Sponge Examples: Energy-Latency Attacks on Neural Networks

Ilia Shumailov*^, Yiren Zhao*, Daniel Bates*, Nicolas Papernot^, Robert Mullins*, Ross Anderson*

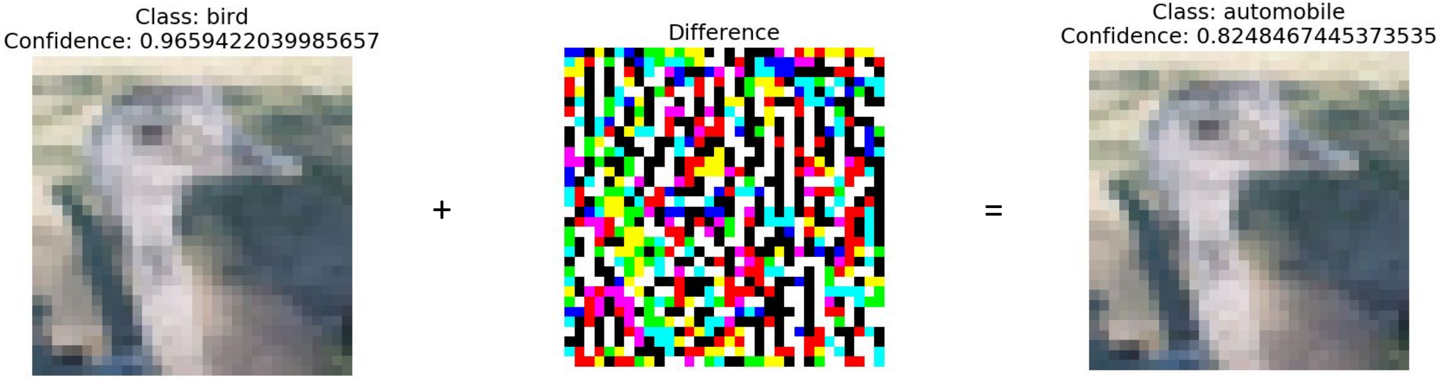
* University of Cambridge
^ University of Toronto, Vector Institute

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

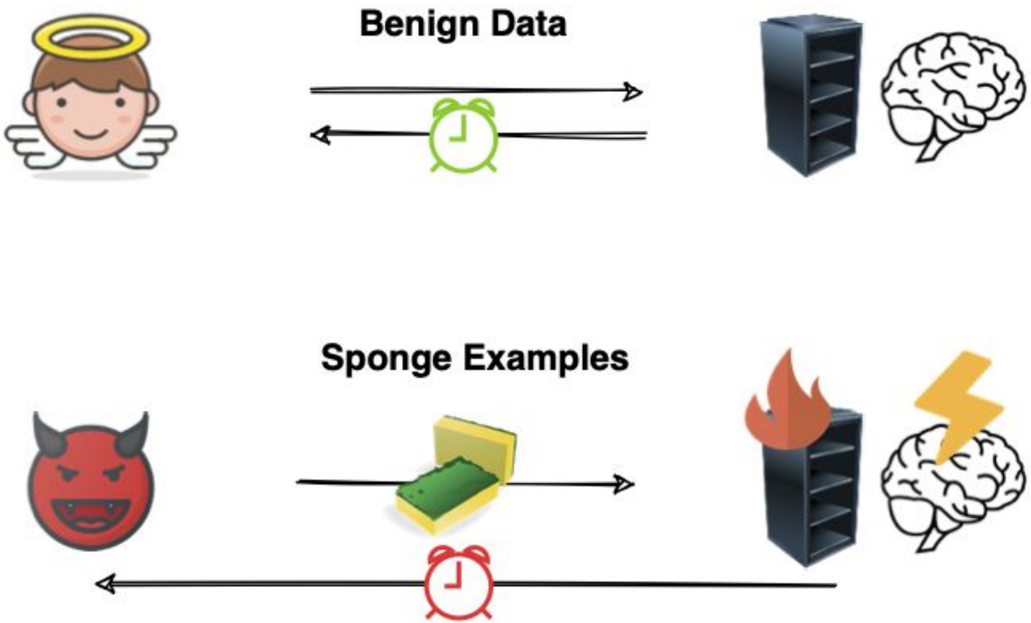


- Adversarial examples exist for all models
- A large taxonomy of confidentiality and integrity attackers
- What about availability?

Availability

Ensuring **timely** and **reliable** access to and use of information.
(NIST Special Publication 800-12)

Availability



Increased latency

Over-heating and over-consumption of energy

Energy Gap

The amount of energy consumed by one inference pass (i.e. a forward pass in a neural network) depends primarily on:

- The overall **number of arithmetic operations** required to process the inputs;
- The **number of memory accesses** e.g. to the GPU DRAM.

Computation Dimensions

Modern networks have a **computational dimension**

- A large number of NLP models are **auto-regressive** e.g. RNNs and GPT2
- **Adaptive** input **dimensions** to help performance e.g. GPT2 uses Byte Pair Encoding
- ML components are **a part of loop**

see more in the paper ...

Computation Dimensions for GPT2

Auto-regressiveness adds an unbounded loop

Algorithm 1: Translation Transformer NLP pipeline

Result: y

$\downarrow O(l_{\text{tin}})$

x_{tin} = Tokenize(x);

$y_{\text{touts}} = \emptyset$;

$\downarrow O(l_{\text{ein}})$

x_{ein} = Encode (x_{tin});

$\downarrow O(l_{\text{tin}} \times l_{\text{ein}} \times l_{\text{tout}} \times l_{\text{eout}})$

while y_{tout} *has no end of sentence token* **do**

    $\downarrow O(l_{\text{eout}})$

    y_{eout} = Encode (y_{tout});

    $\downarrow O(l_{\text{ein}} \times l_{\text{eout}})$

    y_{eout} = model.Inference(x_{ein}, y_{eout}, y_{touts});

    $\downarrow O(l_{\text{eout}})$;

    y_{tout} = Decode(y_{eout});

    y_{touts}.add(y_{tout});

end

$\downarrow O(l_{\text{tout}})$;

y = Detokenize(y_{touts})

Computation Dimensions for GPT2

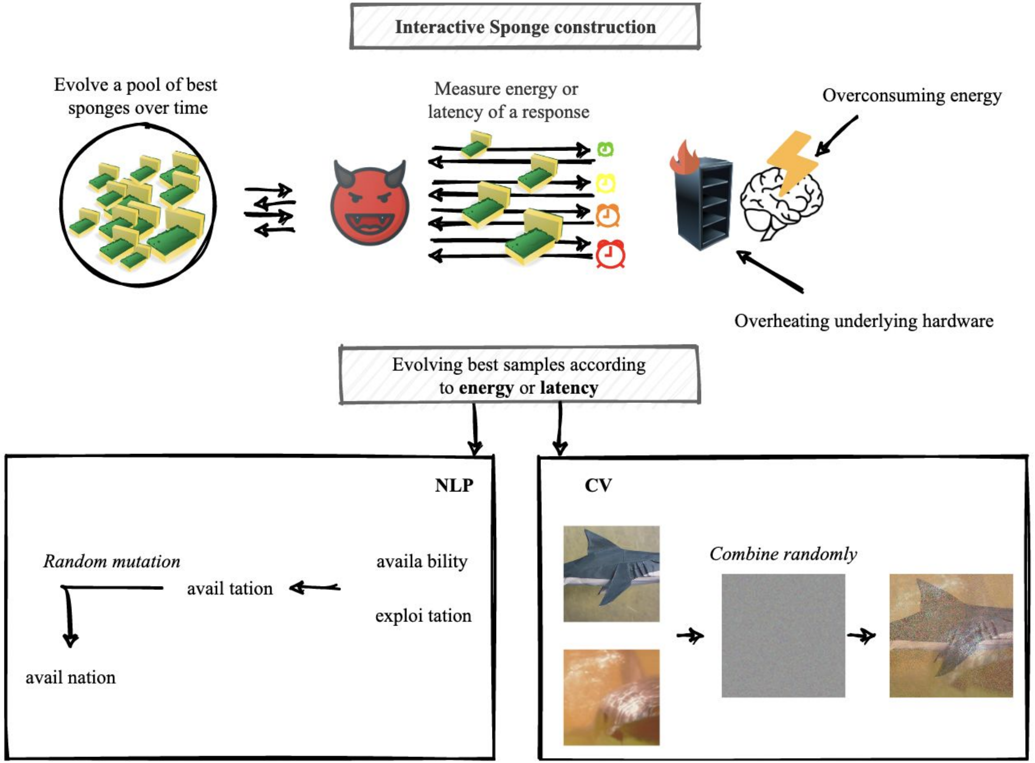
Encoding adds **variable** I/O representation

Benign with 4 tokens for input of size 16:
Athazagoraphobia => ath, az, agor, aphobia

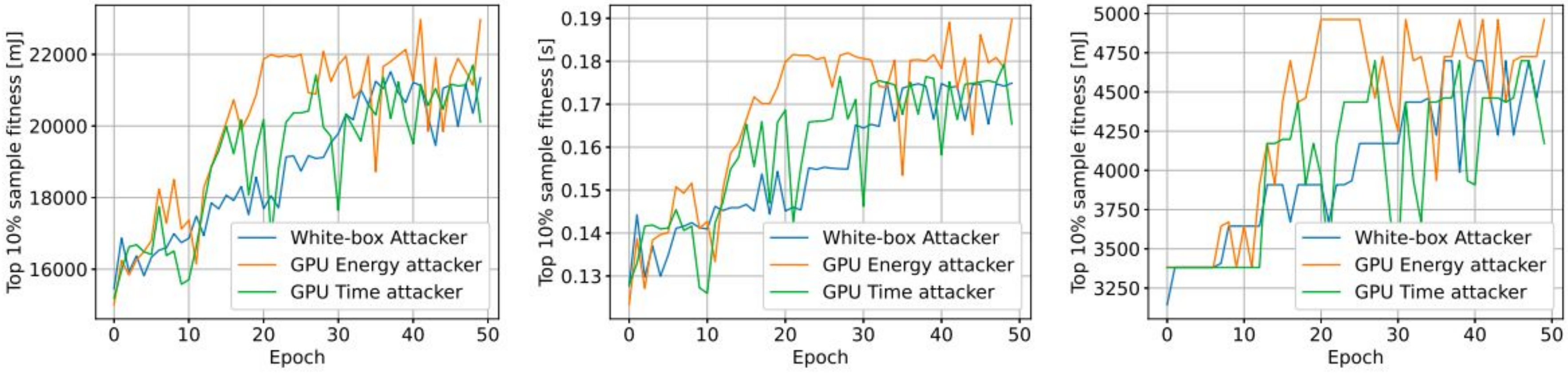
1 error with 7 tokens for input of size 16:
Athazagoraphpbia => ath, az, agor, aph, p, bi, a

Malicious with 16 tokens for input of size 16:
A/h/z/g/r/p/p/i/ => A, /, h, /, z, /, g, /, r, /, p, /, p, /, i, /

Multiple ways to search for Sponge examples



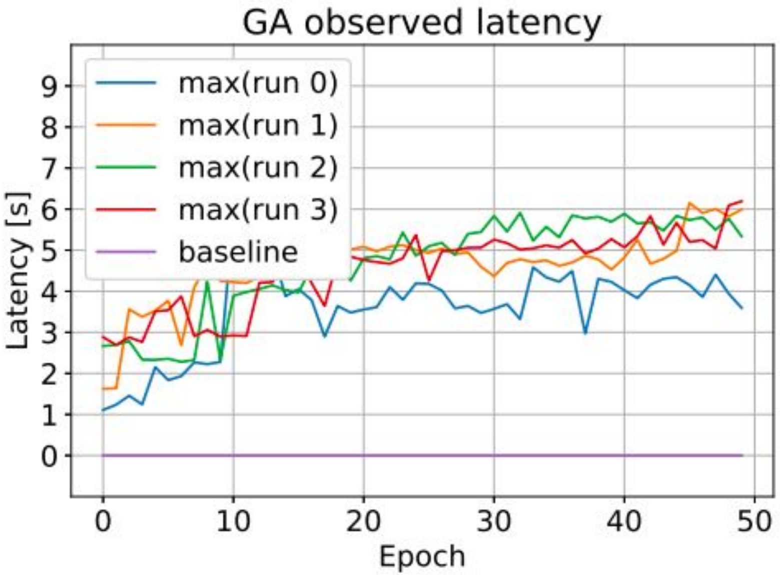
Interactive Black-box attack performance against WMT16 En→Fr



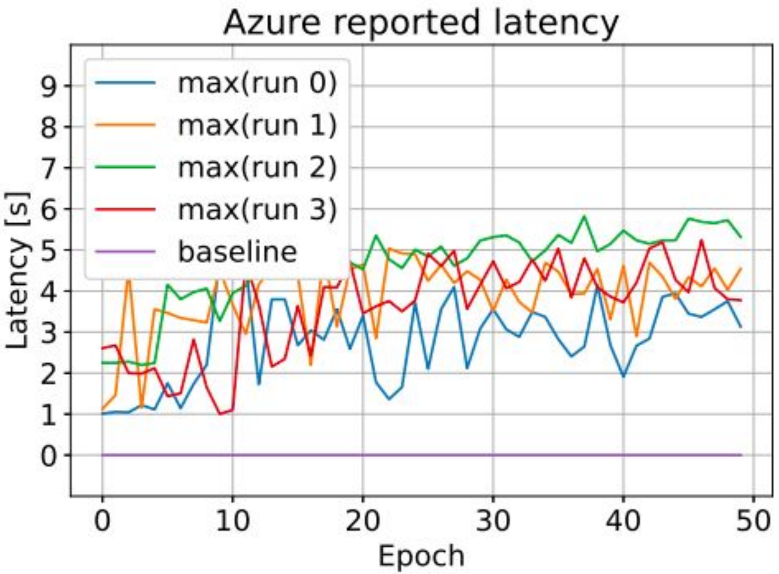
(a) GPU Energy (b) GPU Time (c) ASIC Energy

Attack works equally as well optimising number of ops, energy and latency.

Microsoft Azure



(a) Requesting server measured



(b) Azure reported

Baseline is at 1ms. Attack performs consistently with multiple restarts.

Conclusions

- It is possible to attack model **availability** in both White and Black-box settings

- Attack can target **hardware optimisations**
    - For some CV tasks we fully negated benefits from acceleration
- Attacks can target **algorithmic complexity**
    - For some NLP tasks we managed to get up to **x30** energy consumption and **x27** time

Conclusions

- Average case is very **different** from **worst case** scenario
- Pipeline **complexity matters**
- **Impact** of ML **on climate change** might have been **underestimated**
- It is **not clear how to defend** systems against Sponge examples
- **Real-time systems** with ML components **should model availability** adversary