Neural Network Verification With Vehicle: Chapter 5 - Application Areas and Conclusions

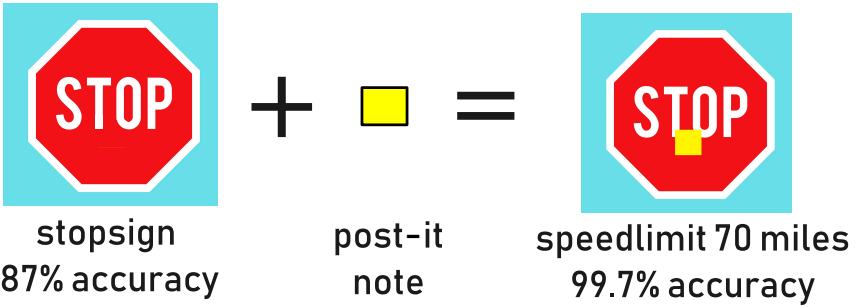
VeTSS Summer School'23

Luca Arnaboldi[1] Ekaterina Komendantskaya[2] Matthew Daggitt (online) [3]

[1]University of Birmingham · [2]University of Southampton · [3]Heriot-Watt University

graphics/artwork-logo

Some More Reasons to Verify AI... (Cars)



Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

EVEN MORE REASONS! (NLP)**

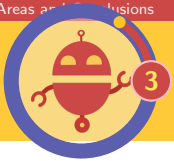
- Adversarial examples in NLP
  - Character perturbations
  - Word perturbations
  - Sentence perturbations

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.

With slide contributions from M. Casadio (Thanks)

EVEN MORE REASONS! (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 REASONS! (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 REASONS! (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 Verification

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 one? (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)....

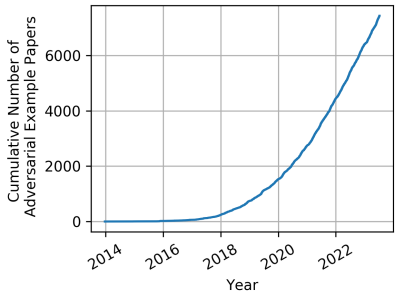
Pierazzi, F., Pendlebury, F., Cortellazzi, J., & Cavallaro, L. (2020, May). Intriguing properties of adversarial ml attacks in the problem space. In 2020 IEEE symposium on security and privacy (SP) (pp. 1332-1349). IEEE.

OK - I promise last one! (ML Network IDS)

- Identification fields: Src IP, Src Port, Dst IP,Dst Port, Protocol,Timestamp
- Features: Flow Duration, Fwd/Bwd Header Length, (Fwd/Bwd) Packet Length Min/Max/Mean/Std/Total, Total Fwd/Bwd Packets, (Fwd/Bwd) Inter- Arrival Time Min/Max/Mean/Std/Total, (Fwd/Bwd) SYN/FIN/ACK/RST/CWR/PSH/URG/ECE flags count, Packets/second, Bytes/second, Flow Active Duration Min/Max/Mean/Std, Subflow (Fwd/Bwd) Packets/Bytes, Up/Down Ratio
- Label: FlowType (should be mapped to 0 - BENIGN or 1 - MALICIOUS)
- **Attacker Objective:** Can packets be manypulated in such a way that the classifcation switches?

Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., & Colajanni, M. (2022). Modeling realistic adversarial attacks against network intrusion detection systems. Digital Threats: Research and Practice (DTRAP), 3(3), 1-19.

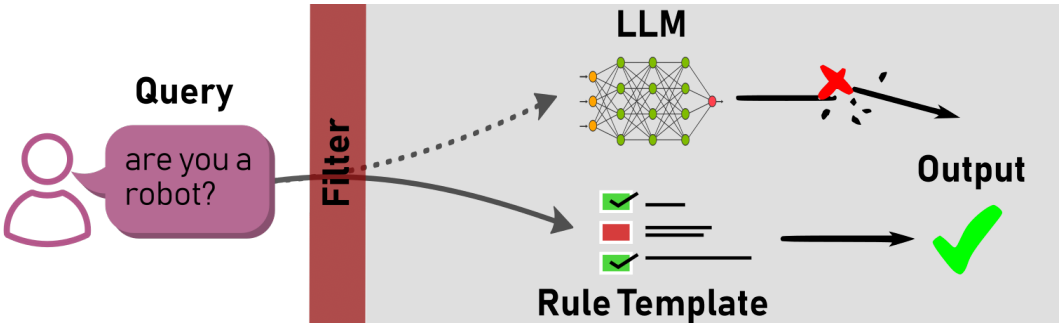
Summary so far



- Adversarial attacks are here to stay
- Verification is a promising way to protect against them
- We have a tool to specify properties and verify them
- So what are the open problems?
- .... **Remember NLP? (Malware, Text, Dialogue etc....**

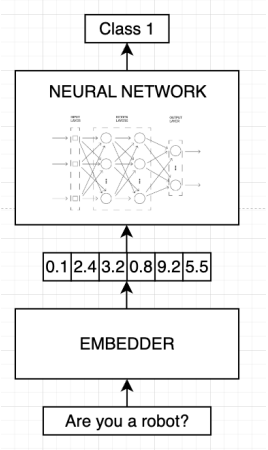
NLP Verification - Our approach

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



NLP Verification - Our approach

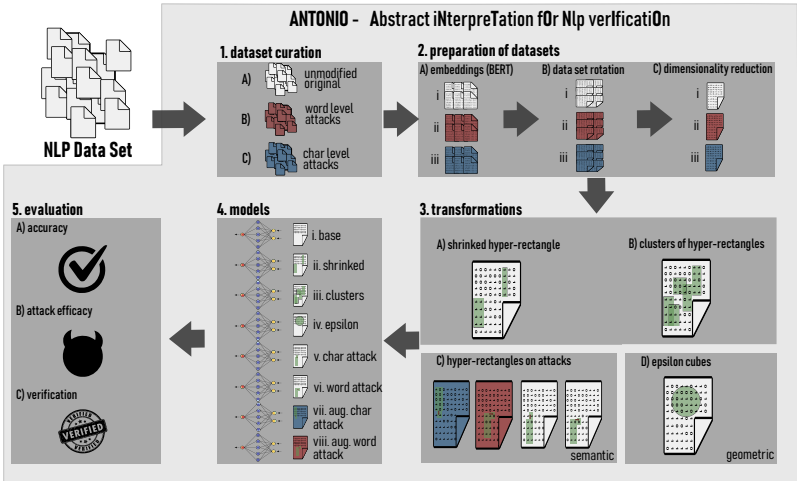
- Verify the NLP system
- ϵ-ball
- Naive approach (ϵ-ball verification)



NLP Verification - Obstacles to Verification

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

NLP Verification - ANTONIO



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.

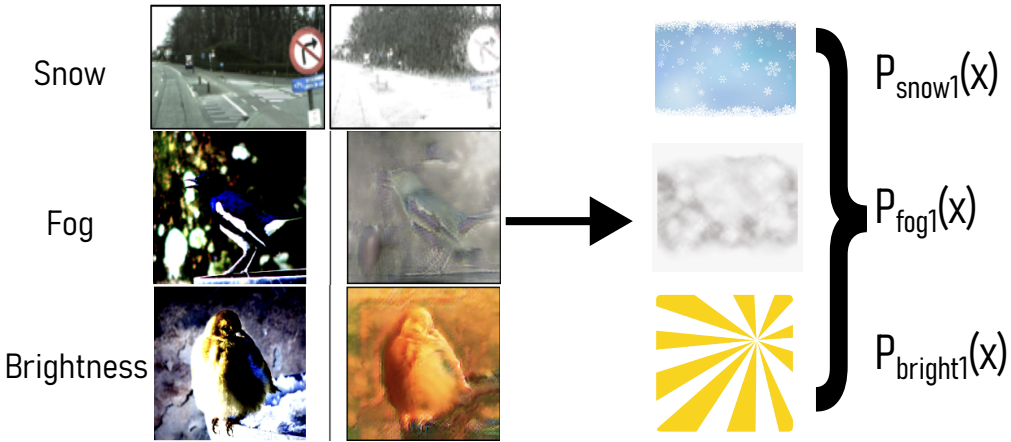
Vehicle Sensor Verification - Reminder

Definition of Verification for a Black Box Model

For a neural network $N : \overline{x} \to \overline{y}$, the input property $P(\overline{x})$ and the output property $Q(\overline{y})$, does there exist an input $\overline{x_0}$ which satisfies $P(\overline{x_0})$ such that its corresponding output $\overline{y_0}$ satisfies $Q(\overline{y_0})$?

- $P(\overline{x})$ **characterises inputs checked**
- $Q(\overline{y})$ characterises the behaviour we DO NOT wish for
- if satisfied, counterexample is returned, else property holds
- the P for traditional adversarial robustness is $\|\overline{x} - \overline{x_0}\| L_\infty \leq \delta$
- the Q is, $\bigvee_i (\overline{y}[i_0] \leq \overline{y}[i])$, where $\overline{y}[i_0]$ is the desired label

Formal Verification of ML/Sensors - For Resilient Autonomy



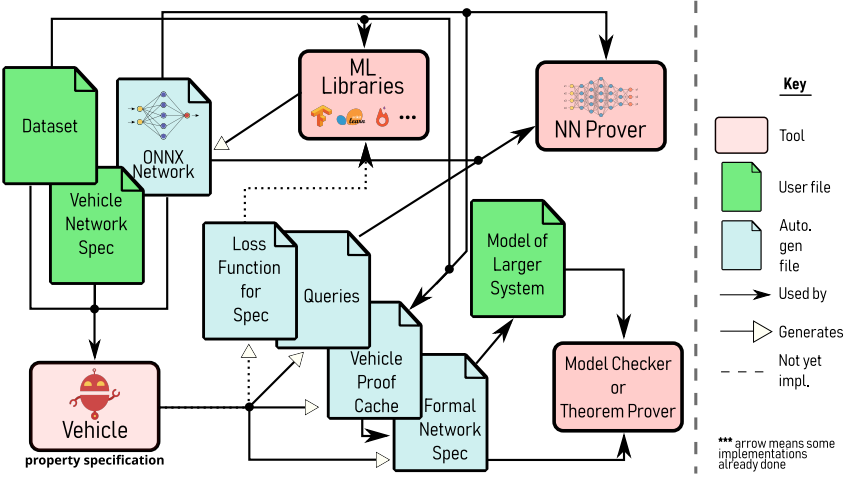
Formally Verified IDS Systems

- Identification fields: Src IP, Src Port, Dst IP,Dst Port, Protocol,Timestamp
- Features: Flow Duration, Fwd/Bwd Header Length, (Fwd/Bwd) Packet Length Min/Max/Mean/Std/Total, Total Fwd/Bwd Packets, (Fwd/Bwd) Inter- Arrival Time Min/Max/Mean/Std/Total, (Fwd/Bwd) SYN/FIN/ACK/RST/CWR/PSH/URG/ECE flags count, Packets/second, Bytes/second, Flow Active Duration Min/Max/Mean/Std, Subflow (Fwd/Bwd) Packets/Bytes, Up/Down Ratio
- Label: FlowType (should be mapped to 0 - BENIGN or 1 - MALICIOUS)
- **Objective:** Given an attacker can perturb these, can we still correctly classify benign and malign traffic?

Panchuk, B., Arnaboldi, L., Daggitt, L., M., & Letychevskyi, O. (2023, Coming Soon). Formal Verification of ML Based Network Intrusion Detection. Work in progress

Vehicle-Tool

One specification, multiple verifications, and more!



Conclusions

- Verification of AI has tons of security case studies to investigate
- Some upcoming research work from the AISEC team:
  1. Create a detailed mathematical representation of different weather events
  2. Formally Verified ML based Network Intrusion Detection
  3. Continue Down NLP path to include Dialogues
     (e.g. Q. Are you a robot? A. No Q2. Are you sure?)
  4. Formal verification of Soundwaves (e.g. Dolphin Attacks)

Thats all folks!