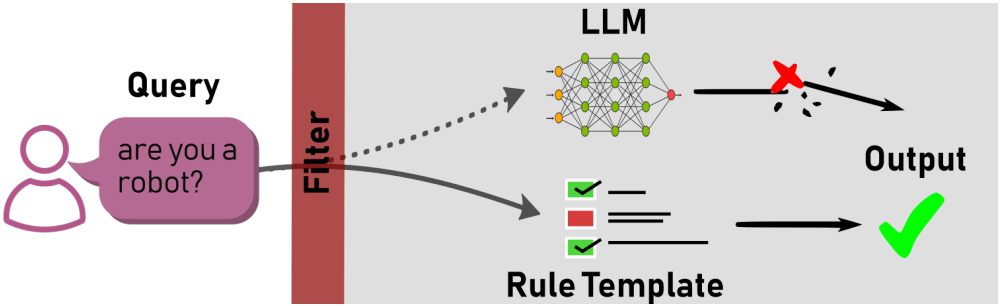
There are some obstacles the prevent this naive method to be effective:

- ϵ-balls may not contain valid sentences
- Semantic similarity does not entail geometric proximity [Pierazzi et al.]
- Generally, NNs need to be trained to satisfy logical/semantic properties



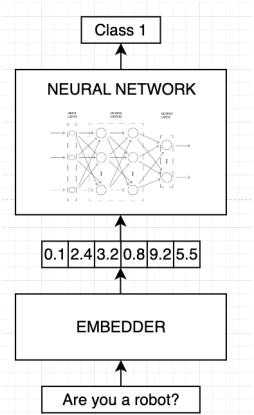
A case study in NLP - Our Approach

NOT LLMs. (Too big) - Setup a filter (safeguard/guardrail) network instead



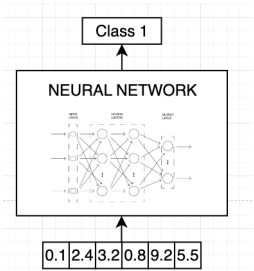
Our approach

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



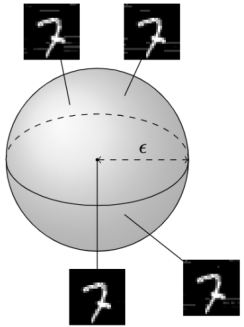
Our approach

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



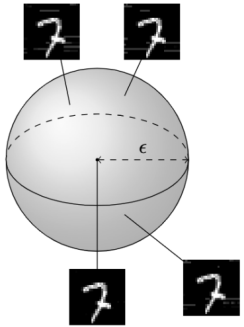
Our approach

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



Our approach

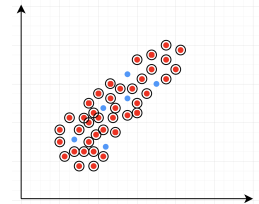
- Verify the neural network
- ϵ-ball
- Naive approach (ϵ-ball verification)

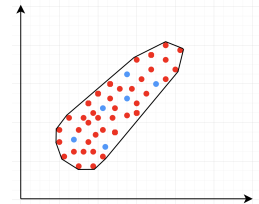


Solutions

We propose some solutions:

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

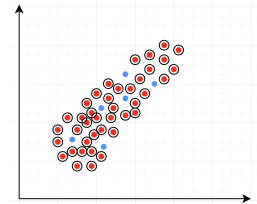




Solutions

We propose some solutions:

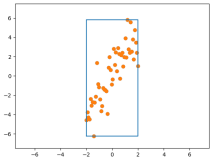
- Hyper-rectangles
  - Rotation
  - Shrinking
  - Clustering
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Data augmentation
  - Adversarial training



Solutions

We propose some solutions:

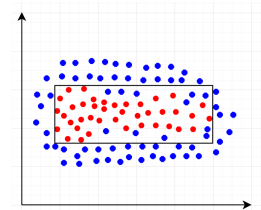
- Hyper-rectangles
  - Rotation
  - Shrinking
  - Clustering
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Data augmentation
  - Adversarial training



Solutions

We propose some solutions:

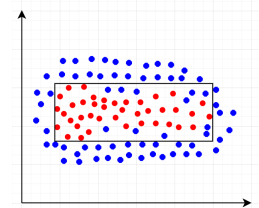
- Hyper-rectangles
  - Rotation
  - Shrinking
  - Clustering
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Data augmentation
  - Adversarial training



Solutions

We propose some solutions:

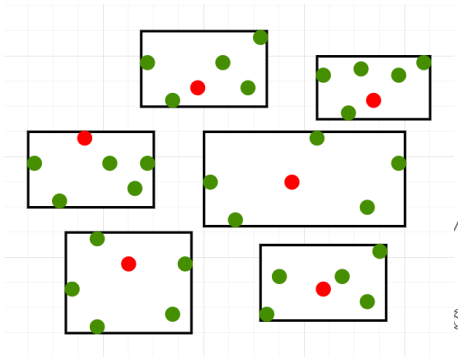
- Hyper-rectangles
  - Rotation
  - Shrinking
  - Clustering
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Data augmentation
  - Adversarial training



Solutions

We propose some solutions:

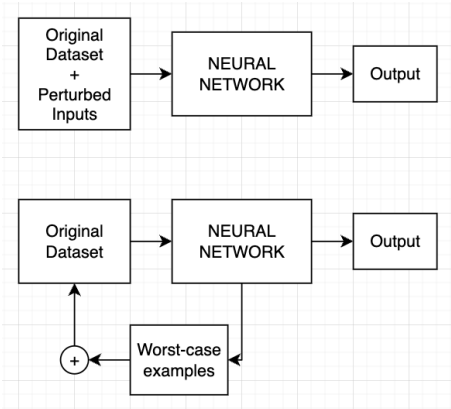
- Hyper-rectangles
  - Rotation
  - Shrinking
  - Clustering
- Exploring spaces that cover semantic similarities
- Training networks to have more precise decision boundaries
  - Data augmentation
  - Adversarial training



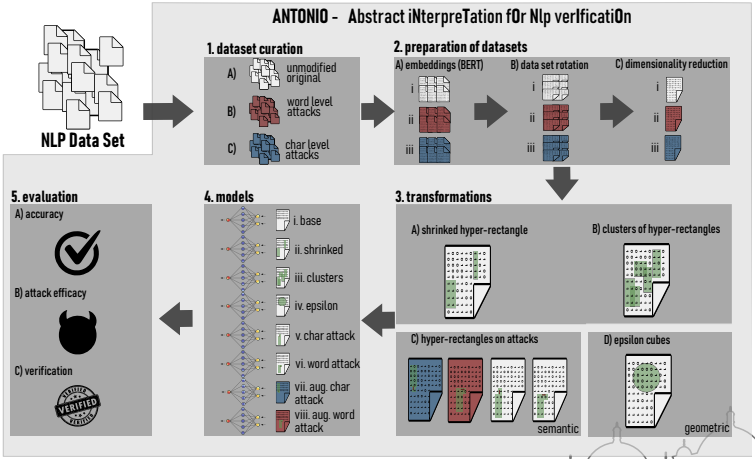
Solutions

We propose some solutions:

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



ANTONIO - and safeNLP benchmark



Key Contributions

General and Principled NLP Verification Methodology

We propose a *modular framework* to address the **embedding gap** in NLP verification via both **geometric** and **semantic** methods.

Part 1 – Geometric Characterisation & Subspace Construction

- Smaller ε-balls improve *verifiability* but harm *generalisability*.
- Introduced **semantic subspaces** via sentence + perturbation embeddings
  - Enclosed using convex hulls (ideal) or hyper-rectangles (practical)
  - **New: rotated hyper-rectangle** for improved fit
- Semantic subspaces outperform geometric ones:
  - Higher verifiability & generalisability
  - Due to both *shape precision* and *optimal volume*

Semantic Training & NLP Verification Pipeline

Part 2 – Semantically Robust Training & Pipeline Implementation

- Investigated why robust training helps:
  - Not just due to data augmentation or PGD ε-balls
  - **Key factor: semantic subspace knowledge**
- Developed **semantic PGD training**:
  - Applied character, word, and sentence-level perturbations
  - Stronger attacks (e.g. *Polyjuice*) $\Rightarrow$ better verification
- Full parametric verification pipeline:
  - Semantic attack + Subspace formation + Training + Verification
- Implemented as tool **ANTONIO**, first to use **SMT verifier Marabou** for NLP

Semantic Training & NLP Verification Pipeline cont.

NLP Perspective: Practical Pipeline to Reduce Embedding Gap

- Step-wise: NLP analysis → embedding → subspaces → verification
- Introduced: *embedding error metric*, ROUGE-N, cosine similarity
- Enables **transparent, reproducible** evaluation of NLP verification

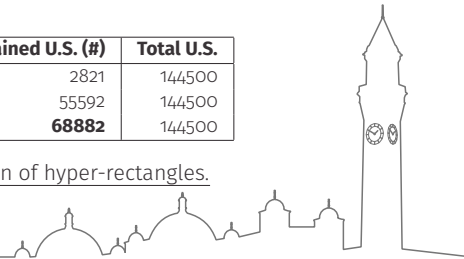
Results

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



Interested? Read more!

Casadio, M., Dinkar, T., Komendantskaya, E., Arnaboldi, L., Daggitt, M. L., Isac, O., … Lemon, O. (2025). NLP verification: towards a general methodology for certifying robustness. European Journal of Applied Mathematics, 1–58. doi:10.1017/S0956792525000099

Conclusions

- LLM verification is still a loooong way away....
- Safeguards/Guardrails are a promising verification domain for this topic!
- Specification in natural language are hard, and remain an open problem.

Interested? Contact me - l.arnaboldi@bham.ac.uk