Bad Characters: Imperceptible NLP Attacks

Nicholas Boucher, Ilia Shumailov, Ross Anderson, Nicolas Papernot
Machine Learning

- Machine learning is everywhere
- We operate based on data, not formal rules
- There’s a lot of non-determinism
- It is suddenly hard to define Security

https://xkcd.com/1838/
Computer Security in context of Machine Learning

- Adversarial examples exist for all models
- A large taxonomy of attackers
- Attacks are scalable because of transferability
Machine Learning in context of Computer Security

- As secure as the weakest component
- Clear threat model
- Safety and Security policies and cases
- Existence of trusted components
- Well defined environment
- ML is a part of a larger pipeline!
ML Pipeline and attacks

Data

Preprocessing

Output

Model

Explanation

Visualisation
ML Pipeline and attacks

- Data Reordering and Reshuffling
- Poisoning Backdoors
ML Pipeline and attacks

- Downsampling attacks (Not sure if compression is covered)
- Anything else?
- Can we do something specific with NLP here? Text full of stop words?
ML Pipeline and attacks

- Adversarial examples
- Bit flips
- Anything else?
ML Pipeline and attacks

- Not sure anything exists here, overflows of sorts?
ML Pipeline and attacks

Attacks on SHAP and LIME from Lakkaraju and Dimanov

Is there anything more we can say?
ML Pipeline and attacks

- Bad Characters!
ML Pipeline and attacks

- Bad Characters
- Sponge examples
- Any other whole-pipeline attacks?
Bad Characters: Homoglyphs

A A

are these the same?
Bad Characters: GTranslate

Avocado
Avocado

a

Translate from: Bulgarian

Avocat
vocado

un
une
Bad Characters: Reorderings

1234
2341

are these the same?
Bad Characters: GTranslate

Send money to account 1234

Envoyer de l'argent sur le compte 4321
And there are many more…

What is the main lesson?
Main lesson

Visualisation and preprocessing pipelines need to be the same
Some quantitative results

- Google services break
- Microsoft services break
- Toxic detectors break
- Search engines break
- Spam detectors break

How come they’re still not fixed?!?
https://www.cl.cam.ac.uk/~is410/Papers/badchars_draft.pdf

https://arxiv.org/abs/2106.09898

*PS* try copying the name of the paper from pdfs and googling it