MITIGATING DENIAL OF SERVICE VULNERABILITIES IN MACHINE LEARNING ALGORITHMS

Problem

Machine Learning (ML) solutions have become part of our lives and are typically able to provide fast and accurate results to everyday problems. They are, however, not infallible and are known to be particularly fragile against a group of attacks inherent to ML, collectively described as Adversarial Machine Learning (AML). There are many types of AML attacks, some of which belong to a class akin to Denial of Service (DoS) attacks known from cyber security. Sometimes referred to as “sponge” attacks in ML literature, these attacks seek to degrade performance by forcing ML solutions to use increased resources to process inputs and generate outputs by exploiting weaknesses intrinsic to the ML algorithms deployed. While their outputs may be correct, the loss of speed and increased calculation cost may render an ML solution unsuitable.

Need and relevance to Defence

Defence will increasingly rely on ML solutions to maintain its competitive edge, and requires tools to verify the suitability, reliability, and safety of these solutions. This research addresses one aspect of this, providing an understanding of, and tools for testing and evaluating weaknesses against, denial of service type attacks on machine learning algorithms, as well as proposing techniques to make ML systems more resilient to these attacks.

Research question

How do we test, evaluate and possibly certify a machine learning algorithm’s resiliency against attacks that aim to slow down its performance? How do we identify what attacks can slow down a deployed ML solution? How do these attacks work/manifest themselves in practice? How do we determine if a solution has been trained in such a way that it is vulnerable to certain denial of service attacks (for example, by specific triggers)? Are there any approaches that may make ML algorithms more resilient to DoS attacks aiming to exploit their inherent weaknesses?

Expected outcomes

The project should answer the research questions and deliver the methodology and prototypes to test for, and identify inputs to, an ML system that can cause degradation of its performance.

The outcomes from this project will serve as a basis for further development of tools for Defence Test and Evaluation to determine the suitability of ML systems procured or developed for use in Defence.

Methodology/approach

To deliver the expected outcomes, the project may require multiple approaches. This includes

- examination of prior art,
- what can cause performance degradation of machine learning algorithms (e.g. tailored training data, specific runtime queries, the choice of ML architecture used etc.),
- how DoS attacks can be triggered against an algorithm (e.g. examining algorithmic behaviour, its structure, implementation etc.), and
- what mitigation strategies may be effective against such attacks.

Research may be focused on, though not limited to, transformer-based models in natural language based applications.

Some hypothetical solutions may involve

- input validation, i.e., implementing mechanisms for filtering out malicious resource-intensive inputs,
- rate limiting, i.e., limiting the number of requests an algorithm can process within a given timeframe and
- resource allocation, i.e., monitoring and efficiently allocating resources.

We anticipate these to involve mostly theoretical research but also hands-on experiments to verify hypotheses and develop techniques, such as methods, scripts and algorithms, to transition into prototype tools to form the basis for further development.