Proposta de análise de desempenho do uso de mecanismos de segurança em sistemas inteligentes usando aprendizado de máquina e IA gerativa adversariais do tipo LLM

Milton P. Pagliuso Neto[1], **Charles C. Miers**[1]

[1]Programa de Pós-Graduação em Computação Aplicada(PPGCAP)
Universidade do Estado de Santa Catarina (UDESC)

milton.neto@edu.udesc.br,

charles.miers@udesc.br

Resumo. *A integração de métodos seguros tornou-se crucial para garantir a segurança e confiabilidade de sistemas de aprendizado de máquina. Este artigo propõe uma análise sobre o impacto da adoção de medidas de segurança no desempenho de LLMs, visando compreender o impacto das medidas de segurança na eficiência e escalabilidade desses sistemas.*

1. Introdução

A cibersegurança é uma área multidisciplinar que tem como objetivo a criação, operação, análise e teste de sistemas computacionais seguros. Uma sub-área da cibersegurança é a área denominada *Adversarial Machine Learning* (AML) que tem como objetivo a proteção de sistemas inteligentes baseados em aprendizado de máquina [Vassilev et al. 2024]. À medida que a adoção de tecnologias emergentes, como a inteligência artificial (IA), se torna mais comum, os esforços de pesquisa em cibersegurança ganham maior importância, dados as novas ameaças e vulnerabilidades. A prevenção da exploração maliciosa de sistemas gerativos baseados em *Large Language Models* (LLMs) se torna uma necessidade como uma solução de segurança, além disso, também se torna necessidade entender o impacto em desempenho que esta solução implica no funcionamento de um modelo de aprendizado de máquina.

O objetivo deste artigo é apresentar uma proposta de análise de desempenho utilizando um *Large Language Model* (LLM) com um módulo de segurança como objeto de estudo, para a compreensão do impacto que a adesão de um método seguro tem em um sistema inteligente baseado em aprendizado de máquina.

2. Fundamentação

Os LLMs são sistemas inteligentes baseados em inteligência artificial que conseguem processar e produzir respostas textuais com uma comunicação coerente, além de generalizar diferentes tarefas [Naveed et al. 2023], e desta forma, mudaram como é visto o que pode ser realizado através de um agente baseado em diálogo. O sucesso destes sistemas também está enraizado em sua escalabilidade, podendo ser expandidos com mais dados ou parâmetros, sem a necessidade de alterar os algoritmos e arquiteturas subjacentes, mas estes modelos utilizam uma quantidade expressiva de recursos para treinamento e implantação, como processadores, GPUs, memória, energia e banda de rede [Xu et al. 2024]. Os mecanismos de cibersegurança em modelos de aprendizado de máquina são

explorados há tempos, mas a proteção dos algoritmos, práticas e ferramentas contra adversários usados para garantir a integridade, confidencialidade e disponibilidade desses sistemas são pouco explorados. O conjunto de ameaças existentes no contexto da área de AML (e.g., envenenamento de dados de treinamento e violação de privacidade através da extração de dados do usuário) chamou a atenção de instituições e órgãos de segurança ao redor do mundo como o NIST, ENISA e MITRE. O uso malicioso de um sistema inteligente pode ter consequências como um consumo elevado de recursos computacionais a ponto de comprometer a disponibilidade das instâncias deste serviço. Logo, é importante poder unicamente identificar as instâncias, os usuários interagindo com estas instâncias e a permissão que diferentes usuários tem ao acessar recursos.

3. Proposta

O crescimento escalável de LLM em tamanho e complexidade resulta em um crescimento no conjunto de recursos necessários para poder atender a demanda, criando um desafio para o desenvolvimento e manutenção de sistemas inteligentes [Xu et al. 2024]. A eficiência destes modelos é crucial para sua utilização, dado a escala da infraestrutura necessária para a operação de diferentes instâncias, cria-se a necessidade de entender o impacto dos mecanismos de segurança. Como proposta, um *benchmark* do consumo computacional da LLM, com captura de informações como: (i) consumo de memória; (ii) consumo de processador; (iii) consumo de GPU; e (iv) tráfego de rede.

Para o plano de testes, os seguintes cenários são definidos: (i) Cenário padrão: o consumo de recursos da operação do modelo de linguagem como um oráculo; e Cenário seguro: o consumo de recursos da operação do modelo de linguagem, incluso de um módulo seguro. Os cenários tem como objetivo permitir a compreensão do impacto que um módulo de segurança teria no desempenho do modelo de linguagem ao prevenir a exploração maliciosa de sistemas inteligentes de maneira abrangente.

4. Considerações

A área de pesquisa sobre aprendizado de máquina tem um esforço de pesquisa para encontrar modelos de linguagens mais eficientes, devido a necessidade crescente de requerimentos de infraestrutura a cada avanço tecnológico atingido nas arquiteturas. O desenvolvimento de uma análise de desempenho realista complementa este esforço, fornecendo informações sobre o impacto das medidas de segurança em sistemas baseados em aprendizado de máquina.

Agradecimentos: Apoiado pelo LARC/USP, LabP2D/UDESC, FAPESP e FAPESC.

Referências

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Vassilev, A., Oprea, A., Fordyce, A., and Andersen, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations.

Xu, M., Yin, W., Cai, D., Yi, R., Xu, D., Wang, Q., Wu, B., Zhao, Y., Yang, C., Wang, S., et al. (2024). A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*.